2005

A Computational Model of Auditory Feature Extraction and Sound Classification

Coath, Martin

http://hdl.handle.net/10026.1/1822

http://dx.doi.org/10.24382/1319

University of Plymouth

A Computational Model of Auditory Feature Extraction and Sound Classification.

Martin Coath

Centre for Theoretical and Computational Neuroscience

University of Plymouth

A thesis submitted to the University of Plymouth in partial fulfilment of the requirements for the degree of

Ph.D.

October 2005

A Computational Model of Auditory Feature Extraction and Sound Classification.

Martin Coath

This thesis introduces a computer model that incorporates responses similar to those found in the cochlea, in sub-cortical auditory processing, and in auditory cortex. The principle aim of this work is to show that this can form the basis for a biologically plausible mechanism of auditory stimulus classification. We will show that this classification is robust to stimulus variation and time compression. In addition, the response of the system is shown to support multiple, concurrent, behaviourally relevant classifications of natural stimuli (speech).

The model incorporates transient enhancement, an ensemble of spectro - temporal filters, and a simple measure analogous to the idea of visual salience to produce a quasi-static description of the stimulus suitable either for classification with an analogue artificial neural network or, using appropriate rate coding, a classifier based on artificial spiking neurons. We also show that the specto-temporal ensemble can be derived from a limited class of 'formative' stimuli, consistent with a developmental interpretation of ensemble formation. In addition, ensembles chosen on information theoretic grounds consist of filters with relatively simple geometries, which is consistent with reports of responses in mammalian thalamus and auditory cortex.

A powerful feature of this approach is that the ensemble response, from which salient auditory events are identified, amounts to stimulus-ensemble driven method of segmentation which respects the envelope of the stimulus, and leads to a quasi-static representation of auditory events which is suitable for spike rate coding.

We also present evidence that the encoded auditory events may form the basis of a representation-of-similarity, or second order isomorphism, which implies a representational space that respects similarity relationships between stimuli including novel stimuli.

Contents

List of Figures

Acknowledgements

All of my colleagues at the Centre for Theoretical and Computational Neuroscience, University of Plymouth, have, without exception, been supportive and helpful. None more so than my supervisor Dr. Susan Denham whose energy and enthusiasm, quite apart from her knowledge of the field, have underpinned every aspect of my work. In addition Prof. Roman Borisyuk and Dr. Thomas Wennekers were generous with their time and were not impatient with my mathematical shortcomings.

Outside the department, other academics have helped with discussions, data, or advice, or merely by expressing an interest. Particular mention must be made of Dr. Jan Schnupp (Oxford) and Dr. Jennifer Linden (UCL) for allowing me access to their hard won spike data.

In addition I would like to thank first, Dr. Raymond Flood, Fellow of Kellogg College, Oxford who said to me, on the subject of doing a Ph.D. (in paraphrase) 'if not now, then when?', and second, Dr. C. D. Coath who, among other things, made me aware of LaTeX.

Absent from the academic fray, but central to the effort that produced this thesis is my partner Milica who keeps me on the straight and narrow, both in terms of sanity and typography.

Author's declaration.

At no time during the registration for the degree of Doctor of Philosophy has the author been registered for any other University award. This study was financed by the European Union open FET IST-2001-33099.

Relevant scientific seminars and conferences were regularly attended at which work was often presented. Three papers have been accepted for publication in refereed journals.

Publications :

Coath, M. & Denham, S.L. *Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience.* Biol. Cybernetics, **2005 ,93, pp 22-30**

Coath, M.; Brader, J.M.; Fusi, S. & Denham, S.L. *Multiple views of the response of an ensemble of spectro-temporal features support concurrent classification of utterence, prosody, sex, and speaker identity.* Network: Neural computation, **2005, 16(2/3) pp 285-300**

Denham, S.L. & Coath, M. *A model based upon response fields derived during early experience can account for the interference effects of synthetically degraded speech signals.* Proc. of ISCA workshop on plasticity in speech perception. **2005**

Coath, M. & Denham, S.L. *The role of onsets in auditory processing.*. Biosystems - accepted for publication **2006**

Posters and conference presentations :

2005: British Society of Audiology, Short papers meeting on experimental studies of hearing and deafness, Cardiff. Poster presentation: *Auditory transient responses and the efficient coding of natural sounds.*

Neural Coding 2005, Marburg. Oral presentation: *The 'why' question of auditory processing.*

British Neuroscience Association, Annual Conference, Brighton. Poster presentation: *Multiple views of the response of an ensemble of spectro-temporal features supports concurrent classification.*

Attend to Learn and Learn to Attend, European team meeting, Plymouth. Oral presentation: *Concerning the possible whitening of auditory stimuli by peripheral processing models.*

2004: Gordon Research Council, Conference on Sensory coding and the natural environment, Queen's College, Oxford. Poster presentation.

British Society of Audiology, Short papers meeting on experimental studies of hearing and deafness, UCL, London. Poster presentation: *Robust sound classification using response fields derived from stimuli during early experience.*

Attend to Learn and Learn to Attend, European team meeting, Rome.

Workshops in Auditory scene analysis and speech perception by human and machine, Hanse-wissenschaftskolleg, Delmenhorst. Poster presentation: *As BSA above.*

2003: British Society of Audiology, Short papers meeting on experimental studies of hearing and deafness, Nottingham. Poster presentation: *Re-synthesis of sounds from onset information only.*

Attend to Learn and Learn to Attend, European team meeting, Barcelona. Oral presentation: *Properties of spectral distributions as possible features for characterising sounds.*

CNS 2003, Alicante.

Word count for the main body of this thesis: 31,230

Signed: ________________

Date: 27- MARCH - 2006

Chapter 1

Introduction and overview.

Objectives and motivation for the work.

The motivation for the work contained in this thesis came from the desire to build an artificial but biologically plausible auditory perceptual system. This system would operate on a restricted but rich set of natural auditory stimuli and would distinguish stimulus classes in a background of noise or other overlapping stimuli. Given that the brain uses networks of spiking neurons it is clear that whatever method is adopted has to be suitable for implementation using a spike based encoding of the signal. Although it is not clear that spike *rate* encoding is the only, or even most efficient, method available to neural systems, neural responses are in the overwhelming majority of cases, reported as changes in the spike rate. This presents difficulties for the generic encoding of features of auditory stimuli in that they are not generally stationary in time. Automatic speech recognition systems (with no pretention to biological plausibility) avoid this problem by assuming that the signal can be represented as a sequence of static states in windows of typically $10 - 20ms$ and this approach has proved very successful. This approach is not, however, 'stimulus driven' in that it makes no attempt to identify salient 'landmarks' or auditory 'events' which may form the basis for a quasi-static representation of the stimulus that is suitable for rate coding. We propose to develop and implement a model of peripheral auditory processing based on a model of the cochlea but which emphasizes the sensitivity to energy change found in biological systems. We hope to use this stimulus representation as the basis for our model of sound classification. It is also hoped that by investigating the properties of this representation some light might be thrown on why transients, that is periods of short term energy change that might be called *onsets* and *offsets*, are increasingly emphasized in the ascending auditory pathway towards cortex.

As the sub-cortical auditory system emphasizes these transient responses at the expense of continuous, or tonic, firing in response to stimuli, then we hypothesize that it should be possible to characterize the response of at least some cortical neurons to patterns of onsets and offsets. We hope to show that this is indeed possible by using, for the first time, the established method of *reverse correlation* to estimate the spectro-temporal preferences of cortical neurons in terms of the transient sensitive representation.

Given that patterns of transients could form the basis of auditory feature extraction, we hope to investigate whether useful patterns could be derived from the stimuli themselves. Based on work using a measure of similarity between 'fragments' of visual images and a library of images from different classes, we hope to adapt and extend this idea to measures of time dependent similarity using ensembles of fragments in the auditory domain.

The next aim is to investigate if the response of these spectro-temporal feature extractors could form the basis for deriving a readout of discrete, salient auditory 'events' as the basis of the sound classification. The goal of this stimulus driven, feature dependent segmentation is a representation which respects the time course of the stimulus (i.e. the prosody of speech stimuli), results in a quasi-static representation suitable for rate coding, and achieves a large reduction in computational overhead. We hope to test the resulting model by comparing its performance with human psychophysics and quantifying its performance in terms of the mutual information between the input and the output classes.

The proposal is to design each of the stages outlined above to be consistent with considerations of biological and developmental plausibility. In this way it is hoped that the resulting representation will be, in common with biological systems, flexible enough to classify a wide range of stimuli, rich enough to support the classification of novel stimulus classes, and robust to moderate manipulations and distortions of the stimuli.

A separate but related objective of the work is that the model should be suitable for integration within a larger model, being developed by a number of partner projects, of learning and attention (ALAVLSI, 2001). Other collaborative projects are developing neuromorphic analogue VLSI circuits consisting of artificial spiking neurons with the eventual aim that all or part of the model should be implemented in hardware. To this end we hope to show that our model of feature extraction provides responses that are suitable for rate coding as the input to such a system.

Peripheral processing.

The nature of stimulus processing in the peripheral auditory centres of the brain is still only partly understood. In this work emphasis is placed on only two well documented aspects. First, many results show that responses have well defined characteristic frequencies and that this 'tonotopic' arrangement is preserved throughout the ascending pathway. This implies that much of the processing is done *within* a frequency channel. Second, there is well documented evidence for sensitivity to (and enhancement of) envelope transients, i.e. onsets and offsets (e.g. Heil, 2001). The principle requirements of the model of peripheral processing adopted are therefore that it should not be inconsistent with these findings and that it should be relatively simple, with a view to its eventual hardware implementation. At the same time it must also be consistent with the other goals of the project such as information preservation, and robustness to noise. In addition it would be desirable for it to exhibit some degree of level independence.

In Chapter 2 the evidence for the importance of envelope transients is reviewed and a novel model of transient enhancement based on the distribution of energy within a time window is introduced. This new representation is shown to have many attractive features and to be consistent with neurophysiological and psychophysical results. In Chapter 3 it is further shown that this representation can be used to characterize response data recorded in auditory cortex.

Spectro-temporal responses.

The principle method used to characterize responses in higher auditory regions, such as thalamus and cortex, is to determine the spectro-temporal preferences of neurons. Described in this way each neuron can be thought of as a filter. An alternative view, however, is that each neuron has a 'preferred stimulus' in response to which it fires vigorously. This preferred pattern can be thought of as a 'feature' of the stimulus and the output of the neuron as signalling the presence or absence of this feature. Both of these ideas are summarized in the concept of a *spectro-temporal response field* (STRF) and this idea forms the basis for the low-dimensional representation implemented in the second stage of the model.

Although the STRF is a widely discussed concept it is not known how these response fields develop. However it is known that cortical organization is much disrupted by aberrant auditory stimuli during early post-natal development and so it is the assumption of this project that STRF formation is, at least in part, driven by exposure to formative stimuli and their geometry is determined by the necessity of distinguishing between the various stimuli in the early auditory environment.

Results detailed in Chapter 4 indicate that useful filters, or features, can be derived from a very limited set of formative stimuli. Their usefulness can be described in terms of the entropy of their responses to the formative stimulus set. These results show that there is a basis for preferring one feature over another. We also examine how ensembles of features may be formed by a selection process based on information theoretic principles.

From stimulus representation to response representation.

The mapping of stimulus to response requires more than just calculating the responses of an ensemble of roughly tuned spectro-temporal features. The outputs of each of these filters is merely a linear transformation of the output from the peripheral model. However, if the features in the ensemble are chosen on the basis of their informativeness with respect to a set of formative stimuli then their responses to new stimuli of a similar type might be expected to be maximal at times that coincide with salient features.

In Chapter 4 we discuss how appropriate ensembles of features might be derived from formative stimuli and in Chapter 5 we go on to show that the response of such an ensemble can be used as a basis for identifying salient events in a stimulus driven manner that are suitable for rate coding. We also show that the pattern of ensemble response during these events preserves information about the stimulus class.

Characterizing the performance.

To gauge the representational quality of the feature space, it was proposed to compute the mutual information between the input and outputs of the system. Mutual information is an information theoretic property related to the correlation between distributions of probabilities and can only be calculated on data that have nominal classes. The input stimuli chosen for the experiments are labelled so the input class for each is known. The representations derived from the array of spectro-temporal features, however, must be assigned to a class before the mutual information can be calculated. One approach to this is to build a classifier based on an artificial neural network which would allow the performance of the model to be quantified in information theoretic terms. Results are also compared with human psychophysics. Also in Chapter 5 it is shown there is a clear systematic difference between the performance of fragments of different temporal extent which suggests a relationship between the time constants for feature extraction and the intelligibility of speech measured in human psychophysics.

The use of mutual information to judge the success of these feature spaces is discussed in Chapter 5. In line with the biologically plausible approach of the project it was a requirement that the representation of auditory stimuli should be suitable for spike rate coding with the intention of using a classifier based on a network of artificial spiking neurons. Results using simulated networks of artificial spiking neurons are contained in Chapter 6.

Concurrent classifications.

A key feature of biological perceptual systems is that the judgements they support are task dependent. Of the many classifications of a given sound that are possible (often called a 'what' judgement) priority is given to the one that is most task relevant by mechanisms of attention. Evidence is emerging that different 'what' judgements are made concurrently in spatially distinct areas of the brain and any plausible model of sensory processing must support this type of multiple concurrent processing.

Experimental results reported in Chapter 6 show that multiple concurrent classifications are supported by the model detailed in this work.

Stimulus set.

The stated aim of using a stimulus set that is *'restricted but rich'* is somewhat constrained by the availability of example corpora so the decision was made to use widely available speech corpora of numerals and letter names. The choice of speech corpora invites comparisons with speech recognition systems because the 'classes' into which these stimuli apparently fall are the familiar English word, or letter, names. However, it should be stated that the goal of the work is not speech recognition *per se* but a biologically informed acoustic processing model suitable for a flexible and robust perceptual system. The speech stimuli also have the advantage that they either meet, or can be manipulated to meet, another experimental requirement, i.e. each stimulus can fall in to more than one class simultaneously. For example, in the corpus of spoken digits half are spoken by male, and half by female subjects. As a result each stimulus can be labelled by its digit name (*one*, *two*, and so on) or by the sex of the speaker. Both are percepts associated with the stimulus and an important test of any model perceptual system, as detailed above, is that is should preserve features that allow distinctions between multiple classifications concurrently (see Chapter 6).

Summary.

The work herein represents a significant and novel departure from the recognized models of sound, particularly speech, classification. In the first instance the model is **developmental** in that it is based on limited numbers of formative stimuli which are used as the basis for a selection of features which are informative with respect to these stimuli. Second, it is **productive** in the sense that the same features are subsequently used to to generalize to new stimulus classes, including novel speakers, intonations and so on. Third, the model of peripheral auditory processing abandons the traditional view of a stimulus with a characteristic spectral and temporal envelope and substitutes **pattern of energy change** which is widely reported as being central to biological peripheral processing. This representation preserves many features of the temporal envelope but is to a large extent independent of the spectral envelope. Fourth, the salient parts of the stimulus are identified in a way which is dependent on both the **stimulus** and the **feature set**. The stimulus dependence ensures that the response of the model respects the rhythm and presentation rate of the stimulus and the dependence on the feature set has implications for cortico-fugal interactions with sub-cortical processing. Last, the responses derived in this way are tested for their ability to support a range of classifications, not just a mapping from the stimulus to a single distinct response, each of which, in the organism, would represent a **behaviourally relevant judgement**. This *must* be a feature of any model of auditory perception rather than simply a system of word recognition or speech transcription.

Chapter 2

Auditory transients.

2.1 Overview.

Principle aims. To develop a representation of sound stimuli based on a cochleagraphic representation but sensitive only to short term increases and decreases in energy within a cochlear channel. There are two important and novel features of this representation. First, the time scale over which these 'onsets' and 'offsets' are detected is frequency dependent. Second, the calculated response reflects the asymmetry of the energy distribution within a time window for each frquency channel. To this end we use a statistical property of the energy distribution over this time - the skewness or third central moment. We hope to examine the properties of this representation using a range of stimuli, and mixtures of stimuli.

Motivation. There is a great deal of evidence that the auditory pathway is arranged tonotopicaly and little evidence to support integration across frequency channels in sub-cortical areas. There is also a great deal of evidence that the auditory system is sensitive to relatively short term changes in energy, i.e. onsets and (to a lesser extent) offsets. In addition physiological measurements suggest that information in different parts of the tonotopicaly arranged auditory system is extracted on different, frequency dependent time scales. There is currently no consistent explanation as to why transient or phasic responses are increasingly emphasized over continuous or tonic responses as signals pass from the auditory periphery through the mid-brain to more central areas. It is possible that they confer some benefit related to the goal of efficiently coding natural stimuli. We hope to shed some light on this question.

Achievements. We show, in the results of psychophysical experiments in Appendix A, that stimuli re-synthesized from onsets alone preserve a great deal of information necessary for comprehension of speech, provided that the stimuli are re-synthesized in a way that allows for the onset to be located tonotopically as well as in time. We also show that our model responds differently to stimuli with a range of spectro-temporal modulation statistics. In mixtures of synthetic noise and speech, it preserves the onset patterns of speech when mixed with some types of noise with statistics that differ from natural stimuli. Comparison with data from human auditory brainstem responses shows that the level, and timing, of maximal activity predicted using this model broadly match physiological data. Finally, we show that the representation produces a de-correlation in both the spectral and temporal domain which amounts to a redundancy reduction or 'whitening' in the stimulus representation. These results support our hypothesis that the onset sensitivity observed in the auditory system is related to the goal of efficient coding of ethological stimuli and may aid figure-ground separation.

2.2 Temporal envelopes in the auditory system.

2.2.1 Onsets.

It is well documented that the auditory system is sensitive to the temporal structure of the amplitude envelope, particularly rising (*onset*) transients. This has been shown both in physiological and psychophysical measurements (eg Heil (1997b); Phillips *et al.* (2002)). This sensitivity increases as measurements are made at successively higher levels in the auditory pathway. Units that detect onsets are found throughout the auditory system; in VCN (Frisina *et al.*, 1985; Rhode & Greenberg, 1994), IC (Heil & Irvine, 1996; Langner & Schreiner, 1988) MGB (thalamus) (Rouiller & de Ribaupierre, 1982; Rouiller *et al.*, 1981), cortex (Eggermont, 2002).

Psychophysics has shown that the manipulation of the time varying amplitude envelope within frequency channels (the *temporal envelope*) affects speech intelligibility (Drullman, 1995; Drullman *et al.*, 1994a, b, 1996). In addition work has been carried out using speech re-synthesized by imposing the temporal envelope extracted from varying numbers of frequency bands on band limited noise (Fu *et al.*, 1998; Shannon *et al.*, 1995, 1998). This preserves the temporal structure of the within-channel information but, because of the limited number of noise bands used, severely disrupts or distorts the profile of energy relationships between frequency bands (the *spectral envelope*). These experiments have shown that speech is intelligible with well defined temporal envelopes in the virtual absence of information about the energy relationships between channels. This was found to be true for experiments carried out in Mandarin as well as English. A particularly interesting result from this work, however, was reported for experiments where the speech is re-synthesized by modulating band limited noise with the temporal envelope from a single, narrow, frequency band. Under these conditions, with *no* spectral information, Mandarin phonemes that exhibit a rising or falling pitch are more easily identified than tonally *'flat'* phonemes, and Mandarin was found to be more intelligible, at a sentence level, than English. This result strongly suggests that it is the more tonally varied stimuli that produce patterns of rising and falling amplitude transients within a frequency channel which aid recognition.

Original psychophysical experiments reported in Appendix A support the idea that, for speech stimuli at least, spectro-temporal patterns of energy change (particularly regions of rising energy or 'onsets') are important for perception. In this work we show that intelligibility is preserved in stimuli re-synthesized from onsets only, by placing a tone burst starting at the spectro-temporal point occupied by the beginning of the onset envelope. Intelligibility increases as the duration of the tone burst increases but the benefit of increasing duration diminishes quickly beyond $\approx 4 - 5$ times the period of the frequency. Note that this is close to the period over which the skewness is calculated in the proposed representation. This study is preliminary in nature and the results can be taken only as being indicative.

Bregman *et al.* (1994) have reported that transients are important for the separation of objects within an auditory scene. In these experiments clusters of four tones were presented with asynchronous onset and offset ramps of different durations. It was found that the ability to judge the order in which the tones were presented depended on the acceleration of the onset ramp chosen and a similar effect was reported for offset rates. One problem for models of onsets, or more generally transient sensitivity (as well as other problems in auditory processing such as pitch sensitivity) is that they all rely implicitly on some form of auditory delay to provide the time constant, or time constants, over which the change in energy is extracted. Evidence for such a mechanism has at least been shown in bats (Hattori & Suga, 1997). The dearth of other evidence has led Shamma (2001) to propose mechanisms for a range of auditory percepts that concentrate on spectral characteristics of the stimulus and template matching. The spectral profile of the stimulus is, of course, supported by the mechanical (cochlear) and neurophysiological (tonotopic) substrate, this approach has other advantages including close parallels with visual processing. However, as has been mentioned above, the weight of evidence suggests relative insensitivity to spectral profile; implying predominantly 'within channel' processing of sound in sub-cortical areas.

The use of onsets and offsets, particularly as a means of sound segmentation has been well documented by Smith (1995) using a convolution between the bandpass filtered sound and an asymmetric kernel. Using a similar onset sensitive convolution, Fishbach *et al.* (2001) have proposed a neural model for onset sensitivity. This uses the amplitude envelope of tone burst stimuli, a neural response model based on inner hair cell potential, followed by a delay layer and convolution with kernel derived from a first order derivative of a Gaussian. This model accurately reproduces a broad range of physiological and psychophysical data including first spike timings in primary auditory cortex (PAC).

2.2.2 Offsets.

Responses to falling amplitude transients (offsets) are less often reported but well attested in the literature, e.g. He *et al.* (1997); Phillips *et al.* (2002); Van-Campen *et al.* (1997). Results from recordings of auditory brainstem responses

(ABR) (VanCampen *et al.*, 1997) show that offset responses are of comparable amplitude, and exhibit similar latencies to onset responses. This is somewhat at odds with the asymmetry of onsets and offsets measured in single unit recordings (Phillips *et al.*, 2002) and the origins of these responses at a cellular level are still obscure. However, onset responses have been shown to be locked to the maximum acceleration of pressure in electrophysiological recordings (Heil, 1997a) and this maximum acceleration is found at the very beginning of the stimulus for both sine-squared and linear onset ramps. Given this and the similar latencies measured for onsets and offsets in ABRs, it is plausible that the offset response is similarly tied to some aspect of the offset ramp. These results are somewhat complicated by the observation that offset latencies are sensitive to onset rise times, i.e. offset latencies are shorter when rise times are longer. Because the plateau duration was not manipulated in these experiments it was necessarily shorter for stimuli with long rise times and it has been suggested (VanCampen *et al.*, 1997) that refractory times for neurons recruited to fire at both onset and offset, reported for example in He (2002), may account for this.

2.2.3 Duration tuning.

An important related issue is the existence of duration tuned neurons in the inferior colliculus and cortex of a number of different species; frogs (Gooler & Feng, 1992), mice (Brand *et al.*, 2000), cats (He *et al.*, 1997), bats (Fuzessery & Hall, 1999) and Guinea pigs (He, 2002). Duration sensitive responses have also been recorded in visual cortex of cats (Duysens *et al.*, 1996). These responses are often characterized by spikes that occur shortly after the offset of the stimulus. Their presence across many classes of vertebrates and in both auditory and visual systems suggest that this is a general property and a valuable extra 'dimension' alongside frequency tuning. Although responses of this type have been found in bats in IC (Faure *et al.*, 2003) they are rarely reported in centres below the thalamus which has led to suggestions (Casseday *et al.*, 2000) that this type of response is generated in the auditory midbrain and above by the integration of onset, offset, and tonic patterns for earlier auditory centres. One model proposed by Ehrlich *et al.* (1997) involves onset inhibition and offset excitation.

It is in any case clear that peripheral auditory processing must preserve envelope timing information and this is clearly related to the preservation of onset and offset information.

2.3 A model of temporal envelope extraction.

2.3.1 Spectral decomposition

The first stage of the model approximates processing in the cochlea. Sounds are processed using a bank of 24 Gammatone filters (Slaney, 1994), with centre frequencies, ranging from 100 to $\approx 8000Hz$ arranged evenly on an ERB scale. The output in each frequency channel is low-pass filtered and half wave rectified.

The result of this pre-processing is a representation of the output of each of a set of filters corresponding to the activity in a set of tonotopic channels or bands (Slaney, 1993). This representation is referred to as the Simple Cochlear Model, or SCM in subsequent sections, see Figure 2.3. The output of each cochlear channel was re-sampled at this stage to $8Khz$ (using the MATLAB `resample` function) for efficiency of storage and to reduce computational expense in subsequent processing. It will on occasion be convenient to refer to this representation as a function of time (t) and channel (x) an in these cases the value of the SCM at (t, x) is referred to as $y_{scm}(t, x)$.

2.3.2 Transient extraction.

We introduce in this section a representation based on the cochlear model described above, calculated from the skewness of the activity in each cochlear channel over successive overlapping time dependent windows. During the course of this work it has become the practice to refer to this representation as "SKewness in Variable time" and in the interests of brevity instances of it are referred to as the **SKV** representation in subsequent sections. The skewness is a statistical measure which reflects the asymmetry in the 'tails' of a distribution and can take positive and negative values depending on whether the asymmetry favours the right or left of the distribution; i.e. rising or falling energy within a channel. Positive values for the SKV represent areas of short term rise in the output from a cochlear channel, or 'onsets', and negative values represent 'offsets'. The value of the SKV at time t in cochlear channel x is referred to as $y_{skv}(x, t)$ where x is the channel and t the time.

The SKV calculation. The activity $y_{scm}(t, x)$ of each cochlear channel is divided into overlapping temporal windows of duration twice the period of the centre frequency (CF) of the channel, but with a minimum window size of $2.5ms$ at high frequency. This in effect means that CFs of over $800Hz$ are treated as if they were $800Hz$. These parameters have been reported by Wiegrebe in the context of a paper dealing with the minimum time necessary for neural pitch extraction (Wiegrebe, 2001). We have used this as an indication of the relative time scales for processing each channel using the values for the centre frequency of each channel in each case.

In this description we will deal with only one cochlear channel as all channels are calculated independently - this simplifies the notation. The length of one window in the cochlear channel δt^w will thus be;

$$\delta t^w = \frac{2}{CF}$$

where CF is the centre frequency of the channel in question. The number of sample values in this window, s^w, which will depend on the sample rate r, in each window is;

$$s^w = \delta t^w.r$$

hence the mean value of the cochlear response w_n within the nth window is;

$$w_n = \frac{1}{s^w}\sum_{j=1}^{s^w} y_{scm}(j)$$

where $y_{scm}(j)$ is the jth value of the cochlear output inside the window.

The degree of overlap between windows for all experiments was set to 10% of the window duration and thus the spacing δt^o between leading edges of successive overlapping windows is;

$$\delta t^o = \frac{1.8}{CF}$$

and the time of the leading edge of the nth window is $n.\delta t^o$ which we will designate t_n. To calculate the value of the SKV at time t_n the means of the nth window and the three preceding windows are normalized so that they sum to unity, and then the formula for skewness applied;

$$y_{skv}(t_n) = \frac{1}{4}\sum_{k=0}^{3}(\frac{(w_{n-k} - \overline{w})^3}{\sigma^3}) \tag{2.1}$$

where 4 is the number of windows, w_{n-j} is the normalized mean of the $n - j$th window, $\overline{w}$ is the mean of the four values of w, and σ the variance. This process is illustrated in outline in Figure 2.1.

The skewness is a sensitive indicator of rising and falling energy and has a value near zero when the distribution is symmetrical about the centre. It is relatively insensitive to individual values within the distribution, most importantly this applies to the initial and final values which would distort measures based on rate of change. An example of the SKV representation is illustrated in Section 2.4.1. Note here that although it is likely that onsets and offsets are detected by separate and distinct pathways (He, 2001), it is often convenient to show both on the same diagram to give a synoptic view of the activity; onsets in red, offsets in blue (see Section 2.4.1). In order to simplify this plot, to store the SKV representations of sounds as matrices, and as a computational convenience for later work, the results from each channel were up-sampled to $8kHz$ using the MATLAB `resample` command.

2.3.3 Level independence, saturation, and adaptive rescaling.

Because the SKV representation depends on the *distribution* of energy within a time window independent of its magnitude, it will display a degree of inherent level independence in its response. In addition it is possible to introduce a sim-

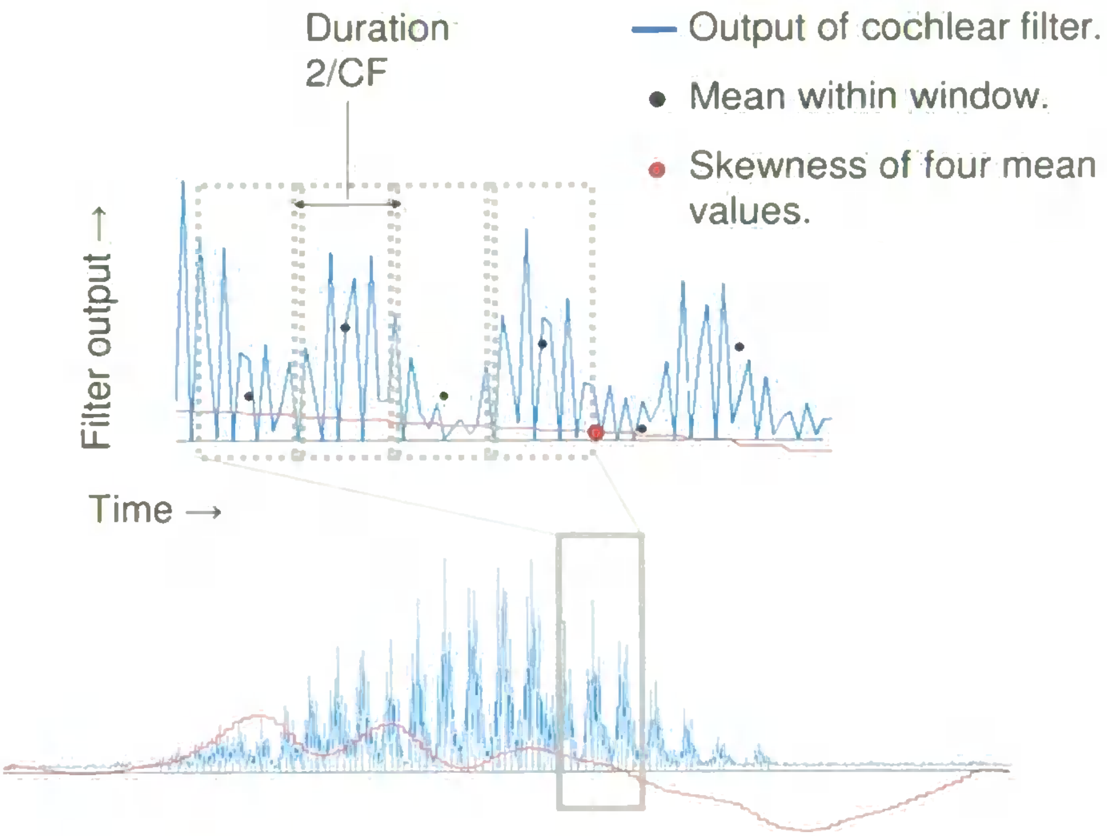


Figure 2.1: *Illustration of skewness calculation.*
In the upper part of the figure the blue line shows the output of one channel of the cochlear filter bank. Four overlapping grey windows are superimposed each of which is of duration $2/CF$, that is twice the period of the centre frequency of the filter. The mean amplitudes of these four successive windows are indicated by black dots inside red circles. The red dot indicates the skewnwess calculated from the normalized values of these four means. The red line joins successive values of the skewness calculated for each window from the mean within that window and the previous three means.
The lower part of the figure shows the time covered by the four windows in the upper part as a grey box and the cochlear output (Blue) and the skewness (Red) over a longer time period. See Section 2.3.2 for the details of the calculation.

ple model of *adaptive re-scaling* (Brenner *et al.*, 2000). This form of adaptive response effectively normalizes the variance by 'stretching' (or compressing) the gain for low intensity signals to match the distribution of intensities of stimuli. An alternative way of looking at this is that the response adapts to match the large dynamic range of the input to the limited dynamic range of neuronal responses. This strategy maximizes the entropy of the distribution of the output, this is explained further in Chapter 4. At higher levels the gain is reduced corresponding to a maximum firing rate or saturation.

Both of these ends, variable gain and saturation, are achieved by a sigmoid response function applied symmetrically to both onsets and offsets (Equation 2.2 and Figure 2.2).

$$y(t) = 0.5 - \left(\frac{1}{1 + \exp(a(t)x(t))}\right) \tag{2.2}$$

In this equation the general symbols $x(t)$ and $y(t)$ represent the time varying input and output. The single parameter $a(t)$ in Equation 2.2 represents adaptation in all the frequency channels of the representation. The value of the parameter a is a function of time and was derived from the output of the SCM. Equation Equation 2.3 indicates the derivation of a, it is inversely proportional to the maximum value in the SCM representation within a short time δt prior to the current time t with the constant of proportionality K. The value of K was found heuristically to be 2.

$$a(t) = \frac{K}{\max(\mathrm{SCM}_{t-\delta t}^{t})} \tag{2.3}$$

This function limits the range of values at the output of each frequency channel. We have chosen to limit these values to between −0.5 and +0.5. See results in Section 2.4.2 for details.

2.3.4 Autocorrelation and whitening.

In this, and subsequent chapters, we will include in our discussions the concept of a stimulus being *'white'* or sometimes *'whitened'*. What follows is not a definition of these terms, as this is outside the range of the current discussion, but a clarification of what is meant in the context of the current work.

White noise is called 'white' because it has equal power at all frequencies. For practical purposes it should be stated that this is only possible if 'all frequencies' is taken to mean all frequencies within a broad, but finite range. It is further necessary to be clear that this can only be true over a stimulus of infinite length as the instantaneous spectral profile is never flat. Can we described stimuli other than noise as white if they meet this simple requirement? This is clearly not enough, for example a tone that rises linearly from the minimum to the maximum frequency in the specified range is also white in this sense. This sort of stimulus

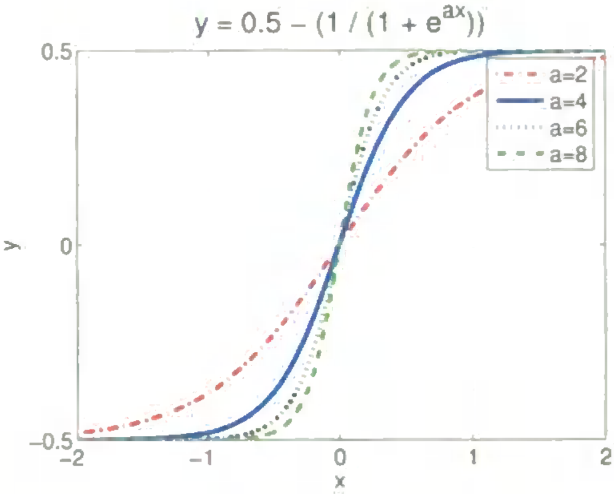


Figure 2.2: *Parametric curves used to implement a simple model of adaptive re-scaling and saturation in the SKV response. The curves illustrate Equation 2.2 Section 2.3.3 for various values of a.*

fails to meet the requirements for whiteness in the wider sense that to be white its spectro-temporal structure must be *uncorrelated*. This implies that the amount of energy present at a particular frequency at a particular time is in no way predictable from knowledge of the value preceding it, or in adjacent channels.

Autocorrelation. To expose the temporal correlations in a signal it is possible to calculate the similarity between that signal and the same signal advanced, or delayed by a range of discrete time steps. It is clear that the signal is most like itself at a time lag of zero. Peaks in similarity at non-zero time lags, indicate temporal structure or regularities in the signal. For a spectro-temporal description of a signal it is possible to use a similar method to expose the spectral structure as well. This type of analysis is called autocorrelation.

The autocorrelation of a matrix (or a stimulus representation such as the SCM or SKV in the form of a matrix) having two dimensions t and x is given by Equation 2.4.

$$C(i,j) = \sum_{\tau=0}^{(T-1)} \sum_{\chi=0}^{(X-1)} y(\tau,\chi).(y(\tau+i,\chi+j)) \tag{2.4}$$

The value of the autocorrelation (C) at the point (i,j) is the double sum of all the products of the value y at point (τ,χ) (where τ and χ represent indices in the dimensions t and x) with the corresponding point distant from it by i in one dimension and j in the other. The power spectral density is, by the Wiener-Kinchin theorem (Hartmann, 1998) the Fourier transform of this autocorrelation. For examples of this type of analysis see Section 2.5.1. For an idealized 'white' signal that has no temporal or spectral structure the autocorrelation is zero everywhere except at $\Delta X = \Delta Y = 0$ (i.e. it is a delta function) and the power spectrum has the same value for all frequencies, in other words it is *'white'*.

The eigenvalues of the covariance matrix. An alternative, but related, way to look at whiteness in a spectro-temporal representation is to construct the covariance matrix. For this purpose we will treat each row, or column of a representation such as the SCM or SKV as a vector $\mathbb{V}$ where $\mathbb{V}_i$ represents the i^{th} row or column. Each element of the covariance matrix $cov_{i,j}$ will then be;

$$cov_{i,j} = \langle \mathbb{V}_i \mathbb{V}_j \rangle - \langle \mathbb{V}_i \rangle \langle \mathbb{V}_j \rangle \tag{2.5}$$

If the representation contains no correlations then the covariance matrix is zero everywhere apart from the diagonal, that is where $i = j$ and the eigenvalues of this matrix will all be equal, and where correlations are present there will be a range of eigenvalues.

In order to quantify the variation we calculate the coefficient of variance C_v in the vector of eigenvalues $\mathbb{E}$ of the covariance matrix. This is defined as the ratio of the standard deviation $\sigma_{\mathbb{E}}$ and the mean $\overline{\mathbb{E}}$ of the elements of this vector as in Equation 2.6.

$$C_v = \frac{\sigma_{\mathbb{E}}}{\overline{\mathbb{E}}} \tag{2.6}$$

And since most of the variation is expressed in the lower order eigenvalues, we calculated the C_v from the first 30 eigenvalues for both the SCM and SKV representations. The percentage difference between the corresponding SCM and SKV values of C_v for each stimulus was then calculated.

$$\Delta C_v\% = \frac{C_{v(SCM)} - C_{v(SKV)}}{C_{v(SCM)}} \times 100 \tag{2.7}$$

Put simply, if the SKV calculations act as a whitening filter then C_v will be higher for the SCM ($\Delta C_v\%$ positive) and if the representation introduces correlations then it will be lower ($\Delta C_v\%$ negative).

The concept of whitening is relevant because statistical independence of one frequency channel from another, or of each spectral contour from those before and after it, implies a lack of redundancy in the representation.

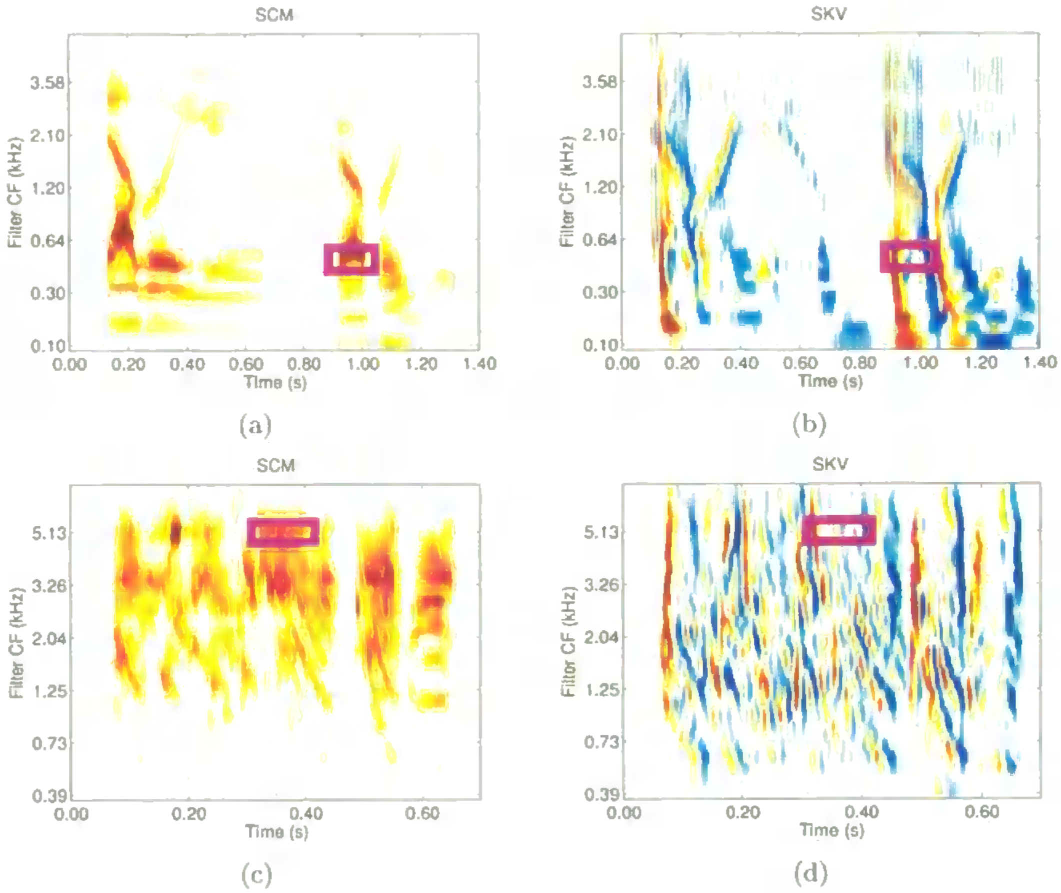
2.4 Results.

2.4.1 Examples of the SCM and SKV representation.

Figure 2.3 shows two examples of each of the representations described above. In the SCM representation (left column) the activity in each channel is always positive and is colour coded orange/red. In the SKV representation (right column) orange/red areas indicate rising energy (onset) and cyan/blue areas falling energy (offsets). The upper row shows a male and female speaker saying the letter name 'w' and the lower row shows a short segment of zebra finch song. Two magenta boxes in corresponding positions in the illustrations show how the SKV representation has values near zero when the energy is near constant in the SCM representation.

2.4.2 Level independence and adaptive non-linearities.

Several short sound files were progressively attenuated to 0, -3, -6, -12, -18, -24 and $-30dB$ and the SCM and SKV representations of each of these calculated. The root mean square (RMS) response of five randomly chosen 2 second segments of each representation was then calculated. The RMS value of *each* of the



(a)

(b)

(c)

(d)

Figure 2.3: *Examples of the SCM and SKV representations as described in Section 2.3. Upper row; (a) SCM and (b) SKV representation of female and male speaker saying the letter name W ('double-you'). Lower row; (c) SCM and (d) SKV representation of a short section of zebra finch song. Boxes indicate regions of (approximately) continuous energy bounded by large onsets and offsets, see Section 2.4.1.*

segments of the representation used in the calculation is

$$y_{RMS} = \left(\frac{\sum_{i=0}^{N-1} \sum_{k=1}^{M} y_{\text{rep}}(t - i\delta t, k)^2}{M \times N} \right)^{0.5} \tag{2.8}$$

for a stimulus representation segment ('rep' is either SCM or SKV) consisting of N discrete time values at spacing δt and M frequency channels. The response is the mean value of y_{RMS} over the five segments. The responses were then normalized such that the $0dB$ result was made equal to unity in each case. This was repeated 10 times and the standard deviation indicated in the error bars. The results, illustrated in Figure 2.4 indicate that the response of the SKV shows compression of the dynamic range. For comparison the results are also shown for the SKV with the adaptive response. This shows only $6.6dB$ attenuation over the entire 30dB range indicating level independence. However it should be noted that there is no realistic limit placed this adaptation which would result in a reduced output at very high attenuation values. The response of the system is, therefore, unrealistically high in these conditions.

2.4.3 Responses to one-over-f noise stimuli.

It has been known since the 1970s that many aspects of natural signals have characteristic spectral and temporal 'signatures' (Voss & Clarke, 1975); a specific example of this is that they show most of their temporal and spectral modulations at low frequencies. In addition they exhibit a decrease in the power of modulation in both of these dimensions with increasing modulation frequency in a way that is characteristic (Singh & Theunissen, 2003). Sounds which exhibit specific modulation power spectra can be created by modulating an ensemble of carrier frequencies to produce noise-like stimuli that exhibit '$\frac{1}{F^{\alpha}}$' characteristics, i.e. the power at each modulation frequency is proportional to the reciprocal of the modulation frequency raised to a power α. It has been reported, e.g. by Yu *et al.* (2005) that neurons in the visual cortex exhibit higher gain, and the spike responses exhibit higher coding efficiency and information transmission rates, for stimuli that exhibit $\frac{1}{F^{\alpha}}$ characteristics with $\alpha = 1$ compared with $\alpha = 0$ or $\alpha = 2$, this is close to the characteristics of natural scenes. It has been suggested by the authors (and by many others including Barlow, 1961) that the statistics of natural signals *'may play an important role in shaping and optimizing the machinery of neurons in their adaptation to the natural environment'* (Yu *et al.*, 2005). In addition there is recent evidence that cells in auditory cortex of ferrets respond more strongly and more reliably to auditory stimuli that exhibit similar

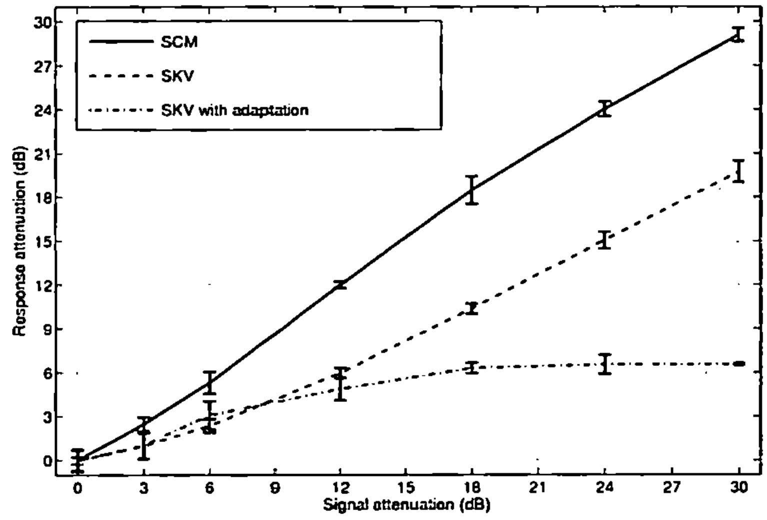
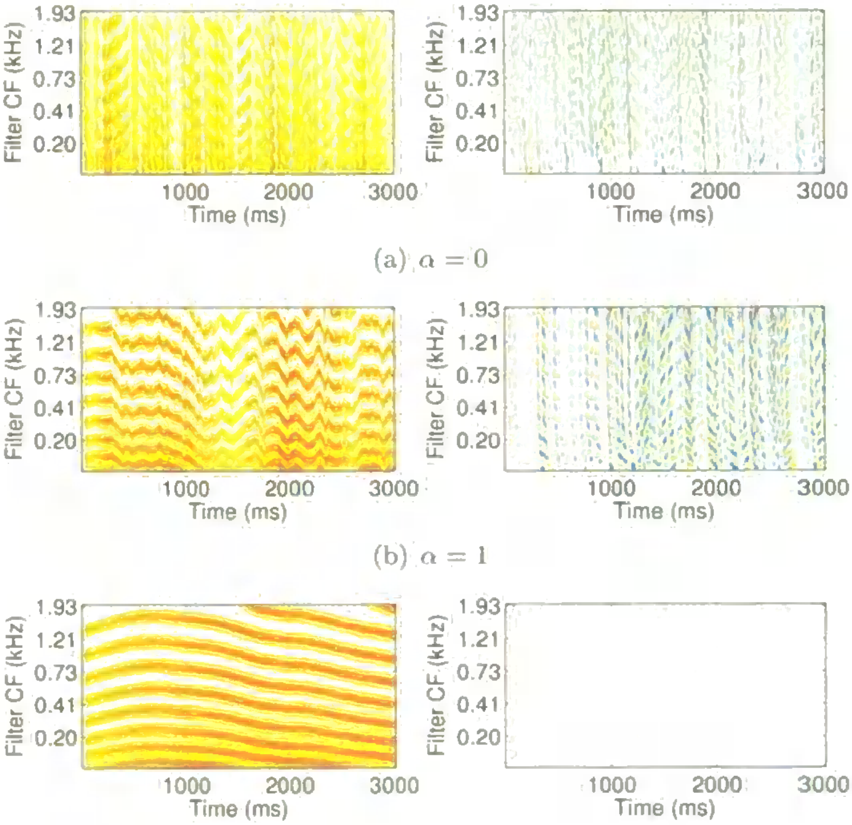


Figure 2.4: *Signal attenuation plotted against the attenuation of the RMS activity in the SCM and SKV representations, with and without adaptation. This shows a degree of level independence, see Section 2.4.2*

spectro-temporal statistics (Schnupp *et al.*, 2005). We thought it possible that stimuli with different power distributions in their spectro-temporal modulations would be represented quite differently when described in terms of short term energy change, i.e. by the SKV representation developed in this chapter. Given the evidence that auditory processing emphasizes changes of this sort it is possible that the origin of the differential response to stimuli with different statistics lies partly in this sensitivity.

To investigate whether the SKV representation responded differentially to sounds with different power spectra, a series of random tone complexes were synthesized with frequency components ranging from 0.2 to 3.5 kHz. In these stimuli the fundamental frequency and the temporal envelope were varied in accordance with independent variables with $\frac{1}{F^{\alpha}}$ distributions with values of α from zero to four. Three examples of short segments of such stimuli in SCM and SKV representation are shown in Figure 2.5. The RMS activity in the SKV and SCM representations was then calculated for five randomly chosen 10 second sections for each of the values for α. This was repeated 10 times in order to calculate error bars. The results are shown in Figure 2.6. It can be seen that the response of the SKV representation peaks at alpha values less than unity. For comparison, data from cortical responses of ferrets (Schnupp *et al.*, 2005) using similar stimuli are included on this graph.

Although the maximum in the SKV response shown in Figure 2.6 occurs for $\alpha < 1$ and the data from cortical responses of ferrets peaks for $\alpha > 1$ (which is closer to the statistics of natural sounds) it is clear that there is a differential response, and that the maximum of this response occurs for low values of α, see discussion and future work sections of this chapter.



(a) $\alpha = 0$

(b) $\alpha = 1$

(c) $\alpha = 2$

Figure 2.5: *Examples of SCM (left column) and SKV (right column) representations of $1/F^{\alpha}$ stimuli. Top row; $\alpha = 0$. Centre row; $\alpha = 1$. Bottom row; $\alpha = 2$. This illustrates the differential response quantified in Figure 2.6.*

2.5 Responses to mixtures of sounds.

This differential response illustrated in the previous section is clearly visible in the SKV representation of mixtures of sounds with different statistics. Figure 2.7 shows speech (top row) speech mixed with $\frac{1}{F^2}$ noise (middle row) and speech mixed with $\frac{1}{F^1}$ noise (bottom row). The SKV representation picks out the essential pattern of changes in the speech signal even when the level of both signals is the same for the $1/F^2$ noise example.

2.5.1 Whitening the representation.

Results from autocorrelations. Figure 2.8 shows autocorrelation diagrams (see Section 2.3.4) for; (a), (b) white noise, and (c), (d) random chord stimuli

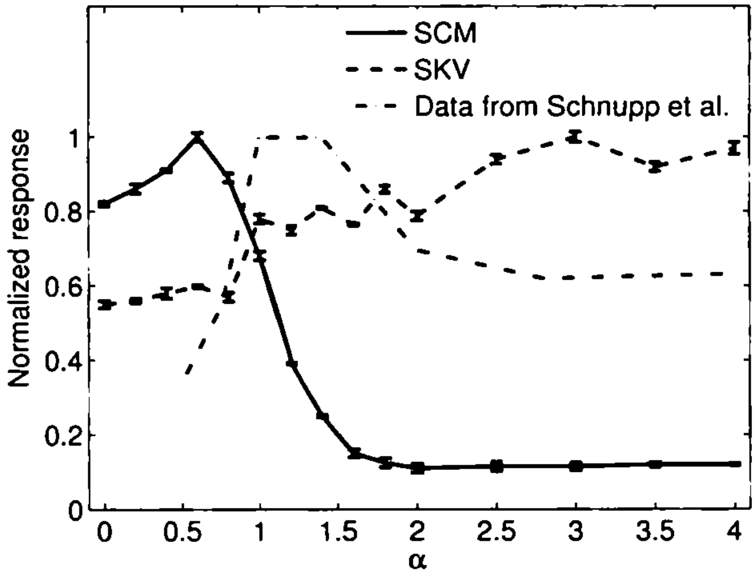
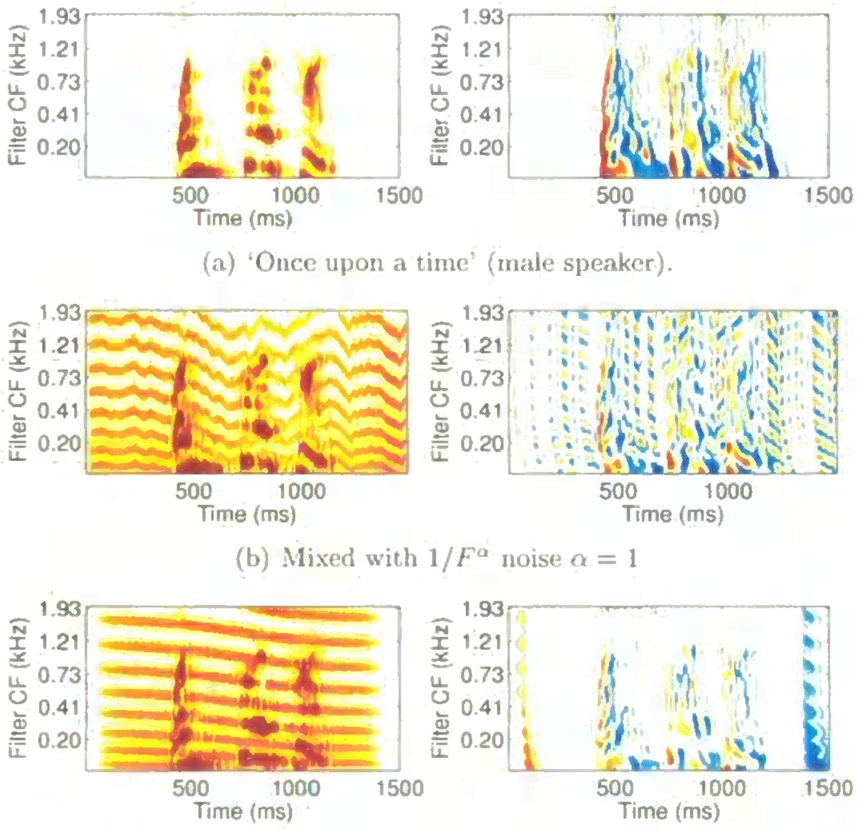


Figure 2.6: *RMS values of the activity in SCM and SKV response to $\frac{1}{F^\alpha}$ stimuli, see Section 2.4.3. For comparison data from cortical responses of ferrets are included. For each line the values have been normalized such that the highest value is unity.*



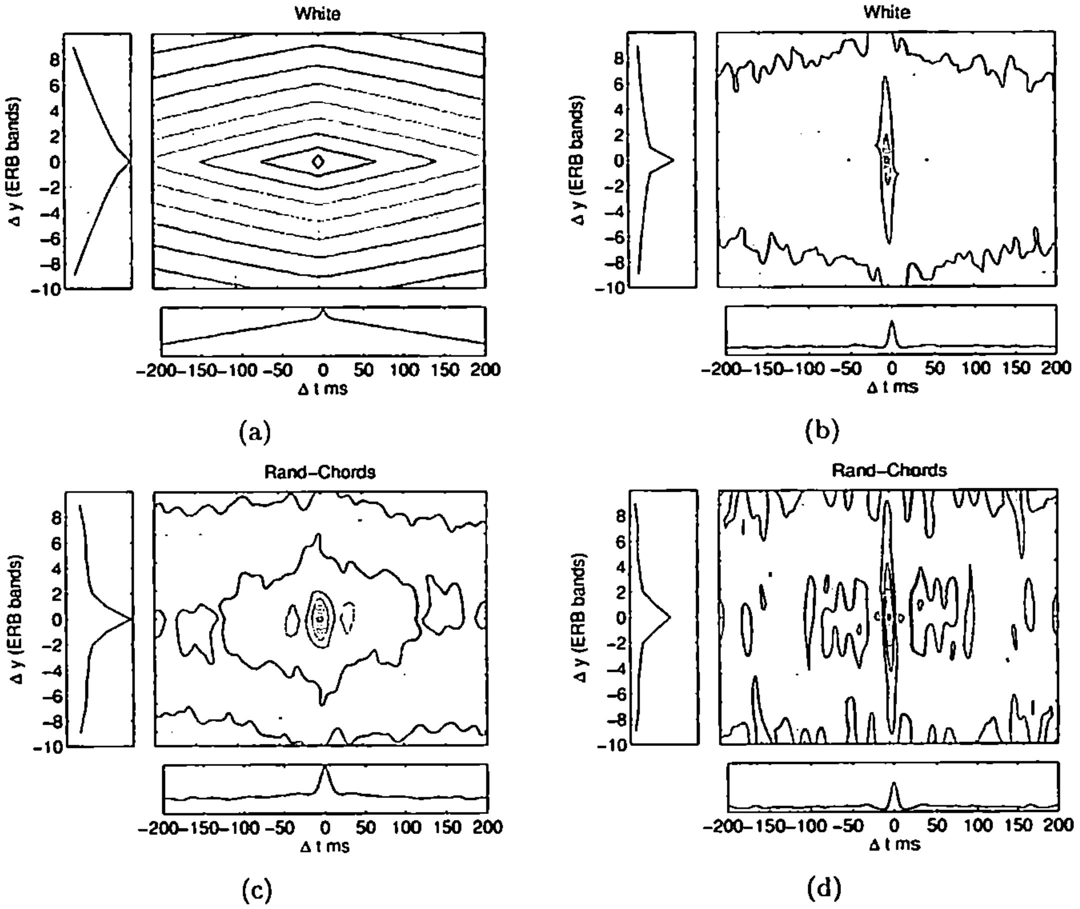
(a) 'Once upon a time' (male speaker).

(b) Mixed with $1/F^{\alpha}$ noise $\alpha = 1$

(c) Mixed with $1/F^{\alpha}$ noise $\alpha = 2$

Figure 2.7: *SCM (left column) and SKV (right column) representations of mixtures of two sounds with different statistics. Speech (top row), speech mixed with $1/F^1$ noise (middle row), and speech mixed with $1/F^2$ noise (bottom row). The patterns of the onsets in the speech stimulus are clearly dominant in the $\alpha = 2$ example when processed using SKV.*

consisting of a continuous sequence of $8ms$ tone complexes, each consisting of 8 sine tones with their frequencies chosen randomly from a set of frequencies at 1/3 octave intervals from 100 to $8000Hz$. Results are shown for both the SCM (left column) and SKV (right column) representations in both cases. Figure 2.8(a)



(a)

(b)

(c)

(d)

Figure 2.8: *Correlations in SCM (left) and SKV (right) representations of synthetic stimuli illustrated using two dimensional autocorrelation diagrams. Both stimuli ((a)-(b) white noise, and (c)-(d) random chords) are 'white' in the sense that they contain no spectral or temporal correlations. The SCM (left column) and SKV (right column) representations of the stimuli however show different patterns of temporal and spectral correlations. Lines in sub-figures show cross sections at* $\Delta t = 0$ *and* $\Delta \mathrm{ERB} = 0$. *For details see Section 2.5.1.*

shows that despite the noise being 'white' there are correlations in the cochlear representation of the stimulus. This is unsurprising as SCM consists of a set of filters with bandpass characteristics where the bands are very wide and overlap to a considerable extent. In addition the model contains a low-pass filter on each channel that introduces temporal correlations to each output band even if there are none in the signal. Figure 2.8(b) shows that in the SKV representation these correlations are markedly reduced. Figures 2.8(c) and 2.8(d) show that the SCM and SKV representations of the random chords stimulus have similar autocorrelations. These two synthetic stimuli, both nominally 'white', represent the extremes of behaviour with respect to the SKV representation. This can more easily be seen in the horizontal and vertical line plots which show a cross-section contour at $\Delta t = 0$ and $\Delta\mathrm{ERB} = 0$ respectively.

Results from covariance matrices. Results of the calculations of the coefficient of variance in the eigenvalues of the covariance matrix for a range of sounds are shown in Figure 2.9. This result for random-chord stimuli, and the result from the previous section, indicate that this type of stimulus, because of its synthetic nature consisting of very closely packed onsets and offsets, might not have greatly different properties in each of the two representations. Results from $\frac{1}{F^\alpha}$ synthetic noises for $\alpha > 1$ are not included in these results as the SKV response is extremely sparse (see Figure 2.5) and correlations occur only on very long time scales.

This measure shows a spectral and temporal de-correlation for all stimuli in the transient representation except the random-chords, with the effect being greatest for mixed speech and music, and greater for $\frac{1}{F^1}$ noise than $\frac{1}{F^0}$.

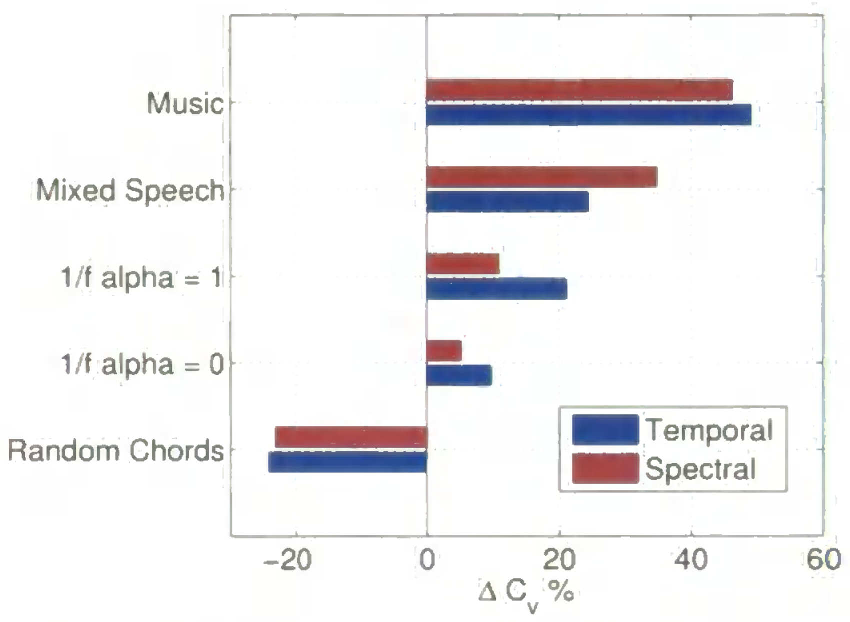


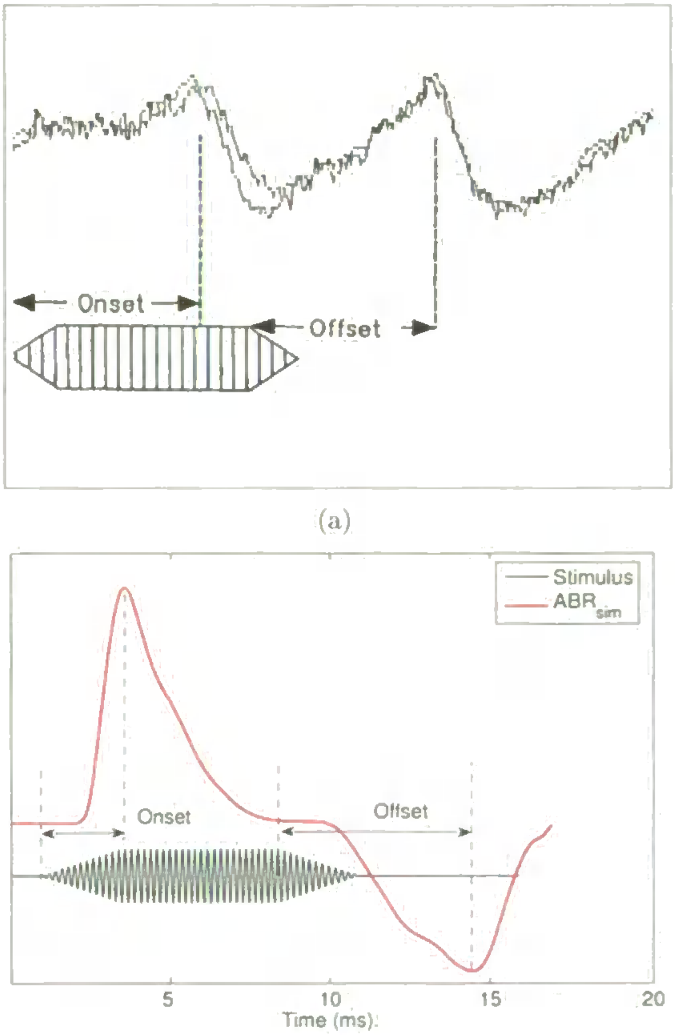
Figure 2.9: *Change in the coefficient of variation of the eigenvalues of the auto-covariance matrix from SCM to SKV. Results shown that both spectral and temporal correlations are reduced in the onset sensitive representation. The exception to this is the random-chords stimulus where temporal and spectral correlations increase on this metric.*

2.5.2 Auditory brainstem response data.

VanCampen *et al.* have reported onset and offset latencies in human Auditory Brainstem Responses (ABRs) measured using tone bursts of different frequencies and a variety of rise and fall times. This was part of a study to exclude the possibility that offset responses were an artifact of acoustic ringing in the transducers used in the experiment. During these experiments they confirmed that the offset ABR was not due to ringing and that its amplitude and latency were principally a function of the offset ramp characteristics. Figure 2.10(a) shows a schematic diagram of the measurements made by VanCampen *et al.* (1997). The envelope of the stimulus is shown in hatched outline in the lower section, and the amplitude of the ABR in the upper section. This diagram illustrates the way the latencies were measured with respect to the start of the onset and offset ramp. To make our own measurements of onset and offset latency based on the SKV representation we first summed the activity in the representation over all filter channels (x) at each time step (t) to obtain a total activity representing the simulated population response of all onset and offset sensitive units which here we will call ABR_{sim};

$$ABR_{sim}(t) = \sum_{x} y_{skv}(t, x) \tag{2.9}$$

This is shown in the red line in Figure 2.10(b) which illustrates the way the latencies are measured with respect to the start of the onset and offset ramp. These measurements are made in the same way as the results reported by VanCampen *et al.* (1997) except that in the model representation offset responses are negative while in ABR wave V measurements both peaks have the same sign representing activity in both onset and offset sensitive units. As has been mentioned in



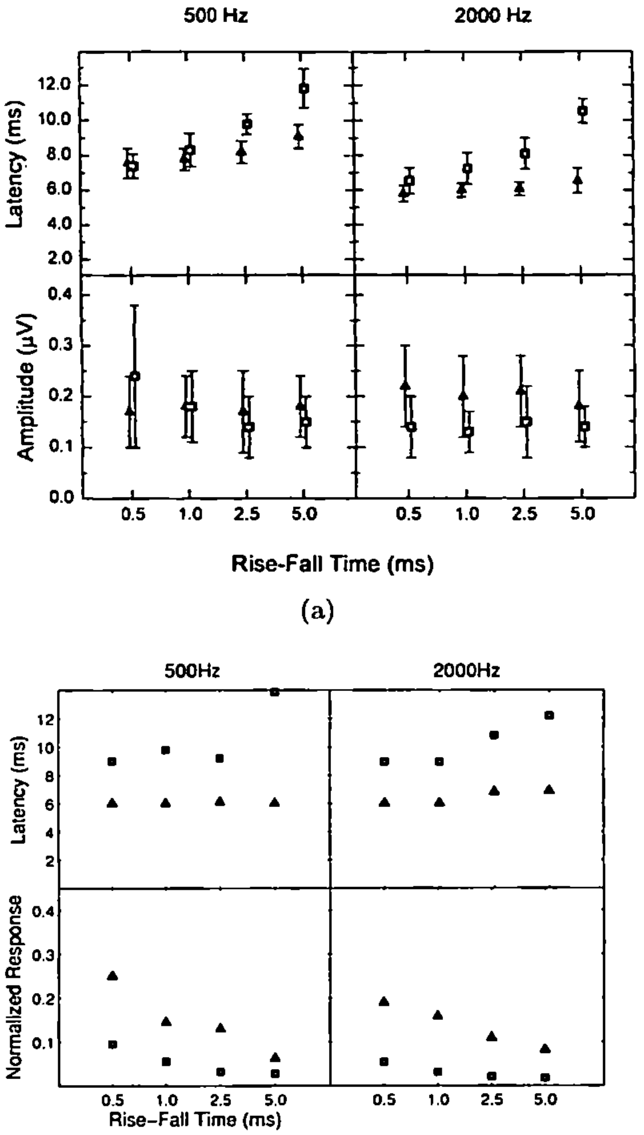
(a)

(b)

Figure 2.10: *(a) Diagram from VanCampen* et al. *(1997) showing a schematic representation of the stimulus envelope (hatched) and the amplitude of the onset and offset ABR peaks. Also illustrated are the methods of measuring the onset and offset latencies. Onset latencies are measured from the beginning of the onset ramp and offset latencies from the beginning of the offset ramp. (b) The stimulus (black) and simulated ABR onset and offset response derived from the SKV representation (red, see Equation 2.9).*

Section 2.3.2 there is some evidence He (2001) that there are two separate populations so both onsets and offsets are both indicated by an increase in activity. Using negative values for offsets in the SKV diagrams is merely a representational convenience so that both can be clearly represented in the same diagram. The results of VanCampen *et al.* (1997) for experiments with no ringing are reproduced in Figure 2.11(a). In these results the rise and fall times were the same for each tone. Tone bursts of identical characteristics were synthesized and the SKV representation calculated. The latencies and amplitudes from the model are illustrated in Figure 2.11(b). For comparison the amplitudes of the model results have been linearly scaled so that the greatest has the same value as that in the experimental data. The model results show a correspondence with the experimental data in a number of respects:

- The latency data for both onsets and offsets show a range of values from approximately $6ms$ to $12ms$.
- The offset latencies are greater than the onset latencies, although the model predicts a greater difference between the two.
- The latencies rise as the ramp times rise with the highest value being for $500Hz$, $5ms$ offset ramp, and the lowest value being for the $2000Hz$, $0.5ms$ onset ramp.
- The onset latencies rise less over the range of ramp times, although in the case of the $500Hz$ results the model predicts little or no rise whereas the physiological data records a rise of $\approx 2ms$.
- Onset and offset amplitudes fall over the range of ramp times, however these cover a greater range of values in the model results.



(a)

(b)

Figure 2.11: *Comparison of (a) response latencies and amplitudes measured from the ABR wave V from VanCampen* et al. *(1997) with (b) those measured using the simulated ABR results as derived from the summed SKV response over all filter channels. The amplitude data derived from the simulations have been normalized such that the largest result has been given the same numerical value as the largest amplitude in the physiological data.*
△ = Onset response; □ = Offset response.

The model does, however, predict a greater disparity between the onset and offset results in both latency and amplitude and fails to predict the slight overall reduction in latency from $500Hz$ to $2000Hz$. Both sets of results show that the ABR onset/offset amplitude decreases with increasing rise/fall time although the trend in the physiological data is not clear and the error bars are large.

2.6 Discussion.

2.6.1 General.

The SKV representation of auditory stimuli developed in this chapter is an attempt at a phenomenological description of auditory sensitivity to energy change that is consistent with the biological data. The psychophysical evidence (Appendix A) supports the idea that, at least, in speech, the time and spectral position of short term energy changes identified in this way, conveys a great deal of information; enough to support speech perception. Although this is consistent with much of what is known of auditory processing it has not previously been demonstrated experimentally.

The SKV representation also predicts that the latency of peak activity in subcortical processing (as measured using ABR) will rise with tone burst ramp time and the amplitude will fall for both onsets and offsets. Significantly it can be seen in Figure 2.10(b) that although the offset response starts at the beginning of the offset ramp, the *maximum* activity is not reached until after the end of the ramp. Thus the model predicts an increase in offset latency with ramp time that is much greater than the increase in onset latency. This is also found in the ABR data. The correspondence between simulated and physiological ABR data is not close enough to be viewed as an explanation but it does match some important features which have yet to be modelled in any other way.

The response of the SKV to $1/F^{\alpha}$ noise like stimuli (Figure 2.6) shows a peak at $\alpha < 1$ and a sharp fall off for $\alpha > 1$ that is not closely matched by the only physiological data available which is measured in cortex. However the fact that there *is* a differential response encourages us to pursue the idea that onset/offset sensitivity in sub-cortical regions may be a key feature in accounting for any sensitivity to natural stimulus statistics that is observed in cortex (Schnupp *et al.*, 2005) or elsewhere. See future work below.

2.6.2 Biological implications.

A great deal of work over many years has led to the characterization of a very large number of response types in the early stages of auditory processing (Trussel, 2002). This work has provided some idea, at a cellular level, of the answer to the important question: *what do cells in the auditory periphery do?* The answer to the more general question: *what does auditory peripheral processing achieve?* is however less clear.

There is evidence that (a) the auditory system performs a spectral decomposition in to a number of frequency bands, (b) that many cells of a large number of different types have well defined characteristic frequencies (He, 2001, 2002; Heil, 2001; Phillips *et al.*, 2002), and (c) there is enhancement of transients. Given this evidence it is certainly plausible that there is a role for within-frequency-band, temporal-edge-sensitive processing in the auditory system. This is consistent with results from psychophysics (e.g. Fu *et al.*, 1998; Noordhoek & Drullman, 1997) which examine the intelligibility of speech with limited numbers of bands and manipulated temporal envelopes. In addition another strand of evidence (Krumbholz *et al.*, 2003; Wiegrebe, 2001) suggests that parallel frequency channels may be processed on different, frequency dependent, time scales. This of course does not represent the whole picture, but if it is at least a partial answer to the *what* question then it is germane to ask the *why* question.

To investigate what the advantages of such processing might be, we have implemented in this chapter a simple model consisting of a cochlear filter bank followed by transient detection based on the skewness of energy distribution in overlapping frequency dependent time windows. The output of this model was found to be broadly consistent with latencies and amplitudes of auditory brainstem responses (VanCampen *et al.*, 1997) which are themselves consistent with experimental first spike timings (Heil, 1997a). Further analysis of the resulting representations of sound also showed that this pre-processing exhibited;

1. considerable rejection of stimuli with characteristics not found in natural stimuli,
2. inherent level independence,
3. de-correlation of energy in both the spectral and temporal domain.

This last property can be described as a 'whitening' of the stimulus. This whitening effect was marked when processing natural sounds, or mixtures of natural sounds. The whitening of the signal implies statistical independence of the outputs from the frequency channels thereby reducing redundancy while preserving information in the representation. That the results show that there *is* de-correlation (whitening) of the stimulus when represented in this way is significant in that it is established that the auditory pathway is sensitive to transients at all stages, and also de-correlation is a widely discussed principle of sensory (mostly visual) processing in the periphery and elsewhere. As far as we are aware no one has previously made the connection between the two.

Of particular interest is the result that whitening is found in 'white noise' stimuli that are already nominally white. This could indicate that onset sensitivity is effective at removing correlations that are introduced by the cochlea. Due to the mechanical nature of sound transduction from the tympanic membrane to the inner hair cells of the cochlea these correlations are inevitable, and it would seem a plausible hypothesis that early stages of auditory processing may be specifically designed to remove them, although this has not yet been demonstrated (cf. Atick & Redlich, 1992, for a similar argument in the visual system). The results in this chapter are consistent with this hypothesis.

These results are also entirely consistent with the information theoretic principle, the currency of which is largely due to Barlow, that:

> *At progressively higher levels in sensory pathways information ... is carried by progressively fewer active neurons. ... in most situations neighbouring points ... are more likely to be similar than distant points ... The argument can be carried on to cover the redundancy-reducing value of movement, edge, or disparity detectors* Barlow (1972).

The ability of a system to transmit information about the signal is degraded if it is encoding information about the noise in the signal as well, therefore it is useful to use a method that reduces the response to stimuli that fall outside of the range found in the environment. A degree of level independence is desirable as it enables the full use of the dynamic range of the system in the context of stimuli that vary in intensity with time.

Our hypothesis based on these results is that within-channel, transient-sensitive processing on multiple frequency-related time scales is related to the goal of efficient coding of naturalistic, behaviourally relevant stimuli. This type of processing will be present in the auditory periphery and midbrain before outputs from disparate frequency channels are integrated in cortical areas (Metherate *et al.*, 2005). It is not clear either that onset sensitivity is 'hard wired' (present in the auditory system before exposure to sound) or that it is a developmental strategy (informally, that we 'learn' to listen to onsets because it is advantageous to do so). As onset sensitivity is widespread at various points in the auditory pathway there is certainly room for both hypotheses to be true but of different areas of the auditory pathway.

Future work. The differential response of the model to stimuli with different statistics is interesting *per se* but does not closely match the only data available for cortical responses to similar stimuli (Schnupp *et al.*, 2005). However onset sensitivity is found throughout the ascending auditory pathway and the time constants in the model are not chosen to explicitly duplicate cortical responses. Two strands of work suggest themselves for future investigation. First, by manipulating the parameters of the model to more closely match cortical responses some light may be thrown on the time constants to be found in cortex. Second, by introducing more of what is known of cortical responses in to the model it is possible that the differential response of the model might more closely match cortical data.

Chapter 3

Spectro-temporal response fields.

3.1 Overview.

Principle aims. In this chapter we review a general method for estimating the spectro-temporal response field (STRF) of a neuron based on reverse correlation between a stimulus representation and an experimentally measured physiological response. Using both simulated responses and recordings made *in vivo* we investigate if the onset sensitive (SKV) representation of sound, introduced in the previous chapter, is suitable as the basis for estimation of neural responses.

Motivation. In reverse correlation estimates of neural responses (or as an alternative view, estimates of the preferred stimulus of a neuron) a choice has to be made as to how the stimulus is to be represented. If the sub-cortical auditory system emphasizes short term energy change within a frequency band then it should be possible to characterize at least some neurons as having 'preferences' for patterns of onsets and offsets. Neural responses as embodied in STRFs are more frequently derived from representations that reflect the instantaneous energy within a cochlear channel. We hope to show that estimates based on preferences for short term energy change are at least plausible as the basis for future experiments. This may show that this approach gives a more useful view of neural responses, or distinguishes between two types of neuron one of which responds to 'onsets'.

Achievements. We show for the first time that STRFs derived from an short term energy change based representation are plausible in that they are consistent with energy based estimates from the same data. We also show that these estimates in many cases have simpler geometries, often resembling Gaussian patches. These results are consistent with our hypothesis that some cortical responses could be described in terms of their sensitivity to patterns of energy change in one or more tonotopic sub-cortical regions.

3.2 Introduction.

Neurophysiologists are faced with a situation which is in some ways similar to that faced by code breakers. Given a coded message, the task of the code breaker is to discover the original message. This necessarily involves discovering the details of the encoding process. In contrast, experimentalists in neurophysiology have access to both the original *and* encoded message, i.e. the signal and the response of the system. The signal (stimulus) is under experimental control and the response can be measured using a variety of electrical and electro-magnetic techniques. This results in the apparently much simpler task of exposing the rule for the transformation of the signal in to the response, i.e. the code. It is the code, not the message, that provides insight in to the way the brain works.

Since Hubel & Wiesel (1962) showed that, for neurons in visual cortex, there were *'preferred stimuli'* which evoked a more vigorous response than all other stimuli, it has become commonplace to think of discrete neural units as having stimulus preferences and it is the quantification of this idea that has led to the concept of the spatio-temporal response field or STRF. For auditory stimuli the principle is similar but the representational dimensions are time and frequency and hence the term spectro-temporal response field (also referred to as STRF) has been adopted.

The STRF can also be viewed as a kernel describing the linear filter that best accounts for the transformation of the chosen representation of the stimulus in to the experimentally observed response.

3.3 Estimating the neural response as an STRF.

3.3.1 Reverse correlation.

The simplest case. If we describe our stimulus as a function of time and we assume that the response is likewise a function of time, then the situation described above can be summarized in Figure 3.1. [1] The problem is to describe in some way the action of the *'black box'* indicated with a question mark in Figure 3.1. We first make two assumptions.

That the system is *causal*: this implies that the response at time t is a function only of the stimulus at times less than t.

[1] For our purposes the stimuli discussed are sounds which are variations of air pressure, or tympanic displacement, in time but the argument generalizes to any function of one dimension. The 'response' means the spikes, or membrane potential measured from a neuron. From a perceptual point of view the response is not, strictly, a function of time. This point is revisited in Chapter 6.

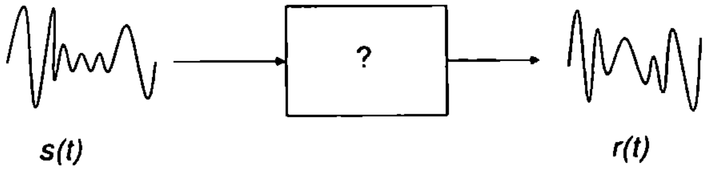


Figure 3.1: *The essence of the reverse correlation problem is how do we estimate what is in the box if we know the stimulus $s(t)$ and the response $r(t)$?*

That the system is *linear*: this limits $r(t)$ to being a weighted sum of the stimulus.

The second assumption is justified only in that there is no general method for the non-linear case. Given these two assumptions the action of the black box can be written as an integral over the interval leading up to t thus:

$$r_L(t) = \quad r_0 + \int_0^\infty h(\tau)s(t-\tau)d\tau \tag{3.1}$$

In this equation $r_L(t)$ means the linear approximation to the response at time t given by $h(\tau)$ which represents the weighting factor that determines how strongly, and with what sign the stimulus affects the response at a time τ before t, that is $t-\tau$. The constant r_0 is the response of the system when there is no signal.

Equation 3.1 actually represents the first two terms of the full expansion of the functional description of our black box in Figure 3.1 of which the first four terms are:

$$\begin{aligned} r(t) = \quad & r_0 + \int_0^\infty d\tau h(\tau)s(t-\tau) \\ & + \int_0^\infty d\tau_1 d\tau_2 h_2(\tau_1,\tau_2)s(t-\tau_1)s(t-\tau_2) \\ & + \int_0^\infty d\tau_1 d\tau_2 d\tau_3 h_3(\tau_1,\tau_2,\tau_3)s(t-\tau_1)s(t-\tau_2)s(t-\tau_3)\ldots \end{aligned} \tag{3.2}$$

This is known as the Wiener expansion and with reference to this h is called the *first Wiener kernel* or sometimes just the *linear kernel.*

Two dimensional kernels. Auditory stimuli are routinely described as functions of time and frequency.[1] We will call this second frequency dimension x. The two dimensional equivalent of Equation 3.1 is Equation 3.3.

$$r_L(t) = r_0 + \int \int h(\tau, x) s(t - \tau, x) d\tau dx \tag{3.3}$$

This can be rewritten for the case when the dimensions are discrete and h is finite in extent:

$$r_L(t) = \sum_{i=0}^{N-1} \sum_{k=0}^{M-1} h[i, k] s[t - i\delta t, k] \tag{3.4}$$

here the index i sums over N time steps. The value of δt is chosen to reflect the desired temporal resolution and the value of N represents the 'memory' of the system $N\delta t$. The index k sums over M parameters spaced over the range of interest in the dimension x, in our case frequency.

The linear kernel h now describes *'a mathematical construct that describes the integrating area of the neuron along time and along the sensory epithelium (i.e. the frequency axis)'* (Escabi & Schreiner, 2002). This is known in the auditory literature as a spectro-temporal response field or STRF.

Equation 3.4 can be conveniently re-written if we express h and s (i.e. the two dimensional matrix representations of the linear kernel and the stimulus) in

[1] This reflects the fact that the cochlea supports the analysis of the sound in to a number of overlapping frequency 'channels' the outputs of which are reflected in the auditory nerve, and that this 'tonotopic' arrangement is preserved through much of the ascending auditory pathway. As before the treatment is the same for stimuli that are functions of any two dimensions.

vector form [1] as $\vec{\mathbf{h}}$ and $\vec{\mathbf{s}}$; for the kernel we use:

$$\vec{\mathbf{h}} = \begin{pmatrix} h_{1,1} \\ \vdots \\ h_{1,M} \\ h_{2,1} \\ \vdots \\ h_{2,M} \\ \vdots \\ h_{N,1} \\ \vdots \\ h_{N,M} \end{pmatrix} \tag{3.5}$$

and we similarly form a column vector $\vec{\mathbf{s}}$ from the matrix representing the spectro-temporal description of the stimulus. Then $r_L(t)$ can be formulated as a simple matrix product:

$$r_L(t) = \vec{\mathbf{h}}_t^T \cdot \vec{\mathbf{s}}_t \tag{3.6}$$

where the vector representations of h and s are shown in bold with an overarrow and the upper-case superscript T indicates the transpose.

Equation 3.6 indicates that the coefficients of h are a function of time. If the linear response of the neuron is time invariant then estimates of h at different times can be averaged to obtain an single estimate (Theunissen *et al.*, 2001). This approach can be seen in Equation 3.7. To find values for the coefficients of $\vec{\mathbf{h}}$ which describe the linear transform between the stimulus and the measured response

[1] The approach presented here and in the rest of Section 3.3.1 is that of Theunissen *et al.* (2001).

$r(t)$ we minimize the mean square difference (MSD) between the estimated and actual response. To do this we first re-state the MSD using Equation 3.6 in order to differentiate with respect to $\vec{\mathbf{h}}$. Note that the angle brackets indicate the mean over some arbitrary time period. This can be the entire time course of the stimulus or some part of it.

$$\begin{aligned} MSD = & \quad \langle (r_L - r)^2 \rangle \\ = & \quad \langle ((\vec{\mathbf{h}}^T \cdot \vec{\mathbf{s}}) - r)^2 \rangle \\ = & \quad \langle \vec{\mathbf{h}}^T \vec{\mathbf{s}}\vec{\mathbf{s}}^T \vec{\mathbf{h}} - \vec{\mathbf{h}}^T \vec{\mathbf{s}} r - r \vec{\mathbf{s}}^T \vec{\mathbf{h}} + r^2 \rangle \\ = & \quad \langle \vec{\mathbf{h}}^T \vec{\mathbf{s}}\vec{\mathbf{s}}^T \vec{\mathbf{h}} \rangle - \langle \vec{\mathbf{h}}^T \vec{\mathbf{s}} r \rangle - \langle r \vec{\mathbf{s}}^T \vec{\mathbf{h}} \rangle + \langle r^2 \rangle \\ = & \quad \vec{\mathbf{h}}^T \langle \vec{\mathbf{s}}\vec{\mathbf{s}}^T \rangle \vec{\mathbf{h}} - \vec{\mathbf{h}}^T \langle \vec{\mathbf{s}} r \rangle - \langle r \vec{\mathbf{s}}^T \rangle \vec{\mathbf{h}} + \langle r^2 \rangle \end{aligned} \tag{3.7}$$

We can rewrite Equation 3.7 using C_{xy} for the correlation between two vectors $\vec{x}$ and $\vec{y}$;

$$MSD = \vec{\mathbf{h}}^T C_{ss} \vec{\mathbf{h}} - \vec{\mathbf{h}}^T C_{sr} - C_{rs} \vec{\mathbf{h}} + \langle r^2 \rangle \tag{3.8}$$

C_{ss} is the (NM)-by-(NM) auto-correlation matrix of the stimulus and C_{sr} is 1-by-(NM) cross-correlation vector of the stimulus and response, both averaged over the time period as indicated in Equation 3.7.

To minimize the MSD we take the derivative of the RHS of Equation 3.8 with respect to $\vec{\mathbf{h}}$ and set it equal to zero: [1]

$$\frac{dMSD}{d\vec{\mathbf{h}}} = 2\vec{\mathbf{h}}C_{ss} - 2C_{sr} = 0$$

therefore:

$$\vec{\mathbf{h}} = C_{ss}^{-1}C_{sr} \tag{3.9}$$

Note that in this formulation $\vec{\mathbf{h}}$ is no longer a function of t because it is estimated over some time period during which it is assumed to be invariant. The result, $\vec{\mathbf{h}}$, which contains the weights of the estimated linear kernel, can be re-formed in to the M-by-N matrix for convenience of plotting and this is the STRF.

The general idea of the linear estimation of neural responses by *reverse correlation* of the response with the stimulus goes back to de Boer & Kuyper (1968). The method presented in this section (due to Theunissen *et al.* (2001)) represents a generalized technique for the estimation of the STRFs of sensory neurons using arbitrary stimuli, and the normalization of the resulting estimate with stimulus autocorrelation matrix (see Equation 3.9). The method places no restriction on the nature of $s(t, x)$, i.e. the spectro-temporal description of the stimulus, and in Section 3.4 we use this method to investigate whether STRFs can be derived from the SKV representation using both simulated and physiological response data.

[1] The function to be minimized; $MSD = \langle(\vec{\mathbf{h}}^T \cdot \vec{\mathbf{s}}) - r)^2\rangle$ is positive everwhere. The solution; $\vec{h} = C_{ss}^{-1}C_{sr}$ is in the form of a general method for finding the coefficients for the solution of a set of linear equations, and as such will have either no solution or only one solution. As there is only one solution to $\frac{dMSD}{d\vec{h}} = 0$ in the linear approximation and this function is everywhere positive it cannot be a maximum.

3.4 Results.

In this section we illustrate two types of results. The first is the stimulus response cross correlation[1] discussed in Section 3.3.1. We refer to this, as in Equation 3.8, as C_{sr}. The second is the spectro-temporal response field or STRF obtained by correcting the weights of the C_{sr} to remove distortions caused by correlations in the stimulus as in Equation 3.9. The STRF, as has been previously mentioned, represents a two-dimensional set of weights, or kernel, that embodies the linear approximation to the response of a neuron.

The first part of this section (Section 3.4.1) illustrates results based on simulated neural responses derived from an arbitrary set of example weights. This serves as an illustration of the general method of Theunissen *et al.* (2001) discussed in Section 3.3.1 and to show that for white noise and random-chord stimuli the corrected (STRF) and uncorrected (C_{sr}) estimates are similar. The second part (Section 3.4.3) shows that estimates based on the SKV and SCM representations using physiological data have, as one would expect, different but consistent geometries.

3.4.1 Results using model neurons.

To investigate the methods described above we used an arbitrary set of weights representing the kernel describing the response of a hypothetical neuron as illustrated in in Figure 3.2. This simple kernel, which has a well defined characteristic frequency and an excitatory region at times close to $10ms$ is similar to many found in the physiological responses reported in this work (see Section 3.4.3). The

[1] The stimulus response cross correlation C_{sr} can be shown to be equivalent to 'the spike triggered average' (STA). (Theunissen *et al.*, 2001)

weights in the kernel are normalized such that the sum of their absolute values is zero.

3.4.2 The 'response' of a kernel.

The kernel is a two dimensional set of weights, $h[i, k]$ and the response $r(t)$ of a neuron that this kernel describes can be calculated by a convolution of this kernel with the signal representation $s[t, k]$.

$$r(t) = \sum_{i=0}^{N-1} \sum_{k=0}^{M-1} h[i, k] s[t - i\delta t, k] \tag{3.10}$$

This is merely a restatement of Equation 3.4. We interpret this response as a variation in the probability of spiking. The gain and offset of the values of $r(t)$ resulting from the convolution were adjusted to ensure a mean probability of 0.05, corresponding to a mean firing rate of 20Hz in 1ms time bins, and a maximum probability of unity. Responses below zero were treated as zero.

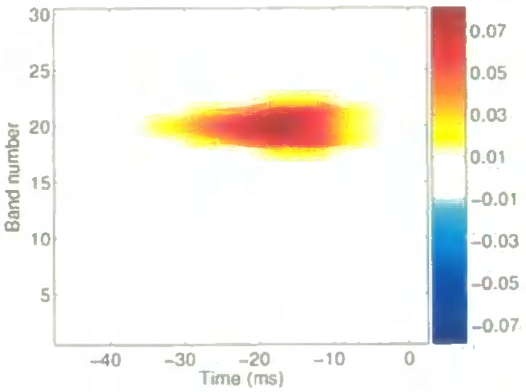
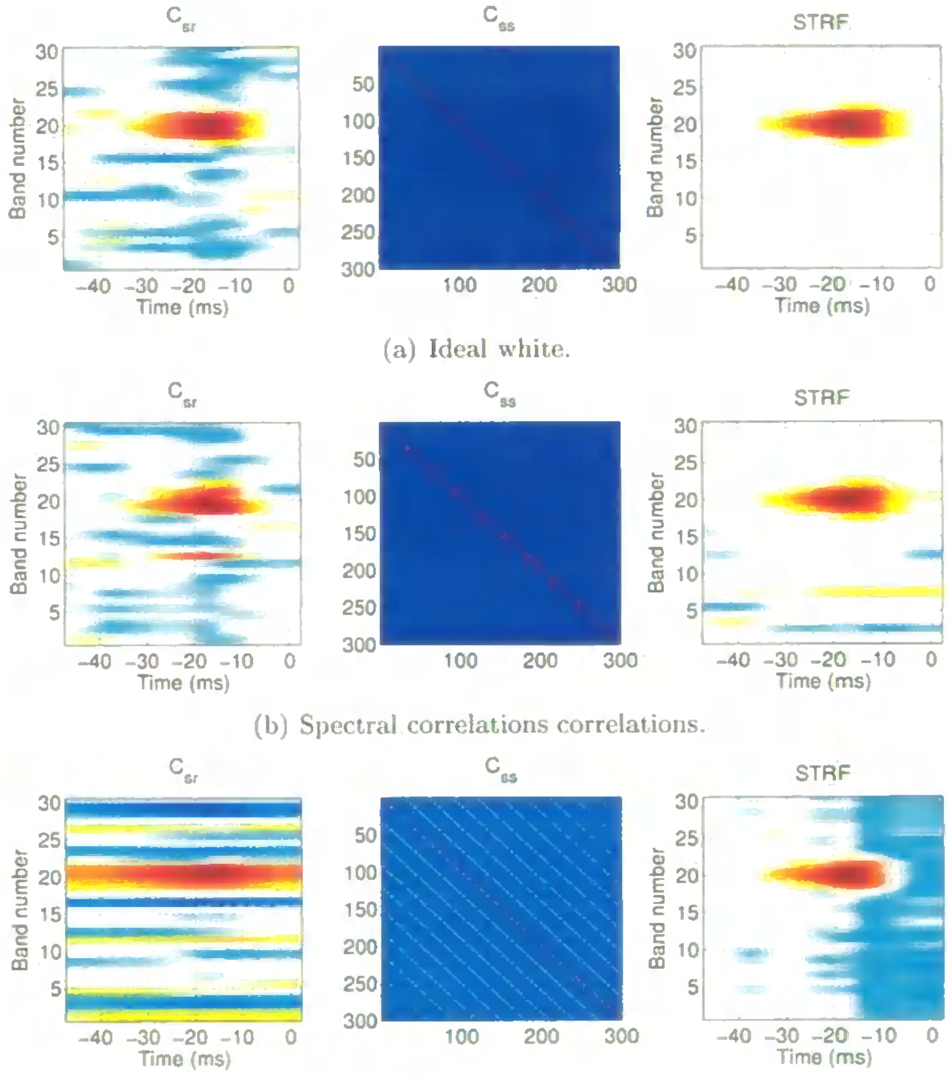


Figure 3.2: *The figure shows the pattern of weights used as an example kernel to illustrate the methods described in Section 3.3.1. Although these weights are arbitrary the spectro-temporal pattern is not unlike many STRFs estimated from physiological data in Section 3.4.3.*

How the estimate is affected by the stimulus representation. Results in this section, shown in Figure 3.3, are not derived from sounds but from two dimensional matrices similar to the specto-temporal representation of sounds engineered to illustrate the effect of correlations in the stimulus representation. Note, that this is not the same thing as *correlations in the stimulus* as the process of deriving representations may introduce or eliminate correlations. The result shown in 3.3(a) is what would be expected from an idealized 'white' stimulus representation. The stimulus response cross correlation (C_{sr} shown in the left panel) is proportional to the estimated STRF (shown in the right panel) because the stimulus has no time or frequency correlations. This can easily be seen in the centre panel as the stimulus autocorrelation (C_{ss}) is a diagonal matrix, that is, it has non-zero values only at $\Delta x = \Delta t =$ zero. With C_{ss} equal to the identity matrix then Equation 3.9 tells us that the coefficients of the stimulus response cross correlation (C_{sr}) and the weights linear STRF are the same.

Figure 3.3(b) (centre panel) shows the C_{ss} of a stimulus that is highly correlated between frequency bands, i.e. the stimulus has a similar temporal structure in one or more pairs of bands. The C_{sr} and the estimated STRF are significantly different. It can be seen that although the diagonal of the C_{ss} has become fragmented, the overall matrix is composed of a 10-by-10 array of sub-matrices and that the diagonal arrangement of these sub-matrices is preserved. This is due to there being no temporal structure to the stimulus.

In contrast Figure 3.3(c) (centre panel) shows the C_{ss} of a stimulus that contains temporal correlations, i.e. the stimulus has a similar frequency profile at a number of different, temporally related time steps. This produces large off-diagonal terms in the sub-matrix structure. The C_{sr} and the estimated STRF



(a) Ideal white.

(b) Spectral correlations correlations.

(c) Temporal correlations.

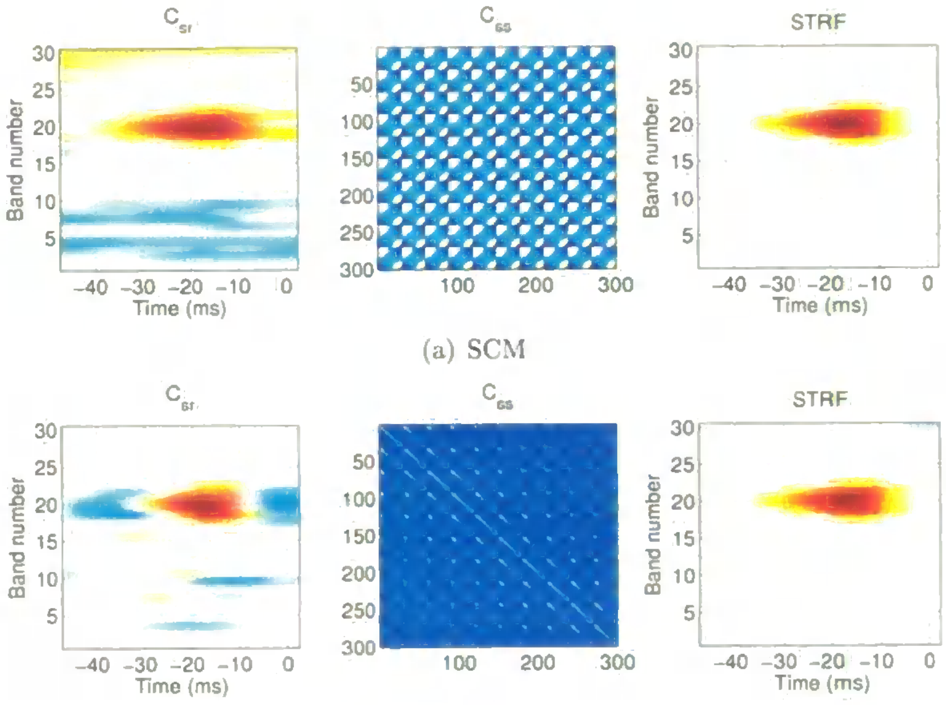
Figure 3.3: *Estimate of the example kernel using illustrative stimuli and the methods outlined in Section 3.3.1. Left: the stimulus/response cross-correlation, Centre: the stimulus autocorrelation, Right: the result of correcting the estimate using the correlations in the stimulus. The stimuli are designed to be: Upper: ideal uncorrelated or 'white', Centre: with correlations between frequency bands, Lower: with correlations in time. See Section 3.4.2 for details.*

are greatly different.

In the cases where the stimulus contains correlations the C_{ss} is not equal to the identity matrix and Equation 3.9 tells us that the coefficients of the stimulus response cross correlation (C_{sr}) and the weights linear STRF cannot be the same. The multiplication of C_{sr} by the inverse of C_{ss} acts as a 'correction' that compensates for the correlations that occur in the stimulus which are responsible for distorting the estimate of the linear kernel.

Estimates using white noise. Figure 3.4(a) shows the result of estimation of the STRF with a 'real' white noise stimulus processed using the SCM and Figure 3.4(b) shows the result using the onset sensitive representation (SKV). It can be seen in Figure 3.4(a) that despite the stimulus being 'white', the stimulus-stimulus autocorrelation matrix C_{ss} of the stimulus representation contains both spectral and temporal correlations. However, the C_{sr} and the STRF are not greatly different despite the correlations and do not show the grosser distortions visible in Figure 3.3(b) and 3.3(c).

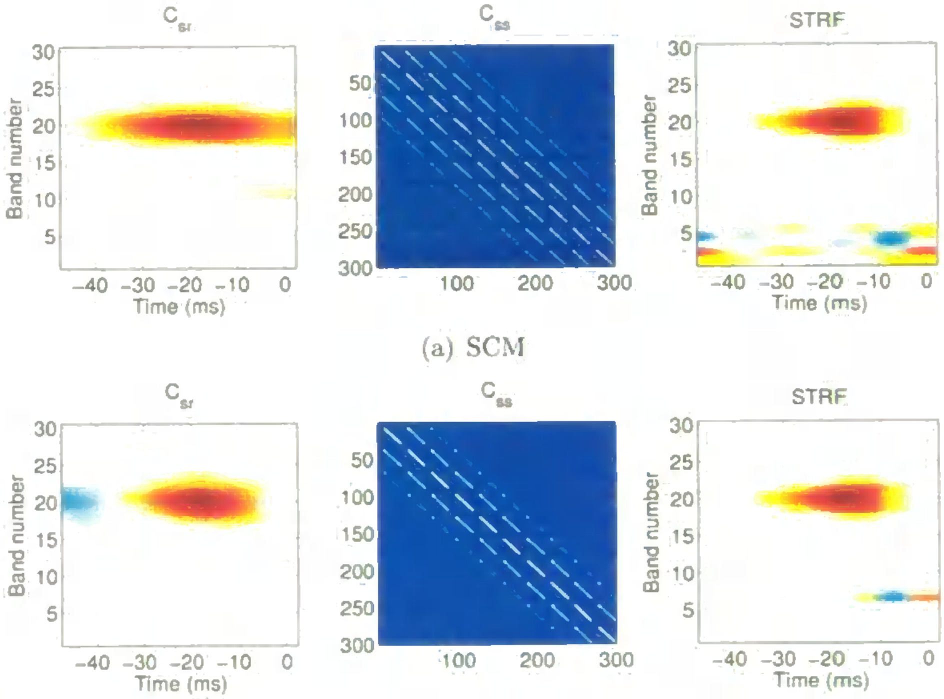
Estimates using random chord stimuli. Results of estimates of the example kernel in this section were made using random chord stimuli synthesized in the same way as those in Section 2.5.1. Stimuli of this type have been widely used in STRF estimates *in vivo* (e.g. deCharms *et al.*, 1998; Sahani & Linden, 2003). Results are shown in Figure 3.5. Again the lower sub-figure shows STRFs derived from the onset sensitive representation. The correlations in the stimulus can again be seen in the C_{ss} (centre panels).



(a) SCM

(b) SKV

Figure 3.4: *Estimate of the example kernel using white noise stimulus. Top row: using the SCM representation. Bottom row: Using the SKV representation. The correlations in the stimulus representation are visible in the stimulus autocorrelation matrix (centre panel) but the uncorrected estimates (C_{sr} left panel) are similar to the STRFs (right panel).*

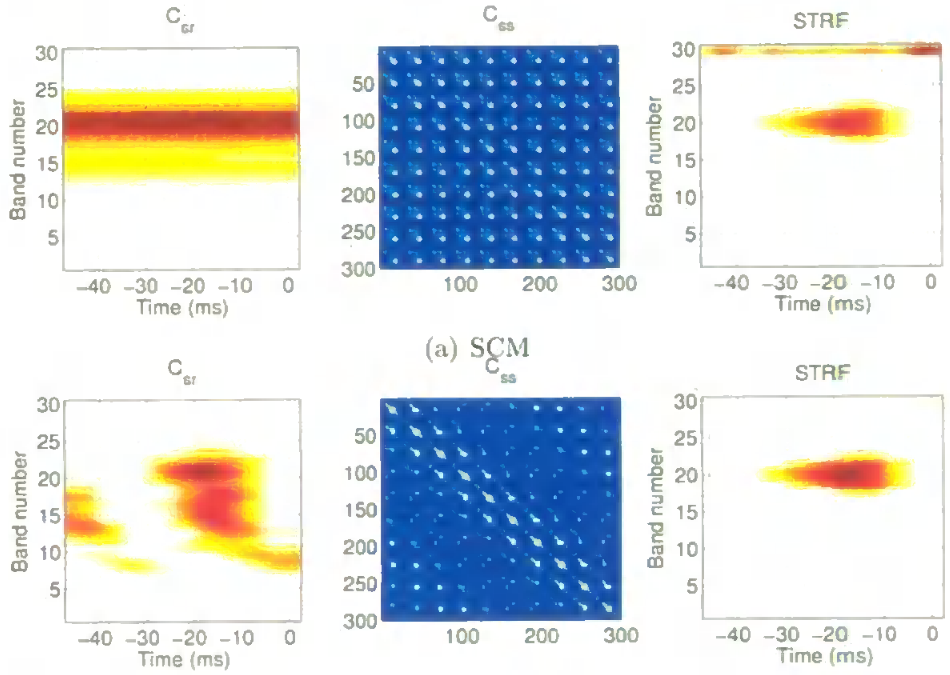


(a) SCM

(b) SKV

Figure 3.5: *Estimate of the example kernel using random chord stimulus. Top row: using the SCM representation. Bottom row: Using the SKV representation.*

Estimates using natural stimuli. Estimate were then obtained using a mixed recording of animal vocalizations (jungle noises) and the results are shown in Figure 3.6. Each of the sub-figures represents an estimate using a different section of the recording and it is clear that the STRFs derived from the SKV representation have the same geometry despite the stimulus at each of these times having very different correlations in frequency and time.



(a) SCM

(b) SKV

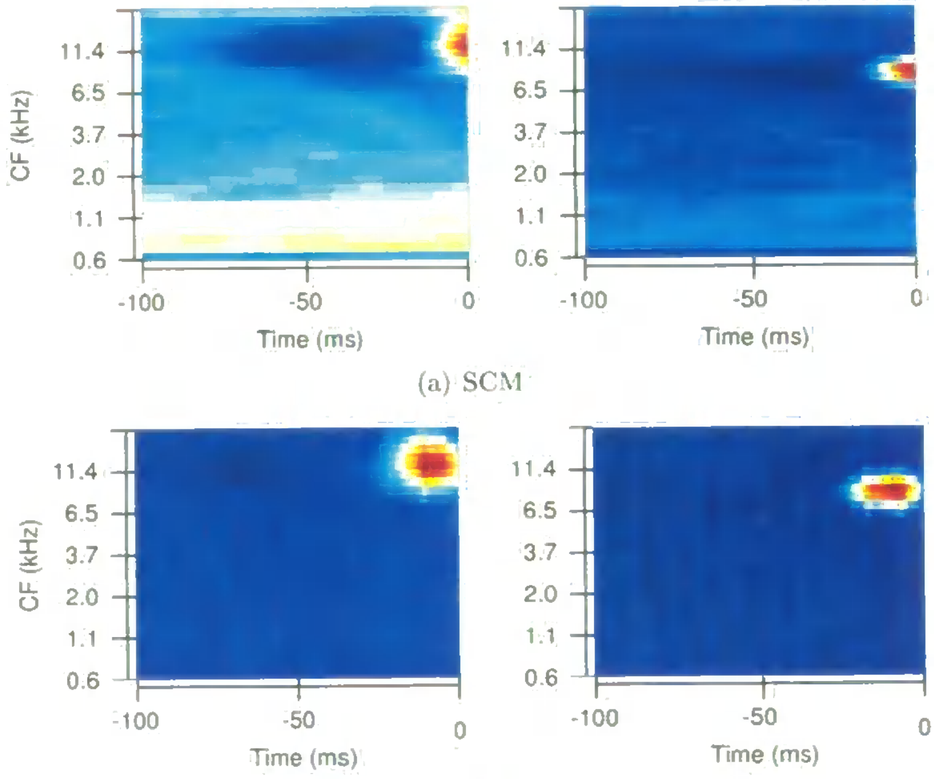
Figure 3.6: *Estimates of the example kernel using natural stimuli (a mixture of animal vocalizations or 'jungle noises'). The correlations in the stimulus representations (centre panel) are different and produce different distortions of the C_{sr} (left panels) but the same STRF after correction (right panels).*

These results indicate that natural stimuli (such as mixed environmental noise) and artificial stimuli that have been designed to contain spectro-temporal correlations, give substantially different estimates of the linear kernel before and after correction using the stimulus autocorrelation. However estimates made with white noise or with random chord stimuli, although they do show some change after correction, are of substantially the correct geometry before the correction is applied. Importantly we have shown this to be true for estimates made using both of our stimulus representations (SCM and SKV).

3.4.3 Results using neurophysiological data.

Here, in Figure 3.7, we present estimates of the linear kernel from physiological data. The results were calculated using spike trains supplied by Jan Schnupp of the University Laboratory of Physiology, University of Oxford. They were obtained from anaesthetized ferrets using random chord stimuli consisting of a number of 7.5 second segments of tone complexes consisting of $2.5ms$ tone bursts presented synchronously and chosen randomly from frequencies from $500Hz$ to $24000Hz$. The spike trains were recorded in a number of areas all in auditory cortex.

The two estimates shown in Figure 3.7are typical of over 50 obtained in different locations within the ferret auditory cortex. The estimates based on the SKV representation (left column) are consistent with those based on the SCM representation (right column) and have simple geometries indicating a well defined spectral and temporal preference. These results demonstrate that estimating the linear kernel which approximates the response of units in cortex using a representation which is sensitive only to energy change within a frequency band yields results that are at least plausible. However, two interesting points emerge from comparisons between the two sets of results. First, evidence suggests that there is a significant delay associated with processing in the ascending auditory pathway which can be measured in human auditory brainstem response (e.g. VanCampen



(a) SCM

(b) SKV

Figure 3.7: *Physiological estimates of the linear kernel approximating the response of two cortical units in ferret (the two different units are in the left column and the right column). Upper row, estimates based on the SCM representation. Lower row, estimated based on the SKV.*

et al., 1997) and in cortex (e.g. Heil, 1997a). The minimum latency recorded for human ABR in the study cited was between five and six milliseconds, and in cortex Heil investigated the effect of rise time function on first spike timing and reported *' ... the different functions appear to converge on a single minimum at* 12.3 *ms'*. The overwhelming majority of kernels estimated from the SCM representation have a maximum excitatory component in a narrow frequency band *and* at zero latency. Put another way, the presence of energy in the stimulus at the best frequency (BF) of the cell is most likely to evoke a spike with zero time delay in the cortex. This is in contrast to those derived from the onset sensitive representation which predict that a rising transient is most likely to evoke a spike at a time offset of between five and twenty milliseconds.

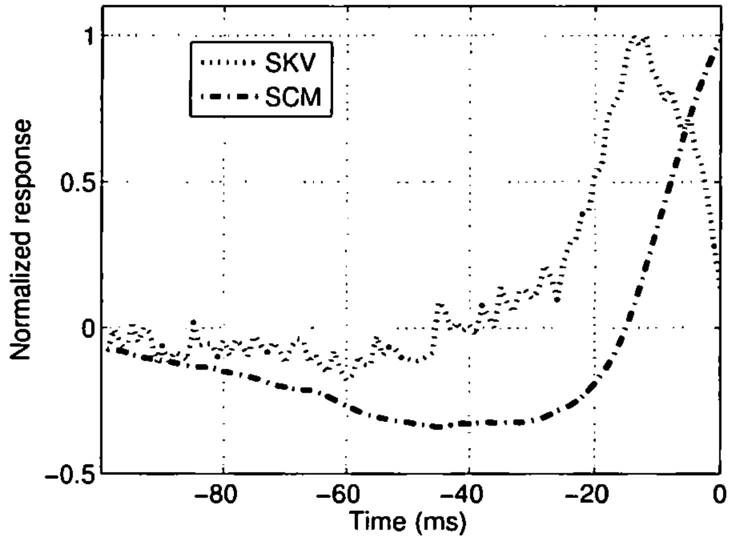


Figure 3.8: *The temporal structure of a typical C_{ST} pair. The two lines illustrate the temporal profile at best frequency of a pair of kernels estimated from the same spike train using the SCM and SKV representations. The SCM based estimate is still rising at time* = 0.

An illustration of this can be seen in Figure 3.8 which shows the temporal structure at best frequency of the pair of kernels shown in Figure 3.7(a). The maximum weight in the energy sensitive kernel occurs at −12ms for the stimulus, or alternatively $+12ms$ spike latency which is comparable with that reported by Heil (1997a).

3.5 Discussion.

3.5.1 General.

In this chapter we have shown for the first time that estimates of spectro-temporal preferences exhibited by cortical units can be understood in terms of patterns of short term energy change. These estimates are meaningful and broadly consistent with those obtained using more traditional, energy sensitive methods. Although not all units may be change sensitive this provides a novel framework for discussion of auditory feature extraction which is consistent with much of what is known about the nature of sub-cortical auditory processing.

3.5.2 Biological implications.

The process of reverse correlation is dependent on choosing a meaningful representation of both the stimulus and the response. STRFs in the auditory centres of the brain have been measured using spectrographic (FFT), cochleographic, and parametric representations, all of which have led to results which give some insight in to the nature of auditory processing. However, there is evidence that much of the ascending auditory pathway is processed within each tonotopic channel, and hence without respect to spectral profile, and that much emphasis is placed in each channel on representing the change of energy rather than the energy level itself. It could be therefore that STRFs could be estimated in terms of a representation such as that introduced in Chapter 2 and that this could lead to a more informative model of processing higher levels of the auditory system. The exact nature of the neural responses immediately antecedent to the unit from which recordings are being made is always going to be a matter of some uncertainty. It is, however, likely to reflect the nature of neural responses measured in more peripheral regions and not *necessarily* bear a close relationship to the parameters used in the synthesis of the stimulus or the simple spectro-temporal picture provided by spectrographic analysis. Thus the picture that emerges from experiments of this sort is indicative of the whole chain of auditory processing and not the response of the unit under investigation. Reverse correlation with representations that respect sub-cortical responses may tell us more about cortical function and may help to bridge the gap between the predicted and measured responses to novel stimuli.

One of the stated aims of project, i.e., to show that cortical responses might couched in terms of sensitivity to energy change within a frequency channel, has been brought one stage closer to being justified biologically. We have used the simple expedient of characterizing a real neural response (in fact a range of responses measured in different cortical locations) in terms of the SKV representation that is developed in this work. In Section 3.4.3 we have calculated the stimulus response cross-correlation C_{sr} using spike trains measured *in vivo* correlated with the onset sensitive representation and also with the cochlear representation and the results show that both methods give meaningful results that are consistent with one another. Many estimates of cortical responses based on energy are obtained by grouping spikes in time bins of up to $20ms$. This is larger than Heil's reported limiting first spike delay and in this case the excitatory region would be located *'in the first 20 ms'*. This is also true in the case of kernels estimated using the SKV representation. However, when the kernel is estimated with much smaller bin sizes (in this chapter bins of $1ms$) the delay introduced by the SKV representation which is a function of the size and number of the windows used in this calculation, is consistent with first spike latencies measured physiologically. In addition these latencies are frequency dependent, a feature which is reported in some cortical responses (e.g. Krumbholz *et al.*, 2003).

An additional point of interest is that the estimates based on the SKV representation have, in many cases, simpler geometries resembling simple two dimensional gaussian distributions. The precise implications of this can only be established by future work as outlined below.

3.5.3 Future work.

If some cortical responses are, as we hypothesize, dominated by sensitivity to the energy changes emphasized in sub-cortical processing this makes a prediction that is, in principle, easy to test. Physiological recordings from cortex could be made using stimuli with parametrically controlled onset and offset distributions. These could then be reverse correlated with a suitable stimulus representations and this should distinguish neurons that respond to transients from those that respond to the instantaneous energy within a frequency band based on tonic firing, rather than phasic firing, in sub-cortical regions. It is our intention in the near future to use our established links with physiologists to design such an experiment.

Chapter 4

Fragments and ensembles.

4.1 Overview.

Principle aims. It has been shown using small section of images (image *fragments*) that similarity between these small patches of a picture and a large range of other pictures, carries information about whether the picture in question belongs to the same 'class' as the picture from which the fragment is drawn.

In this, and in the next chapter, we hope to adapt and extend this idea to measures of time dependent similarity between *ensembles* of fragments and classes of stimuli in the auditory domain. These spectro-temporal patches, which are small areas excised from stimulus representations, can be seen as fragments, stimulus preferences, or STRFs, and the time varying similarity between each patch and the stimulus representation can be interpreted as the 'response' of each to that stimulus. Such patterns are the best way we currently have of describing the features which support auditory perception and predicting the response of a neuron to a novel stimulus.

We will examine a large range of fragments derived from a small set of stimuli, and we hope to show that the responses of some of these are more likely to convey class information than others. This distinction may form the basis of a developmental process through which neural responses are refined.

Motivation. It is widely believed that auditory perception is based on the responses of cortical neurons that have spectro-temporal preferences, as described in the previous chapter. This has sometimes been called 'feature extraction' although it is not clear at this stage what these features might be or how they might come in to existence. There is evidence that cortical responses develop to reflect the nature of stimuli in the early post-natal period. We hope to show that the patterns found in a limited number of stimuli which reflect some putative early auditory environment, may bootstrap the formation of these responses.

It has also been shown, in the work on fragments of visual images, that the size of the fragments affects the amount of class information carried in the measure of similarity. If this is true of auditory fragments we hope to gain some insight in to the time period over which auditory features may be extracted in the brain.

Achievements. In preliminary results gained by selecting groups of fragments at random we show that there is a difference in the entropy of the responses from ensembles of different maximum temporal extent. This we interpret as evidence that, at least for speech, STRFs of different temporal extents may may be more or less suitable for auditory feature extraction. This is further pursued in the next chapter. In addition, using a more sophisticated (and previously reported) procedure designed to select ensembles of features whose responses are highly correlated with class information but not with each other, we show that ensembles chosen using these criteria have temporal properties that are consistent with this time scale and with those of STRFs measured *in vivo*.

4.2 Representation in response patterns.

A powerful idea that first came to the fore with the theory of trichromatic colour vision (Young, 1802) was that a particular stimulus quality was characterized by an *across-neuron response pattern* (ANRP). The idea was that there are three types of receptors in the retina that respond differentially to different wavelengths of light and that a single wavelength of light, or combination of wavelengths, produces a characteristic ratio of responses across the three receptors. This idea was taken up by Helmholtz (1860) and later developed by Erickson (1974) as a comprehensive theory applicable to all areas of sensory coding.

The basis for this idea is that although responses of single cells are typically broad, organisms are capable of making behavioural judgements based on very fine distinctions. There is no contradiction here because a population of broad, overlapping responses can encode a specific stimulus in the pattern of activity across the population; the ANRP. These ideas form the framework for the proposal introduced in Chapter 3, that auditory coding can be understood as the response of a set of broadly tuned spectro-temporal filters, and leads us to an important but difficult set of questions about how useful the responses of one, or more, of these filters might be and from a developmental point of view, how they come to have the properties that they do.

4.2.1 Entropy and Mutual information.

If the output of a filter, or feature extractor, is essentially flat (monotonic) over the time course of a stimulus, or over the entire range of stimuli with which it is presented, then the output of this filter is not 'interesting' (Dayan & Abbot, 2001). [1] The response of a filter might be also be uninteresting because its output bears a simple relationship to that of another filter and so no additional benefit is gained from this response. A third way a filter might be deemed uninteresting is if its output is varied and unique but bears no relationship to the change of the input stimulus. These three characteristics of the output can all be quantified using measures from information theory. The first, the information capacity due to variability, is the **entropy** of the response; the second, the difference between the response and other responses, is the **redundancy**; and the last, the amount of information about the stimulus that is contained in the response (of a single filter or an ensemble of filters) is the **mutual information**. These three quantities are all interrelated.

The context within which these issues are often discussed is that proposed by Shannon (1948), whose work forms the basis for the understanding of how information is preserved by systems of coding. For this approach to be applicable both the signal and the responses must consist of nominal classes, or symbols, within a corresponding time window.

The probability with which the symbol r occurs in the signal is written $P[r]$ and the element of 'novelty' or 'interestingness' associated with each symbol as $-log_2P[r]$, the log measure is chosen so that the value is additive for independent

[1] The introductory treatment given here relies principally on that given in Dayan & Abbot (2001).

sources, and the negative sign ensures that the level of surprise goes down as the frequency of occurrence goes up. This quantity is then averaged over all symbols weighted by the probability of occurrence $P[r]$ (Equation 4.1) to yield the **entropy** H.

$$H = -\sum_{r} P[r] \log_2 P[r] \tag{4.1}$$

The effect of this calculation is that symbols that almost never occur, and symbols that occur frequently, both contribute little to the entropy. The highest contribution will be from those symbols which occur not too often so as to reduce the surprise associated with them, and not so rarely that they carry a low weight, and the maximum entropy will be when all symbols occur with equal frequency. The logarithm in Equation 4.1 can be calculated to any base but conventionally base two is used and the entropy stated in *bits*, although it is in fact dimensionless.

For a response to provide information about the stimulus its variability must be related in some way to the stimulus variability. To quantify this the entropy of the response when the stimulus does *not* change (known as the noise entropy) is subtracted from the total entropy. This is known as the mutual information (I) between the stimulus and response:

$$I = H - H_{noise} = -\sum_{r} P[r] \log_2 P[r] + \sum_{s,r} P[s] P[r|s] \log_2 P[r|s] \tag{4.2}$$

which can be rearranged to give:

$$I = \left\langle \sum_{s} P(s|r) \log_2 \left[\frac{P(s|r)}{P(s)} \right] \right\rangle_r \tag{4.3}$$

where $\langle \ldots \rangle_r$ denotes the expected value over all response symbols.

The maximum value of I is just $log_2(n_C)$ where n_C is the number of symbols or classes. In order to compare results from investigations involving a variety of values for n_C, results in this work will often be quoted as normalized mutual information I_n which is simply:

$$I_n = I/log_2(n_C) \tag{4.4}$$

4.2.2 Visual fragments.

Ullman *et al.* (2002) have shown that visual features of intermediate extent and complexity are optimal for classifying images. In this work, a library of fragments were extracted from a large number of images each belonging to a restricted set of classes. Each fragment was then assigned a class informativeness based on the frequency of its occurrence in databases of images that either did, or did not, contain images of that class. The presence or absence of a fragment in an image was judged by a measure of similarity using a threshold level. This leads to a binary result representing the presence or absence of the fragment in an image. The resulting frequencies of occurrence of a fragment in a database of images of a particular class allow the calculation of the mutual information between the fragment and the class. It was shown that the fragments that were most informative about the class were of intermediate size and resolution. It was further shown that when classifying images using a combination of fragment results were improved if the fragments were drawn from the most informative set.

The idea that a fragment that is not so large as to be too specific, and not so small as to be too general, might provide information about a stimulus based on its presence or absence in that stimulus is not conceptually complex. However, the results reported by Ullman *et al.* have shown not only that fragments of intermediate size and complexity are superior, but they have demonstrated that the selection of these features as the basis of neural representations might be driven by maximization of mutual information. This is a useful idea that might be applied to auditory stimulus classification and form the basis of a developmental argument.

4.2.3 Fragments of auditory representations.

The approach reviewed above recommends itself for investigation in auditory classification for a number of reasons.

- As has been discussed in Chapter 3 the time dependent response of an STRF to a stimulus can be approximated using a convolution. The convolution can also be interpreted as a measure of similarity; its output is at a maximum at times when the stimulus representation matches the STRF. In this chapter we introduce the idea that spectro-temporal patterns which are fragments of stimulus representations can be viewed an *'candidate STRFs'* that is, we examine whether an STRF *would* be useful if it had the same spectro-temporal geometry as a particular fragment of the response representation. So we define a *'candidate STRF'* as a two dimensional (i.e. spectro-temporal) pattern of weights derived by subdividing a spectro-temporal description of some stimulus in to fragments of limited spectral and temporal extent.

- In contrast with visual representations, auditory representations do not normally suffer from the same variations of scale or orientation. In order to

> compare images with fragments it is necessary to compensate for the inherent variation of extent exhibited by visual stimuli on the physiological substrate (the retina). Also, for the comparisons to be valid the orientation of both has to be assumed to be the same. Auditory stimuli are extremely complex and many problems surrounding their processing and representation are currently intractable. However, it is, at least as a useful approximation, true to say that the same auditory stimulus manifests itself at the same scale along the physiological substrate (the cochlea) at each presentation, and that the time dimension which is the principle degree of freedom for *scaling* and *displacement* has a consistent *orientation*.

Ullman *et al.* (2002) also show that features derived from e.g. faces when used for classification generalized well to both novel faces and *face-like-objects* such as paintings of faces. They also show that these features form the basis for a perceptually convincing reconstruction of face stimuli. These results form the basis of an argument for a neural representation that is essentially a first order isomorphism or *representation-by-similarity*. It has been suggested (Edelman, 2002; Shepard & Chipman, 1970) that a more plausible mechanism for biological systems would be a second-order isomorphism or *representation-of-similarity*, see Section 5.3.

4.2.4 How interesting is a fragment?

The representation used in this work as the basis for deriving fragments (the SKV) is considered to be a matrix of scalar values. A fragment of this representation is therefore is a rectangular sub-matrix drawn from contiguous values of the representation. How interesting is each of these fragments as a candidate STRF?

In some senses this is an analogous question to that asked by Ullman *et al.* about their visual fragments. They investigated how informative a fragment was in respect of the class from which it was drawn. A much more general question would be to ask *'how interesting is this fragment in its response to a range of different stimuli'*. The answer to this question can be framed in terms of the entropy of the response of the fragment to a limited set of stimuli. Note that in this context 'response' means the convolution between the fragment and the stimulus representation, i.e. the time varying similarity. If this response is uniform across all classes then it tells us nothing about them. If on the other hand the fragment responds in a distinctly different way to each of the test classes then it is 'interesting' as described in the previous section, i.e. it is 'of relatively high entropy' and '*potentially* information bearing with respect to stimulus class'. As ever the devil is in the detail, and the devil in this scheme lies firstly in the phrase *distinctly different.* For a real valued continuous measure of similarity it is necessary to decide how different a response has to be before it is regarded as distinct. The second problem is that there are any number of ways in which the response can be characterized. The approach adopted in this work is discussed in detail in Section 4.3 below.

4.2.5 How interesting is an ensemble of fragments?

In a similar way to that discussed above, it is possible to calculate the response entropy of an ensemble of fragments. It is therefore possible in principle to compare all possible ensembles and pick the one which has the most interesting response, or at least pick one of the subset which have equal and maximal response entropy. This is, however, not practical because it involves calculating the response for $\frac{k!}{n_E!(k-n_E)!}$ ensembles, where k is the number of fragments and n_E the ensemble size, and subsequently comparing each pairwise.

In Section 4.3.2 an algorithm is described which seeks to add fragments to an ensemble on the basis that each new fragment has a response that contains more information about a range of stimuli than it does about any of the responses of previously chosen fragments. This correlation based filter provides a fast way to identify relevant responses, and redundancy among responses, without pairwise comparisons (Yu & Liu, 2003b).

4.2.6 Auditory salience.

The issue of the symbols which are to be used to calculate the entropy and mutual information between the stimulus and response remains unaddressed. For the stimulus we are interested only in its class, which is a static label attached in the preparation of the experiments. The class of a stimulus represents some *a priori* knowledge the experimenter has about a stimulus. This allows it to be labelled in a way which is relevant to the experiment. For example, if the speech corpus has 500 utterances that are labelled as being the letter *'A'* 250 of which are recorded from male speakers and the remaining 250 are recorded from female speakers then these 500 stimuli belong either to the same class *'A'*, or two different classes 'Male' and 'Female' depending on the context of the experiment.

The class of the response has to be handled differently. For each individual fragment the response is a convolution between the fragment, interpreted as a kernel, and a spectro-temporal description of the stimulus, See Section 3.4.2, Equation 3.10. These individual responses are continuous, real valued functions of time and are hence doubly unsuitable for the purpose of representing response symbols. What is required is a quasi-static description, representing a view of the stimulus within a finite time window representing an event. This problem is addressed in Section 4.3.3.

4.3 Methods.

4.3.1 Fragment extraction.

In order to explore the possibility that the formation of STRFs may be bootstrapped by fragments of activity patterns in response to acoustic stimuli, several libraries of fragments derived from small sets of sounds, the *formative stimuli* each representing one of a small number of classes, referred to as the *formative classes*, were created. The first set contained samples spoken stimuli, i.e. the numerals; 'one', 'two' ...'nine', 'Oh', and 'Zero' making the number of classes $n_C = 11$. The second set of formative stimuli consisted of stimuli belonging to eleven non-speech classes; wind noise, rain noise, bird calls, frog calls, whale calls, isolated engine noises, traffic, telephone bells, simple collisions (e.g. pool balls, plates etc), electronic sweeps and glides, and breaking bottles. All of the sounds were pre-processed using the SKV representation described in Section 2.3.2.

From each of these sets of formative stimuli four separate libraries of fragments with a range of durations from 10 to 200 milliseconds, were created. Within each library fragments were 4, 8, 12, 16, 20, or 24 frequency bands wide. Each fragment, therefore, represents a two dimensional, rectangular patch excised from the SKV representation of one of the formative stimuli at intervals of one third of a fragment horizontally and two bands vertically, this is illustrated in Figure 4.1. An example of fragments drawn from the 100ms library of both the speech (upper

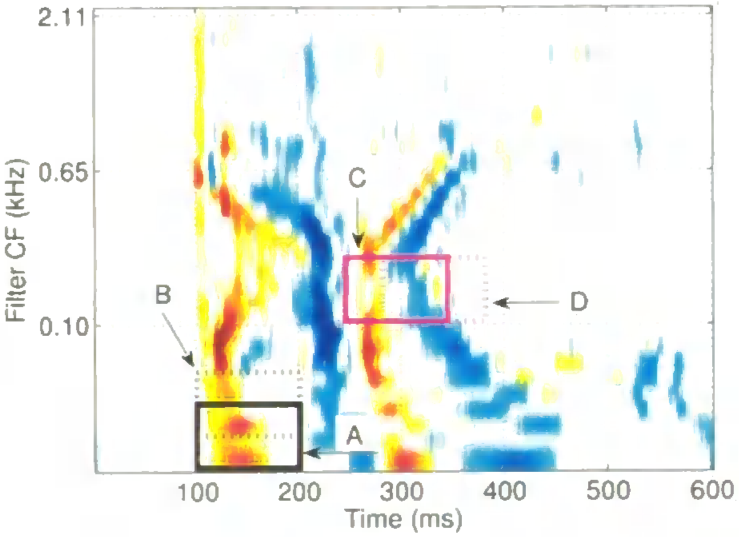


Figure 4.1: *Illustration of fragment extraction. The solid black box A represents a* $100ms$ *fragment covering 4 frequency bands. The black dotted box B represents the next fragment in a vertical tiling shifted by two bands. The solid magenta box C represents a similar fragment and the magenta dotted box D represents the next fragment in a horizontal tiling shifted by one third of the fragment length. The entire representation is treated in this way.*

row) and non-speech (lower-row) libraries is shown in Figure 4.2.

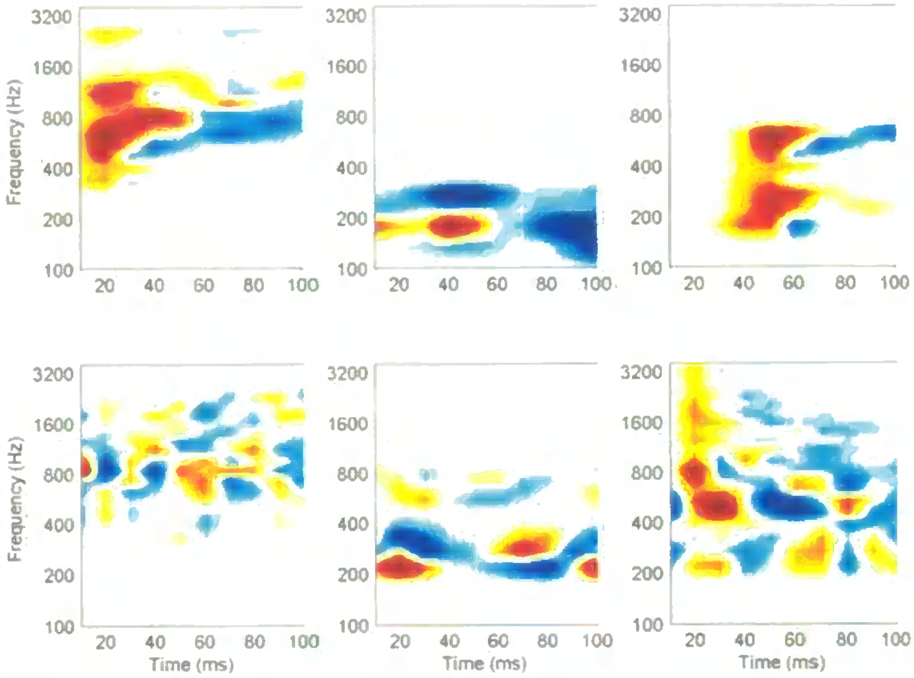


Figure 4.2: *Examples of fragments derived from speech (top row) and environmental noise samples (bottom row).*

4.3.2 Ensemble selection.

Maximal entropy. In contrast to Ullman *et al.* (2002) it was not sought to optimize the fragment choices with respect to the set of all stimuli. From a developmental point of view, responses must be useful with respect to the formative stimuli only, as this is the basis on which the developing perceptual system might prefer them. If the formative stimuli are sufficiently rich, and are representative of the statistics of ethological sounds in general, it is to be hoped that the patterns derived from them will be useful for classes outside the developmental experience.

Results given in Section 4.4 show that this is indeed the case.

As the basis for a decision as to which fragments ensembles were likely to perform better, large numbers of random ensembles of fragments (n_E = 2, 4, 8, and 16) were generated and the entropy (H) of each of their responses to the set of formative stimuli from which they were drawn was calculated. Ensembles from the highest and lowest entropy bins were saved for subsequent use in the classification experiments, see Chapter 5, Section 5.5.

Ensemble selection using a correlation measure. A more sophisticated method of ensemble selection was adopted for the experiments in Chapter 6. Essentially the aim is to select a set of features which convey as much information with respect to stimulus class as possible, whilst at the same time ensuring that their mutual information is minimized, i.e. a feature is 'good' if its response is highly correlated to the class vector but not to the responses of other features in the ensemble. The problem of feature selection, therefore, can be reduced to finding a suitable measure of correlations between features, and between features and classes.

A feature selection procedure based on the Fast Correlation Base Filter (FCBF) (Yu & Liu, 2003b) which uses an information-theoretic correlation measure has been adopted here. The FCBF method addresses the twin problems of a) removing both irrelevant and redundant features and b) reducing the computational overhead of the search in high dimensional space. According to Yu & Liu there are two types of feature selection algorithm, which they refer to as *feature-weighting* and *subset-search.* The first evaluates the usefulness of individual features. This approach is fast but does not remove redundant features, i.e.

it retains features that provide essentially the same information as others already chosen. The second type of search, based on evaluating the usefulness of subsets of features (which is equivalent to an 'ensemble') has a high computational cost. The FCBF algorithm is an attempt to produce a feature selection algorithm that takes in to account redundancy of features as well as the usefulness of individual features at moderate computational cost. The correlation between a feature response (r) and the class vector of stimuli (s), called *symmetrical uncertainty* (SU) by Yu & Liu, is defined in Equation 4.5.

$$SU(s,r) = \frac{2 \cdot IG(s|r)}{H(s) + H(r)} \tag{4.5}$$

In this equation H is the entropy and IG is the information gain defined in terms of the entropy in Equation 4.6.

$$IG(s|r) = H(s) - H(s|r) \tag{4.6}$$

As has previously been mentioned (Section 4.2.1) the entropy calculation (Equation 4.1) requires the response to be symbolic, or if the response is numeric, then it must be made discrete. The discrete numeric values can then be treated as symbols.

The FCBF algorithm starts with a feature which is most correlated to the class vector and removes all *'redundant peers'* from the set, that is it removes all features that are closely correlated to the chosen one. The chosen feature is designated a *'predominant feature'*. This is then repeated with the next most highly correlated feature remaining and so on. With each new choice of predominant feature a great many redundant peers are eliminated and the algorithm halts when there are no more features to be considered.

The FCBF selection method is not highly parametric but is necessary to fix the threshold relevance value T_{su}, that is the value of correlation between the response and the class vector below which a feature might be regarded as irrelevant. Setting this value too high means that useful features are not considered for inclusion. Setting it too low means that a great many useless features are considered for inclusion which simply increases the time taken by the process to halt. For all subsequent experiments this was fixed at 0.59. This figure was arrived at after a number of trials to ensure that the algorithm chose $\approx$ 300 fragments as this was the target maximum ensemble size (n_E) for experiments. The decision to use 300 fragments was based on two considerations. First, the limitations on the computational power available for the experiments combined with size of the corpus of stimuli, required a compromise to ensure manageable run times; of the order of tens of hours not hundreds of hours. Second, we hoped to analyze various properties of the chosen fragments in order to characterize them and make a comparison with properties of STRFs estimated from physiological measurements (see Section 4.4). To this end, it was important not to have too few fragments but to ensure that a broad range of useful fragments was included in the ensemble.

The FCBF algorithm used in this way gives a repeatable ensemble selection method from a library of fragments given the value for T_{su}, which in any case determines only the size of the resulting ensemble and not the order of the fragments chosen.

4.3.3 Event detection and responses vectors.

The basis of event detection in the model presented here is the presence of a coherent response across an ensemble of feature detectors. The response of a single fragment, or ensemble of fragments is calculated (see Section 3.4.2, Equation 3.10) and the summed response over all fragments normalized such that the maximum is unity for the range of stimuli in the experiment. In the experiments reported here the window corresponding to an event was defined as the period during which the summed response exceeded 0.2 i.e. 20% of the maximum response. The response vector that characterizes this event is formed from the maximum values of each individual response within the event window, this process is summarized in Figure 4.3. Note that the values of the elements in the vector are rounded up to two decimal places effectively placing each in to one of one hundred equally spaced bins.

This method provides the basis for an asynchronous, stimulus–ensemble driven event detector which triggers a readout of the population response pattern within a time window, the length of which is determined by the duration of the coherent ensemble response. The result is a short time scale context for the extraction of a pattern of responses that characterizes a distinct auditory event. This idea has its roots in work by Erickson (1986) and the notion of *neural mass*, which is in effect a measure of the vigour of the neural response. Erickson argued that this coded for the intensity of the stimulus, but in the current context this idea is equivalent to a saliency map in the temporal domain (Koch & Ullman, 1985), where the signal is analyzed locally with respect to a range of properties (the ensemble response) and the results integrated, in this simple model, by summation.

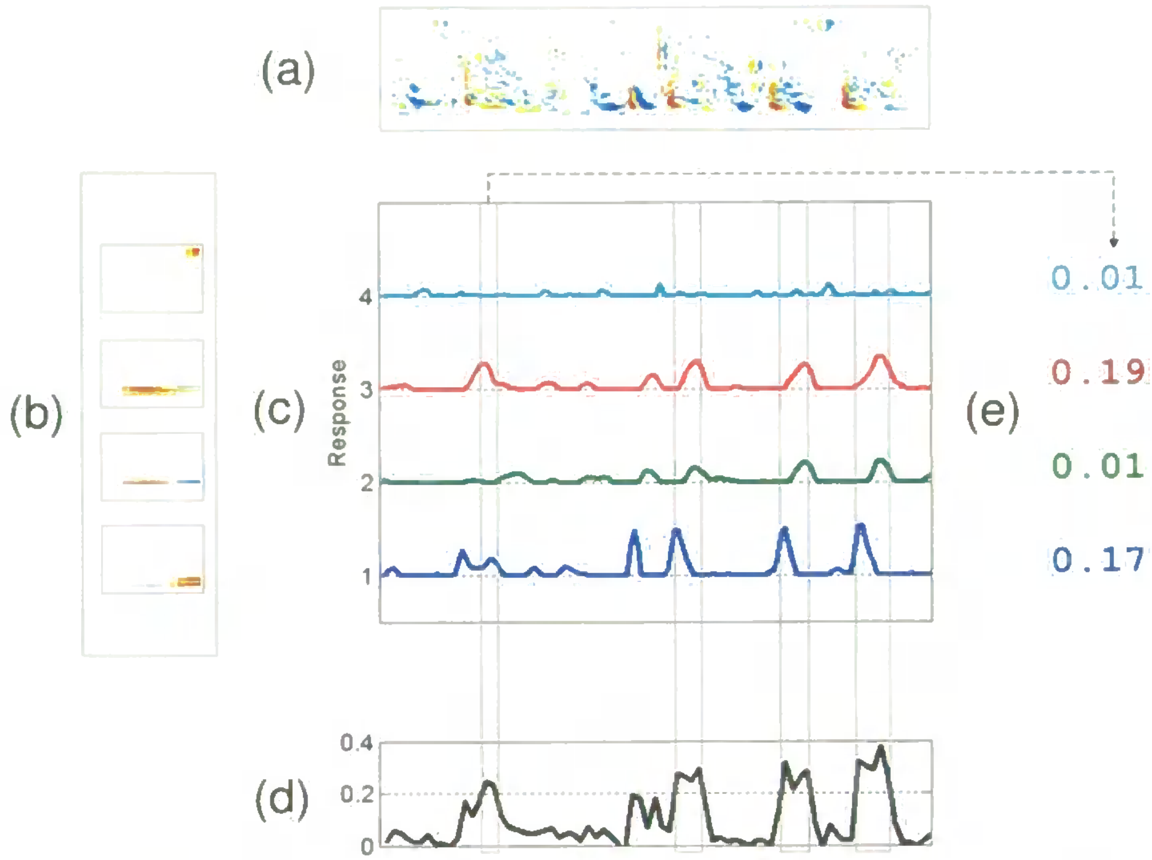


Figure 4.3: *This figure shows; (a) The SKV representation of a sample speech stimulus* 'Once upon a time there was a girl called Cinderella', *female speaker;(b) An ensemble of four spectro-temporal fragments (these are not shown on the same horizontal scale as the stimulus); (c) The response of each of these fragments to the stimulus; (d) The summed response of all fragments in the ensemble with the* 20% *salience level indicated; (e) The numerical values of the first vector corresponding to the ensemble response to the first detected event.*

4.4 Results.

Random ensembles. We first investigated the entropy of responses from a large number of randomly chosen fragments from the four speech fragment libraries (i.e. those of 10, 50, 100 and 200 milliseconds). The aim was to investigate whether this provided a basis on which to prefer one time scale for fragments over another. Forty thousand ensembles from each library were selected, 10,000 each of ensemble size $n_E = 2, 4, 8, \text{and} 16$. The entropy of the response (H) of each was then calculated and assigned to one of seven bins: $H < 0.5$, $0.5 \leq H < 1.0$, $1.0 \leq H < 1.5$, $1.5 \leq H < 2.0$, $2.0 \leq H < 2.5$ $2.5 \leq H < 3.0$ $3.0 \leq H \leq$ Maximum ($\log_2(11) = 3.46$) The probability of an ensemble, composed of each number and extent of fragment, appearing in each of these seven bins was then calculated and the results are shown in Figure 4.4.

This shows that fragments of $100ms$ extent are more likely to be present in the group of highest entropy for all ensemble sizes investigated, including $n_E = 2$ which is difficult to see in Figure 4.4 due to the scale. Further results which suggest that fragments of maximum extent $100ms$ might be preferred is provided in Section 5.4.3 based on their performance in a classification task.

Ensembles chosen using a correlation based filter. The FCBF algorithm was allowed to run to completion using $T_{su} = 0.59$ which yielded an ensemble of 303 feature extractors which are shown in Figure 4.5. For the sake of compactness in the figure, only fragments 1-100 and 201-300 are shown. The properties of the chosen ensemble were then analyzed in order to compare them to the properties of the complete set of all $100ms$ fragments. The best temporal modulation (BTM) was calculated for all fragments and the distribution of BTMs plotted

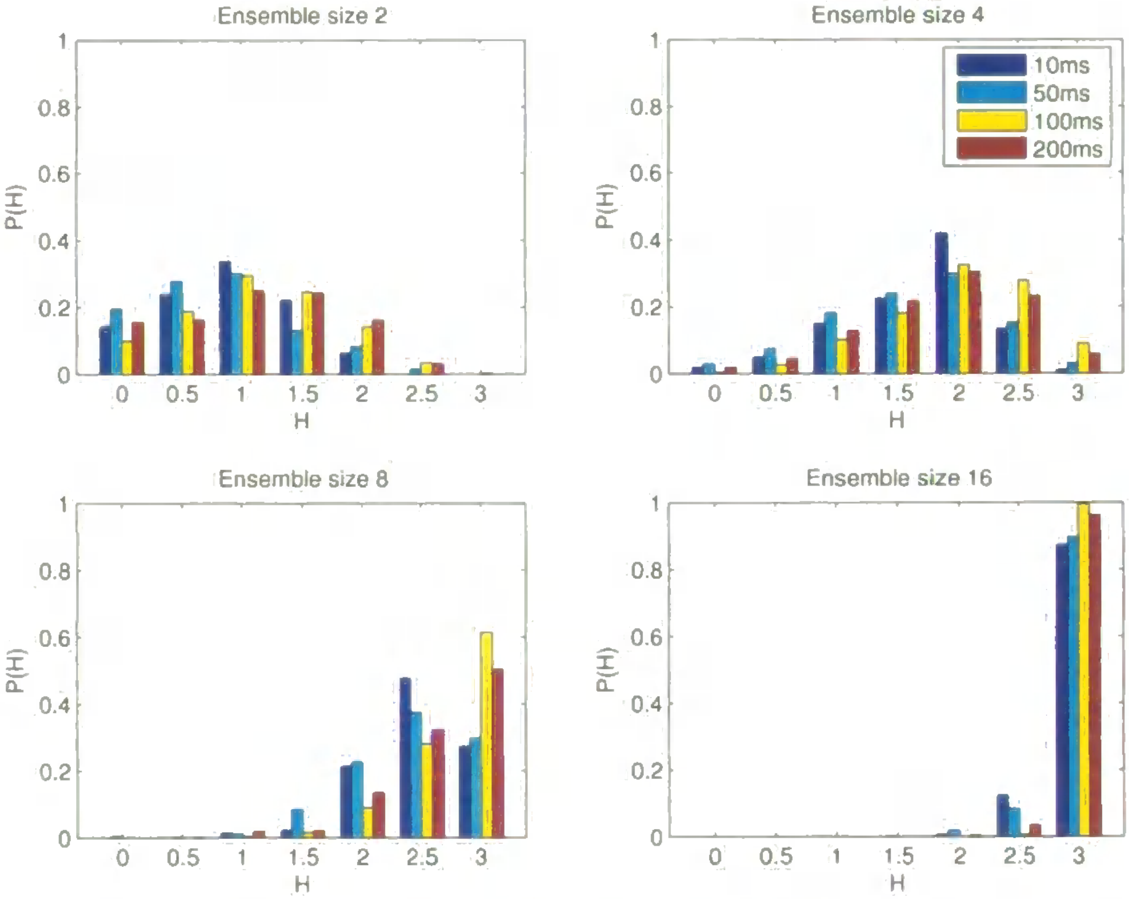
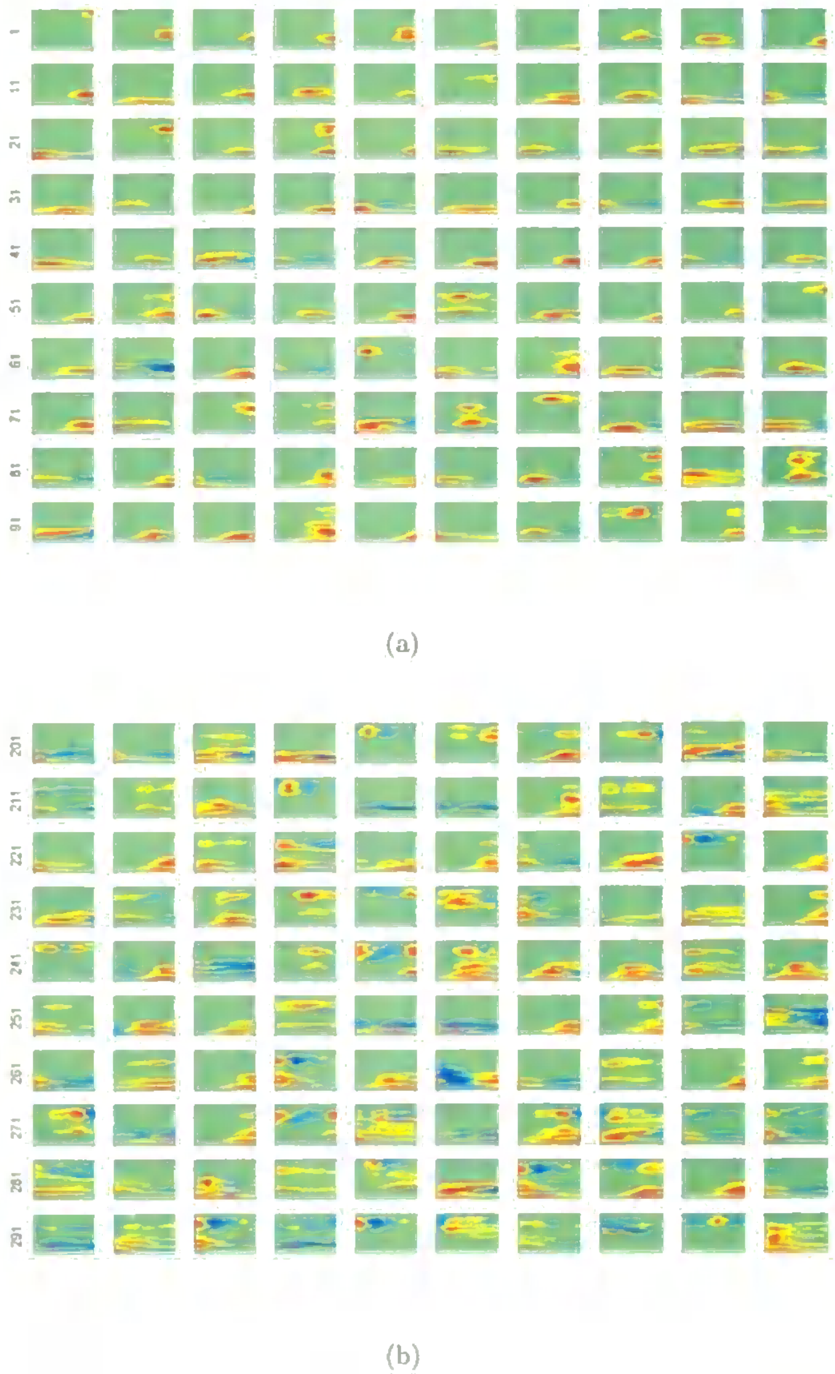


Figure 4.4: *Distribution of response entropy of different ensemble sizes from each of the four fragment libraries. The figures show the lower limit of the seven entropy bins (abscissa) and the probability of finding a random ensemble in each of the bins - for each temporal extent, colour coded (ordinate). For each ensemble size there are more ensembles of* $100ms$ *fragments in the highest entropy bin.*



(a)

(b)

Figure 4.5: *Example of ensemble chosen using FCBF. (a) Fragments 1-100. (b) Fragments 201-300. The lower numbered, most highly informative fragments have mostly very simple spectro-temporal characteristics in common with most of the estimates of STRFs in ferrets (Section 3.4.3). The higher numbered, less useful fragments have more complex distributions in time and frequency.*

for the entire fragment set and the ensemble chosen by the FCBF algorithm, see Figure 4.6. This shows that the ensemble has a greater proportion of fragments with BTMs in the range 10 to $15Hz$ than the fragment set from which they were drawn. The predominance of fragments with BTMs $\approx 10Hz$ in the complete fragment set is, of course, explained by the fact that the fragments all have a length of $100ms$. However the enrichment of the ensemble with fragments in the $10 - 15Hz$ is further justification for choosing fragments of this length and evidence that features on this scale are useful.

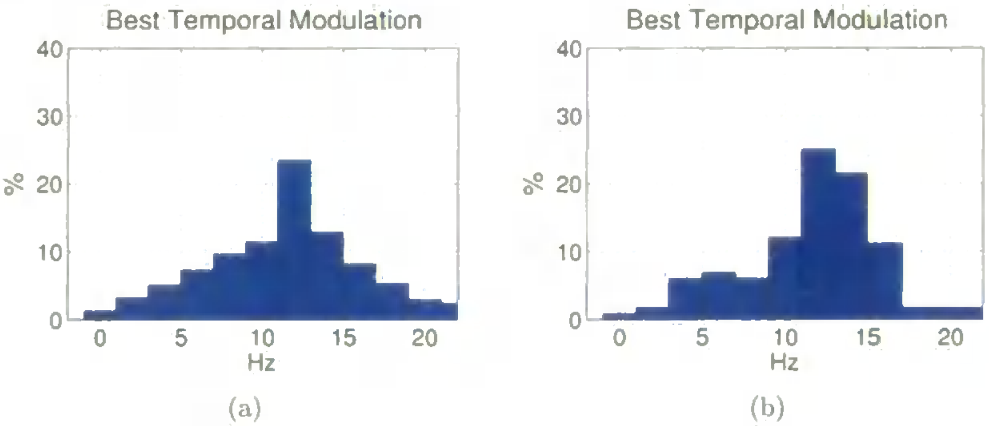


Figure 4.6: *Characterization of (a) the complete library of* $100ms$ *fragments,* $n \approx 22,000$ *and (b) an example fragment ensemble chosen using the FCBF algorithm* $n_E = 303$ *in terms of the best temporal modulation. The distribution of best temporal modulations in the chosen ensemble has changed with respect to the distribution found in the library of fragments from which it was drawn.*

Comparison with BTMs of cortical STRFs *in vivo* is complicated by the lack of human results. However, results from cortical STRFs in cats reported by Miller *et al.* (2002) show a peak in BTMs at $\approx 12Hz$.

4.5 Discussion.

4.5.1 General.

Two methods of selecting fragments, or kernels, to form an ensemble were explored, both based on information theoretic principles. The first simply combines fragments at random in small ensembles. The results (Figure 4.4) show that even for these small ensembles the response of some proportion of them is interesting, in the sense of having high entropy. Also, that in the highest entropy group, fragments of 100ms occur more frequently than those of 10ms, 50ms, or 200ms. This is a result that will be returned to in Chapter 5. In the case of larger ensembles chosen using the FCBF algorithm, Figure 4.5 shows that the most informative fragments consist of simple onset sensitive patches with moderately well defined best frequencies. These are not unlike those found in auditory cortex of ferrets in Section 3.4.3. Only one of the first 50 (number 24) has an obvious bi-modal spectral distribution. In contrast the higher numbered fragments have, on the whole, much broader spectral sensitivity and more complex geometries.

4.5.2 Biological implications.

It has long been assumed that one of the principle goals of neural coding must be to efficiently preserve information about environmental stimuli (e.g. Barlow, 1960). It is also a widely held view that this information is conveyed in patterns in an ensemble of neural responses, and that the ontogeny of these responses is in some way influenced by formative experience. This is a complex issue (for a review see Illing, 2004) but it has been shown that the development of auditory cortex is dependent on exposure to a 'normal', and 'rich' set of environmental sounds and that abnormal auditory environments produce long term atypical cortical organization (Zhang *et al.*, 2001, 2002). It has also been shown that the mechanisms of auditory perception remain sufficiently plastic to compensate for early organizational disarray (Wang, 2004) provided the subsequent stimuli are 'rich'. This result strongly suggests not only that formative stimuli could be the source of the patterns found in spectro-temporal descriptions of neural responses, but that this process remain plastic for some proportion of an organisms life after the early post natal period.

4.5.3 Future work

The investigations in this chapter represent the first steps in addressing the following questions. Can useful response fields be derived from a limited number of sounds? What temporal and spectral extent of acoustic feature is best for extracting meaningful information from the rich variety of ethological stimuli? In Chapter 5 these questions are revisited and we investigate if the ensemble response preserves information with respect to stimulus class.

A further question; 'what is the neural substrate which implements the spectro-temporal weighting which leads to STRF measurements?' may be answered by greater understanding of cortex, or thalamus, or more likely in thalamo-cortical interactions. Work on modelling such micro-circuitry is already in progress.

Chapter 5

A model of word classification using fragments.

5.1 Overview.

Principle aims. The work in this chapter develops and extends that in the previous chapter on the responses of ensembles of spectro-temporal fragments. We aim to show that vectors derived from the response of an ensemble of spectro-temporal fragments to a large number of stimuli contain information about the class to which each stimulus belongs. We can obtain a lower bound for the mutual information between the classes of the stimuli and responses using the method in the previous chapter only if we also assign the responses to a class; we do this with an artificial neural network (ANN). We aim to investigate how this measure changes if the fragments are derived from stimuli that are dissimilar from those used in the classification experiments. We also investigate how this changes as the stimuli are distorted by time compression without spectral modification.

Motivation. First, we wish to confirm that the responses of the chosen ensemble are not just 'interesting' in the sense that they are chosen on the basis of their entropy, but that they convey useful information about the classes from which the stimuli were drawn. We also hope to show that this generalizes to large numbers of stimuli that are different examples of the same classes. The stimuli used in these experiments are speech sounds and so this shows that the classification is robust to the variability inherent in speech. Second we investigate ensembles of fragments with different temporal extents using time compressed stimuli. These will be compared with psychophysical tests using similarly compressed stimuli. It is hoped that this might throw further light on the desirability of different time scales for auditory feature extraction. Last, as part of the overall aims of the project we wish to show that the projection in to a space spanned by the ensemble response is suitable as an input to a classifier based on a network of artificial spiking neurons. We hope to confirm that this is feasible by first using a conventional analogue ANN.

Achievements. Results show that the response vectors derived from a large corpus of speech stimuli contain information about the class of the stimuli when classified with an ANN. We also show that this information is reduced if the ensembles used have low entropy, or are chosen from fragments of formative stimuli that are dissimilar from speech.

Graphs showing how the mutual information changes with time compression of the stimuli (up to five times the original speed) show that fragments of $100ms$ show a falling off in performance similar to that reported in results from psychophysical experiments based on speech comprehension with similarly time compressed stimuli. Shorter and longer fragments do not show this effect. This supports the hypothesis, introduced in the previous chapter, that the time scale of feature extraction for speech stimuli could be $\approx 100ms$

5.2 Difficulties in models of auditory processing.

Models of auditory processing, particularly of speech, face many difficulties. Included in these are variability among speakers, variability in speech rate, and robustness to moderate distortions such a time compression. In this chapter we construct a model based on ensembles of feature detectors derived from fragments of an onset sensitive sound representation. This method is based on the idea of 'spectro-temporal response fields' (Chapter 3) and uses convolution to measure the degree of similarity through time between the feature detectors and the stimulus. The output from the ensemble is used to derive segmentation cues and patterns of response which are used to train an artificial neural network (ANN) classifier. This allows us to estimate a lower bound for the mutual information between the classes of the inputs and the classes of the outputs. The results suggest that there is significant information in the output of the system, and that this is robust with respect to the exact choice of feature set, time compression in the stimulus, and speaker variation. In addition the robustness to time compression in the stimulus has features in common with human psychophysics. Similar experiments using feature detectors derived from fragments of non-speech sounds performed less well. This result is interesting in the light of results showing aberrant cortical development in animals exposed to impoverished auditory environments during the developmental phase (Zhang *et al.*, 2002) and reinforces the hypothesis that spectro-temporal characteristics of cortical responses reflect the patterns in ethological stimuli. A speech database was used containing a large number of examples of a small number of classes (words). Response fields were constructed from acoustic fragments of varying temporal and spectral extent extracted from the utterances of a single speaker. These response fields were then convolved with utterances from a large number of different speakers, and the mutual information between the ensemble response and actual stimulus class was characterized. In this way the robustness of the approach to the variability inherent in speech, and to the later inclusion of novel classes, was assessed. The sensitivity of the system to the choice of fragments within an ensemble, to abnormalities in early experience, and to temporal manipulations of the stimuli was also investigated. On the basis of the experiments described below it is concluded that useful response fields can be derived from a limited number of sounds. Although it is important that individual members of an ensemble convey different information, provided that there are sufficient response fields, their precise form is not critical and a rather small number of response fields acting in parallel can convey class information. Finally, if the effects of temporal manipulations were taken into account, it was found that acoustic fragments of intermediate temporal extent conveyed class information most effectively, in line with the findings for visual object classification.

5.3 Second order isomorphisms.

It is part of our every day experience that our interpretation of the world through sensory data is flawed; in vision, for example, we are all familiar with various optical illusions. If the correspondence between the physical world (the *distal* stimulus) and our internal perceptual order (the *proximal* representation) were exact then our internal representation could be described as *veridical* or truthful (Edelman, 1998). In the absence of a veridical representation it is at least necessary to achieve a representation that is *principled* in that it manifests some standard of rightness.

It has been proposed that organisms achieve a principled representation not by the closeness of comparison between stimuli and a set of internal archetypes, i.e. *representation-by-similarity* (an idea that can be traced back to Aristotle) but by building a representational space where similar objects occur closely together, i.e. a *representation-of-similarity* (Shepard & Chipman, 1970). Edelman has argued that such a representation could be implemented by measurement of similarity of the stimulus to a set of references, plus dimensionality reduction to achieve a low dimensional representational space. Within this space each object would have, at the very least, a unique position, i.e. it would be distinct. However, a representation of distinctness is not enough. To achieve a principled representation, objects that closely resemble one another in the distal space (objects that are similar) must also be close to one another in the proximal representation. In the limiting case where the identity of the kth nearest neighbour of each point in proximal space is preserved for all values of k (assuming a finite number of objects) the correspondences in the distal space can be recovered in their entirety by examining those in the proximal space and we have achieved *similtude.* In this chapter we will look for evidence that such a second order isomorphism arises from the vector representation of stimuli derived from ensembles of spectro-temporal filters as described in previous chapters.

5.4 Methods.

Figure 5.1 summarizes the stages of the model from sound file to response vector and is a summary of the stages previously described. The first two stages are as described in detail in Chapter 2.

Spectral decomposition. The first stage approximates processing in the cochlea, Figure 5.1(b). Sounds are processed using a bank of 30 gammatone filters using the SCM representation as described in Section 2.3.1.

Transient extraction. The next stage of processing identifies envelope transients within each frequency channel using the SKV representation, Figure 5.1(c), for details see Section 2.3.2.

Convolution using an ensemble of STRFs. In the next stage we use an ensemble of spectro-temporal filters derived using the methods in Chapter 4. Each STRF in the ensemble is specified in terms of a pattern of onsets and/or offsets extending over a specified spectral range and duration. Each member of the ensemble of n_E STRFs is convolved with the pre-processed incoming signal, thereby generating a set of n_E 'temporal signatures', which indicate the degree of similarity between the incoming pattern and the STRF at each point in time. This is illustrated in Figure 5.1(d) for an ensemble of 16 STRFs. This was the

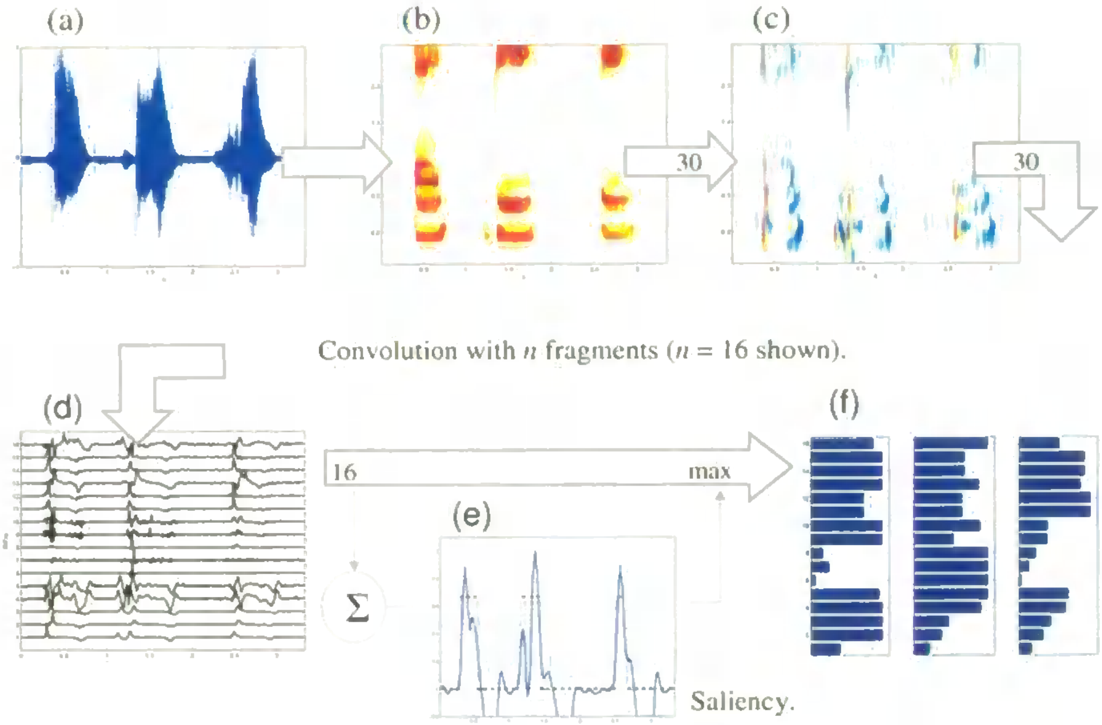


Figure 5.1: *Summary of processing stages in the model of sound classification. (a) Waveform, (b) SCM (Section 2.3.1, (c) SKV(Section 2.3.2), (d) Response of an ensemble of STRFs (Section 4.2.6), (e) The summed response of the ensemble used for event detection (Section 4.2.6), and (f) The resulting response vectors (Section 4.2.6).*

maximum ensemble size for results in Section 5.4.2 and also the ensemble size for the results in Section 5.4.3.

Event detection and mapping to response space. The same procedure is used here as is described in Section 4.3.3. In brief, the selected ensemble of fragments (from Section 4.3.2) represent our candidate STRFs and each is specified in terms of a pattern of onsets and offsets extending over a specified spectral range and temporal duration. The response of each to any stimulus, or set of stimuli can be calculated by convolving the candidate STRFs with the SKV representation of the incoming sound, thereby generating a 'temporal signature', which indicates the degree of similarity between the incoming pattern and the STRFs at each point in time. The summed response of all STRFs in the ensemble provides an indication of the presence of an acoustic event, the timing and duration of which is determined both by the stimulus *and* by the ensemble used. This process is illustrated in Figure 4.3. It is possible for a single stimulus to generate more than one such event, but in the experiments described below when this occurred only the first event was characterized.

5.4.1 The analogue artificial neural network.

Topology. The purpose of the analogue artificial neural network (analogue ANN) was to act as a classifier. The architecture of the ANN is shown in Figure 5.2 and consisted of three layers. The first layer has one input node for each element of the response vectors, this is number of STRFs in the ensemble n_E, see Section 5.4. In Figure 5.2 $n_E = 4$ nodes are shown (the results reported in Section 5.5.2 use $n = 16$, and those in next chapter use $n_E = 303$). The second layer consists of pools of five nodes, one pool for each class. Figure 5.2 shows 2 such pools (there are 11 classes in the results reported in Section 5.5 so 55 nodes in 11 pools). These nodes are fully interconnected with the input layer, not all connections are shown in the diagram. These nodes have log-sigmoidal transfer functions. The third layer consists of one node for each class with connections to all of the nodes in the appropriate pool. These nodes are also log-sigmoidal.

Training and testing. The response vectors derived from all stimuli were divided in to three subsets obtained by combining random choices from each of the classes. For the training phase two sets were used; the 'training set' (60% of

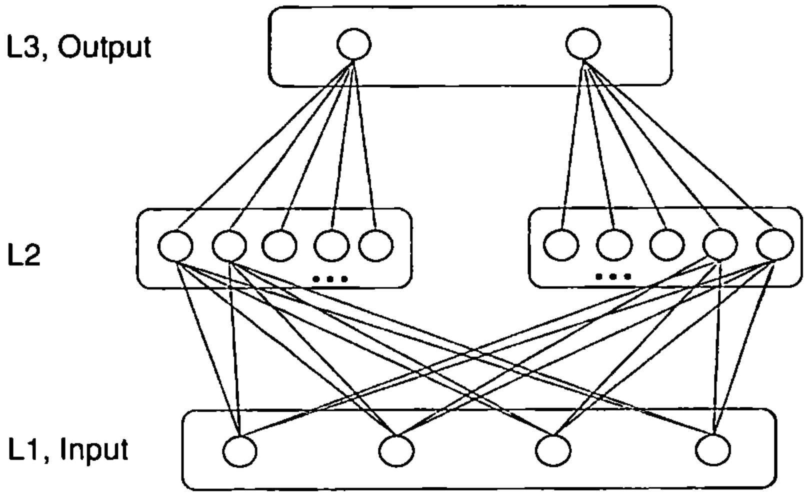


Figure 5.2: *The analogue artificial neural network. The input layer at the bottom of the diagram consists of one node for each element of the response vector to be classified, 4 are shown. The central layer consists of pools of 5 log-sigmoidal nodes fully interconnected with the input layer. There is one pool for each class. The upper, output layer consists of one log-sigmoidal node per class connected only to the appropriate pool.*

stimuli), and the 'validation set' (15% of stimuli). 25% of stimuli were withheld for testing in each experiment (the 'test set'). To obtain error bars this process was repeated 10 times. Each of the 10 training and validation subsets were used to train the network 10 times from random starting weights, hence 100 results in total. In all cases the subset population was balanced with respect to the 11 classes. Each pool of five units in the central layer was trained independently by setting the target values in layer 3 to unity or zero depending on whether stimulus was a member of the class represented by the pool.

After training the network was presented with each of the vectors of the test set and the class recorded as the number of the output unit with the highest value in a 'winner takes all' fashion. These results were used together with the known classes of the stimuli to establish a lower bound on the mutual information between the input and output (see Section 4.2.1).

5.4.2 Generalization.

First, the generalization capabilities of the proposed approach was investigated, and the ability of the model to cope with the natural variability in speech sounds considered. Ensembles of STRFs derived from fragments of speech (Section 4.3.1) and chosen using the FCBF algorithm(Section 4.3.2) were used to derive response vectors from stimuli consisting of spoken digits. The corpus represented over 300 male and female speakers and used recordings with signal to noise ratios between 8 and $25dB$. The mutual information between the classifier output and the stimulus class was calculated. The results are plotted in Figure 5.4.

5.4.3 Robustness to time compression.

It has been shown that for long stimuli thalamo-cortical responses adjust their response frequencies to that of the stimulus (Ahissar *et al.*, 2000) making it essential that stimuli fall within the effective operational range of these circuits. Evidence to support this view has come from the result that poor comprehension of speech that has been time compressed appears to correlate closely with the inability of responses in auditory cortex, as measured using magnetoencephalographic (MEG) responses, to 'lock' to the stimulus envelope amplitude and phase (Ahissar *et al.*, 2001). We were interested to to see if this effect was visible in our model using single word stimuli and response fields of different temporal extents. We repeated the generalization results using the same speech stimuli after time compression (without distortion of the spectral or pitch content) and compared the results with those reported by Ahissar *et al.* (2001) see Figure 5.5.

5.5 Results.

5.5.1 Generalization results.

In this sequence of experiments fragment lengths were limited to a maximum of $200ms$ as this represents the approximate length of a syllable in normal speech, and the maximum ensemble size chosen was $n_E = 16$. Results for large ensemble sizes up to $n_E = 128$ were obtained for fragment lengths of $100ms$ and $200ms$ only and these are shown in Figure 5.3. These show a trend towards smaller increases in mutual information with increasing n_E and so for computational convenience n_E was limited to 16. The results of preliminary investigations comparing fragments ensembles of different sizes, fragments of different temporal extents, and fragments from speech and non-speech sources are are shown in

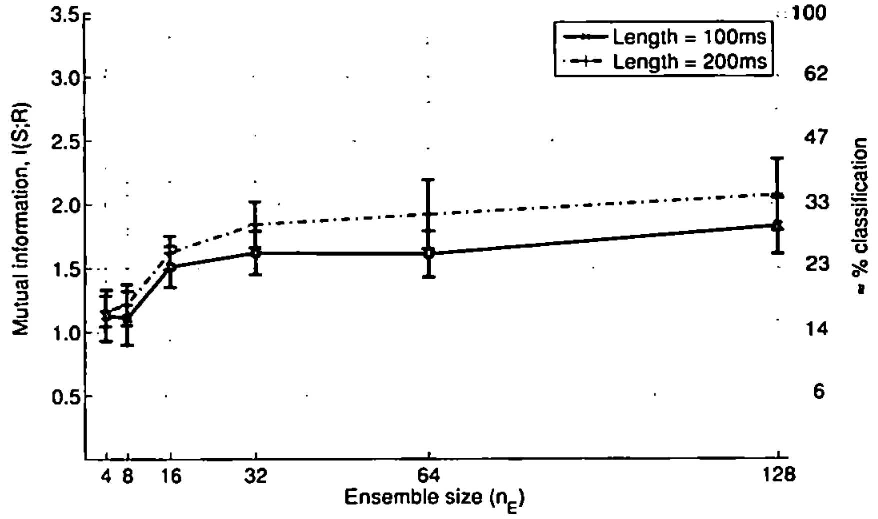
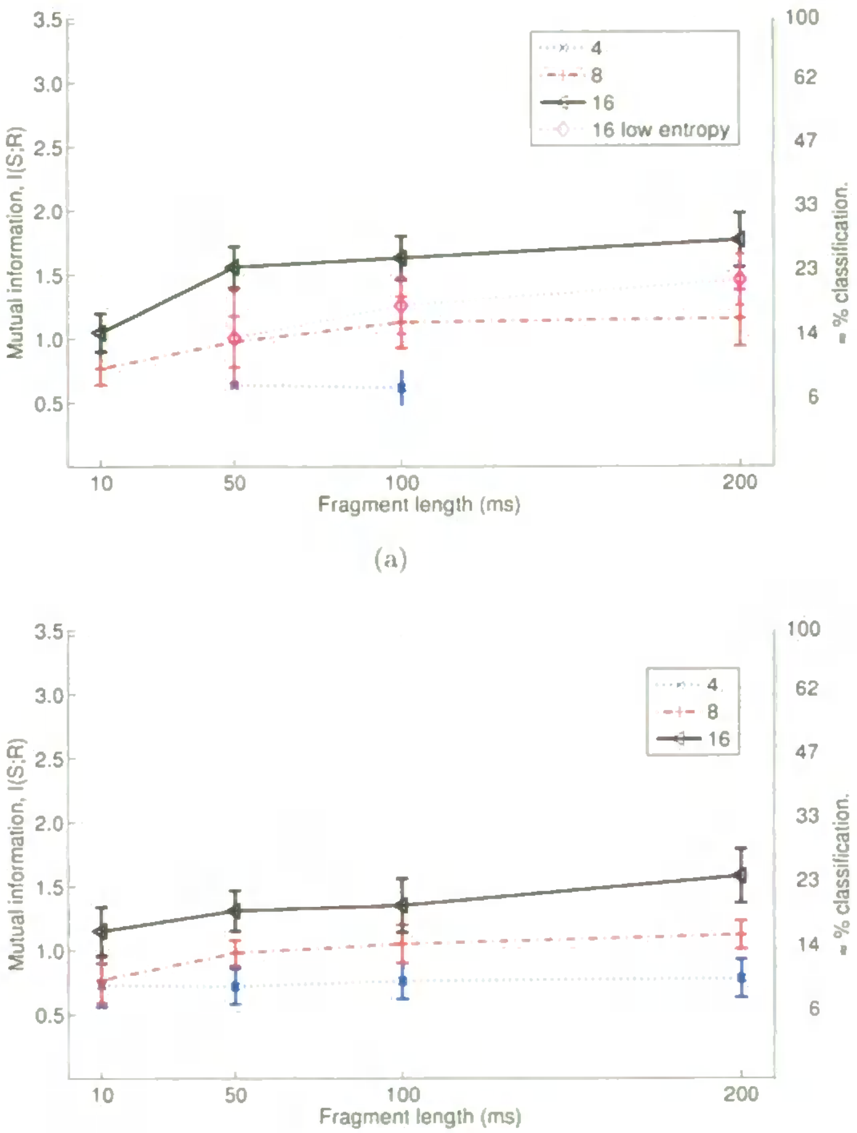


Figure 5.3: *This figure shows the performance of larger ensembles up to $n_E = 128$. It is clear that the performance continues to improve with ensemble size although the performance gain is less for each new member of the ensemble. The decision was made to limit the ensemble size to 16 for initial experiments to reduce computational overhead, see Section 5.5.1.*

Figure 5.4.

It was found that there is significant mutual information between stimulus class and model classification, and that mutual information improves with fragment duration up to 200ms, and with the number of fragments in the ensemble up to $n_E = 16$. This suggests that there is some form of clustering which is robust to the variability present in normal speech. Included in these results are data from the 'low-entropy' ensembles (see Section 4.3.2) showing that these perform less well on generalization.

In order to discover whether the model was very sensitive to the precise nature of the formative stimuli from which the fragments were formed, results from ensembles of STRFs derived from the set of environmental noises (Section 4.3.1) were derived and once again we trained the system to classify the digit utterances. The results, also shown in Figure 5.4, are very interesting. Their performance, although lower than that of the speech fragments, is comparable. For example in the case of $100ms$ fragments, not less than 15% lower. This suggests that the classification of sounds on the basis of projections into a response space spanned by a set of STRFs, is perhaps surprisingly, not very sensitive to the precise nature of the receptive fields used. However, the fact that they perform less well is consistent with experimental findings (Chang & Merzenich, 2003; Zhang *et al.*, 2001) showing that in an extremely restricted early auditory environment the auditory cortex fails to develop properly. This result also establishes the 'productivity' of the system in that the responses of fragments can be used to classify sounds very different from the ones from which they were derived.



(a)

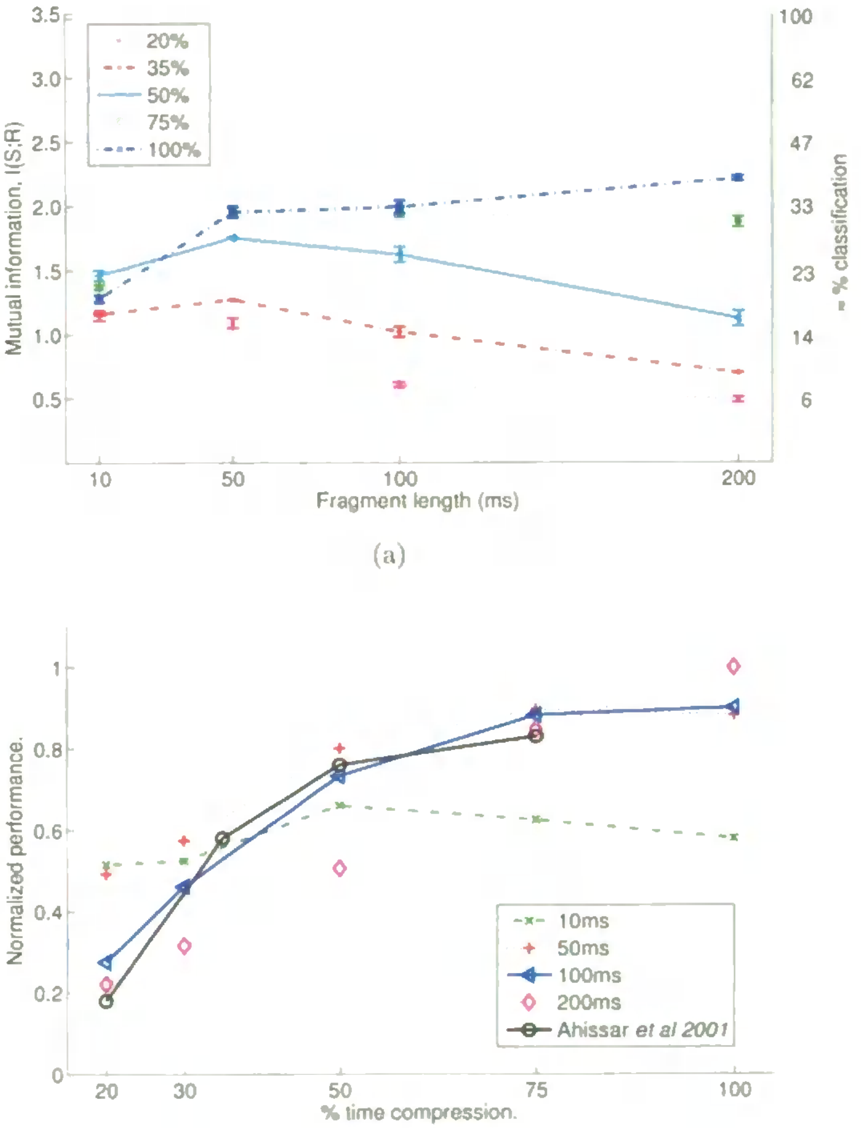
(b)

Figure 5.4: *Mutual information between classifier output and stimulus class for fragment ensembles of size 4 (blue), 8 (red) and 16 (black), and varying temporal extent (abscissa). This is plotted for both (a) speech and (b) non-speech fragments. Dotted line on (a) shows results for 'low-entropy' ensembles of 16. Performance for speech fragment is between* 15 *and* 20% *better than non-speech fragments.*

5.5.2 Results with time compressed stimuli.

Figure 5.5 summarizes the results of the experiments using time compressed speech. The upper sub-figure shows that for longer fragments, which perform well with un-compressed speech, the performance penalty for compressed speech is great. For shorter fragments which perform less well overall their performance suffers less as the stimuli are compressed. To make this clear, the data are re-plotted in the lower sub-figure with the degree of compression in the stimuli rather than the fragment length on the abscissa and the best performance normalized to unity; the error bars are omitted for clarity. These show that the $100ms$ and $50ms$ fragments, although they perform less well than the $200ms$ fragments on uncompressed speech, perform best in the 75% and 50% compressed experiments. In a recent experiment it was found that speech comprehension measure in human psychophysics is quite robust to compression up to about 50% and thereafter degrades quickly as time compression increases to 20% (Ahissar *et al.*, 2001). These results are plotted for comparison in Figure 5.5(b) (black line).

The authors of this work found that whenever the comprehension of time compressed speech was degraded so was the phase locking of the speech envelope to the cortical response as measured by MEG. One possible explanation for the phase-locking could be the degree to which the STRFs in auditory cortex are able to respond to incoming spectro-temporal patterns, i.e. the observed phase-locking may simply be a by-product of the degree of similarity between the STRFs of cells in auditory cortex and the spectro-temporal pattern of the sounds. The performance of the model in this experiment is consistent with the suggestion from the previous chapter that fragments with temporal extent between 50 and $100ms$ correlate best with human performance. This is also consistent with the



(a)

(b)

Figure 5.5: *Mutual information between stimulus and response class using time compressed speech. (a) Plotted against fragment length and (b) plotted against time compression. The* $200ms$ *fragments perform best with uncompressed stimuli but their performance drops quickly with successive compressions. The fragments of* $100ms$ *have a performance curve that most closely matches the results of human psychophysics with compressed sentences from Ahissar* et al. *(2001) (solid black line).*

suggestion that the phase-locking is best within the range of spontaneous and evoked cortical oscillations ($\approx 14Hz$) i.e. a period of $70ms$ (Ahissar *et al.*, 2001).

5.6 Discussion.

5.6.1 General.

In this chapter we based our results on the idea, introduced in the previous chapter, of using a convolution as a measure of similarity between a stimulus representation and an ensemble of roughly tuned, spectro-temporal detectors. Using this approach sounds were represented by the patterns of activity present within the ensemble during a time window. The response of the ensemble can be understood as a projection into a low dimensional space spanned by the outputs of the detectors. The ANN classifier serves to label the output as belonging to a single class, and hence acts as a mechanism for estimating the lower bound of the mutual information between the stimulus and response classes. The transformation in to low dimensional space may be understood as essentially a second order isomorphic mapping, which may be organized in a hierarchical fashion to extend to longer duration stimuli.

Importantly, although performance is robust to the precise choice of filters we have proposed, in the previous chapter, a basis for preferring some fragments and some ensembles over others. This is based on the entropy of their responses to the formative sounds (Figure 4.4) and ensembles that had low entropy responses performed poorly (Figure 5.4(a)) in the generalization tests.

We have also found evidence that, given that the formation of the candidate STRFs was stimulus driven, the mutual information between the input and output classes is greater if the formative stimuli were to some extent representative of the sounds to which the system is subsequently exposed. In these experiments STRFs were abstracted from a limited set of speech and non-speech sources, however, because the non-speech sources were quite 'rich' (not just tone bursts for instance) they still produced significant information preserving representations. There is a parallel here with the formation of auditory cortex based on the early auditory environment.

Crucial to the success of this model is the ability to segment the incoming stimulus, that is to identify the salient auditory events which provide the basis for classification. To achieve this we did not use properties of the signal, or properties of the onset sensitive representation of the signal, but properties of the response of an ensemble of detectors. This makes the segmentation dependent on the choice of detectors and provides a mechanism whereby segmentation can become an active part of of the perceptual process under adaptive control and has biological implications.

5.6.2 Biological implications.

We have, in an earlier chapter introduced the idea of using patterns of activity in our ensembles of candidate STRFs to provide a window representing a salient auditory event. In this chapter we also use the level of activity in the ensemble during this window to characterize this event. This results in features being integrated over a 'context' period (Nelken *et al.*, 2003) provided by the event window and we have shown that this preserves information about the stimulus class. In speech some form of segmentation is necessary to match discrete percepts and to make speech perception robust to rate variation. In the absence of interference humans can also do this in the presence of cross channel asynchrony (Arai & Greenberg, 1998) i.e. they are capable of integrating cues identifying speech sounds which do not occur simultaneously. This could be as a result of using the context period, or window length, provided by an event detection method based on the response of a range of auditory feature detectors. It is clear that patterns of outputs from STRFs in auditory cortex could be used for event detection, segmentation, and classification by one or more of the many areas of the brain to which it is connected.

It has been reported that the formation of response properties in auditory cortex is dependent on the richness of the early auditory environment, (e.g. Zhang *et al.*, 2001), and that the cortex fails to organize effectively if it is subject to an impoverished, or aberrant set of stimuli. This could be partly explained if the spectro-temporal patterns abstracted during the developmental period are based on the formative stimuli themselves. The model results presented here suggest that features that provide useful responses to one set of stimuli might not generalize so well to stimulus sets that have different statistics. In the extreme case of formative stimuli that consist of broadband noise (Chang & Merzenich, 2003) few features useful in distinguishing a range of ethological stimuli can be abstracted.

The results using time compressed speech stimuli reflect results from human psychophysics when features are extracted on time scales of $\approx 100ms$ (Figure 5.5). This experiment suggests that, as in vision (Ullman *et al.*, 2002), fragments of intermediate extent may be optimal. This temporal extent is intermediate in the sense that it falls into an range between phonemes of roughly $40ms$ and syllables of typically $200ms$. It is considerably longer than the acoustic models typically used in automatic speech recognition systems.

Although it may be true that features extracted on this time scale provide robustness to the variation in the stimulus rate, an alternative view is that the degree to which our perception is robust to time compression may be limited by the extent of STRFs. Rates of spontaneous and evoked cortical oscillations may help to explain the psychophysics (Ahissar *et al.*, 2001) and the temporal extent of cortical STRFs by establishing the perceptual time scale on which auditory events are identified. Although there is no electrophysiology from humans, STRFs of $\approx 100ms$ are broadly consistent with results from animals such as mice (Linden *et al.*, 2003), rats (Machens *et al.*, 2004) and ferrets (Fritz *et al.*, 2003).

5.6.3 Future work.

In pursuing experiments based on patterns of onsets and offsets based on the SKV representation we are making a judgement, based on extensive evidence, about the predominant nature of sub-cortical auditory processing. However, we have not established whether the performance of the system in terms of mutual information, or percentage classification, would be better or worse if the same process were applied to patterns based on an assumption of tonic firing, i.e. fragments of cochleagraphic or short term fourier transform descriptions of the stimulus. It has certainly not been established that the energy within a tonotopic region is *not* available beyond the auditory periphery. In this light further effort should be made to repeat results using these alternative approaches and to establish if performance could improved if a mixture of representations were used in parallel.

The parallels with results from Ahissar *et al.* (2001) provide some supporting evidence for the hypothesis, introduced in the previous chapter, that useful features of speech may be extracted on time scales of $\approx 100ms$. However these results were gained on isolated utterances and it is clear that in order to support this argument fully these experiments should be repeated using continuous speech stimuli.

Chapter 6

Multiple, concurrent classifications.

6.1 Overview.

Principle aims. In this chapter we seek to establish whether the response vectors, as derived in previous chapters, can be used with more than one classifier to derive multiple class information from a stimulus. We also aim to introduce here, as one of the stated aims of this work, results obtained from simulations of networks of artificial spiking neurons.

Motivation. Any model, however simple, of auditory classification (or perception) must ultimately address the fact that organisms can derive more than one type of information from a single stimulus. For a human being listening to speech, for example, information is available not only about which words are being used, but also additional prosodic information (e.g. is it a question or a statement?) and additional information about the speaker (male or female? - large or small?).

The current hypothesis is that these judgements are handled by multiple 'what' pathways and there is evidence to support this. What is not clear is whether our representation is capable of supporting multiple classifications.

Achievements. We show for the first time that the same responses derived from ensembles of spectro-temporal feature extractors are capable of supporting qualitatively different classifications including those based on pitch track distinctions and spectral profile as well as the more complex spectro-temporal course of the stimuli.

6.2 'What' pathways in auditory cortex.

Complex sounds can be perceived in a number of qualitatively different ways. For example, voice communication conveys information that can be perceived independently of verbal content; this includes the speaker's identity, sex, emotional state *etc.*, as well as semantic information such as whether the utterance is a question or a statement. Judgements of this type are collectively known as 'what' judgements, that is judgements of class to distinguish them from 'where' or localization judgements. These two types of judgement are, it seems, functionally segregated in specialized streams for auditory, visual and somatosensory stimuli (Alain *et al.*, 2001; Haxby *et al.*, 1991; Kaas & Hackett, 1999; Pons *et al.*, 1992). There is also evidence to support the idea that different 'what' judgements are made in spatially separated areas of the brain. It has been observed for example that lesions in the right temporo-parietal cortex impair speaker recognition, but not speech comprehension (e.g. Lancker *et al.*, 1989)). Also it has been found that although both superior temporal sulci (STS) are responsive to voice stimuli, the right anterior STS is not involved in processing the verbal content of speech (Belin *et al.*, 2000, 2004; Kriegstein *et al.*, 2003). These results are also consistent with the recent finding, using MEG, that there is differential task-dependent modulation of parallel processing maps within the auditory 'what' pathway in phonological and speaker identity classification tasks (Obleser *et al.*, 2004). In this paper results which suggest that extraction of different abstract invariants, specifically phonological information and speaker identity, take place in spatially separated areas of the neural substrate. The relative activation of these areas changes as attention is moved from one task to the other.

Since essentially all information about the acoustic world entering cortex passes through areas of primary auditory cortex (PAC), representations in PAC must be sufficiently rich to support a wide range of judgements, including identifying the source and nature of the stimulus. Higher centres in auditory cortex, with different functionality, could then subsequently abstract different properties for use in various aspects of object classification (Griffiths *et al.*, 2004). Nevertheless, the way in which sounds are represented and processed in primary auditory cortex remains controversial (Griffiths *et al.*, 2004). A significant problem, when it comes to understanding the processing of speech, is the lack of any data regarding the nature of receptive fields in human PAC.

It has been shown in Chapter 5 that ensembles of spectro-temporal response fields (STRFs) derived from speech stimuli can preserve information about utterance class. In this chapter further experiments on this putative model of processing in PAC establish whether the same representation can support multiple, qualitatively different, classifications. It should be stressed that it is not at all clear *a priori* whether such an ensemble of STRFs should be capable of extracting and conveying information useful for speaker identification, sex, or prosody classification. There is no clear understanding of how humans perform these tasks and they are all thought to involve pitch, a feature which is not explicitly represented in this model.

In Section 6.5 we present results obtained using the methods described in Chapter 5 and also from a slightly modified approach where the analogue ANN classifier is replaced with an ANN that is spike driven. The purpose of this is to meet one of the stated aims of the project; to demonstrate that the response vectors derived from stimuli using an ensemble of STRFs are suitable for spike rate encoding.

6.3 Methods.

6.3.1 The spike-Driven Network.

Response vectors derived from large numbers of stimuli (described in Section 6.4 below) were divided in to training, validation, and test sub-sets (see Section 5.4.1) and supplied to Joe Brader at the Institute of Physiology, University of Bern, Switzerland who encoded them as spike trains (Section 6.3.1), trained the network, and returned the output classes of the test set from which the mutual information and classification percentage results reported in this chapter were calculated.

The spike-driven network architecture used, described in more detail in Brader *et al.* (2004) and DelGiudice *et al.* (2003), consists of a single feed forward layer in which the input neurons are fully connected to the output layer by plastic synapses. Neurons in the output layer have no lateral connections and are subdivided into pools of equal size, each selective for a particular class of stimuli. In addition to the signal from the input layer the output neurons receive signals from inhibitory and teacher populations. The inhibitory population serves to balance the excitation coming from the input layer. The teacher population is active during training and reinforces the selectivity of the output pools by means of an additional excitatory or inhibitory signal. A schematic view of this network architecture is shown in Figure 6.1.

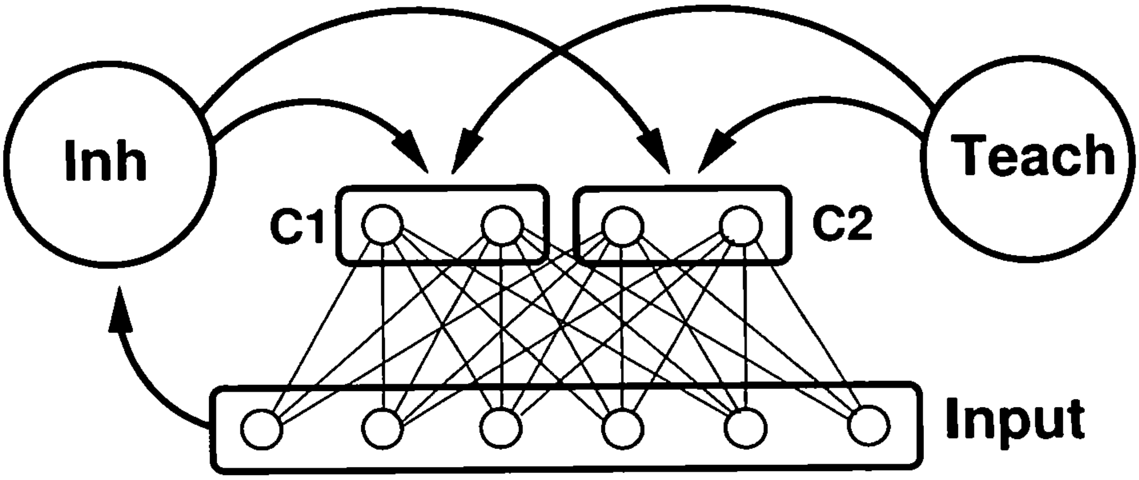


Figure 6.1: *A schematic view of the spike-driven network architecture. When considering two classes of stimuli the output units are grouped into two pools each selective to a given class. Additional signals are provided by external inhibitory and teacher populations.*

Learning within the network is spike driven, and takes place within the synapses using information local to each synapse. A novel bistable synaptic model (Fusi, 2002), designed to ensure memory maintenance on long time scales, while retaining sensitivity on short time scales, is used. This model takes advantage of the finding that memory capacity can be maximized by making stochastic rather than deterministic synaptic transitions (Amit & Fusi, 1992, 1994; Fusi, 2002). If the probability of these transitions is small then only a small fraction of the stimulated synapses is changed upon each stimulus presentation. This extends the memory span of the system and prevents it from forgetting previously learned memories too quickly. Furthermore, by exploiting the inherent irregularity of the input spike trains (Fusi, 2003; Fusi *et al.*, 2000), stochastic transitions between the synaptic states are easily achieved, making the model particularly suitable for VLSI implementation (Chicca & Fusi, 2001; Fusi *et al.*, 2000; Indiveri., 2002).

The particular synaptic dynamics we employ are designed to be Hebbian with an additional stop-learning mechanism which makes synaptic transitions increasingly unlikely if the response of the relevant output neuron becomes either too low or too high Fusi (2003) see (see Brader *et al.*, 2004, for a detailed description of the dynamics). Extreme responses are an indication that the output neurons have already learned to classify the stimulus, and that it is unnecessary to modify the synapses to improve the performance (Senn & Fusi, 2004). This modification enables the model to learn highly correlated input patterns.

Spike rate encoding of response vectors.

Each stimulus is encoded as a 128 element feature vector within which each element is a continuous value, ξ between zero and unity, thus there are 128 neurons in the input layer. When presented with a stimulus each input neuron emits a Poisson spike train at a rate 50ξHz. The output neurons are grouped into pools, one for each class, with 10 neurons per pool. Although the output neurons will all see the same input patterns, the stochasticity of learning will create different representations for each output neuron. A similar technique has been exploited in Amit & Mascaro (2001) where the authors use random receptive fields. 70% of the dataset was used for training and the remaining 30% for testing.

In order to assess the classification performance following training, a fixed frequency threshold was defined (the same for all output neurons); an output neuron was regarded as active or inactive depending upon whether it fired at a mean rate above or below this threshold when presented with a test stimulus. The class of the stimulus was then determined by counting the number of active neurons within each pool and finding the one which expresses the largest number of votes. This network architecture therefore allows for two possible types of error when presented with a test stimulus: (i) no output neurons express a vote and the stimulus is non-classified or (ii) the wrong output pool expresses the largest number of votes and the stimulus is misclassified. Non-classifications are preferable to misclassifications because the network simply expresses no preference and leaves open the possibility that such cases could be sent to subsequent networks for further analysis or that the stimulus is simply ignored.

6.4 Stimuli.

6.4.1 The ISOLET corpus.

The data used to obtain the results in this chapter were from the ISOLET corpus of spoken letter names from the Oregon Health and Science University (OGI, 1999). This consists of ≈ 8000 spoken letter names from 150 speakers, male and female. These data provided the basis for the digit classification results, the male/female classification results, and the question/statement classification results (all in Section 6.5).

6.4.2 The question/statement manipulation.

In English the primary cue which distinguishes a question from a statement is the pitch trajectory; questions have pitches which rise towards the end of the word or phrase, and statements ones which are flat or falling. The ISOLET corpus was pre-processed using PRAAT (Boersma, 2001; Boersma & Weenink, 2005) in order to manipulate the pitch tracks and to introduce a question or statement prosody. First, a time stretching algorithm was used to ensure that all stimuli had a standard duration of $500ms$. Next, the pitch tracks were adjusted using;

$$F_0(t) = \overline{f}_0.[1 + 0.3\sin(6\pi t + \alpha)]$$

In this equation $F_0(t)$ is the time-varying fundamental frequency or pitch trajectory of the stimulus and $\overline{f}_0$ is the mean pitch of the original utterance; for a statement, $\alpha = 4$ and for a question, $\alpha = 1$. Each stimulus was processed with both question and statement pitch tracks, giving $\approx$ 16000 stimuli. The precise form of the pitch manipulation was chosen so that we could compare the model performance with those of human subjects in a recent psychophysics study (Head & Denham, 2004). The results of these manipulations are illustrated in Figure 6.2. The standardization of the stimulus length to $500ms$ was to ensure that the pitch track equation, which is time dependent, produced the same pitch track for all stimuli.

6.4.3 The speaker recognition set.

The stimuli for this experiment were not drawn from the ISOLET corpus but from a subset of the Speaker Recognition v1.1 corpus (OGI, 1996). This consisted of

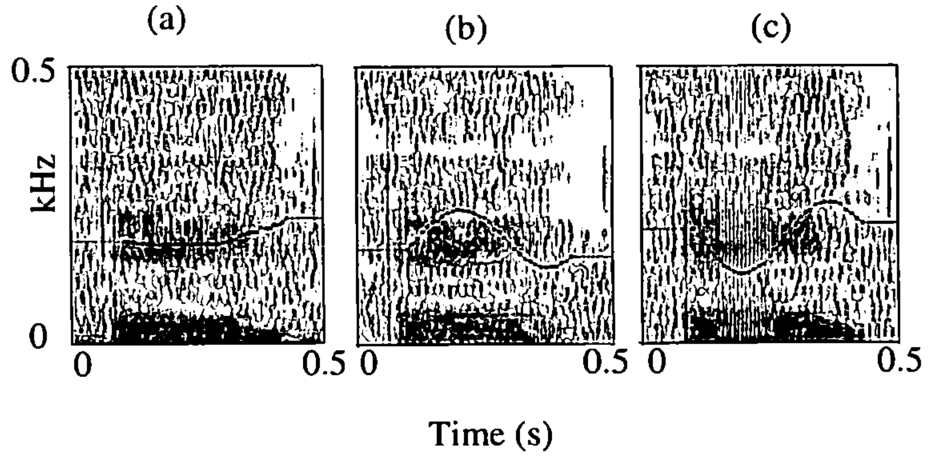


Figure 6.2: *Question/statement processing example, showing spectrograms with pitch tracks superimposed in blue. Left: Original utterance (letter 'a', female speaker, mean pitch 190 Hz). Centre: Question form. Right: Statement form. The principle difference between the spectrograms here is in the periodicity of the signal which is visible only in the spacing of the vertical stripes in the spectrogram.*

four speakers, two male and two female, answering questions such as *'What is your eye colour?'*, and *'Where do you live?'* with most answers given more than once. There are approximately 100 answers for each speaker. Longer answers were truncated at two seconds to save pre-processing time.

6.5 Results

The results for each of the four experiments using the ISOLET database classified with the analogue ANN and the network of simulated spiking neurons are shown in Figure 6.3 in terms of the normalized mutual information as described in Chapter 5, and for comparison in terms of correct classification percentage. For each result the number of classes n_C is also indicated.

The mean normalized mutual information for letter-names classification using the spike-driven network was 0.741. This represents a classification accuracy

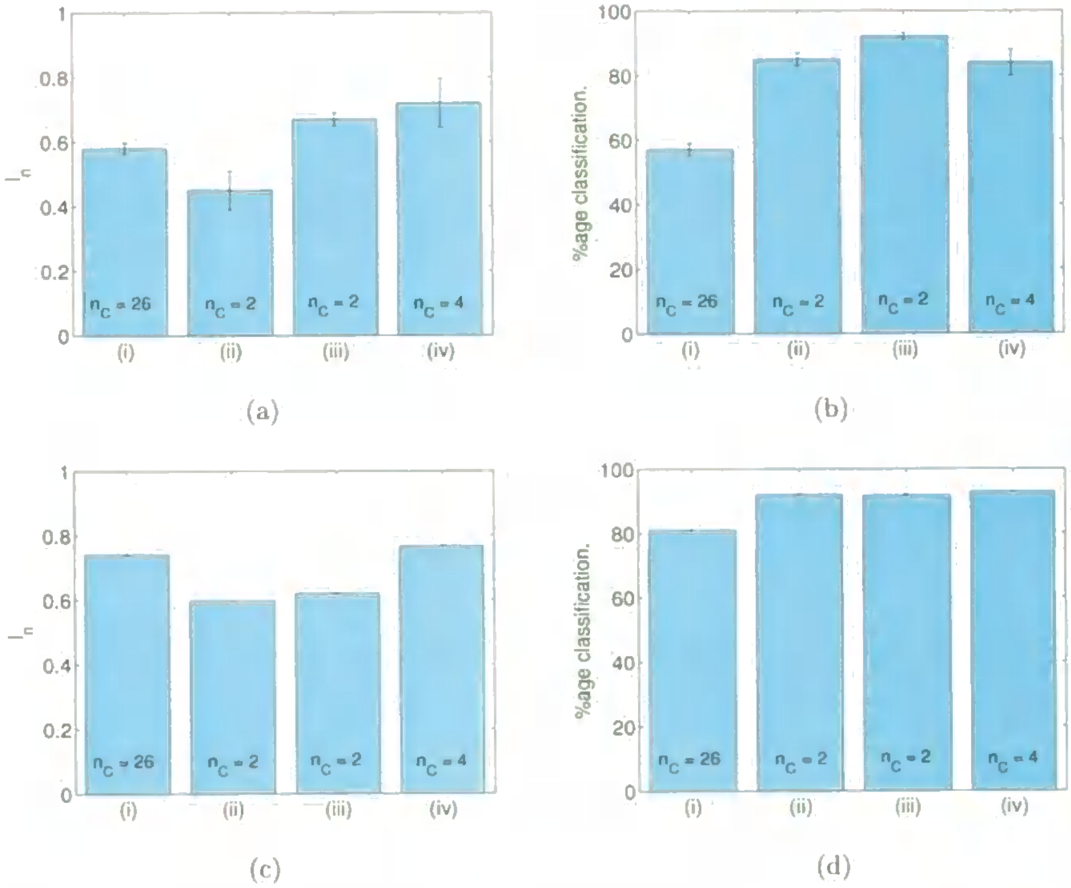


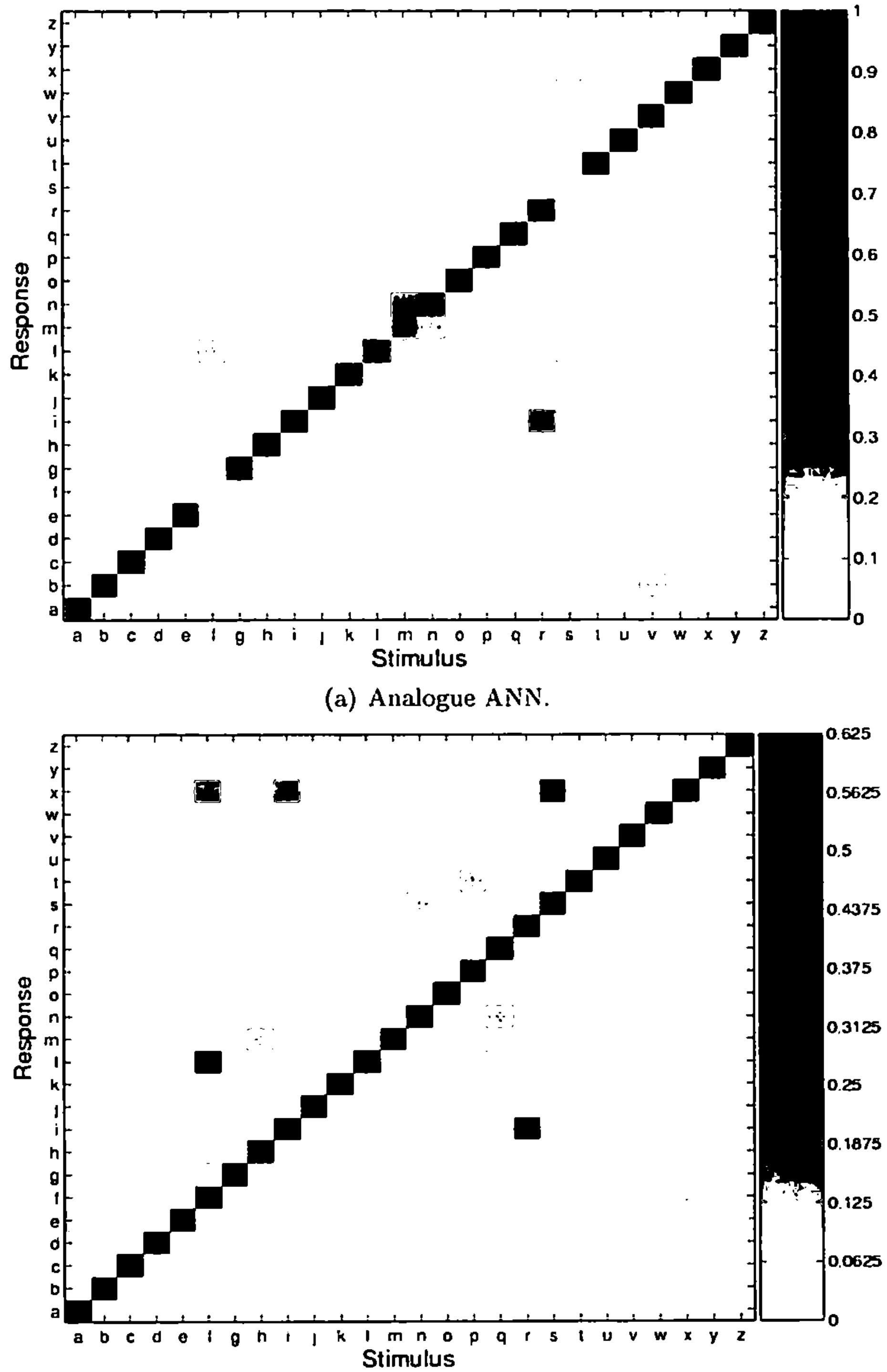
Figure 6.3: *Results of the three ISOLET experiments and speaker identification task. (i) ISOLET letter names, (ii) Question/Statement, (iii) Male/Female, (iv) Speaker identity. (a)-(b) Results from the analogue ANN in terms of normalized mutual information (left) and percentage correct classification (right). (c)-(d) Results from the network of spiking neurons in terms of normalized mutual information (left) and percentage correct classification (right). Also shown for each result is n_C the number of classes.*

of over 80% which compares favourably with that reported for other machine learning algorithms (Yu & Liu, 2003b). Plots of the misclassifications for the two classifiers are compared in Figure 6.4, note that because the majority of errors in the spiking network results were non-classifications the gray scale in Figure 6.4(b) covers a limited range of values to make the misclassifications visible.

6.5.1 Letter name results.

Misclassifications. Figure 6.5(a) shows the pattern of experimental misclassifications. These experimental confusions account for less than 6% of the total stimulus presentations, but among the most frequent are *f→[lx]*, *r→i* and *s→x* which all share an initial phoneme. Some interesting features emerge from a comparison of the pattern of experimental misclassifications with the pattern of misclassifications from human psychophysics shown in Figure 6.5(b) (Hull, 1973). To better compare Figure 6.5(a) and Figure 6.5(b), Figure 6.5(c) is plotted as a percentage change of the within-class error rate between Figure 6.5(a) and Figure 6.5(b). In Figure 6.5(c) white areas represent classes that are not confused by the model nor in human psychophysics. Green areas represent agreement between the model and the psychophysics as to how easy or difficult it is to distinguish the two letters. Red areas are those where the model has more success in differentiating the classes, and blue areas are those where humans outperform the model. The vast majority of the map is either white or green.

Red areas (those where the model results compare favourably) are found in the *d→e*, *k→a* and *v→[dbep]* misclassifications. These pairs are distinguished by their initial phonemes. The dark blue areas (those where model results compare unfavourably) include *r→i*, and *s→x*. These pairs share an initial phoneme.



(a) Analogue ANN.

(b) Spike-driven network.

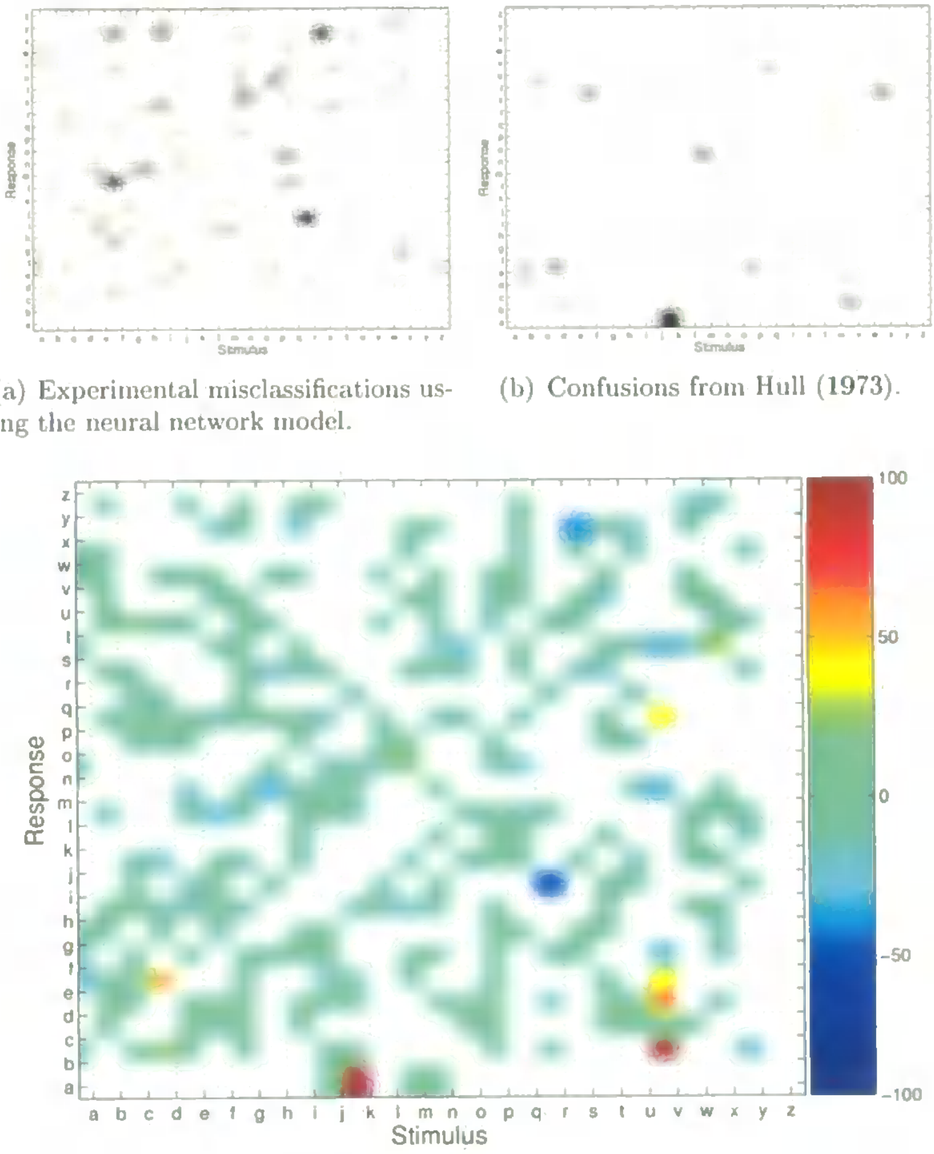
Figure 6.4: *Results from letter name classification. NB the much reduced range of values covered by the grey scale in 6.4(b) to bring out the detail in the small number of misclassifications.*

It is likely therefore that performance could be improved still further by incorporating events other than the first event in each stimulus presentation. Note that the ISOLET database uses $Z =$ *'zee'* (US) whereas the experiments in Hull (1973) use $Z =$ *'zed'* (UK) so the results for this letter name are omitted in this comparison.

PCA analysis of network weights. In order to investigate the contribution of each feature to the classification of each of the letters, we performed a principal component analysis (PCA) of the neural network weights obtained in each of the training sessions. A composite loading vector was obtained for each letter in the stimulus set by combining the eigenvectors corresponding to all eigenvalues greater than 0.7. The resulting matrix, illustrated in Figure 6.6, shows that there is a sparse representation of the data set; with each feature contributing significantly to only a few classes, and each class being primarily defined by a rather small set of features. This is encouraging as it shows that the FCBF fragment selection algorithm successfully chooses features that are de-correlated, and also means that the ensemble can in principal encode a very wide range of classes. If the weights were not sparse then there would be less room for new codings, for new classes, with different patterns of significant contributions.

6.5.2 Question/statement classification.

The average correct classification achieved by the model (88%) is comparable with the average performance of human subjects (80%) (Head & Denham, 2004). This may seem rather surprising since the classes are defined by the pitch trajectories and the feature ensembles are chosen from a spectro-temporal envelope



(a) Experimental misclassifications using the neural network model.

(b) Confusions from Hull (1973).

(c) Percentage change from Figure 6.5(b) to Figure 6.5(a)

Figure 6.5: *The plot in (c) shows differences between (a) and (b). White: agreement, i.e. low levels of misclassifications in the model or in psychophysics. Green: agreement, the model and the psychophysics agree as to the confusability of letter names. Red: the model finds these distinction easier than human subjects. Blue: The model misclassifies where human subjects rarely do.*

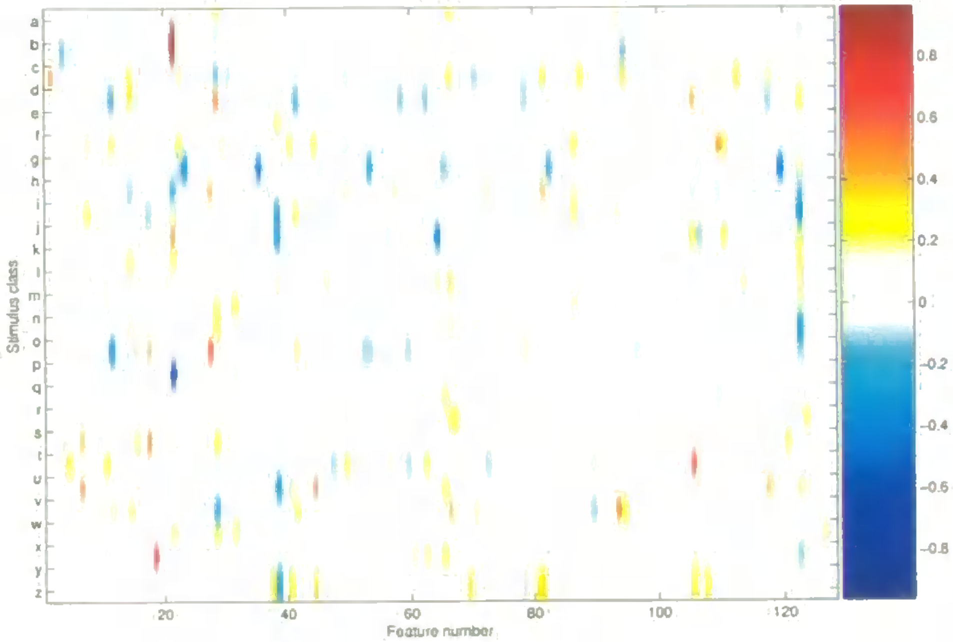
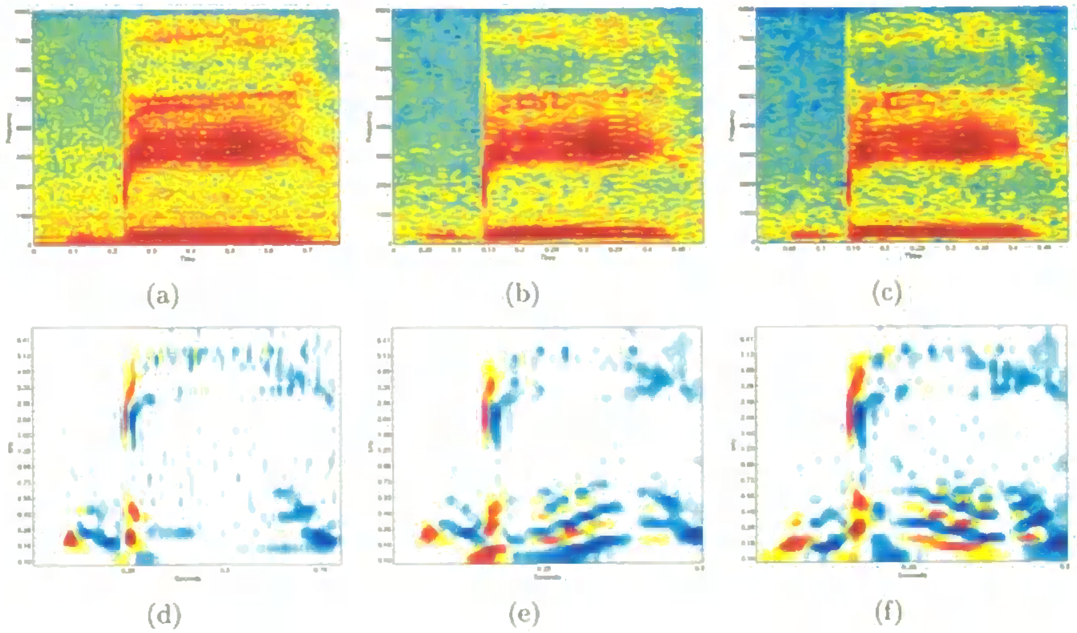


Figure 6.6: *Sparse coding of the stimulus set: the image shows the significant contributions of features to each class derived from a PCA analysis of neural network weights.*

representation; pitch is not explicitly extracted in the model. However, on closer examination it seems that in the onset/offset representation a rising or falling pitch track creates a characteristic pattern of onsets and offsets as the energy moves from one frequency channel to another (as illustrated in Figure 6.7) and this could allow stimuli from the two classes to be distinguished. To illustrate this enhanced difference Figure 6.8 shows the pairwise summed cross-correlation between time slices of stimuli with manipulated and un-manipulated pitch tracks. From this diagram it can be seen that there is very little similarity between pairs of stimuli in the SKV representation.

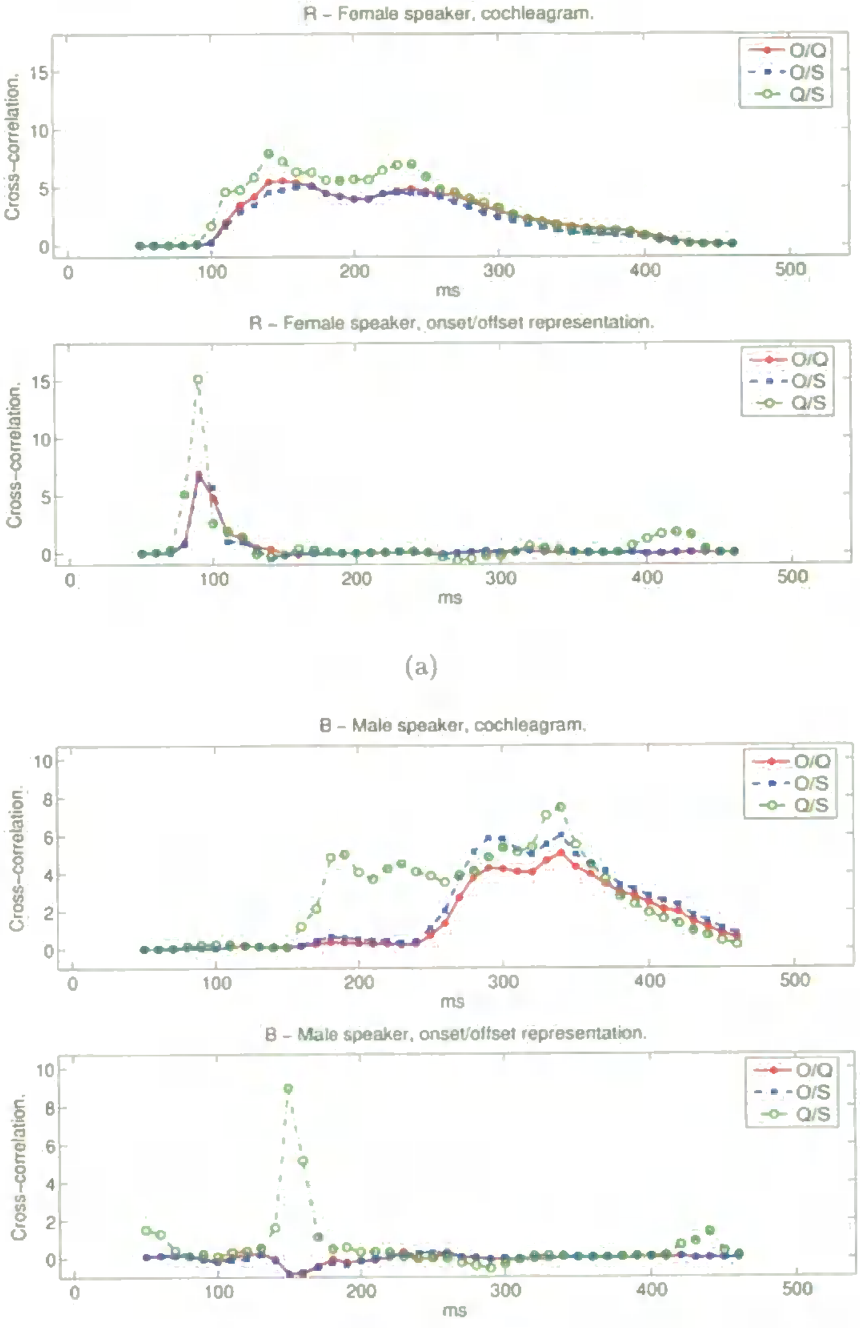
Another important aspect to note is that the mean pitches vary widely across the stimulus set, from low male pitches, typically ≈ 80Hz to high female pitches of ≈ 350Hz, which implies that the representations derived from the projections into feature response space support the abstraction of *pitch trajectory shape.* The ability of this model to classify the shape of pitch trajectories in complex sounds perhaps sheds some light on the somewhat contradictory data for amusics. In a recent experiment it was found that amusics' ability to detect and classify continuous pitch changes in sounds was almost as good as that for normals, while their ability to detect differences in discontinuous pitch sequences was much worse (Foxton *et al.*, 2004). The result demonstrates that ensembles of STRFs similar to those measured in PAC of animals, are capable of classifying pitch trajectories which can be represented within a single event. However, recognizing a pattern of discrete pitches would require the system to learn the sequence of projections of separate events within the feature responses space: a different problem involving higher order processing, perhaps the locus of impairment in amusics?



(a) (b) (c)

(d) (e) (f)

Figure 6.7: *Top row: The letter B, normal, question, and statement. Bottom row: each processed using the onset/offset representation. Rising and falling pitch trajectories are clearly visible in the SKV representation.*



(a)

(b)

Figure 6.8: *Showing correlations between pairs of utterance types in different representations. (a): 'R'-Female. (b): 'B'-Male. Upper: SCM representation. Lower: SKV representation.*

The high points on each line illustrate times when a pair of stimuli are similar. Red, the correlation between the origian and the question form (O/Q). Blue, between the original and the statement form (O/S). Green, between the question and statement form (Q/S). Correlations are reduced in the SKV representations.

6.5.3 Male/female results.

Classification success for the male/female discrimination task was ≈ 95%, which is broadly consistent with data from human psychophysics (eg Whiteside, 1998) with a reported mean success of 98.9% in an experiment using short vowel segments. Since clear differences in vocal tract length and vocal tract morphology between males and females are known to exist (Fitch & Giedd, 1999), it is perhaps not surprising that the model was able to perform this classification task. Nevertheless, the problem is not trivial as changes in vocal tract length result in quite small changes in the positions of formant peaks, and it is necessary to detect these in the presence of much larger changes in formant position characterizing the different speech sounds. In a recent PCA analysis of the variability of spoken vowel sounds, it was found that 80% of the variability was accounted for by differences between vowels, and of the 20% of intra-vowel variability, 90% was explained by changes in vocal tract length; i.e. 18% of the total variability (Turner & Walters, 2004). The model of vocal tract length (VTL) estimation presented in that study matched experimental data very well, but was restricted to the single vowel sound *'aa'*. The current model on the other hand is able to learn to classify speaker sex for arbitrary utterances, and as far as we are aware may be the first biologically plausible model of voice gender classification.

6.5.4 Speaker identification results.

This was the only experiment that did not use the ISOLET corpus (see 6.4.3). The model was able to identify correctly each of the four speakers with an accuracy of ≈ 89% using short segments of randomly chosen utterances. For comparison, in the recent study of Obleser *et al.* (2004) subjects were able to identify two speakers with an accuracy of $\approx$ 95%. As the number of speakers in the experiment was small the result is only suggestive, but it was achieved in a text independent experiment using the same feature extractors as the other experiments reported here. This establishes, at least in principle, that information about speaker identity can be preserved in the pattern of responses of such an ensemble, and that responses of the same ensemble can be used in parallel for a number of different perceptual classifications; as found in the human MEG study for phonological and speaker classifications in Obleser *et al.* (2004).

6.6 Discussion.

6.6.1 General.

These results show that the response vectors support classifications of many different, qualitatively different types. The performance of the system in letter name correct classification ($\approx$ 80%) may be comparable to other machine learning algorithms (e.g. Yu & Liu, 2003b), and the success in male/female classification (Whiteside, 1998) and question/statement distinctions Head & Denham (2004) close to results from human psychophysics - but what is important is that these results are based on a single set of feature vectors. This is the first time this type of concurrent task has been simulated using diverse tasks in a comparable way.

6.6.2 Biological implications.

Given that it is widely reported that responses in PAC can, at least to a first approximation, be characterized by their spectro-temporal characteristics, it is not unreasonable to ask whether an ensemble of spectro-temporal feature extractors might provide a representation sufficiently rich to be biologically useful. We have shown using biologically plausible pre-processing, a modestly sized ensemble, and a spike-rate encoding, that salient features of the stimulus can be simply extracted and used as the basis for judgements that in a living organism would be behaviourally relevant. Moreover, the same ensemble can support many qualitatively different judgements concurrently. This is consistent with evidence that 'what' processing in auditory cortex can be viewed as a set of parallel processes in which concurrent phonological classifications are made in spatially separated areas (Obleser *et al.*, 2004) and implicit semantic processing continues when attention is directed to non-verbal input analysis (Kriegstein *et al.*, 2003).

The range of classifications supported by the model includes those distinguished primarily by spectral profile (male/female), solely by pitch trajectory (question/statement), as well as those characterized by more complex spectro-temporal relationships (letter-names, speaker identity). The question/statement result in particular demonstrates that a representation of pitch change can be abstracted from the output of the system in which there is no explicit representation of pitch *per se*. Furthermore the performance of the model in each of the tasks shows some similarities with human psychophysics. One of the strengths of the spiking neural network is its ability to provide non-classifications. This implies that the characterization of the stimulus by the model using a single event is unclear. Such stimuli account for $\approx 14\%$ of the test set in the current results; most frequently in classes *[flmns]* i.e. classes that are not resolved by their initial phonemes. Work is already underway to use subsequent events, when they occur, to reinforce the classification judgement raising the probability above the threshold for an unambiguous assignment of class. However, in the first instance the confusions *validate the approach* inasmuch as they are similar to confusions measured in psychophysics. If there were no confusions then we may have achieved a representation of *distinctness* but

> *a representation whose fidelity is limited to distinctness provides no basis for generalization because it does not contain information concerning relations among stimuli* (Edelman, 1998).

In some ways the more confusions exist the better e.g. *v* → *[dbep]* provided the rank order of similarity is preserved. In the limiting case if all stimuli are similar *and* in rank order of similarity (for a finite number of points in the representational space) then the isomorphism is perfect, what Edelman calls a *similtude.* This argument is independent of classification success and mutual information.

6.6.3 Future work.

If measures such as classification and mutual information are not the measures of success that we want then further work must seek to establish whether the mapping of stimulus to response is principled. This is a profound point. The mutual information (or normalized mutual information) gives us confidence that our response vectors do contain information about the nature of the stimulus. What is more important is that these vectors can form the basis of a representation of similarity, or second order isomorphism (Shepard & Chipman, 1970) which allows generalization *and* misclassifications that respect the relations among stimuli. The pattern of misclassifications then assumes great importance as is implied by the quote from Edelmann above and we have, in a preliminary way, moved towards such an analysis in Section 6.5.1. More work needs to be done on the letter name misclassifications and it would also be valuable, for example, to analyze the stimuli, and response vectors to see if those data that resulted in Male/Female misclassifications had some features in common, and to conduct psychophysics on these stimuli to see if Male/Female misclassifications followed a similar pattern in human subjects.

Chapter 7

Concluding discussion.

7.1 A model of auditory processing.

The auditory neural pathway is a complex, dynamic system that starts with the imperfectly understood mechanics of cochlear motion and the complexities of inner hair cell (IHC) transduction.[1] The resulting signals propagate through a network of recurrently connected neural way-stations culminating in a number of 'higher' centres involved in a wide range of behaviourally important judgements about auditory stimuli such as *what*, *where*, *how big*, *how near*, and so on. The same system, in humans, also serves as the precursor to the perception and comprehension of language even though speech appears to differ markedly from unsophisticated stimuli such as raindrops on leaves and a twig snapping on the forest floor.

Many years of careful research have shown that the responses typical of various

[1] Of course we have to be careful what we mean when we use words like 'starts' in the context of system with efferent connections aplenty at all levels. In this case it is justified by the IHCs being part of a mechanical chain of events that can be traced all the way back to the stimulus.

important parts of this system can be expressed in spectro-temporal terms. In this thesis we have endeavoured to take a broad view and to build a model of auditory processing and classification based on some of these results and as part of this endeavour we have been forced to address a number of important questions about auditory processing:

- Why are peripheral and mid-brain characteristics dominated by transient responses?
- Given that transients are enhanced in the periphery and mid-brain, is it plausible to characterize cortical responses (STRFs) in these terms, i.e. as patterns of onsets and offsets?
- Can we begin to understand, in developmental terms, how STRFs come in to being?
- Are the resulting responses from STRFs rich enough to support a range of behaviourally relevant judgements?

In addition there are some additional questions more closely linked to the model itself:

- Can we reconcile the fast changing nature of auditory stimuli with the hypothesis that responses are spike-rate coded?
- Can we identify 'salient' parts of stimuli which might form the basis of a quasi-stationary stimulus response? This might consist of a series of discrete events.

- Is the response derived from within the salient window *information bearing*, i.e. can it be used as the basis for a classifier as a model of perceptual categorization?

The 'success' of the current project is, therefore, to be judged not only in terms of the performance of the model but by the insights which may have been gained in to the answers to all of these questions.

The success of the model. In the process of addressing the questions outlined above we have developed a novel 'test-bed' system that functions as a universal auditory classifier. The formation of the representation is stimulus driven on the basis of patterns found in fragments of a limited number of 'formative' stimuli. Results suggest that these patterns can be rated for there usefulness on the basis of the entropy of their responses to these stimuli. We have used this rating, and a selection procedure based on information theoretic principles to derive an ensemble of spectro-temporal patterns that is a model of the range of STRFs found in cortical electrophysiology. This process can be automated and involves very few free parameters, principally used to control the eventual size of the ensemble (n_E) for the purposes of computational convenience. The selected ensemble was then used to support a stimulus driven asynchronous event detection method based on coherence of the response across the *whole* ensemble. The responses of *each member* of the ensemble was then used to derive a vector that characterized the stimulus when used as the input to a classifier. This simple method of producing a response vector as a quasi-static view of the salient part of a time varying stimulus proved to be successful. We have used the responses to successfully classify a speech corpus in a number of diverse ways, including by the sex of the speaker which we believe to be an entirely novel result.

The system is a 'test-bed' as each part of it, preprocessing, feature selection, response generation, salience detection, and classifier, can be freely exchanged with others to test alternative methods and compare hypotheses on the basis of results.

The role of transient responses in the periphery and mid-brain. We have shown in Sections 2.4.2 and 2.5.1 that the SKV representation compresses the dynamic range within each frequency channel and whitens the spectro-temporal representation with respect to the cochlear response. Both of these results have parallels in studies of the responses of neurons in the visual system (Atick & Redlich, 1992) and de-correlation, or redundancy reduction, has been proposed as a strategy for cortical processing (e.g. Barlow & Foldiak, 1989). It is interesting to note from the results in Chapter 2 that a white-noise stimulus has a highly correlated cochlear representation which is profoundly whitened by the SKV representation. It could be, therefore, that in addition to other possible roles for transient sensitivity in the *larger strategy*, its *utility* may also lie in compensating for the correlations inevitably introduced by mechanical transduction in the cochlea.

Transient sensitivity also has an impact on the 'sparseness' of the response. Taking speech stimuli as an illustrative example, they are highly modulated in both time and frequency, and consequently areas of high energy in the spectro-temporal representation are relatively few. However, areas of fast energy change, as identified by the SKV representation, are considerably fewer and are separated by gaps (see Figure 2.3) representing the short periods of relatively unchanging energy within a frequency channel. These offer 'glimpses' of any interfering stimuli which may aid auditory figure-ground separation (Cooke, 2003).

Results in Section 2.4 also indicate a differential response for sounds having different distributions of spectro-temporal modulations, precisely because both types of modulation produce within-channel energy transients. This may be related to, and may even be a partial explanation of, results reported by Schnupp *et al.* (2005) which indicate preferences in *cortical* responses to different classes of $1/F^{\alpha}$ noise. In addition the results also show that the patterns of transients exhibited by speech stimuli are not greatly disrupted when mixed with non-preferred noise interference, even at low signal to noise ratio (SNR). This may shed some light on why SNR is, by itself, such a poor predictor of intelligibility of speech in noise, or speech mixed with other interfering signals (Brunghart & Simpson, 2002; Cooke, 2003).

The ability of the model to predict interference effects has already yielded preliminary results (Denham & Coath, 2005) and there is much scope here for further investigation.

Spectro-temporal responses in terms of transients. As is discussed in Chapter 3 the characterization of cortical and mid-brain responses as linear kernels derived from reverse correlation with spectrographic, or cochleagraphic representations has yielded some interesting results. In many cases, however, these kernels fail to predict the response to novel stimuli, and the degree to which they fail depends strongly on the sounds used in the experiment (Machens *et al.*, 2004). This is precisely what we would expect if peripheral processing of the sound profoundly altered the spectral and temporal envelopes found in the sub-cortical representation available to cortical units.

We have established that STRFs obtained by reverse correlation with transient sensitive spectrograms, are plausible in that they are, as we would expect, consistent those derived from the cochleagraphic representation for a limited class of stimuli. The stimuli used in the collection of the data presented in this work (random chord stimuli) are not rich enough in the range of modulations they contain to expose differences. We predict that measurements with parametric stimuli designed to explore onset sensitivity may lead to kernels that better predict responses to novel stimuli and may help to explain the cortical preference for stimuli with different spectral and temporal envelope statistics (Schnupp *et al.*, 2005).

In addition, our results show that at least some energy derived STRF estimates are inconsistent with first spike latencies in that they predict a maximum in response at $\delta t = 0$, whereas the corresponding estimates based on the SKV representation are broadly consistent with the experimental latency data. This result demands further investigation. A great many STRF estimates in the published literature are obtained by pooling spikes in bins of up to 20ms. If these linear kernels predict that the maximum excitatory regions are in the first bin then the minimum latency is obscured and this could be hiding an important piece of evidence.

The ontogeny of STRFs. It has been reported that the organization of auditory cortex is affected by formative stimuli (e.g. Zhang *et al.*, 2001) and that reorganization of cortical responses can continue in to adult life to reflect the nature of the auditory environment (Wang, 2004). It seems possible therefore that when seeking the spectro-temporal patterns (equivalent to 'preferred stimuli' Hubel & Wiesel, 1962) to which auditory neurons are sensitive, that these may be found in the stimuli themselves. Our experiments have shown that the response of a candidate STRF, or feature extractor, can be judged more or less 'interesting' with respect to a limited number of formative stimuli based on the entropy of its response. We have also tried to optimize, in some sense, response of an ensemble by ensuring that responses of each element are uncorrelated, thus reducing the overall redundancy. These principles form a basis for preferring one STRF over another, and also for preferring one *ensemble* of STRFs over another, and as a consequence they form the basis for a developmental pressure. Importantly for this argument we have shown that the exact choice of STRF geometry is not critical, at least when dealing with a limited number of formative stimulus classes. This allows for gradual optimization from a sub-optimal, or random, set of responses. We have also shown that if the formative stimuli are sufficiently rich then the resulting ensemble responses are productive, i.e. they are suitable for a wide range of stimuli of disparate types, not just those that are similar to the formative stimulus set, see Section 5.5.1.

On a different level, the argument about the ontogeny of STRFs has led us to develop and introduce a novel way in which an artificial system might select features of a set of stimuli that are useful in classification. These features can be formed on the basis of patterns found in the stimuli themselves. We have also shown that these features, even if selected from a limited set of formative stimuli, are productive in the sense that they can be useful in classifying a much wider range of sounds. This is true if the additional stimuli are statistically similar to the formative stimuli, but also, although to a lesser extent, if they are not.

Auditory salience. Auditory stimuli change continuously. However there are discrete percepts associated with them. This is a difficulty for any model of auditory perception. Part of the answer may lie in results that suggest that perception and classification require integration over relatively short time scales around those parts of the signal that exhibit change (Furui, 1986). The SKV representation introduced here emphasizes change (Section 6.5.2) and our experiments show that STRFs of $100ms$ duration or less have properties which make them desirable, at least for speech stimuli in that they have a higher response entropy than other fragment lengths (Figure 4.4), their performance reflects human psychophysics with compressed stimuli (Figure 5.5), and the information theoretic selection procedure shows a preference for features with best temporal modulations of $\approx 10Hz$. This provides the temporal context. The response of an array of transient pattern detectors therefore can be viewed as an indication of auditory salience which allows the stimulus to be viewed as a series of quasi-stationary salient states that can be represented as a response vector and can be spike rate coded. This novel approach was an important and necessary step for the hardware implementation project of which this work forms a part, but it may also provide some insight in to neural auditory processing.

The 'salience window' in the current model is based on the presence of a coherent response across the ensemble of feature detectors. This provides an asynchronous, stimulus-ensemble driven event detector which triggers a read-out of the population response pattern within a time window. The length of this window is determined by the duration of the coherent ensemble response and the result is a short time scale context for the extraction of a pattern of responses that characterizes a distinct auditory event. These events are likely to be represented by population responses which, because of the time window and the asynchronous read-out, are not likely to bear a simple relationship to the temporal structure of the stimulus. It has been suggested that this type of post-primary cortical processing might be found in the *planum temporale* (Griffiths & Warren, 2002) where responses that are not closely coupled to the time course of the stimulus do occur (Steinschneider *et al.*, 1999). We have used the summed response of our ensembles and a simple threshold to define the salience window and this approach has proved fruitful. However, in future work it may be that some other property of the ensemble response, such as the entropy, may prove to be a more effective indicator of salience. We have shown that the response vectors derived from the salience window contain information about a wide range of classes in to which the stimulus might fall, and that it is plausible that they could form the basis of perceptual categorizations (Section 6.5).

Classification based on onset sensitive kernel responses. As stated in Section 6.2 there is no reason to assume in advance that the quasi-static, rate coded, response of a limited number of STRFs would contain sufficient information about the stimulus to allow its classification in a number of disparate ways. We have shown response vectors derived in this way *do* contain enough information for multiple, concurrent, behaviourally relevant classifications in speech stimuli despite the 'bottlenecks' imposed by the preceding processing stages of the model. Our hypothesis is that the vectors, when used as the input to a classifier, form the basis of a 'representation of similarities' (Edelman, 2002) mapping the distal stimulus space to a low dimensional proximal vector space that preserves the distance between adjacent representations is in some way related to their similarity. We have shown that such mappings are possible for a variety of similarities depending on the view taken of the stimuli. The range of classifications supported by the model (Section 6.5) includes those distinguished primarily by spectral profile (male/female), solely by pitch trajectory (question/statement), as well as those characterized by more complex spectro-temporal relationships (letter-names, speaker identity). The question/statement result in particular demonstrates that a representation of pitch change can be abstracted from the output of the system in which there is no explicit sense of pitch *per se*. Furthermore, the performance of the model in each of the tasks shows some similarities with human psychophysics. This is the first time, as far as we are aware, that a model of auditory classification has reported success in male/female classifications that is close to that reported in human psychophysics.

It has been reported that different perceptual categories are processed in distinct areas of auditory cortex anterior to PAC (consistent with the 'what' pathway) and also distinct from regions involved in decisions that are correlated with reaction times (Binder *et al.*, 2004). It is possible that there is competition between these perceptual judgements which is subject to a task-dependent attentional bias originating in other cortical areas. The aspect that is attended to is the one most likely to be task-relevant. This is consistent with evidence that 'what' processing in auditory cortex can be viewed as a set of parallel processes in which concurrent phonological classifications are made in spatially separated areas (Obleser *et al.*, 2004) and implicit semantic processing continues when attention is directed to non-verbal input analysis (Kriegstein *et al.*, 2003).

7.2 Epilogue.

Around the year 400, the man later known as St. Augustine of Hippo wrote:

> *Deus Creator omnium: this verse of eight syllables alternates between short and long syllables . . . I affirm this and report it, and common sense perceives that this indeed is the case. . . . But when one sounds after another, if the first be short and the latter long, how can I hold the short one and how can I apply it to the long one as a measure, so that I can discover that the long one is twice as long, when, in fact, the long one does not begin to sound until the short one leaves off sounding? That same long syllable I do not measure as present, since I cannot measure it until it is ended; but its ending is its passing away. What is it, then, that I can measure? . . . I could not do this unless they both had passed and were ended. Therefore I do not measure them, for they do not exist any more. But I measure something in my memory which remains fixed* (Augustine of Hippo, c.400).

This is a profound point that has resonances throughout the field of auditory science. The question of how we come to have internal representations of anything, objects, rhythm, meaning, class, etc. is still not at all well understood in *any* field. In auditory science there is the further complication that the stimuli are often fleeting. Augustine's answer is, surely, plausible at least. It is not, he says, from a perceptual point of view, the stimulus that has timing or rhythm ('*. . . I do not measure them . . .* '). Rather it is the rhythm of the stimulus that evokes a memory of previous similar rhythmic experiences ('*. . . something in my memory which remains fixed.* ').

This interpretation is very close to the idea of the 'representation of similarities' drawn on in Chapter 5 of this work. A neural representation of similarities is a sort of memory, but not one that is evoked by a narrow class of closely related stimuli, which would be a representation of similarities. To be useful, all stimuli, even completely novel stimuli, have to map to some point in this representational space. In this way Augustine would be able to spot the short-long-short-long (*iambic*) pattern in languages he could not understand, and in Irish dance music as well as the familiar patterns of devotional Latin verses.

On this basis, it could be argued that the primary role of sensory coding is to ensure that similar stimuli are *perceived as similar*, independent of volume, interference, pitch changes, and so on. But the number of ways a sound can be interpreted is very large, i.e. it can be mapped to more than one similarity space. Sensory coding must support these multiple mappings and features that do not contribute to the separation are likely to be de-emphasized

But not all stimuli are equal. Those that are most important are those most likely to be encountered, particularly during the early developmental phases, and those that form the basis of behaviourly important judgements. It is to be expected then that as a secondary role of sensory processing it will have a preference for such stimuli, maximizing the information transmission in these cases. It is possible that the results presented in this work may help to provide insights as to how both of these ends are achieved.

Appendix A

Psychophysics using stimuli re-synthesized from onsets.

A.1 Abstract.

Onsets extracted from the time varying output of a model of cochlear responses on a channel by channel basis can be used to reconstruct experimental stimuli that retain intelligibility (for speech samples) or features which allow sounds to be correctly classified in psychophysical experiments.

These onsets are extracted by calculating the skewness of a distribution derived from output of each channel of the cochlear filter bank. The output is divided in to windows, the lengths of which are varied with the period of the central frequency of each filter.

The experimental stimuli are reconstructed from the onset data by replacing each onset with a shaped tone burst of uniform maximum amplitude at the central frequency of the filter. The length of the tone burst is in accordance with analysis

of repeated period noise experiments and autocorrelation models of neural pitch extraction (Wiegrebe, 2001). The envelope of the tone burst ramps gently upwards from a non-zero floor and more gently downwards to the same minimum level. This minimizes onset and offset transient effects. In keyword recognition experiments subjects were able to identify up to 50% of the key words from stimuli derived from onset information alone.

A.2 Cochlear model.

The model adopted for this work is based on Michael Slaney's equivalent rectangular bandwidth (ERB) filter bank in his MATLAB auditory toolbox (Slaney, 1994). This is a model of the response of the basilar membrane and is based on independent gammatone bandpass filters as suggested by Patterson *et al.* (1992). The design of these filter banks is based on psychoacoustic measurements of 'critical bands', indeed Slaney uses the terms ERB and 'critical band' interchangeably.

Beyond an approximation of the critical bands the model also includes half wave rectification of the output of each bandpass filter. This approximates the response of the inner hair cells (IHC). More complex models of IHC response, such as the MeddisHairCell routine available in the auditory toolbox, include automatic gain control, adaptation, and onset sharpening. All of these factors are, in this work, dealt with by subsequent processing.

A.3 Neural pitch extraction.

The time constants used as parameters for feature extraction and re-synthesis are based on work by Wiegrebe (2001). This work examines the strength of the perceived pitch in samples of repeated period noise. Comparing this perceived pitch strength with a running summary autocorrelogram of the same waveform Wiegrebe concludes that

> *The current data strongly suggest that the temporal window of pitch extraction depends on the pitch itself, i.e. the minimal integration time is longer for lower pitches. Qualitatively useful results were obtained with the time constant being fixed at* 2.5*ms for correlation lags smaller than or equal to 1.25ms and the time constant being twice the correlation lag for periods larger that 1.25ms (Wiegrebe, 2001).*

A.4 Skewness and onsets.

For this work the outputs of the filters used in the cochlear model were treated as independent of each other. The output from each was divided in to time windows the length of which varied with the centre frequency of the filters. The minimum window size was 2.5ms for centre frequencies of 800Hz and above and for frequencies below $800Hz$ the size of the window was twice the period of the centre frequency. Each window was then normalized and the skewness of the distribution within the window calculated (Figure A.1d). Skewness is a measure of how symmetrical the 'tails' of are distribution are. It is the third central moment of the distribution calculated from Equation A.1

$$\beta_1 = \frac{1}{N}\sum_{j=1}^{N}(\frac{(x_j - \overline{x})^3}{\sigma^3}) \tag{A.1}$$

The skewness data was further processed to produce data that represented the rising and falling of each peak in each of the spectral bands. Areas of skewness that represented an increase in energy were assigned a value of +1, and those that represented a decrease in energy were assigned a value of -1. Values less than 5% of the maximum skewness were assigned a value of zero. This process identifies regions of rising, falling, and approximately constant output from the filter bands. The onset times were taken to be the positions of the first rise in energy of any group.

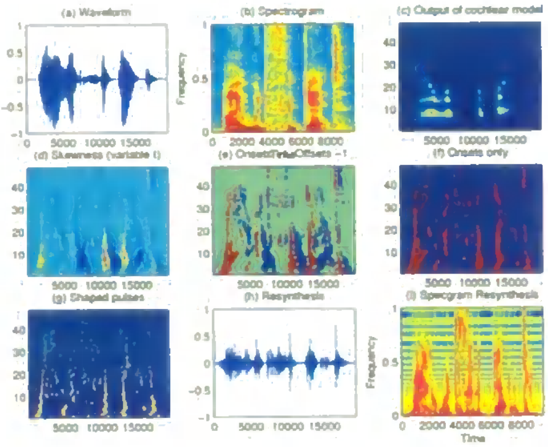


Figure A.1: *The various stages from original to re-synthesized waveform. Spectrograms of both are included for comparison. NB, the spectrograms are plotted on a vertical axis that is linear, whereas the outputs of the cochlear model are plotted on a non-linear vertical axis.*

A.5 Re-synthesis.

To produce experimental stimuli it was necessary to re-synthesize sounds from the extracted features. This re-synthesis is illustrated in Fig A.1. Figs A.1(a) and (b) show the waveform and spectrogram of the original sound ('I would forget'), (c) shows the output of the cochlear filter bank, (d) the skewness extracted from each filter band, (e) shows areas of onset (red) and offset (blue), (f) shows onsets only, (g) shows how the first instance of each group of onsets has been replaced with a shaped sine wave pulse, (h) and (i) show the waveform and spectrogram of the reconstructed sound.

The durations of the pulses used to re-synthesize the experimental stimuli from the position of the onsets are controlled by a time constant which can be varied by the experimenter. The shortest pulse in these experiments, for frequencies over $800Hz$ is $2.5ms$. Below $800Hz$ the length of the pulses is twice the period. A $2.5ms$ burst of sound represents a very short pulse indeed which is perceived as a 'click' at most frequencies. The stimuli prepared for the experiments were a selection of 100 sentences with lengths varying from three to eleven words (average 5.3 words) were chosen at random.[1] Examples showing obscure or archaic English usage were rejected. These stimuli were then re-synthesized using the method described above, using a number of different time constants. The resulting sound files were used as the basis of the psychophysical experiments (see below).

A.6 Experiments.

EXPERIMENT 1. Five subjects were chosen from native English speakers. After a short orientation session in which subjects were provided with on-screen text feedback, each was asked to listen to the fifty sentences presented in a variety of different orders and re-synthesized using the shortest time constant. Their responses were typed on a touch sensitive screen and compared with a list of words in the sentence to give a *'percentage of words correct'* score for each response. Subjects were not limited in the number of times they could listen to the stimulus but were encouraged to move on quickly if they could not recognise any of the words. After a short rest the experiment was then repeated using the longer time

[1] Mostly from the Oxford Concise Dictionary of Quotations.

constants.

EXPERIMENT 2. The second set of experiments used a different subset of the prepared stimuli. Each subject was presented with sets of sixteen or twenty sentences. The sentences were presented only once in each set and results were collated from subjects that had listened to a minimum of three sets. The stimuli in each set were drawn equally from each time constant group and each sentence was presented only once.

A.7 Results.

Experiment 1. Figure A.3 shows the performance of the group in identifying key words in the example sentences. The results show that a significant number of keywords can be identified from the re-synthesis using the shortest time constant. The percentage of words identified correctly increases as the time constant for the re-syntheses increases with the greatest difference between the first and second group. Figure A.2 shows that the results of the individual subjects mirror those of the whole group.

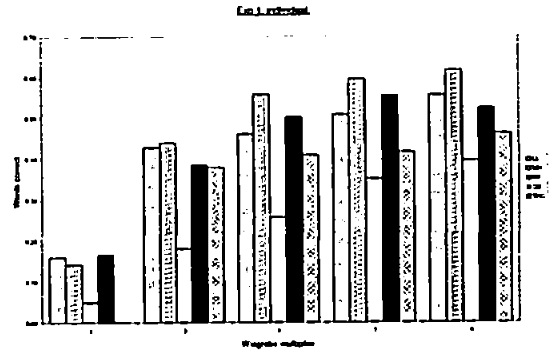


Figure A.2: *Onset re-synthesis, individual performance 1.*

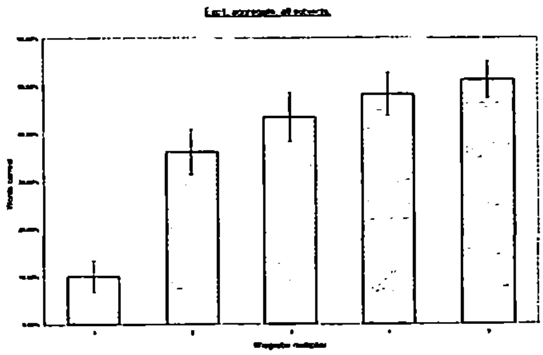


Figure A.3: *Onset re-synthesis, Group performance 1.*

Experiment 2. The results are shown in Figure A.4. These exhibit a similar pattern of results to those in Experiment 1.

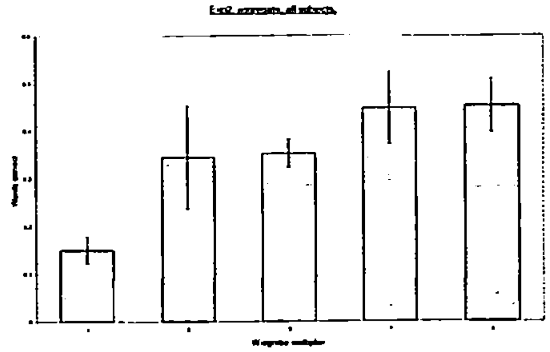


Figure A.4: *Onset re-synthesis, Group performance 2.*

A.7.1 Discussion

The results indicate that at the shortest time constant which is in line with that suggested by Wiegrebe (2001) as a minimum for neural pitch extraction, it is difficult for subjects to extract information from the spoken stimuli, although 10% of the words in the sentence were still correctly identified. As the time constant increases, the ratios between successive performance figures closely follows the ratio between successive time constants. This could indicate that the recognition of words from onsets only becomes easier as the onsets become more clearly tonal.

There is a great deal of analysis yet to be done which may throw light on which features of the original sounds are being preserved, and which of these are most useful in identifying words. The re-synthesis method preserves much of the rhythm of the sentence but in addition, the longer time constant stimuli preserve pitches within the frequency bands.

A.8 Percussive stimuli.

Subjects were first asked to listen to a variety of percussive sounds generated with different materials and told by the experimenter to which categories they belonged. This was continued until the subjects were happy that they could hear the differences in the original recordings.

Subjects were then presented with each re-synthesized stimulus once only and asked to assign each to one of three categories in a forced choice experiment. After a short rest the experiment was then repeated using the second, longer, time constant and so on for all five sets of re-syntheses. The results of the preliminary trial using stimuli re-synthesized from percussive stimuli are shown in Figure A.5. These results show the aggregate responses for all subjects. In this diagram the correct categories run horizontally and the responses run vertically; as a result the correct responses are on the lower-left to upper-right diagonal. The circles have an area which is proportional to the number of responses and the time constants are represented by colours in rainbow order (Red, Yellow, Green, Blue, Violet). Circles anywhere other than in the lower-left to upper-right diagonal represent

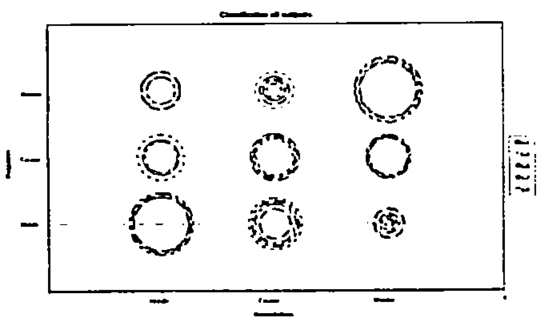


Figure A.5: *Aggregate performance, all subjects, for percussive stimuli.*

incorrect responses.

A.8.1 Discussion.

The largest circles are placed on the diagonal indicating correct classification but there are a large number of misclassifications. Of particular interest are:

Ceramic misclassified as metallic. These increase with time constant. A possible explanation is that peoples expectation of ceramic percussive noises is less tonal than their expectation of metallic percussive noises.

Ceramic misclassified as wooden. The significant group of ceramic samples misclassified as wooden at short time constants decreases rapidly at long time constants. This is perhaps consistent with people's expectation that wooden percussive sounds will be shorter and less tonal than ceramic percussive sounds.

Wooden misclassified as ceramic. There are a group of wooden samples misclassified as ceramic, which is of a similar size for all time constants.

Nomenclature

Acronyms, abbreviations and symbols.

C_{sr} The stimulus-response cross-correlation, see Section 3.3.1 Page 52. The C_{sr} can be shown to be equivalent to 'the spike triggered average' (STA) (Theunissen *et al.*, 2001).

C_{ss} The stimulus autocorrelation, see Section 3.3.1 Page 52.

C_{xy} The correlation between two vectors $\vec{x}$ and $\vec{y}$, see Section 3.3.1C_{xy} is defined as $\langle \vec{x} \cdot \vec{y}^T \rangle$ i.e. the expected, or mean value of the outer product of $\vec{x}$ and the transpose of $\vec{y}$ over a series of values. Depending on the dimensions of $\vec{x}$ and $\vec{y}$ this can result in a matrix, vector, or scalar.

n_C The number of nominal classes in to which a group of stimuli fall. For example in experiments using letter names as the class label then $n_C = 26$. Section 4.3.1, Page 77.

n_E The size of the ensemble of spectro-temporal filters used to derive the response to a stimulus. Section 4.2.5, Page 76.

SU Symmetrical uncertainty, an information-theoretic measure of correlation (Yu & Liu, 2003b). Section 4.3.2, Page 81.

T_{su} In the FCBF selection algorithm this is the minimum value of the symmetrical uncertainty (SU) below which a fragment is not considered for inclusion in an ensemble. This parameter controls the number of fragments considered and indirectly the number of fragments chosen. Section 4.3.2, Page 82.

ABR Auditory brainstem response. Section 2.2.2, Page 15.

ANN Artificial Neural Network. Section 5.2, Page 94

BTM Best temporal modulation. Section 4.4, Page 88

ERB Equivalent rectangular bandwidth. Section 2.3.1, Page 17

FCBF Fast correlation based filter (Yu & Liu, 2003b). Section 4.3.2, Page 80

FFT Fast fourier transform. Section 3.5, Page 65

IC Inferior colliculus. Section 2.2.1, Page 12.

MGB Medial geniculate body of the thalamus. Section 2.2.1, Page 12.

PAC Primary auditory cortex. Section 2.2.1, Page 14

PCA Principle component analysis. Section6.5.1, Page123.

PRAAT This is a large, sophisticated suite of software tools developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences, University of Amsterdam, for (among many other things) manipulating and synthesizing acoustic stimuli. Section 6.4.2, Page 118.

RMS Root mean square value. Section 2.4.2, Page 27.

SCM Simple cochlear model. The model of cochlear response implemented in MATLAB taken from Slaney (1994). Section 2.3.1, Page 17

SKV Skewness in variable time. Section 2.3.2, Page 19.

SNR Signal to Noise Ratio. Section 7.1, Page 138.

STA Spike triggered average, in reverse correlation experiments. This can be shown to be equivalent to the stimulus response cross correlation or C_{sr} (Theunissen *et al.*, 2001). Section 3.4, Page 53.

STRF Spectro-temporal (or spatio-temporal) response field. Section 1, Page 5.

VCN Ventral cochlear nucleus. Section 2.2.1, Page 12.

List of references.

AHISSAR, E., SOSNIK, R. & HAIDARLIU, S. (2000). Transformation from temporal to rate coding in a somatosensory thalamocortical pathway. *Nature*, **406**, 302–306. 5.4.3

AHISSAR, E., NAGARAJAN, S., AHISSAR, M., PROTOPAPAS, A., MAHNCKE, H. & MERZENICH, M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proc Natl Acad Sci U S A*, **98**, 13367–72. 5.4.3, 5.5.2, 5.5, 5.5.2, 5.6.2, 5.6.3

ALAIN, C., ARNOTT, S.R., HEVENOR, S., GRAHAM, S. & GRADY, C.L. (2001). "What" and "where" in the human auditory system. *Proc Natl Acad Sci U S A*, **98**, 12301–6. 6.2

ALAVLSI (2001). EU Open FET IST-2001-38099. 1

AMIT, D. & FUSI, S. (1992). Constraints on learning in dynamics synapses. *Network*, **3**, 443–464. 6.3.1

AMIT, D. & FUSI, S. (1994)). Learning in neural networks with material synapses. *Neural Computation*, **6**, 957–982. 6.3.1

LIST OF REFERENCES.

AMIT, Y. & MASCARO, M. (2001). Attractor networks for shape recognition. *Neural Computation*, **13 (6)**, 1415–1442. 6.3.1

ARAI, T. & GREENBERG, S. (1998). Speech intelligibility in the presence of cross-channel asynchrony. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 933–936. 5.6.2

ATICK, J.J. & REDLICH, A.N. (1992). What does the retina know about natural scenes. *Neur.Comp.*, **4**, 196–210. 2.6.2, 7.1

AUGUSTINE OF HIPPO (c.400). *Confessions*, vol. 11.27.34. 7.2

BARLOW, H.B. (1960). *The coding of sensory messages*, chap. XIII, 331–360. CUP. 2.6.2, 4.5.2

BARLOW, H.B. (1972). Single units and sensation: a neuron doctrine for perceptual psychology? *Perception*, **1**, 371–394. 2.6.2

BARLOW, H.B. & FOLDIAK, P. (1989). *The computing neuron*. Addison-Wesley New York. 7.1

BARLOW, H.D. (1961). *Current Problems in Animal Behaviour*, chap. The coding of sensory messages, 331–360. Cambridge University Press. 2.4.3

BELIN, P., ZATORRE, R.J., LAFAILLE, P., AHAD, P. & PIKE, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, **403**, 309–12. 6.2

BELIN, P., FECTEAU, S. & BDARD, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends Cogn Sci*, **8**, 129–35. 6.2

BINDER, J.R., LIEBENTHAL, E., POSSING, E.T., MEDLER, D.A. & WARD, B.D. (2004). Neural correlates of sensory and decision processes in auditory object identification. *Nature Neuroscience*, **7**, 295–301. 7.1

BOERSMA, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5:9/10**, 341–345. 6.4.2

BOERSMA, P. & WEENINK, D. (2005). Praat: doing phonetics by computer (version 4.4.07) [computer program] retrieved july 10, 2005. http://www.praat.org/. 6.4.2

BRADER, J., SENN, W. & FUSI, S. (2004). Learning real world stimuli in a neural network with spike driven synaptic dynamics. *Neural Computation*. 6.3.1, 6.3.1

BRAND, A., URBAN, R. & GROTHE, B. (2000). Duration tuning in the mouse auditory midbrain. *J Neurophysiol*, **84**, 1790–9. 2.2.3

BREGMAN, A., AHAD, P. & KIM, J. (1994). Resetting the pitch-analysis system. 2. Role of sudden onsets and offsets in the perception of individual components in a cluster of overlapping tones. *J Acoust Soc Am*, **96**, 2694–703. 2.2.1

BRENNER, N., BIALEK, W. & DE RUYTER VAN STEVENINCK, R. (2000). Adaptive rescaling maximizes information transmission. *Neuron*, **26**, 695–702. 2.3.3

BRUNGHART, D.S. & SIMPSON, B.D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *J. Acoust. Soc. Am.*, **112**, 2985–2995. 7.1

CASSEDAY, J.H., EHRLICH, D. & COVEY, E. (2000). Neural measurement of sound duration: control by excitatory-inhibitory interactions in the inferior colliculus. *J Neurophysiol*, **84**, 1475–87. 2.2.3

CHANG, E.F. & MERZENICH, M.M. (2003). Environmental noise retards auditory cortical development. *Science*, **300**, 498–502. 5.5.1, 5.6.2

CHICCA, E. & FUSI, S. (2001). Stochastic synaptic plasticity in deterministic avlsi networks of spiking neurons. In F. Rattay, ed., *Proc. of the World Congress on Neuroinformatics*, 468–477, ASIM Verlag, Vienna,. 6.3.1

COOKE, M. (2003). Glimpsing speech. *Journal of Phonetics*, **31**, 579–584. 7.1

DAYAN & ABBOT (2001). *Neural Coding*. MIT Press. 4.2.1, 1

DE BOER, R. & KUYPER, P. (1968). Triggered correlation. *IEEE Trans Biomed Eng*, **15**, 169–79. 3.3.1

DECHARMS, R., BLAKE, D. & MERZENICH, M. (1998). Optimizing sound features for cortical neurons. *Science*, **280**, 1439–43. 3.4.2

DELGIUDICE, P., FUSI, S. & MATTIA, M. (2003). Modeling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses. *J. Phys. Paris*, **97**, 659–681. 6.3.1

DENHAM, S.L. & COATH, M. (2005). A model based upon response fields derived during early experience can account for the interference effects of synthetically degraded speech signals. In *Proceedings of ISCA workshop on plasticity in speech perception.*. 7.1

DRULLMAN, R. (1995). Temporal envelope and fine structure cues for speech intelligibility. *J Acoust Soc Am*, **97**, 585–92. 2.2.1

DRULLMAN, R., FESTEN, J. & PLOMP, R. (1994a). Effect of reducing slow temporal modulations on speech reception. *J Acoust Soc Am*, **95**, 2670–80. 2.2.1

DRULLMAN, R., FESTEN, J. & PLOMP, R. (1994b). Effect of temporal envelope smearing on speech reception. *J Acoust Soc Am*, **95**, 1053–64. 2.2.1

DRULLMAN, R., FESTEN, J. & HOUTGAST, T. (1996). Effect of temporal modulation reduction on spectral contrasts in speech. *J Acoust Soc Am*, **99**, 2358–64. 2.2.1

DUYSENS, J., SCHAAFSMA, S.J. & ORBAN, G.A. (1996). Cortical off response tuning for stimulus duration. *Vision Res*, **36**, 3243–51. 2.2.3

EDELMAN, S. (1998). Representation is representation of similarities. *Behav Brain Sci*, **21**, 449–67; discussion 467–98. 5.3, 6.6.2

EDELMAN, S. (2002). Constraining the neural representation of the visual world. *Trends Cogn Sci*, **6**, 125–131. 4.2.3, 7.1

EGGERMONT, J.J. (2002). Temporal modulation transfer functions in cat primary auditory cortex: separating stimulus effects from neural mechanisms. *J Neurophysiol*, **87**, 305–21. 2.2.1

EHRLICH, D., CASSEDAY, J.H. & COVEY, E. (1997). Neural tuning to sound duration in the inferior colliculus of the big brown bat, Eptesicus fuscus. *J Neurophysiol*, **77**, 2360–72. 2.2.3

ERICKSON, R.P. (1974). *The neurosciences. Third study program*, chap. Parallel population coding in feature extraction, 155–69. MIT Press, Cambridge MA. 4.2

ERICKSON, R.P. (1986). A neural metric. *Neurosci Biobehav Rev*, **10**, 377–86. 4.3.3

ESCABI, M.A. & SCHREINER, C.E. (2002). Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. *J. Neuroscience.*, **22**, 4114–4131. 3.3.1

FAURE, P.A., FREMOUW, T., CASSEDAY, J.H. & COVEY, E. (2003). Temporal masking reveals properties of sound-evoked inhibition in duration-tuned neurons of the inferior colliculus. *J Neurosci*, **23**, 3052–65. 2.2.3

FISHBACH, A., NELKEN, I. & YESHURUN, Y. (2001). Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. *J Neurophysiol*, **85**, 2303–23. 2.2.1

FITCH, W.T. & GIEDD, J. (1999). Morphology and development of the human vocal tract: a study using magnetic resonance imaging. *J Acoust Soc Am*, **106**, 1511–1522. 6.5.3

FOXTON, J.M., DEAN, J.L., GEE, R., PERETZ, I. & GRIFFITHS, T.D. (2004). Characterization of deficits in pitch perception underlying 'tone deafness'. *Brain*, **127**, 801–10. 6.5.2

FRISINA, R.D., SMITH, R.L. & CHAMBERLAIN, S.C. (1985). Differential encoding of rapid changes in sound amplitude by second order auditory neurons. *Exp Brain Res*, **60**, 417–422. 2.2.1

FRITZ, J., SHAMMA, S., ELHILALI, M. & KLEIN, D. (2003). Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. *Nat Neurosci*, **6**, 1216–23. 5.6.2

FU, Q.J., ZENG, F.G., SHANNON, R.V. & SOLI, S.D. (1998). Importance of tonal envelope cues in Chinese speech recognition. *J Acoust Soc Am*, **104**, 505–10. 2.2.1, 2.6.2

FURUI, S. (1986). On the role of spectral transition for speech perception. *J Acoust Soc Am*, **80**, 1016–25. 7.1

FUSI, S. (2002). Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates . *Biological Cybernetics*, **87**, 459–470. 6.3.1

FUSI, S. (2003). Spike-driven synaptic plasticity for learning correlated patterns of mean firing rates. *Reviews in the Neurosciences*, **14**, 73–84. 6.3.1

FUSI, S., ANNUNZIATO, M., BADONI, D., SALAMON, A. & AMIT, D. (2000). Spike-driven synaptic plasticity: theory, simulation, vlsi implementation. *Neural Computation*, **12**, 2227–2258. 6.3.1

FUZESSERY, Z.M. & HALL, J.C. (1999). Sound duration selectivity in the pallid bat inferior colliculus. *Hear Res*, **137**, 137–54. 2.2.3

GOOLER, D.M. & FENG, A.S. (1992). Temporal coding in the frog auditory midbrain: the influence of duration and rise-fall time on the processing of complex amplitude-modulated stimuli. *J Neurophysiol*, **67**, 1–22. 2.2.3

GRIFFITHS, T.D. & WARREN, J.D. (2002). The planum temporale as a computational hub. *Trends Neurosci*, **25**, 348–53. 7.1

GRIFFITHS, T.D., WARREN, J.D., SCOTT, S.K., NELKEN, I. & KING, A.J. (2004). Cortical processing of complex sound: a way forward? *Trends Neurosci*, **27**, 181–5. 6.2

HARTMANN, W.M. (1998). *Signals, Sound, and Sensation.* AIP, Springer. 2.3.4

HATTORI, T. & SUGA, N. (1997). The inferior colliculus of the mustached bat has the frequency-vs-latency coordinates. *J Comp Physiol [A]*, **180**, 271–84. 2.2.1

HAXBY, J.V., GRADY, C.L., HORWITZ, B., UNGERLEIDER, L.G., MISHKIN, M., CARSON, R.E., HERSCOVITCH, P., SCHAPIRO, M.B. & RAPOPORT, S.I. (1991). Dissociation of object and spatial visual processing pathways in human extrastriate cortex. *Proc Natl Acad Sci U S A*, **88**, 1621–5. 6.2

HE, J. (2001). On and off pathways segregated at the auditory thalamus of the guinea pig. *J Neurosci*, **21**, 8672–9. 2.3.2, 2.5.2, 2.6.2

HE, J. (2002). OFF responses in the auditory thalamus of the guinea pig. *J Neurophysiol*, **88**, 2377–86. 2.2.2, 2.2.3, 2.6.2

HE, J., HASHIKAWA, T., OJIMA, H. & KINOUCHI, Y. (1997). Temporal integration and duration tuning in the dorsal zone of cat auditory cortex. *J Neurosci*, **17**, 2615–25. 2.2.2, 2.2.3

HEAD, P. & DENHAM, S.L. (2004). Perceptual interference between fine structure and spectrotemporal envelope in complex sounds. *Perception and Psychophysics - under review.* 6.4.2, 6.5.2, 6.6.1

HEIL, P. (1997a). Auditory cortical onset responses revisited. I. First-spike timing. *J Neurophysiol*, **77**, 2616–41. 2.2.2, 2.6.2, 3.4.3, 3.4.3

HEIL, P. (1997b). Auditory cortical onset responses revisited. II. Response strength. *J Neurophysiol*, **77**, 2642–60. 2.2.1

HEIL, P. (2001). Representation of sound onsets in the auditory system. *Audiol Neurootol*, **6**, 167–72. 1, 2.6.2

HEIL, P. & IRVINE, D. (1996). On determinants of first-spike latency in auditory cortex. *Neuroreport*, **7**, 3073–6. 2.2.1

HELMHOLTZ, H.L.F. (1860). *Handbuch der physiologischen Optik*. Leopold Voss:Hamburg. 4.2

HUBEL, D.H. & WIESEL, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J Physiol*, **160**, 106–54. 3.2, 7.1

HULL, A. (1973). A letter-digit matrix of auditory confusions. *Br J Psychol*, **64**, 579–85. 6.5.1, 6.5(b)

ILLING, R.B. (2004). Maturation and plasticity of the central auditory system. *Acta Otolaryngol Suppl*, **552**, 6–10. 4.5.2

INDIVERI., G. (2002). *Advances in Neural Information Processing Systems , Vol. 15*, chap. Neuromorphic bistable VLSI synapses with spike-timing-dependent plasticity. MIT Press., Cambridge, MA. 6.3.1

KAAS, J.H. & HACKETT, T.A. (1999). 'What' and 'where' processing in auditory cortex. *Nat Neurosci*, **2**, 1045–7. 6.2

Koch, C. & Ullman, S. (1985). Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, **4**, 219–27. 4.3.3

Kriegstein, K., Eger, E., Kleinschmidt, A. & Giraud, A.L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. *Brain Res Cogn Brain Res*, **17**, 48–55. 6.2, 6.6.2, 7.1

Krumbholz, K., Patterson, R.D., Seither-Preisler, A., Lammertmann, C. & Ltkenhner, B. (2003). Neuromagnetic evidence for a pitch processing center in Heschl's gyrus. *Cereb Cortex*, **13**, 765–72. 2.6.2, 3.5.2

Lancker, D.R.V., Kreiman, J. & Cummings, J. (1989). Voice perception deficits: neuroanatomical correlates of phonagnosia. *J Clin Exp Neuropsychol*, **11**, 665–74. 6.2

Langner, G. & Schreiner, C. (1988). Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms. *J Neurophysiol*, **60**, 1799–822. 2.2.1

Linden, J.F., Liu, R.C., Sahani, M., Schreiner, C.E. & Merzenich, M.M. (2003). Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. *J Neurophysiol*, **90**, 2660–75. 5.6.2

Machens, C.K., Wehr, M.S. & Zador, A.M. (2004). Linearity of cortical receptive fields measured with natural sounds. *J Neurosci*, **24**, 1089–100. 5.6.2, 7.1

Metherate, R., Kaur, S., Kawai, H., Lazar, R., Liang, K. & Rose, H.J. (2005). Spectral integration in auditory cortex: mechanisms and modulation. *Hear Res*, **206**, 146–158. 2.6.2

MILLER, L.M., ESCAB, M.A., READ, H.L. & SCHREINER, C.E. (2002). Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. *J Neurophysiol*, **87**, 516–27. 4.4

NELKEN, I., FISHBACH, A., LAS, L., ULANOVSKY, N. & FARKAS, D. (2003). Primary auditory cortex of cats: feature detection or something else? *Biol Cybern*, **89**, 397–406. 5.6.2

NOORDHOEK, I. & DRULLMAN, R. (1997). Effect of reducing temporal intensity modulations on sentence intelligibility. *J Acoust Soc Am*, **101**, 498–502. 2.6.2

OBLESER, J., ELBERT, T. & EULITZ, C. (2004). Attentional influences on functional mapping of speech sounds in human auditory cortex. *BMC Neurosci*, **5**, 24. 6.2, 6.5.4, 6.6.2, 7.1

OGI (1996). Oregon health and science university: The speaker recognition corpus v1.1. 6.4.3

OGI (1999). Oregon health and science university: The isolet corpus v1.3. 6.4.1

PATTERSON, R.D., ROBINSON, K., HOLDSWORTH, J., MCKEOWN, D., CZHANG & ALLERHAND, M.H. (1992). *Auditory physiology and perception.*, 429–446. Oxford. A.2

PHILLIPS, D., HALL, S. & BOEHNKE, S. (2002). Central auditory onset responses, and temporal asymmetries in auditory perception. *Hear Res*, **167**, 192–205. 2.2.1, 2.2.2, 2.6.2

PONS, T.P., GARRAGHTY, P.E. & MISHKIN, M. (1992). Serial and parallel processing of tactual information in somatosensory cortex of rhesus monkeys. *J Neurophysiol*, **68**, 518–27. 6.2

RHODE, W. & GREENBERG, S. (1994). Encoding of amplitude modulation in the cochlear nucleus of the cat. *J Neurophysiol*, **71**, 1797–825. 2.2.1

ROÜILLER, E. & DE RIBAUPIERRE, F. (1982). Neurons sensitive to narrow ranges of repetitive acoustic transients in the medial geniculate body of the cat. *Exp Brain Res*, **48**, 323–6. 2.2.1

ROUILLER, E., DE RIBAUPIERRE, Y., TOROS-MOREL, A. & DE RIBAUPIERRE, F. (1981). Neural coding of repetitive clicks in the medial geniculate body of cat. *Hear Res*, **5**, 81–100. 2.2.1

SAHANI, M. & LINDEN, J.F. (2003). How linear are auditory cortical responses? 3.4.2

SCHNUPP, J.W.H., GARCIA-LAZARO, J.A. & AHMED, B. (2005). Tuning to natural stimulus dynamics in primary auditory coretx, poster for SFN 2004. 2.4.3, 2.4.3, 2.6.1, 2.6.2, 7.1, 7.1

SENN, W. & FUSI, S. (2004). Slow stochastic learning with global inhibition: a biological solution to the binary perceptron problem. *Neurocomputing*, **58-60**, 321–326. 6.3.1

SHAMMA, S. (2001). On the role of space and time in auditory processing. *Trends Cogn Sci*, **5**, 340–348. 2.2.1

Shannon, C.E. (1948). A mathematical theory of communication. *AT&T Bell Labs Tech. J.*, **27**, 379–423. 4.2.1

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J. & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, **270**, 303–4. 2.2.1

Shannon, R.V., Zeng, F.G. & Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues. *J Acoust Soc Am*, **104**, 2467–76. 2.2.1

Shepard, R. & Chipman, S. (1970). Second-order isomorphism of internal-representations: shapes of states. *Cognitive Psychology*, **1**, 1–17. 4.2.3, 5.3, 6.6.3

Singh, N.C. & Theunissen, F.E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *J Acoust Soc Am*, **114**, 3394–411. 2.4.3

Slaney, M. (1993). An efficient implementation of the patterson-holdsworth auditory filter bank. apple technical report 35. Tech. rep., Apple Computer Inc. 2.3.1

Slaney, M. (1994). Auditory toolbox documentation. technical report 45. Tech. rep., Apple Computers Inc. 2.3.1, A.2, A.8.1

Smith, L. (1995). Using an onset-based representation for sound segmentation. In *Proceedings of NEURAP95*. 2.2.1

STEINSCHNEIDER, M., VOLKOV, I.O., NOH, M.D., GARELL, P.C. & HOWARD, M.A. (1999). Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. *J Neurophysiol*, **82**, 2346–57. 7.1

THEUNISSEN, F.E., DAVID, S.V., SINGH, N.C., HSU, A., VINJE, W.E. & GALLANT, J.L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, **12**, 289–316. 3.3.1, 1, 3.3.1, 3.4, 1, A.8.1

TRUSSEL, L.O. (2002). *Integrative functions in the mammalian auditory pathway.*, chap. Cellular mechanisms for information coding in auditory brainstem nuclei., 72–98. Springer. 2.6.2

TURNER, R.E. & WALTERS, T.C. (2004). *BSA Short papers meeting*. 6.5.3

ULLMAN, S., VIDAL-NAQUET, M. & SALI, E. (2002). Visual features of intermediate complexity and their use in classification. *Nat Neurosci*, **5**, 682–7. 4.2.2, 4.2.3, 4.2.4, 4.3.2, 5.6.2

VANCAMPEN, L.E., HALL, J.W. & GRANTHAM, D.W. (1997). Human offset auditory brainstem response: effects of stimulus acoustic ringing and rise-fall time. *Hear Res*, **103**, 35–46. 2.2.2, 2.5.2, 2.5.2, 2.5.2, 2.10, 2.11, 2.6.2, 3.4.3

VOSS & CLARKE (1975). 1/f noise in music and speech. *Nature*, **258**, 317–318. 2.4.3

WANG, X. (2004). The unexpected consequences of a noisy environment. *Trends Neurosci*, **27**, 364–366. 4.5.2, 7.1

WHITESIDE, S. (1998). Identification of a speaker's sex: a study of vowels. *Percept Mot Skills*, **86**, 579–84. 6.5.3, 6.6.1

WIEGREBE, L. (2001). Searching for the time constant of neural pitch extraction. *J Acoust Soc Am*, **109**, 1082–91. 2.3.2, 2.6.2, A.1, A.3, A.7.1

YOUNG, T. (1802). On the theory of light and colours. *Philo Trans R Soc Lond*, **92**, 12–48. 4.2

YU, L. & LIU, H. (2003a). Efficiently handling feature redundancy in high-dimensional data. In *The Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-03), Washington D.C., August 2004*. 4.3.2

YU, L. & LIU, H. (2003b). Feature election for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington, 2003*. 4.2.5, 4.3.2, 6.5, 6.6.1, A.8.1

YU, Y., ROMERO, R. & LEE, T.S. (2005). Preference of sensory neural coding for 1/f signals. *Phys Rev Lett*, **94**, 108103. 2.4.3

ZHANG, L., BAO, S. & MERZENICH, M. (2001). Persistent and specific influences of early acoustic environments on primary auditory cortex. *Nat Neurosci*, **4**, 1123–30. 4.5.2, 5.5.1, 5.6.2, 7.1

ZHANG, L.I., BAO, S. & MERZENICH, M.M. (2002). Disruption of primary auditory cortex by synchronous auditory inputs during a critical period. *Proc Natl Acad Sci U S A*, **99**, 2309–14. 4.5.2, 5.2

Bound copies of published papers.

Biol. Cybern. (2005) 93: 22–30
DOI 10.1007/s00422-005-0560-4

ORIGINAL PAPER

Biological
Cybernetics

Martin Coath · Susan L Denham

Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience

Received: 10 August 2004 / Accepted: 18 March 2005 / Published online: 8 June 2005

Abstract Models of auditory processing, particularly of speech, face many difficulties. Included in these are variability among speakers, variability in speech rate, and robustness to moderate distortions such as time compression. We constructed a system based on ensembles of feature detectors derived from fragments of an onset-sensitive sound representation. This method is based on the idea of 'spectro-temporal response fields' and uses convolution to measure the degree of similarity through time between the feature detectors and the stimulus. The output from the ensemble was used to derive segmentation cues and patterns of response, which were used to train an artificial neural network (ANN) classifier. This allowed us to estimate a lower bound for the mutual information between the class of the input and the class of the output. Our results suggest that there is significant information in the output of our system, and that this is robust with respect to the exact choice of feature set, time compression in the stimulus, and speaker variation. In addition, the robustness to time compression in the stimulus has features in common with human psychophysics. Similar experiments using feature detectors derived from fragments of non-speech sounds performed less well. This result is interesting in the light of results showing aberrant cortical development in animals exposed to impoverished auditory environments during the developmental phase.

1 Introduction

How sounds are represented in auditory cortex is still unclear. The prevailing view of visual object classification is that the visual system is organized hierarchically and it is within this hierarchy that features of increasing complexity and spatial extent are analyzed. In this context, it has been shown that visual fragments of intermediate complexity and intermediate extent are optimal for the classification of visual objects (Ullman et al. 2002). An analogous hierarchical organization in auditory cortex has not yet been identified, and the representations employed in auditory cortex are poorly understood (Nelken et al. 2003). We were interested to discover whether acoustic fragments of intermediate spectral and temporal extent might similarly be useful for the classification of auditory events, and whether such fragments could be derived from natural sounds.

A popular method for characterizing the response fields of cortical cells is the spectro-temporal response field (STRF), e.g. (Kowalski et al. 1996a,b; deCharms et al. 1998; Depireux et al. 2001; Theunissen et al. 2001; Miller and et al. 2002; Elhilali et al. 2004). Some time ago it was shown how reverse correlation in response to white noise could be used to characterize STRFs (Aertsen and Johannesma 1981), and this approach has been extended allowing STRFs to be derived from natural stimuli (Theunissen et al. 2001; Miller and et al. 2002). In principle, given an assumption of linearity, the response of a cell to a novel stimulus can be predicted by convolving its STRF with a spectro-temporal representation of the stimulus. Although the linearity assumption has been called into question (Bar-Yosef et al. 2002), and seems to be stimulus-dependent (Machens et al. 2004), STRFs can often predict neuronal responses very well (Elhilali et al. 2004) and this representation of a cell's response field can provide valuable insights into the factors influencing neuronal behaviour.

There is a great deal of evidence to show that the auditory system is interested in change (Phillips et al. 2002), or spectral and temporal 'edges' (deCharms et al. 1998). In addition, it has been shown that a model based upon derivatives of the stimulus envelope can successfully account for many aspects of neural and perceptual responses to amplitude transients (Fishbach et al. 2001). Here, we use a representation which emphasizes changes in the stimulus envelope, such as onsets and offsets, and de-emphasizes unchanging activity. To do this, we extract the short-term third-order moment or 'skewness' of the signal within each frequency channel. In this model, the term STRF refers to a spectrogram-like response

M. Coath (✉) · S. L. Denham
Centre for Theoretical and Computational Neuroscience,
University of Plymouth, Drakes Circus, PL4 8AA, UK
E-mail: martin.coath@plymouth.ac.uk,
Tel.: +44-01752-232704)

field consisting of the short-term skewness within a range of frequency channels; a form of higher-order spectrogram (Nikias and Athina 1993).

In what sense can an ensemble of STRFs represent a sound? A useful way of thinking about this problem first proposed by Shepard (Shepard and Chipman 1970), and later elaborated by Edelman (Edelman 1998), is in terms of a second-order isomorphic mapping from the physical world to the response space. This is a different and far more powerful representation than the more commonly used first-order isomorphism in which a stimulus is represented in terms of its similarity to some prototype. In a second-order isomorphic mapping, all that is required is that the similarities between stimuli in the world are preserved in the similarities between their projections into the low-dimensional space spanned by the outputs of a small number of roughly tuned detectors. As Edelman stresses, this is a representation *of* similarity, not *by* similarity. The convolution of an STRF with an incoming sound can be interpreted as a measure of the similarity between that sound and the response field. So, in principal, any sound can be positioned in the response space spanned by an ensemble of STRFs. The question is whether the similarities in the physical world are preserved in this projection, and to what extent the mapping is robust to the variability inherent in natural sounds.

It has recently been shown that sounds experienced during an early critical period influence the organization of primary auditory cortex and the response fields of cortical cells (Zhang et al. 2001, 2002). It has previously been suggested (Terhardt 1974) that the perception specifically of pitch might be influenced by development. However, the work of Merzenich and colleagues has suggested that a broad range of useful neuronal response fields might develop through experience of a limited number of environmental sounds. In humans, an important source of early acoustic experience is speech, but the number of speakers and words heard in early life are likely to be rather limited. Therefore, an important aspect of the proposed representation is that it is 'productive', in the sense that it is possible to represent a novel class within the response space. All that is required to do this is that the projections from exemplars of the novel class cluster appropriately within the response space. As a putative model of cortical processing, this is important since it is obviously necessary to be able to learn to classify novel sounds after the critical, plastic period of development.

In summary, our investigations were aimed at addressing the following questions. Can useful response fields be derived from fragments of a limited number of sounds? What size of acoustic fragment, in terms of spectral and temporal extent, is best for classification? Can the responses of an ensemble of fields be understood in terms of a second-order isomorphism between stimulus class and ensemble response? Can the ensemble response convey significant information with respect to stimulus class? In order to address these questions, we used speech data, partly because of the overwhelming importance of speech for human audition, (analogous to that of faces for vision) and partly because the classification of speech is far better understood than that of other sounds. We used a speech database containing a large number of examples of a small number of classes (words). Response fields were constructed from acoustic fragments of varying temporal and spectral extent extracted from the utterances of a single speaker. These response fields were then convolved with utterances from a large number of different speakers, and the mutual information between the ensemble response and actual stimulus class was characterized.

2 Method

The model, whose operation is illustrated in Fig. 1, consists of six principal processing stages: spectral decomposition, extraction of envelope transients, convolution using a bank of STRFs, event detection, mapping to response space, and classification.

2.1 Spectral decomposition

The first stage approximates processing in the cochlea. Sounds are processed using a bank of 24 Gammatone filters (Slaney 1994), with centre frequencies, ranging from 100 Hz to ~ 4000 Hz arranged evenly on an ERB scale (Glasberg and Moore 1990). The output in each frequency channel is low-pass filtered, half wave rectified, and compressed using a sigmoidal function (Eq. 1);

$$y = 1 - \left(\frac{1}{1 + \exp -\alpha(1/2 - x)} \right), \tag{1}$$

where x is the input (the instantaneous energy in a filter channel), y the compressed output, and α a constant which controls the degree of compression. We used $\alpha = 7$, which provides moderate compression (between 2:1 and 6:1) for moderately high values of x.

2.2 Transient extraction

The next stage of processing enhances envelope transients within each frequency channel, as is found in the subcortical auditory system (Phillips et al. 2002). In this simple model, we do not consider the extraction of any other acoustic features. The mean level of activity within each channel is calculated in overlapping temporal windows of duration twice the period of the centre frequency but with a minimum window size of 2.5 ms at high frequency (Wiegrebe 2001). The overlap for all experiments was set to 10% of the window duration. The skewness, z, of the distribution of energy across four successive windows is then calculated;

$$z = \frac{1}{N} \sum_{j=1}^{N} \left(\frac{x_j - \overline{x}}{\sigma} \right)^3, \tag{2}$$

where N is the number of windows, x_j is the energy in the jth window, $\overline{x}$ is the mean, and σ is variance. Short-term

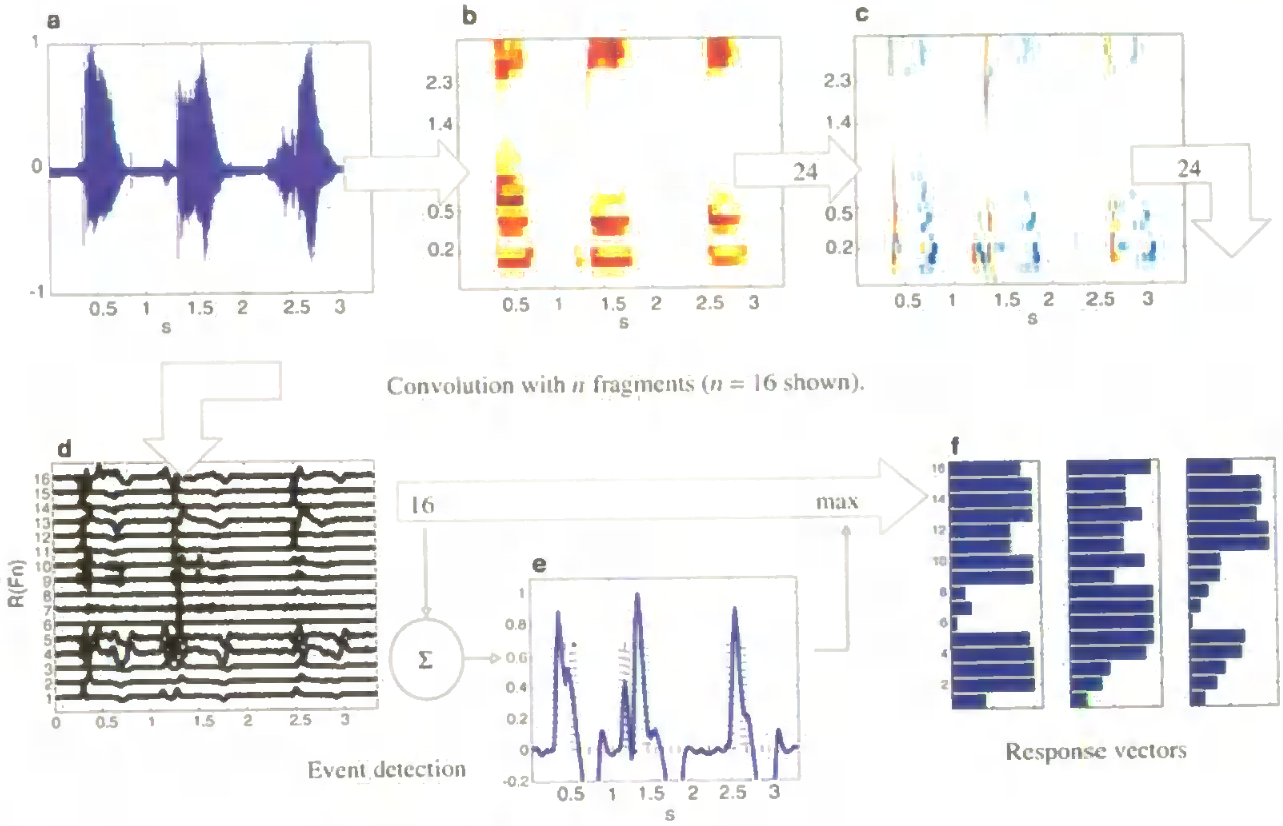


Fig. 1 A summary of the key processing stages in the model. See Sects. 2.1 to 2.5

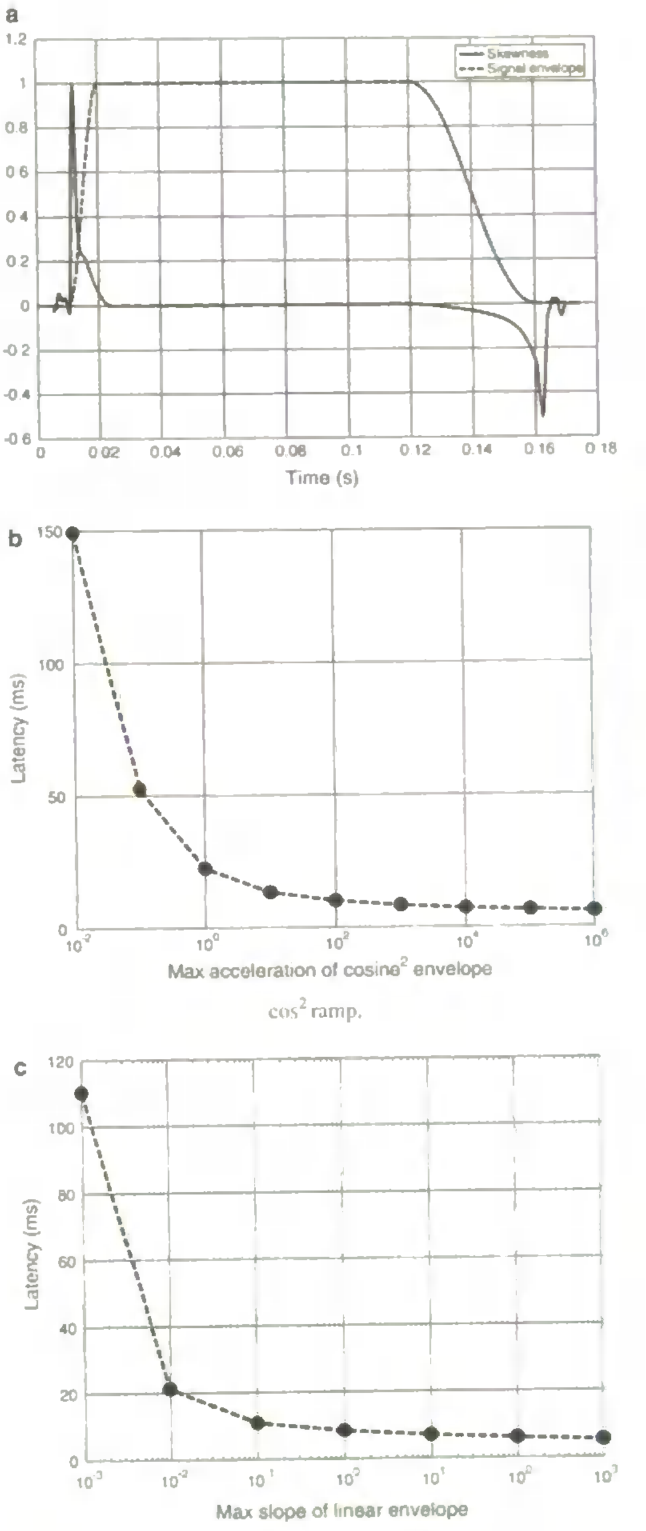
skewness is a sensitive indicator of rising and falling energy and has a value near zero when the energy is approximately unchanging, as illustrated in Fig. 2. One advantage of calculating energy distributions within short time windows is that of a rapidly adapting threshold, and hence a roughly level independent representation (Phillips et al. 2002). Furthermore, the maximum skewness is locked to the onset of the transient and its timing depends upon the dynamics of the onset envelope, as found experimentally (Heil 1997). In effect, this processing amounts to edge detection in the temporal domain and the result is a spectro-temporal map of envelope transients in response to the processed sound as illustrated in Fig. 1c. Furthermore, the growth of short-term skewness depends on the dynamics of the transients, and varies with the maximum rate of change and acceleration in a way which is consistent with neural behaviour (Heil and Irvine 1997). Specifically, the latency at which the integrated short term skewness exceeds some threshold can be related to the maximum acceleration (for $\cos^2$ ramps) or to maximum rate of change (for linear ramps) in a way which is very similar to the first spike latencies measured in auditory cortex (Heil and Irvine 1997; Heil and Neubauer 2001).

2.3 Convolution using a bank of STRFs

Each STRF is specified in terms of a pattern of onsets and/or offsets extending over a specified spectral range and duration. Each member of the ensemble is convolved with the transient incoming sound pattern, thereby generating a 'temporal signature', which indicates the degree of similarity between the incoming pattern and the STRF at each point in time; as illustrated in Fig. 1d for an ensemble of 16 STRFs.

2.4 Event detection and mapping to response space

The response of all STRFs in the ensemble (Fig. 1e) provides an indication of the presence of an acoustic event, the timing and duration of which is determined both by the stimulus *and* by the ensemble used. These events are detected on the scale of tens of milliseconds and this is distinct from the short time scale event detection such as that used by Irino et al. (2005) to identify glottal pulses and hence aid the estimation of the fundamental frequency. The stimulus-ensemble-driven event detection results in a method of segmentation where auditory events are marked by coherence in the response of the



cos^2 ramp.

Linear ramp.

Fig. 2 **a** An example of the short-term skewness in response to rising and falling cos^2 ramp. **b, c)** The latency of integrated short-term skewness depends on envelope shape in a very similar way to that measured experimentally (Heil 1997)

ensemble and not wholly by properties of the stimulus. In these experiments, we summed the output of the ensemble and recorded the maximum response of each STRF within the period during which the summed response exceeded a threshold value (20% of the maximum). The result is a vector defining a point in the N-dimensional space spanned by the responses of the N STRFs (Fig. 1f). It is possible for a sound to generate more than one such event, but in the experiments described below when this occurred only the first event was classified.

2.5 Classification

In order to assign a class to each response, we trained 11 separate ANN classifiers each with N inputs (where N was the ensemble size), five hidden units, and one output unit. Log-sigmoidal units were used for hidden and output nodes. For each training, the data were divided 70, 15, and 15% in to training, validation, and test sets, respectively. We employed early stopping based on the validation set to avoid over-fitting. The output from each of the 11 classifiers formed the input to a winner-take-all stage, which assigned the stimulus to a class based on the classifier with the highest output. Although the ANN formed no part of the model, it provided a convenient way to estimate the mutual information between the stimulus class and the response by assigning each response a class. In order to measure the effectiveness of the model, we quantified the mutual information $I(S; R)$ between the classes of the stimuli S and the classifications made by the ANN; these can be thought of as the 'responses' R;

$$I(S; R) = \left\langle \sum_{s} P(s|r) \log_2 \left[\frac{P(s|r)}{P(s)} \right] \right\rangle_r , \qquad (3)$$

where $P(s|r)$ is the conditional probability of the stimulus class s given the response class r, $P(s)$ is the probability of class s, and $< \cdots >_r$ represents the average over the (unconditional) response distribution (Golomb et al. 1997). It is important to note that we are not characterizing the mutual information between the stimulus and the response, but between the *class of the stimulus* and the *class of the response*.

3 Fragment extraction

In order to explore the possibility that the formation of STRFs may be bootstrapped by fragments of activity patterns in response to acoustic stimuli, several libraries of fragments derived from small sets of sounds were created. The first set contained samples of a single speaker saying each of the numerals; 'one', 'two' ... 'nine', 'Oh', and 'Zero' making 11 classes. The second set consisted of eleven non-speech sounds such as environmental noises (wind and rain, bird and frog calls) and some mechanically or electronically produced sounds (engine noise, dialling telephones, colliding pool balls). From each of these sets, four separate libraries of fragments, with durations 10, 50, 100, and 200 m, were created. Fragments within each library were either 4, 8, 12,

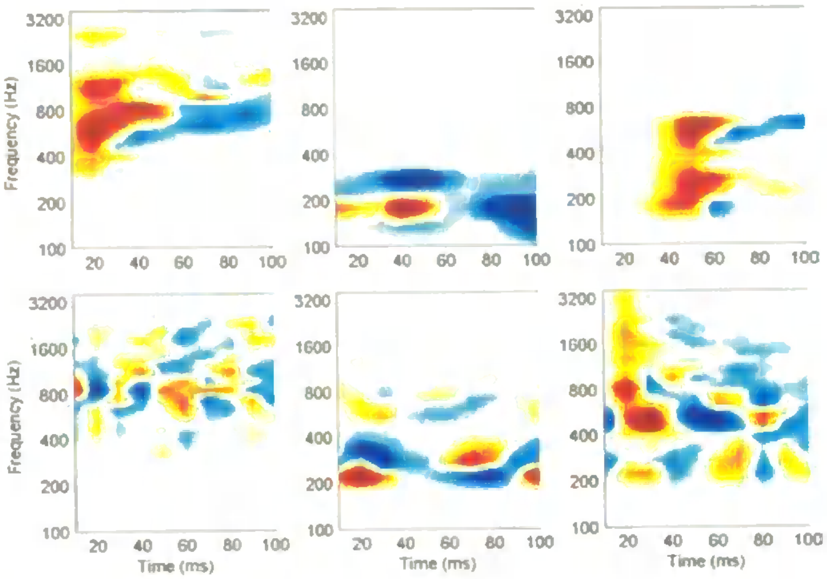


Fig. 3 Examples of fragments contained in the libraries. *Top row* – examples from the speech fragments library. *Bottom row* – examples from the library derived from non-speech sounds

16, 20, or 24 frequency bands wide. See Fig. 3 for examples of speech and noise fragments.

In contrast to Ullman et al. (2002), we did not seek to optimize the fragment choices with respect to the set of all stimuli. As the basis for a decision as to which fragment ensembles were likely to perform better, random ensembles of 2, 4, 8, and 16 fragments were generated (5000 of each) and the entropy (H) of each of their responses to the 11 stimuli from which they were drawn was calculated. The entropy gives some measure of how 'interesting' the response is (Dayan and Abbot 2001), and we want to keep only those ensembles that are maximally 'interesting'. Entropy is calculated based on the log of the probability of a response, $-\log P[r]$, averaged over all responses (Eq. 4).

$$H = -\sum_r P[r] \log_2 P[r]\,. \tag{4}$$

Distributions of the entropy of random ensembles containing 2, 4, 8, and 16 fragments from each of the four speech fragment libraries of varying temporal extent are shown in Fig. 4. The entropy values are divided into seven bins, the rightmost bin representing ensembles with the most interesting responses, i.e. those with entropies of $\sim \log_2(11)$ (3.46 bits). Ensembles from the highest and lowest entropy bins were saved for subsequent comparison (see Sect. 4.1). It is interesting to note that there is a slight bias towards 100-ms fragments in the higher-entropy bins.

4 Experiments

4.1 Generalization

We first investigated the generalization capabilities of the proposed approach, and considered the ability of the model to cope with the natural variability in speech sounds. The responses of ensembles of speech-derived STRFs were tested using utterances from over 300 male and female speakers, using recordings with signal-to-noise ratios between 8 and 25 dB. The mutual information between the classifier output and the stimulus class was calculated. The results are plotted in Fig. 5. We found that there is significant mutual information between the stimulus class and model classification, and that mutual information improves with fragment duration, and with the number of fragments in the ensemble. This suggests that there is some form of clustering which is robust to the variability present in normal speech. Included in these results are data from the 'low-entropy' ensembles (see Sect. 3) showing that these perform less well on generalization.

In order to discover whether the model was very sensitive to the precise nature of early experience, we used ensembles of STRFs derived from the set of environmental noises described previously and once again trained the system to classify the digit utterances from a large number of speakers. The results, also shown in Fig. 5, are very interesting, for although these fragments convey less information, they nevertheless do rather well. This suggests that the classification of sounds on the basis of projections into a response space spanned by a set of STRFs is, perhaps surprisingly, not very sensitive to the precise nature of the receptive fields used. However, the fact that they perform less well is consistent with experimental findings (Zhang et al. 2001; Chang and Merzenich 2003) showing that in an extremely restricted early auditory environment the auditory cortex fails to develop properly. This result also establishes the 'productivity' of the system in that the responses of fragments can be used to classify sounds very different from the ones from which they were derived (see Discussion, Sect. 5).

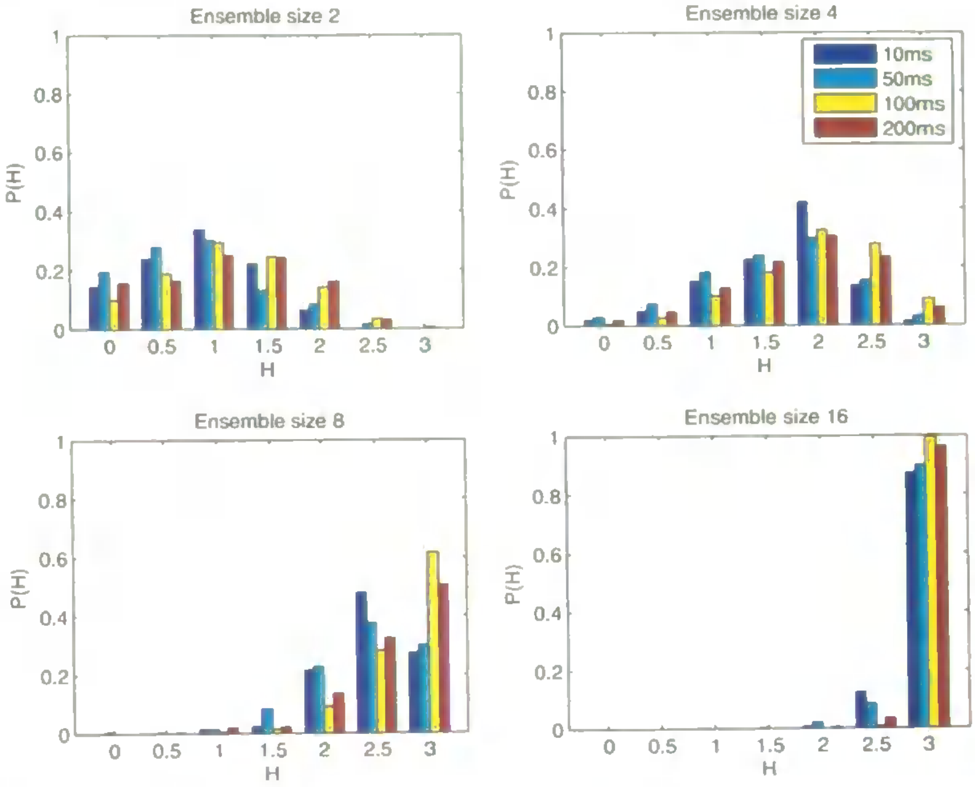


Fig. 4 Distribution of response entropy of different ensemble sizes from each of the four libraries

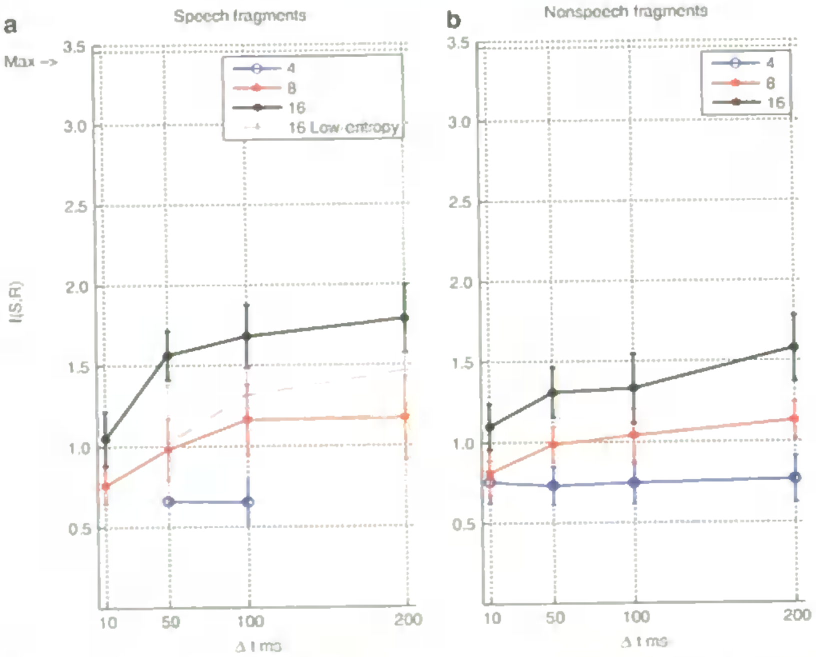


Fig. 5 Mutual information between classifier output and stimulus class for fragment ensembles of size 4 (*blue*), 8 (*red*), and 16 (*black*), and varying temporal extent (abscissa). This is plotted for both **a** speech and **b** non-speech fragments. *Dotted line* on **a** shows results for 'low-entropy' ensembles

4.2 Robustness to time compression

Finally, we considered the robustness of the STRF response to time compressed speech. The results are plotted in Fig. 6. We were interested to discover whether the performance of the model would parallel that shown in human psychophysics. In a recent experiment, it was found that speech comprehension in people is quite robust to compression up to about 50% and thereafter degraded substantially (plotted in Fig. 6b) (Ahissar et al. 2001). This work was interesting in that it showed that speech comprehension could be predicted by the degree of phase locking between cortical activity measured by MEG and the temporal envelope of the speech. One explanation for the phase locking could be the degree to which the STRFs in auditory cortex are able to respond to incoming spectro-temporal patterns, i.e. the observed phase locking may simply be a by-product of the degree of similarity between the STRFs of cells in auditory cortex and the spectro-temporal pattern of the sounds. An interesting aspect of the model's performance is that it suggests that fragments with temporal extent between 50 ms and 100 ms correlate best with human performance. This is consistent with the suggestion that the phase locking is best within the range of spontaneous and evoked cortical oscillations (~ 14 Hz) (Ahissar et al. 2001).

5 Discussion

We have built a simple representation of sounds based on an onset-, or change-sensitive measure. The change detection is level independent, as the skewness measure provides a rapid adaptive level adjustment. Change is also detected independently within each frequency channel of the cochlear model. For this reason, it does not rely on synchrony across many, or all, channels to characterize an onset; this allows us to use different time scales in each channel, and to detect onsets that occur in a narrow spectral domain. The result is a response which is not only independent of signal level but is also independent of the differences in energy between channels (the spectral profile), which is characteristic of the cochlear stage of the model. The method is robust to noise of any type that is stationary within the time constant of the band within which it is present.

Using convolution as a measure of similarity between this representation of the stimulus and an ensemble of roughly tuned, spectro-temporal detectors, sounds were represented by means of patterns of activity within the ensemble. The response of the ensemble can be understood as a projection into a low-dimensional space spanned by the outputs of the detectors. Our ANN classifier serves to demonstrate that this isomorphism is not arbitrarily complex and, by labelling the output as belonging to a single class, to act as a mechanism for estimating the lower bound of the mutual information. The transformation into low-dimensional space may be understood as essentially a second-order isomorphic mapping – which may be organized in a hierarchical fashion to extend to longer duration stimuli.

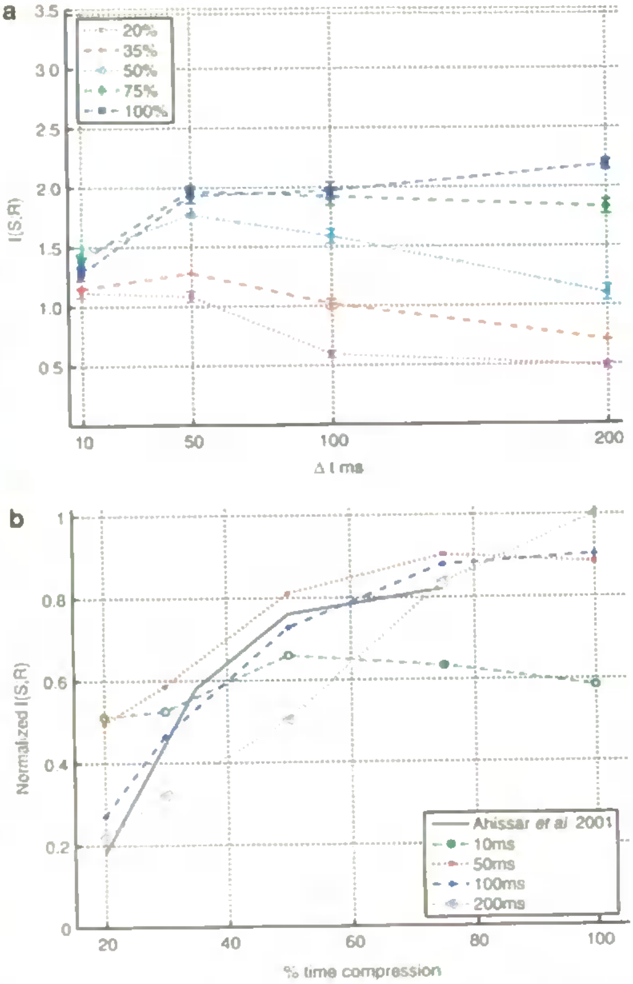


Fig. 6 Mutual information between model classification and stimulus class in response to time-compressed speech

Importantly, although performance is robust to the precise choice of fragments, there is a basis for preferring some ensembles over others i.e. the entropy of their responses to the formative sounds (Fig. 5). This provides a developmental pressure for the refinement of the ensemble choice that does not require experience of the complete set of all possible stimuli. Our ensembles were generated randomly but in subsequent work (not reported here) using a larger set of formative classes, we have established that ensembles may perform better if composed of fragments whose individual responses to the formative classes is of intermediate (neither high nor low) entropy. We found that, given that the formation of the candidate STRFs was stimulus driven, the mutual information between the input and output classes was greater if the formative stimuli were to some extent representative of the sounds to which the system is subsequently exposed. There is a parallel here with results which show that the formation of the auditory cortex is dependant on the richness of the early auditory environment, e.g. (Chang and Merzenich

2003; Zhang et al. 2001). In our experiments, STRFs were abstracted from a limited set of speech and non-speech sources; however, because our non-speech sources were quite rich (not just tone bursts for instance), they still produced significant information preserving representations.

Crucial to the success of this model is the ability to segment the incoming stimulus, that is to characterize the auditory events which provide the basis for classification. To achieve this we used properties, not of the signal, or the onset sensitive representation of the signal, but properties of the response of an ensemble of detectors. This makes the segmentation dependant on the choice of detectors and provides a mechanism whereby segmentation can become an active part of audition under conscious or attentive control. In speech, some form of segmentation is necessary to distinguish discrete percepts and to make speech perception robust to rate variation. In the absence of interference, humans can do this in the presence of significant 'cross channel asynchrony' (Arai and Greenberg 1998) and are also capable of integrating cues identifying speech sounds which do not occur simultaneously (Buss et al. 2003). This implies that there is a window during which almost-synchronous events are grouped. Using patterns of activity in our candidate STRFs to provide this window results in features being integrated over a 'context' period (Nelken et al. 2003) of typically 200–350 ms. Patterns of outputs from STRFs in AI could be used for event detection, segmentation, and object identification by one or more of the many areas of the brain to which it is connected.

Our results using time-compressed speech stimuli closely parallel human psychophysics when features are extracted on time scales of ~ 100 ms (Fig. 6). This experiment suggests that, as in vision (Ullman et al. 2002), fragments of intermediate extent may be optimal. In auditory processing, this may be because they provide robustness to the variation in the stimulus rate; alternatively the degree to which our perception is robust to time compression may be limited by the extent of STRFs. One hundred milliseconds is considerably longer than the acoustic models typically used in automatic speech recognition systems, and falls into an intermediate position between phonemes of roughly 40 ms and syllables of typically 200 ms. Rates of spontaneous and evoked cortical oscillations may help to explain the psychophysics (Ahissar et al. 2001) and the temporal extent of cortical STRFs by establishing the perceptual time scale on which auditory events are identified. Although there is no electrophysiology from humans, STRFs of ~ 100 ms are broadly consistent with results from animals such as mice (Linden et al. 2003), rats (Machens et al. 2004), and ferrets (Fritz et al. 2003).

Acknowledgements This work was supported by EU Open FET IST-2001-38099. Many thanks to Ehud Ahissar for providing the data for re-plotting in Fig 6b.

References

Aertsen A, Johannesma P (1981) A comparison of the spectro-temporal sensitivity of auditory neurons to tonal and natural stimuli. Biol Cybern 42(2):145–56

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich M (2001) Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. Proc Natl Acad Sci USA 98(23):13367–13372

Arai T, Greenberg S (1998) Speech intelligibility in the presence of cross-channel asynchrony. In: IEEE international conference on acoustics, speech and signal processing, pp 933–936

Bar-Yosef O, Rotman Y, Nelken I (2002) Responses of neurons in cat primary auditory cortex to bird chirps: effects of temporal and spectral context. J Neurosci 22(19):8619–8632

Buss E, Hall JW, Grose JH (2003) Spectral integration of synchronous and asynchronous cues to consonent identification. J Acoust Soc Am 115(5):2278–2285

Chang EF, Merzenich MM (2003) Environmental noise retards auditory cortical development. Science 300(5618):498–502

Dayan, Abbot (2001) Neural Coding. MIT, Cambridge

deCharms R, Blake D, Merzenich M (1998) Optimizing sound features for cortical neurons. Science 280(5368):1439–1443

Depireux DA, Simon JZ, Klein DJ, Shamma S (2001) Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85(3):1220–1234

Edelman S (1998) Representation is representation of similarities. Behav Brain Sci 21(4):449–467; discussion 467–498

Elhilali M, Fritz JB, Klein DJ, Simon JZ, Shamma SA (2004) Dynamics of precise spike timing in primary auditory cortex. J Neurosci 24(5):1159–72, DOI 10.1523/JNEUROSCI.3825-03.2004, URL `http://dx.doi.org/10.1523/JNEUROSCI.3825-03.2004`

Fishbach A, Nelken I, Yeshurun Y (2001) Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. J Neurophysiol 85(6):2303–2323

Fritz J, Shamma S, Elhilali M, Klein D (2003) Rapid task-related plasticity of spectrotemporal receptive fields in primary auditory cortex. Nat Neurosci 6(11):1216–1223

Glasberg BR, Moore BC (1990) Derivation of auditory filter shapes from notched noise data. Hear Res 47(1):103–138

Golomb D, Hertz J, Panzeri S, Treves A, Richmond B (1997) How well can we estimate the information carried in neuronal responses from limited samples? Neural Comput 9(3):649–665

Heil P (1997) Auditory cortical onset responses revisited. II. Response strength. J Neurophysiol 77(5):2642–2660

Heil P, Irvine D (1997) First-spike timing of auditory-nerve fibers and comparison with auditory cortex. J Neurophysiol 78(5):2438–2454

Heil P, Neubauer H (2001) Temporal integration of sound pressure determines thresholds of auditory-nerve fibers. J Neurosci 21(18):7404–7415

Irino T, Patterson RD, Kawahara H (2005) In Speech separation by humans and machines., Kluwer Academic, Massachusetts, chap Speech segregation using an event synchronous auditory image and STRAIGHT, pp 153–165

Kowalski N, Depireux D, Shamma S (1996a) Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. J Neurophysiol 76(5):3503–3523

Kowalski N, Depireux D, Shamma S (1996b) Analysis of dynamic spectra in ferret primary auditory cortex. II. Prediction of unit responses to arbitrary dynamic spectra. J Neurophysiol 76(5):3524–3534

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM (2003) Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol 90(4):2660–2675

Machens CK, Wehr MS, Zador AM (2004) Linearity of cortical receptive fields measured with natural sounds. J Neurosci 24(5):1089–100, DOI: 10.1523/JNEUROSCI.4445-03.2004, URL http://dx.doi.org/10.1523/JNEUROSCI.4445-03.2004

Miller L et al, ME (2002) Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87:516–527

Nelken I, Fishbach A, Las L, Ulanovsky N, Farkas D (2003) Primary auditory cortex of cats: feature detection or something else? Biol Cybern 89(5):397–406, URL `http://dx.doi.org/10.1007/s00422-003-0445-3`

Nikias C, Athina P (1993) Higher-order spectra analysis. Prentice Hall Signal Pocessing Series

Phillips D, Hall S, Boehnke S (2002) Central auditory onset responses, and temporal asymmetries in auditory perception. Hear Res 167(1-2):192–205

Shepard R, Chipman S (1970) Second-order isomorphism of internal-representations: shapes of states. Cogni Psychol 1:1–17

Slaney M (1994) Auditory toolbox documentation. technical report 45. Tech. rep., Apple Computers Inc.

Terhardt E (1974) Pitch, consonance, and harmony. J Acoust Soc Am 55(5):1061–1069

Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Netw Comput Neural Syst 12:289–316

Ullman S, Vidal-Naquet M, Sali E (2002) Visual features of intermediate complexity and their use in classification. Nat Neurosci 5(7):682–687

Wiegrebe L (2001) Searching for the time constant of neural pitch extraction. J Acoust Soc Am 109(3):1082–1091

Zhang L, Bao S, Merzenich M (2001) Persistent and specific influences of early acoustic environments on primary auditory cortex. Nat Neurosci 4(11):1123–1130

Zhang LI, Bao S, Merzenich MM (2002) Disruption of primary auditory cortex by synchronous auditory inputs during a critical period. Proc Natl Acad Sci USA 99(4):2309–2314

Network: Computation in Neural Systems
June/September 2005; 16(2/3): 285–300

Multiple views of the response of an ensemble of spectro-temporal features support concurrent classification of utterance, prosody, sex and speaker identity

M. COATH[1], J. M. BRADER[2], S. FUSI[2], & S. L. DENHAM[1]

[1] *Centre for Theoretical and Computational Neuroscience, University of Plymouth, Plymouth, UK, and* [2] *Institute of Physiology, University of Bern, Bern, Switzerland*

(Received 12 December 2004; revised 11 July 2005; accepted 26 July 2005)

Abstract
Models of auditory processing, particularly of speech, face many difficulties. These difficulties include variability among speakers, variability in speech rate and robustness to moderate distortions such as time compression. In contrast to the 'invariance of percept' (across different speakers, of different sexes, using different intonation, and so on) is the observation that we *are* sensitive to the identity, sex and intonation of the speaker.

In previous work we have reported that a model based on ensembles of spectro-temporal feature detectors, derived from onset sensitive pre-processing of a limited class of stimuli, preserves significant information about the stimulus class. We have also shown that this is robust with respect to the exact choice of feature set, moderate time compression in the stimulus and speaker variation. Here we extend these results to show a) that by using a classifier based on a network of spiking neurons with spike-driven plasticity, the output of the ensemble constitutes an effective rate coding representation of complex sounds; and b) that the same set of spectro-temporal features concurrently preserve information about a range of qualitatively different classes into which the stimulus might fall. We show that it is possible for multiple views of the same pattern of responses to generate different percepts. This is consistent with suggestions that multiple parallel processes exist within the auditory 'what' pathway with attentional modulation enhancing the task-relevant classification type.

We also show that the responses of the ensemble are sparse in the sense that a small number of features respond for each stimulus type. This has implications for the ensembles' ability to generalise, and to respond differentially to a wide variety of stimulus classes.

Keywords: *Auditory transients, spectro-temporal responses, auditory cortex, models, multiple 'what' pathways*

Introduction

Complex sounds can be perceived in a number of qualitatively different ways. For example, voice communication conveys information that can be perceived independently of verbal content; this includes the speaker's identity, sex, emotional state etc., as well as semantic information such as whether the utterance is a question or a statement. Since most information

Correspondence: M. Coath, Centre for Theoretical and Computational Neuroscience, University of Plymouth, Drakes Circus, Plymouth PL4 8AA, UK. Tel: +44(0)1752 232611. Fax: +44 (0)1752 233349. E-mail: mcoath@plymouth.ac.uk

ISSN: 0954-898X print / ISSN 1361-6536 online © 2005 Taylor & Francis
DOI: 10.1080/09548980500290120

about the acoustic world entering cortex passes through primary auditory cortex (PAC), representations in PAC must be sufficiently rich to support a wide range of judgments, including identifying the source and nature of the stimulus. Higher centres in auditory cortex, with different functionality, could then subsequently abstract different properties for use in various aspects of object classification (Griffiths & Warren 2004). This idea is consistent with results showing that verbal and non-verbal analysis of stimuli are handled in parallel by different areas of cortex (Kriegstein et al. 2003). It is also consistent with the recent finding, using MEG, that there is differential task-dependent modulation of parallel processing maps within the auditory 'what' pathway in phonological and speaker identity classification tasks (Obleser et al. 2004).

Nevertheless, the way in which sounds are represented and processed in primary auditory cortex remains controversial (Griffiths & Warren 2004). A significant problem, when it comes to understanding the processing of speech, is the lack of any data regarding the nature of receptive fields in human PAC. However, data describing spectro-temporal response fields (STRFs) in cortex and midbrain of animals (Escabi & Schreiner 2002; Linden et al. 2003) is available and it would seem plausible that there are similarities across species. In previous work (Coath & Denham 2005), we have shown that ensembles of STRFs derived from speech stimuli can preserve significant information about utterance class. The STRFs were derived from fragments of an onset/offset enhanced representation of a very limited set of utterances. We then investigated the information transmitted by this representation using a speech corpus containing utterances from a wide variety of speakers. The results showed that the preservation of class information was robust with respect to the exact choice of feature set, moderate time compression in the stimulus and speaker variation. We found, as for vision (Ullman et al. 2002), that ensembles of fragments of intermediate spectral and temporal extent conveyed most class information.

Here, we extend our investigations of this putative model of processing in PAC, by considering firstly, whether the same representation can support multiple qualitatively different types of classification, and secondly, whether the representation provides a suitable basis for spike train encoding so that a network of biologically plausible spiking neurons with synaptic plasticity (Del Giudice et al. 2003) could learn to recognise and classify acoustic stimuli. It should be stressed that it is not at all clear *a priori* whether such an ensemble of STRFs should be capable of extracting and conveying information useful for speaker identification, sex or prosody classification. There is no clear understanding of how humans perform these tasks and they are all thought to involve pitch, a feature which is not explicitly represented in this model. Here we adopt a similar approach to our previous work (Coath & Denham 2005) but extend the classifications of the stimuli to encompass utterance class, sex, speaker identity and prosody; all classified on the basis of exactly the same representation.

Methods

The model

The model, whose operation is illustrated in Figure 1, consists of three principal processing stages: spectral decomposition, extraction of envelope transients and convolution using a bank of STRFs. This is followed by event detection which leads to a mapping of each event to a response space, and subsequent classification. The stages are described in detail in (Coath & Denham 2005).

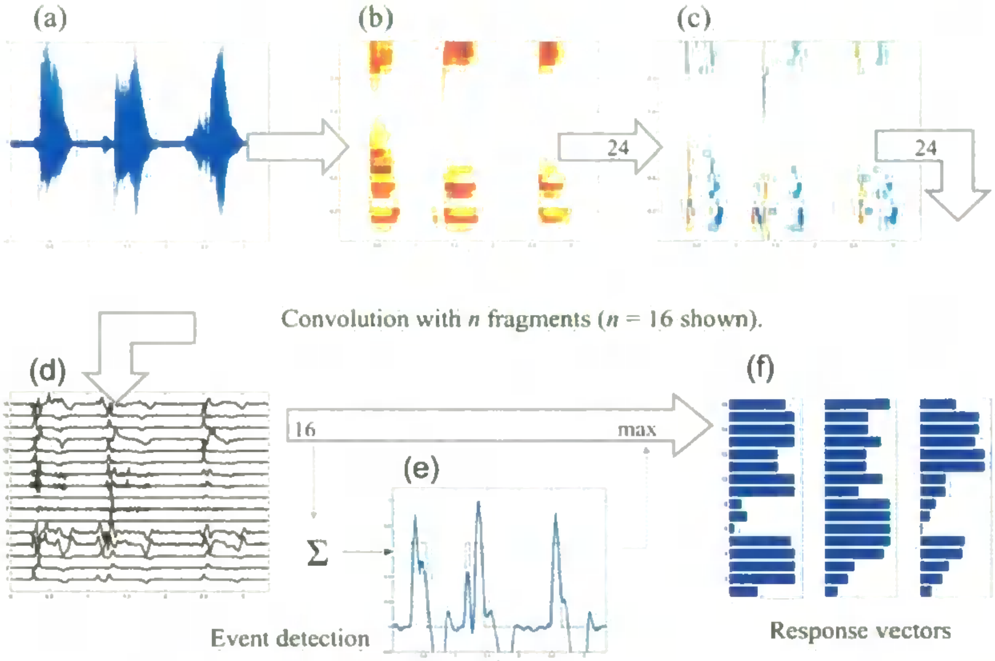


Figure 1. Stages of the process: The waveform (a) is processed by a cochlear model (b) and the within-channel envelope transients extracted (c). For each element in the ensemble of STRFs (ensemble size 16 illustrated) a time-varying response (d) is derived using a convolution of each STRF with (c). The output of the ensemble is segmented using the derived temporal saliency map (e). This results in a series of response vectors (f).

Spectral decomposition. The first stage approximates processing in the cochlea. Sounds are processed using a bank of 30 Gammatone filters (Slaney 1994), with centre frequencies, ranging from 100 to ~8000 Hz arranged evenly on an ERB scale (Glasberg & Moore 1990), see Figure 1b.

Transient extraction. The next stage of processing enhances envelope transients within each frequency channel. Responses of this type have been reported in the subcortical auditory system (Phillips et al. 2002) including the cochlear nucleus. The mean level of activity within each channel is calculated in overlapping temporal windows of duration twice the period of the centre frequency but with a minimum window size of 2.5 ms at high frequencies (Wiegrebe 2001). The overlap for all experiments was set to 10% of the window duration. The third central moment, or skewness of the distribution of energy across four successive windows is then calculated. In effect this processing amounts to edge detection in the temporal domain and the result is a spectro-temporal map of envelope transients in response to the processed sound as illustrated in Figure 1c. This approach is in some ways similar to onset/offset detection by means of a convolution with an asymmetric kernel (Smith 1996; Fishbach et al. 2001).

Convolution using an ensemble of STRFs. Each STRF in the ensemble is specified in terms of a pattern of onsets and/or offsets extending over a specified spectral range and duration. Each member of the ensemble of *n* STRFs is convolved with the pre-processed incoming signal, thereby generating a set of *n* 'temporal signatures', which indicate the degree of similarity between the incoming pattern and the STRF at each point in time. This is illustrated in Figure 1d for an ensemble of 16 STRFs. In the experimental results that follow, 128 STRFs were used.

Event detection and mapping to response space. The summed response of all STRFs in the ensemble (Figure 1d) provides an indication of the presence of an acoustic event, the timing and duration of which is determined both by the stimulus *and* by the ensemble used. Analysing the ensemble response in this way and looking for a coherent response across the whole ensemble, amounts to a bottom-up temporal saliency map providing 'interesting locations in complex scenes' (Einhusel & King 2003). This results in a method of segmentation which is not only stimulus driven but also 'detector driven', i.e., salient auditory events are marked by coherence in the response of the ensemble and not wholly by properties of the stimulus. In these experiments, we summed the output of the ensemble and recorded the maximum response of each STRF within the period during which the summed response (Figure 1e) exceeded a threshold value (20% of the maximum). The result is a vector defining a point in the *n*-dimensional space spanned by the responses of the *n* STRFs (Figure 1f). It is possible for a sound to generate more than one such event, but, in the experiments described below, when this occurred only the first event was classified.

Classifiers

Analogue classifier. In order to assign a class to each response, we trained an artificial neural network (ANN) classifier each with *n* inputs (where *n* was the ensemble size), 7 hidden units and one output unit for each class. Log-sigmoidal units were used for hidden and output nodes. For each training, the data were divided 70%, 15% and 15% into training, validation and test sets, respectively. We employed early stopping based on the validation set to avoid over-fitting. The output vector from the network formed the input to a winner-take-all stage which assigned the stimulus to an output class based on the classifier with the highest output.

Spike-driven network. The spike-driven network architecture we consider, described in more detail in Del Giudice et al. (2003) and Brader et al. (2005), consists of a single feed forward layer in which the input neurons are fully connected to the output layer by plastic synapses. Neurons in the output layer have no lateral connections and are subdivided into pools of equal size, each selective for a particular class of stimuli. In addition to the signal from the input layer, the output neurons receive signals from inhibitory and teacher populations. The inhibitory population serves to balance the excitation coming from the input layer. The teacher population is active during training and entrains the selectivity of the output pools by means of an additional excitatory or inhibitory signal. A schematic view of this network architecture is shown in Figure 2.

Learning within the network is spike driven, and takes place within the synapses using information local to each synapse. A novel bistable synaptic model (Fusi 2002), designed to ensure memory maintenance on long time scales, while retaining sensitivity on short time scales, is used. This model takes advantage of the finding that memory capacity can be maximized by making stochastic rather than deterministic synaptic transitions (Amit & Fusi 1992, 1994; Fusi 2002). If the probability of these transitions is small then only a small fraction of the stimulated synapses is changed upon each stimulus presentation. This extends the memory span of the system and prevents it from forgetting previously learned memories too

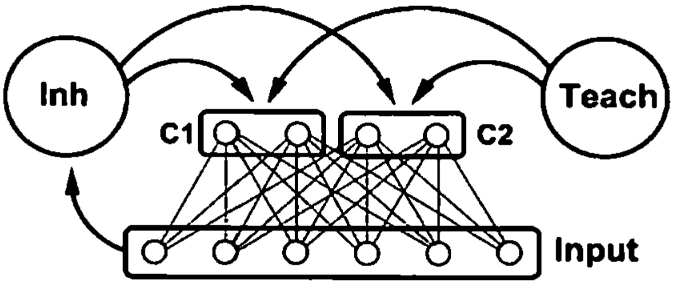


Figure 2. A schematic of the spike-driven network architecture. When considering two classes of stimuli the output units are grouped into two pools each selective to a given class. Additional signals are provided by external inhibitory and teacher populations.

quickly. Furthermore, by exploiting the inherent irregularity of the input spike trains (Fusi et al. 2000; Fusi 2003), stochastic transitions between the synaptic states are easily achieved, making the model particularly suitable for VLSI implementation (Fusi et al. 2000; Chicca & Fusi 2001; Indiveri 2002). The particular synaptic dynamics we employ are designed to be Hebbian with an additional stop-learning mechanism which makes synaptic transitions increasingly unlikely if the response of the relevant output neuron becomes either too low or too high (Fusi 2003.) (see Brader et al. 2005) for a detailed description of the dynamics). Extreme responses are an indication that the output neurons have already learned to classify the stimulus, and that it is unnecessary to modify the synapses to improve the performance (Senn & Fusi 2004). This modification enables the model to learn highly correlated input patterns.

The spike-driven classifier is implemented as follows. Each stimulus is pre-processed using 128 STRF responses, and encoded as a 128 element feature vector within which each element is a continuous value, ξ between zero and unity, thus there are 128 neurons in the input layer. When presented with a stimulus each input neuron emits a Poisson spike train at a rate 50 ξHz. The output neurons are grouped into pools, one for each class, with 10 neurons per pool. Although the output neurons will all see the same input patterns, the stochasticity of learning will create different representations for each output neuron. A similar technique has been exploited in Amit and Mascaro (2001) where the authors use random receptive fields. 70% of the dataset was used for training and the remaining 30% for testing.

In order to assess the classification performance following training, a fixed frequency threshold is defined (the same for all output neurons); an output neuron is regarded as active or inactive depending upon whether it fires at a mean rate above or below this threshold when presented with a test stimulus. The class of the stimulus is then determined by counting the number of active neurons within each pool and finding that which expresses the largest number of votes. This network architecture therefore allows for two possible types of error when presented with a test stimulus: (i) no output neurons express a vote and the stimulus is non-classified or (ii) the wrong output pool expresses the largest number of votes and the stimulus is misclassified. Non-classifications are preferable to misclassifications because the network simply expresses no preference and leaves open the possibility that such cases could be sent to subsequent networks for further analysis or that the stimulus is simply ignored.

Measuring performance

In order to measure the effectiveness of the model, we quantified the mutual information $I(S; R)$ between the classes of the stimuli S and the outputs of the classifiers, these can be thought of as the 'responses' *R*. The mutual information is calculated from Equation 1:

$$I(S; R) = \left\langle \sum_{s} P(s|r) \log_2 \left[\frac{P(s|r)}{P(s)} \right] \right\rangle_r \tag{1}$$

where $P(s|r)$ is the conditional probability of the stimulus class s given the response class r, $P(s)$ is the probability of class s, and $\langle \cdots \rangle_r$ represents the average over the (unconditional) response distribution (Golomb et al. 1997). It is important to note that we are not characterising the mutual information between the stimulus and the response, but between the *class of the stimulus* and the *class of the response*. As the maximum mutual information, $I_{\max}$ depends on the number of classes M,

$$I_{\max} = \log_2(M) \tag{2}$$

in order to compare results from experiments with differing numbers of classes the results are given as a percentage of the maximum mutual information, the normalised mutual information N_I.

$$N_I = \frac{100 \times I}{I_{\max}} \tag{3}$$

Ensemble selection

Using the method described earlier, we can derive the 'response' of any candidate feature extractor to a small set of formative classes. In order to combine these features into an ensemble of manageable size we need a measure of 'goodness' which selects the 'best' feature and allows us to add further features to the ensemble in such a way that their responses are not redundant. Essentially the aim is to select a set of features which convey as much information with respect to stimulus class as possible, whilst at the same time ensuring that their mutual information is minimised, i.e., a feature is 'good' if its response is highly correlated to the class vector but not to the responses of other features in the ensemble. The problem of feature selection, therefore, can be reduced to finding a suitable measure of correlations between features, and between features and classes.

We have adopted a feature selection procedure based on the Fast Correlation Based Filter (FCBF) (Yu & Liu 2003) which uses an information-theoretic correlation measure. The method starts with a feature which is highly correlated to the class vector (normally the most correlated feature) and removes all 'redundant peers' of this feature. The chosen feature is designated a 'predominant feature'. This is then repeated with the most highly correlated feature remaining and so on. For our experiments, we take the first *n* features selected rather than let the process come to a conclusion. This selects one predominant feature, and the n-1 features that are successively less informative of the classes, but maximally de-correlated from the previous choices. The FCBF selection was performed 10 times starting from different random positions within the top 50 rated fragments. In Experiment 1 (letter classification) (see Results) we used all ten ensembles, but in the subsequent experiments we used only the best performing ensemble from this set.

The properties of the features comprising the best performing ensemble were analyzed in order to compare them to STRFs measured experimentally. We used the same analysis procedures as described in Miller et al. (2002) in order to calculate the best frequency

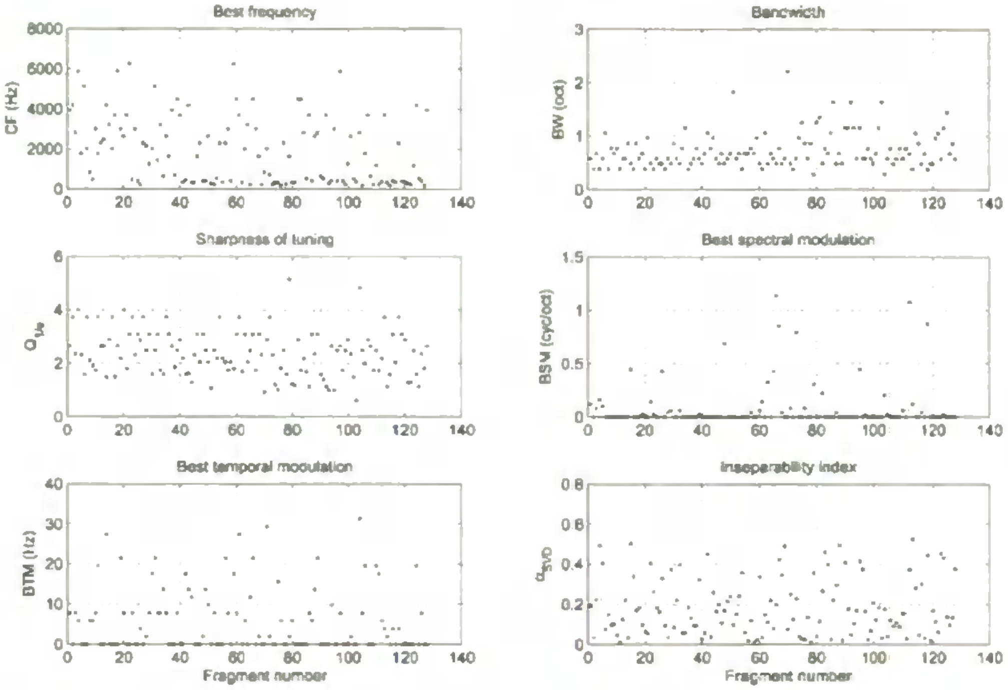


Figure 3. Distributions of the characteristics indicated across the ensemble of selected fragments.

(CF), bandwidth (BW), sharpness of tuning (Q), best spectral modulation (BSM) and best temporal modulation (BTM). We also calculated the spectrotemporal asymmetry or non-separability (α_{SVD}) index (Depireux et al. 2001). The results, illustrated in Figure 3, are broadly consistent with experimental findings in animals (Depireux et al. 2001; Miller et al. 2002). Of particular interest is the measure of separability, since, given the prominence of formant transitions in human speech, it may have been expected that STRFs with much higher α_{SVD} scores would have been selected. Clearly this is not the case, and the distribution across the ensemble is not very dissimilar from that in ferrets (Depireux et al. 2001); see Figure 4.

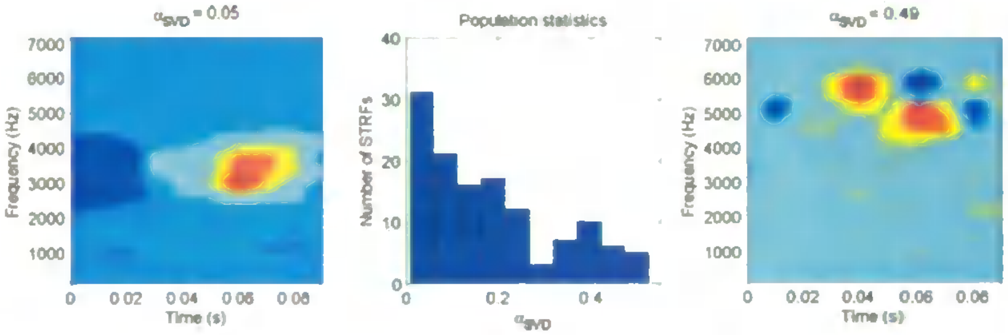


Figure 4. Distribution of α_{SVD} with examples of separable and inseparable fragments; see Figure 13 of Depireux et al. (2001) for comparison.

Stimuli

The ISOLET and male/female sets

The stimuli consist of ≈8000 spoken digits (150 speakers, male and female) contained in the ISOLET database (OGI 2002a). The same data were used in the male/female classification experiment.

The question/statement set

In British English, the primary cue which distinguishes a question from a statement is the pitch trajectory; questions have pitches which rise towards the end of the word or phrase, and statements ones which are flat or falling. The ISOLET corpus was pre-processed using PRAAT (Boersma & Weenink 1996) in order to manipulate the pitch tracks and to introduce a question or statement prosody. Firstly, a time stretching algorithm was used to ensure that all stimuli had a standard duration of 500 ms. Next, the pitch tracks were adjusted using;

$$F_0(t) = \bar{f}_0.[1 + 0.3\sin(6\pi t + \alpha)] \tag{4}$$

In Equation 4, $F_0(t)$ is the time-varying fundamental frequency or pitch trajectory of the stimulus and $\bar{f}_0$ is the mean pitch of the original utterance; for a statement, $\alpha = 4$ and for a question, $\alpha = 1$. Each stimulus was processed with both question and statement pitch tracks, giving ≈16000 stimuli. The precise form of the pitch manipulation was chosen so that we could compare the model performance with those of human subjects in a recent psychophysics study (Denham & de Thornley Head 2005). The results of these manipulations are illustrated in Figure 5.

The speaker recognition set

The stimuli for this experiment were not drawn from the ISOLET corpus but from a subset of the Speaker Recognition v1.1 corpus (OGI 2002b). This consisted of four speakers, two male and two female, answering questions such as *'What is your eye colour?'*, and *'Where do you live?'* with most answers given more than once. There are approximately 100 answers for each speaker. Longer answers were truncated at 2 seconds to save pre-processing time.

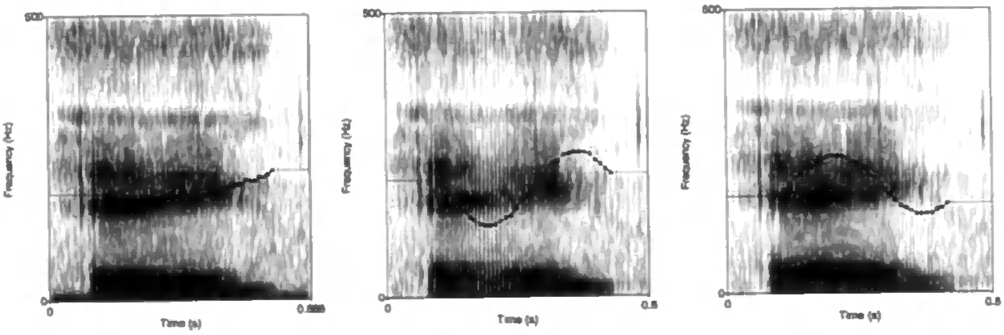


Figure 5. Question/statement processing example, showing spectrograms with pitch tracks superimposed in blue. Left: Original utterance (letter 'a', female speaker, mean pitch 190 Hz). Centre: Question form. Right: Statement form.

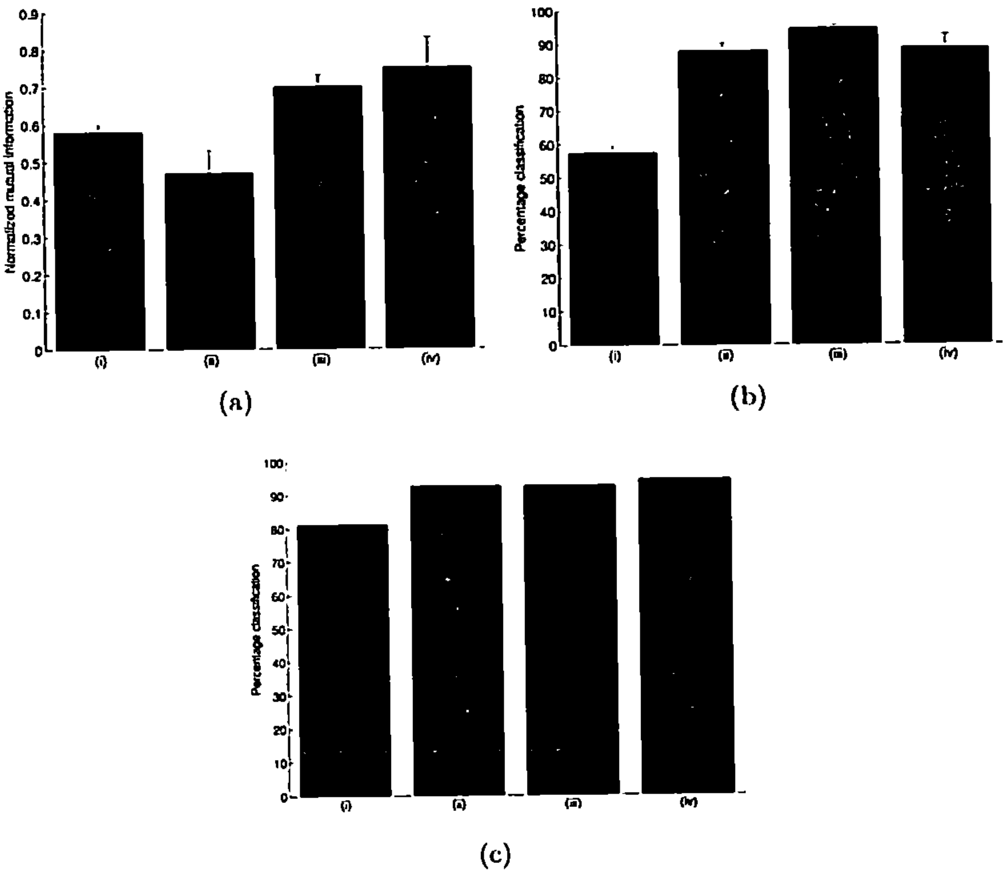


Figure 6. Results of the three ISOLET tasks. (i) ISOLET letter names. (ii) Question/statement. (iii) Male/female. (iv) Speaker identity. Error bars represent ± one S.D. (a) Results in terms of normalized mutual information. (b) Results in terms of percentage correct classification. (c) Results in terms of percentage correct classification using the spiking network.

Results

The results for each of the four experiments using the ISOLET database classified with the analogue ANN and spiking networks are shown in Figure 6. Results for the analogue ANN are shown in terms of classification percentage and the normalised mutual information as described in the 'Measuring Performance' Section. Because of the simple nature of the analogue classifier used to obtain these results, the mutual information should be seen as a lower bound; the results for the classifier built of spiking neurons show that more information is present in the output of the model. For the spike-driven network, only classification percentages were available at the time of submission. The mean classification accuracy for letter-names using the spike-driven network was over 80% which compares favourably with that reported for other machine learning algorithms (Yu & Liu 2003). Plots of the misclassifications for the two classifiers are compared in Figure 7; note that because the majority of errors in the spiking network results were non-classifications the gray scale is greatly compressed to show the misclassifications.

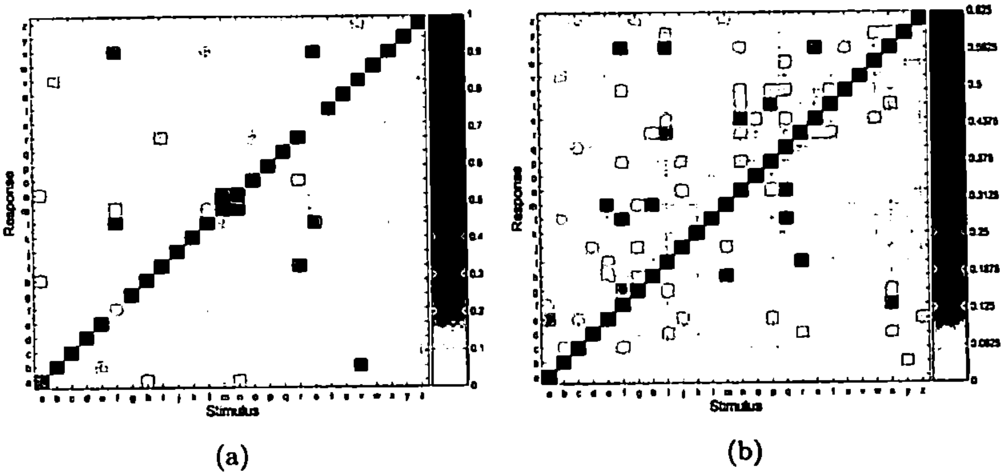


Figure 7. Results from letter name classification. NB grey scale greatly compressed in 7(b). (a) Analogue ANN, (b) Spike-driven network.

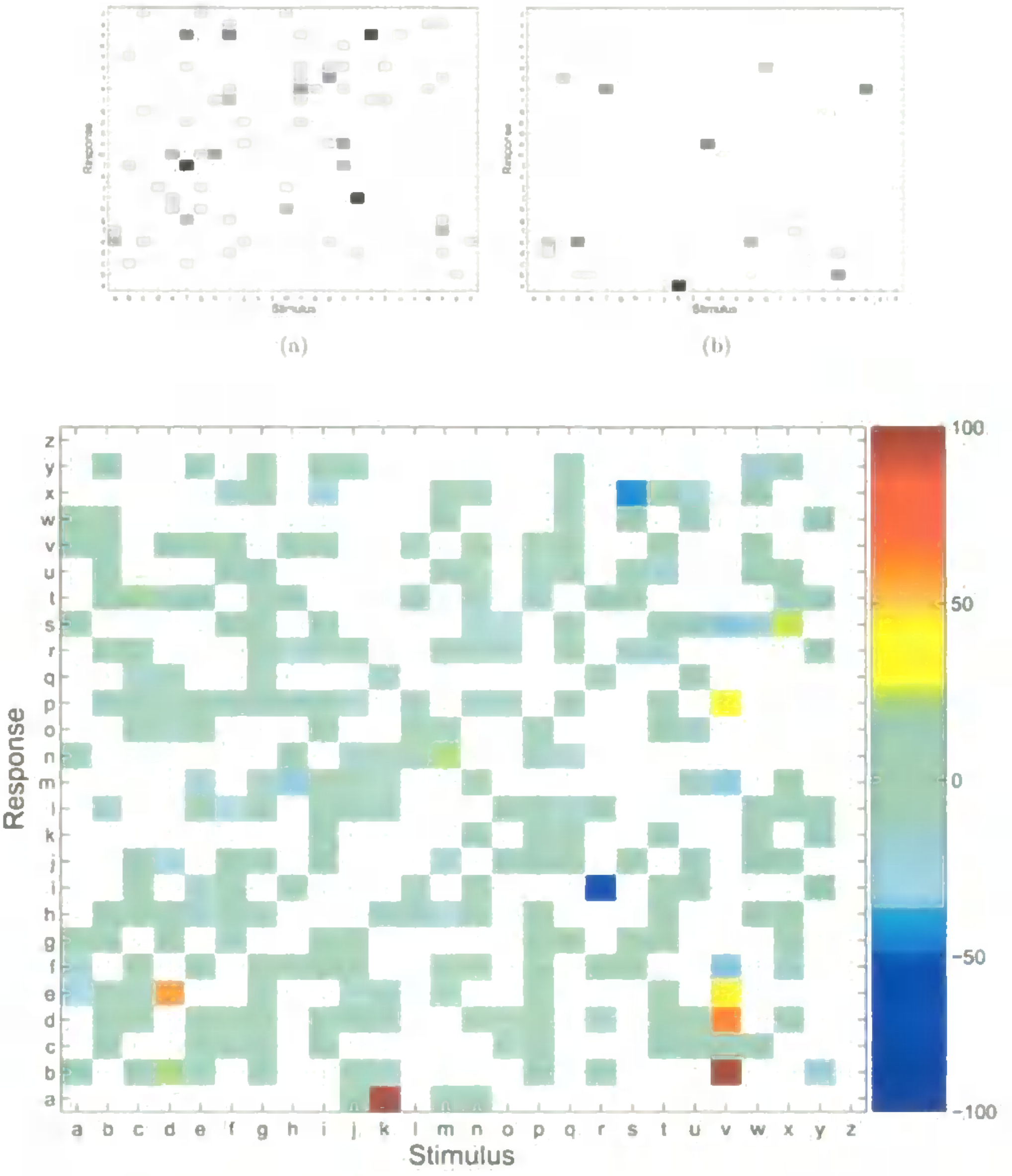
Letter name classification

Figure 8a shows the pattern of experimental misclassifications. These experimental confusions account for less than 6% of the total stimulus presentations, but among the most frequent are *f* → [*lx*], *r* → *i* and *s* → *x* which all share an initial phoneme. Some interesting features emerge from a comparison of the pattern of experimental misclassifications with the pattern of misclassifications from human psychophysics shown in Figure 8b (Hull 1973). To better compare Figure 8a and Figure 8b, Figure 8c is plotted as a percentage change of the within-class error rate between Figure 8a and Figure 8b. In Figure 8c white areas represent classes that are not confused by the model nor in human psychophysics. Green areas represent agreement between the model and the psychophysics as to how easy or difficult it is to distinguish the two letters. Red areas are those where the model has more success in differentiating the classes, and blue areas are those where humans outperform the model. The vast majority of the map is either white or green.

Red areas (those where the model results compare favourably) are found in the *d* → *e*, *k* → *a* and *v* → [*dbep*] misclassifications. These pairs are distinguished by their initial phonemes. The dark blue areas (those where model results compare unfavourably) include *r* → *i*, and *s* → *x*. These pairs share an initial phoneme. It is likely therefore that performance could be improved still further by incorporating events other than the first event in each presentation; a subject of our current investigations. Note that the ISOLET database uses *Z* = '*zee*' (US) whereas the experiments in Hull (1973) use *Z* = '*zed*' (UK) so the results for this letter name are omitted in this comparison.

PCA analysis of network weights

In order to investigate the contribution of each feature to the classification of each of the letters, we performed a principal components analysis of the neural network weights obtained in each of the training sessions. A composite loading vector was obtained for each letter in the stimulus set by combining the eigenvectors corresponding to all eigenvalues greater than 0.7. The resulting matrix, illustrated in Figure 9, shows that there is a sparse representation



(a)

(b)

(c)

Figure 8. The plot in (c) shows differences between (a) and (b). White: agreement, i.e., no significant misclassifications in the model or in psychophysics. Green: agreement, the model and the psychophysics agree as to the confusability of letter names. Red: the model finds these distinction easier than human subjects. Blue: the model misclassifies where human subjects rarely do. (a) Experimental misclassifications using the spike-driven network model. (b) Confusions (from Hull (1973)). (c) Percentage change from Figure 8(b) to Figure 8(a).

of the data set; with each feature contributing significantly to only a few classes, and each class being primarily defined by a rather small set of features. This is encouraging as it shows that the fragment selection algorithm successfully chooses features that are de-correlated, and also means that the ensemble can in principal encode a very wide range of classes.

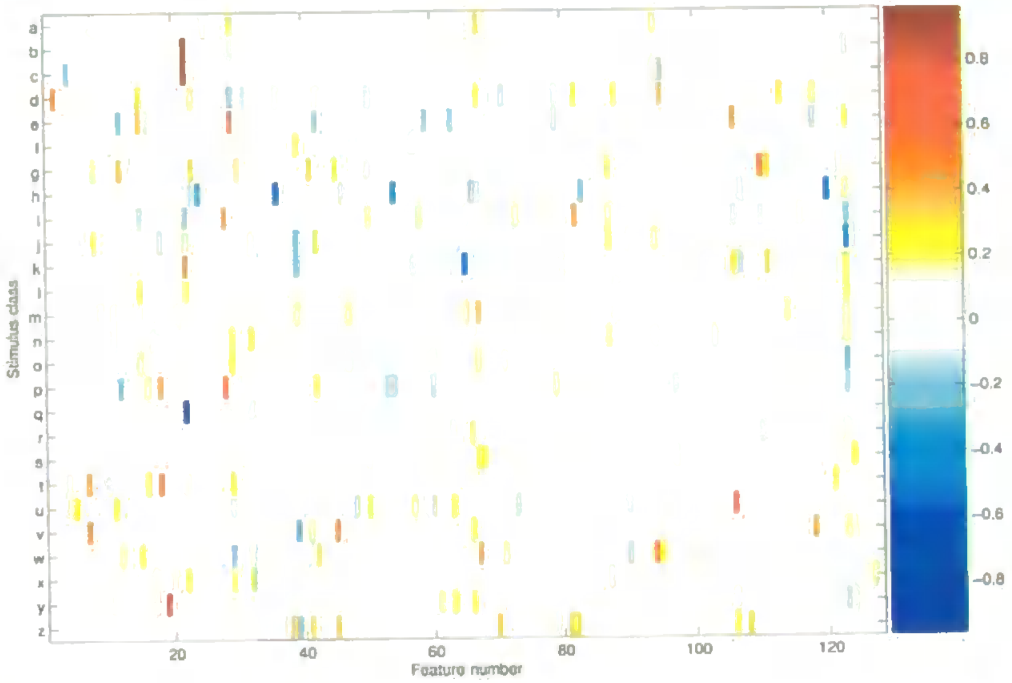


Figure 9. Sparse coding of the stimulus set; the image shows the significant contributions of features to each class derived from a PCA analysis of neural network weights.

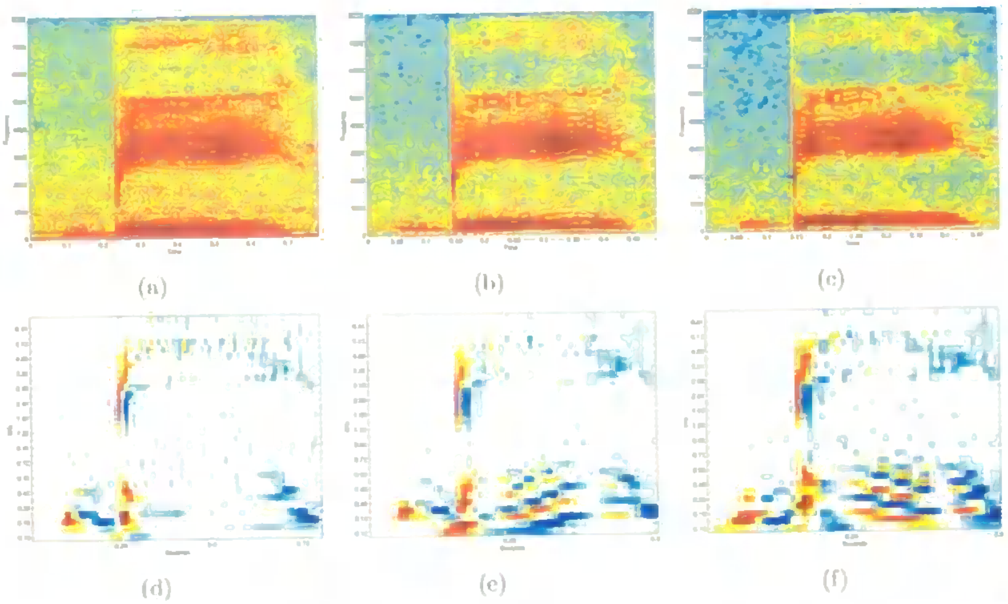


Figure 10. Top row: the letter B, normal, question, and statement. Bottom row: each processed using the on-set/offset representation.

Question/statement classification

The average correct classification achieved by the model (88%) is comparable with the average performance of human subjects (80%) (Denham & de Thornley Head 2005). This may seem rather surprising since the classes are defined by the pitch trajectories and the feature ensembles are chosen from a spectro-temporal envelope representation; pitch is not explicitly extracted in the model. However, on closer examination it seems that in the onset/offset representation, a rising or falling pitch track creates a characteristic pattern of onsets and offsets as the energy moves from one frequency channel to another (as illustrated in Figure 10) and this could allow stimuli from the two classes to be distinguished. Another important aspect to note is that the mean pitches vary widely across the stimulus set, from low male pitches, typically ≈80 Hz, to high female pitches of ≈350 Hz, which implies that the representations derived from the projections into feature response space support the abstraction of pitch trajectory shape. The ability of this model to classify the shape of pitch trajectories in complex sounds perhaps sheds some light on the somewhat contradictory data for amusics. In a recent experiment, it was found that amusics' ability to detect and classify continuous pitch changes in sounds was almost as good as that for normals, while their ability to detect differences in discontinuous pitch sequences was much worse (Foxton et al. 2004). Our model demonstrates that ensembles of STRFs similar to those measured in PAC of animals, are capable of classifying pitch trajectories which can be represented within a single event. However, recognising a pattern of discrete pitches would require the system to learn the sequence of projections of separate events within the feature responses space; a different problem involving higher order processing, perhaps the locus of impairment in amusics?

Male/female results

Classification success for the male/female discrimination task was ≈95% which is broadly consistent with data from human psychophysics (e.g., Whiteside (1998)) with a reported mean success of 98.9% in an experiment using short vowel segments. Since clear differences in vocal tract length and vocal tract morphology between males and females are known to exist (Fitch & Giedd 1999), it is perhaps not surprising that the model was able to perform this classification task. Nevertheless, the problem is not trivial as changes in vocal tract length result in quite small changes in the positions of formant peaks, and it is necessary to detect these in the presence of much larger changes in formant position characterising the different speech sounds. In a recent PCA analysis of the variability of spoken vowel sounds, it was found that 80% of the variability was accounted for by differences between vowels, and of the 20% of intra-vowel variability, 90% was explained by changes in vocal tract length; i.e., 18% of the total variability (Turner & Walters 2004). The model of VTL estimation presented in that study matched experimental data very well, but was restricted to the single vowel sound *'aa'*. Our model on the other hand is able to learn to classify speaker sex for arbitrary utterances, and as far as we are aware may be the first biologically plausible model of voice gender classification.

Speaker identification results

This was the only experiment that did not use the ISOLET corpus. The model was able to correctly identify each of the four speakers with an accuracy of ≈89% using short segments of randomly chosen utterances. For comparison, in a recent study (Obleser et al. 2004) subjects were able to identify two speakers with an accuracy of ≈95%. As the number of speakers in our experiment was small, our result is only suggestive, but it was achieved in a text independent experiment using the same feature extractors as the other experiments reported here. This establishes, at least in principle, that information about speaker identity can be preserved in the pattern of responses of such an ensemble, and that responses of the same ensemble can be used in parallel for a number of different perceptual classifications; as found in the human MEG study for phonological and speaker classifications in (Obleser et al. 2004).

Discussion

Given that it is widely reported that responses in PAC can, at least to a first approximation, be characterized by their spectro-temporal characteristics, it is not unreasonable to ask whether an ensemble of spectro-temporal feature extractors might provide a representation sufficiently rich to be biologically useful. Our model attempts to incorporate some of the physiological evidence for processing in the ascending auditory pathway and feature extractors in PAC. The approach adopted is complimentary to work that seeks to model the integrated activity of neural populations. One recent study by Husain et al. (2004) for example has shown that large scale, neurobiologically plausible modelling of auditory processing provides results consistent with studies of cerebral activity measured using optical and MRI techniques, during tasks involving simple, synthetic stimuli.

In contrast, we have shown using biologically plausible pre-processing, a modestly sized ensemble, and a spike-rate encoding, that salient features of ethological stimuli can be simply extracted and used as the basis for behaviourally important judgements. Moreover the same ensemble response can support many qualitatively different judgements concurrently. We assume that there is competition between these perceptual judgements which is subject to a top-down task-dependent attentional bias. The aspect that is attended to is the one most likely to be task-relevant. This is consistent with evidence that 'what' processing in auditory cortex can be viewed as a set of parallel processes in which concurrent phonological classifications are made in spatially separated areas (Obleser et al. 2004) and implicit semantic processing continues when attention is directed to non-verbal input analysis (Kriegstein et al. 2003).

The basis of feature extraction in the current model is the presence of a coherent response across the ensemble of feature detectors such as those found in PAC. This is equivalent to a saliency map in the temporal domain (Koch & Ullman 1985), where the signal is analysed locally with respect to a range of properties (the ensemble response) and the results integrated (summed). This provides the basis for an asynchronous, stimulus-ensemble driven event detector. This triggers a readout of the population response pattern within a time window, the length of which is determined by the duration of the coherent ensemble response. The result is a short time scale context for the extraction of a pattern of responses that characterizes a distinct auditory event. These events are likely to be represented by population responses which, because of the time window and the asynchronous read out, are not likely to bear a simple relationship to the temporal structure of the stimulus. It has been suggested that this type of post-primary cortical processing might be found in the planum temporale (Griffiths & Warren 2002) where responses that are not closely coupled to the time course of the stimulus do occur (Steinschneider et al. 1999).

The range of classifications supported by the model includes those distinguished primarily by spectral profile (male/female), solely by pitch trajectory (question/statement), as well as those characterised by more complex spectro-temporal relationships (letter-names, speaker identity). The question/statement result in particular demonstrates that a representation of pitch change can be abstracted from the output of the system in which there is no explicit sense of pitch *per se.* Furthermore the performance of the model in each of the tasks shows some similarities with human psychophysics. It has been reported that perceptual categories such as these are processed in distinct areas of auditory cortex anterior to PAC (consistent with the 'what' pathway) and also distinct from regions involved in decisions that are correlated with reaction times (Binder et al. 2004).

One of the strengths of the spiking neural network is its ability to provide non-classifications. This implies that the characterisation of the stimulus by the model using a single event is unclear. Such stimuli account for ≈14% of the test set in the current results; most frequently in classes [flmns] i.e., classes that are not resolved by their initial phonemes. Work is already underway to use subsequent events, when they occur, to reinforce the classification judgement raising the probability above the threshold for an unambiguous assignment of class.

We have chosen to use spoken letter names for three of the current experiments and a wider range of spoken stimuli for the fourth. This was due to the ready availability of large and well characterised corpora. But it must be emphasized that the principle goal of our research is not speech recognition or speaker identification, although both may be informed by this approach, rather we aim to understand the representation and processing of complex sounds in general within the auditory system.

Acknowledgements

This work was supported by EU Open FET IST-2001-38099.

References

Amit D, Fusi S. 1992. Constraints on learning in dynamics synapses. Network 3:443–464.

Amit D, Fosi S. 1994. Learning in neural networks with material synapses. Neural Computation 6:957-982.

Amit Y, Mascaro M. 2001. Attractor networks for shape recognition. Neural Computation 13(6):1415–1442.

Binder JR, Liebenthal E, Possing ET, Medler DA, Ward BD. Mar 2004. Neural correlates of sensory and decision processes in auditory object identification. Nature Neuroscience 7(3):295–301.

Boersma P, Weenink D. 1996. Report 132. Institute of Phonetic Sciences, University of Amsterdam.

Brader J, Senn W, Fusi S. Learning real world stimuli in a neural network with spike driven synaptic dynamics. Neural Computation - Accepted.

Chicca E, Fusi S. 2001. Stochastic synaptic plasticity in deterministic avlsi networks of spiking neurons. In: F. Rattay, editor. *Proc. of the World Congress on Neuroinformatics*, ASIM Verlag, Vienna, pp. 468–477.

Coath M, Denham SL. 2005. Robust sound classification through the representation of similarity using response fields derived from stimuli during early experience. Biol Cybernetics 3.

Delgiudice P, Fusi S, Mattia M. 2003. Modeling the formation of working memory with networks of integrate-and-fire neurons connected by plastic synapses. J Phys Paris 97:659–681.

Depireux DA, Simon JZ, Klein DJ, Shamma SA. Mar 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. J Neurophysiol 85(3):1220–1234.

Einhauser W, Konig P. Mar 2003. Does luminance-contrast contribute to a saliency map for overt visual attention? Eur J Neurosci 17(5):1089–1097.

Escabi MA, Schreiner CE. May 2002. Nonlinear spectrotemporal sound analysis by neurons in the auditory midbrain. J Neuroscience 22(10):4114–4131.

Fishbach A, Nelken I, Yeshurun Y. June 2001. Auditory edge detection: a neural model for physiological and psychoacoustical responses to amplitude transients. J Neurophysiol 85(6):2303–2323.

Foxton JM, Dean JL, Gee R, Peretz I, Griffiths TD. Apr 2004. Characterization of deficits in pitch perception underlying 'tone deafness'. Brain 127, Pt 4:801–810.

Fusi S. 2002. Hebbian spike-driven synaptic plasticity for learning patterns of mean firing rates. Biological Cybernetics 87:459–470.

Fusi S. 2003. Spike-driven synaptic plasticity for learning correlated patterns of mean firing rates. Reviews in the Neurosciences 14:73–84.

Fusi S, Annunziato M, Badoni D, Salamon A, Amit D. 2000. Spike-driven synaptic plasticity: theory, simulation, vlsi implementation. Neural Computation 12:2227–2258.

Glasberg BR, Moore BC. 1990. Derivation of auditory filter shapes from notched noise data. Hear Res 47(1):103-138.

Golomb D, Hertz J, Panzeri S, Treves A, Richmond B. Apr 1997. How well can we estimate the information carried in neuronal responses from limited samples? Neural Comput 9(3):649-665.

Griffiths TD, Warren JD. July 2002. The planum temporale as a computational hub. Trends Neurosci 25(7):348-353.

Griffiths TD, Warren JD, Scott SK, Nelken I, King AJ. Apr 2004. Cortical processing of complex sound: a way forward? Trends Neurosci 27(4):181-185.

Head P, Denham SL. 2004. Perceptual interference between fine structure and spectrotemporal envelope in complex sounds. Perception and Psychophysics - under review.

Heil P. 2001. Representation of sound onsets in the auditory system. Audiol Neurootol 6(4):167-172.

Hull A. Nov 1973. A letter-digit matrix of auditory confusions. Br J Psychol 64(4):579-585.

Husain F, Tagamets M-A, Fromm S, Braun A, Horwitz B. Apr 2004. Relating neuronal dynamics for auditory object processing to neuroimaging activity: a computational modeling and an fMRI study. Neuroimage 21(4):1701-1720.

Indiveri G. 2002. *Advances in Neural Information Processing Systems*. Vol. 15. MIT Press, Cambridge, MA, ch. Neuromorphic bistable VLSI synapses with spike-timing-dependent plasticity.

Koch C, Ullman S. 1985. Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4(4):219-227.

Kriegstein K, Eger E, Kleinschmidt A, Giraud AL. June 2003. Modulation of neural responses to speech by directing attention to voices or verbal content. Brain Res Cogn Brain Res 17(1):48-55.

Linden JF, Liu RC, Sahani M, Schreiner CE, Merzenich MM. Oct 2003. Spectrotemporal structure of receptive fields in areas AI and AAF of mouse auditory cortex. J Neurophysiol 90(4):2660-2675.

Miller LM, Escab MA, Read HL, Schreiner CE. Jan 2002. Spectrotemporal receptive fields in the lemniscal auditory thalamus and cortex. J Neurophysiol 87(1):516-527.

Nelken I. Aug 2004. Processing of complex stimuli and natural scenes in the auditory cortex. Curr Opin Neurobiol 14(4):474-480.

Obleser J, Elbert T, Eulitz C. Jul 2004. Attentional influences on functional mapping of speech sounds in human auditory cortex. BMC Neurosci 5(1):24.

OGI. 1996. Oregon health and science university: The speaker recognition corpus v1.1.

OGI. 1999. Oregon health and science university: The isolet corpus v1.3.

Phillips D, Hall S, Boehnke S. May 2002. Central auditory onset responses, and temporal asymmetries in auditory perception. Hear Res 167(1-2):192-205.

Senn W, Fusi S. 2004. Slow stochastic learning with global inhibition: a biological solution to the binary perceptron problem. Neurocomputing 58-60:321-326.

Slaney M. 1994. *Auditory toolbox documentation, technical report 45*. Tech. rep, Apple Computers Inc.

Smith LS. 1996. Onset-based sound segmentation. In: Touretzky DS, Mozer MC, Hasselmo ME, editors. *Advances in Neural Information Processing Systems*, vol. 8. The MIT Press, pp. 729-735.

Steinschneider M, Volkov IO, Noh MD, Garell PC, Howard MA. Nov 1999. Temporal encoding of the voice onset time phonetic parameter by field potentials recorded directly from human auditory cortex. J Neurophysiol 82(5):2346-2357.

Turner RE, Walters TC. 2004. *BSA Short papers meeting*.

Ullman S, Vidal-Naquet M, Sali, E. July 2002. Visual features of intermediate complexity and their use in classification. Nat Neurosci 5(7):682-687.

Whiteside S. Apr 1998. Identification of a speaker's sex: a study of vowels. Percept Mot Skills 86(2):579-584.

Wiegrebe L. Mar 2001. Searching for the time constant of neural pitch extraction. J Acoust Soc Am 109(3):1082-1091.

Yu L, Liu H. 2003. Feature election for high-dimensional data: a fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington.*