Draft of a chapter for "Listening to speech: An auditory perspective," S. Greenberg & W. Ainsworth, eds.

Modeling the auditory organization of speech - a summary and some comments

Daniel P.W. Ellis <dpwe@icsi.berkeley.edu>
International Computer Science Institute
Berkeley CA

Introduction

The preceding three chapters have been concerned with the issues arising as a result of the inconvenient fact that our ears are rarely presented with the sound of a single speaker in isolation, but more often with a combination of several speech and nonspeech sounds which may also have been further altered by the acoustic environment. Faced with such a mixture, the listener evidently needs to consider each source separately, and this process of information segregation is known as *auditory organization* or *auditory scene analysis* (Bregman, 1990). Pure curiosity as well as the possibility of applications in automatic signal inerpretation drive us to investigate auditory scene analysis through psychological experiments and computational modeling.

Since these papers bear on several different aspects of auditory organization, we may begin the discussion by sketching a view of the entire problem. Starting with the broadest question of the purpose of hearing and our ability to separate different sources, we note that sound – the transmission of air-pressure variations in a certain frequency range – offers a potential source of information concerning events in the physical world which might lead to an adapatation in the listener's behavior. There are particular sources and attributes that the listener cares about (starting, perhaps, with threats and including subtle nuances of communication), and thus to be useful, auditory perception needs to recover this information in the widest range of circumstances i.e. to separate interesting sources from one another and other interference.

We see immediately that the definition of a source in this context is not so much based on some objective physical quality, but on far more elusive subjective considerations of a listener's internal conception of the world, which can depend on many incidental factors. The sound of a car passing in the street is perceived as one source, if we distinguish it from the background at all, but the sound of my own car may be analyzed into separate percepts for the sound of the engine, tyres, and that worrying rattle from the rear suspension. Conversely, the physically distinct speech sounds made by the vibrating vocal chords (in voiced portions) and turbulant constrictions of the vocal tract (as in fricatives) are almost irresistably heard as a single source, presumably because of our strong and benficial inclination to treat the communication sounds of a given speaker as a single, co-ordinated source.

The overall problem, then, is to separate arbitrary information from unconstrained mixtures of subjectively-defined sources. The fact that we have a sense of hearing at all demonstrates that this is something that we accomplish quite successfully, in spite of the broad requirements. By the same token, most situations will afford only a limited 'view' of any individual source, and the auditory system has evolved both to distinguish the information that can be separated, and to tolerate the absence of the information that cannot be recovered in a particular situation. The complement of the auditory system's ability to separate sources is its ability to 'degrade gracefully' (Marr, 1982) in demanding situations, manifested as a preconscious inference of hidden features; both these functions are required for a useful perceptual modality.

Figure 1 shows the basic structure we use to model the auditory system. The incoming physical (sound) stimulus is converted by the transducers and processing in the front end to generate the intermediate representation in terms of various feature values. The value of such a mid-level representation comes from meeting the dual constraints of being directly and unambiguously computable from the explicit physical properties of the original signal, while at the same time constituting a more suitable domain for the more abstract and uncertain organization that is to follow (Ellis & Rosenthal, 1998). This representation can benefit from displaying multiple features in parallel 'maps' as advocated by Brown (1992); one particularly useful feature map is the correlogram (Slaney & Lyon, 1983; Meddis & Hewitt, 1991; strongly related to the processing described in the chapter by Todd & Lee, and used as the input to the system of Brown & Wang), which displays the physical property of amplitude modulation on an autocorrelation axis (as a function of time and channel), upon which

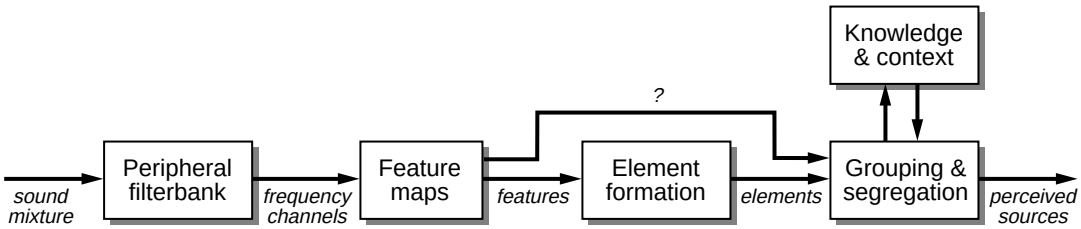

Figure 1: Block-diagram of the possible structure auditory organization, as typically assumed by modellers. The bypass around "element formation" acknowledges the debate over whether segmentation can be done in a purely bottom-up fashion, or whether it must depend on top-down knowledge.

the abstract grouping principle of common periodicity may be based, and from which the source attribute of pitch may be calculated (Ellis, 1997).

The sources perceived as present in the scene by the higher levels of cognitive processing are derived from the intermediate representation through the process of grouping & segregation, a catch-all category which might be better described as several subcomponents, but whose structure is not clear. In the most straightforward conception of auditory organization, suggested by Bregman and realized, for example, in the models of Cooke (1991), Brown (1992), Mellinger (1991) and Ellis (1994), the intermediate representation is a collection of discrete atomic 'elements', such as sine tones and voice formants, each describing a contiguous region of energy in time-frequency, and each related to a single source. To construct the perceived sources from these is literally a question of grouping together elements on the basis of their properties, according to rules such as common onset, common periodicity and perhaps more equivocal gestalt principles such as 'good continuation'. Although combining these rules, which might be in opposition, is a computational challenge, the overall process of finding an exhaustive disjoint partition of the intermediate elements constitutes a well-formed definition of the scene analysis task.

Unfortunately, the experience of these implementations casts doubt on the assumption that it would be possible to find a unique analysis of an input sound into discrete, atomic elements. Firstly, real sounds include numerous overlaps in the time-frequency plane between the energy of different sources, and even the third axis of the correlogram (periodicity) cannot always separate the features arising from different sources. Secondly, a number of experimental phenomena – the restoration phenomena discussed in Warren's chapter being the most important – show that high-level perceptions are not always based directly on signal features, but can arise from knowledge, memory and inference. This leads to a more nebulous conception of the grouping & segregation function, in which rather than operating on ready-segmented elements, continuously-valued features are used as the basis for hypotheses concerning which sound sources might be present, which can then be used both to accumulate evidence from the features, and as a locus for the constraints and predictions of abstracted knowledge and memory. Models that have taken this approach in the analysis of real sounds into perceived sources include Ellis (1996), Klassner (1996), Nakatani *et al*. (1998) and Godsmark & Brown (1997).

Having sketched this framework and the current limits to our understanding of the process of auditory organization, we can now examine the material of each of the three chapters in more detail, seeing how it fits into this framework and also where the framework may be inadequate. Following these discussions, we will conclude with some remarks suggested by the particular combination of results in this section.

Todd & Lee: Multiscale Representations & Rhythm

The first paper we consider is the "sensory-motor" theory of Todd & Lee. This paper contains very many ideas, among which the principle themes are:

- An outline of an auditory front-end model inspired by a theory of primary visual processing;
- A critique of the way in which current speech recognition systems ignore speech variability and prosodic information;
- A discussion of how the novel features of the front end could support the perception of prosidy and rhythm, which are seen as critical to speech understanding, particularly through sensory integration.

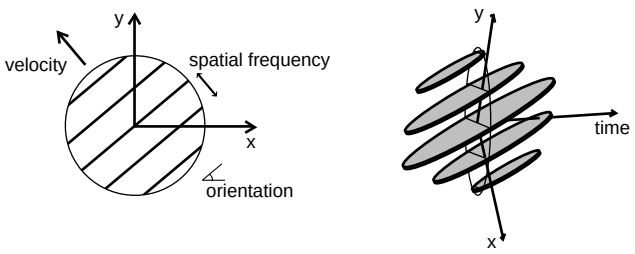

Figure 2: The three-dimensional kernels of Heeger (1987) are sensitive to grids with a particular spacing and orientation, moving in the image plane at a particular velocity.

The authors argue passionately for the importance of the prosodic information in speech, and discuss how their approach could resolve many problems, from speech variability to language acquisition. To examine these intriguing claims, we first look at the front-end model in some detail, then go on to consider the more general question of assessing alternative representations.

The Todd & Lee auditory model

The front-end processing presented in the paper starts from an appealing premise: since hearing and vision are the two most important sensory modalities, might we not expect considerable similarities in how they operate? In particular, they take a model of the detection of visual flow in the primary cortex (Heeger, 1987) and see what could be expected from a similar structure employed in the hearing system.

The Heeger model posits the existence of families of neurons specifically tuned to detect moving textures (visual flow). Each neuron responds to information from a particular area of the retina, and is tuned to respond to a 'grid' of parallel lines at a particular spatial frequency and orientation, moving at a particular velocity. Heeger models this receptive field with linear filters whose reversed impulse response (in a three-dimensional space formed by extending the two-dimensional image grid along time) is an oriented, slanted set of planes corresponding to the tuned velocity, as illustrated in figure 2. Since motion of the grid parallel to its gratings has no visible effect, there is only one velocity axis corresponding to the motion of the pattern perpendicular to the grid. Thus each neuron has three degrees of freedom – spatial frequency, orientation and velocity – which map to the ω_x, ω_y and ω_t of eqn. (1) in Todd & Lee. The 'size' or locality of the filter in time and space give another three dimensions corresponding to σ_x, σ_y and σ_t in the equation, typically held constant over a family of tunings. In the equation, the filter is implicitly centered at the origin in all three dimensions, although there are presumably different families for different locations in the image plane.

The novelty in this model is that it addresses the perception of image dynamics – the movement of spatial-frequency features – as a direct function of the time-varying image, rather than as a secondary function calculated from the output of static, two-dimensional detectors for oriented grids. Todd & Lee are concerned with the perception of dynamics in sound, which explains their interest in this model. There are, however, profound differences to the physics governing information carried by light and by sound waves, so the first question to be answered in transferring the model from vision to hearing is over the interpretation to be used for the two spatial dimensions in the vision model. Todd & Lee answer this by augmenting the natural cochleotopic axis (arising from the array of tuned basilar membrane filters in the cochlea) with a periodotopic axis, in which the envelope within each peripheral frequency channel has been analyzed by a second array of band-pass filters to expose modulations in the pitch range of 10-1000 Hz. Similar two-dimensional frequency-period displays have become increasing popular in models of auditory processing, in particular for their utility in modeling pitch-perception phenomena. Detecting periodic modulation by autocorrelation was first suggested by Licklider (1951) as part of his duplex pitch-perception model, and has recently been popularized by Meddis & Hewitt (1991) and Slaney & Lyon (1993), who call the three-dimensional frequency-lag-time display the *correlogram*; the close relationship between autocorrelation and band-pass filtering of subband envelopes is investigated in (Slaney, 1997).

Employing oriented features to detect dynamics in sound has also appeared in other models; Brown (1992) used oriented filters to build a map of frequency transitions in a time-frequency cochleogram, and Mellinger (1991) even employed three-dimensional oriented filters in a correlogram domain. However, because the fil-

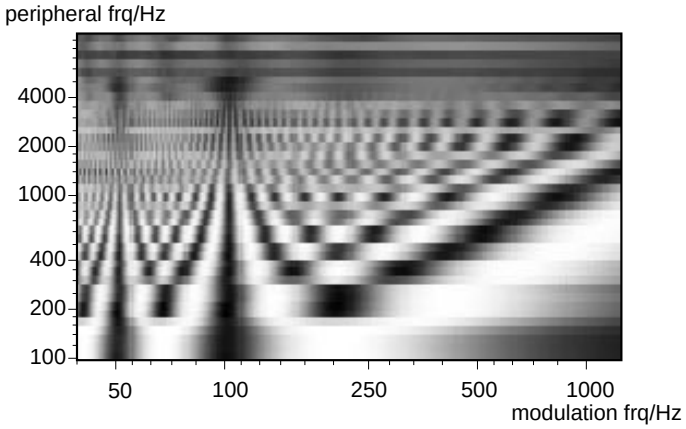

Figure 3: An example of a peripheral frequency/modulation frequency display, showing typical features for a periodic signal. Both axes are logarithmic. Because this example is autocrrelation-based, it displays subharmonics (e.g. the ridge at 50 Hz for this 100 Hz signal; a display based on band-pass filters would instead show modulation overtones or superharmonics, e.g. at 200 Hz).

ters in Todd & Lee are oriented sinusoidal grids, they have the interesting property, not seen in other models, of being band-pass tuned to a particular temporal frequency, which emerges as the ridges of the 3D grid cross the time axis. This third level of band pass tuning (after the cochlea and the periodic-modulation detectors) gives an elegant symmetry to the model.

There is, however, a negative aspect. The periodicity in time arises from the motion of a pattern which is periodic in space. But on the auditory frequency-period plane which has been substituted for the image x-y plane, it is difficult to imagine how such space-periodicity (expressed as cycles per octave along the carrier or modulation frequency axes) could be significant or useful. After defining the general form of their detectors in their eqn. (2) (which now has 9 degrees of freedom, i.e. orientation, locality and center in each of time, peripheral frequency, and modulation frequency), they describe four 'types' distinguished on the basis of whether they are truly band-pass or simply low-pass (ω=0) in either or both of the temporal axis and 'spatial' plane. Type I is low-pass in all dimensions and corresponds to the multiscale blurring described in Todd & Brown (1996) as well as the visual edge-detectors proposed by Marr (1982). Type II detectors are similarly low-pass in the 'space' plane, but have tunings to particular modulation rates in the time domain. It is types III and IV that are truly exotic, with oriented grids in the space domain that are either smoothed in time (type III) or that actually move in time to give a sensitivity that is periodic in both time and space (type IV, corresponding to the grid-motion detectors of Heeger, 1987).

Figure 3 illustrates the kind of output generated on a frequency-periodicity display (it is actually adapted from the autocorrelation-based system of Ellis (1996), but the differences are immaterial for this discussion – see Slaney, 1997). This snapshot is from a voiced segment of speech, and shows the typical patterns generated by such wideband periodic signals. It is in such a domain that Todd & Lee employ their oriented-grid detectors, and the question is whether convolving such a display with grid patterns of different periods and orientations will yeild any useful information. Although patterns are clearly present, they do not have a fixed period in any dimension: Along the modulation period dimension (horizontal axis), a given frequency channel will be periodic in the carrier frequency, but this turns into a skewed spacing on the log-modulation-frequency (linear-octave) display; similarly, the resolved lower harmonics appear as separate stripes in the lower half of the peripheral frequency (vertical) axis, but again compressed by the log scaling. Using logarithmic rather than linear axes is strongly supported by many physiological and psychological results, but even switching to linear axes would not much help the usefulness of fixed-period grids, since motion in the display (reflecting the slow source modulations that are of primary interest to this model) would appear as dilation rather than as the rigid translations assumed by the detectors (and present in the log-axis display).

Todd & Lee start with an interest in both multiscale smoothing and slow temporal modulation, and pursue the observation that both are special cases of their nine-dimensional general model. But their investigations of

other corners of that space, where the temporal periodicity gets rotated into combinations of the carrier and modulation frequency axes yeild less interesting variants. It is surely valuable to consider oriented filters as a mechanism for detecting frequency motion in correlogram-style displays, but this is more simply achieved through the two-dimensional oriented kernels of Mellinger (1991) and Brown (1992). Similarly, long-period band-pass filters may provide a useful characterization of slow modulation in speech and other signals (as well as explaining our acute sense of rhythm), but there is no apparent benefit in going to the computational expense of calculating them for a particular region of frequency-periodicity space rather than in a separate, subsequent stage (as in Scheirer, 1997, for example).

How to judge a representation?

Starting from the very wide range of behaviors that may be constructed from their general model, Todd & Lee go on to discuss how many different kinds of important signal features would be reflected by the detectors. This raises a slightly subtle point that might be overlooked in an illustrative discussion (but which would rapidly become quite clear in any kind of functional modeling), namely the distinction between *reflecting* a particular attribute and *separating* it from other confounding effects. Thus they suggest that grids oriented perpendicular to the carrier frequency = modulation frequency axis (their types IIId and IVd) will be the principle basis for detecting formants and their motion. While the appropriately-placed units will certainly detect energy associated with formants, the same is true for other grid orientations at the same spot, and, more seriously, these 'formant-detecting' units may also respond somewhat to aperiodic noise and even pure tones in the same frequency channels. The process of perception is concerned with recovering 'important' features from the sensory information; the location of formant peaks in voiced speech may be an important intermediary in this task. But as such, the representation that is useful is one that responds exclusively to particular features and not to other aspects of the signal. The vast majority of changes in an acoustic signal will be reflected in the overall energy of the signal (or perhaps on two dimensions as the energy in two overlapping frequency bands); however, because one change will look very much like many others in such a representation, it is not a particularly useful basis for signal interpretation.

It is, therefore, not enough to note that a particular attribute will influence the output of a particular detector; to be useful, the detector must transform that attribute in a way that facilitates subsequent processing, typically by disentangling its influence from that of other variations in the signal. In most cases, however, it is difficult to make this argument in the abstract, and the best demonstration of utility comes from the success of some functional model (a speech enhancement system for instance) that employs the representation in question. Although Todd & Lee show a number of suggestive illustrations of how their processing responds to several aspects of speech prosidy, we should not underestimate the challenge of translating such interesting images into information of benefit in a task such as speech recognition.

Brown & Wang: Neural Oscillator models

The second paper in the section also describes a computational model, but rather than modeling the kind of signal features that may be used in hearing, Brown & Wang address the subsequent problem of grouping together features into coherent sources perceived as present in the scene – the second box in our overview figure 1. This focus is complementary to the treatment in the Todd & Lee paper, which offers a variety of features and images suggesting their utility in scene organization, but no algorithms for actually constructing sources. More complete scene analysis systems (such as Weintraub, 1985; Brown, 1992; Ellis 1996) have addressed this source-formation problem, but usually through the tools of symbolic computer science e.g. lists and other data structures. Brown & Wang note that such algorithms bear no relation to our current understanding of the kinds of processing performed in the brain, and they therefore seek an alternative approach to source grouping founded in a more neurally-plausible architecture.

This they find in neural oscillator models, a structure that has been proposed in a variety of perceptual contexts as a way of using the time dimension to circumvent the fixed-size limitations we expect in neural circuits. If the brain has a single representational structure – say a cochleotopic array of neurons indicating energy present in different frequency bands – but needs to represent several different objects at the same time – say the different spectra of simultaneously-present sound sources – one solution is to have a rotating cycle

of time slices, with each of the different objects allocated its own slice during which its information is represented in the structure. Thus each element in the structure oscillates with the same period, but with a phase that is specific to the particular time-slice (and hence object) to which it belongs. Proponents of this model point to the extensive evidence of oscillatory activity in the brain from electromagnetic imaging studies, in particular correlations between so-called 40-Hz oscillations and perceptual phenomena.

von der Malsburg et al (1986) used a stylized auditory example to introduce their 'correlation theory', but emphasized that the general algorithm for refining and reinforcing weights linking groups of neurons could be employed in many different perceptual grouping domains. In this spirit, Brown & Wang have taken the same basic idea, but applied it the problem of separating pairs of simultaneous, static vowels distinguished by their fundamental period (f_0). They start with the Meddis & Hewitt (1991) model which employed an autocorrelation front-end to reproduce human identification scores very accurately, but whereas that model used a symbolic labeling scheme to choose the channels allocated to each vowel, Brown & Wang show how the same function can be achieved by a neural oscillator structure which is much more plausibly present in the brain. We will now examine how this structure actually reproduces the labelling in the Meddis & Hewitt model, then go on briefly to consider the relative advantages and disadvantages of 'emergent' neural models over conventional 'explicit' symbolic models.

Correlogram-based double-vowel separation

Common periodicity across frequency – perceived as pitch – is an extremely powerful cue to source formation and integration; an illustration of the benefits of periodicity differences to source separation is the experiment of Assmann & Summerfield (1990, following Scheffers, 1983), in which listeners had to identify both of a pair of simulataneous 'static' vowels (rigidly periodic signals whose spectra matched one of five canonical vowel formant patterns). Even when both vowels had the same pitch, listeners performed well above chance, identifying both vowels about 57% of the time. But as a pitch difference was introduced between the vowels, identification rapidly improved to over 70% at a frequency shift of just 6% (one semitone).

This simple demonstration the benefits of periodicity for auditory scene analysis has become a popular focus of computational models (although the assumption that it is truly 'static' and free of time-course effects has been challenged by recent studies showing the strong influence of slow 'beating' between lower resolved harmonics by Culling & Darwin, 1994). For simple vowel pairs, Meddis & Hewitt (1992) achieved an extremely close match to subjective results with a model that autocorrelated the 'neural firing probability' emerging from each cochlea frequency channel to give a time-varying function of frequency and autocorrelation period (lag) known as the *correlogram* (Slaney & Lyon, 1993).

Meddis & Hewitt's algorithm first summed across frequency channels to form a summary autocorrelation as a function of modulation period. The largest peak in this summary was taken as the dominant period, and assumed to correspond to one of the vowels. The frequency channels were then divided according to whether their *individual* autocorrelations reflected this period or not. Summary autocorrelations for all the matching channels were then matched against templates to give vowel identification responses.

Brown & Wang use exactly the same model in terms of the correlogram representation and the template matching, but replace the channel selection procedure with a continuously-running neural-oscillator separation network (although note that the input signal is considered static i.e. the correlogram display is not updated in the described system).

Neural oscillator grouping schemes that follow the 'correlation theory' of von der Malsburg (1986) have four essential pieces:

- A spatial array of units corresponding to some feature axis of the objects being represented
- Some input providing a raw indication that certain parts of the array belong together
- For each unit, an oscillator whose frequency is more or less fixed but whose phase relative to the other units may vary
- An algorithm for improving or maintaining the the phase-synchrony between elements that appear to belong together, and promoting a phase-difference between distinct groups of elements.

In Brown & Wang's system, the feature axis is cochleatopic frequency, the indication of grouping is a flag marking channels that have an autocorrelation peak at the current 'period of focus', and the oscillator update

algorithm modifies the oscillators' phase space to accelerate the silent half-cycle and prolong the active half-cycle of simultaneously-stimulated units, thereby 'pulling them in' to a synchronized firing phase. (Unstimulated oscillators experience the reverse influence via the 'global inhibitor', pushing other groups into a different phase).

Most interesting is the way in which the Brown & Wang system sequentially identifies and groups each periodicity detected in a signal. Like Meddis & Hewitt, they have a summary autocorrelation whose largest peak (indicating the periodicity dominant in the frequency channels being summarized) determines the 'period of focus' used to flag which channels 'belong' to this object by virtue of sharing this major peak. But where Meddis & Hewitt picked just a single period, and then used whatever was left to estimate the second vowel, Brown & Wang's system goes on to estimate *secondary* dominant pitches in the signal by removing the channels assigned to the first pitch from the summary autocorrelation. This subtract-and-reestimate approach has been used elsewhere (as described in Summerfield & Culling, 1995), but this neuronal implementation is novel and elegant: By gating each frequency channel's contribution to the summary autocorrelation through its oscillator phase (active or silent), the channels contributing the overall dominant period are all simultaneously removed when their (synchronized) oscillators finish their active phase and cycle into silence. During their silent phase, the summary autocorrelation examines only the channels not dominated by this period, and hence a second, weaker period may be identified and its channels selected. Depending on the ratio of active-to-silent cycle phase, the system may have the opportunity to detect still further weak periodicities in the signal when these secondary-period channels also become silent. Eventually, the oscillators in the primary dominant group will complete their silent phase, and, in becoming active, restore their channels to the summary autocorrelation. At this stage, channels identified as belonging to secondary periods will presumably be in *their* silent phases, meaning that the summary autocorrelation now consists of only the dominant-period channels, possibly giving a more accurate estimate of that period. Such an iterative remove-and-reestimate algorithm is quite common in signal separation systems, but this 'emergent' implementation, exploiting the inevitable silent phase of even the most strongly-supported oscillations, is surprising and very satisfying.

Emergent and explicit models of grouping

Elegance alone is not the only motivation for interest in such models, so in this section we consider the work by a variety of other criteria. Neural oscillator models highlight the broader dichotomy between such *emergent* models, which reproduce phenomena using deceptively simple (and often biologically-motivated) systems, and the alternative approach of constructing systems that address collections of phenomena in a deliberate and decoupled fashion – here termed *explicit* models. We will discuss the distinction between these approaches, and examine the arguments for oscillator models made by Brown & Wang.

The defining feature of emergent systems is their ability to exhibit a particularly wide range of behaviors in comparison to their complexity. This kind of optimized efficiency is also typical of the biological systems produced by evolution, and given that auditory modeling is concerned with duplicating such a living system it is very likely that the 'right' model, at some level, will be a relatively simple structure from which auditory function 'emerges'. Autocorrelation models of pitch perception show a degree of 'emergence' in that they explain a wide range of phenomena within a single simple structure, in contrast with older 'explicit' spectral-pattern-recognition models of pitch perception, which could often only be made to agree with curiosities like the pitch of inharmonic tones through complex modifications.

Neural oscillator models go one step further, however, by purporting to model certain aspects of auditory processing in terms of neuronal circuits. In the terms used by Marr (in his 1982 discussion of vision), they go below the *algorithm layer*, considered in previous models, to address the bottom, *implementation layer*. There are three main advantages to this kind of modeling:

- Compared to the symbolic, sequential processing of models that ignore neurons, they are at least plausible as descriptions of systems that might be found in the brain. This, of course, is Brown & Wang's main reason for investigating oscillators.
- Emergent models are parsimonious, since by definition they explain more with less. In particular, certain puzzling results may become obvious as natural but insignificant by-products (i.e. epiphenomena) of the true implementation; the pitch of inharmonic tones could be seen as an example. Brown &

Wang's model os certainly much simpler than complete symbolic separation systems (such as Brown, 1992 or Ellis, 1996), although the scope is far smaller. This particular model does not provide any novel explanations of epiphenomena.

- Neural models generally deal with lowest-common-denominator representations such as nerve firings and connections weights. As such, all kinds of information (from different signal cues etc.) are in an interchangeable format, suitable for combination and integration. Brown & Wang argue that since both grouping and memory occur in the cortex, models of cortical circuitry are the right approach to connecting them.

The majority of auditory models, however, have not sought sought this kind of 'emergence', leaving a plausible implementation as a problem secondary to the fundamental puzzle of explaining auditory function at all, irrespective of the computational substrate. Asking why other work in auditory modeling may have placed less emphasis on emergence leads to the following disadvantages of the approach:

- Since emergent models have behaviors that are in some respect surprising or unexpected, building such systems requires a particular combination of dilligence, inspiration and good fortune. Brown & Wang note that their model took many years of investigation and yet offers no functional advantage (as distinct from its plausibility advantage) over the well-established model upon which it is based. Symbolic systems have been the first resort of computational models, from which we may conclude that they are easier to design and build, allowing the researcher to address a wider domain of functional phenomena.
- Marr was critical of work that slavishly reproduced certain simple properties of neurons in the early visual systems without considering the overall significance or role of such behaviors i.e. 'mimicing' the implementation layer without specifying the algorithm or the overall 'computational theory'of perception. Models that succeed in duplicating the structure and behavior of neural systems could conceivably leave us little wiser about how the system actually works or why it is so constructed. If the goal of computational modeling is to discover processing principles that can be abstracted and extended in other applications, a literal model may not be the most useful kind.
- In comparison to models whose correspondence to the biological prototype is at a higher level of abstraction, systems claiming to model actual neurons bear the mixed blessing of being far easier to test and, perhaps, refute. The negative side of this is that a promising algorithm may be disregarded for the wrong reasons; neural oscillator models of visual binding enjoyed wide interest but are currently less in favor. It may turn out that the "40 Hz" brain-waves cited in support of the oscillator models have some largely unrelated explanation – which would in no way prove that the models were irrelevent, merely that they were not reflected by those measurements. Testability is a scientific virtue, but there may be specific risks inherent in attempting direct modeling of neural structures.

We conclude this section by considering some of the specific arguments made by Brown & Wang. They offer oscillator models as a "principled" foundation for auditory scene analysis and feature binding. While the models may be principled in the sense of being mathematically well-defined (as well as the specific proof of capacity for convergence in Terman & Wang, 1995), they have little to offer the problem of describing and understanding the underlying physical principles which the auditory system is exploiting to make scene analysis possible – the principles which would underly a Marrian 'computational theory'. Indeed, at the level of abstraction appropriate to a computational theory, the question of how the grouping between different spectral regions is represented or implemented is not even a consideration.

Brown & Wang also note that, excepting the work of Todd & Lee, which discusses grouping without proposing an algorithm, there are few if any credible alternative models for the neuronal implementation of grouping. This is true, but it is not hard to find cognitive functions for which no neural-level algorithm has been proposed or elucidated. Our lack of choice in neuronal models of grouping reflects our very rudimentary understanding of the processing involved and of brain function in general, rather than indicating that a particular solution has emerged as a clear winner.

Indeed, the scope of this model is still quite modest, and to expand such a system out to a full model of signal organization would probably still rely on symbolic-style algorithms. Recall that the primary advantage of the oscillator grouping was that the neurally implausible sets of channels and procedural algorithm of Meddis & Hewitt was replaced by a fixed-space, continuously-running neural circuit. Yet there still has to be something 'looking' at the output of oscillator display which notices, records and tracks that particular pattern of stim-

ulation. In the discussion of 'duplex' perception (in which they are really describing the commonplace experience of perceiving two sources as contributing to a single peripheral channel, as examined in Ellis, 1997), they show that their model's parameters can be adjusted to have each channel participate in more than one oscillation phase. But the interesting question is what happens next, how the brain uses a single representation of a given channel to participate in two percepts without confusing the properties of one with the other. In terms of the low-level, bottom-up properties, this problem may be solved by the phase difference between the two groupings. But how are top-down properties such as continuity and inference attached to just one of these phases without spilling over onto the other also? Such are the exacting problems incumbent upon modelers of neural circuits.

To return to the three levels of analysis proposed by Marr – implementation, algorithm and computational theory – the overall message is that analysis and understanding must proceed at all three levels. While modeling the precise firing pattern of a particular neuron in certain laboratory conditions may not tell us much about the purpose of the larger structure of which it is a part, it would be wilfull and foolish to ignore the increasingly detailed results coming from neurophysiology when trying to understand the auditory system. In particular, the ability of implementational models to explain curious epiphenomena which make no sense from any other perspective is ample reason to pursue such models. By the same token, however, the implausibility of symbolic, procedural algorithms should not be equated with irrelevance for non-neural models of sound organization; they remain the most powerful approach to investigating complex and abstract aspects of audition. As ever, the best scenario is for a full spectrum of approaches, since breakthroughs will occur in the least expected places.

Warren: The nature of high-level perception

Unlike the two previous papers, Warren is not directly concerned with building computational models, but instead sheds light on the overall process of auditory organization by investigating some surprising perceptual effects. As with any complex system, the behavior in response to extreme or anomolous inputs can be very revealing of the internal computation, and Warren's results provide some very telling indications and strong constraints to guide any theories of auditory processing we may construct.

Warren discusses two phenomena that are particularly relevant to speech perception. The first is the treatment of certain sequences of brief sounds as 'temporal compounds'. Sounds made by concatenating a sequence of static vowel sounds, each lasting just a few tens of milliseconds, can be reliably identified and discriminated by listeners who, however, are unable to specify the identity or order of the constituant vowels, but instead hear a sequence of syllable-like 'compounds' each corresponding to several of the vowel segments. We would perhaps expect that as the duration of each segment in a sequence was reduced there would come a point at which listeners would be unable to identify them, but this transition to syllable-like perceptual units seems to give a very strong indication of the nature of the mid-level representation being used to combine the signal-derived auditory information with the internal constraints of language. Listeners report hearing these kinds of looped sequences as repeated nonsense words, whose constituant syllables are however limited to those that actually occur in the listener's native tongue. Further evidence for a 'syllabary' as the intermediate representation of speech sounds comes from situations in which listeners hear two, simultaneous nonsense words for sounds which, presumably, could not be interpreted as a single word.

This last result has wider implications for the nature of speech-processing in the auditory system. Remez *et al.* (1994) have posit the existence of a 'speech module' which can process sound independently of general-purpose auditory organization. Such a module would be activated by certain signal characteristics that are distinctively indicative of speech. In support, they cite the ability of listeners to understand the words in "sine-wave speech" which consists of just three or four sinusoids copying the formant motions of a real utterance; by their reasoning, this highly distorted version of speech still contains the crucial 'speech characteristics' that permit it to be processed by the speech module. This account, however, does not explain the extremely context-sensitive perception of such stimuli, in which listeners typically hear nothing but a combination of whistles until they are 'primed' to listen for speech.

By contrast, the phonetic 'temporal compounds' do seem to suggest the existence of signal features that specifically invoke linguistic interpretation: the compounds do not correspond to real utterances, and in many

cases cannot even be heard as a single nonsense word, but rather than perceiving a single non-speech sound (perhaps the most 'accurate' perception), the auditory system constructs an explanation in terms of two combined nonsense words, or a word plus a nonspeech sound. It is as if there are aspects of the signal – the presence of format peaks, the regular pitch pulses – that make the auditory system determined to find a linguistic account, even if this can only be done as the seemingly unlikely combination of two sounds – rather than accepting the sound as nonspeech. This would appear as stronger evidence than the sine-wave speech phenomena for a speech-processing module invoked by lower-level signal features.

Warren's second discussion covers the many forms of auditory restoration, in which a missing portion of a sound (that can, hoever, be inferred from the context) will be perceived just as if it were present provided there is a masking signal able to obscure the missing sound. (The presence of the masking signal can make it immaterial whether or not the masked sound is 'missing', by rendering it mathematically indetectible in the mixture). As Warren notes, this kind of perceptual inference is a very important feature of any hearing system useful in our cluttered, noisy world, in which a great many of the sounds we care about will be obscured to some degree. The speech signal in particular contains much redundancy, which we can exploit as listeners only because we have the inference and restoration capabilities revealed in these experiments.

Warren gives details of three classes of phenomena that fall into this general category. *Temporal induction* applies to the classic 'phonemic restoration' experiments, in which a phoneme-sized segment of speech, removed and masked by a cough-like burst of noise, will be perceived as present to such an extent that the relative alignment of the cough and the speech can be judged only very crudely, even though the identity of the restored speech sound can be governed by semantic context provided several words *after* the obliteration. In this case, the information over a certain time window has been completely removed, but is restored on the basis of information from preceding and/or following times. *Contralateral induction* acts to perceive the speech signal from one ear as also present but masked in the other ear, thereby centralizing the perceived spatial location when the criterion of plausible masking is met. *Spectral induction* is suggested by experiments in which intelligibility was improved by adding band-pass masking noise between two frequency bands carrying highly reduced views of a speech signal – as if the additional noise allowed the perceptual system to infer the presence of a full-band, but masked, speech signal, rather than the more unnatural situation of a speech spectrum with a big silent hole in the middle.

The relationship between restoration phenomena and intelligibility is noteworthy. Restoration is primarily defined by a perception of continuity, in which the masked sound is perceived as in tact; this is not always associated with any improvement in the ability to determine which words were present i.e. the intelligibility of the speech. If we view the masked signal as complete speech to which some masker (e.g. a brief cough) has been added, then intelligibility has been somewhat reduced. If, however, we compare the speech-plus-masker to plain deletion (where the obliterated segment is left as silence), the addition of the noise energy, containing no information beyond the maximum time and amplitude extents of the unobserved sound, serves to improve the intelligibility of the speech. One way to understand this effect is in terms of the removal of a distraction: with a silent gap, the listener is confronted with an unusual and inexplicable situation, since real-world masking does not take this form. By contrast, filling the silence with a plausible masker renders the speech easier to understand, since although the signal contains essentially the same information, the confusing distraction of the silent gap has been eliminated, and the noise-masked signal is easily interpreted as the common real-world case of a transient masking sound. This point of view is suggested by Warren's results for interruption thresholds, in which a 60 ms *silent* gap was perceived as an interruption in normal continuous speech, continuous speech in which the semantics had been disrupted by reversing the word order, and isolated monosyllables; Alternating these speech types with *noise bursts* instead showed tolerances that were both 3-6 times longer than for silent gaps, and that varied with the speech material, being longest for the normal dialog which provided the greatest contextual basis for inference. It is as if silent gaps are detected by an independent non-speech-related gap detection mechanism, whereas noise bursts are only heard as interruptions when the lexical inference mechanisms fail to patch over them, an ability that varies in proportion to the speech information available.

Implications for Todd & Lee's front-end model

The most significant message of results of the kind described by Warren is to remind us that perception is enormously dependent on the powerful and specific constraints provided by high-level knowledge and ex-

pectations. In temporal restoration phenomena, listeners have the experience of hearing the deleted sound just as if it had been present, yet considering only the information in the front-end, none of the features of that sound could be disentangled from the masker. Todd & Lee propose a variety of front-end features, and discuss how they might contribute to a complete hearing system, but no model will be complete without accounting for the very powerful and abstract contextual inferences exhibited in restoration. Indeed, depending on the degree to which perception is truly 'guided hallucination' (Churchland et. al, 1994), inference and restoration may be the perceptual rule rather than the exception.

One notable detail of the temporal compound phenomena the requirement to listen to several seconds of a short, repeating stimilus before the verbal percept arises. This is a feature of a wide range of auditory phenomena (for example, the harmonic capture of Bregman & Pinker, 1978, or the looped noise phenomena of Guttman & Julesz, 1963), but has also attracted some suspicion as a highly unnatural kind of sound that may, therefore, provide limited insights into real-world hearing (Darwin, 1984). It does, however, tie in with Todd & Lee's idea of resonant structures with time-constants on the order of syllable durations and longer. One possibility is that such structures exist purely for their ability to map the dynamic features of nonrepeating transients into a spatial dimension more appropriate to slower brain processes, but this band-pass resonant nature gives rise to a number of epiphenomena such as peculiar sensitivities to periodically-repeating stimuli, and, perhaps, our sense of musical rhythm.

Implications for Brown & Wang's oscillator grouping model

Again, the abstract processing indicated by high-level perceptual phenomena demonstrates that relatively simple models of low-level effects can never provide a complete account of auditory perception. The model of Brown & Wang is specifically concerned with separating pairs of vowel sounds on the basis of their differences in fundamental frequency, yet the 'temporal compounds' described by Warren may be perceived as two simultaneous words (containing different vowels) despite containing but a single pitch; the perceptual fissioning into two voices results from the dynamic properties of the sound in conjunction with the top-down speech-specific constraints. (Although Brown & Wang's results show that their model is performs above 30% at identifying both vowels in a pair when the fundamental-frequency difference is zero, this probably reflects that five of the 15 possible unordered pairs have both vowels the same; it is open to question whether such cases should be regarded as identifying 'both' vowels).

In its favor, the neural oscillator approach holds the possibility of top-down influence on the connection weights between different units, along the lines of the 'associative recall' (Wang *et al.*, 1990) mentioned by Brown & Wang, or the vowel templates explicitly encoded into the oscillator networks of Liu *et al.* (1994). There is no obvious opening for this information in the particular model presented by Brown & Wang, but it could perhaps be included through variable weights between units that predispose them to group together, or though some kind of *a priori* weighting of the different frequency channels in their contribution to the summary autocorrelation or as they are considered for gating into the active group.

Summary and conclusions

The chapters in this section have demonstrated the very wide range of functions involved in auditory organization, and have touched upon surprisingly many of them. However, the overall message is that we must resist the natural tendency to underestimate the complexity of perceptual systems. Helmholtz unravelled a deep mystery when he discovered the relationship between the strength of partials and 'tone color', but, like him, we cannot assume that the revalation of some new and complex process indicates that we have gotten to the bottom of hearing. Each new level of understanding simply enables us to see a little further into the gloom.

Todd & Lee addressed the question of front-end transformations and representations, and although this is the best-understood part of the auditory system, their novel ideas for secondary and tertiary spectral-style analyses out to the timescales of syllables and phrases certainly show that there are plenty of avenues still to pursue in this area. Brown & Wang looked at the subsequent question of how features derived from the data can be formed into coherent perceptual groups, and in particular how such an abstract process may none-the-less be implemented in neural circuitry as a structure that could quite plausibly exist within the brain. Starting from the other end – the resulting behavior of human experimental subjects – Warren described investigations into

several surprising but illuminating phenomena concerning the perception of short, artificial 'temporal compounds' which listeners interpret as nonsense words, and the conditions under which a deleted portion of a longer sound, disguised by a suitable masker sound, is perceived just as if it were literally present. Warren shows that high-level phenomena of this kind ultimately impose strong constraints on the nature of the lower levels of processing at work in the auditory system, such as the way in which neural correlates of sound energy are accounted for and distributed between different perceived sources.

Despite this coverage, there are a number of very significant issues that have not been represented. Todd & Lee follow Pisoni & Nygaard (in identifying 'perceptual constancy' as a central problem for curent speech recognition technologies, yet the question of *what* it is that we count as 'the same' between two different people uttering the same word has not been answered; the complementary question of how we separate, represent and employ the information that allows us to distinguish between individual instances of a common class (be it speakers, pianists or vehicles) is similarly unknown. The entire area of learning, knowledge and memory, and its influences, over both the long and short terms, on perceptual organization are likely to make the aspects of grouping that we currently model appear ultimately as a rather trivial corner of the overall grouping problem – just as Helmholtz's static spectral envelopes really only scratch the surface of speech perception.

It is exciting, however, to see the rate at which new ideas and models are being developed. The enormous potential of computer modeling to help us understand complex information processing systems, coupled with the dual accelerators of rapidly increasing computational power and a wealth of new insights from neurophysiolical and psychoacoustical investigations, ensure that models and results of the kind we have discussed will continue to develop and multiply as we draw closer to an awed understanding of hearing.

References

Assmann, P.F. & Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88(2), 680-697.

Bregman, A.S. (1990). *Auditory Scene Analysis: the perceptual organization of sound*, MIT Press.

Bregman, A.S. & Pinker, S. (1978). "Auditory streaming and the building of timbre," Canadian Journal of Psychology 32, 19-31.

Brown, G.J. (1992). *Computational auditory scene analysis: A representational approach*, unpublished doctoral thesis (CS-92-22), Department of Computer Science, University of Sheffield.

Churchland, P, Ramachandran, V.S. & Sejnowski, T. (1994). "A critique of pure vision," in *Large scale neuronal theories of the brain*, C. Koch & J. Davis (eds.), MIT Press.

Cooke, M.P. (1991). *Modelling auditory processing and organisation*, doctoral thesis, published by Cambridge University Press, 1993.

Culling, J.F. & Darwin, C.J. (1994). "Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating," J. Acoust. Soc. Am. 95(3), 1559-1569.

Darwin, C.J. (1984). "Perceiving vowels in the presence of another sound: Constraints on formant perception," J. Acoust. Soc. Am. 76(6), 1636-1647.

Ellis, D.P.W. (1994). "A computer model of psychoacoustic grouping rules," Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem.

Ellis, D.P.W. (1996). *Prediction-driven computational auditory scene analysis*, unpublished doctoral dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.

Ellis, D.P.W. (1997). "The Weft: A representation for periodic sounds," Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Munich, 1307-1310.

Ellis, D.P.W. & Rosenthal, D.F. (1998). "Mid-level represenetations for sound: The Weft element," in *Readings in Computational Auditory Scene Analysis*, D. Rosenthal & H. Okuno (eds.), Lawrence Erlbaum.

Godsmark, D. & Brown, G.J. (1997). "Modelling the perceptual organization of polyphonic music," Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Nagoya.

Guttman, N. & Julesz, B. (1963), Lower limits of auditory periodicity analysis, J. Acoust. Soc. Am. 35, 610.

Heeger, D. (1987). "A model of visual image flow," J. Opt. Soc. Am. 4(8), 1454-1470.

Klassner, F. (1996). *Data reprocessing in signal understanding systems*, unpublished Ph.D. dissertation, Department of Computer Science, University of Massachusetts Amherst.

Licklider, J.C.R. (1951). "A duplex theory of pitch perception," Experientia 7, 128-133, reprinted in *Physiological Acoustics*, D. Schubert (ed.), Dowden, Hutchinson & Ross, Inc. (1979).

Liu, F., Yamaguchi, Y. & Shimizu, H. (1994). "Flexible vowel recognition by the generation of dynamic coherence in oscillator neural networks: speaker-independent vowel recognition," Biological Cybernetics, 71, 105-114.

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

Marr, D. (1982). *Vision*, W.H. Freeman.

Meddis, R. & Hewitt, M.J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. 89(6), 2866-2882.

Meddis, R. & Hewitt, M.J. (1992), Modelling the identification of concurrent vowels with different fundamental frequencies, J. Acoust. Soc. Am. 91(1), 233-245.

Mellinger, D.K. (1991). *Event formation and separation in musical sound*, unpublished doctoral dissertation, Department of Music, Stanford University.

Nakatani, T., Okuno, H.G., Goto, M. & Ito, T. (1998). "Multiagent based binaural sound stream segregation," in *Readings in Computational Auditory Scene Analysis*, D. Rosenthal & H. Okuno (eds.), Lawrence Erlbaum.

Nygaard, L. & Pisoni, D. (1995). "Speech perception: New directions in research and theory," in *Speech, language and communication: Handbook of perception and cognition*, J. Millar & P. Eimas (eds.), 2nd ed., 63-96.

Remez, R.E., Rubin, P.E., Berns, S.M., Pardo, J.S. & Lang, J.M. (1994). "On the perceptual organization of speech," Psychological Review 101(1), 129-156.

Scheffers, M.T.M. (1983). *Sifting vowels: auditory pitch analysis and sound segregation*, unpublished doctoral thesis, University of Groningen.

Scheirer, E.D. (1998). "Tempo and beat analysis of acoustic musical signals," J. Acoust. Soc. Am. 103(1) , 588-601.

Slaney, M. (1997). "Connecting Correlograms to Neurophysiology and Psychoacoustics," Proc. XIth International Symposium on Hearing, Grantham, UK.

Slaney, M. & Lyon, R. F. (1993). "On the importance of time--a temporal representation of sound," in *Visual Representations of Speech Signals*, M. Cooke, S. Beet & M. Crawford (eds.), Wiley.

Summerfield, Q. & Culling, J.F. (1995). "Auditory computations which separate speech from competing sounds: a comparison of binaural and monaural processes," in *Fundamentals of speech synthesis and speech recognition*, E. Keller (ed.), J. Wiley.

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

Todd, N.P.M. & Brown, G.J. (1996). "Visualization of rhythm, time and metre," AI Review 10, 253-273.

Wang, D.L., Buhmann, J. & von der Malsburg, C. (1990). "Pattern segmentation in associative memory," Neural Computation 2, 94-106.