PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

Phil. Trans. R. Soc. B (2008) **363**, 1011–1021
doi:10.1098/rstb.2007.2156
Published online 7 September 2007

Listening to speech in the presence of other sounds

C. J. Darwin*

Department of Psychology, University of Sussex, Brighton BN1 9QG, UK

Although most research on the perception of speech has been conducted with speech presented without any competing sounds, we almost always listen to speech against a background of other sounds which we are adept at ignoring. Nevertheless, such additional irrelevant sounds can cause severe problems for speech recognition algorithms and for the hard of hearing as well as posing a challenge to theories of speech perception. A variety of different problems are created by the presence of additional sound sources: detection of features that are partially masked, allocation of detected features to the appropriate sound sources and recognition of sounds on the basis of partial information. The separation of sounds is arousing substantial attention in psychoacoustics and in computer science. An effective solution to the problem of separating sounds would have important practical applications.

Keywords: speech perception; cocktail-party problem; auditory grouping; auditory scene analysis; auditory perception; auditory localization

1. PERCEPTION OF MULTIPLE SOUND SOURCES

Despite early work by Cherry (1953) and Cherry & Taylor (1954) drawing attention to the 'cocktail-party problem', all of the major strands of research on the perception of speech have largely ignored the problems caused by the fact that most of the time we listen to speech against a background of, often intense, irrelevant sounds. For example, acoustic phonetics (see Diehl 2008) has concentrated on mapping the relationship between acoustic features (like formants) and linguistic categories (such as phonemic features or syllables), but has been much less concerned with how the acoustic features produced by the required talker might be extracted from a sound mixture. Similarly, psycholinguists have studied the processing of connected speech but with little attention to the problem of how we follow the speech of a particular talker in a mixture of sounds. Such neglect is understandable given the complexity of the problem posed by recognizing the speech of even an isolated talker, yet the presence of additional sound sources—especially the speech of another talker—can have significant practical and theoretical implications.

The practical application of speech recognition algorithms is limited by the problems raised by additional sounds. Speech recognition algorithms, though increasingly successful in a naturally quiet environment (or one that has been artificially quietened with a noise-cancelling microphone), fail in the presence of background sound which does not trouble a human listener (Lippmann 1997; Sroka & Braida 2005). Additional sounds create problems for speech recognition in a number of ways. Simple strategies for automatic speech recognition, like spectral template matching, can be seriously disrupted by additional sounds since the overall spectrum of the mixture differs substantially from that of the target sound. This disruption is especially marked for non-stationary masking sounds, which are difficult to remove by adaptive filtering. More analytic recognition strategies can also be disrupted since parts of a sound can be masked by a background sound, and the background sound can also provide additional acoustic features that need to be discounted.

Another practical consequence of the presence of additional sounds involves hearing-impaired listeners who typically find noisy environments disproportionately difficult for understanding speech, whether they wear a hearing aid or not (Killion 1997; Moore 1998). A major reason for noise causing additional problems for hearing-impaired listeners is that their auditory-filter bandwidths are wider than normal as a result of outer hair cell loss (see Moore 2008). The wider bandwidths lead to both spectral smearing of auditory features and reduced signal-to-noise ratio in the presence of background noise. Similar problems are encountered by users of cochlear implants, again owing to the limited frequency resolution available with implants (Clark 2003).

The problem of additional irrelevant sounds is not of course the one that is specific to speech perception; it applies to the recognition of any sound in a natural environment. Our conscious experience of sound is typically of a number of separate sources, each with their appropriate location, pitch and timbre, apparently unaltered by whatever sounds they occur with. These percepts are the result of a variety of remarkable processing strategies by the brain.

Two of the problems created by sound mixtures are detection and allocation. For speech masked by high-level noise, the main problem is detection—actually

*cjd@sussex.ac.uk

One contribution of 13 to a Theme Issue 'The perception of speech: from sound to meaning'.

hearing the component frequencies of the sound. For speech masked by, say, a single other talker, the additional problem of allocation arises: which detected components belong to which sound source. I deal with the latter problem extensively in §3. After both of these problems have been overcome, an additional problem arises—how to recognize a particular category of speech sound on the basis of only partial information (Cooke *et al.* 2001).

2. LISTENING TO SPEECH IN NOISE

To return to Cherry's cocktail party: Plomp (1977) has calculated that when everyone at a well-attended party talks at the same level, the speech of the attended talker at a distance of 0.7 m has a signal-to-noise (S/N) ratio of approximately 0 dB—the background is as intense as the target talker. Since speech fluctuates substantially in level, an average of 0 dB implies that many of the weaker sounds of the target speaker will be masked. Nonetheless, 0 dB is sufficient to give adequate intelligibility with normal pronunciation and redundancy for listeners with normal hearing (Miller 1947). With a steady masking sound in non-reverberant conditions, intelligibility can be well predicted using the speech intelligibility index (SII; ANSI 1997), which is calculated from a weighted sum of the contributions of different frequency bands according to their S/N ratio. If the speech environment is reverberant, the speech transmission index (STI; Houtgast & Steeneken 1973), which is based on the modulation transfer function, can be used to calculate the predicted intelligibility. However, neither the SII nor the STI has a principled basis for modelling the short-term properties of either speech or noise that are involved in perceptual separation, since these indices are based on the long-term average properties of the speech and noise (but see Rhebergen *et al.* 2005).

These measures all work best when the interfering noise is stationary, so that there are no pronounced temporal fluctuations in level or spectrum. But the speech of a single background talker is far from stationary. For masking sounds that are equated for their overall average level, speech is considerably more intelligible when in the presence of a different-sex talker than with a steady noise with the same overall level and spectral composition produced by adding together many different talkers (Miller 1947). Intermediate levels of intelligibility occur for noise that has a spectrum similar to the long-term average spectrum of speech but whose instantaneous amplitude fluctuates like that of speech (Duquesnoy 1983; Festen & Plomp 1990; Peters *et al.* 1998) or for a competing speaker who is identical to or of the same sex as the target speaker (Stubbs & Summerfield 1990; Drullman & Bronkhorst 2000). Such differences between different maskers are not only partly attributable to the spectral and temporal gaps in speech and in modulated noise (for more details see the review by Bronkhorst 2000), but also require an understanding of how the brain deals in general with the separation of simultaneous sound sources across frequency and time.

3. AUDITORY SCENE ANALYSIS

The general problem of how the brain is able to separate component sound sources in mixtures has received considerable attention from psychologists over the last 30 years under the general name of auditory scene analysis—the title of an influential book summarizing the field (Bregman 1990). Bregman's book makes the case for the brain using two types of information to group together sound components that have originated from a common source: heuristics based on general properties of sound sources, and schematic knowledge about specific sounds. General properties such as the common onset and (for periodic sounds) the harmonic relations between frequency components from a single source can help partition sounds from different sources that occur simultaneously; other properties, such as continuity of pitch, timbre, overall level and spatial location can help to track a single sound source across time. For a more recent review of such low-level auditory grouping, see Darwin & Carlyon (1995). Bregman also recognizes that, as well as general heuristics, the brain may employ schema-based grouping. Here, the knowledge about specific sounds that schemata contain can be used to select from a mixture those components that form a schematized sound. Schema-based selection might be particularly important in the perception of speech. For example, if two steady vowels are synthesized with the same fundamental frequency and played strictly simultaneously, listeners can still identify them to some extent. There are no general grouping cues to help them to separate the two sounds of the mixture, only the experience of hearing particular vowels, or perhaps more abstract knowledge about how vocal tracts can shape sound (Darwin 1984).

(a) *Auditory scene analysis and speech*

The relationship between auditory scene analysis and speech perception has been a contentious issue. One view builds on the idea that speech is perceived by special processing mechanisms (Liberman *et al.* 1967; Liberman 1982) which are quite separate from those involved in the perception of other sounds—the 'speech is special' view. The other view (Bregman 1990; Darwin 1991) is that all sound input is indiscriminately subject to low-level grouping mechanisms.

The former view maintains (Remez *et al.* 1981) that speech sounds are not the subject of general auditory scene analysis heuristics. Rather, speech perception pre-empts scene analysis, using speech schemata to cherry-pick from mixtures of sounds those components that can form speech (Whalen & Liberman 1987). One argument for this view is that general auditory heuristics could not put together into a single source the varied sounds that make up the speech stream, as it rapidly switches between voiced, aspirated and fricated sounds and silence. Another argument concerns the intelligibility of sine-wave speech. Sine-wave speech is synthesized by frequency modulating (FM) and amplitude modulating (AM) three sine waves so that they follow the frequencies and amplitudes of the first three formants (the resonant frequencies of the vocal tract—see Diehl 2008), producing the speech equivalent of a line drawing. Individually, each FM

sine-wave component sounds like a whistle that rises and falls in pitch; it does not sound like speech. But together, the three whistles gain, in addition to their whistliness, a speech-like quality that renders them, after a little exposure, moderately intelligible. Sine-wave speech not only lacks the harmonic structure that helps to bind together the voiced portions of normal speech, but the frequency movements of the three sine waves are largely uncorrelated; so, it is argued that the only factor which can perceptually integrate them is that their movement relates to the dynamics of vocal tract movement—a speech-specific, schematic constraint.

Various points have been made against this argument. First, it ignores the fact that the default condition for the auditory system is to treat everything as coming from a single source, and only to segregate different sources if the evidence is sufficient (Bregman 1978). Second, the computational analysis of two simultaneous sine-wave speech sentences indicates that the sine waves from a single talker can be grouped at least partially by low-level cues such as onset time (Barker & Cooke 1999). Third, even if the argument were correct, it is irrelevant (Darwin 1991). The scope for general auditory heuristics in the perception of sine-wave speech may be limited, but there are numerous examples where grouping by harmonicity (Darwin 1981), amplitude modulation or onset time (Darwin 1984) change the perception of speech sounds. These examples are unlikely to be due to speech-specific mechanisms, since the low-level grouping cues can lead to the percept of a less natural speech sound than pre-emptive schema-based grouping would have provided. Finally, it is a rather curious, if not perverse, argument to suppose that general auditory mechanisms which have gradually evolved to help organisms separate multiple sound sources should be ignored in the particularly difficult job of perceiving speech in a mixture of other sounds (Darwin 1991).

The related question of whether there are areas of the brain that respond selectively to speech sounds has been addressed by a number of imaging studies. For example, a recent review claims that 'speech perception emerges from the connectivity between (generic) auditory areas and … frontal lobe regions' (Price *et al.* 2005). For more details of this flourishing new area see the paper in this issue by Patterson & Johnsrude (2008).

(b) *Intermediate representations in speech*

Speech perception is remarkably resistant to distortion. For instance, speech remains moderately intelligible under extremes of high-pass or low-pass filtering, infinite peak clipping, or the addition of high levels of noise. It is also remarkably intelligible when its spectral information is reduced either to three FM sine waves, as described above for sine-wave speech, or to only four broad frequency channels each excited by a noise that is slowly amplitude modulated to match the energy distribution in those bands of the original speech (Shannon *et al.* 1995)—a transformation that is similar to that produced over a larger number of frequency channels by modern cochlear implants. More bizarrely, thanks to the automatic separation of the contributions to speech of the larynx (source) and the vocal tract (filter) made possible by linear-predictive coding (Atal & Hanauer 1971), the buzz and hiss of the larynx can be replaced by an arbitrary broadband signal—such as an orchestra—and the speech articulated by the moving vocal tract can then be heard as a ghostly modulation of the orchestral sound (Hunt *et al.* 1989).

Such resistance to distortion arises through speech being redundant at many levels, and the perceptual system being adept at perceiving sound that has been filtered and masked by its environmental context. At the acoustic–phonetic level, there are many sufficient acoustic cues to a phoneme but no necessary ones (see Diehl 2008). Moreover, in interpreting a particular sound, listeners are able to compensate for the way that a sound has been filtered in travelling from its source to the listener (Darwin 1990; Watkins & Makin 1994). The sheer variability of sounds that can be effectively perceived as speech makes it very difficult to discover what intermediate representations of sound are used in perceiving speech.

Although formants are used to describe the way that sounds are synthesized in speech perception experiments, and so have become a de facto acoustic description of at least the vocal portion of speech, their status as a perceptual entity is surprisingly vague. On the one hand, formant frequencies do seem to be perceptually salient: vowels change their perceived category when formant frequencies are changed, but usually only change their general timbre (e.g. becoming more or less muffled) when formant amplitudes are changed (Klatt 1985). For example, the amplitude of the second formant in a two-formant synthesis of a front vowel can be reduced by as much as 28 dB without the phonetic identity of the vowel being changed (Ainsworth & Miller 1972). On the other hand, listeners find it surprisingly hard to match single-formant sounds according to their formant frequency when they differ in fundamental frequency (Dissard & Darwin 2000, 2001) and when the harmonics close to the formant peak are resolved (see Moore 2008). In addition, formants are notoriously difficult to track automatically in a way that is reliable enough to be the sole metric for recognition, and so speech recognition algorithms tend to use spectral metrics that are more global than formant frequencies, such as cepstral coefficients (derived from a Fourier transform of the log power spectrum), despite the fact that formants are more resistant to the overall changes to the spectrum that occur in natural listening situations (Hunt 1987).

There is thus no consensus about the form of the auditory information that is used to access the brain's acoustic–phonetic knowledge in order to categorize the speech sounds that we hear. Whatever form it takes, it is likely to be in a source-specific form. It is very unlikely that the brain would store information about, say, the vowel /a/ mixed with each of the background sounds with which we have ever encountered it. Rather, the stored description of /a/ must capture salient properties that allow us to recognize it when it is mixed with other sounds. The frequencies of spectral peaks, like formants, since they represent the locally highest energy, will be the most likely to escape masking by other sounds. It seems probable that one of the functions of the early stages of auditory

processing is to get the incoming sound mixture into a form where it can sensibly make contact with stored source-specific information (Gutschalk *et al.* 2005).

4. SEPARATING SPEECH SIGNALS

(a) *The nature of the problem*

One interesting property of speech is that, on a frequency–time plot like a spectrogram, it is a sparse signal, which mostly consists of discrete harmonics, whose amplitudes vary gradually with frequency, being maximal near the formant peaks. There are also periods of silence, or near-silence, for example when the vocal tract is closed during production of a stop consonant. Consequently, when two different speech signals are mixed together with similar overall levels, any particular local frequency–time region is substantially dominated by one signal or the other. Such dominance is exaggerated by the fact that the auditory system codes intensity approximately logarithmically, since if $a \gg b$ then $\log(a+b) \sim \log(a)$. In other words, in a mixture of two speech signals, each local frequency–time region will reflect predominantly the value of one of the speech signals. Put another way, the log-amplitude spectrogram of the mixture is almost the same as the (frequency/time) element-wise maximum of the component spectrograms (Moore 1986; Varga & Moore 1990). If knowledge of the two signals is used to extract just those frequency–time regions which are dominated by one particular signal using an 'ideal binary mask' (Hu & Wang 2004), then a good, intelligible version of that signal can be resynthesized from just those regions (Cooke 2003; Roweis 2004). Such independent knowledge is, of course, not available to someone listening to a mixture of two voices, or to a computer program attempting to separate a mixture. Both the listener and the program must use whatever knowledge they have in order to attempt to identify which frequency–time regions come from the signal of interest (Wang 2004).

The problem of separating two talkers is thus very different from the problem of trying to listen to one talker against a background of steady noise. Such noise provides a level of masking that is constant over time (apart from statistical fluctuations) in each frequency band. Consequently, only the most intense parts of the signal will be detectable against this background, and the perceptual (or computational) problem is one of detection. Anything that is detectably different from the background belongs to the signal. Thus, with noise, the problem is one of detecting parts of the signal, whereas with two talkers the problem is one of assigning readily detectable frequency–time elements to the appropriate sound source. This difference has been characterized as the difference between energetic masking (see Moore 2008) and 'informational' masking (Watson 1987; Durlach *et al.* 2003). Experiments which have formally assessed the intelligibility of speech produced by the ideal binary mask from a multi-talker background (Brungart 2005) have shown that energetic masking plays a very small role when there is only a single competing talker, but that its role increases rapidly when additional interfering voices are added. In the limit, a multi-talker babble becomes very similar to noise that has the same average spectrum as speech; the sparseness of the single interfering talker is lost.

Although speech masked by a single talker can be convincingly resynthesized by attributing each time–frequency element to either the target or the masking speech, the auditory system does not always conform to this principle of *disjoint* allocation. Listeners can *conjointly* allocate the energy in a particular frequency channel to more than one sound source—an application of Bregman's (1990) 'Old+New' heuristic. This heuristic states that we try to maintain perceived continuity of an initial sound when a new sound starts. What is then heard as the new sound corresponds to the total sound that is now present *minus* the old sound. Complementing the additivity of sound mixtures, this Old+New heuristic is subtractive within a frequency channel, so that, for example, when a noise burst alternates with a more intense one, not only does the less intense one sound continuous, but the loudness of the more intense one is reduced compared with its loudness in isolation (Warren *et al.* 1972; McAdams *et al.* 1998). A similar effect can be shown on the timbre of complex sounds when part of the complex precedes the remainder (Darwin 1995). Although the precise metric used for such subtraction is not clear, what is clear is that a single frequency channel is not exclusively allocated to one source or another. A minimalist demonstration of non-exclusive allocation can be made with a single frequency. Play the same sine wave at equal amplitude to each ear, and then put a pulsed increment just on the left ear. You hear a continuous steady tone in the middle of the head, with an additional pulsing tone at the same frequency on the left ear. Incidentally, this demonstration argues against Kubovy's (1981) notion that for two sounds to be heard they must differ in frequency (an 'indispensable attribute' of a sound source).

(b) *Auditory grouping and the perception of speech*

The perception of noisy speech has been characterized as 'glimpsing' (Miller & Licklider 1950), where the listener is given occasional, relatively undistorted views of the speech scattered across the frequency–time plane (Cooke 2003; Assmann & Summerfield 2004). The auditory system seems to be well adapted to deal with such glimpses since it can tolerate the temporal interruption of speech by noise, even asynchronously in different frequency channels (Howard-Jones & Rosen 1993; Carlyon *et al.* 2002; Buss *et al.* 2004) as illustrated in figure 1. Intelligibility is improved if the interrupting noise is intense enough to be capable of masking the speech that it replaces, especially if what is being said is relatively predictable (Miller & Licklider 1950; Cherry & Wiley 1967; Powers & Wilcox 1977; Verschuure & Brocaar 1983). The presence of such noise also leads to the illusion that the speech is continuous and to an inability to identify which speech sounds are physically present, and which replaced by noise (Warren 1970, 1984; Warren *et al.* 1972; Samuel 1981).

As we saw in §4*a*, the perceptual problem is more complex when speech is heard at the same time as a small number of other talkers. How are the 'glimpses' from a particular talker perceptually separated from

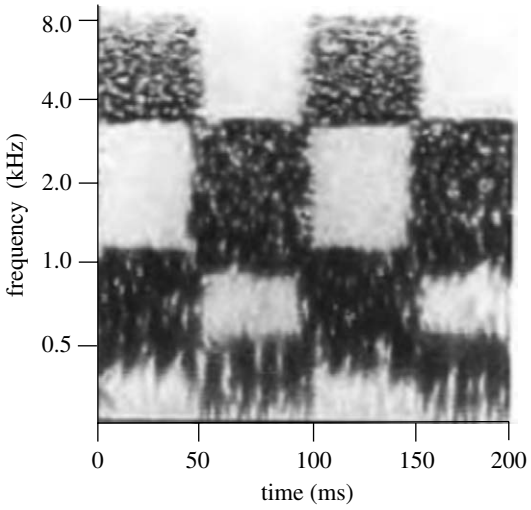

Figure 1. Spectrogram of 'checkerboard' noise with four channels logarithmically spaced in frequency. Redrawn from Howard-Jones & Rosen (1993).

those coming from other talkers? This complex field brings together the research areas of speech perception and auditory scene analysis. The present paper can only cherry-pick some issues; for a more detailed discussion, the interested reader is referred to two excellent recent reviews (Bronkhorst 2000; Assmann & Summerfield 2004).

It is convenient to follow Bregman's (1990) conceptual structure by distinguishing between simultaneous and sequential grouping of speech glimpses. Simultaneous grouping determines whether frequency–time elements that are presented at the same time belong to the same sound source, whereas sequential grouping refers to the process of following a particular sound source across time. Such a distinction may be a simplification, since it may be useful for the listener to track across time an individual element like a harmonic, rather than an entire sound source.

(i) *Harmonicity*

Much of speech is voiced, and normal voicing results in a quasi-periodic signal that shows considerable harmonic structure with a perceptible pitch corresponding to the fundamental frequency (F0; see Moore 2008). This harmonic structure is used by the auditory system to help separate simultaneous sounds. For example, two steady, synthetic simultaneous vowels are more intelligible if they are played at slightly different F0s than if they are played at the same F0 (Scheffers 1979, 1983; Assmann & Summerfield 1989, 1990). Although identification is well above chance when the sounds have the same F0, it increases by approximately 20% as the F0 difference increases to one semitone, but then asymptotes with further F0 increases. Although, this result stimulated subsequent research into both the psychology and physiology (Palmer 1988) of the perception of simultaneous sounds, the likely explanation of the intelligibility increase at very small F0s (approx. one semitone) is not one that would be very useful for more natural sounds. When two harmonic sounds differ in F0 by, say, a semitone, corresponding pairs of harmonics are too close together in frequency (approx. 6% separation) to be resolved by the cochlea

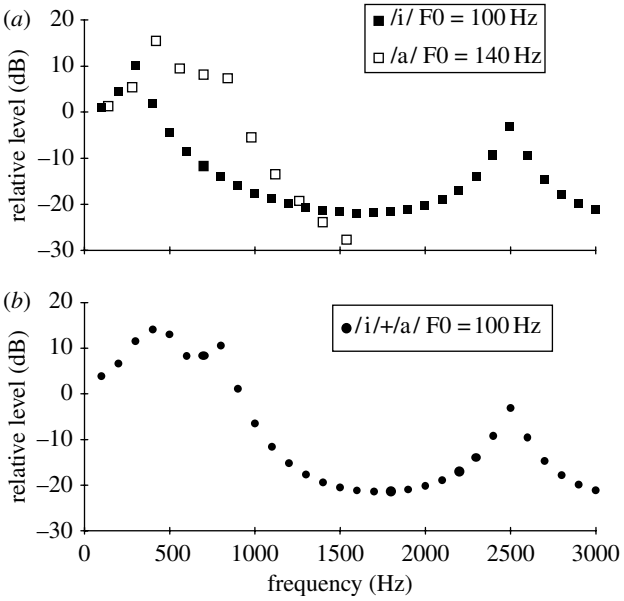

Figure 2. Schematic magnitude spectrum showing levels of individual harmonics in a mixture of vowel sounds. (*a*) The two vowels have different fundamental frequencies, forming two different harmonic series. (*b*) The two vowels are on the same fundamental frequency so only a single harmonic series is present. Note that the different first-formant frequencies of the two vowels are not well represented in (*b*).

(see Moore 2008). Since the two corresponding harmonics excite essentially the same region of the basilar membrane, its vibration reflects the physical addition of the sounds to give beats—relatively slow fluctuations in the level of the summed sound. It is probable that in a sufficiently long sound (Assmann & Summerfield 1994) these slow fluctuations allow the perceptual system to 'glimpse' one vowel or the other at different instants, as the spectral envelope systematically changes (Culling & Darwin 1993, 1994). Natural speech is not sufficiently stationary to benefit from this mechanism. Indeed, if the F0 of natural speech is manipulated to give consistent differences between the F0s of two competing speakers, intelligibility shows relatively little improvement at the one semitone F0 difference at which the double-vowel data asymptotes (Brokx & Nooteboom 1982; Bird & Darwin 1998). However, there are two other reasons why a difference in F0 improves the intelligibility of natural speech in the presence of a competing talker (Bird & Darwin 1998).

One reason is that a difference in F0 improves the definition of the first-formant (F1) frequencies of the two speakers compared with when the F0s are the same. Figure 2*a* superimposes the spectra of two different vowels that have different F0s. The spectrum of each vowel consists of a harmonic series, and the peak of the amplitude envelope of each harmonic series corresponds to the vowel's F1 frequency. Provided that the perceptual mechanism is able to separate sufficiently well the two harmonic series in figure 2*a*, the lower F1 could be recovered. Figure 2*b* shows the spectrum of the sum of the two vowels when they have the same F0. Note that the spectral envelope of the summed sound does not have a peak corresponding to the lower of the two F1s and so it would be perceptually invisible.

The other reason, which in practice only applies for F0 differences greater than approximately four semitones (Bird & Darwin 1998), is that a common harmonic series helps to group together the different formant regions that make up a voiced sound such as a vowel (as proposed in Broadbent & Ladefoged 1957) and to separate them from sounds on different F0s (Darwin 1981; Darwin & Gardner 1986). Allocating the appropriate upper formants to the appropriate speaker is an important use of the pitch mechanism originally proposed by Schouten (1940) that operates on unresolved harmonics (see Moore 2008).

(ii) *Onset time*

Frequency components from a single sound often share a common onset time, and this property is used by listeners to group frequency components together in the perception of timbre (Bregman & Pinker 1978), including vowel quality in steady vowels and simple syllables (Darwin 1981, 1984). For example, if the harmonics of a complex tone start at different times, say 100 ms apart, the frequencies of the individual harmonics are clear, but the timbre of the complex is less clear than if they had all started simultaneously. Similarly, if one of the harmonics of a steady vowel starts earlier than the others, it makes less of a contribution to the vowel quality than if it had been simultaneous. Such effects may be an advantageous consequence of the way that sounds are coded in auditory nerve fibres, which are known to adapt rapidly to steady sounds (Smith 1979; Yates *et al.* 1985; Chimento & Schreiner 1991). Rapid adaptation will reduce the neural response to a leading sound by the time the remaining sounds start. That this adaptation might not be a complete explanation was indicated first by offset asynchrony producing a similar, though somewhat smaller effect (Darwin & Sutherland 1984; Roberts & Moore 1991); and second by the fact that the contribution of the leading sound could be partly reinstated by making just its leading portion group with another sound that started with the leading sound but stopped when the remaining sounds started (Darwin & Sutherland 1984). However, this grouping explanation has itself been challenged recently, and an alternative explanation proposed, based on across-frequency interactions between sounds in the cochlear nucleus (Roberts & Holmes 2006). The relative contribution of relatively peripheral coding mechanisms and more central cognitive grouping mechanisms to the problem of sound source separation is likely to remain a topic of interest for sometime.

Onset time also provides a useful grouping cue for connected speech. For example, it has been used in computational auditory scene analysis, along with harmonicity, to group together frequency components from one of the two harmonic sound sources (Cooke 1993) and also has been found effective at grouping together the frequency-modulated sine waves of sine-wave speech (Barker & Cooke 1999). Its usefulness in connected speech is due to the interruptions to the continuous speech signal associated with stop and voiceless consonants. In the absence of onset-time differences between two voices, there is more scope for pitch-difference cues to be effective. This increased scope is probably the reason why the improvement in intelligibility with increasing difference in F0 of two simultaneous sentences is very much larger than normal when the sentences mainly contain continuant consonants (compare Brokx & Nooteboom (1982) and Bird & Darwin (1998)).

(iii) *Spatial direction*

Different natural sound sources usually come from different directions in space, and localization cues have been used very extensively for the machine segregation of different talkers (Bodden 1996). There are various different advantages that accrue to the listener from having sounds come from different directions (for a recent review see Bronkhorst (2000)); however, the way that the brain uses directional cues in auditory grouping is not straightforward and reflects the problems that difficult listening environments provide for sound localization.

The intelligibility of speech masked by a single continuous noise is higher when the speech and noise come from different directions. The two main reasons (Plomp 1976) are (i) for the higher-frequency components, the head casts an acoustical shadow which can benefit the ear on the side of the speech and (ii) the different relative phases at the two ears of the low-frequency components of the speech and noise make the speech easier to detect (the so-called binaural masking level difference, Licklider 1948). The relative contribution of these two mechanisms to intelligibility depends on the type of speech material used, the former being more important for monosyllables, and the latter for spondees (Dirks & Wilson 1969). With speech, rather than noise, as the interfering sound, masking is reduced owing to the sparsity of the speech signal (see above), but there are the additional problems of simultaneously grouping together the components that make up a particular speech sound, and of tracking one of the sources across time.

The dominant cue for localizing speech signals in the horizontal plane (or azimuth) is the inter-aural time differences (ITDs) of the sound's low-frequency (less than 1.5 kHz) components (Wightman & Kistler 1992), at least in non-reverberant environments. Curiously, though, most listeners are very poor at selectively grouping together simultaneous frequency components on the basis of common ITDs. Culling & Summerfield (1995) constructed four different vowel-like sounds by pairing in different combinations of four different narrow band-pass noises close in frequency to the first two formant frequencies of each vowel. For example, bands 1 and 4 together give a percept of /i/ while 2 and 3 together give /a/. The alternative grouping of 1 with 3, and 2 with 4 gives two different vowels. When one pair of noise bands (say 1 and 4) was played to one ear and the other pair to the other ear, listeners had no difficulty in identifying the vowel on the left ear. However, when the noise bands were all led to both ears but given different ITDs, listeners were, surprisingly, unable to identify the vowel heard on the left side.

We have then the apparently paradoxical result that the cue that is dominant for the localization of complex sounds is quite ineffective for the grouping of

simultaneous sounds. This unexpected finding also seems to go against one's introspective experience of attending spatially to one sound source rather than another. Is this simply an illusion, although one supported by experimental evidence (Spence & Driver 1994; Munte *et al.* 2001), or is the relationship between spatial attention and auditory localization more complex than implied by the simple notion of attending to frequency channels that share the same ITD?

Evidence supports the more complex relationship. The basic idea is that auditory grouping (based at least on primitive cues such as harmonicity and onset time) occurs prior to the localization of complex sounds. According to this idea, which was initially proposed by Woods & Colburn (1992) and subsequently pursued by Hill & Darwin (1996), the interaural time (and intensity) differences of individual frequency channels are computed independently, in parallel with a separate process which assigns these frequency channels to separate sound sources. The two types of information are then brought together so that the localization of an auditory object can be constructed from the ITDs of the frequency components that make up that auditory object. In the case of Culling & Summerfield's noise bands, there are no grouping cues other than ITD to pair off the bands, and consequently there is no segregation into two perceived vowels. By contrast, if the members of a pair of ordinary voiced vowels with a small difference in F0 are additionally given different ITDs, identification of the vowels improves (Shackleton *et al.* 1994).

Subsequent work on this lack of grouping by ITD has qualified the original conclusion. With practice, some listeners *can* learn to perform segregation by ITD (Darwin 2002; Drennan *et al.* 2003), though it is not clear how well this ability generalizes outside of the set of sounds that the listeners have been exposed to. Another important qualification is that if one uses more complex, natural sounds, rather than steady-state vowel-like sounds, listeners can much more easily identify two spatially separate sounds simply on the basis of ITD cues. Darwin & Hukin (1999) showed that natural sounds, modified to have the same F0 and that differed only in ITD could be selectively attended very easily. For example, if two simultaneous monosyllabic words ('bead' and 'globe') are given ITDs of $+90$ and $-90\,\mu s$, respectively, they are readily perceived as two spatially distinct auditory objects which can be readily attended to. For natural speech, there are many cues (e.g. harmonicity, onset-time differences) which can help the auditory system to allocate individual frequency channels to the two different sound sources. What is more surprising is that the impression of two separate objects survives when the two words are resynthesized on the same F0. For simpler sounds, such as steady vowels, the impression of two distinct sources with separate locations is destroyed by a common F0.

The upshot of these experiments is to provide broad support for the model outlined by Woods & Colburn (1992). ITD is a surprisingly weak cue to segregation, at least for unpractised listeners and in the absence of other grouping cues. When the listener has some independent way of grouping together the frequency components that make up different sound sources, then ITD differences between the sources give improved identification. One reason why the auditory system may not use ITD as a strong primary grouping cue is that in normal, difficult listening situations (with multiple sound sources, sound reflections and reverberation) the ITD information in a single frequency channel is not robust. The auditory system may be able to produce a more stable estimate of the location of sound sources by pooling ITD information across those frequencies, and only those frequencies (Hill & Darwin 1996), that make up a sound source. Bearing out this point is the interesting observation that the perceived location of sound sources is very stable: for example, the high- and low-frequency parts of a sound never appear to come from different directions. Such spatial fragmentation might be a more common experience if the brain did use ITD as a primary grouping cue.

Once an auditory object is spatially localized, this location provides a powerful cue for tracking it across time. Spatial cues become particularly important if other cues such as semantic redundancy (Treisman 1960), pitch and voice differences (Darwin & Hukin 2000; Darwin *et al.* 2003) or level (Egan *et al.* 1954; Brungart 2001) are ineffective.

The perceptual separation of two interleaved melodies by spatial differences is strikingly illustrated in Bregman & Ahad's demonstration of West African xylophone (amadinda) music (Bregman & Ahad 1995, demonstration 41). Experimental data illustrating a similar effect of spatial segregation (using interleaved monotone rhythms) is provided by Sach & Bailey (2004), who also distinguish between the effects of perceived location and the effect of individual cues. They traded-off ITD and interaural level cues to show that perceived spatial separation, rather than an ITD difference itself, is responsible for the perceptual segregation of the two rhythmic streams.

An ingenious experiment by Freyman *et al.* (1999) reached a similar conclusion concerning the intelligibility of speech masked by similar speech. They exploited the precedence or Haas effect (Haas 1972) to produce the impression that sound sources were coming from different azimuthal directions without using either of the conventional binaural cues (ILD or ITD). They measured listeners' identification of nonsense sentences spoken by a female talker in the presence of either speech-spectrum noise or of similar sentences spoken by a second female talker. They used two spatial arrangements: one where the target and masker both came from straight ahead, and one in which the distractor also came from a second loudspeaker, 45° to the right, but its signal led the distractor signal in the straight-ahead loudspeaker by 4 ms. Thanks to the precedence effect, the latter configuration gives the clear impression that the distractor is coming from the right-hand loudspeaker, while the target remains straight ahead. This spatial separation substantially improved performance when the distractor was the other female talker. In the absence of other cues (the female talkers were similar and the sentences were nonsense), the perceived spatial difference allowed the listener to selectively follow the

target talker. However, the additional version of the distractor from straight ahead complicates the interaural time and intensity differences between the targets and the distractors so that there is little binaural masking level difference. Consequently, the impression of a large spatial difference produced by the precedence effect produced little intelligibility increase with the speech-spectrum noise. Freyman's experiment thus demonstrates that a perceived spatial difference between talkers can increase intelligibility under conditions where there is no energetic masking advantage from the spatial separation.

Although, as we have seen, the natural spatial separation of attended and unattended sources improves detection and tracking of the attended source, a spatially separated distractor sound *can* cause a remarkable amount of disruption to selective attention. Brungart & Simpson (2002) asked listeners to respond to a target speech signal spoken by one of two competing talkers in one ear while ignoring a simultaneous masking sound in the other ear. When the masking sound in the unattended ear was noise, listeners were able to segregate the competing talkers in the target ear nearly as well as they could with no sound in the unattended ear. But when the masking sound in the unattended ear was speech, speech segregation in the target ear was very substantially worse than with no sound in the unattended ear. The presence of speech-like (Brungart *et al.* 2005) sounds in the unattended ear makes the separation of sounds in the attended ear much harder.

5. CONCLUDING REMARKS

The difficult problem of how speech can be recognized against a background of other sounds is receiving increased attention from both psychologists and computer scientists. Speech can be remarkably well recognized by human listeners under a wide variety of distortions and against both random and structured noise backgrounds in anechoic and reverberant surroundings. The brain uses a wide range of perceptual mechanisms to achieve a level of recognition that is presently beyond computer systems. Some of these mechanisms have probably arisen as a response to the general problem of recognizing sounds in noisy and reverberant environments, while others may rely on specific knowledge about speech. The types of mechanisms involved vary widely depending on the characteristics of the noise. For random noise, the problem is mainly one of detection and also requires recognition mechanisms that can operate on the basis of partial information, tolerating 'missing data'. For more structured noise, such as another talker, additional problems arise of allocating sensory fragments to one or other sound source, and of tracking an individual sound source over time. The effectiveness of various parameters in allowing this perceptual grouping has begun to be studied, although this knowledge has not yet been integrated into the mainstream of work on speech perception. One particular gap in our knowledge is an understanding of what are the intermediate representations of sound between the sensory coding of the auditory nerve and the human brain's representation of phonetic knowledge. It is probable that those representations will be intimately linked to the problem of how the brain deals with multiple sound sources.

REFERENCES

Ainsworth, W. A. & Miller, J. B. 1972 The effect of relative formant amplitude on the identity of synthetic vowels. *Lang. Speech* **15**, 328–341.

ANSI 1997 *ANSI S3.5-1997: methods for calculation of the speech intelligibility index*. New York, NY: American National Standards Institute.

Assmann, P. F. & Summerfield, Q. 1989 Modelling the perception of concurrent vowels: vowels with the same fundamental frequency. *J. Acoust. Soc. Am.* **85**, 327–338. (doi:10.1121/1.397684)

Assmann, P. F. & Summerfield, Q. 1990 Modelling the perception of concurrent vowels: vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* **88**, 680–697. (doi:10.1121/1.399772)

Assmann, P. F. & Summerfield, Q. 1994 The contribution of waveform interactions to the perception of concurrent vowels. *J. Acoust. Soc. Am.* **95**, 471–484. (doi:10.1121/1.408342)

Assmann, P. F. & Summerfield, Q. 2004 The perception of speech under adverse conditions. In *Speech processing in the auditory system* (eds S. Greenberg, W. A. Ainsworth, A. N. Popper & R. R. Fay), pp. 231–308. New York, NY: Springer.

Atal, B. S. & Hanauer, S. L. 1971 Speech analysis and synthesis by linear prediction of the acoustic wave. *J. Acoust. Soc. Am.* **50**, 637–655. (doi:10.1121/1.1912679)

Barker, J. & Cooke, M. 1999 Is the sine-wave speech cocktail party worth attending? *Speech Commun.* **27**, 159–174. (doi:10.1016/S0167-6393(98)00081-8)

Bird, J. & Darwin, C. J. 1998 Effects of a difference in fundamental frequency in separating two sentences. In *Psychophysical and physiological advances in hearing* (eds A. R. Palmer, A. Rees, Q. Summerfield & R. Meddis), pp. 263–269. London, UK: Whurr Publishers.

Bodden, M. 1996 Auditory demonstrations of a cocktail-party processor. *Acustica* **82**, 356–357.

Bregman, A. S. 1978 Auditory streaming is cumulative. *J. Exp. Psychol. Hum. Percept. Perform.* **4**, 380–387. (doi:10.1037/0096-1523.4.3.380)

Bregman, A. S. 1990 *Auditory scene analysis: the perceptual organisation of sound*. Cambridge, MA: Bradford Books/MIT Press.

Bregman, A. S. & Ahad, P. A. 1995 *Compact disc: demonstrations of auditory scene analysis*. Montreal, Canada: Department of Psychology, McGill University.

Bregman, A. S. & Pinker, S. 1978 Auditory streaming and the building of timbre. *Can. J. Psychol.* **32**, 19–31.

Broadbent, D. E. & Ladefoged, P. 1957 On the fusion of sounds reaching different sense organs. *J. Acoust. Soc. Am.* **29**, 708–710. (doi:10.1121/1.1909019)

Brokx, J. P. L. & Nooteboom, S. G. 1982 Intonation and the perceptual separation of simultaneous voices. *J. Phonet.* **10**, 23–36.

Bronkhorst, A. W. 2000 The cocktail party phenomenon: a review of speech intelligibility in multiple-talker conditions. *Acustica* **86**, 117–128.

Brungart, D. S. 2001 Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* **109**, 1101–1109. (doi:10.1121/1.1345696)

Brungart, D. S. 2005 A binary masking technique for isolating energetic masking in speech perception. *J. Acoust. Soc. Am.* **117**, 2484. (doi:10.1121/1.1835509)

Brungart, D. S. & Simpson, B. D. 2002 Within-ear and across-ear interference in a cocktail-party listening task. *J. Acoust. Soc. Am.* **112**, 2985–2995. (doi:10.1121/1.1512703)

Brungart, D. S., Simpson, B. D., Darwin, C. J., Arbogast, T. L. & Kidd Jr, G. 2005 Across-ear interference from parametrically-degraded synthetic speech signals in a dichotic cocktail-party listening task. *J. Acoust. Soc. Am.* **117**, 292–304. (doi:10.1121/1.1835509)

Buss, E., Hall III, J. W. & Grose, J. H. 2004 Spectral integration of synchronous and asynchronous cues to consonant identification. *J. Acoust. Soc. Am.* **115**, 2278–2285. (doi:10.1121/1.1691035)

Carlyon, R. P., Deeks, J., Norris, D. & Butterfield, S. 2002 The continuity illusion and vowel identification. *Acta Acustica–Acustica* **88**, 408–415.

Cherry, E. C. 1953 Some experiments on the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **25**, 975–979. (doi:10.1121/1.1907229)

Cherry, E. C. & Taylor, W. K. 1954 Some further experiments upon the recognition of speech, with one and with two ears. *J. Acoust. Soc. Am.* **26**, 554–559. (doi:10.1121/1.1907373)

Cherry, C. & Wiley, R. H. 1967 Speech communication in very noisy environments. *Nature* **214**, 1164. (doi:10.1038/2141164a0)

Chimento, T. C. & Schreiner, C. E. 1991 Adaptation and recovery from adaptation in single fiber responses of the cat auditory-nerve. *J. Acoust. Soc. Am.* **90**, 263–273. (doi:10.1121/1.401296)

Clark, G. 2003 *Cochlear implants: fundamentals and applications*. Berlin, Germany; London, UK: Springer.

Cooke, M. 2003 Glimpsing speech. *J. Phonet.* **31**, 579–584. (doi:10.1016/S0095-4470(03)00013-5)

Cooke, M. P. 1993 *Modelling auditory processing and organisation*. Cambridge, UK: Cambridge University Press.

Cooke, M. P., Green, P. D., Josifovski, L. & Vizinho, A. 2001 Robust automatic speech recognition with missing and unreliable acoustic data. *Speech Commun.* **34**, 267–285. (doi:10.1016/S0167-6393(00)00034-0)

Culling, J. F. & Darwin, C. J. 1993 Perceptual separation of simultaneous vowels: within and across-formant grouping by F_0. *J. Acoust. Soc. Am.* **93**, 3454–3467. (doi:10.1121/1.405675)

Culling, J. F. & Darwin, C. J. 1994 Perceptual and computational separation of simultaneous vowels: cues arising from low-frequency beating. *J. Acoust. Soc. Am.* **95**, 1559–1569. (doi:10.1121/1.408543)

Culling, J. F. & Summerfield, Q. 1995 Perceptual separation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay. *J. Acoust. Soc. Am.* **98**, 785–797. (doi:10.1121/1.413571)

Darwin, C. J. 1981 Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Q. J. Exp. Psychol. A* **33**, 185–208.

Darwin, C. J. 1984 Perceiving vowels in the presence of another sound: constraints on formant perception. *J. Acoust. Soc. Am.* **76**, 1636–1647. (doi:10.1121/1.391610)

Darwin, C. J. 1990 Environmental influences on speech perception. In *Advances in speech and language processing* (ed. W. A. Ainsworth), pp. 219–241. London, UK: JAI Press.

Darwin, C. J. 1991 The relationship between speech perception and the perception of other sounds. In *Modularity and the motor theory of speech perception* (eds I. G. Mattingly & M. G. Studdert-Kennedy), pp. 239–259. Hillsdale, NJ: Erlbaum.

Darwin, C. J. 1995 Perceiving vowels in the presence of another sound: a quantitative test of the "Old-plus-New" heuristic. In *Levels in speech communication: relations and interactions: a tribute to Max Wajskop* (eds C. Sorin, J. Mariani, H. Méloni & J. Schoentgen), pp. 1–12. Amsterdam, The Netherlands: Elsevier.

Darwin, C. J. 2002 Auditory streaming in language processing. In *Genetics and the function of the auditory system: 19th Danavox symposium* (eds L. Tranebjaerg, T. Andersen, J. Christensen-Dalsgaard & T. Poulsen), pp. 375–392. Brøndby, Denmark: Holmens Trykkeri.

Darwin, C. J., Carlyon, R. P. 1995 Auditory grouping In *The handbook of perception and cognition*. Vol. 6 *Hearing* (ed. B. C. J. Moore), pp. 387–424. 2nd edn. London, UK: Academic Press.

Darwin, C. J. & Gardner, R. B. 1986 Mistuning a harmonic of a vowel: grouping and phase effects on vowel quality. *J. Acoust. Soc. Am.* **79**, 838–845. (doi:10.1121/1.393474)

Darwin, C. J. & Hukin, R. W. 1999 Auditory objects of attention: the role of interaural time-differences. *J. Exp. Psychol. Hum. Percept. Perform.* **25**, 617–629. (doi:10.1037/0096-1523.25.3.617)

Darwin, C. J. & Hukin, R. W. 2000 Effectiveness of spatial cues, prosody and talker characteristics in selective attention. *J. Acoust. Soc. Am.* **107**, 970–977. (doi:10.1121/1.428278)

Darwin, C. J. & Sutherland, N. S. 1984 Grouping frequency components of vowels: when is a harmonic not a harmonic? *Q. J. Exp. Psychol. A* **36**, 193–208.

Darwin, C. J., Brungart, D. S. & Simpson, B. D. 2003 Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.* **114**, 2913–2922. (doi:10.1121/1.1616924)

Diehl, R. L. 2008 Acoustic and auditory phonetics: the adaptive design of speech sound systems. *Phil. Trans. R. Soc. B* **363**, 965–978. (doi:10.1098/rstb.2007.2153)

Dirks, D. D. & Wilson, R. H. 1969 The effects of spatially separated sound sources on speech intelligibility. *J. Speech Hear. Res.* **12**, 5–38.

Dissard, P. & Darwin, C. J. 2000 Extracting spectral envelopes: formant frequency matching between sounds on different and modulated fundamental frequencies. *J. Acoust. Soc. Am.* **107**, 960–969. (doi:10.1121/1.428277)

Dissard, P. & Darwin, C. J. 2001 Formant frequency matching between sounds with different bandwidths and on different fundamental frequencies. *J. Acoust. Soc. Am.* **110**, 409–415. (doi:10.1121/1.1379085)

Drennan, W. R., Gatehouse, S. & Lever, C. 2003 Perceptual segregation of competing speech sounds: the role of spatial location. *J. Acoust. Soc. Am.* **114**, 2178–2189. (doi:10.1121/1.1609994)

Drullman, R. & Bronkhorst, A. W. 2000 Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *J. Acoust. Soc. Am.* **107**, 2224–2235. (doi:10.1121/1.428503)

Duquesnoy, A. J. 1983 Perceptual segregation of competing speech sounds: the role of spatial location. *J. Acoust. Soc. Am.* **74**, 739–743. (doi:10.1121/1.389859)

Durlach, N. I., Mason, C. R., Kidd Jr, G., Arbogast, T. L., Colburn, H. S. & Shinn-Cunningham, B. G. 2003 Note on informational masking. *J. Acoust. Soc. Am.* **113**, 2984–2987. (doi:10.1121/1.1570435)

Egan, J. P., Carterette, E. C. & Thwing, E. J. 1954 Some factors affecting multi-channel listening. *J. Acoust. Soc. Am.* **26**, 774–782. (doi:10.1121/1.1907416)

Festen, J. M. & Plomp, R. 1990 Effects of fluctuating noise and interfering speech on the speech–reception threshold for impaired and normal hearing. *J. Acoust. Soc. Am.* **88**, 1725–1736. (doi:10.1121/1.400247)

Freyman, R. L., Helfer, K. S., McCall, D. D. & Clifton, R. K. 1999 The role of perceived spatial separation in the unmasking of speech. *J. Acoust. Soc. Am.* **106**, 3578–3588. (doi:10.1121/1.428211)

Gutschalk, A., Micheyl, C., Melcher, J. R., Rupp, A., Scherg, M. & Oxenham, A. J. 2005 Neuromagnetic correlates of streaming in human auditory cortex. *J. Neurosci.* **25**, 5382–5388. (doi:10.1523/JNEUROSCI.0347-05.2005)

Haas, H. 1972 The influence of a single echo on the audibility of speech. *J. Audio Eng. Soc.* **20**, 145–149.

Hill, N. I. & Darwin, C. J. 1996 Lateralisation of a perturbed harmonic: effects of onset asynchrony and mistuning. *J. Acoust. Soc. Am.* **100**, 2352–2364. (doi:10.1121/1.417945)

Houtgast, T. & Steeneken, H. J. M. 1973 The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acustica* **28**, 66–73.

Howard-Jones, P. A. & Rosen, S. 1993 Uncomodulated glimpsing in "checkerboard" noise. *J. Acoust. Soc. Am.* **93**, 2915–2922. (doi:10.1121/1.405811)

Hu, G. & Wang, D. L. 2004 Monaural speech segregation based on pitch tracking and amplitude modulation. In *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPA)*, pp. 79–82.

Hunt, M. J. 1987 Delayed decisions in speech recognition—the case of formants. *Pattern Recognit. Lett.* **6**, 121–137. (doi:10.1016/0167-8655(87)90093-6)

Hunt, M. J., Zwierzynski, D. A. & Carr, R. C. 1989 Issues in high-quality LPC analysis and synthesis. In *Proc. Eur. Conf. on Speech Communication and Technology (Euro-Speech-89)*.

Killion, M. C. 1997 Hearing aids: past, present, future: moving toward normal conversations in noise. *Br. J. Audiol.* **31**, 141–148.

Klatt, D. H. 1985 The perceptual reality of a formant frequency. *J. Acoust. Soc. Am.* **78**, S81–S82. (doi:10.1121/1.2023019)

Kubovy, M. 1981 Concurrent pitch segregation and the theory of indispensible attributes. In *Perceptual organization* (eds M. Kubovy & J. R. Pomerantz), pp. 55–98. Hillsdale, NJ: Erlbaum.

Liberman, A. M. 1982 On the finding that speech is special. *Am. Psychol.* **37**, 148–167. (doi:10.1037/0003-066X.37.2.148)

Liberman, A. M., Cooper, F. S., Shankweiler, D. S. & Studdert-Kennedy, M. 1967 Perception of the speech code. *Psychol. Rev.* **74**, 431–461. (doi:10.1037/h0020279)

Licklider, J. C. R. 1948 The influence of interaural phase relations upon the masking of speech by white noise. *J. Acoust. Soc. Am.* **20**, 150–159. (doi:10.1121/1.1906358)

Lippmann, R. P. 1997 Speech recognition by machines and humans. *Speech Commun.* **22**, 1–15. (doi:10.1016/S0167-6393(97)00021-6)

McAdams, S., Botte, M. C. & Drake, C. 1998 Auditory continuity and loudness computation. *J. Acoust. Soc. Am.* **103**, 1580–1591. (doi:10.1121/1.421293)

Miller, G. A. 1947 The masking of speech. *Psychol. Bull.* **44**, 105–129. (doi:10.1037/h0055960)

Miller, G. A. & Licklider, J. C. R. 1950 The intelligibility of interrupted speech. *J. Acoust. Soc. Am.* **22**, 167–173. (doi:10.1121/1.1906584)

Moore, B. C. J. 1998 *Cochlear hearing loss*. London, UK: Whurr Publishers.

Moore, B. C. J. 2008 Basic auditory processes involved in the analysis of speech sounds. *Phil. Trans. R. Soc. B* **363**, 947–963. (doi:10.1098/rstb.2007.2152)

Moore, R. K. 1986 Signal decomposition using Markov modelling techniques. Royal Signals Research Establishment (UK) Memorandum no. 3931.

Munte, T. F., Kohlmetz, C., Nager, W. & Altenmuller, E. 2001 Superior auditory spatial tuning in conductors. *Nature* **409**, 580. (doi:10.1038/35054668)

Palmer, A. R. 1988 The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers. In *Basic issues in hearing* (eds H. Duifhuis, J. W. Jorst & H. P. Wit), pp. 244–251. London, UK: Academic Press.

Patterson, R. D. & Johnsrude, I. S. 2008 Functional imaging of the auditory processing applied to speech sounds. *Phil. Trans. R. Soc. B* **363**, 1023–1035. (doi:10.1098/rstb.2007.2157)

Peters, R. W., Moore, B. C. J. & Baer, T. 1998 Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *J. Acoust. Soc. Am.* **103**, 577–587. (doi:10.1121/1.421128)

Plomp, R. 1976 Binaural and monaural speech intelligibility of connected discourse in reverberation as a function of a single competing sound source (speech or noise). *Acustica* **34**, 200–211.

Plomp, R. 1977 Acoustical aspects of cocktail parties. *Acustica* **38**, 186–191.

Powers, G. L. & Wilcox, J. C. 1977 Intelligibility of temporally-interrupted speech with and without intervening noise. *J. Acoust. Soc. Am.* **61**, 195–199. (doi:10.1121/1.381255)

Price, C., Thierry, G. & Griffiths, T. D. 2005 Speech-specific auditory processing: where is it? *Trends Cogn. Sci.* **9**, 271–276. (doi:10.1016/j.tics.2005.03.009)

Remez, R. E., Rubin, P. E., Pisoni, D. B. & Carrell, T. D. 1981 Speech perception without traditional speech cues. *Science* **212**, 947–950. (doi:10.1126/science.7233191)

Rhebergen, K. S., Versfeld, N. J. & Dreschler, W. A. 2005 Release from informational masking by time reversal of native and non-native interfering speech. *J. Acoust. Soc. Am.* **118**, 1274–1277. (doi:10.1121/1.2000751)

Roberts, B. & Holmes, S. B. 2006 Asynchrony and the grouping of vowel components: captor tones revisited. *J. Acoust. Soc. Am* **119**, 2905–2918. (doi:10.1121/1.2190164)

Roberts, B. & Moore, B. C. J. 1991 The influence of extraneous sounds on the perceptual estimation of first-formant frequency in vowels under conditions of asynchrony. *J. Acoust. Soc. Am.* **89**, 2922–2932. (doi:10.1121/1.400731)

Roweis, S. 2004 Automatic speech processing by inference in generative models. In *Speech separation by humans and machines* (ed. P. L. Divenyi), pp. 97–134. New York, NY: Kluwer Academic Publishers.

Sach, A. J. & Bailey, P. J. 2004 Some characteristics of auditory spatial attention revealed using rhythmic masking release. *Percept. Psychophys.* **66**, 1379–1387.

Samuel, A. G. 1981 The role of bottom-up confirmation in the phonemic restoration illusion. *J. Exp. Psychol. Hum. Percept. Perform.* **7**, 1124–1131. (doi:10.1037/0096-1523.7.5.1124)

Scheffers, M. T. 1979 The role of pitch in perceptual separation of simultaneous vowels. Institute for Perception Research, Annual progress report 14, pp. 51–54.

Scheffers, M. T. 1983 Sifting vowels: auditory pitch analysis and sound segregation. PhD thesis, University of Groningen, The Netherlands.

Schouten, J. F. 1940 The residue and the mechanism of hearing. *Proc. Kon. Akad. Wetenschap* **43**, 991–999.

Shackleton, T. M., Meddis, R. & Hewitt, M. J. 1994 The role of binaural and fundamental frequency difference cues in the identification of concurrently presented vowels. *Q. J. Exp. Psychol. A* **47**, 545–563.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J. & Ekelid, M. 1995 Speech recognition with primarily temporal cues. *Science* **270**, 303–304. (doi:10.1126/science.270.5234.303)

Smith, R. L. 1979 Adaptation, saturation, and physiological masking in single auditory nerve fibers. *J. Acoust. Soc. Am.* **65**, 166–178. (doi:10.1121/1.382260)

Spence, C. J. & Driver, J. 1994 Covert spatial orienting in audition: exogenous and endogenous mechanisms. *J. Exp. Psychol. Hum. Percept. Perform.* **20**, 555–574. (doi:10.1037/0096-1523.20.3.555)

Sroka, J. J. & Braida, L. D. 2005 Human and machine consonant recognition. *Speech Commun.* **45**, 401–423. (doi:10.1016/j.specom.2004.11.009)

Stubbs, R. J. & Summerfield, A. Q. 1990 Algorithms for separating the speech of interfering talkers: evaluations with voiced sentences and normal-hearing and hearing-impaired listeners. *J. Acoust. Soc. Am.* **89**, 1383–1393. (doi:10.1121/1.400539)

Treisman, A. 1960 Contextual cues in selective listening. *Q. J. Exp. Psychol.* **12**, 242–248.

Varga, A. P. & Moore, R. K. 1990 Hidden Markov model decomposition of speech and noise. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 845–848.

Verschuure, J. & Brocaar, M. P. 1983 Intelligibility of interrupted meaningful and nonsense speech with and without intervening noise. *Percept. Psychophys.* **33**, 232–240.

Wang, D. L. 2004 An ideal binary mask as the computational goal of auditory scene analysis. In *Speech separation by humans and machines* (ed. P. L. Divenyi), pp. 181–197. New York, NY: Kluwer Academic Publishers.

Warren, R. M. 1970 Perceptual restoration of missing phonemes. *Science* **167**, 392–393. (doi:10.1126/science.167.3917.392)

Warren, R. M. 1984 Perceptual restoration of obliterated sounds. *Psychol. Bull.* **96**, 371–383. (doi:10.1037/0033-2909.96.2.371)

Warren, R. M., Obusek, C. J. & Ackroff, J. M. 1972 Auditory induction: perceptual synthesis of absent sounds. *Science* **176**, 1149–1151. (doi:10.1126/science.176.4039.1149)

Watkins, A. J. & Makin, S. J. 1994 Perceptual compensation for speaker differences and for spectral-envelope distortion. *J. Acoust. Soc. Am.* **96**, 1263–1282. (doi:10.1121/1.410275)

Watson, C. S. 1987 Uncertainty, informational masking and the capacity of immediate auditory memory. In *Auditory processing of complex sounds* (eds W. A. Yost & C. S. Watson), pp. 267–277. Hillsdale, NJ: Erlbaum.

Whalen, D. M. & Liberman, A. M. 1987 Speech perception takes precedence over nonspeech perception. *Science* **237**, 169–171. (doi:10.1126/science.3603014)

Wightman, F. L. & Kistler, D. J. 1992 The dominant role of low-frequency interaural time differences in sound localization. *J. Acoust. Soc. Am.* **91**, 1648–1661. (doi:10.1121/1.402445)

Woods, W. A. & Colburn, S. 1992 Test of a model of auditory object formation using intensity and interaural time difference discriminations. *J. Acoust. Soc. Am.* **91**, 2894–2902. (doi:10.1121/1.402926)

Yates, G. K., Robertson, D. & Johnstone, B. M. 1985 Very rapid adaptation in the guinea-pig auditory nerve. *Hear. Res.* **17**, 1–12. (doi:10.1016/0378-5955(85)90124-8)