Pergamon

Neural Networks, Vol. 10, No. 9, pp. 1547–1558, 1997
© 1997 Elsevier Science Ltd. All rights reserved
Printed in Great Britain
0893–6080/97 $17.00+.00

PII: S0893-6080(97)00046-4

CONTRIBUTED ARTICLE

Modelling the Perceptual Segregation of Double Vowels with a Network of Neural Oscillators

GUY J. BROWN[1] AND DELIANG WANG[2]

[1]University of Sheffield and [2]The Ohio State University

(*Received* 8 *October* 1996; *accepted* 11 *April* 1997)

Abstract—*The ability of listeners to identify two simultaneously presented vowels can be improved by introducing a difference in fundamental frequency (F0) between the vowels. We propose an explanation for this phenomenon in the form of a computational model of concurrent sound segregation, which is motivated by neurophysiological evidence of oscillatory firing activity in the auditory cortex and thalamus. More specifically, the model represents the perceptual grouping of auditory frequency channels as synchronised (phase-locked zero phase lag) oscillations in a neural network. Computer simulations on a vowel set used in psychophysical studies confirm that the model qualitatively matches the performance of human listeners; vowel identification performance increases with increasing difference in F0. Additionally, the model is able to replicate other findings relating to the perception of harmonic complexes in which one component is mistuned. © 1997 Elsevier Science Ltd. All rights reserved.*

Keywords—Auditory model, Auditory scene analysis, Neural network, Neural oscillator, Perceptual grouping, Vowel perception, Correlogram.

1. INTRODUCTION

In his influential book, Bregman (1990) likens the perceptual organisation of sound to an *auditory scene analysis*, which takes place in two conceptual stages. In the first stage, the acoustic mixture reaching the ears is broken down into a collection of ''sensory elements''. Secondly, elements that are likely to have arisen from the same environmental source are grouped to form a perceptual structure (stream) which can be interpreted by higher-level processes.

Although auditory scene analysis explains why certain acoustic features are combined to form perceptual wholes, it does not address the issue of how such groups of features are represented and formed in the brain. A similar question has been debated in the context of other modalities (particularly the visual system) and is known as the *binding problem.* A solution to the binding problem has been proposed by von der Malsburg, 1981; von der Malsburg & Schneider, 1986; see also Milner, 1974). He suggests that the responses of feature detecting cells might be bound together by the synchrony of their firing activity. In this scheme, neuronal groups with temporally synchronised activity represent a single perceptual object, whereas neuronal groups with desynchronised activity represent different objects. Von der Malsburg's theory has gained support from physiological studies, which report synchronised oscillations in the olfactory and visual systems (Eckhorn et al., 1988; Gray et al., 1989; Freeman, 1991; Singer & Gray, 1995; Gray & McCormick, 1996).

In addition, there is growing evidence that acoustic stimuli evoke synchronised oscillations in the auditory system. Galambos et al. (1981) have demonstrated that 40 Hz oscillations occur in auditory potentials evoked by a tone (see also Madler & Pöppel, 1987; Mäkelä & Hari, 1987). Using magnetic field tomography, Llinás and his colleagues have demonstrated that acoustic stimuli elicit synchronised oscillations in the auditory cortex and thalamus (Ribary et al., 1991; Llinás & Ribary, 1993; see also Barth & MacDonald, 1996). Indeed, direct evidence for the role of neural oscillations in auditory

Acknowledgements: This work was mainly undertaken while DLW was a visiting research fellow of the Department of Computer Science, University of Sheffield. GJB was supported by EPSRC grants GR/H53174 and GR/K18962. DLW was supported in part by NSF grant IRI-9423312, ONR grant N00014-93-1-0335, and ONR Young Investigator Award N00014-96-1-0676. Our thanks to Martin Cooke for useful discussions, and to Ray Meddis and Lowel O'Mard for making available the set of synthetic vowels used in the Assmann and Summerfield (1990) and Meddis and Hewitt (1992) studies.

Requests for reprints should be sent to Dr G. J. Brown, Department of Computer Science, University of Sheffield, Regent Court, 211 Portobello Street, Sheffield, S1 4DP, UK; Tel: 0114 222 1821; Fax: 0114 278 0972; e-mail: g.brown@dcs.shef.ac.uk.

grouping has recently come from a study by Joliot et al. (1994). They presented awake human subjects with a stimulus consisting of two clicks, separated by a time interval. For clicks presented less than 12–15 ms apart, listeners reported a single source and simultaneous magnetoencephalography (MEG) recordings from the auditory cortex showed a single 40 Hz response. The perceived timbre in this condition was different to that of an isolated click, indicating that the two stimuli had been fused into a single percept. For interstimulus intervals greater than 12–15 ms, listeners reported two clicks and MEG recordings showed a second 40 Hz wave. Hence it appears that 40 Hz oscillatory activity in the auditory cortex is associated with temporal grouping of auditory stimuli.

These findings have prompted a number of computational modelling studies which address the role of neural oscillations in auditory organisation (Wang, 1994, 1996, 1997; Baird, 1996; Brown & Cooke, 1997). Here we present a neural oscillator model with an emphasis on one particular aspect of auditory perceptual grouping; the segregation of concurrent vowel sounds ("double vowels") with different fundamental frequencies (F0s). The following section of this paper gives a detailed description of the neural model. In Section 3, simulations are described which replicate findings from psychophysical studies of double vowel perception, and other studies relating to the perception of harmonic complexes in which one component is mistuned. We conclude with a discussion of the limitations of the model, and its relationship with other computational studies.

2. STRUCTURE OF THE MODEL

The model consists of four main stages: peripheral auditory processing, periodicity detection, neural oscillator network and vowel recognition (Figure 1). Digitally recorded signals, sampled at a frequency of 10 kHz with 16 bit resolution, provide the input to the model.

2.1. Peripheral Auditory Processing

The frequency selective properties of the basilar membrane are modelled by a bank of 64 gammatone filters (Patterson et al., 1988), each of which simulates the frequency response of a point along the cochlear partition. The impulse response of a gammatone filter of order n and centre frequency f_0 Hz is given by

$$g(t) = t^{n-1}\exp(-2\pi bt)\cos(2\pi f_0 t + \phi) \tag{1}$$

where ϕ is phase and b is related to bandwidth. Here, we use a digital implementation of the fourth order gammatone filter proposed by Cooke (1993). The filters are distributed in frequency according to their bandwidths, which increase quasi-logarithmically with increasing centre frequency. Specifically, the centre frequencies of the filters are equally spaced on the equivalent rectangular bandwidth (ERB) scale of Glasberg and Moore (1990) between centre frequencies of 80 Hz and 4000 Hz. Outer and middle ear resonances are modelled by adjusting the gains of the gammatone filters according to the BS 3383 standard for equal loudness contours (BSI, 1988).

Subsequently, the output of each filter channel is processed by the Meddis model of hair cell transduction (Meddis, 1986, 1988). The Meddis model converts the gammatone filter output into a probabilistic representation of firing activity in the auditory nerve, incorporating well-known effects such as saturation and adaptation.

2.2. Periodicity Analysis

A key finding in psychophysical studies of double vowel perception is that identification performance improves when there is a difference in F0 between the two vowels

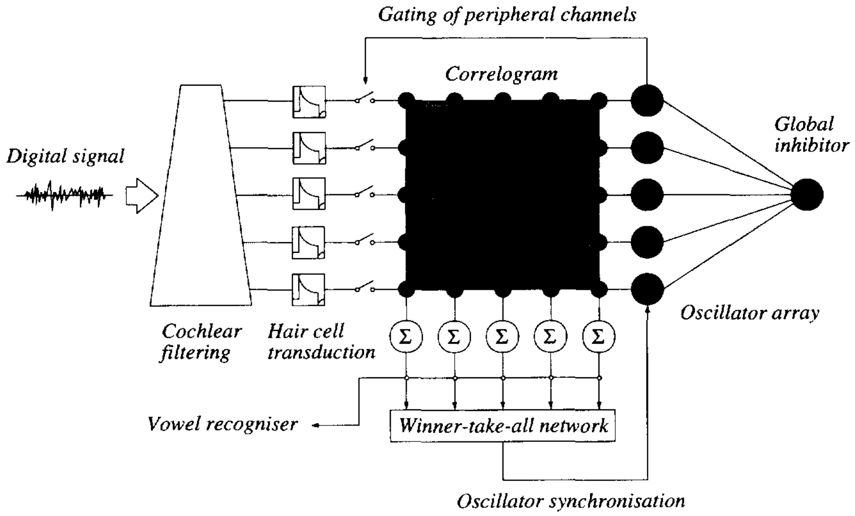

FIGURE 1. Schematic diagram of the auditory model. The synchronisation of neural oscillators is determined by the periodicity information in the correlogram; in turn, oscillator activity gates the input to the correlogram from each peripheral channel.

(Scheffers, 1983; Assmann & Summerfield, 1990; Chalikia & Bregman, 1989). This is usually interpreted as evidence that the auditory system groups together spectral regions that exhibit the same periodicity (Bregman, 1990). Accordingly, the second stage of our model involves the extraction of periodicity information from simulated auditory nerve firing patterns.

Periodicity analysis is achieved in the model by computing a *correlogram* (Slaney & Lyon, 1990; Meddis & Hewitt, 1992; Brown & Cooke, 1994). The correlogram is one of a class of pitch models in which periodicity information is combined across resolved and unresolved harmonic regions (see also Moore, 1997). Models of this type are able to account for many classical pitch phenomena, and have previously been applied to the segregation of concurrent periodic sounds with some success (Summerfield et al., 1990; Brown & Cooke, 1992; Meddis & Hewitt, 1992).

A correlogram is formed by computing a running autocorrelation of the simulated auditory nerve firing patterns at each centre frequency. For an auditory filter channel with centre frequency f, the running autocorrelation $a(t,f,\tau)$ at time t and lag τ is given by

$$a(t,f,\tau)=\sum_{i=0}^{\infty} p(t-T,f)p(t-T-\tau,f)\mathrm{e}^{-T/\Omega} \qquad (2)$$

where the time constant Ω is 25 ms, $p(t,f)$ is the probability of a spike occurring in auditory channel f at time t (derived from the Meddis hair cell model) and

$$T=i\Delta t \qquad (3)$$

Autocorrelation functions are computed for values of τ between 0 ms and 12.5 ms in steps of the sampling period Δt. The correlogram is therefore a two-dimensional representation, in which channel centre frequency and autocorrelation lag are represented on orthogonal axes (Figure 2). Since the spectral profile of a vowel is roughly constant over time, subsequent stages of the model are based on the processing of a single correlogram frame. Section 4.3 discusses possible extensions to the model which would enable it to segregate time-varying sound sources.

2.3. Oscillator Network

Grouping is represented by the correlated pattern of firing activity in a network of neural oscillators; specifically, a group of frequency channels is indicated by synchronisation of the corresponding group of oscillators. The building block of the neural network is a single oscillator, which consists of a reciprocally connected excitatory unit x and inhibitory unit y (Wang & Terman, 1995; Terman & Wang, 1995; Wang, 1996, 1997). Formally,

$$\frac{\mathrm{d}x}{\mathrm{d}t}=3x-x^3+2-y+I+S+\rho \qquad (4a)$$

$$\frac{\mathrm{d}y}{\mathrm{d}t}=\varepsilon[\gamma(1+\tanh(x/\beta))-y]. \qquad (4b)$$

Here, I represents the external input to the oscillator, S is the overall coupling from other oscillators in the

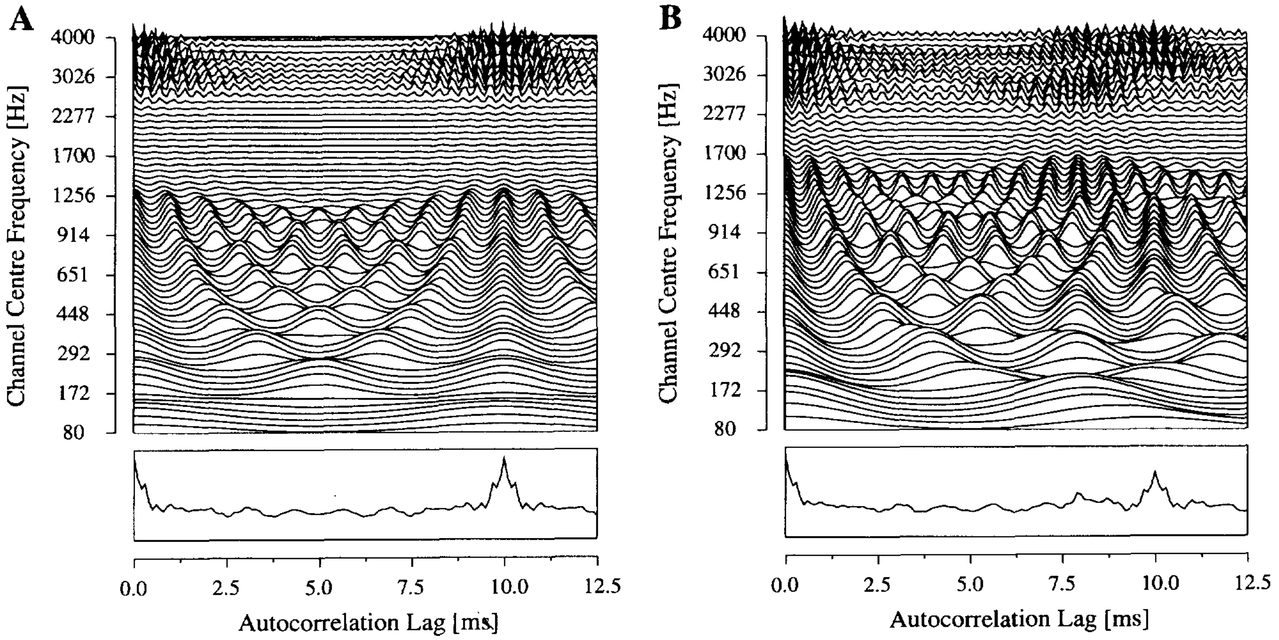

FIGURE 2. (A) Correlogram (top) for the vowel /ah/ (F0 = 100 Hz). The summary correlogram (bottom) exhibits a large peak at the autocorrelation lag corresponding to the period of the vowel (10 ms). (B) Correlogram (top) and summary correlogram (bottom) for a mixture of two vowels, /ah/ (F0 = 100 Hz) and /er/ (F0 = 126 Hz). Peaks occur in the summary function at the period of each vowel (10 ms and 7.9 ms, respectively).

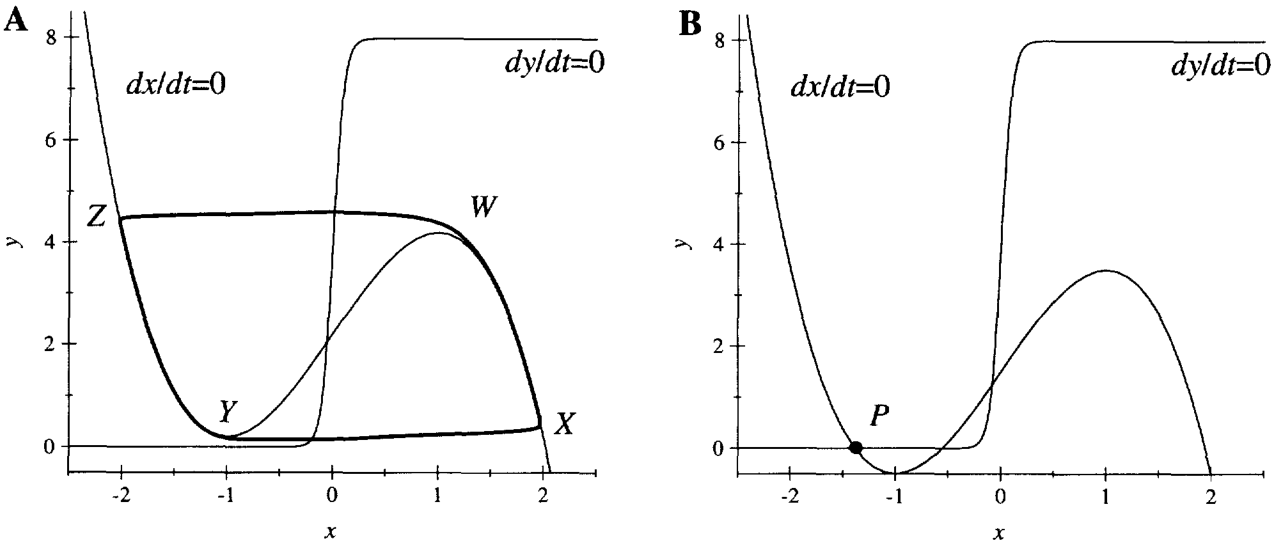

FIGURE 3. Behaviour of a single oscillator in the (*x*,*y*) phase plane. (A) When *I* > 0, the system gives rise to a stable limit cycle (*WXYZ*). The limit cycle consists of rapid transitions between an active phase (*WX*) and a silent phase (*YZ*) of near steady state behaviour. (B) When *I* < 0 the system is not oscillatory; it exhibits a stable fixed point (*P*).

network, ε, γ and β are parameters and ρ is a Gaussian noise term. The principal function of the noise is to assist the desynchronisation of different input patterns.

The behaviour of a single oscillator is best illustrated by considering the solutions of the system (4) in the (x,y) phase plane. The x-nullcline (i.e. dx/dt = 0) of (4) is a cubic curve, while the y-nullcline (i.e. dy/dt = 0) is a sigmoid curve. For $I > 0$ and with ε sufficiently small, (4) gives rise to a stable limit cycle (Figure 3A). This periodic solution alternates between a phase of relatively high values of x and a phase of relatively low values of x, called the *active phase* and *silent phase* respectively. These two phases exhibit near steady state behaviour, whereas the transition between the phases is rapid. Periodic motion of this type is known as relaxation (or discontinuous) oscillation (van der Pol, 1926). We use relaxation oscillators because it has been found recently that a network of such oscillators with local excitatory coupling and global inhibitory coupling rapidly achieves both synchronisation within a block of oscillators that may correspond to a stream or object, and desynchronisation between different blocks (Terman & Wang, 1995; Wang & Terman, 1995). For $I < 0$, there is no periodic solution and the system exhibits a stable fixed point (Figure 3B).

Physiologically, the system (4) may be interpreted as a model of action potential generation in a single neuron, where the state variable x corresponds to the membrane potential of the neuron, and y corresponds to a state variable which describes the activation or inactivation of ion channels and evolves on a slow time scale. Indeed, (4) is closely related to the simple model of action potential generation proposed by FitzHugh (1961) (see also Nagumo et al., 1962). However, the form of (4b) allows the relative durations of the active and silent phases to be specified; a larger value of γ gives a relatively shorter active phase. Thus, (4) has a dimension of flexibility that is absent from the FitzHugh–Nagumo equations (Terman & Wang, 1995). Alternatively, the oscillator model may be interpreted as a mean field approximation to a network of interacting excitatory and inhibitory neurons. In this case, x and y represent the average firing activity of a group of excitatory cells and a group of inhibitory cells, respectively (Wilson & Cowan, 1972).

Each oscillator receives an excitatory input from its own frequency channel. In addition, each oscillator is coupled with a global inhibitor, which receives excitation from every oscillator and feeds back inhibition. The purpose of the global inhibitor is to ensure that weakly coupled groups of oscillators desynchronise to form different streams; in other words, it introduces an element of competition in the oscillator network so that only relatively strong coupling leads to phase synchronisation.

A key concept in our model is the use of neural oscillators as "gates" on their corresponding peripheral channels (Figure 1). Specifically, the activity in a peripheral channel contributes to the percept of a sound source only when the corresponding oscillator is in its active phase (i.e. the "gate" on the channel is "open"). Separation of concurrent sounds is therefore achieved by selectively gating groups of peripheral channels. Initially, all oscillators are synchronised, so that the gates on all channels are open. Starting from this state, simulation proceeds as an autonomous, dynamical process; however, it is convenient to consider the behaviour of the network in three stages, as described below.

2.3.1. *Formation of a Summary Correlogram from Channels with Open Gates.* For a periodic stimulus, each channel of the correlogram exhibits a peak at an autocorrelation lag corresponding to the stimulus period. These peaks line up across frequency, generating a distinctive "spine" (upper panel of Figure 2A). A convenient way of emphasising this pitch-related structure is to sum the channels of the correlogram across frequency, giving a "pooled" or "summary" correlogram. The summary correlogram exhibits a large peak at a lag corresponding to the period of the stimulus (lower panel of Figure 2A). Indeed, Slaney and Lyon (1990) and Meddis and Hewitt (1991) have demonstrated that there is good agreement between the position of the peak in the summary correlogram and perceived pitch. Further, the height of the peak in the summary correlogram can be interpreted as a measure of pitch strength. For mixtures of concurrent sound sources with different F0s, the summary correlogram exhibits a peak at the period of each source (Figure 2B).

In our model, a summary correlogram is formed only from those channels of the correlogram whose gates are currently open (i.e. whose corresponding oscillators are in their active phase). Formally, the summary correlogram $s(t, \tau)$ at time t and lag τ is given by

$$s(t,\tau) = \sum_{f} \left[H(x_f - \theta_f)a(t,f,\tau)\right] \qquad (5)$$

where x_f is the activity of the oscillator connected to peripheral channel f, θ_f is a threshold and H represents the Heaviside step function.

2.3.2. *Pitch Determination.* Since the activity in the correlogram may reflect the presence of more than one pitch, we assume that the largest peak in the summary correlogram corresponds to the pitch period of the most dominant source. Selection of the largest peak in the summary correlogram can be easily computed by a winner-take-all network, as shown in Figure 1 (see Grossberg, 1976, for one formulation of winner-take-all dynamics). However, to reduce the computational load in the current implementation we simply compute the dominant pitch period $d(t)$ by applying a maximum selector. Note that when applying winner-take-all, the search is confined to autocorrelation lags between 12.5 ms and 4.5 ms (corresponding to a pitch range of 80 Hz to 222 Hz). This is acceptable for the vowel stimuli used in this study, which have F0s near to 100 Hz.

2.3.3. *Channel Selection.* In the last stage of a simulation cycle, groups of correlogram channels are identified which are likely to belong to the same vowel, and the input to the oscillator network is updated. This promotes the phase synchronisation of the oscillators which represent the group. For example, consider Figure 2B which shows the correlogram for a mixture of vowels /ah/ (F0 = 100 Hz) and /er/ (F0 = 126 Hz). Peaks occur in the summary correlogram at the period of each vowel, and groups of channel autocorrelation functions are visible which are dominated by one of these two periods. Hence, the spectral components of the dominant vowel can be isolated by selecting those channels of the correlogram which exhibit a peak at the period identified in Section 2.3.2 above. Formally, let P be the set of channels such that for each channel f, $f \in P$ if and only if

$$a(t,f,d(t))/a(t,f,0) > \zeta \qquad (6)$$

The autocorrelation function $a(t,f,0)$ corresponds to the energy in channel f of the correlogram. Hence, (6) represents "peak detection" on the basis of an energy threshold; if the peak height in the autocorrelation function at the delay $d(t)$ exceeds some proportion of the energy in the channel then it is considered a peak. Here we use $\zeta = 0.8$. Our observations suggest that (6) is more reliable than other methods of peak detection, since the low-frequency channels of the correlogram tend to exhibit very broad peaks.

Oscillators corresponding to P receive an additional input, which maintains their synchrony and causes (via the global inhibitor) their desynchronisation from oscillators which do not receive an additional input. After a short time, the group of oscillators that are currently in their active phase move to their silent phase, causing the "gates" to close on the corresponding channels of the correlogram. Similarly, other oscillators that were previously inhibited (in their silent phase) now jump to their active phase, causing a different set of gates in the correlogram to open. The three stages above continue autonomously so long as input is presented to the network. Essentially, then, the dynamics of our neural network embody a process in which initial estimates of the peripheral channels belonging to each sound source are refined during alternating simulation cycles. At the end of this grouping process, the pattern of synchronisation in the oscillator network indicates the frequency channels that belong to each vowel. For a stimulus consisting of a single vowel, nearly all oscillators are phase synchronised.

Regarding the number of oscillation cycles needed to achieve synchronisation, Terman and Wang (1995) have proven that the number of oscillation cycles a general network of the relaxation oscillators described in (4) requires to achieve synchronisation within each segment and desynchronisation between different segments is less than or equal to the number of segments. In our oscillator array, which is a simplified version of theirs, their result implies that the system takes at most two cycles to segregate two groups of frequency channels, each corresponding to the components of a different F0. Our simulation results presented in Section 3 are consistent with their analysis.

2.4. Vowel Recogniser

In the last stage of the model, the groups of correlogram channels represented in the oscillator network are classified by a simple vowel recogniser. This allows the performance of the model to be compared with the recognition performance of human listeners. The segregation mechanism used here does not provide a suitable input for conventional vowel recognition schemes which match observed spectra against templates, since holes appear in the auditory spectrum when the gates for a channel are closed. Instead, a modified recognition scheme proposed by Meddis and Hewitt (1992) is employed, which uses periodicity information from the summary correlogram rather than spectral information.

Meddis and Hewitt make a distinction between the timbre region and the pitch region of a summary correlogram. The timbre region lies between autocorrelation delays of 4.5 ms and 0.1 ms (222 Hz–10 kHz), and contains information about the high frequency components (''timbre'') of the stimulus. This region of the summary correlogram is used for vowel recognition. The pitch region extends from autocorrelation delays of 12.5–4.5 ms (80–222 Hz), and indicates the likely pitches in the stimulus. The pitch region is excluded from the recognition process in order to prevent the F0 of a vowel from influencing its categorisation by the recogniser.

A template is formed for each vowel by computing the correlograms for six presentations of the vowel with different F0s. The six timbre regions obtained in this way are averaged to form the template, which is then normalised to have zero mean and unity variance. Vowel recognition proceeds by matching an input pattern against each template, using an inverse Euclidean distance metric

$$m = \left[1 + \sum_i (r_i - v_i)^2\right]^{-1} \qquad (7)$$

Here, r_i is the ith element of the normalised template and v_i is the ith element of the timbre region to be classified, which is also normalised to have zero mean and unity variance. The best matching template (i.e. the template that gives the largest value of m) is taken to be the vowel identity.

After the network has converged to a stable pattern of oscillations, two successive bursts of oscillations are classified by the vowel recogniser. If the stimulus consists of two vowels with the same F0, the two bursts of oscillations will be identical, and therefore represent the same group of auditory channels. Otherwise, the two bursts of oscillations represent two different groups of channels (the same vowel with two different F0s, or two different vowels with different F0s).

2.5. Example

Before presenting the results of simulations using the neural network oscillator model, we consider an example. Figure 4 shows the output from the model at some of the processing stages described above, for an input consisting of two concurrent vowels, /ah/ (F0 = 100 Hz) and /er/ (F0 = 126 Hz). For this mixture, in which the F0s of the vowels are separated by four semitones, listeners would be expected to perform well in identifying the two constituent vowels.

Figure 4A shows the activity x in each channel of the oscillator network. Oscillators initially have the same phase, but quickly segregate into a repeating pattern of two phase-synchronised groups. Each group of oscillations represents a set of channels in the correlogram, shown in panels B and C of the figure. For each group of channels, the position of the largest peak in the pitch region of the summary correlogram indicates the period of the vowel, and the profile of the timbre region indicates the vowel identity. The channels shown in Figure 4B have centre frequencies close to the five formant frequencies of the /ah/ (F1 = 650 Hz, F2 = 950 Hz, F3 = 2950 Hz, F4 = 3300 Hz, F5 = 3850 Hz). Note also that the summary correlogram for this group of channels has a large peak at 10 ms in its pitch region, which

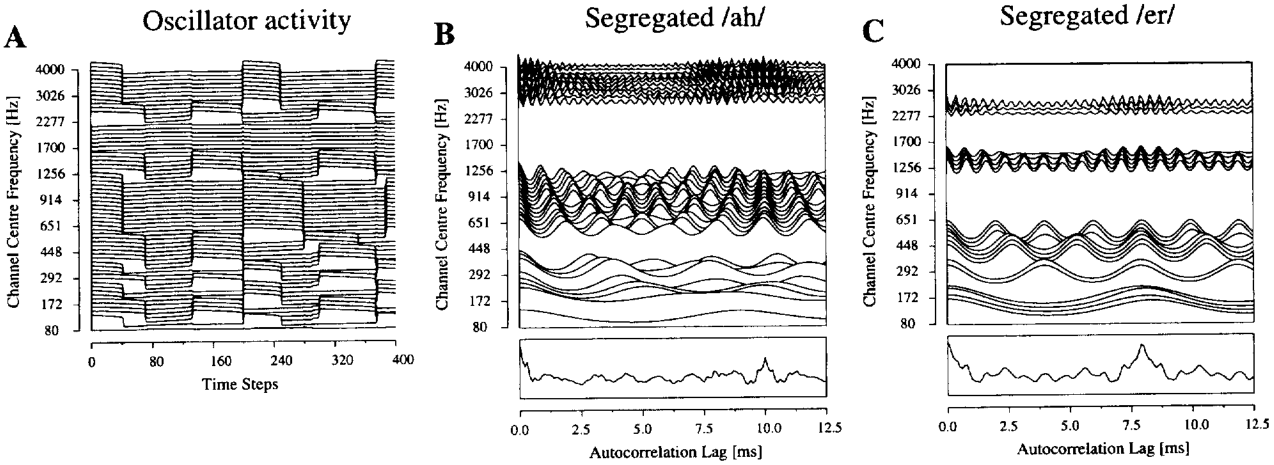

FIGURE 4. Behaviour of the model for a stimulus consisting of the double vowel /ah/ (F0 = 100 Hz) and /er/ (F0 = 126 Hz). Neurons in the oscillator network (A) phase synchronise to form two groups, corresponding to correlogram channels which define the formants of the two vowels (B and C).

correctly indicates that the F0 of the vowel /ah/ is 100 Hz. Similarly, the channels in Figure 4C correspond to the first three formants of the vowel /er/ (F1 = 450 Hz, F2 = 1250 Hz, F3 = 2560 Hz), and the F0 is correctly indicated as 126 Hz.

3. SIMULATIONS

The evaluation described below compares the performance of the model with that of human listeners for the same double vowel identification task, using listeners' data from a study by Assmann & Summerfield (1990). Two versions of the model were investigated; one in which auditory channels were allocated to a single vowel, and a version in which the same auditory channel could be allocated to both vowels. We also demonstrate that the model can replicate findings from harmonic mistuning experiments (Moore et al., 1985).

3.1. Segregation of Double Vowels

Assmann & Summerfield (1990) studied the perception of double vowels using a set of five British English monophthongs; /ɑ/, /i/, /ɛ/, /u/ and /ɔ/ (here, we refer to them as /ah/, /ee/, /er/, /oo/ and /or/ respectively. The vowels had a duration of 200 ms, and were synthesised using the Klatt (1980) cascade formant synthesis formant algorithm at a sampling frequency of 10 kHz. Each vowel was synthesised on six F0s, corresponding to differences of 0, 0.25, 0.5, 1, 2 and 4 semitones from 100 Hz. Listeners were presented with pairs of vowels, and asked to identify both vowels in the mixture. In each trial, one vowel always had a F0 of 100 Hz, and the other constituent could be any vowel on any F0. In total, there were 150 stimuli (5 vowels × 5 vowels × 6 fundamental frequencies).

The results from the Assmann and Summerfield (1990) study are shown in Figure 5. For vowel pairs with the same F0, listeners were able to correctly identify both vowels at well above the chance level of 6.7%. Identification performance increased with increasing difference in F0 between the two vowels, reaching a plateau at a separation of two semitones.

Figure 5 also shows the vowel identification performance of the neural oscillator model, for the same set of 150 stimuli used by Assmann and Summerfield. The vowels were presented in pairs, and the response of the model was accepted as correct only if it correctly identified both constituent vowels.With the exception of the 0 and 0.25 semitone conditions, the vowel identification performance of the model is comparable to human listeners. More importantly, however, it shows the required pattern of response; an increase in vowel identification performance with increasing difference in F0.

In the simulations described so far, the oscillations in the neural network partition the channels of the correlogram into mutually exclusive sets; in other words, the energy in each frequency channel is assigned only to one source. The model therefore exploits a "principle of exclusive allocation" (Bregman, 1990). Although there are many examples of the application of this principle in perceptual organisation there are also some exceptions, many of which occur in the perception of speech sounds (e.g. the "duplex" perception of speech; Rand, 1974; Liberman et al., 1981). It is therefore instructive to consider the vowel identification performance of our model when joint allocation of channels is permitted.

In the oscillator network we have explained so far, the

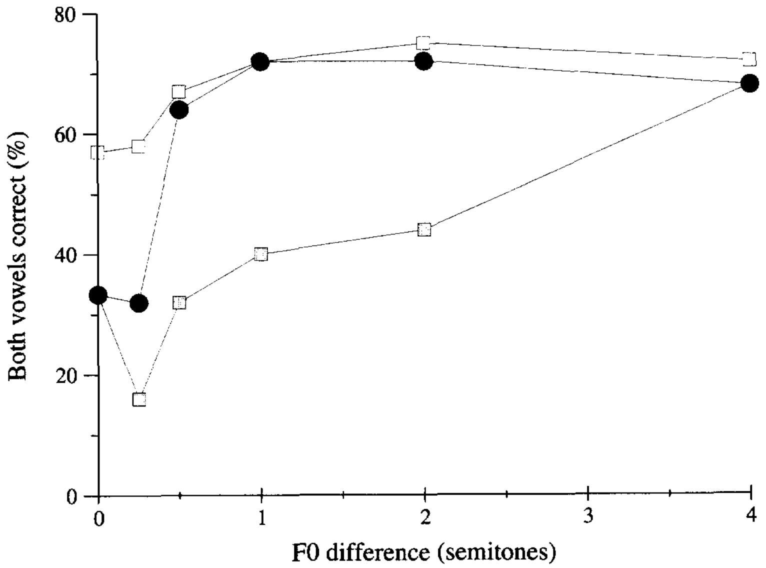

FIGURE 5. Percentage of vowel pairs for which both vowels were correctly identified. Results are shown for model simulations with exclusive allocation (●) and joint allocation (▨) of channel energy, together with data for listeners (□) from the study by Assmann and Summerfield (1990).

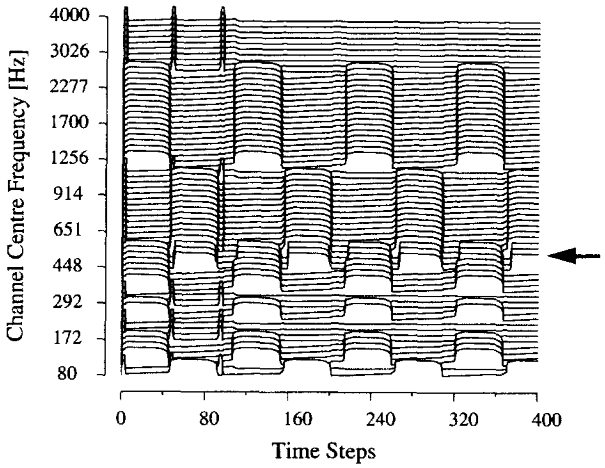

FIGURE 6. Activity in the oscillator network elicited by the double vowel /ah/ (F0 = 100 Hz) and /er/ (F0 = 126 Hz) for a version of the model which allows the shared allocation of auditory channels. Two groups of oscillations emerge, but they are not mutually exclusive. The 5th harmonic of the /ah/ (500 Hz) and the 4th harmonic of the /er/ (504 Hz) are nearly coincident, and hence the model allocates energy in the 500 Hz region (arrowed) to both vowels.

oscillators oscillate with the same frequency, and their phases signal the identity of each group (vowel). To allow joint allocation of peripheral channels, we require that the oscillators stimulated by shared channels oscillate with a frequency double that of the oscillators stimulated by non-shared channels (Figure 6). Such behaviour can be achieved by introducing a stable fixed point on the higher branch of the y-nullcline of (4) (see Figure 3) and the use of a slow inhibition mechanism. The purpose of the new fixed point is to reduce the phase variations between oscillators corresponding to shared channels and those corresponding to non-shared channels. On the other hand, slow inhibition is stimulated when any oscillator jumps to the active phase, but it takes some time to be activated. When the slow inhibition mechanism is activated, it causes all of the oscillators in the active phase to simultaneously jump down to the silent phase. We note that slow inhibition has been used frequently in biologically realistic modelling (for example, see Wang & Rinzel, 1992; Terman & Lee, 1996).

The performance of the joint allocation model on Assmann and Summerfield's (1990) stimuli is shown in Figure 5. Although the model shows the correct pattern of response, its vowel identification is worse than that of listeners, and is also below the performance of the exclusive allocation version of the model. This problem is most likely due to spurious peaks in the high-frequency channels of the correlogram, which are not related to the period of either vowel (the problem of "overlapping harmonics"; see also Summerfield et al., 1990; Brown, 1992). With exclusive allocation, there is a 50% chance that channels with spurious peaks will be inappropriately grouped depending on which F0 is detected first (if the channel is first allocated to the appropriate F0, it cannot then be allocated to the second F0). When joint allocation is permitted, however, such channels will always be inappropriately grouped.

3.2. Segregation of a Mistuned Harmonic from a Complex Tone

The segregation of concurrent harmonic sounds has been investigated by Moore et al. (1985) using a mistuning paradigm. They presented listeners with a harmonic complex in which one component was mistuned by a percentage of its frequency, so that it was no longer an integer multiple of the F0 of the complex. For small amounts of mistuning (up to 3% of the harmonic frequency), the mistuned harmonic made a normal contribution to the perceived pitch of the complex. However, the contribution of the mistuned harmonic decreased with progressively larger mistuning, until at 8% it made no contribution to the pitch of the complex.

Since our model is intended to explain the grouping of acoustic components by common F0, we would expect it to be able to replicate the findings from mistuning experiments. Figure 7 shows the activity in the oscillator network elicited by two stimuli used in the study of Moore et al. (1985). The stimuli had a duration of 90 ms, and consisted of the first 12 harmonics of a 155 Hz fundamental. In some conditions the 4th partial was mistuned by a percentage of its harmonic frequency. For a stimulus with no mistuning (Figure 7A), the responses of most oscillators in the network are phase synchronised. The only channels excluded from the group are those in which there is no significant energy in the stimulus; specifically, those above 2200 Hz and channels in the region of 230 Hz (which lie between the first and second harmonics of the complex). However, in the 8% mistuned condition (Figure 7B) channels close to the frequency of the mistuned harmonic (670 Hz) form a separate perceptual group, as indicated by the different phase of their oscillations. The switch from one group of oscillations to two groups occurs at a mistuning of 4%, which is in good agreement with psychophysical findings; Moore et al. (1985) report that listeners hear a harmonic that is mistuned by more than 3% as a separate sound source.

In addition, the model partially accounts for the pattern of change in perceived pitch reported by listeners for increasing amounts of mistuning. For mistunings of 3% or less, one group of oscillations is observed; therefore, the channels close to the frequency of the fourth harmonic are included in the computation of the summary correlogram. This causes the position of the peak in the summary function to shift, which corresponds to the change in pitch reported by listeners. For a mistuning of 8%, the 4th harmonic is represented by a separate group of oscillations (Figure 7b). Therefore, the mistuned harmonic makes no contribution to the pitch

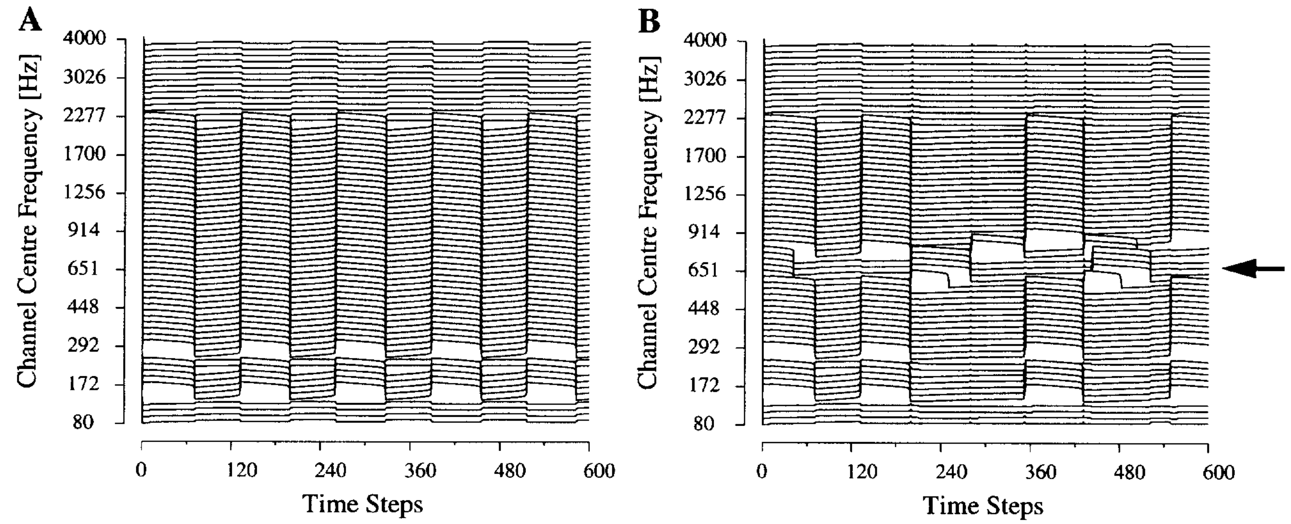

FIGURE 7. (A) Activity in the oscillator network elicited by a stimulus consisting of the first 12 harmonics of a 155 Hz fundamental. (B) Oscillator activity for the same stimulus, but with the 4th harmonic mistuned by 8% of its frequency. Neurons which receive input from channels close to the frequency of the mistuned harmonic (670 Hz) do not synchronise with the remaining oscillators, indicating that the mistuned harmonic is segregated from the other components of the stimulus.

of the complex (i.e. it is excluded from the computation of the summary correlogram), as reported by listeners.

However, the 4th harmonic is also represented by a separate pattern of oscillations for mistunings between 4% and 8%. Since exclusive allocation is enforced in the version of the model used here, the model predicts that the 4th harmonic will make no contribution to the pitch of such stimuli; it will be heard as a separate source. In contrast, listeners report that the mistuned harmonic is heard as a separate source *and* makes a contribution to the pitch; this is therefore a violation of the principle of exclusive allocation. A version of the model which permits the sharing of channels between sources would be compatible with the experimental findings; however, it was noted in Section 3.1 that such a scheme performed poorly in our vowel segregation experiments. Clearly, the relative merits of exclusive and joint allocation schemes should be a topic for further investigation.

4. GENERAL DISCUSSION

We have presented a neural oscillator model of auditory grouping, which is able to replicate listeners' performance in double vowel identification and mistuning experiments with good accuracy. In this section we discuss the compatibility of the model with other psychophysical findings. Additionally, we consider the limitations of the model, and discuss its relationship with other models of double vowel segregation.

4.1. Compatibility with Psychophysical Findings

An important aspect of our model is the manner in which neural oscillators act as gates for their corresponding frequency channels; when the oscillator moves to its active phase , the "gate" opens and allows the energy in the channel to contribute to the percept of the sound source. Such gating mechanisms have been criticised as too rigid by Bregman (1992), on the grounds that they only allow the energy in an auditory channel to be allocated to a single perceptual stream. We have demonstrated in Section 3.1 that an explanation of auditory grouping based on gating does *not* imply that groups of channels must be mutually exclusive; the relaxation oscillators used here may be permitted to synchronise with more than one group of oscillations. However, we also found that the joint allocation version of our model provided a poorer match to listeners' vowel identification performance than one in which exclusive allocation of channels is enforced, a finding which is compatible with the results of other modelling studies (Assmann & Summerfield, 1990; Meddis & Hewitt, 1992; Summerfield et al., 1990). This finding is intriguing, given that joint allocation of auditory channels is required to model other well known effects in speech perception, such as the "duplex" perception of speech sounds (Rand, 1974; Liberman et al., 1981).

Recently, McKeown and Patterson (1995) have investigated the effect of stimulus duration on the identification of double vowels. They presented listeners with double vowels varying in duration between one and eight waveform cycles, and found that the ability of listeners to identify both vowels improved with increasing stimulus duration. However, this improvement was almost completely due to improved identification of one of the vowels (the "nondominant" vowel); even for a duration of one waveform cycle, identification of the other ("dominant") vowel in the mixture was near ceiling performance. Since very short stimuli do not elicit a clear pitch percept, this result appears to be incompatible with models of double vowel segregation which require F0 estimates for both vowels (Assmann & Summerfield, 1990; de Cheveigné, 1993) or for the dominant vowel only (Meddis & Hewitt, 1992).

Although our model employs information about the F0 of both vowels, it is not in conflict with McKeown and Patterson's data. In conditions where the stimulus duration is very short, there is no meaningful pitch information in the summary correlogram; as a result, all oscillators remain synchronised and every channel of the correlogram contributes to the computation of the timbre region. Hence, the oscillator network reports a single group, and the timbre region for this group will resemble the dominant vowel rather than the nondominant vowel. In other words, for very short stimulus durations only the dominant vowel is correctly identified. For stimuli of more than one cycle duration, pitch information is available in the summary correlogram. Since the pitch estimates are more accurate for longer stimuli, the model predicts that segregation of the nondominant vowel should improve with increasing stimulus duration. Indeed, simulations on a subset of Mckeown and Patterson's stimuli indicate that our model is able to replicate their findings; the dominant vowel is correctly identified for all stimulus durations, whereas the nondominant vowel is correctly identified only for durations of five or six waveform cycles.

4.2. Relationship with Other Modelling Studies

There are some similarities between our model and that of Meddis and Hewitt (1992). In particular, we use their vowel recognition scheme, which is based on template matching of the short-lag periodicities (''timbre-region'') of the summary correlogram. However, there are also many differences in our two approaches. First, the Meddis and Hewitt scheme is an algorithmic description of the vowel segregation process, which does not appeal to specific neural mechanisms; in contrast, the neural oscillator model proposed here is an autonomous, dynamical system supported by recent findings in neurobiology. A second difference is that F0 estimation is a one pass mechanism in the Meddis and Hewitt scheme, whereas in our model the F0 of each source is progressively refined. Further, the Meddis and Hewitt scheme only determines the F0 of the dominant vowel, whereas our model determines the F0s of both vowels. Our approach finds better support in psychophysical literature. Summerfield et al. (1990) note that listeners can often hear both pitches in a double vowel, and are able to indicate which vowel has the higher pitch and which has the lower pitch. Similarly, Beerends and Houtsma (1989) have found that listeners are able to correctly identify the pitches of concurrent two-tone complexes, for differences in F0 of two semitones or more. It seems unlikely, then, that segregation is based only on the F0 of the dominant vowel.

As noted above, the model achieves the segregation of sound sources by a dynamical process; the set of channels and F0 are refined for each source in alternation, until there is no further change in the state of the network. Other approaches to source separation which progressively refine the F0 estimate for each source have been described by a number of workers (Parsons, 1976; Scheffers, 1983; Lea, 1992; de Cheveigné, 1993), although not in the context of a neural network architecture. These approaches share the same rationale as the model described here; namely, that it is easier to estimate the characteristics of one source in an acoustic mixture when the other competing sources are attenuated. In the model proposed here, attenuation occurs when the group of oscillators representing a source move to their silent phase; this closes the ''gates'' on the corresponding group of auditory frequency channels.

4.3. Limitations of the Model

Currently, the model does not address the representation of time in the auditory system; the input to the neural oscillator network is a single correlogram frame. While this is acceptable for the limited situations with steady state-stimuli considered here, it is clearly inadequate for the segregation of acoustic signals which vary in time-frequency. One possible approach to time representation is to create a time axis in a two-dimensional oscillator network using a system of delay lines (Wang, 1996, 1997). Alternatively, an inertia could be introduced to the oscillator network, which would maintain a pattern of oscillations after the pattern was turned off (Horn & Usher, 1992); such a mechanism would act as a short term memory.

Another limitation of the model is the arbitrary division between the pitch and timbre regions of the summary correlogram (see also Meddis & Hewitt, 1992). For example, the voice pitch of a female speaker is likely to exceed the 222 Hz limit of the pitch region, and cause a peak in the timbre region. A possible solution to this problem is to perform vowel recognition in the spectral domain with a template matching algorithm which performs robustly when there are gaps in the input spectrum (caused by the gates on some channels being closed). Recent work on the partial matching of speech signals (Cooke et al., 1994) suggests that this is a viable approach.

5. CONCLUSION

A computational model of the perceptual segregation of concurrent vowels has been proposed. Groups of auditory channels which define the spectrum of each vowel are represented by the pattern or temporal synchronisation in a network of neural oscillators. The model is able to replicate the finding that listeners' identification of concurrent vowels improves with increasing difference in F0 between the vowels. Additionally, the model is able to replicate the results of harmonic mistuning experiments. Future work will extend the model to include a representation of time, thus allowing the segregation of acoustic signals that vary in time-frequency.

REFERENCES

Assmann, P. F., & Summerfield, Q. (1990). Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies. *Journal of the Acoustical Society of America, 88,* 680–697.

Beerends, J. G., & Houtsma, A. J. M. (1989). Pitch identification of simultaneous diotic and dichotic two-tone complexes. *Journal of the Acoustical Society of America, 85,* 813–819.

Baird, B. (1996). *A cortical network model of cognitive attentional streams, rhythmic expectation and auditory stream segregation* (CPAM Technical report 173-96). Berkeley, CA: Department of Mathematics, University of California.

Barth, D. S., & MacDonald, K. D. (1996). Thalamic modulation of high-frequency oscillating potentials in auditory cortex. *Nature, 383,* 78–81.

Bregman, A. S. (1990). *Auditory scene analysis.* Cambridge, MA: MIT Press.

Bregman, A. S. (1992). How does physiology support auditory scene analysis? In Y. Cazals, L. Demany & K. Horner (Eds.), *Advances in the Biosciences Volume 83: Auditory Physiology and Perception* (pp. 417–427). Pergamon Press.

Brown, G. J. (1992). *Computational auditory scene analysis: A representational approach.* Unpublished Ph.D. thesis, University of Sheffield.

Brown, G. J., & Cooke, M. P. (1992). Grouping sound sources using common pitch contours. *Proceedings of the Institute of Acoustics, 14,* 439–446.

Brown, G. J., & Cooke, M. P. (1994). Computational auditory scene analysis. *Computer Speech and Language, 8,* 297–336.

Brown, G. J. & Cooke, M. P. (1997). Temporal synchronisation in a neural oscillator model of primitive auditory stream segregation. In H. Okuno & D. Rosenthal (Eds.), *Readings in Computational Stream Analysis.* Lawrence Erlbaum, in press.

BSI (1988). *Normal equal loudness level contours for pure tones under free-field listening conditions* (BS 3383). London: British Standards Institution.

Chalikia, M. H., & Bregman, A. S. (1989). The perceptual separation of simultaneous auditory signals: Pulse train segregation and vowel segregation. *Perception and Psychophysics, 46,* 487–496.

Cooke, M. P. (1993). *Modelling auditory processing and organisation.* Oxford University Press.

Cooke, M. P., Green, P. D. & Crawford, M. D. (1994). Handling missing data in speech recognition. *Proceedings of the International Conference on Spoken Language Processing* (ICSLP) (pp. 1555–1558). Yokohama.

de Cheveigné, A. (1993). Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *Journal of the Acoustical Society of America, 93,* 3271–3290.

Eckhorn, R., Bauer, R., Jordan, W., Brosh, M., Kruse, W., Munk, M., & Reitboeck, H. J. (1988). Coherent oscillations: A mechanism of feature linking in the visual cortex. *Biological Cybernetics, 60,* 121–130.

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophysical Journal, 1,* 445–466.

Freeman, W. J. (1991). Nonlinear dynamics in olfactory information processing. In J. L. Davies, & H. Eichbaum (Eds.), *Olfaction* (pp. 225–249). Cambridge, MA: MIT Press.

Galambos, R., Makeig, S., & Talmachoff, P. J. (1981). A 40 Hz auditory potential recorded from the human scalp. *Proceedings of the National Academy of the Sciences of the USA, 78,* 2643–2647.

Glasberg, B. R., & Moore, B. C. J. (1990). Derivation of auditory filter shapes from notched-noised data. *Hearing Research, 47,* 103–138.

Gray, C. M., König, P., Engel, A. K., & Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronisation which reflects global stimulus properties. *Nature, 338,* 334–337.

Gray, C. M., & McCormick, D. A. (1996). Chattering cells: superficial pyramid neurons contributing to the generation of synchronous oscillations in the visual cortex. *Science, 274,* 109–113.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics, 23,* 121–134.

Horn, D. & Usher, M. (1992). Oscillatory model of short term memory. In J. E. Moody, S. J. Hanson, & R. P. Lippmann (Eds.), *Advances in Neural Information Processing Systems 4* (pp. 125–132). Morgan Kaufmann.

Joliot, M., Ribary, U., & Llinás, R. (1994). Human oscillatory brain activity near to 40 Hz coexists with cognitive temporal binding. *Proceedings of the National Academy of the Sciences of the USA, 91,* 11748–11751.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America, 67,* 971–995.

Lea, A. (1992). *Auditory models of vowel perception.* Unpublished Ph.D thesis, University of Nottingham.

Liberman, A. M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics, 30,* 133–143.

Llinás, R., & Ribary, U. (1993). Coherent 40 Hz oscillation characterises dream state in humans. *Proceedings of the National Academy of the Sciences of the USA, 90,* 2078–2082.

Madler, C., & Pöppel, E. (1987). Auditory evoked potentials indicate the loss of neuronal oscillations during general anesthesia. *Naturwissenschaften, 74,* 42–43.

Mäkelä, J. P., & Hari, R. (1987). Evidence for the cortical origin of 40 Hz evoked response in man. *Electroencephalographical and Clinical Neuropsychology, 66,* 539–546.

McKeown, J. D., & Patterson, R. D. (1995). The time course of auditory segregation: Concurrent vowels that vary in duration. *Journal of the Acoustical Society of America, 98,* 1866–1877.

Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *Journal of the Acoustical Society of America, 79,* 702–711.

Meddis, R. (1988). Simulation of auditory-neural transduction: Further studies. *Journal of the acoustical society of America, 83,* 1056–1063.

Meddis, R., & Hewitt, M. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery: I. Pitch identification. *Journal of the Acoustical Society of America, 89,* 2866–2882.

Meddis, R., & Hewitt, M. (1992). Modelling the identification of concurrent vowels with different fundamental frequencies. *Journal of the Acoustical Society of America, 91,* 233–245.

Milner, P. M. (1974). A model for visual shape recognition. *Psychological Review, 81,* 521–535.

Moore, B. C. J. (1997). *An introduction to the psychology of hearing* (4th ed.). San Diego: Academic Press.

Moore, B. C. J., Glasberg, B., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America, 77,* 1853–1860.

Nagumo, J., Arimoto, S., & Yoshizawa, S. (1962). An active pulse transission line simulating nerve axon. *Proceedings of the Institute of Radio Engineers, 50,* 2061–2070.

Parsons, T. W. (1976). Separation of speech from interfering speech by means of harmonic selection. *Journal of the Acoustical Society of America, 60,* 911–918.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I. & Rice, P. (1988). *SVOS final report, part B: Implementing a gammatone filterbank* (Applied Psychology Unit report 2341). Cambridge.

Rand, T. C. (1974). Dichotic release from masking for speech. *Journal of the Acoustical Society of America, 55,* 678–680.

Ribary, U., Ionnides, A. A., Singh, K. D., Hasson, R., Bolton, J. P. R., Lado, F., Mogilner, A., & Llinás, R. (1991). Magnetic field tomography of coherent thalamocortical 40 Hz oscillations in humans. *Proceedings of the National Academy of the Sciences of the USA, 88,* 11037–11041.

Scheffers, M. T. M. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation.* Unpublished Ph.D thesis. University of Gröningen.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience, 18,* 555–586.

Slaney, M. & Lyon, R. F. (1990). A perceptual pitch detector, *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 357–360).

Summerfield, Q., Lea, A., & Marshall, D. (1990). Modelling auditory scene analysis: Strategies for source segregation using autocorrelograms. *Proceedings of the Institute of Acoustics, 12,* 507–514.

Terman, D., & Lee, E. (1996). Partial synchronization in a network of neural oscillators. *SIAM Journal on Applied Mathematics, 57,* 252–293.

Terman, D., & Wang, D. L. (1995). Global competition and local cooperation in a network of neural oscillators. *Physica D, 81,* 148–176.

van der Pol, B. (1926). On ''relaxation oscillations''. *Philosophical Magazine,* 2(11), 978–992.

von der Malsburg, C. (1981). *The correlation theory of brain function* (Internal report 81-82). Göttingen, FRG: Max Planck Institute for Biophysical Chemistry.

von der Malsburg, C., & Schneider, W. (1986). A neural cocktail party processor. *Biological Cybernetics, 54,* 29–40.

Wang, D. L. (1994). Auditory stream segregation based on oscillatory correlation. In *Proceedings of the IEEE 1994 Workshop on Neural Networks for Signal Processing* (pp. 624–632).

Wang, D. L., & Terman, D. (1995). Locally excitatory globally inhibitory oscillator networks. *IEEE Transactions on Neural Networks, 6,* 283–286.

Wang, D. L. (1996). Primitive auditory segregation based on oscillatory correlation. *Cognitive Science, 20,* 409–456.

Wang, D. L. (1997). Stream segregation based on oscillatory correlation. In H. Okuno, & D. Rosenthal (Eds.), *Readings in Computational Auditory Scene Analysis.* Laurence Erlbaum. in press.

Wang, X. J., & Rinzel, J. (1992). Alternating and synchronous rhythms in reciprocally inhibitory model neurons. *Neural Computation, 4,* 84–97.

Wilson, H. R., & Cowan, J. D. (1972). Excitatory and inhibitory interactions in localised populations of model neurons. *Biophysical Journal, 12,* 1–24.