INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

ICSI Technical Report TR-98-016

THE AUDITORY ORGANIZATION OF SPEECH IN LISTENERS AND MACHINES

Martin P. Cooke
Department of Computer Science
University of Sheffield
<m.cooke@dcs.shef.ac.uk>

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

1. Introduction

Speech is typically perceived against a background of other sounds. The acoustic mixture reaching the ears is processed to enable constituent sources to be heard and recognized as distinct entities. The auditory system may not always succeed in this goal, but the range of situations in which spoken communication is possible in the presence of competing sources highlights the flexibility and robustness of human speech perception. The background against which a conversation is carried out is made up of acoustic intrusions which may overlap temporally and spectrally with the target speech. The background may consist of other utterances, with fundamental frequency and formant contours occupying similar regions to those of the target. Target and background may contain similar ranges of envelope modulations, and can arrive from similar locations in space. Sometimes, the background will be characterized by high-intensity onsets which completely mask the target conversation, albeit temporarily. Figure 1 depicts a mixture of two digit sequences whose constituents differ in onset time, fundamental frequency contour and formant structure but which are nevertheless sufficiently similar in these properties as to make (visual) separation and identification difficult.

Robust automatic speech recognition (ASR) remains an important unsolved engineering problem. For example, Lippmann (1997) has compared error rates obtained by listeners and machines, finding that while automatic speech recognition systems suffer an order of magnitude more errors than listeners for clean speech, the margin widens to two orders of magnitude for noisy speech. An appreciation of the mechanisms employed by listeners to select a target conversation in a noisy background could lead to progress in robust ASR.

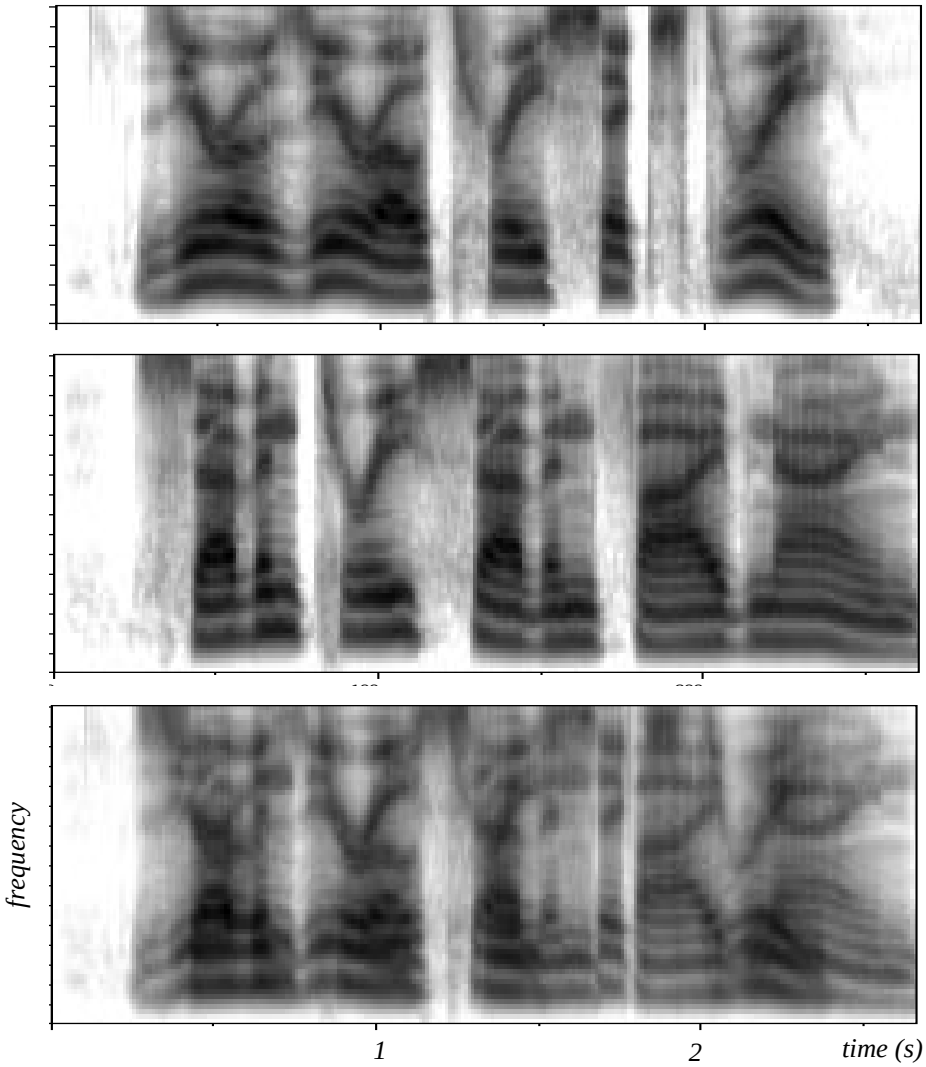

Figure 1: Auditory spectrograms of spoken digit sequences.Upper: "zero zero three six three". Middle: "seven three seven five nine". Lower: auditory spectrogram of the mixed signal. Grey-levels are proportional to log-energies at the output of a bank of 64 gammatone filters, equally spaced on an auditory scale (ERB-rate) from 50 to 6500 Hz.

Some of our ability to handle complex sound environments arises from familiarity with the patterns of spoken language. These regularities manifest themselves at a number of levels, from the sub-syllabic to the sentence and above. Speech represents a rich and redundant encoding of information, and it is not at all surprising that knowledge gained from prior experience can help to fill in those parts of the signal that are masked or otherwise distorted. Such top-down processes have been termed *schema-driven* mechanisms (Bregman, 1990).

Apart from familiarity with speech, there are other mechanisms which have the potential to explain how speech can be perceived in a background of other sounds. Sound sources may differ in location, or in instantaneous fundamental frequency, or in the patterns of energy envelope modulation in different frequency bands. If it is possible to reliably extract these 'cues' sufficiently often, and to *group* those parts of the mixture possessing similar values of each property, then listeners have the basis for organizing into a coherent whole those components which have a common origin. Such

mechanisms would complement those based on prior knowledge of spoken language. They are often described as *bottom-up* or *primitive* processes.

The purpose of this chapter is to describe the evidence for auditory organization in listeners and to explore the computational models which have been motivated by such evidence. The primary focus is on speech rather than on sources such as polyphonic music or nonspeech ambient backgrounds, although these other domains may be equally amenable to auditory organization.

The remainder of this section introduces some of the terminology of auditory grouping, summarizes those features of the stimulus currently thought to promote grouping, and provides a chronological perspective on the development of the field.

1.A A perspective on auditory organization

Bregman (1990) draws a distinction between the concrete, physical manifestation of a sound wave and the abstract, conceptual effect it has in the mind of the listener. At the concrete level, sound is generated by physical processes. A physical system which is regarded as a single sound source may be termed an *acoustic source*. Sound waves reaching the eardrum of a listener will include the output of many physical processes mixed together and colored by the acoustic properties of the environment – which might be a mountainside or a concert hall.

On entering the ear, the signal undergoes several stages of transduction, leaving the periphery as patterns of nerve-firings which may be considered as *representations* of all or part of the sound. Features of these representations which are used to achieve a particular end (such as organizing the sound) are called *cues*. Different theories for the organization of sound may have varying assumptions of which features are actually employed as cues.

The function of auditory organization is find an account for the acoustic mixture as a collection of independent sources, which are the abstract mental constructions such as “Jim” or “car” that we use to relate to the world. Although these sources are usually identified with distinct physical systems, such auditory identification is not always achieved or desirable: the sound coming from a stereo system is frequently perceived as the original recorded instruments but not as the pair of vibrating speaker cones that are the immediate physical source.

Representations

In addition to the terminological distinctions introduced above, it will be useful to clarify some issues relating to the level of description at which auditory phenomena and associated models are phrased. This perspective is influenced by the work of Marr (1982), who argued that vision research was torn between detailed descriptions of the neural circuitry involved, and more abstract accounts of the effective function of those circuits. Marr distinguished among computational theory, algorithm and implementation. The first term corresponds to the basic physical principles underlying a perceptual task which allow it to be solved. The algorithmic level identifies one or more possible constructive approaches to the solution. Details of the computational hardware used to implement the algorithm are distinguished from the earlier levels of description.

The relevance of a representational perspective to the computational problem posed by hearing has been noted by a number of researchers (Green & Grace, 1981; Darwin, 1984; Green & Wood, 1986;

Green *et al.*, 1990; Cooke, 1991/1993; Brown, 1992; Ellis, 1996). Our current synthesis is contained in Table 1 below.

Table 1: Representational perspective on auditory organization

level	problem	solution possibilities
Computational theory	Sound source organization	Employ characteristics that define distinct sources: independence, continuity and source features (periodicity, spatial location etc.).
Algorithm	Auditory grouping	Decompose acoustic signal across time and frequency; reassemble into complete sources on the basis of grouping principles via particular cues.
Implementation	Feature calculation & feature binding	Calculation of individual features in auditory maps; combination of features represented neurally.

Most previous work in auditory organization has adopted the descriptive vocabulary of the algorithmic level in Table 1, namely, the nature of grouping and the cues that promote it in listeners. This level is discussed further after examining the computational theory level below. Models at the implementation layer are a relatively recent development and are covered in section 6.

Computational theory

The notion of an underlying computational theory emphasizes that any perceptual process must be based upon reliable characteristics of the physical world, which may be exploited to obtain information of value to the organism. The intimate dependence between perceptual processing and the specific characteristics of the environment, regardless of any idealizations or particular mechanisms, was forcefully argued by Gibson (1966), who referred to this perspective as "ecological perception." The ecological constraints in sound are so basic that they can escape our notice, but their central role in perceptual organization must be recognized. The most important such constraints employed by the auditory system are the *independence* and *continuity* of sources.

Independence refers to the observation that changes in the properties of one source in a mixture will be largely independent of changes in the others. Although this appears to be an observation about the properties of sources, it can also be viewed as one of the best bases we have for *defining* a source, i.e. as a physical system whose acoustic emissions are highly coherent and correlated in a way that listeners can immediately apprehend. Thus a car, despite consisting of many different acoustic processes, can be perceived as a single source because of a certain correlation between all these sounds (arising from the mechanical coupling, and the motion of the car in the environment) rather than because of any more profound physical relationship, albeit that the correlation generally arises because of such relationships. When this independence of distinct sources becomes blurred – as in ensemble music performance – the perceptual organization becomes unusual and ambiguous, which may be one of music's peculiar attractions.

Continuity conveys the idea that properties of a given source will tend to change smoothly, and will not undergo an abrupt change to a completely different sound. While certain properties may change abruptly (consider for instance the effect on the speech spectrum of a major articulator motion such as the parting of the lips), others, such as voicing, will be piecewise-continuous (i.e. they will not

exhibit jumps during the episodes when they are present). The converse of this constraint is that if all source properties change abruptly, the sound on either side of the change is likely to be heard as two distinct sources.

Independence and continuity are so basic in their role of defining sources that there is little point in trying to identify which aspects of auditory processing reflect those particular constraints; rather, they underly the entire function. For this reason, we propose that they serve as elements of the "computational theory" of hearing. By contrast, the other cues used in auditory organization, such as pitch and interaural parameters, are related to specific feature calculations within the hearing system. We regard them as belonging to Marr's "algorithm" layer, to which we now turn.

Auditory grouping

The auditory system represents just one approach to getting information out of sound, albeit one that an artificial computational device would be hard-pressed to better. Clues to the functional partitioning of auditory processing have been obtained by both physiological and psychological experimentation, permitting some inferences concerning the algorithmic layer.

Early auditory signal processing involves at least two forms of decomposition. First, the signal is subject to a spectral decomposition in the cochlea – an organizational axis maintained throughout many later processing stages. Second, it appears that different properties are extracted in distinct auditory maps (Moore, 1987). Consequently, information arising from a single acoustic source finds itself distributed both across cochleotopic frequency and between several auditory nuclei. For instance, a voiced speech sound gives rise to a series of harmonically-related peaks in the low-frequency portion of an excitation pattern. The higher frequencies might contain envelope modulations at the voicing fundamental frequency (f_0) as reflected in the full-band temporal envelope (or equivalently caused by the interaction of neighboring harmonics in the response area of the auditory filter). The fine time response at the output of each such filter would also contain periodicities related to the fundamental and its harmonics. Moore (1997, fig 5.6) depicts some of these properties of the auditory filterbank response to periodic sounds. It is possible that further processing of harmonic peaks, envelope and fine structure is carried out in distinct auditory maps.

This two-fold separation (by frequency channel and cue class) is understandable: since different sources in an acoustic mixture may dominate distinct spectral regions, spectral decomposition is an elementary first step in signal separation. Functional decomposition – processing in distinct auditory maps – allows the deployment of relevant processing hardware to extract different signal properties such as f_0 and location, including the possibility of using several complementary processing approaches for each of these properties.

Given this fragmentation of the original sound waveform into several features defined over multiple dimensions, the problem of deducing a description of a particular event in the physical world is now dominated by the question of which portions of this distributed representation belong together as relating to that event. Referring to Table 1, the top-level problem of "source organization" becomes the algorithmic/representational issue of "auditory grouping."

In describing the different forms of grouping that arise, it is tempting to make short-hand statements such as "sound components with a common pitch are grouped together." However, the possibility of multiple mechanisms for sound organization based on even a single factor such as f_0 demands a more precise discussion. In fact, it is possible to distinguish between at least three types of grouping:

- grouping of local features within auditory maps, that is the assembly of locally-consistent regions of the maps that presumably reflect a single source,

Table 2: Summary of grouping cues

Source property		Potential grouping cue	Illustrations	Notes
Starts & ends of events (common onset/offset)		Synchrony of transients across frequency regions	Effect of onset asynchrony on syllable identification (Darwin, 1981) and pitch perception (Darwin & Ciocca, 1992)	Onsets and offsets can also be considered as slow amplitude modulations. Offset generally weaker than onset.
Temporal modulations	slow	Correlation among envelopes in different frequency channels	Comodulation masking release (Hall *et al.*, 1984)	Common frequency modulation may lead to common amplitude modulation as energy shifts channels (Saberi & Hafter, 1995)
	fast, periodic	Channel envelopes with periodicity at f_0 (unresolved harmonics)	Segregation of two-tone complex by AM phase difference (Bregman *et al.*, 1985)	
		Harmonically-related peaks in the spectrum (resolved harmonics)	Mistuning of resolved harmonics (Moore *et al.*, 1985); effect on phonetic category (Darwin & Gardner, 1986)	
		Periodicity in fine structure (resolved & unresolved harmonics)	Perception of ‘double vowels’ (Scheffers, 1983, etc.)	Basis for autocorrelation models (Patterson, 1987; Meddis & Hewitt, 1991)
Spatial location		Interaural time difference due to differing source-to-pinna path lengths	Vowel identification (Hukin & Darwin, 1995). Strongest effect if direction is previously cued.	Evidence that suggests role of ITD is limited (Shackleton & Meddis, 1992) or absent (Culling & Summerfield, 1995b)
		Interaural level difference due to head shadowing	Noise-band vowel identification (Culling & Summerfield, 1995b)	
		Monaural spectral cues due to pinna interaction	Localization in the sagittal plane (Zakarauskas & Cynader, 1993)	Has not been investigated for complex, dynamic signals such as speech.
Event sequences		Across-time similarity of whole-event attributes such as pitch, timbre etc.	Sequential grouping of tones (Bregman & Campbell, 1971); sequential cueing (Darwin *et al.*, 1989, 1995)	
		Long-interval periodicity	Perception of rhythm	By-product of very-low-frequency ‘spectral’ analysis (e.g. Todd 1996)?
Source-specific		Conformance to learned patterns	Sine-wave speech (Remez *et al.*, 1981)	

- grouping of features corresponding to the same source represented in different maps, such as a pitched source whose low and high harmonics may be grouped in separate maps by spatial pattern and temporal structure respectively, and

- grouping based on the acquired expectations of prior knowledge ("schema-driven" grouping) as distinct from "primitive grouping" involved in earlier processing stages (Bregman, 1990).

Having situated the concept of 'auditory grouping' within the entire perceptual problem, the next section summarizes current understanding of the kinds of grouping at work in the hearing system.

1.B Summary of grouping cues

Table 2 is an attempt to summarize the many experimental investigations of grouping using the framework expressed above. The organization of the table reflects the idea that each property of an acoustic source produces a number of auditory consequences, each of which represents a potential grouping cue. Darwin & Carlyon (1995) provide a quantitative tabulation of some of these investigations and demonstrate that grouping is not "all-or-nothing", but occurs at different degrees of feature prominence depending on the measure used.

Having numerous cues for sound organization respects the fact that any one of them may fail to indicate the correct grouping, but it simultaneously presents higher auditory levels with the possibility of inconsistent or conflicting cues. Investigations of conflicts such as frequency proximity vs. ear of presentation (Deutsch, 1975) or onset asynchrony and mistuning (Darwin & Ciocca, 1992; Ciocca & Darwin, 1993) can provide valuable insight into high-level audition.

Some signal features have been proposed as potential grouping cues but do not appear in Table 2. Foremost amongst these is the common frequency modulation imposed on harmonics in voiced speech. There is little evidence for an independent effect grouping by common FM over and above that provided by instantaneous harmonicity (Gardner & Darwin, 1986; Summerfield & Culling, 1992), although the presence of FM can make vowels more prominent against a background of unmodulated sounds (McAdams, 1984).

This introductory section concludes with a chronological review of developments in the field of auditory organization in listeners and machines. Many of these results will be discussed in more detail in sections 2 through 5.

1.C Historical overview

Listeners

One of the earliest accounts of the problem faced by listeners when presented with simultaneous utterances was described by Cherry (1953). Considering the task he termed the "cocktail party problem," he speculated on the possible cues to its solution – location, lip-reading, mean pitch differences, different speeds, male/female speaking voice, accents etc. Cherry highlighted the relative ease with which one of a pair of simultaneous sentences could be repeated when the messages were sent to different ears. In a refinement of this strategy, Broadbent & Ladefoged (1957) employed synthetic, two-formant speech to examine the roles of both ear of presentation and fundamental frequency on perceptual fusion, as reflected by the number of voices heard by listeners. They found that fusion occurred even when the two formants were sent to different ears, but that giving the two formants sufficiently different fundamental frequencies prevented fusion. Their findings not only demonstrated a clear role for fundamental frequency differences in perceptual

organization, but were an early anticipation of the interactions that occur when multiple cues for grouping are placed in opposition which each other, a recurrent theme in studies of grouping and segregation. Broadbent & Ladefoged were amongst the first authors to recognize the computational problem posed by hearing, noting that perception in the presence of other sounds represents the normal, everyday mode for spoken language processing.

A different approach to the study of speech perception in such everyday acoustic backgrounds came with the finding by Warren (1970) that listeners were unaware of the absence of short segments of sentences which had been replaced by a louder noise. This phenomenon was termed the *phonemic restoration effect*. Later work (Warren *et al*., 1972) generalized its application to non-speech signals and phonemic restoration is now considered as a special instance of a collection of "auditory induction" effects, including induction between ears and across frequencies. Auditory induction appears to reflect a desire for coherent explanations in sensory processing.

Warren's work was an important demonstration that the auditory system was not simply a passive conduit for sensory information, but was engaged in an active interpretation of the signal, with illusory percepts as a side-effect. Bregman & Campbell (1971) showed that, dependent upon stimulus parameters such as frequency separation and repetition time, an alternating sequence of high and low frequency tones would be perceived as a single sound source alternating between high and low frequencies (the veridical percept) or as two sources, consisting of repeated high tones and low tones respectively (the illusory percept). Bregman referred to these percepts as "auditory streams" to distinguish them from more objective physical entities, and the rules governing this description-forming process have been extensively investigated by Bregman and his colleagues since the early 1970s.

Much of this early work on streaming employed simple tonal stimuli, although some studies used speech-like sounds and demonstrated similar effects of factors such as spectral dissimilarity on streaming in a temporal order identification task (Cole & Scott, 1973) and pitch and formant continuity on speech coherence (Darwin & Bethell-Fox, 1977). These studies used repeated sequences to induce segregation, which raises questions over whether the grouping cues uncovered in such experiments can be usefully employed in everyday speech perception. Darwin's (1981) attempt to find evidence for grouping in speech was a turning point. Darwin used single presentations of synthetic vowels and consonant-vowel (CV) syllables in which formants differed in either onset times or f_0. Earlier, Cutting (1976) has shown that listeners were able to identify syllables whose formant resonances had been divided between ears: The lowest, first formant (F1) was presented to one ear; the other ear received the higher formants (F2 and F3) but with a different fundamental. Darwin failed to find an effect of onset asynchrony or difference in f_0 on phonetic category except in one condition in which grouping could result in two equally-plausible syllables. Here, a synthetic four-formant syllable was constructed which would be perceived as /ru/ if all formants were played together, or as /li/ if F2 were omitted. This innovative paradigm enabled Darwin to manipulate f_0 and relative onset times of the second formant (F2), and to demonstrate an effect of perceptual organization on phonetic categorization.

The conclusion of Cutting (1976) and Darwin (1981) that phonetic interpretations could easily override conflicting cues for perceptual organization led to the realization that explorations of grouping need to be performed in a phonetically-neutral context. Over the next few years, a series of refinements and new paradigms enabled a much closer analysis of the role of perceptual grouping in speech, with the spotlight on the identification of synthetic stationary vowels. Darwin (1984) exploited the fact that a vowel continuum from [ɪ] to [ɛ] could be constructed by varying F1 between 375 Hz and 500 Hz to provide a sensitive indicator of whether tones at harmonics close to F1 were perceptually integrated into the vowel under various conditions. These experiments demonstrated that onset or offset asynchrony could reduce the contribution that a harmonic makes to vowel quality. Darwin & Gardner (1986) employed a harmonic mistuning paradigm (Moore *et al.*, 1985)

and the [ɪ]-[ɛ] continuum to show that, just as a mistuned component could be excluded from computation of pitch, it could similarly contribute less to vowel quality.

An alternative approach to the study of grouping in speech was introduced by Scheffers (1983). He asked listeners to identify both constituents of pairs of concurrent synthetic vowels. This double vowel task, as it came to be known, has proved to be a fertile paradigm for the study of auditory perceptual organization. One early finding was that a difference in fundamental frequency between the constituent vowels leads to a significant improvement in identification scores. Subsequent experimental and modeling studies have resulted in several quite distinct explanations for this effect.

The links between perceptual organization in audition and other modalities, such as vision, were made explicit by Bregman (1984), who coined the term "auditory scene analysis" to describe the goal of processes attempting to form coherent explanations of the external sound field. Darwin (1984) also drew an analogy with Marr's (1982) work in vision, pointing out the distinction between low-level properties, directly evident in the waveform, which are used to assign features to different sound sources, and those more abstract properties which should be allowed in contact with phonetic categories.

By 1990, a significant body of perceptual studies of auditory fusion and segregation had accumulated, consolidated by Bregman's (1990) comprehensive monograph. Many properties of sound sources considered as potential features for organization had been investigated. One finding has been the failure of grouping under circumstances which might otherwise have been thought to promote it. For example, changes in f_0 lead to correlated changes in harmonic frequencies, known as common frequency modulation (FM). In investigating whether common FM causes perceptual fusion, it is necessary to rule out cues based on instantaneous mistuning caused by FM phase differences, and the detection of incoherent FM. Gardner *et al*. (1989), using the /ru/-/li/ paradigm, found no effect of incoherent FM in segregating F2 from the remainder of the syllable.

Other recent trends in the study of auditory perceptual organization include:

- explorations of the relationship between grouping and other phenomena such as comodulation masking release (Hall & Grose, 1990; Grose & Hall, 1992, 1993), modulation detection interference (Yost & Sheft, 1989; Hall & Grose, 1991; Moore & Shailer, 1992), binaural interference (Stellmack & Dye, 1993) and informational masking (Kidd *et al.*, 1994, 1995).

- investigations of the relationship of grouping to other aspects of auditory function, such as the determination of pitch, location or phonetic quality of a sound source. A careful quantitative analysis of this task-dependent influence of grouping is provided in Darwin & Carlyon (1995), who document the way in which the size of the cue manipulation required to reveal an effect varies according to the task involved. Thus, for the tasks of detection, identification as a separate source, determination of pitch, vowel classification, speech separation, and literalization, the degree of mistuning required of a single harmonic varies from 1% to 10%. Similarly, the amount of onset or offset asynchrony required in a similar range of tasks can vary from a few milliseconds for detection to several hundreds of milliseconds for tasks involving pitch and vowel identification.

- developmental studies, for example the examination of perception of inharmonicity in infants (Clarkson & Clifton, 1995; Clarkson & Rogers, 1995), showing a decrease of pitch salience with inharmonicity similar to that of adults.

Models

One of the earliest computational attempts at speech separation was the signal-processing approach of Parsons (1976). Although Parson was not motivated by auditory findings, his system served to define – and partially solve – some of the issues which have since become central for computational auditory scene analysis (CASA) systems operating on voiced speech, namely the resolution of overlapped harmonics, the determination of multiple pitches, and the tracking of fundamental frequency contours which may cross. Parsons described the separation of voiced speech as the "principal subproblem", and his system set about solving it by identifying two sets of harmonic peaks in a standard fixed-bandwidth Fourier-transform spectrum, estimating their pitches and tracking their evolution through time.

In 1983, Scheffers described an algorithm for concurrent vowel separation to explain data from his perceptual tests in which listeners demonstrated an improvement in their identification of both vowels as the fundamental frequency difference between them increased (Scheffers, 1983). The concurrent vowel task has since become an important proving-ground for ASA, and a number of models of listeners' performance in this domain have been proposed since Scheffers' pioneering studies (Assmann & Summerfield, 1990; Meddis & Hewitt, 1991, 1992; Lea, 1992; de Cheveigné, 1993; Brown & Cooke, 1994; Culling & Darwin, 1994; Berthommier & Meyer, 1997; Varin & Berthommier, 1997; Brown & Wang, 1997).

In the same year, Lyon (1983) – influenced by Jeffress' (1948) proposal for an interaural delay line mechanism – presented a computational model of binaural localization and separation which performed a cross-correlation of the outputs of cochlear simulations for opposing ears. Lyon used the term "correlagram" to describe the cross-correlation representation (the term "correlOgram" has since come to refer primarily to an *auto*correlation analysis) and demonstrated separation of a short speech signal from an impulsive sound generated by striking a ping-pong ball.

Weintraub (1985) was the first to design a system with an explicit auditory motivation to tackle the more difficult problem of sentence separation. His pitch-based separation system was inspired by the neural autocoincidence model of Licklider (1951).

These early demonstrations illustrated the engineering potential of cues such as pitch and interaural differences, but they did not provide quantitative measures of algorithm performance. One of the first studies to do so was the evaluation by Stubbs & Summerfield (1988) of two algorithms for the separation of voices based on a difference in fundamental frequency in a single channel. One approach operated by attenuating the pitch peak corresponding to the interfering voice through filtering the cepstrum of the mixed signal. The other was similar to Parsons' (1976) harmonic selection scheme. By resynthesizing the target voice, possible speech enhancement benefits of these approaches could be evaluated. Stubbs & Summerfield used synthetic vowel pairs in one task and CV words masked by synthetic vowels in another to show that the enhanced speech was more intelligible to listeners with normal hearing and with hearing impairments.

The decade since Weintraub's system have witnessed a proliferation of modeling attempts. Cooke (1991/1993) described a system for computational auditory scene analysis which operated by seeking organization amongst time-frequency tracks representing the evolution of spectral dominances in the outputs of a model of the auditory periphery. His approach employed grouping rules based on principles such as harmonicity, common amplitude modulation, common fundamental and frequency proximity. Mellinger's system (1991), while concerned with modeling the auditory organization of musical rather than speech sources, looked for groups with onset synchrony and common frequency modulation. Brown (1992) used a collection of computational maps as a substrate for automatic grouping. These maps were designed to represent possible

tonotopically-organized computation in the auditory brainstem and cortical regions, and included maps describing onsets, offsets, local periodicities and frequency movement.

The fact that these systems employed multiple grouping cues led to the issue of how best to combine evidence from factors such as pitch and onset synchrony. Kashino & Tanaka (1993) attempted to handle such integration using Dempster's combination law; later work (Kashino *et al*., 1998) has exploited Bayesian networks. A popular alternative (Cooke *et al*., 1993; Ellis, 1996; Klassner, 1996; Godsmark & Brown, 1997) has been the use of blackboard architectures (Carver & Lesser, 1992; Nii, 1986), which consist of independently-defined *knowledge sources* or *experts* (e.g. a tracking expert, or a harmonic grouping expert), cooperating via a global data structure (the blackboard).

More recently, concern has focussed on the role of schemas in auditory scene interpretation. Blackboard systems also cater well for the interaction of stored knowledge or *expectations* about sound sources with evidence arriving from low-level grouping processes. Ellis (1996) implemented an expectation-driven system for CASA whose goal reflects the fundamental purpose of auditory scene analysis – to reconcile the evidence grouped out of arbitrary sound mixtures with the predictions made by prior knowledge about sound sources.

A potential attraction of computational auditory scene analysis is the prospect of an approach to robust automatic speech recognition. Weintraub's system attempted to recognize separated speech, but the difficulties of interfacing CASA and ASR have only recently received closer attention. Cooke *et al*. (1994) proposed the development of missing data theory to allow ASR systems to attempt recognition on the basis of partial evidence, as an alternative to the recognition of CASA-enhanced speech by an unmodified recognizer. Ellis (1997) envisaged a speech "knowledge source" to be used in conjunction with nonspeech elements within a single prediction-reconciliation framework to construct complete explanations for sound mixtures.

Chapter organization

The chapter has two levels of organization. The sequencing of sections 2 to 5 reflects a systematic progression from lower to higher levels of stimulus complexity. Section 2 deals primarily with simple tonal configurations used to demonstrate the streaming effect. Section 3 examines some of the extensive experimental and modeling work which has employed simultaneous synthetic vowels ('double vowels'). Section 4 considers the processing of natural utterances using only primitive grouping cues. Section 5 describes the additional role played by schemas for instance in understanding speech in noise. The organization of material within each section reflects the title of the chapter: first, relevant perceptual evidence for organization in listeners is considered, followed by details of algorithms which attempt to replicate the effects in machines. Some of these models are motivated by neurophysiology, others by the desire to match listeners' data in perceptual tests, while some are predominantly inspired by the notion that a good engineering solution to the problems posed by hearing can be obtained by following, at some level of abstraction, the laws of auditory organization. After surveying these functional models, section 6 looks briefly at models of the underlying neural machinery. The chapter concludes with a discussion of the major issues facing computational auditory scene analysis.

2. The streaming effect

2.A Listeners

A sequence of alternating high and low frequency tones can result in the perception of either one or two coherent patterns or *streams* (Miller & Heise, 1950; Bregman & Campbell, 1971). Factors

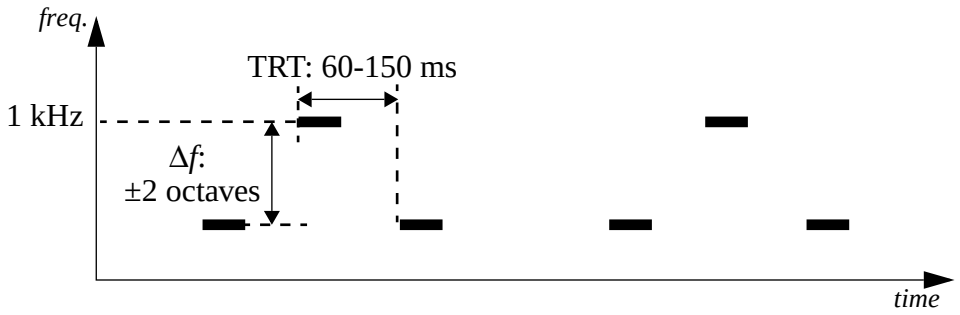

Figure 2: Stimulus configuration for the streaming experiments of van Noorden (1975). The sequences of alternating sinusoidal signals are presented with differing frequency separations (Δf) between the tones and differing overall repetition periods (TRT).

influencing segregation into streams are discussed at length in Bregman (1990, chapter 2) and summarized below:

- *frequency separation*: if the frequency difference between alternating high and low tones is progressively increased, the perception of a continuously alternating pitch (the 'trill') changes to that of two interrupted tones. The frequency separation at which this occurs was termed the "trill threshold" by Miller & Heise (1950). Using a different measure of streaming based on rhythm, van Noorden (1975) demonstrated that the streaming effect could better be described by two thresholds, one (which he called the "temporal coherence boundary") located at the smallest frequency separation which was too large for the tones to be heard as one coherent stream, the other marking the upper limit of tones that always formed a single stream (the "fission boundary", below which two streams could not be heard). In the intervening range of frequency separations, listeners could alternate between hearing one or two streams.

- *rate of alternation*: van Noorden (1975) mapped out the fission and temporal coherence boundaries as a function of tone onset-to-onset interval. At short tone repetition times (60 ms), the boundaries are quite close, while for larger intervals (150 ms), the boundaries are far apart. However, the fission boundary remains low and is largely unaffected by tone repetition time, suggesting that while it is relatively easy to try to hear two streams, it is very difficult to hold on to a single stream at high repetition speeds.

- *duration*: the default tendency of a stream to be heard as coherent until sufficient evidence to split it has been mentioned. Bregman (1978) found the segregation effect to be cumulative, with evidence accumulating over a period of a few seconds.

Cyclic sequences of somewhat greater timbral complexity have been also been used. Bregman & Pinker (1978) used an alternating sequence of a single tone with a pair of tones to reveal a trade-off between onset asynchrony and frequency separation in streaming: constituents of synchronous tone pairs are more difficult to capture into a competing stream than asynchronous pairs. Bregman & Levitan (1983) put into opposition streaming-by-fundamental and streaming-by-timbre, demonstrating the efficacy of both factors, albeit with a stronger effect of the fundamental.

Stream segregation has also been demonstrated using non-cyclic sequences. Deutsch (1975) used musical scales to demonstrate the dominance of grouping by frequency proximity over grouping by ear of presentation, while Hartmann & Johnson (1991) used an interleaved melody identification task.

Several theories have been proposed to account for aspects of the streaming effect. Three of these are discussed in Rogers & Bregman (1993), to which can be added the peripheral channelling interpretation of Hartmann & Johnson (1991). Rogers & Bregman contrast Bregman's (1990) auditory scene analysis account – which favors sequential grouping by the Gestalt principle of frequency proximity – with those of van Noorden (1975) and Jones (1976). van Noorden's suggestion was that hypothetical frequency jump detectors become adapted and unable to follow the alternating pattern of tones. Jones proposed a theory based on rule-based predictability of sequences.

Rogers & Bregman attempted to distinguish between the three accounts by measuring the effect of preceding 'induction' tones on the streaming of a test sequence of the form HLH_HLH_ (where H and L signify high and low frequency tones, and _ indicates a pause). All induction conditions led to an improvement in streaming effectiveness in comparison to a control condition which used low-intensity white noise. All induction sequences consisted solely of high frequency tones, ruling out van Noorden's proposed adaptation of frequency jump detectors. Induction sequences which differed only in the predictability of inducer tones performed no better than those containing irregular patterns of tones, in contrast to the predictions of Jones' theory.

A second experiment, using inducer sequences which varied in number and total duration of tone elements, demonstrated that segregation improved with the total number of tone onsets rather than the summed tone durations in the inducer sequence. This finding runs counter to Bregman's original hypothesis that the inducer would set up a cumulative frequency bias for the higher tone, but was interpreted by Roger & Bregman as an example of sequential grouping by similarity of the number of tone onsets in inducer and test sequences.

A challenge to the grouping account of the streaming effect comes from the work of Hartmann & Johnson (1991). They used an interleaved melody identification task (Dowling, 1973) to look for streaming effects which could not be explained by the simpler process of peripheral channeling. Peripheral channels were defined as those established in the auditory periphery, and include tonotopic and lateral channels. Elements of one of the interleaved melodies were manipulated in each of 12 different conditions designed to favor explanations in terms of peripheral channelling or grouping (or both). Example manipulations included those that produced differences in frequency range, level differences or duration between the two melodies. They interpreted their results as indicating that "those tone differences that lead to the excitation of different peripheral channels promote stream segregation much more effectively than tone differences that do not excite different channels but which might well evoke the images of different sources, based on other source-grouping grounds." However, Hartmann and Johnson point out that a source-grouping model might contain peripheral channelling as an early component.

2.B Models

A number of models which seek to explain streaming as an emergent consequence of early, low-level, auditory computations have been built, starting with the simple excitation integration model of Beauvois & Meddis (1991, 1996). They sought to explain the perceptual coherence of tone sequences alternating in frequency, as used by van Noorden (1975), noting that listeners tend to hear more than one stream if the tone repetition time is sufficiently short, or if the frequency separation of the tones is sufficiently large. Beauvois & Meddis addressed these findings with a three-channel model, with bandpass channels centered at each of the tone frequencies and at their geometric mean. Noise was added to the rectified output of each channel, and the summed signal formed the input to a leaky integrator. The channel with the highest output was selected, and activity in the other two channels was attenuated by 50%. Temporal coherence was indicated when the short-term averaged level in response to each tone was roughly equal. Beauvois & Meddis showed that temporal coherence could be obtained when the two tones were close in frequency, since in this condition the dominant channel is the middle one, preventing either of the other channels to predominate. Thus,

their average levels of channels at the tone frequencies are roughly the same. They also showed that temporal coherence would occur for larger frequency separation, as long as the tone repetition time was sufficiently long for the excitation in the most-recently stimulated channel to decay over the time course of the interval (this requires tone duration to be short relative to the tone repetition time). Conversely, streaming occurs in the model when the tone repetition interval is short. In this situation, the most-recently activated channel does not suffer a sufficient decay in activity during the tone interval, and the internal noise tends to favor the dominance of one or other channel, leading to an imbalance and hence the model criterion for streaming is obtained. The noise level plays a crucial role in determining the precise balance between coherence and streaming. Beauvois & Meddis demonstrate that a single setting of this parameter allows the model to explain grouping by frequency and temporal proximity, as well as the build up of streaming over time (Anstis & Saida, 1985). However, they acknowledge that the model cannot explain across-channel grouping phenomena such as that of Bregman & Pinker (1978).

McCabe & Denham (1997) extended the Beauvois & Meddis model to include multichannel processing and inhibitory feedback signals, whose strength they related to frequency proximity in the input. This mechanism leads to the suppression of any subsequent stimulus components which are different from those responsible for the suppression. In fact, this residual activity is processed in a separate 'background' map, which in turn has the potential to inhibit components in the foreground map. McCabe & Denham (1997) suggest that their model can be viewed as an implementation of Bregman's old-plus-new heuristic, in which 'new' organization appears in the residual left after subtraction of 'old' components, based on the assumption of continuity. In addition to the streaming data accounted for by Beauvois & Meddis, their model caters for the influence of organization in the background on the perception of the foreground as found by Bregman & Rudnicky (1975).

Recently, Todd (1996) has demonstrated an alternative mechanism to explain some of these primitive streaming phenomena. His physiologically-inspired model computes a 2-dimensional map of activity as a function of best modulation frequency and tonotopic frequency. A slice through this map at a given tonotopic frequency can be understood as an amplitude modulation spectrum in which temporal patterning in the stimulus can be encoded. Todd showed that a cross-correlation of pairs of such amplitude modulation spectra provides a reasonable (and simple) representation of streaming and coherence, with high cross-correlation values indicating coherence. Grouping by frequency proximity occurs in this model for stimuli which are sufficiently close in frequency and which have similar temporal patterns. Grouping by temporal proximity was shown to have the required dependence on the repetition rate of the stimuli. It arises in the model as a consequence of the separation of AM harmonics. At high repetition rates, excitation at the repetition frequency is well separated from its harmonics, leading to a sensitive cross-correlation measure. At lower repetition rates, there is a greater overlap between AM spectra due to a larger numbers of peaks, leading to a higher cross-correlation value and hence reduced sensitivity.

Most of the streaming mechanisms described above require cyclic repetition in order to produce a correlate of fission or fusion. An exception is the model of Godsmark & Brown (1997), which is based on maintaining multiple grouping hypotheses until sufficient information arrives to disambiguate potential organizations. Consequently, their model can handle a wide range of streaming phenomena including context-dependent and retroactive effects (Bregman, 1990). The approach taken by Godsmark & Brown involves training the model to produce streaming effects observed in simple tonal configurations, then observing the more complex emergent grouping behavior on tasks such as polyphonic music transcription. For example, the model produced good matches to listeners' performance in the interleaved melody identification tasks described earlier in this section (Hartmann & Johnson, 1991).

2.C Discussion

Fusion and streaming

Although we have taken streaming as the starting point for our discussion of auditory organization, it presupposes the formation of distinct 'events', possibly requiring the *fusion* of energy in multiple frequency bands. Indeed, Bregman & Pinker (1978) set up a conflict between the formation of single events from simultaneous tones and conventional streaming factors. Factors governing fusion, such as harmonic relations and synchronous onset, have been further investigated and modeled through double-vowel stimuli, as discussed in the next section.

The relevance of streaming phenomena to speech organization

Cyclically-repeated tonal configurations are hardly a common feature of the sound mixtures which listeners typically process. Consequently, it may be unwise to make inferences about the perceptual organization of everyday signals such as speech on the basis of streaming experiments. Bregman's rationale for the use of cyclic sequences (Bregman, 1990, p.53) is largely one of experimental pragmatism, and he urges the use of other methods to verify effects found using cyclic presentation. Since many explanations of listeners' responses to repeated stimuli would be difficult to apply to the general problem of auditory organization, it is conceivable that different mechanisms are invoked to those which apply in more natural settings.

An alternative way to explore grouping is to use stimuli that are somewhat closer to those present in a listener's environment, yet still sufficiently simple to be controllable in an experimental setting. Double vowels are single-presentation stimuli which satisfy these constraints, and the next section looks at their perceptual organization and at models which attempt to account for listeners' identification performance.

3. Double vowels

3.A Listeners

The finding that listeners are able to recognize simultaneously presented synthetic vowels at levels well above chance (Scheffers, 1983) has led to a large number of perceptual studies utilizing this so-called double vowel or concurrent vowel paradigm. Part of the attraction comes from the ease with which stimulus manipulations thought to promote perceptual organization can be performed on vowel pairs. For example, constituent vowels can be synthesized on different fundamental frequencies, modes of excitation, relative intensities and interaural time or level differences.

In the 'standard' double vowel experiment, listeners have to identify both constituents of synthetic concurrent vowel pairs (usually drawn from a set of 5) of a given duration (typically 200 ms).

Key findings for a variety of double vowel manipulations are:

- Concurrent vowels synthesized with the same f_0 can be identified at a level well above chance (Lea, 1992). When the choice is between 5 vowels, a typical result is correct identification of both constituents in 55% of trials.

- Pairs of whispered vowels are identified at about the same rate as vowels with a common f_0 (Scheffers, 1983; Lea, 1992). Whispered vowels may be constructed to contain no clear

grouping cues, so performance in this task is usually taken as the baseline upon which improvements due to grouping are made.

- A difference in fundamental frequency between pairs of concurrent vowels leads to an absolute improvement of 10-15% in vowel identification performance, starting with a difference as small as a quarter of a semitone and asymptoting between 1-2 semitones. This basic finding of Scheffers (1983) has been replicated by several researchers (Assmann & Summerfield, 1990; Culling & Darwin, 1993; Lea, 1992; Meddis & Hewitt, 1992; de Cheveigné, 1997).

- A difference in mode of excitation (voiced/whispered) between the constituent vowels leads to an identification improvement of around 10% (Lea, 1992). Further, the whispered constituent of a voiced/whispered vowel pair was identified significantly more accurately than when both vowels were whispered, but the voiced component was no more intelligible than when both vowels were voiced and on the same f_0 (Lea, 1992).

- Identification performance varies with the harmonicity or inharmonicity of vowel pair constituents (de Cheveigné *et al*., 1995). An inharmonic target vowel presented 15 dB below a harmonic masker vowel was significantly better identified than a harmonic target behind a stronger inharmonic masker.

- When the f_0s of vowel formants are swapped such that the first formant (F1) of one vowel has its higher formants synthesized with the f_0 of the other vowel, and vice versa, or when an f_0 difference is applied only to the F1s of the two vowels, listeners show the same improvement as in the standard condition up to a f_0 difference of 0.5 semitones (Culling & Darwin, 1993). Culling hypothesized that listeners used the time-varying excitation pattern caused by *beating* in the F1 region to identify constituents at times favorable to one or other vowel (Culling & Darwin, 1994).

- Identification improvement with f_0 difference is smaller for brief (50 ms) stimuli than for longer (200 ms) stimuli (Assmann & Summerfield, 1990). Repeating the same 50 ms segment 4 times with 100 ms silent intervals did not lead to any improvement, but performance did improve when successive 50 ms segments were presented with the same silent intervals (Assmann & Summerfield, 1994). Some of this improvement has been attributed to waveform interactions which allow better *glimpses* of one or other vowel at difference times.

- One vowel of the pair (the ‘dominant’ vowel) can be identified at near 100% accuracy for stimuli as short as one pitch period, while identification of the non-dominant vowel improves with an increasing number of pitch periods (McKeown & Patterson, 1995). Introducing a difference in f_0 reduces the number of pitch periods required to reach maximum performance. As well as showing a clear effect of stimulus duration on identification of the non-dominant vowel, these results suggest that f_0 differences are not required for identification of the dominant vowel. The dominance effect can be removed by adjusting levels of constituents in each pair (de Cheveigné *et al*., 1995), a manipulation which may be necessary to allow the conditions of interest to surface.

- Shackleton & Meddis (1992) found that spatial separation of vowels resulted in no increase in identification performance for vowels with the same f_0s. For different f_0s, spatial separation led to a small improvement.

- In a simulated reverberant environment, Culling *et al.* (1994) explored the robustness of binaural and f_0 difference cues, concluding that the latter continued to be useful in reverberant fields that had removed the benefits of the former.

- Culling & Summerfield (1995b) used a reduced form of double vowel stimulus, in which each vowel was represented by two noise bands, to demonstrate an absence of across-frequency grouping by common interaural delay (ITD), although leading the noise bands to different ears did permit recognition. They went on to show that introducing an interaural decorrelation (as opposed to a delay) also allowed identification of the vowels.

- No effects of common, across-frequency, patterns of frequency modulation on double vowel identification have been found (Darwin & Culling, 1990; Culling & Summerfield, 1995a).

Useful reviews of concurrent vowel segregation can be found in Lea (1992), Summerfield & Culling (1995) and de Cheveigné (1993, 1997).

Taken together, these findings suggest that listeners make use of a variety of stimulus properties conveyed by the detailed time-frequency structure of the auditory response to identify double vowels. Some of these can be cast as cues for primitive perceptual grouping, but the role of factors which enable the engagement of vowel schema (e.g. locally-favorable target-to-background level; see Assmann & Summerfield, in press) need to be carefully assessed. In fact, no firm conclusions about mechanisms can be drawn at present, although a number of detailed proposals have been made. These are discussed below.

3.B Models

The first computational model of double vowel segregation was constructed by Scheffers (1983) himself. Scheffers' model employed a harmonic sieve algorithm (Duifhuis *et al*., 1982) in which each f_0 estimate generated a sequence of frequency intervals around each harmonic frequency for that f_0. Peaks in the excitation pattern of the stimulus which fall through these sieve intervals contribute to the evidence for that f_0, and the f_0 which has the largest weight of evidence is chosen. Scheffers introduced a two-vowel procedure which finds the pair of f_0s which together best explain the excitation pattern. His model consistently underperformed listeners (e.g. 27% versus 45% for Δf_0=0), but showed a small improvement with a Δf_0 of 1 semitone (38% versus 62% for listeners). However, this improvement disappeared at 4 semitones difference (27%) while listeners' performance remained at 62%.

Scheffers' harmonic sieve model can be classified as a place domain approach since it operates on a narrowband spectral representation. An alternative strategy is to compute correlates of f_0 by time-domain processing. If this computation takes place on signals filtered by peripheral frequency channels, such approaches are termed place-time processes. A review of place, place-time and pure-time models for double vowel pitch estimation and segregation can be found in de Cheveigné (1993).

One process well-suited to detecting signal periodicities is autocorrelation. Several different autocorrelation-like models have been proposed for auditory computation. In 1951, Licklider suggested a structure for periodicity enhancement consisting of a series of delays, each of which fed a multiplier and integrator, which in turn received an undelayed input. The series of delay elements thus maps out uniformly increasing delays, and the integrated multiplication at any place along this delay axis represents a running autocorrelation with the lag given by the number of delays which the signal passes through to reach that place.

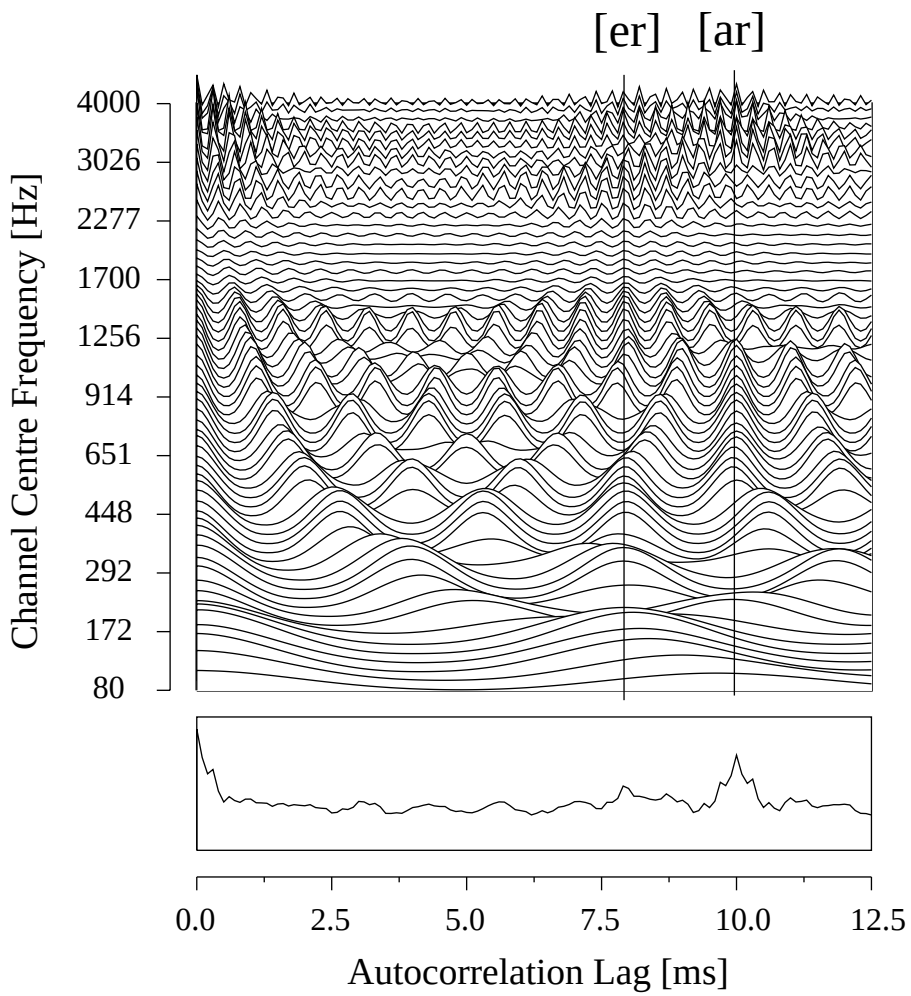

Figure 3: Autocorrelogram of a synthetic double vowel pair ([er] on a fundamental of 126 Hz and [ar] with a fundamental of 100 Hz). The summary correlogram (lower panel) shows a strong peak at an autocorrelation lag of 10 ms, corresponding to periodicities in the signal at harmonics of 100 Hz. A smaller peak at 7.9 ms corresponds to harmonics of 126 Hz.

Assmann & Summerfield (1990) compared two models on the concurrent vowel segregation task. One was a place model similar to that used by Scheffers. The other involved a place-time analysis based on detecting periodicities using an autocorrelation of the output of each channel of a periphery model. Their place model estimated vowel spectra by sampling the excitation pattern at harmonics of the f_0s found by their implementation of Scheffers' sieve. The place-time model estimated vowel pitches as corresponding to the delays with the two largest peaks in a summary autocorrelation function. This summary was created by summing individual autocorrelation functions across channels. Figure 3 depicts an autocorrelogram of a vowel pair together with its summary. Vowel spectra were then formed by taking slices through the autocorrelation functions at lags corresponding to the two pitches. Assmann & Summerfield evaluated the performance of the place and place-time models (and other variants of these involving an optional nonlinear compression stage) and found that the place-time model came much closer to accounting for listeners performance on the same task.

Meddis & Hewitt (1992) also used an autocorrelogram analysis, but chose a different segregation strategy. They first determined the lag of the largest peak in the summary autocorrelogram. They then selected those channels whose individual autocorrelation functions possessed a large peak at this lag. The remaining channels were deemed to belong to the other voice. A further innovation concerned the choice of vowel template. Meddis & Hewitt computed another summary autocorrelation function based solely on those channels selected as belonging to one of the vowels. The lower-order lag coefficients in the summary encode information about periodicities at high frequencies (the lag being inversely proportional to frequency), and they reasoned that spectral information suitable for vowel identification would be encoded in the short-lag section of the summary – which they termed the "timbre region." They repeated this analysis with the unselected channels to get a timbre region vector for the second vowel. Their vowel recognition results, based on channel selection and timbre regions, were very close to the results of subjective tests performed by Assmann & Summerfield.

One issue which has been explored with the aid of double vowel stimuli is the question of whether listeners use an estimate of the fundamental of the target vowel to enhance or select that vowel, or whether the f_0 of the interfering vowel is used to attenuate or cancel it – or indeed whether a combination of both strategies is used. An f_0-based enhancement strategy is advantageous when the target signal is periodic and dominant, since f_0 estimates will be more accurate. Conversely, cancellation ought to favor situations with a periodic and stronger interfering sound.

A number of authors have considered this question in detail (Lea, 1992; de Cheveigné, 1993, 1997). Lea argued that an enhancement mechanism should favor a voiced vowel over a whispered vowel regardless of whether the other vowel was voiced or whispered. By contrast, a cancellation model predicts that a vowel is easier to pick out if the interference is voiced. Lea's experimental results suggests that listeners use a perceptual strategy which can exploit the periodicity of a interfering vowel to help identify a target sound, but that they cannot use target periodicity to extract a vowel from a mix.

More recently, Berthommier & Meyer (1997) have shown how amplitude modulation information can be used as a basis for double vowel segregation. Their "AM map" is computed by performing a 'pitch range' spectral analysis of the envelope at the output of a bank of auditory filters. The resulting representation conveys envelope modulation information as a function of spectral frequency, and can be used in this raw form to group channels which possess a peak at the same envelope modulation frequency. However, Berthommier & Meyer note that the presence of harmonics in the AM spectrum can cause spurious peaks, and propose a further transformation using a harmonic sieve to group these harmonics together prior to vowel classification.

De Cheveigné (1993) proposed a time-domain cancellation model based on a cascade of two comb filters. A comb filter has the property of producing zero output for periodic input signals whose period matches the lag coefficient of the filter. Of course, it is necessary to know the lag parameter in order to actually effect the cancellation; however, the comb filter can be used to find the period of an input signal by searching in filter lag space for a minimum output. Similarly, minimizing the output from a cascade of two such filters by searching over a two-dimensional lag space leads to a time-domain procedure for the estimation of both fundamentals of a pair of concurrent voiced sounds. De Cheveigné compared the performance of this dual-f_0 estimator with a scheme similar to that used in the place-time model of Assmann & Summerfield (1990), described above, based on choosing the two largest peaks in the summary autocorrelogram. His test data consisted of voiced tokens of natural speech. Using the criterion of the percentage of estimates falling further than 3% away from the correct f_0, he found that the comb filter cascade scheme resulted in 10% errors, while the summary correlogram method produced 62% error estimates. De Cheveigné went on to test a neurally-inspired comb filter on auditory-nerve fibre responses to concurrent vowel stimuli,

recorded from guinea pigs by Palmer (1990), demonstrating that it successfully isolated the periodicities of either vowel.

The concurrent vowel paradigm has been used to explore the role of interaural cues in perceptual grouping: Culling and Summerfield (1995b) found that distinguishing noise bands on the basis of interaural time differences was inadequate to convey vowel identities to listeners, whereas interaural decorrelation was, by contrast, sufficient. They accounted for this success with a computational model based on Durlach's (1963) equalization-cancellation procedure.

3.C Discussion

Interplay between pitch and grouping

One issue which models of double vowel segregation have highlighted is the interplay between grouping and pitch: does grouping depend on pitch identification, or does grouping determine pitch, or does each influence the other? It is known, for instance, that onset asynchronies amongst partials of a tonal complex can influence pitch (Darwin & Ciocca, 1992). The very different models of Meddis & Hewitt (1992) and de Cheveigné (1993, 1997) both rely on an initial pitch determination. For Meddis & Hewitt, this allows the grouping of channels, but subsequently, the remaining channels indirectly determine a second pitch. In a sense, this model embodies a bidirectional interaction between pitch and grouping. This interplay should not be too surprising, since grouping should not be considered as a single mechanism, but rather as a set of processes which jointly find coherent structure in the auditory scene.

The time course of double vowel segregation

Some models of double vowel segregation typically operate over short time windows and have difficulty accounting for perceptual findings which involve a wider temporal context (e.g. the results of Assmann & Summerfield, 1994, and McKeown & Patterson, 1995, described in section 3A). Culling & Darwin (1994) have showed that it is not necessary to adopt a time-domain periodicity process to account for listeners' double vowel identification for small f_0 differences (0.25 semitone). Their model used a temporally-smoothed excitation pattern as input to a single-layer perceptron trained to recognize one of 5 vowels, and demonstrated an increase in identification with increasing f_0. They attributed this result to the possibility of glimpsing the changing spectrum arising from the low-frequency beating caused by the small f_0 difference. These results are considered further in the discussion of extending cues across time in the next section.

4. Accumulating grouping information across time

In this section we consider how the auditory system combines information received at different times. It is easy to recognize a temporal aspect to grouping in the many 'buildup' phenomena (discussed above in relation to streaming) where the perception of a stimulus depends on its duration. Many of these phenomena might be explained as no more than sluggishness in the calculation of low-level features, but some may require a separate, central process for integrating a 'grouping' attribute, abstracted from any specific cue. We now examine some of the evidence for such a mechanism.

4.A Listeners

The double-vowel paradigm combined sounds whose properties (fundamental frequency and spectrum) did not vary beyond the scale of their pitch cycles, and in this respect they are unlike most real-world sounds for which the coherent changes in different spectral regions offer a very powerful

indication of common origin. The theoretical account of grouping presented by Bregman (1990) describes the treatment of local, distinct sound elements such as harmonics. These elements are grouped into sources on the basis of various cues; implicit in this account is a central reckoning in which each element is tracked over its period of existence, and evidence for grouping is gathered, stored, and applied over the whole element – even though that evidence may arise from a limited time interval. This subsection considers the experimental evidence for the way that grouping information is used in time at different levels, starting from low-level cues and going on to more abstract inferences; subsection 4B will consider models in which algorithms for combining evidence from different times and different cues form a major part.

Extending a single cue across time

A single cue may influence grouping at times remote from its own temporal focus. Thus, although onset information is present only at the beginning of a tone, the segregation of a harmonic that starts 40-80 ms before the rest of a cluster will persist for many hundreds of milliseconds – as judged from its contribution to the timbre (Darwin, 1984) or pitch (Moore *et al.*, 1986). Thus, a single cue can exert an influence long after it has occurred.

An equally important role for time in low-level grouping is that certain cues may need a significant signal duration for their determination. A detailed pitch judgement, for instance, needs to be averaged across time to reduce internal noise. This may be a factor in the increasing perceptual delay with decreasing pitch difference noted by McKeown & Patterson (1995). Other cues are intrinsically dependent on time, such as the detection of cyclic repetition in iterated frozen-noise stimuli (Guttman & Julesz, 1963; Kaernbach, 1992). Another example, described in Mellinger (1991), is the Reynolds-McAdams oboe signal in which a small degree of frequency modulation is applied to just the even harmonics of a signal that initially has the character of an oboe, but subsequently splits into a clarinet-like tone (formed from the unmodulated odd harmonics) and something like a soprano at an octave above (corresponding to the modulated harmonics). The frequency modulation may take several hundred milliseconds of accumulated observation before it is sufficient to separate the sound into two percepts, but once the threshold has been reached, the influence is much like an instantaneous cue, in that it applies immediately to the tracked continuations of the sound.

Mistuning in double-vowel segregation and harmonic clusters provides an interesting case. In both situations, identification (of the different vowels, or of the presence of a mistuned harmonic) becomes more difficult as the signal duration is reduced from 200 to 50 ms (for vowels; see Assmann & Summerfield, 1994) or 400 to 50 ms (for harmonics; see Moore *et al*., 1986). This suggests a time-integration process able to make finer distinctions when given more of the signal. The alternative explanation, proposed by Culling & Darwin (1994) is that in both kinds of stimulus phase interactions between slightly mistuned harmonics give rise to 'beating' modulations. This may be a cue to discrimination in itself, or it may provide offer 'glimpses' – moments when the signal interactions make the identification task briefly much easier. A longer stimulus has a greater chance of spanning such a glimpse, giving, on average, better identification. If the benefits of glimpsing relied solely on the single best glimpse, a shorter stimulus that happened to contain such a glimpse would be equally well segregated. This is partially supported by the result that certain 50 ms segments give better identification scores than others (Assmann & Summerfield, 1994). However, in that study no 50 ms segment allowed the level of discrimination that occurred with the 200 ms segments, suggesting a benefit from low-level temporal integration available only in the longer stimuli.

Glimpsing has also been proposed to explain the phenomenon of comodulation masking release (CMR), in which the threshold for a sinusoidal target beneath a narrowband noise masker can be *reduced* by *adding* noise bands separate from the target/masker band if the added bands share the amplitude-modulation envelope of the on-band masker (Hall *et al*., 1984). Although there are a

variety of possible cues to this detection (Schooneveldt & Moore, 1989), at least some of the effect appears to result from a comparison between the envelopes in the on-band and flanking frequency channels. For instance, the auditory system could monitor the flanking noise envelopes to detect instants when the on-band masker was briefly at a very low amplitude, giving the most favorable opportunity for 'glimpsing' the target tone, or it could apply processing similar to Durlach's (1963) equalization-cancellation (EC) model (Buus 1985). Before doing this, however, the auditory system must have confirmed that the noise bands are co-modulated; this implies low-level integration along time, either of repeated synchrony between features (such as amplitude peaks), or a more direct calculation of the running cross-correlation (Richards 1987).

In these examples the temporal integration relates to only a single cue, and hence they do not require a central reckoning of an abstract grouping property; the integration can be a direct part of the cue calculation, and the grouping could be rigidly determined on the basis of the single strongest cue. In the next section, however, we look at circumstances where the interaction between different cues is investigated, implying a more complicated process of grouping.

Integrating different cues

Combining different kinds of evidence is one of the most intriguing aspects of auditory organization, and experiments in cue competition form an important paradigm. As we have seen, the Bregman & Pinker (1978) stimuli investigated the competition between the fusion of (near) simultaneous sine tones with the streaming of sequential tones close in frequency. Other experiments have related onset asynchrony to mistuning (Darwin & Ciocca, 1992; Ciocca & Darwin, 1993) or spatial location (Hill & Darwin, 1993). In each case, the result that the effect on grouping of reducing one cue can be compensated for by increasing a different cue implies that, at some level, both cues are mapped to a single perceptual attribute, and thereby become interchangeable.

In fact, the organization of any signal involves the combination of different cues: any simple signal exhibits numerous attributes known to influence grouping such as common onset, harmonicity and common interaural properties. Although a particular experiment may only investigate a single cue, other aspects of the signal, even though they are held constant, will still contribute factors to be integrated into the overall organization. Thus the reduced threshold for detecting mistuned harmonics in longer signals could indicate the kind of integration-along-time discussed above, but it may also reflect a dynamic balance between a continuously-present mistuning cue and the decaying influence of the onset cue. This was directly demonstrated by Pierce (1983), who used a harmonic complex with individual components which abruptly increased in level. At the moment of the change, the boosted harmonic is perceived as separate from the others, but over a timescale of seconds it will 'merge' back into the harmonic complex as the step-change in amplitude becomes increasingly remote in time, and the harmonicity cue regains dominance.

Many experiments have used onset manipulations to investigate other grouping principles such as harmonicity (Darwin & Ciocca, 1992), linguistic formants (Darwin, 1984) and lateralization (Woods & Colburn, 1992). The paradigm typically assumes that a degree of onset asynchrony can preemptively remove the contribution of a particular spectral region from the derived properties of the larger percept. In practice, however, the interaction between onset and other cues may have a more complex temporal development, which can be minimized (but not eliminated) by employing very short stimuli; in contrast, the long stimuli used by Pierce expose these interactions to the full.

The numerous factors influencing the integration of evidence derived from different processes is apparent in experiments concerning the segregation of speech on the scale of sentences. Brokx & Nooteboom (1982) resynthesized nonsense sentences using a monotone pitch different from the constant pitch of continuous interfering speech. This task is unlike double-vowel identification, in that, in addition to f_0 differences, monotone utterances may be distinguished by the common

temporal modulations within each voice, and are subject to wider linguistic-semantic constraints. This greater complexity reveals an interesting trend: whereas segregation of static vowels has plateaued at 12% difference in f_0 (Assmann & Summerfield, 1990), Brokx & Nooteboom saw an approximately linear benefit of pitch separation on intelligibility out to a pitch difference of 20%. More recent studies by Bird & Darwin (1997) have followed this trend out to 60% differences in f_0.

Results of these kind, showing that the organization of sounds depends on a complex interaction between the lowest-level cues, indicate the activity of single, abstract grouping process that depends on a variable combination of basic features, rather than, say, groupings based on single cues which then vie for control of the overall organization.

4.B Models

Although the time dimension provides grouping mechanisms with extra information, it adds a great deal of complexity to the computational task when compared to the essentially time-invariant problem posed by double vowels. We will now look at some of the models that have dealt with these issues by emulating aspects of the organization performed by human listeners on sound scenes at the scale of utterances.

It was not until the mid-1980s that the increasing power of computers allowed researchers to contemplate building algorithmic models of the more sophisticated aspects of auditory perception; at the same time, the principles ultimately described in Bregman (1990) were reaching a wider audience. Weintraub (1985) described the first computational model explicitly motivated by experimental studies of auditory organization. His goal was to separate mixtures of two simultaneous voices, with a view to improving automatic speech recognition applied to each voice. His system used auto-coincidence (a low-complexity version of autocorrelation) of simulated auditory nerve impulses to separate signals of different periodicities in different peripheral frequency bands. Context dependence was included in the form of a Markov model tracking the states (silent, voiced, unvoiced or transitional) of each speaker; the optimal labelling provided by this model controlled a dual-pitch tracking algorithm and guided the division of the signal energy into spectra for each of the two voices. Although the benefits of his system (measured through speech recognition scores) were equivocal, he prepared the ground for subsequent modeling work, particularly in identifying the weaknesses of working solely from local features without the influence of top-down factors.

Cooke's (1991/1993) system decomposed the acoustic mixture into a set of time-frequency tracks called "synchrony strands", then grouped these components using harmonicity (for the lower frequency resolved partials) and common amplitude modulation (for the mid-high frequency unresolved partials). Harmonic grouping employed a temporally-extended form of Scheffers' harmonic sieve, illustrated in figure 4. The main advantage of this scheme lies in the fact that tracking decisions are made locally in frequency. Since grouping relies on identifying each distinct element correctly, situations where features collide and cross can lead to catastrophic mislabellings if the wrong continuations are tracked after the collision. Cooke's algorithm handles sounds with crossing fundamental frequency contours because attributes such as pitch are calculated *after* the tracking of partials, which themselves are less likely to manifest crossing due to the local spectral dominance of one or other source. A further benefit is that the likelihood of a partial falling into an incorrect sieve 'groove' decreases rapidly with the duration of the sieve. To illustrate the generality of the approach, Cooke's model was tested on 100 mixtures of sentence material combined with other acoustic sources, including other sentences. In each case, a worthwhile improvement in signal-to-noise ratio was found. (Different approaches to evaluation are discussed in section 7).

Similar considerations motivated Mellinger (1991) in his study of musical separation. His model tracked spectral peaks across time, grouping peaks with similar onset times or with common

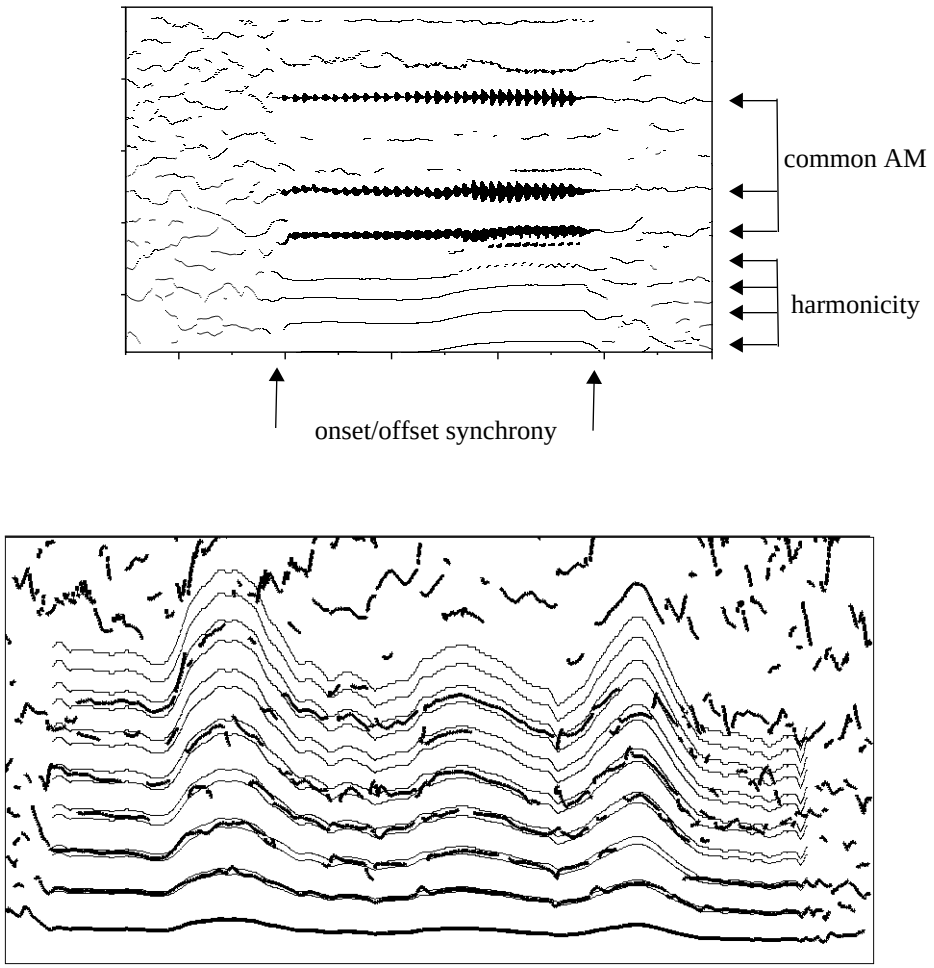

Figure 4: Time-frequency representation and grouping used in Cooke (1991/1993). Upper: synchrony strands and grouping indications for a natural syllable. Strands corresponding to resolved harmonics are visible in the low frequency region. In the mid-high frequency region, strands represent formants F2-F4. The line width encodes instantaneous amplitude, and a clear pattern of amplitude modulation is visible. Lower: synchrony strand representation of the lower spectral region for a completely-voiced utterance, overlaid by a time-frequency harmonic sieve (thin lines). Strands which fall between pairs of sieve lines are deemed to belong to the same source.

frequency modulation. Mellinger's system, like real listeners, maintained an evolving organization, in contrast to Cooke's approach which left all processing until the end of the signal. Newly-detected harmonics had a fixed 'grace period' to build up affinity with existing harmonics, after which they were added to a group, or used as the basis for a new group. Mellinger used the Reynolds-McAdams oboe as one of his test signals; the sudden change in perception from one to two sources in that sound is reflected in an abrupt change in his model's organization, when the initial single source loses the even harmonics to a newly-spawned group (corresponding to the soprano) which has a greater internal coherence of frequency modulation.

Brown (1992) also used a decomposition into partials, and introduced two further innovations. First, he computed a local pitch for each partial by combining the summary autocorrelation function (see figure 3 of the previous section) with the local autocorrelation function in the spectral region occupied by the partial. This has the effect of emphasizing the relevant pitch peak in the summary,

which is used to define the underlying pitch contour for each partial. Second, Brown employed a tonotopically-organized computational map of frequency movement to predict the local movement of partials. His system searched for groups of elements with common pitch contours, favoring sets with common onset times. Brown compared this approach to that obtained using frame-by-frame autocorrelation-based segregation and found that the use of temporal context produced a substantially larger increase in SNR for the target sentence in a mixture.

4.C Discussion

Defining an element

The dominant paradigm for auditory organization, presented by Bregman (1990), involves an analysis of the sound signal into basic elements, defined by their locally coherent properties, from which grouping cues may be calculated and for which grouping decisions can be made. In simple experimental stimuli consisting of sine tones and regular noise bursts, the circumscription of such elements is unambiguous; unfortunately, this is not the case for the noisy, complex sound scenes encountered in the real world. Modelers have often dealt with this problem by limiting their elements to be those defined by strong spectral peaks, but the ability of listeners to organize all kinds of noisy signals may demand a more comprehensive approach. Recent modeling work has attempted to cover a wider range of sounds. Ellis (1996) suggests that a simple vocabulary of tonal, noisy and impulsive elements may encompass most perceptually-salient signals, and Nakatani *et al.* (1997) present a detailed ontology of the signal attributes characteristic of different classes of sound such as speech and music. However, more sophisticated elements tend to be harder and more ambiguous to fit to a particular signal.

Different groupings for different attributes?

Darwin & Carlyon (1995) have cautioned that grouping should not be considered an 'all-or-none' process. Certainly, the interaction of cues in grouping make it misleading to search for a single threshold at which a feature such as mistuning or asynchrony will lead to segregation: these thresholds depend on the contributions of the other cues in a particular experimental paradigm. The deeper point, however, relates to results where, for a single stimulus continuum, measurements based on different attributes give different grouping boundaries. Thus, when a resolved harmonic is mistuned relative to the others in a complex, subjects perceive the harmonic as distinct for detunings of 2%; however, it continues to have an influence on the pitch they perceive for the remaining complex out to mistunings of 8% or more. Darwin & Carlyon see this as evidence for separate grouping processes simultaneously at play – one for the perception of the number of sources, and a different one for the calculation of pitch. There may be an alternative explanation of this as an artifact of the pitch-calculation mechanism's limited ability to respond to differences in organization: even when the harmonic is fully distinct at the abstract percept level, some of its signal characteristics still 'spill' into the pitch calculation of other percepts. This explanation is at odds, however, with the results of Ciocca & Darwin (1993) showing that a sufficiently large onset-time difference can completely remove the contribution of the mistuned harmonic from the pitch of the residual, a phenomenon not attributable to low-level adaptation since it can be released by providing an 'alternative' group to capture the leading portion of the harmonic.

Expectation as the mechanism for combining information along time

Thus far we have been concerned with the grouping of individual 'atomic' elements. There is, however, a higher level at which information could be combined along time: via the influence of 'expectations' – short-term biases towards entire interpretations. Thus, in the experiments of Hukin & Darwin (1995), a harmonic is partially removed from a complex because it is captured by a stream

set up in a preceding sequence of isolated harmonics. This grouping is altered not by any change in the local features of the target harmonic, but by the context of the preceding captor harmonics predisposing the auditory system to treat the harmonic as part of the stream and not the complex. The captor set up an expectation that energy in a certain frequency region formed a continuation of the captor stream; The existence of a gap between the context and the stimulus fragment implies a process operating above the level of elements discussed so far. However, the demonstration that information can exert influence beyond the boundaries of a single region of energy suggests that the model underlying the this section may be unneccessarily narrow: It is possible that onset asynchrony sets up an 'expectation' to affect the harmonics whose beginning it marks, without being specifically attached to those harmonics. This raises the questions of how such 'expectations' are represented, and how they exert their influence. The following section considers the action of such top-down influences in more detail.

5. Context, expectations and speech

Detailed and reliable perceptions of the world turn out to be based upon surprisingly slender and imprecise stimulus information – such as the very limited angle of view of the fovea, or heavily-masked speech in a crowded room. We are able to operate with such limited information in part because our perceptual system is extremely efficient at exploiting and integrating constraints concerning what we 'know' to be the plausible range of alternatives in any given situation. Thus, implicit assumptions of constancy make it unnecessary to scan continuously every item in a visual scene. Similarly, when listening to partially-masked speech, our experience of what comprises a 'reasonable' utterance (in a grammatical or semantic sense) may provide just enough information to construct an impression of how the original speech might have sounded. These aspects of cognitive function involving knowledge and expectation are poorly understood and difficult to research, yet they of are central importance to auditory perception.

Progress in automatic speech recognition in the last decade has been due in a large part to successful techniques for combining 'bottom-up' information derived from the input signal with 'top-down' constraints imposed by the recognizer's knowledge of vocabulary and grammar. Speech perception is a specialized instance of the principle that expectations are used to facilitate perceptual organization; later in this section, we will discuss some of the emerging work on integrating models of auditory scene analysis with speech recognition systems. First, we look at some of the experimental results demonstrating this principle in action.

5.A Listeners

Local context and "old-plus-new"

An 'expectation' is a state of the auditory processing system that will substantially affect the interpretation of a subsequent stimulus. A classic illustration of such an effect is the way in which listeners compensate for the spectral coloration imposed on a signal by the transmission channel. Thus a simple filter can convert the vowel sound in an utterance of "bit" so that, when heard alone, a listener will hear it as "bet" (Watkins, 1991, as discussed by Assmann & Summerfield, in press). However, if the altered word is prefixed with a carrier phrase ("Please repeat this word: bit") modified by the same static coloration, the word is restored to its original phonetic identity: Through exposure to the longer sample, the auditory system has separated the effects of source speech and channel coloration, and has compensated for the latter in the interpretation of the target word. This is an *expectation* because the inference of channel characteristics from the carrier phrase makes a categorical difference to the perception of the target word; the expectation that the channel will continue to color the speech has altered the treatment of the stimulus.

Expectation encompasses the general principle of auditory perception termed “old-plus-new” by Bregman (1990), related to the powerful real-world constraint of the independence of sound sources. Any abrupt change in the properties of the signal probably reflects a change in only one source, and a change in the source spectrum that consists of only an energy *increment* will be interpreted as the *addition* of a “new” source, while all the existing “old” sources continue unchanged – the signal following the change is interpreted as being old-plus-new, and the properties of the new source are effectively calculated by finding the difference between the signal before and after the change.

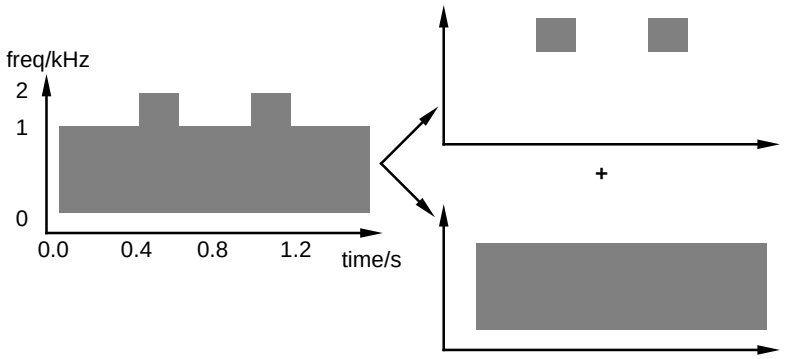

Figure 5: Schematic representation of the alternating narrow- and broad-band noise stimuli, and its perceptual organization, illustrating the principle of old-plus-new.

The old-plus-new idea is illustrated in figure 5 (after Bregman, 1990, p. 344). The alternation between narrow and broader bands of noise is heard not as switching between two different signals but as a continuous low noise to which high noise bands (the difference between the narrow and the broad) are periodically added. Physically, the two interpretations are equally valid, but the auditory system irresistibly chooses division in frequency because it meets the old-plus-new criterion. The interpretation as the alternation between the two noise bands would require the (less likely) coordination of the narrow band of noise turning off at the very instant that the broader band turns on.

Continuity and induction

The most dramatic consequences of expectations in the auditory system occur when an object or source is perceived in the absence of any direct, local cues to its sound. In such situations, the perceived object is ‘induced’ from expectations set up by its context.

The simplest illustration of induction is the continuity illusion (Bregman, 1990, p.28, studied earlier as the “pulsation threshold” e.g. in Houtgast, 1971). If a steady tone has a brief burst of wideband noise added to it, the energy of the noise may mask the tone, leaving the auditory system without direct evidence that the tone is present during the noise (indeed, for increasingly intense and/or brief noise bursts, it is impossible to say if a tone is present with any certainty *a postieri*). In these circumstances, the percept is typically of the tone continuing during the noise despite the absence of tonal features from the stimulus during the burst. The auditory system rejects the interpretation that the tone has ceased during the noise burst because, although it is an adequate explanation of the stimulus, it violates the principle of old-plus-new.

More complex examples of auditory induction are provided by the phonemic restoration phenomena investigated by Warren (1970) and others. In the original demonstration, a single phoneme (the first /s/ in “legislatures”) was attenuated to silence then masked by the addition of a cough. Not only were listeners unaware of the deleted phoneme (the speech was heard as complete), but they were unable

to specify the exact timing of the cough, making a median error of 5 phonemes. Evidently, auditory processing had exploited the redundant information in the speech signal (co-articulatory, phonotactic and semantic) to 'induce' the identity of the masked (missing) segment, a process so complete that, at the level of conscious introspection, it was indistinguishable from 'direct' (non-restored) hearing. Subsequent experiments showed that a keyword occurring several syllables *after* the masked segment could provide the semantic constraint to restore the deleted phoneme, since listeners would reliably perceive *different* restorations for stimuli that differed only in the final keyword (Warren & Warren, 1970). These results demonstrate not only the very powerful effect of expectation in the perception of speech, but also that such 'expectations' can operate backwards in time. Induction also appears to operate between ears ("contralateral induction", Warren & Bashford, 1976) and across the spectrum ("spectral induction", Warren *et al.*, 1997). In the latter study, the spectrum is reduced down to two narrow signal bands with a commensurate reduction in intelligibility. The introduction of an intervening spectral band of noise then modestly increases intelligibility.

Speech information can be combined across regions disjoint in both time and frequency, as demonstrated by "checkerboard noise" masking experiments of Howard-Jones and Rosen (1993). They used stimuli in which speech was alternated with noise in several frequency bands, such that half the bands carried unobstructed speech while masking noise was added to the interspersed remainder, and the pattern of noisy and clear channels flipped every 50 ms to give noise interference that resembled a checkerboard on a log-frequency spectrogram. They found that for a two-channel division (above and below 1.1 kHz), listeners were able to tolerate a level of checkerboard noise 10 dB higher than control conditions of noise gated in one channel but continuous in the other, demonstrating that information from separate frequency regions was being integrated across time. (For wideband pink noise gated at 10 Hz – i.e. simultaneous 'glimpses' in high and low channels – a further 7 dB of SNR decrease was acceptable). Their result supports the notion of a central speech hypothesis (another kind of 'expectation') that gathers information from any available source, rather than more local processes acting to integrate information only within frequency channels. There are numerous other unnatural manipulations of speech from which listeners recover intelligibility; see Cooke & Green (in press) and Assmann & Summerfield (in press) for further discussions.

Speech as the best explanation

The capacity to infer the presence (and identity) of speech with limited evidence is well demonstrated by sine-wave speech (Bailey *et al.*, 1977; Remez *et al.*, 1981, 1994), in which the time-varying frequencies and levels of the first three of four speech formants are resynthesized as pure sine-tones, removing cues to the excitation source present in the original. Although listeners hear such sinewave utterances as a combination of whistles (the interpretation that might be expected), they are often able to interpret them as speech when so instructed.

The combined perception of whistles and speech make sine-wave utterances similar to so-called "duplex" phenomena (Rand, 1974; Liberman, 1982), in which some portion of the stimulus (e.g. an isolated formant transition) is interpreted both as part of speech and as an additional source. For instance, Gardner & Darwin (1986) showed that the application of frequency modulation to a harmonic near to a formant in a synthetic vowel caused the harmonic to stand out perceptually but at the same time to contribute to the vowel percept.

A third example of the very powerful predisposition of the auditory system to interpret the most tenuous of stimuli as speech comes from the description of "temporal compounds" by Warren *et al.* (1990, 1996). The later study used random arrangements of six, 70 ms synthetic vowels made from real glottal bursts, concatenated to form a single repeated token in which listeners could no longer identify the individual vowels or their order; the sequence fused into a "temporal compound" which was perceived as syllables. The resulting signal did not resemble any real utterance, but rather than

perceiving it as a nonspeech sound with some speech-like qualities, listeners often heard *two* simultaneous voices pronouncing syllable sequences. The auditory system appears to reconcile the contradictory speech cues by relaxing the constraint that they be interpreted as a single voice, rather than abandoning a speech-based interpretation. The syllables were invariably drawn from the set commonly used within the native language of the subject, with the result that even given that inter-subject agreement of the perceived syllables was not very strong, speakers of different languages would interpret the same stimulus very differently. Compare these results to phonemic restoration, which can be seen as an interplay between the local cues of context, and the underlying linguistic constraints; in these artificial vowel stimuli, the local cues are largely invalid (since the signal is not, in fact, real speech), so the interpretation relies primarily upon the long-term constraints, expressed as the acceptable 'syllabary' for the listener's native tongue.

Studies such as these reveal the auditory system's strong tendency to interpret any credible signal as speech, invoking a wide range of constraints derived from language structure and the content of the message. These constraints can form a very powerful basis for overcoming distortions and masking in the original signal. In the next section, we describe computational models that have addressed the application of expectations and other high-level constraints in the interpretation of auditory scenes.

5.B Models

Blackboards and explanation-based systems

The perceptual phenomena described above highlight the importance of stored knowledge and expectations in permitting the interpretation of sound. A popular approach in modeling has been to use collections of *knowledge sources* encapsulating specific, limited aspects of the necessary knowledge, and able to act independently to solve the larger explanation problem. Knowledge sources typically co-operate through a common data structure, called a *blackboard*. Several systems for computational auditory scene analysis have been built around blackboard architectures (Carver & Lesser, 1992; Nawab & Lesser, 1992; Cooke *et al.*, 1993; Nakatani *et al.*, 1998; Ellis, 1996; Klassner, 1996; Godsmark & Brown, 1997). Blackboards support an arbitrary combination of data-driven and hypothesis-driven activity, making them suitable for incorporating higher-level knowledge of use in the source separation task. For example, the highest representational level of Klassner's system is a set of "source-scripts", which embody the temporal organization of source sequences such as the regular patterning of footfalls.

One common feature of the blackboard models is the importance placed on generating consistent explanations for *all* of the acoustic evidence. Nakatani *et al*. (1998) call their system a *residue-driven* architecture. Events (in their case groups of harmonically-related elements) are continuously tracked, and predictions about the immediate future are made. These predictions are compared with the actual outcome and the discrepancy, or *residue*, is computed by subtracting the prediction from the remaining mixture. Residues require explanation, often by the creation of new trackers. In this way, their scheme embodies Bregman's old-plus-new principle.

Klassner's (1996) blackboard system also focuses on discrepancies between the observed signal features and those that would be consistent with the current explanation. In his case, however, the discrepancies may be resolved either by modifying the explanation or by changing the parameters of the front-end signal-processing algorithms used to generate the features. Since the optimal values for factors such as filter bandwidth and energy thresholds depend on the detailed conjunction of sources present, his system places those parameters within the control of the blackboard procedures – in sharp contrast to the fixed single-pass signal-processing employed in other models. His system comprises a dual search in explanation space and signal-processing parameter space to find the best explanation for a given sound scene in terms of 39 abstract templates for everyday sounds such as "car engine" and "telephone ring."

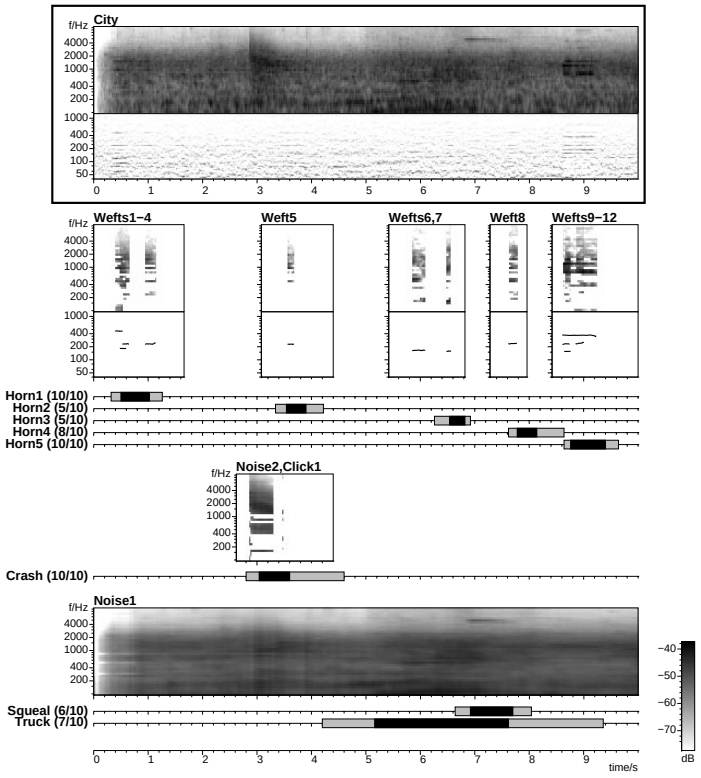

Figure 6: Example figure from Ellis (1996). The top panel shows a 10 s excerpt of "city-street ambience" represented by its time-frequency energy envelope and summary periodogram. Elements below are the 'explanation' of the scene in terms of generic sound elements, along with the distinct sound events reported by listeners in a subjective test.

Ellis's (1996) thesis presents "prediction-driven CASA" as an alternative to the data-driven systems described in section 4. Motivated more closely by auditory realism than the other blackboard systems, his system constructs accounts of the input sound in terms of "generic sound elements" to act as the link between raw signal properties and abstract source descriptions. Most earlier systems for CASA were limited to the separation of voiced sounds, and their choice of representations (e.g. tracked partials) reflected that fact. Ellis's system sought to model unvoiced sources such as noise bursts or impulses, through an expansion of its representational vocabulary. The uncertainty implicit in modeling noise signals further led to a system tolerant of hypotheses for which direct evidence might be temporarily obscured, a framework consistent with the induction phenomena mentioned in section 5A. In Ellis's system, periodic sounds are treated as a special case, with a correlogram-based pitch tracker triggering the creation of "wefts" (i.e. coherent sets of parallel threads; Ellis, 1997a) that estimate the energy at a given modulation period in each frequency channel. The number and timing of events identified by Ellis's system were in good agreement with the sources identified by listeners in the ambient sound examples such as "city street".

Motivated by the goal of reproducing complex perceptual phenomena like ambiguity and restoration, blackboard-based systems have the potential to exhibit very complex behavior arising from the interaction of their abstract rules. However, crafting the knowledge bases is a slow and difficult art, which offers no obvious solution to unrestricted, full-scale problems. Although this may not be a direct concern, progress in fields such as speech recognition suggests the superiority of 'fuzzier' techniques in modeling perceptual interpretation tasks, and in particular the value of,

exploiting training data to tune system parameters. There are also more rigorously-motivated approaches to the problem of integrating widely disparate sources of knowledge; the OPTIMA system of Kashino *et al.* (1998) approaches the problem analyzing complex acoustic signals – in their case, polyphonic music – through the probabilistic-theoretic framework of Bayesian networks.

Integration with speech recognition

Computational auditory scene analysis offers a possible solution to the serious challenges of robust automatic speech recognition. Lippmann (1997) has argued that current approaches to robust ASR (reviewed in Gong, 1995; Junqua & Haton, 1996) are far less flexible than those employed by listeners. In addition to the variability caused by reverberation and channel distortion, recognizers in real-life environments have to cope with the nonstationarity of both target and interfering sources and the fact that the number of sources active at any moment is generally unknown. CASA is attractive because it makes few assumptions about the nature and number of sources present in the mixture reaching the ears, relying only on general properties of acoustic sources such as spectral continuity, common onset of components, harmonicity, and the various other potential grouping cues described in earlier sections.

Several attempts have been made to integrate CASA with ASR. The most common approach uses CASA as a sophisticated form of speech enhancement, relying on an unmodified speech recognizer to do the rest. For instance, Weintraub (1985) passed separate resynthesized signals to a hidden Markov model speech recognizer. Similarly, Bodden (1995) used binaural preprocessing prior to ASR. The main attraction of the speech enhancement route is that it allows use of existing criteria in assessing the performance of a CASA system: As well as SNR improvements and ASR recognition rates, the intelligibility and naturalness of CASA-enhanced speech can be measured through listening tests.

The enhancement-only interpretation of CASA has been much criticized of late (see, for example, Bregman, 1995; Slaney, 1995; Cooke, 1996; Ellis, 1996) – although the weakness was certainly recognized even by Weintraub (1985). Slaney (1995) presents a "critique of pure audition" in which he argues against a purely data-driven approach to auditory scene analysis, inspired by an analysis of top-down pathways and processes in vision (Churchland *et al.*, 1994). Bregman (1995) too has warned against the "airtight packaging" of segregation as a preliminary to recognition, invoking duplex perception of speech as an instance where recognition overrides segregation, thereby "defeating the original purpose of bottom-up ASA".

An alternative approach to the integration of CASA and ASR has been proposed by Cooke *et al.* (1994). This scheme relies on CASA to produce an estimate of spectro-temporal regions dominated by one or other source in a mixture, and applies missing data techniques to recognize the incomplete pattern. It fits naturally with channel selection schemes such as that of Meddis & Hewitt (1992) described earlier in the context of double-vowel identification. Channel selection is further inspired by neurophysiological oscillator models which rely on synchronous activity in a subset of channels to signal grouping of elements (see section 6).

The missing data strategy works on the assumption that redundancy in the speech signal allows successful recognition with moderate degrees of missing data. Robust recognition performance in the face of missing data can be obtained, and further improvements are possible when models of auditory spectral induction (Warren *et al.*, 1997) are incorporated (Green *et al.*, 1995; Morris *et al.*, 1998).

Auditory induction – or, more generally, the effect of perceived auditory continuity – has motivated a number of CASA systems. Ellis (1993) argued that restoration would be necessary to overcome obscured features in data-driven system, and his system makes the inference of masked regions a

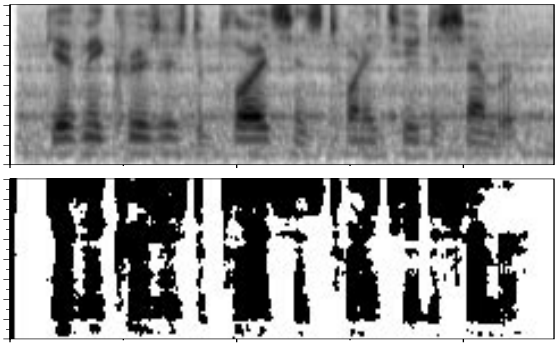

*Figure 7: The upper panel shows an auditory spectrogram for the utterance "*GIVE ME CRUISERS DEPLOYED SINCE TWENTY TWO DECEMBER*" mixed with Lynx helicopter noise at a global SNR of 18 dB. Dark regions of the lower panel indicate those areas where the local SNR is positive. Attempts to recognize the mixture with a conventional recognizer yielded "*IS HORNE+S FOUR DECEMBER*" while use of first-generation missing data techniques via the lower mask produced "*GIVE CRUISERS DEPLOYED SEVENTH DECEMBER*".*

central part of the prediction-reconciliation analysis (Ellis, 1996). Okuno *et al*. (1997) described a scheme in which the residue remaining after extracting harmonically-related regions is substituted in those temporal intervals in which no harmonic structure could be extracted, arguing that this residual is a better guess for the continuation of the voicing than silence would be – since, at the very least, it will permit induction in listeners faced with the resynthesized signal.

Ellis (1997b) makes a specific proposal for incorporating speech recognition within scene analysis. Extending his prediction-driven approach, he includes a conventional speech-recognition engine as one of the "component models" that can contribute to the explanation of a scene. An estimate of the speech spectrum, based on the labeling from the speech recognizer, is used to guide the analysis of the remainder of the signal by nonspeech models; this re-estimation of each component can be iterated to obtain stable estimates.

5.C Discussion

The significance of expectations

This section has focussed on the role of expectations and abstract knowledge in auditory perception, and on efforts to model these effects. Although some of the stimuli involved are contrived, there are important implications from the demonstration that, in the absence of adequate direct cues, the auditory system will employ information from elsewhere to build its interpretation of a scene – and, as seen in the original Warren (1970) experiments, that such 'restored' information is consciously indistinguishable from 'direct' evidence. Given the enormous power of high-level constraints to restrict the range of interpretations that need be considered, listeners might be inclined to rely on inference in many circumstances besides those in which information has been obscured. Clearly, perception exists as a compromise between finding direct evidence of particular sources and the mere absence of contradictory evidence.

Retroactivity

Certain perceptual phenomena, starting with the phonemic restorations which depended on a later keyword (Warren & Warren, 1970), but including much simpler signals such as noise bands of abruptly alternating bandwidths (Bregman, 1990), show that the interpretation of a sound must

sometimes wait for as much as several hundred milliseconds or longer before it can be finally decided. Examples such as the Reynolds-McAdams oboe (Mellinger, 1991) illustrate an initial organization which is consciously revised i.e. the listener is aware of the change in organization. Blackboard systems such as those of Klassner (1996) and Ellis (1996) that maintain multiple alternative hypotheses can exhibit backwards influence in certain circumstances; the system of Godsmark & Brown (1997) explicitly grows its "decision window" until ambiguity can be resolved. Ultimately, models may need an exceptional ability to return to and revise decisions that were previously considered complete, although it is not clear at what level of representation this reassessment might apply.

Duplex perception, masking, and auditory induction

The idea that a single speech fragment can simultaneously be both perceptually segregated (i.e. exist as a separate source) and perceptually integrated (i.e. contribute to a phonetic judgement) may be tied up with the notion of auditory induction. It is easy to conceive of an architectural arrangement in which primitive cues such as differences in harmonicity give rise to assignments of harmonics to different streams, but which co-exist with top-down expectations looking for evidence of speech. Since differences in harmonicity for a single formant, for instance, only serve to redistribute rather than to remove energy in a given spectral region, it is possible that the mistuned harmonics appear as suitable material to 'complete' a phonetic hypothesis. Speech is readily identifiable with large spectral regions removed (Fletcher, 1953; Steeneken, 1992; Warren *et al*., 1997; Lippmann, 1996), thus it is hardly surprising that identification is possible when otherwise missing regions (perceptually segregated harmonics) contain some energy. This argument can be extended to cover other duplex phenomena as long as auditory induction is allowed to operate on the source mixture, since the duplex fragment is likely to provide a credible masker for the missing structure.

6. The neurophysiological substrate for grouping

The notion of an "implementation level" for auditory organization was introduced in section 1, but intervening sections have mainly addressed the higher levels. In a biological system, how are features which originate from the same source marked as belonging together? von der Malsburg & Schneider (1986) called this the "binding problem", and suggested a computational solution in which neurons which encode a common environmental cause are grouped by synchrony of their temporal response. This elegant proposal allows grouping to be represented 'in place', without the need for separate neural structures dedicated to representing the results of grouping. Their implementation models networks of neurons whose output is characterized by an oscillatory pattern. They demonstrate binding of responses, marked by a common phase of oscillation, in a simple auditory example in which common onset and simultaneous activity in different frequency bands give rise to grouping between the channels. Their proposal also allows an attentional mechanism to 'strobe' the temporal pattern and get an unobstructed, if incomplete, view of the attended source (Crick, 1984).

These ideas have been actively researched in vision, where a similar binding problem exists for object segregation. Such investigations have received added impetus from physiological studies which appear to show that visual stimuli can elicit synchronized oscillations across disparate regions of the visual cortex (Gray *et al.*, 1989). Although specific evidence of visual binding through oscillations has failed to appear, the mechanism retains its attraction.

A study by Liu *et al*. (1994) is one of the few attempts to apply neural oscillator models to speech recognition. Strictly, their model does not address auditory grouping, but can nevertheless be interpreted as a mechanism for schema-driven grouping. The model encodes local peaks in a sharpened mel-scale LPC spectrum as independent sets of oscillations which they assume correspond to vowel formants. These oscillations interact with an associative memory in which

formant-vowel associations are hard-wired. Reciprocal top-down and bottom-up activation leads to synchronized oscillations in those spectral regions which globally correspond to a known vowel.

Recently, a number of studies have directly addressed the search for an account of auditory grouping phenomena in terms of neural oscillators (see Brown *et al.*, 1996, for a review). Brown & Cooke (1998) presented an oscillator model which can account for a number of streaming phenomena, including grouping by frequency and temporal proximity, the temporal build-up of streaming, grouping by common onset, and grouping by smooth frequency transitions. The same model, operating on a different input representation, can also account for grouping by common fundamental (Brown & Cooke, 1995), and at the same time provide an adequate explanation for the interaction of onset asynchrony and harmonicity (Ciocca & Darwin, 1993).

Their model consists of 3 stages: an auditory filterbank, hair cell and onset cell simulation, feeding a fully-connected network of neural oscillator units. Units are coupled to each other with a strength defined by a matrix of weights. These weights adapt dynamically during stimulus presentation; they also incorporate a degree of temporal integration. Oscillator dynamics are such that units with a high coupling weight tend to produce similar responses. Coupling strengths are modified in such a way that inputs which undergo common changes lead to an increase in coupling. This embodies the gestalt principle of common fate, discussed by Bregman (1990, p. 248).

When presented with stimuli such as those used by van Noorden (1975) and by Beauvois & Meddis (1991), the model displays the required sensitivity to tone repetition time. Coupling strength is initially low because of the onset asynchrony of the low and high frequency tones. However, coupling strength recovers during the tone repetition interval. If the interval is short, little recovery is possible, and coupling strength is driven lower still by the next asynchronous onset. For longer intervals, more recovery in coupling strength is possible. Frequency proximity effects in the model follow from the overlapping filter response areas in the periphery simulation. When the higher and lower tones are close in frequency, filters with center-frequencies (CFs) near to the lower frequency are excited when the higher frequency is present, and vice versa. Since adaptation of the coupling between units at these CFs is dependent on response similarity, there will be a spectral region within which coupling is maintained. Figure 8 depicts the time-course of oscillator responses and illustrates model and listener responses in a simple two-tone streaming task.

Brown & Cooke's model also accounts for the greater coherence of sinusoidal FM stimuli over square wave FM, as noted by Anstis & Saida (1985). Filters close to the maximal FM frequency are strongly stimulated for half a cycle of square wave FM, thus driving down the coupling between those units at that frequency and those at the minimum FM frequency. By contrast, such filters receive strong activation for a smaller fraction of the FM cycle for sinusoidal FM. This corresponds to the gestalt principle of good continuation (Bregman, 1990, p. 133).

Brown & Cooke (1998) adapted coupling weights between channels using differences in onset activity across channels. Similarly, modification of coupling weights using differences in autocorrelation strengths across channels enabled the model to encode streaming by mistuned harmonics (Moore *et al*., 1985). Further, the segregating effect of onset asynchrony possessed by a mistuned partial (Darwin & Ciocca, 1992) was present in the model.

Wang (1996) presented a model consisting of a two-dimensional time-frequency grid of relaxation oscillators. Local excitatory connections between units in both time and frequency endows sensitivity to temporal and frequency proximity on the network. Brown & Wang (1997) extended this model to incorporate a simulation of the auditory periphery, and applied it to the problem of double-vowel segregation. Their system uses relaxation oscillators, whose output is characterized by a repeating sequence of active and silent phases. Brown & Wang used an autocorrelogram representation, which is processed by the autonomous iteration of the following procedure: First,

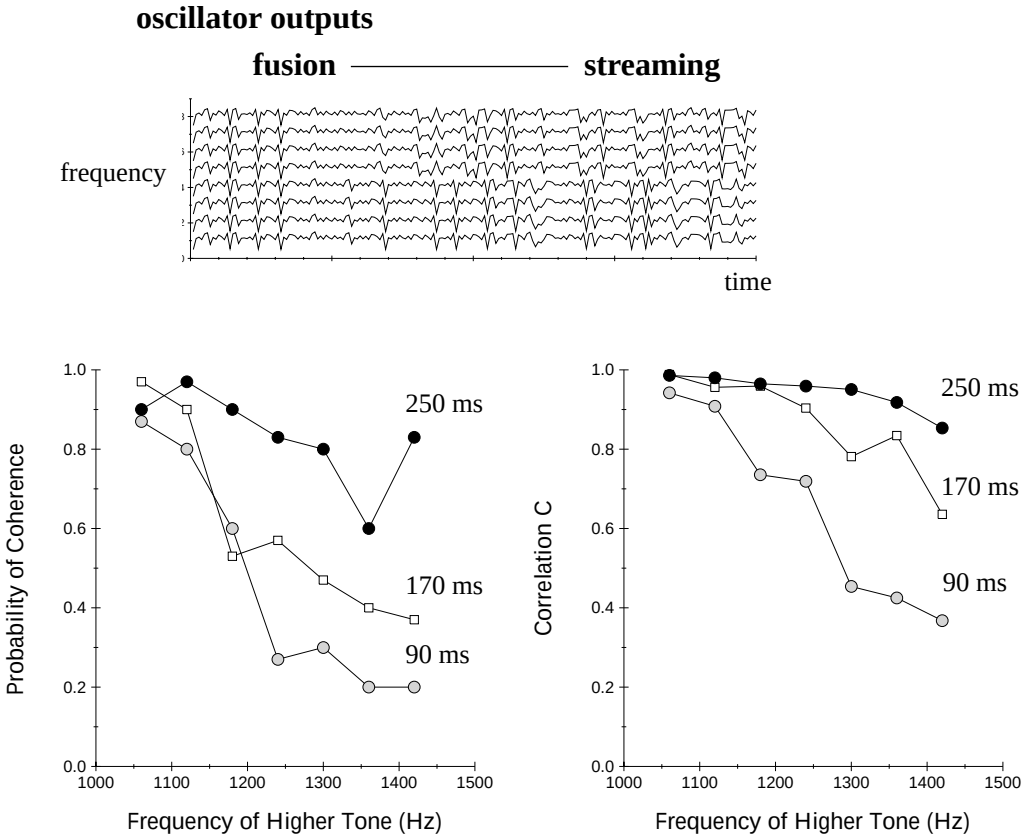

Figure 8: A neural oscillator model of streaming (Brown & Cooke, 1998). Upper: individual oscillator outputs, showing the transition from all channels synchronized (i.e. fused) to fission into two separate streams. Lower panels: comparison of model performance (right) with subjective streaming results (left) for the two-tone sequence depicted in figure 2.

those channels whose associated oscillators are in their active phase are summed to produce a selective summary autocorrelation function. Then, the largest peak in the summary is then used to promote synchronization of all oscillators whose channels have a peak at the corresponding delay. Finally, as a consequence of oscillator dynamics, these synchronized oscillators move from an active to a silent phase. As a consequence, other oscillators move to their active phase and the iteration repeats. This time, a different set of channels makes up the summary, leading to the selection of a different pitch peak. The process continues until all channels are synchronized to one or other periodicity present in the stimulus.

Thus, neural oscillator models have been particularly successful at providing accounts of the interaction of cue combinations, such as common onset and proximity; their ability to support the exploration of such questions is attractive. This may be due to the limited vocabulary of neural architectures, in which information can only be represented as activations and weights, and thus different cues are necessarily expressed in forms that can be combined. By contrast, a traditional symbolic model of grouping might represent periodicity and onset time attributes quite separately, requiring both to be further mapped to some ‘grouping strength’ axis before their interaction could be considered.

The oscillatory aspect of these models confers neural architectures with additional advantages. Most obviously, it provides the dimension of synchrony to indicate time-varying ‘bindings’ between

different features, representing the abstract property of grouping that makes symbolic organization models seem so neurally implausible. The second benefit of the oscillator systems comes from the way they establish equilibria between extremes of grouping. Whereas an explicit model that incorporates 'grouping strengths' among different components (such as Mellinger, 1991) requires an arbitrary threshold to convert strengths into groups, the emergence of synchrony groups in oscillator arrays adjusts automatically to track the weight strengths.

The common onset cue to auditory grouping can map directly to a synchronized 'kick' which establishes a relationship between several oscillators. In other respects, however, neural oscillators should be seen more as a general-purpose technique applicable to any problem of choosing element subsets based of a variety of cues. Their main attraction when compared with more procedural algorithms is that they could plausibly be present in the brain; at the same time, they may be harder to diagnose and modify than less neurally plausible approaches (Ellis, in press). As indicated above, they can account for a variety of auditory streaming phenomena such as temporal and frequency proximity, common onset, good continuation and the temporal build-up of streaming. When applied to representations which encode stimulus periodicities, they have been used to denote streaming by harmonic mistuning and by leading partials in a harmonic complex.

7. Current issues in models of auditory organization

This concluding section identifies some of the distinctions between the various approaches to modeling auditory organization, and describes unresolved issues for the future.

What is the goal of computational auditory scene analysis?

The common goal of CASA systems is the intelligent processing of sound mixtures, but individual systems differ both in the kind of sounds that are handled and in the information about them which is to be extracted. Some approaches seek to pluck a particular signal out of an interference whose properties are essentially ignored (e.g. the enhancement of the target voice in Brown, 1992), while others are concerned with making a complete explanation of *all* components in the acoustic mixture (e.g. Ellis, 1996). The former 'target enhancement' approach pursues algorithms with broad applicability by making the fewest assumptions (e.g. only that the interference will be lower in energy that the target over a significant portion of the time-frequency plane). By contrast, 'complete explanation' accepts the added complexity of characterizing portions of the signal that are to be discarded, in the belief that this is necessary to reproduce human-style context-adaptive processing in which the interpretation of a target is influenced by non-target components. Such influences include the requirement of a plausible masker (Warren *et al.*, 1972). Generally, a signal is influenced by the interpretation of others via top-down influences, and hence only in systems the employ such constraints.

A second debate over the fundamental problem structure concerns the output of organization systems. Auditory scene analysis ought to result in an abstract description of the sources identified (Darwin, 1984), but the nature of this description depends crucially on the particular application domain. One attractive paradigm is a system that converts a single acoustic mixture into several output signals, each consisting of a mixture component heard in isolation. This resynthesis strategy of systems such as those of Weintraub (1985) and Brown (1992) may, however, be unnecessarily demanding, especially for applications such as speech recognition (Cooke, 1996). Instead, an intermediate representation describing the identified signal along with the confidence in each parameter (thereby indicating the 'missing data') provides the results of signal organization in a form more appropriate for input to a subsequent processor able to take advantage of this added information.

Evaluation

Resynthesis of an enhanced target in a mixture permits system evaluation via listening tests. Most CASA systems possess one or more internal source representations which can be used for resynthesis. Other researchers have argued that an adequate model should represent all the perceptually-significant information about a sound, and be able to resynthesize sources without further reference to the original mixture (Ellis, 1996). While this latter approach escapes the problems with overlap in time and frequency, the distortions associated with highly nonlinear analysis and resynthesis techniques present formidable challenges in creating high-quality output. Mistakes in grouping assignments often become very prominent in resytheses; although this can be uncomfortable for the modeler, it also carries a diagnostic benefit.

The systems of Cooke (1991/1993) and Brown (1992) were both evaluated through a calculation of the SNR improvement on test mixtures. Since energy in an output signal cannot be directly associated with a single input component, both evaluations posed a correspondence problem. Cooke classified his "strand" elements for closeness to representations of the separate input components, whereas Brown was able to calculate the attenuation from his time-frequency mask for target and interference presented in isolation. Ellis (1996) sought a more perceptual measure of separation success by conducting listening tests in which subjects were asked to rate, on a subjective scale, the resemblance of resynthesized components to the individual sources they heard in the full original mixture.

Other approaches to evaluation include speech recognition and intelligibility scores (Weintraub, 1985; Bodden, 1995; Okuno *et al.*, 1997), and simulations or equivalents of psychoacoustic tests. Thus the streaming models of section 6 have a particular interpretation put on their state which is equated to the formation of a stream in a listener, and the environmental sound mixtures of Klassner (1996) are analyzed as combinations of known types – a kind of 'forced choice'.

Unlike large-vocabulary automatic speech recognition or message understanding, computational auditory scene analysis lacks a formal evaluation infrastructure at present. This makes it difficult to gauge strengths and advances both within the CASA community and between the various alternative approaches to the problem of understanding sound mixtures. Besides traditional signal processing methods (e.g. Denbigh & Zhao, 1992), more recent innovations have included independent component analysis (Bell & Sejnowski, 1995), which works by minimizing mutual information between the outputs of a neural network, and HMM decomposition (Varga & Moore, 1990), which attempts to find the most probable collection of speech models to explain the mixture.

One suggestion for evaluation comes from Okuno *et al.* (1997), who propose a "challenge problem" for CASA – a task whose solution will stimulate progress in organization models while also providing an objective comparison among different approaches. The problem they propose is the simultaneous transcription of three speakers – choosing three because it guarantees that the average SNR will be below zero (i.e. there is less energy in one target voice than in the combination of the two competing voices), and because they feel that the two-speaker problem has been adequately solved. They propose that the objective assessment of this problem be based on a relaxed form of the word error rate used in speech recognition. This "challenge problem" is interesting because it will clearly reward the integration of scene analysis with speech recognition systems, although its focus on speech may bypass the issues of 'environmental sound' recognition that certain researchers see as more fundamental (Ellis, 1996).

Degree of perceptual plausibility

An important issue is the question of what makes computational auditory scene analysis systems distinctive when considered in relation to all possible approaches to sound mixture understanding.

The simple answer, that CASA systems are really models of processes thought to operate in listeners, disguises the widely differing opinions over the degree to which such correspondence is required or indeed possible. For some, the distinctive aspect of CASA systems is provided at an abstract functional level, in terms of principles of organization. For others, greater importance is placed on finding a neurophysiologically plausible mechanism. Intermediate between these two are perhaps the bulk of existing systems which aim to model both the relatively well-understood processing of the cochlea and the psychoacoustical manifestations of grouping. These activities all represent valid approaches to modeling auditory organization in listeners; the multi-level perspective detailed in section 1 accommodates them all.

As will be clear from the organization of this chapter, those systems which hope to fully explain arbitrary auditory scenes are forced to adopt the most abstract of connections with what is currently known of auditory organization in listeners. In the extreme, some such systems abandon even the assumed frequency selectivity of the cochlea for an initial narrowband fast Fourier transform (FFT) representation (e.g. Klassner, 1996).

Adapting to context and handling ambiguity

A single fragment can serve widely differing roles depending on its surroundings and other predispositions of the interpreter. Auditory organization models must ultimately include a stage of processing that varies according to some notion of context, but there is a wide range of practice in where this stage is placed. Ambiguous signals, whose correct interpretive context is not immediately clear, form an interesting test of context-adaptation.

Double-vowel identification models may have a simple processing sequence with no adaptation or feedback. However, once the time dimension is incorporated, the organization of the acoustic information at each instant will depend on the immediately preceding context. At the very least, the top-level groupings must reflect the accumulation of grouping cues between the different sound elements generated by the lower levels of processing, as in Cooke (1991/1993) and Mellinger (1991).

Other systems have intermediate representations, which, for an identical signal, can vary in response to contextual factors. In Weintraub (1985), these factors are the inferred presence of one or two voiced or unvoiced speakers, which determines how many pitches will be extracted and how their associated spectra will be derived. The system of Ellis (1996) is concerned with signals that may lack any periodicity cues, in which case the division of energy into representational units can only be made according to the prevailing scene interpretation.

The IPUS approach of Klassner (1996) incorporates an even greater degree of adaptation by extending the influence of abstract hypotheses right down to the numerical signal processing. According to their criterion of finding the most efficient and appropriate processing for each particular situation, the internal representation of the same signal – even when interpreted as the same object – may vary considerably depending on the other signals from which it had to be distinguished during analysis.

Greater degrees of context-adaptation imply more sophisticated approaches to ambiguity. The rigid signal models and powerful signal processing of Nakatani *et al.* (1998) permit each signal frame to be incorporated into the representation as soon as it is acquired, subject only to pruning of spurious creations. Other systems can delay making grouping decisions for newly-detected energy to allow the accumulation of disambiguating information: In Mellinger (1991), the delay is a fixed latency before a new harmonic as assigned to a cluster. The systems of Cooke (1991/1993) and Brown (1992) operated in two passes, with the grouping decisions made upon the intermediate elements only when they were completely formed, and all information was available. Weintraub (1985) had

a different two-pass structure, with the voice extraction depending on the overall best path from the initial dynamic-programming double-voice-state determination.

Rather than waiting for a unique solution to appear, some systems handle ambiguity by pursuing multiple alternative hypotheses (Ellis, 1996; Klassner, 1996; Godsmark & Brown, 1997). Although this approach is computationally expensive, it perhaps resembles listeners by maintaining a set of 'current beliefs' for a partially-observed signal; in real-world situations, one may not have the luxury of waiting for signal to end before commencing analysis. Listeners' interpretation of complex signals might be best understood via the incremental influence of each additional signal cue (as in the alternating noise bands of figure 5); ultimately, a correct understanding of human sound organization will probably include a combination of deferral, alternate hypotheses and hypothesis revision. These issues are also discussed in Cooke & Brown (1994).

Representing and employing constraints

Since the problem of separating one signal into multiple subcomponents has, in its simplest form, infinitely many solutions, the problem of auditory scene analysis may be viewed as defining and applying suitable constraints to choose a preferred alternative. The nature of these constraints, and the ways in which they are encoded and applied, forms a final axis on which to distinguish between the computational models.

Each of the cues in the summary of Table 2 corresponds to a certain constraint, i.e. an assumption of restrictions on the form of sound emitted by real-world sources. Thus the cue of harmonicity arises because many sound sources generate matched periodic modulation across wide frequency ranges, and the consequent constraint is that frequency bands exhibiting matched modulation patterns should be regarded as carrying energy from a single source.

In a system such as Brown (1992) which relies upon them, cues such as harmonicity and synchronized onset are directly expressed in the intermediate representation, and thus the 'knowledge' of the constraint is implicit in the computational procedure rather than being explicitly represented. By contrast, many perceptually important constraints – such as characteristic patterns of an individual's native tongue – are more arbitrary, and must be acquired and recalled, rather than simply computed. This is seen in the templates of Klassner (1996), which allow his system to have a somewhat abstracted idea of what, for instance, a telephone ring or a hairdryer sounds like. The system then uses the constraint that any scene must be explained in terms of known objects as a way to overcome the intrinsic uncertainty of a complex mixture.

Although symbolic systems with explicit representation of knowledge provide a natural way for researchers to implement their intentions, connectionist systems such as neural oscillator models provide an alternative approach to capturing and using knowledge, as for instance in the vowel-recognition oscillator network of Liu *et al*. (1994), where the constraints of typical vowel spectra were both represented and applied through the pattern of interconnections between the layers of 'neurons'.

One glaring difference between computational models and real listeners is the ability of the latter to acquire many of their constraints simply through exposure to the world. Ultimately, computer models must exhibit this kind of learning, but our current ignorance even as to the nature of these constraints puts such a system some way into the future.

Attention

When we talk of auditory scene analysis in listeners, we imagine them able to pick out any one of several sound sources in a mixture – but only *one* at any given time. Few models can be said to truly

choose from among the partially-processed sources the one which will be treated as the target to the exclusion of the others. Auditory scene analysis in listeners may well provide the most flexible means to extract a single target from a mixture, but it is possible that if no model of attentional focus is available, these benefits will not be easily realized for problems such as robust automatic speech recognition. This too is a topic for which good computational approaches have yet to be demonstrated.

Conclusion

It has taken three decades for auditory organization to develop from a problem that few researchers recognized to a fertile area of computational modeling. In this chapter we have surveyed both the experimental results that inform us about how listeners handle complex sound scenes, and the wide range of modeling efforts inspired by those results. The increasingly credible promise of benefits, for example in automatic speech recognition, serve to attract more attention to the field, yet at the same time our improving understanding and knowledge of human performance continually reveals new subtleties and capabilities yet to be modeled.

Resources for auditory scene analysis

In addition to Bregman's (1990) book, useful reviews of auditory organization can be found in Darwin & Culling (1990), Darwin & Carlyon (1995), Moore (1997, ch. 7) and Handel (1989). In addition, Volume 336 (1992) of the Philosophical Transactions of the Royal Society of London, Series B is devoted to the psychophysics of concurrent sound perception.

In 1995, the first international conference specifically concerned with computational models of auditory scene analysis processes was held in Montreal as a research workshop associated with the International Joint Conference on Artificial Intelligence. The proceedings of that meeting (Montréal, 1995) provide an illustrative cross-section of the diverse approaches to CASA which now prevail. A second CASA Workshop (Nagoya, 1997) documents further recent advances in this area. A special issue of the Speech Communication journal based on that meeting is due to be published in early 1999.

Other computational perspectives can be found in Cooke & Brown (1994), Summerfield & Culling (1995), Duda (1994), Bregman (1995) and Slaney (1995).

Demonstrations: A CD containing many audio examples demonstrating principle governing auditory scene analysis (Bregman & Ahad (1995) *Demonstrations of auditory scene analysis*; the CD can be ordered from The MIT Press, 55 Hayward Street, Cambridge, MA 02142, USA).

Corpora: To date, computational auditory scene analysis has not required corpora of the scale typically used in automatic speech recognition. Existing speech and noise corpora have been used to create acoustic mixtures suitable for computational auditory scene analysis. For instance, the NOISEX database (Varga *et al.*, 1992) provides a limited set of noise signals. Corpora produced by post-hoc signal combination are less than ideal, and demonstrate none of the conversational effects or compensations which occur in real spoken communication. Two corpora of conversational speech which address this limitation are available. The Map Task corpus (Thompson *et al.*, 1993) provides recordings of several two-person conversations and contains a limited amount of overlapping speech. The ShATR (Sheffield-ATR) corpus (Karlsen *et al.*, 1998), designed specifically for research in computational auditory scene analysis, involves five participants solving two crossword puzzles in pairs (the fifth person acts as a hint-giver). This task generates overlapped speech for nearly 40% of the corpus duration. Eight microphones provides simultaneous digital recordings from a binaurally-wired mannikin, an omnidirectional pressure zone mike and 5 close-talking

microphones, one for each participant. This corpus is available on CDROM; for more information, see the URL below.

More information is available on these databases at the following web addresses:

NOISEX: http://svr-www.eng.cam.ac.uk/comp.speech/Section1/Data/noisex.html

Map Task: http://www.hcrc.ed.ac.uk/dialogue/maptask.html

ShATR: http://www.dcs.shef.ac.uk/research/groups/spandh/pr/ShATR/ShATR.html

A collection of examples and other links relevant to this chapter is available at the following address, which we intend to maintain for the foreseeable future:

http://www.icsi.berkeley.edu/audorg

ACKNOWLEDGEMENTS

Guy Brown, Stuart Cunningham, Phil Green and Steve Greenberg provided useful comments on earlier drafts.

REFERENCES

Anstis, S. & Saida, S. (1985), Adaptation to auditory streaming of frequency modulated tones, Journal of Experimental Psychology: Human Perception and Performance, 11, 257-271.

Assmann, P.F. & Summerfield, Q. (1990), Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies, Journal of the Acoustical Society of America, 88(2), 680-697.

Assmann, P.F. & Summerfield, Q. (1994), The contribution of waveform interactions to the perception of concurrent vowels, Journal of the Acoustical Society of America, 95(1), 471-484.

Assmann, P.F. & Summerfield, Q. (in press), The perception of speech under adverse acoustic conditions, in: The Auditory Basis of Speech Perception (eds: S. Greenberg & W. Ainsworth), Springer.

Bailey, P.J., Summerfield, A. & Dorman, M. On the identification of sine-wave analogues of certain speech sounds. Report no: SR-51/52, Haskins Labs, 1977.

Beauvois, M.W. & Meddis, R. (1991), A computer model of auditory stream segregation, Quarterly Journal of Experimental Psychology, 43A(3), 517-541.

Beauvois, M.W. & Meddis, R. (1996), Computer simulation of auditory stream segregation in alternating-tone sequences, Journal of the Acoustical Society of America 99 (4), Pt. 1, 2270-2280.

Bell, A.J. & Sejnowski, T.J. (1995), An information maximisation approach to blind separation and blind deconvolution, Neural Computation, 7(6), 1129-1159.

Berthommier & Meyer (1997), A model of double-vowel segregation with AM map and without F0 tracking, Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Nagoya.

Bird, J. & Darwin, C.J. (1997), Effects of a difference in fundamental frequency in separating two sentences, Proceedings of the 11th International Conference on Hearing, Grantham, UK.

Bodden, M. (1995), Binaural Modeling and Auditory Scene Analysis, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk.

Bregman, A.S. (1978), Auditory streaming is cumulative, Journal of Experimental Psychology: Human Perception and Performance, 4, 380-387.

Bregman, A.S. (1984), Auditory scene analysis, Proceedings of the 7th International Conference on Pattern Recognition, Silver Spring MD.

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

Bregman, A.S. (1995), Use of psychological data in building ASA models, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk.

Bregman, A.S., Abramson, J., Doehring, P. & Darwin, C.J. (1985), Spectral integration based on common amplitude modulation, Perception & Psychophysics, 37, 483-493.

Bregman, A.S. & Campbell, J. (1971), Primary auditory stream segregation and perception of order in rapid sequences of tones, Journal of Experimental Psychology, 89, 244-249.

Bregman, A.S. & Levitan, R. (1983), Stream segregation based on fundamental frequency and spectral peak. 1: Effects of shaping by filters, Unpublished manuscript, Psychology Department, McGill University.

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

Bregman, A.S. & Rudnicky, A. (1975), Auditory segregation: Stream or streams?, Journal of Experimental Psychology: Human Perception and Performance, 1, 263-267.

Broadbent, D.E. & Ladefoged, P. (1957), On the fusion of sounds reaching different sense organs, Journal of the Acoustical Society of America, 29, 708-710.

Brokx, J.P.L. & Nooteboom, S.G. (1982), Intonation and the perceptual separation of simultaneous voices, Journal of Phonetics, 10, 23-36.

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

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

Brown, G.J. & Cooke, M.P. (1995), A Neural Oscillator Model of Primitive Audio Grouping, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk.

Brown, G.J. & Cooke, M.P. (1998), Temporal synchronization in a neural oscillator model of primitive auditory stream segregation, in: Readings in Computational Auditory Scene Analysis (eds: D. Rosenthal & H. Okuno), Lawrence Erlbaum.

Brown, G.J., Cooke, M.P. & Mousset, E. (1996), Are neural oscillations the substrate of auditory grouping? ESCA Tutorial and Workshop on the Auditory Basis of Speech Perception, Keele University, July 15-19.

Brown G.J. & Wang, D.L. (1997), Modelling the perceptual segregation of double vowels with a network of neural oscillators, Neural Networks, 10, 1547-1558.

Buus, S. (1985), Release from masking caused by envelope fluctuations, Journal of the Acoustical Society of America, 78(6), 1958-1965.

Buus, S. & Pan, C. (1994), Discrimination of envelope frequency in one spectral region in the presence of modulation in another, Journal of the Acoustical Society of America, 96(3), 1445-1457.

Carver, N. & Lesser, V. (1992), Blackboard systems for knowledge-based signal understanding, in: Symbolic and knowledge-based signal processing (eds: A.V. Oppenheim & S.H. Nawab), Prentice Hall.

Cherry, E.C. (1953), Some experiments on the recognition of speech with one and with two ears, Journal of the Acoustical Society of America, 25, 975-979.

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.

de Cheveigné, A. (1997), Concurrent vowel identification III: A neural model of harmonic interference cancellation, Journal of the Acoustical Society of America, 101, 2857-2865.

de Cheveigné, A., McAdams, S., Laroche, J. & Rosenberg, M. (1995), Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement, Journal of the Acoustical Society of America, 97(6), 3736-3748.

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

Ciocca, V. & Darwin, C.J. (1993), Effects of onset asynchrony on pitch perception: Adaptation or grouping?, Journal of the Acoustical Society of America, 93(5), 2870-2878.

Clarkson, M.G. & Clifton, R.K. (1995), Infants' pitch perception: Inharmonic tonal complexes, Journal of the Acoustical Society of America, 98(3), 1372-1379.

Clarkson, M.G. & Rogers, E.C. (1995), Infants require low-frequency energy to hear the pitch of the missing fundamental, Journal of the Acoustical Society of America, 98(1), 148-154.

Cole, R.A. & Scott, B. (1973), Perception of temporal order in speech: The role of vowel transitions, Canadian Journal of Psychology, 27, 441-449.

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

Cooke, M.P. (1996), Auditory organisation and speech perception: Arguments for an integrated computational theory, ESCA Workshop on the Auditory Basis of Speech Perception, Keele University.

Cooke, M.P. & Brown, G.J. (1994), Separating simultaneous sound sources: Issues, challenges and models, in: Fundamentals of speech synthesis and speech recognition, (ed: E. Keller), J. Wiley, 295-312.

Cooke, M.P., Brown, G.J., Crawford, M.D & Green, P.D. (1993), Computational auditory scene analysis: Listening to several things at once, Endeavour, 17, 186-190.

Cooke, M.P. & Green, P.D. (in press), Recognizing occluded speech, in: Listening to speech: An auditory perspective (eds: S. Greenberg & W. Ainsworth), 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 Speech and Language Processing, Yokohama, 1555-1558.

Crick, F. (1984), Function of the thalamic reticular complex: The searchlight hypothesis, Proceedings of the National Academy of Sciences, 81, 4586-90.

Culling, J.F. & Darwin, C.J. (1993), Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0, Journal of the Acoustical Society of America, 93(6), 3454-3467.

Culling, J.F. & Darwin, C.J. (1994), Perceptual and computational separation of simultaneous vowels: Cues arising from low-frequency beating, Journal of the Acoustical Society of America, 95(3), 1559-1569.

Culling, J.F., Summerfield, Q. & Marshall, D.H. (1994), Effects of simulated reverberation on the use of binaural cues and fundamental-frequency differences for separating concurrent vowels, Speech Communication, 14, 71-95.

Culling, J.F. & Summerfield, Q. (1995a), The role of frequency modulation in the perceptual segregation of concurrent vowels, Journal of the Acoustical Society of America, 98(2), 837-846.

Culling, J.F. & Summerfield, Q. (1995b), Perceptual segregation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay, Journal of the Acoustical Society of America, 98(2), 785-797.

Cutting, J.E. (1976), Auditory and linguistic processes in speech perception: Inferences from six fusions in dichotic listening, Psychological Review, 83, 114-140.

Darwin, C.J. (1981), Perceptual grouping of speech components different in fundamental frequency and onset-time, Quarterly Journal of Experimental Psychology, 3A, 185-207.

Darwin, C.J. (1984), Perceiving vowels in the presence of another sound: Constraints on formant perception, Journal of the Acoustical Society of America, 76(6), 1636-1647.

Darwin, C. J. & Bethell-Fox, C. E. (1977), Pitch continuity and speech source attribution, Journal of Experimental Psychology: Human Perception and Performance,3, 665-672.

Darwin, C.J. & Carlyon, R.P. (1995), Auditory Grouping, in: The Handbook of Perception and Cognition, Vol 6, Hearing (ed: B.C.J. Moore), Academic Press, 387-424.

Darwin, C.J. & Ciocca, V. (1992), Grouping in pitch perception: Effects of onset asynchrony and ear of presentation of a mistuned component, Journal of the Acoustical Society of America, 91, 3381-3390.

Darwin, C.J. & Culling, J.F. (1990), Speech perception seen through the ear, Speech Communication, 9,

Darwin, C.J. & Gardner, R.B. (1986), Mistuning a harmonic of a vowel: Grouping and phase effects on vowel quality, Journal of the Acoustical Society of America, 79(3), 838-845.

Darwin, C.J., Hukin, R.W. & Al-Khatib, B.Y. (1995), Grouping in pitch perception: evidence for sequential constraints, Journal of the Acoustical Society of America, 98(2), 880-885.

Darwin, C.J., Pattison, H. & Gardner, R.B. (1989) Vowel quality changes produced by surrounding tone sequences, Perception and Psychophysics, 45, 333-342.

Deutsch, D. (1975), Two-channel listening to musical scales, Journal of the Acoustical Society of America, 57, 1156-1160.

Denbigh, P.N. & Zhao, J. (1992), Pitch extraction and the separation of overlapping speech, Speech Communication, 11, 119-125.

Dowling, W.J. (1973), Rhythmic groups and subjective chunks in memory for melodies, Perception & Psychophysics, 14, 37-40.

Duda, R.O. (1994) Connectionist models for auditory scene analysis, in: Advances in Neural Information Processing Systems 6 (eds: J.D. Cowan, G. Tesauro & J. Alspector), Morgan Kaufmann.

Duifhuis, H., Willems, L.F. & Sluyter, R.J. (1982), Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception, Journal of the Acoustical Society of America, 71, 1568-1580.

Durlach, N.I. (1963), Equalization and cancellation theory of binaural masking-level differences, Journal of the Acoustical Society of America, 35, 1206-1218.

Ellis, D.P.W. (1993), Hierarchic models of hearing for sound separation and reconstruction, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk.

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. (1997a), 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. (1997b), Computational auditory scene analysis exploiting speech-recognition knowledge, Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, Mohonk.

Ellis, D.P.W. (in press), Modeling the auditory organization of speech - a summary and some comments, in: Listening to speech: An auditory perspective (eds: S. Greenberg & W. Ainsworth), Oxford University Press.

Fletcher, H. (1953), Speech and Hearing in Communication, Van Nostrand.

Gardner, R.B., Gaskill, S.A. & Darwin, C.J. (1989), Perceptual grouping of formants with static and dynamic differences in fundamental frequency, Journal of the Acoustical Society of America, 85(3), 1329-1337.

Gardner, R.B. & Darwin, C.J. (1986), Grouping of vowel harmonics by frequency modulation: Absence of effects on phonemic categorization, Perception & Psychophysics, 40(3), 183-187.

Gibson, J.J. (1966), The senses considered as perceptual systems, Houghton Mifflin.

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.

Gong, Y. (1995), Speech recognition in noisy environments: A survey, Speech Communication, 16, 261-291.

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

Green, P.D., Cooke, M.P. & Crawford, M.D. (1995), Auditory scene analysis and hidden Markov model recognition of speech in noise, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 401-404.

Green, P.D., Brown, G.J., Cooke, M.P., Crawford, M.D. & Simons, A.J.H. (1990), Bridging the gap between signals and symbols in speech recognition, in: Advances in speech, hearing and language processing (ed: W.A. Ainsworth), JAI Press.

Green, P.D. & Grace, P.J. (1981), A descriptive approach to computer speech understanding, Proceedings of the Institute of Acoustics, 261-264.

Green, P.D. & Wood, A.R. (1986), A representational approach to knowledge-based acoustic-phonetic processing in speech recognition, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Tokyo, paper 23.4.

Guttman, N. & Julesz, B. (1963), Lower limits of auditory periodicity analysis, Journal of the Acoustical Society of America 35, 610.

Grose, J.H. & Hall, J.W. (1992), Comodulation masking release for speech stimuli, Journal of the Acoustical Society of America, 91, 1042-1050.

Grose, J.H. & Hall, J.W. (1993), Comodulation masking release: Is comodulation sufficient?, Journal of the Acoustical Society of America, 93(5), 2896-2902.

Hall, J.W. & Grose, J.H. (1990), Comodulation masking release and auditory grouping, Journal of the Acoustical Society of America, 88(1), 119-125.

Hall, J.W. & Grose, J.H. (1991), Some effects of auditory grouping factors on modulation detection interference (MDI), Journal of the Acoustical Society of America, 90(6), 3028-3035.

Hall, J.W., Haggard, M.P. & Fernandes, M.A. (1984), Detection in noise by spectro-temporal pattern analysis, Journal of the Acoustical Society of America, 76, 50-56.

Handel, S. (1989), Listening: An Introduction to the Perception of Auditory Events, MIT Press.

Hartman, W.M. & Johnson, D. (1991),Stream segregation and peripheral channeling, Music Perception, 9(2), 155-184.

Hill, N.J. & Darwin, C.J. (1993), Effects of onset asynchrony and of mistuning on the lateralization of a pure tone embedded in a harmonic complex, Journal of the Acoustical Society of America, 93, 2307-2308.

Houtgast, T. (1971), Psychophysical evidence for lateral inhibition in hearing, Journal of the Acoustical Society of America, 51(6), 1885-1894.

Hukin, R.W. & Darwin, C.J. (1995), Effects of contralateral presentation and of interaural time differences in segregating a harmonic from a vowel, Journal of the Acoustical Society of America, 98(3), 1380-1387.

Howard-Jones, P.A. & Rosen, S. (1993), Uncomodulated glimpsing in "checkerboard" noise, Journal of the Acoustical Society of America, 93(5), 2915-2922.

Jeffress, L.A. (1948), A place theory of sound localization, Journal of Comparative and Physiological Psychology, 41, 35-39.

Jones, M.R. (1976), Time, our lost dimension: Toward a new theory of perception, attention and memory, Psychological Review, 83, 323-355.

Junqua, J.-C. & Haton, J.P. (1996), Robustness in Automatic Speech Recognition, Kluwer Academic Publishers.

Kaernbach, C. (1992), On the consistency of tapping to repeated noise, Journal of the Acoustical Society of America 92, 788-793.

Karlsen, B.L., Brown, G.J., Cooke, M.P., Crawford, M.D., Green, P.D. & Renals, S.J. (1998), Analysis of a simultaneous-speaker sound corpus, in: Readings in Computational Auditory Scene Analysis (eds: D. Rosenthal & H. Okuno), Lawrence Erlbaum.

Kashino, K. & Tanaka, H. (1993). A sound source separation system with the ability of automatic tone modeling, Proceedings of the 1993 International Computer Music Conference, 248-255.

Kashino, K., Nakadai, K., Kinoshita, T. & Tanaka, H. (1998), Application of the Bayesian probability network to music scene analysis, in: Readings in Computational Auditory Scene Analysis (eds: D. Rosenthal & H. Okuno), Lawrence Erlbaum.

Kidd, G., Mason, C.R., Deliwala, P.S., Woods, W.S. & Colburn, H.S. (1994), Reducing informational masking by sound segregation, Journal of the Acoustical Society of America, 95(6), 3475-3480.

Kidd, G., Mason, C.R. & Dai, H. (1995), Discriminating coherence in spectro-temporal patterns, Journal of the Acoustical Society of America, 97(6), 3782-3789.

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

Lea, A. (1992), Auditory modeling of vowel perception, unpublished doctoral thesis, University of Nottingham.

Liberman, A.M. (1982), On the finding that speech is special, American Psychologist, 37(2), 148-167, reprinted in: Handbook of Cognitive Neuroscience (ed: M.S. Gazzaniga), Plenum Press, 169-197 (1984).

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

Lippmann, R.P. (1996), Accurate consonant perception without mid-frequency speech energy, IEEE Transactions on Speech and Audio Processing, 4(1), 66-69.

Lippmann, R.P.(1997), Speech recognition by machines and human, Speech Communication, 22(1), 1-16.

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.

Lyon, R.F. (1983), A computational model of binaural localization and separation, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Boston, 1148-1151.

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

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

McAdams, S. (1984), Spectral fusion, spectral parsing and the formation of auditory images, unpublished doctoral dissertation, Stanford University.

McCabe, S.L. & Denham, M.J. (1997), A model of auditory streaming, Journal of the Acoustical Society of America, 101(3), 1611-1621.

McKeown, J.D. and Patterson, R.D. (1995), The time course of auditory segregation: Concurrent vowels that vary in duration Journal of the Acoustical Society of America, 98(4), 1866-1877.

Meddis, R. & Hewitt, M.J. (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(6), 2866-2882.

Meddis, R. & Hewitt, M.J. (1992), Modelling the identification of concurrent vowels with different fundamental frequencies, Journal of the Acoustical Society of America, 91(1), 233-245.

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

Miller, G.A. & Heise, G.A. (1950), The trill threshold, Journal of the Acoustical Society of America, 22, 637-638.

Montreal (1995) Proceedings of the 1st Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Montreal.

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

Moore, B.C.J., Glasberg, B.R. & 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.

Moore, B.C.J., Glasberg, B.R. & Peters, R.W. (1986), Thresholds for hearing mistuned partials as separate tones in harmonic complexes, Journal of the Acoustical Society of America, 80, 479-483.

Moore, B.C.J. & Shailer, M.J. (1992), Modulation discrimination interference and auditory grouping, Philosophical Transactions of the Royal Society London B, 336, 339-346.

Moore, D.R. (1987), Physiology of the higher auditory system, British Medical Bulletin, 43(4), 856-870.

Morris, A.C., Cooke, M.P., Green, P.D. (1998), Some solutions to the missing feature problem in data classification, with application to noise robust ASR, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Seattle.

Nagoya (1997) Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Nagoya.

Nakatani, T., Kashino, K. & Okuno, H. G. (1997), Integration of speech stream and music stream segregation based on ontology, Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Nagoya.

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

Nawab, S.H. & Lesser, V. (1992), Integrated processing and understanding of signals, in: Symbolic and knowledge-based signal processing (eds: A.V. Oppenheim & S.H. Nawab), Prentice Hall.

Nii, H.P. (1986), Blackboard systems part two: Blackboard application systems from a knowledge engineering perspective, The AI Magazine, 7(3), 82-106.

van Noorden, L.P.A.S. (1975), Temporal coherence in the perception of tone sequences, Ph.D. dissertation, Eindhoven University of Technology.

Okuno, H.G., Nakatani, T., Kawabata, T. (1997), Challenge problem: Understanding three simultaneous speakers, Proceedings of the 2nd Workshop on Computational Auditory Scene Analysis, Int. Joint Conf. Artificial Intelligence, Nagoya.

Palmer, A.R. (1990), The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibres, Journal of the Acoustical Society of America, 88(3), 1412-1426.

Parsons, T.W. (1976), Separation of speech from interfering speech by means of harmonic selection, Journal of the Acoustical Society of America, 60(4), 911-918.

Patterson, R.D. (1987), A pulse ribbon model of monaural phase perception, Journal of the Acoustical Society of America, 82(5), 1560-1586.

Pierce, J.R. (1983), The science of musical sound. Freeman.

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

Remez, R.E., Rubin, P.E., Pisoni, D.B. & Carrell, T.D. (1981), Speech perception without traditional speech cues, Science, 212, 947-950.

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.

Richards, V. (1987), Monaural envelope correlation perception, Journal of the Acoustical Society of America, 82(5), 1621-1630.

Rogers, W.L. & Bregman, A.S. (1993), An experimental evaluation of three theories of auditory stream segregation, Perception and Psychophysics, 53(2), 179-189.

Saberi, K. & Hafter, E.R. (1995), A common neural code for frequency and amplitude modulated sounds. Nature, 374 (6522), 537-539.

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

Schooneveldt, G.P. & Moore, B.C.J. (1989), Comodulation masking release (CMR) as a function of masker bandwidth, modulator bandwidth, and signal duration, Journal of the Acoustical Society of America, 85(1), 273-281.

Shackleton, T.M. & Meddis, R. (1992), The role of interaural time difference and fundamental frequency difference in the identification of concurrent vowel pairs, Journal of the Acoustical Society of America, 91, 3579-3581.

Slaney, M. (1998), A critique of pure audition, in: Readings in Computational Auditory Scene Analysis (eds: D. Rosenthal & H. Okuno), Lawrence Erlbaum.

Steeneken, H.J.M. (1992), On measuring and predicting speech intelligibility, unpublished Ph.D. thesis, University of Amsterdam.

Stellmack, M.A. & Dye, R.H. (1993), The combination of interaural information across frequencies: The effects of number and spacing of components, onset asynchrony and harmonicity, Journal of the Acoustical Society of America, 93(5), 2933-2947.

Stubbs & Summerfield (1988), Evaluation of 2 voice-separation algorithms using normal-hearing and hearing-impaired listeners, JASA, 84(4), 1236-1249.

Summerfield, Q. & Culling, J.F. (1992), Auditory segregation of competing voices: absence of effects of FM or AM coherence, Philosophical Transactions of the Royal Society London B, 336, 415-22.

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, (ed: E. Keller), J. Wiley.

Thompson, H., Bard, E., Anderson, A. & Doherty-Sneddon, G. (1993), The HCRC Map Task Corpus: A Natural Spoken Dialogue Corpus, Proceedings of the International Symposium on Spoken Dialogue, Tokyo, 33-36.

Todd, N.P.M. (1996) An auditory cortical theory of primitive auditory grouping, Network: Computation in Neural Systems, 7, 349-356.

Varga, A.P., Steeneken, H.J.M., Tomlinson, M. & Jones, D. (1992), The NOISEX-92 study on the effect of additive noise on automatic speech recognition, in: Technical Report, Speech Research Unit, Defence Research Agency, Malvern, U.K.

Varga, A. P. & Moore, R. K. (1990), Hidden markov model decomposition of speech and noise, Proceedings of the International Conference on Acoustics, Speech and Signal Processing, 845-848.

Varin, L. & Berthommier, F. (1997), A probabilistic model of double-vowel segregation, Proceedings of Eurospeech 97, Rhodes, 2791-2794.

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

Warren, R.M. (1970), Perceptual restoration of missing speech sounds, Science, 167, 392-393.

Warren, R.M. & Bashford, J.A. (1976), Auditory contralateral induction: an early stage in binaural processing, Perception & Psychophysics, 20(5), 380-386.

Warren, R.M., Bashford, J.A. & Gardner, D.A. (1990), Tweaking the lexicon: Organization of vowel sequences into words, Perception & Psychophysics, 47(5), 423-432.

Warren, R.M., Hainsworth, K.R., Brubaker, B.S., Bashford, J.A. & Healy, E.W. (1997), Spectral restoration of speech: Intelligibility is increased by inserting noise in spectral gaps, Perception & psychophysics, 59(2), 275-283.

Warren, R.M., Healy, E.W & Chalikia, M.H. (1996), The vowel-sequence illusion: Intrasubject stability and intersubject agreement of syllabic forms, Journal of the Acoustical Society of America, 100(4), 2452-2461.

Warren, R.M., Obusek, C.J. & Ackroff, J.M. (1972). Auditory induction: perceptual synthesis of absent sounds. Science, 176, 1149-1151.

Warren R.M. & Warren, R.P. (1970), Auditory illusions and confusions, Scientific American, 223(12), 30-36.

Watkins, A.J. (1991), Central, auditory mechanisms of perceptual compensation for spectral-envelope distortion, Journal of the Acoustical Society of America, 90(6), 2942-2955.

Weintraub, M. (1985), A theory and computational model of auditory monaural sound separation, unpublished doctoral dissertation, Department of Electrical Engineering, Stanford University.

Woods, W.S. & Colburn, H.S. (1992), Test of a model of auditory object formation using intensity and interaural time difference discrimination, Journal of the Acoustical Society of America, 91(5), 2894-2902.

Yost, W.A. & Sheft, S. (1989), Across-critical-band processing of amplitude modulated tones, Journal of the Acoustical Society of America, 85, 848-857.

Zakarauskas, P. & Cynader, M.S., (1993), A computational theory of spectral cue localization, Journal of the Acoustical Society of America, 94(3), 1323-1331.