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).

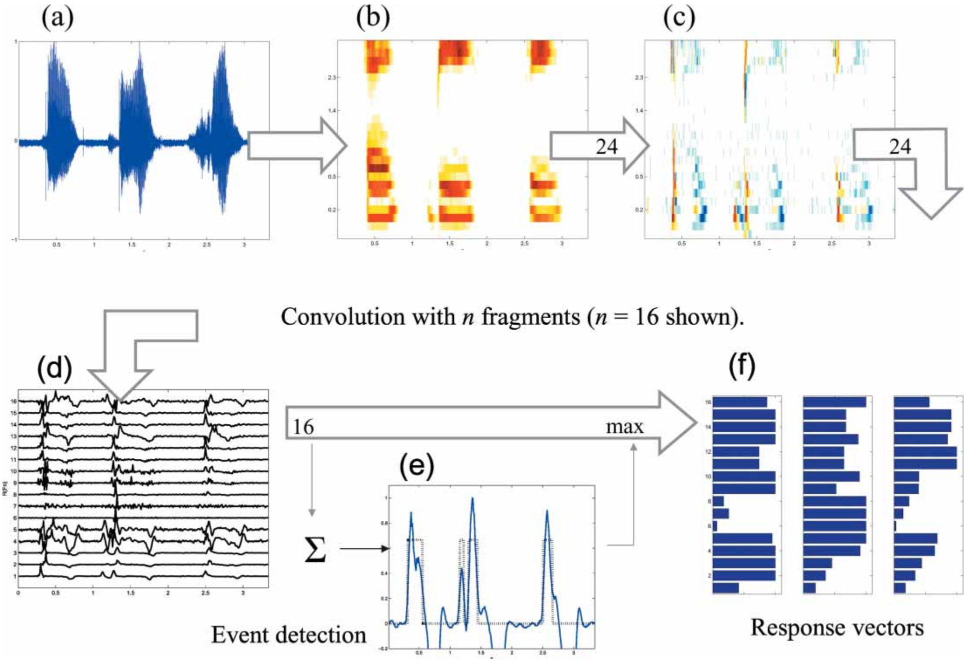


Figure 1. Stages of the process: The waveform (a) is processed by a cochlear model (b) and the within-channel envelope transients extracted (c). For each element in the ensemble of STRFs (ensemble size 16 illustrated) a time-varying response (d) is derived using a convolution of each STRF with (c). The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

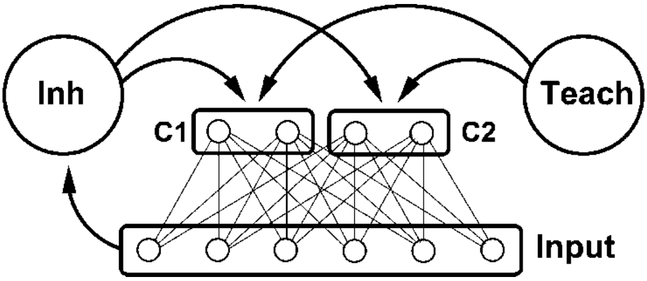


Figure 2. A schematic of the spike-driven network architecture. When considering two classes of stimuli the output units are grouped into two pools each selective to a given class. Additional signals are provided by external inhibitory and teacher populations.

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:

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

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;

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

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.

$$N_I = \frac{100 \times I}{I_{\max}} \qquad (3)$$

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

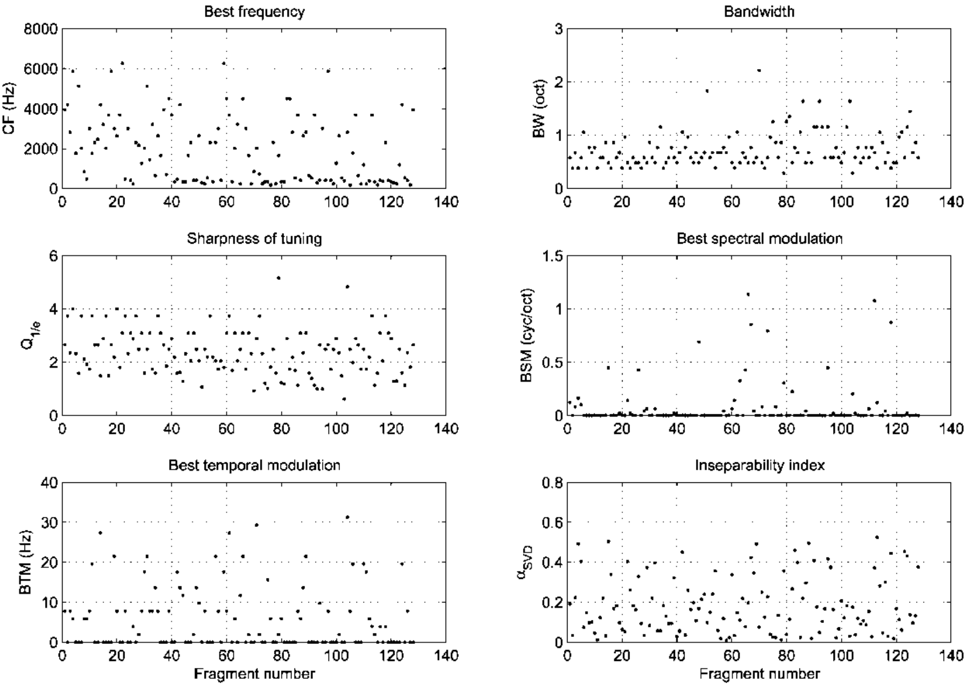


Figure 3. Distributions of the characteristics indicated across the ensemble of selected fragments.

(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.

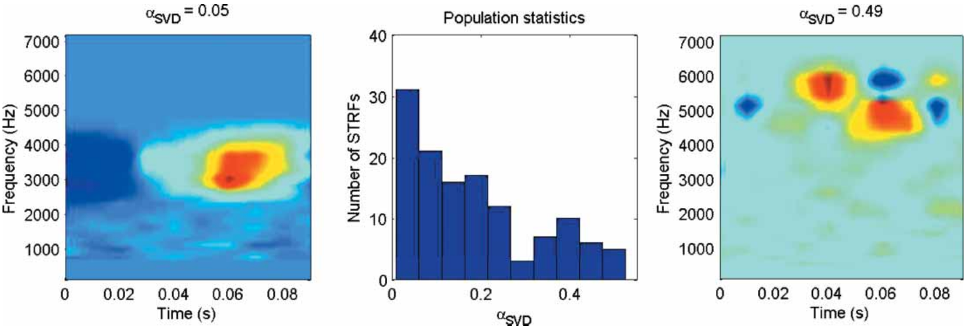


Figure 4. Distribution of α_{SVD} with examples of separable and inseparable fragments; see Figure 13 of Depireux et al. (2001) for comparison.

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.

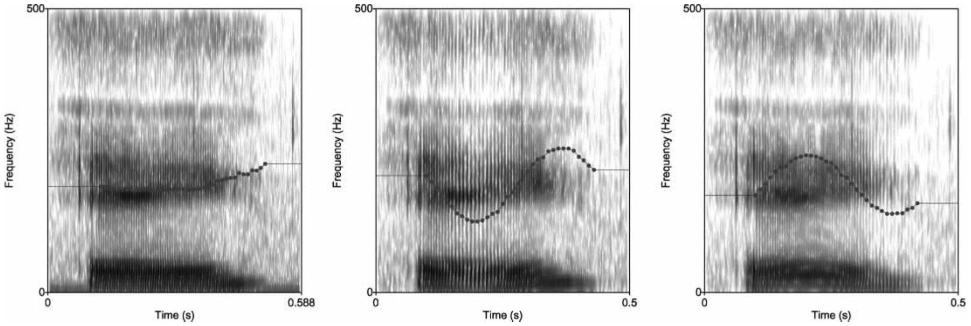


Figure 5. Question/statement processing example, showing spectrograms with pitch tracks superimposed in blue. Left: Original utterance (letter ‘a’, female speaker, mean pitch 190 Hz). Centre: Question form. Right: Statement form.

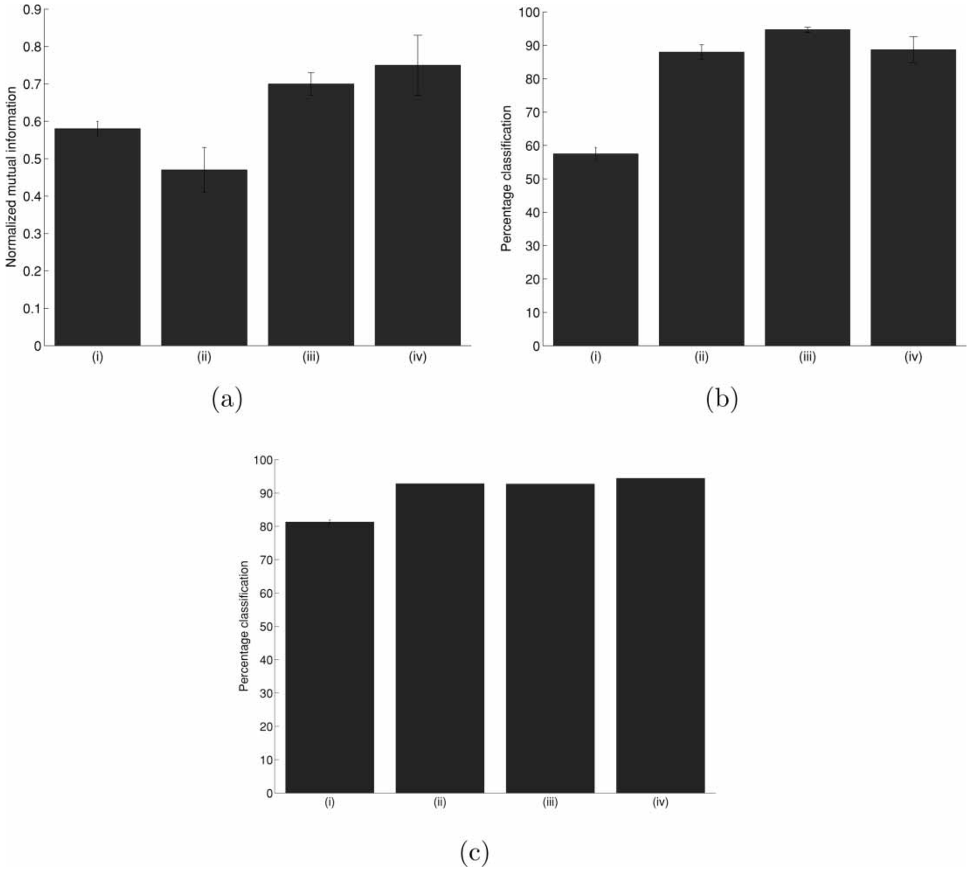
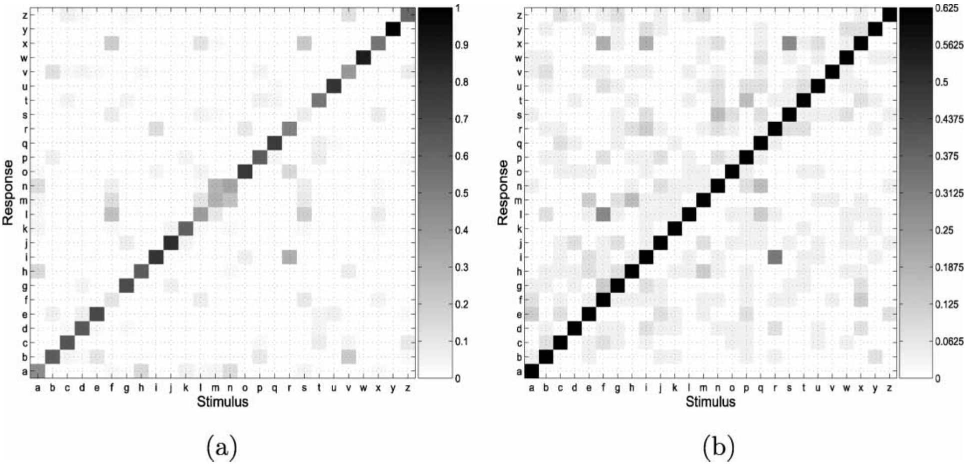


Figure 6. Results of the three ISOLET tasks. (i) ISOLET letter names, (ii) Question/statement, (iii) Male/female, (iv) Speaker identity. Error bars represent ± one S.D. (a) Results in terms of normalized mutual information. (b) Results in terms of percentage correct classification. (c) Results in terms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



(a)

(b)

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

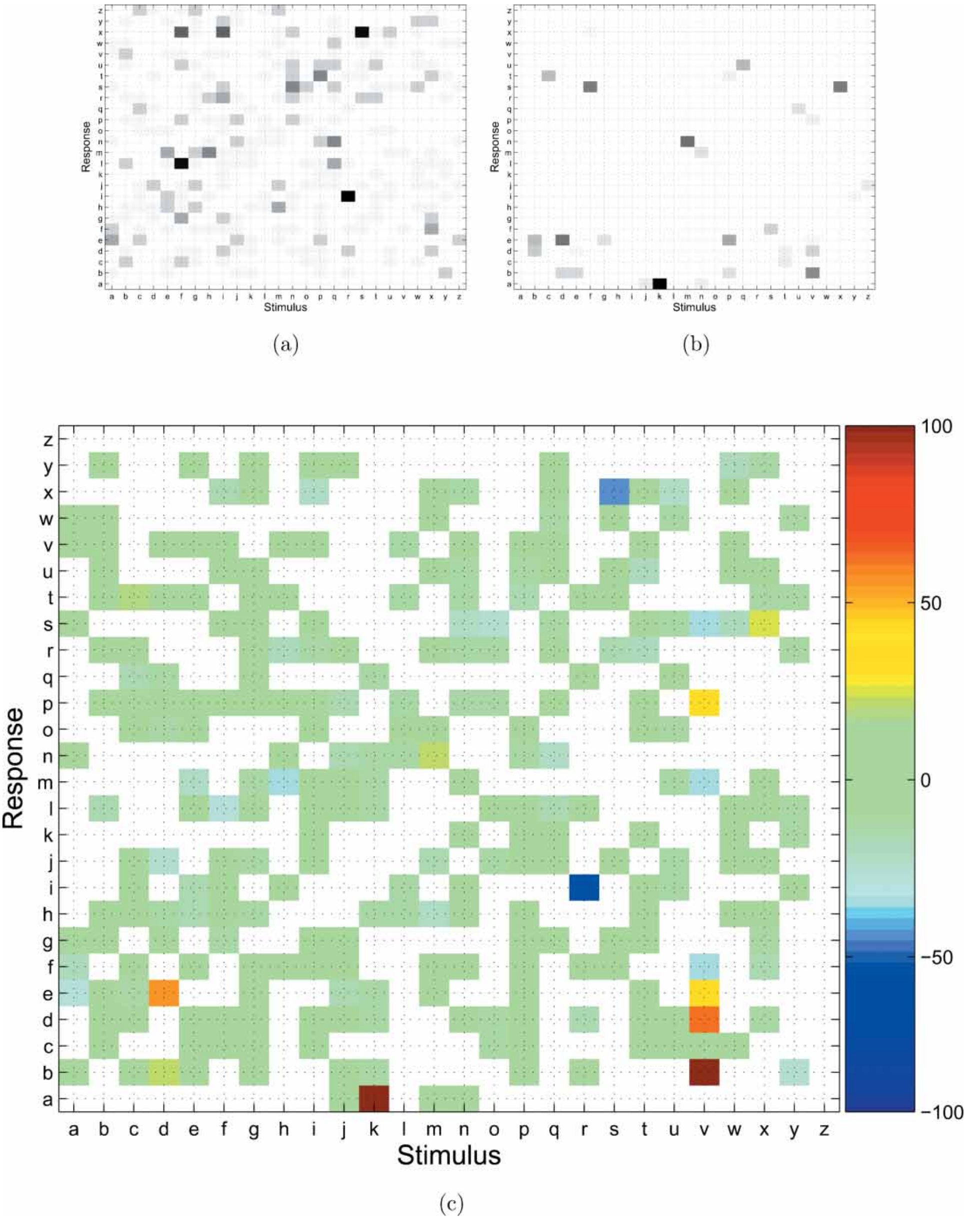
Letter name classification

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 $f \rightarrow [lx]$, $r \rightarrow i$ and $s \rightarrow x$ 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 $d \rightarrow e$, $k \rightarrow a$ and $v \rightarrow [dbep]$ misclassifications. These pairs are distinguished by their initial phonemes. The dark blue areas (those where model results compare unfavourably) include $r \rightarrow i$, and $s \rightarrow x$. 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.

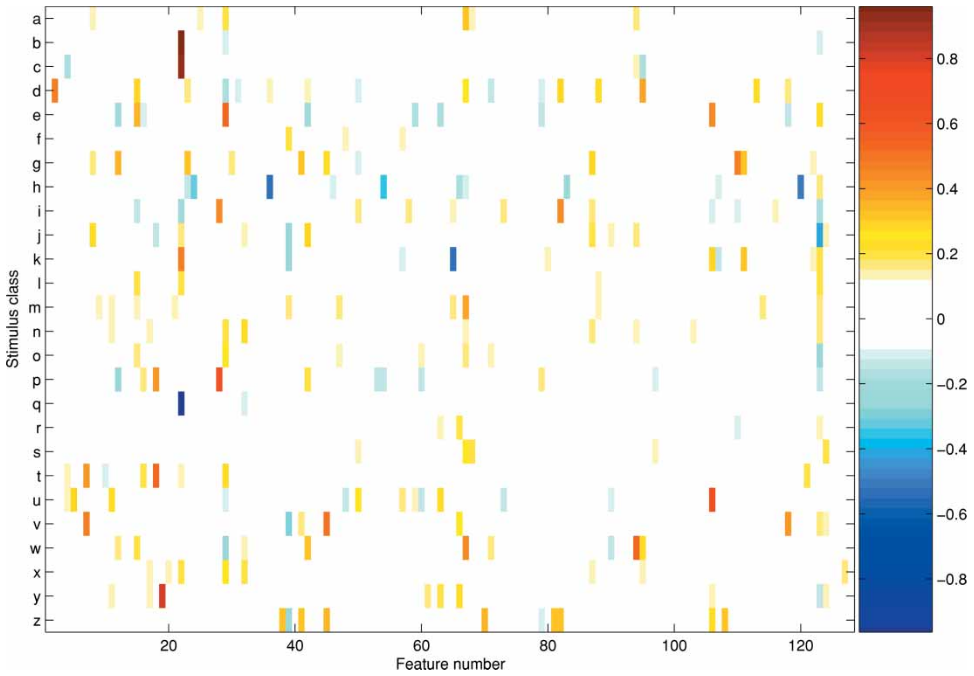
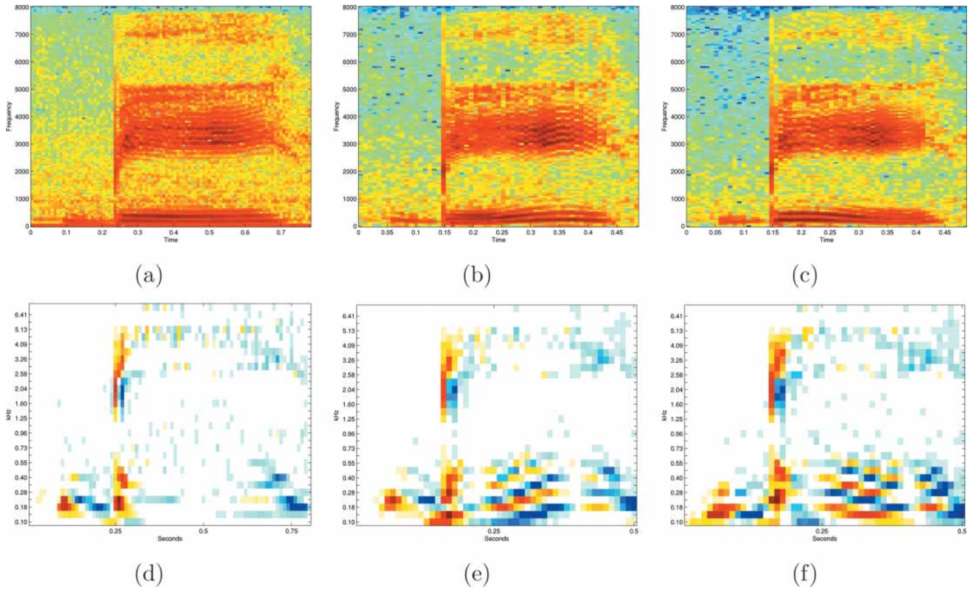


Figure 9. Sparse coding of the stimulus set; the image shows the significant contributions of features to each class derived from a PCA analysis of neural network weights.



(a) (b) (c)

(d) (e) (f)

Figure 10. Top row: the letter B, normal, question, and statement. Bottom row: each processed using the onset/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*.