Journal of Experimental Psychology:
Human Perception and Performance
1999, Vol. 25, No. 3, 617-629

0096-1523/99/$3.00

Auditory Objects of Attention: The Role of Interaural Time Differences

C. J. Darwin and R. W. Hukin
University of Sussex

The role of interaural time difference (ITD) in perceptual grouping and selective attention was explored in 3 experiments. Experiment 1 showed that listeners can use small differences in ITD between 2 sentences to say which of 2 short, constant target words was part of the attended sentence, in the absence of talker or fundamental frequency differences. Experiments 2 and 3 showed that listeners do not explicitly track components that share a common ITD. Their inability to segregate a harmonic from a target vowel by a difference in ITD was not substantially changed by the vowel being placed in a sentence context, where the sentence shared the same ITD as the rest of the vowel. The results indicate that in following a particular auditory sound source over time, listeners attend to perceived auditory objects at particular azimuthal positions rather than attend explicitly to those frequency components that share a common ITD.

This article addresses a paradox. On the one hand, both everyday experience and experimental evidence (Spence & Driver, 1994; Teder & Näätänen, 1994) show that auditory attention can be directed toward sounds that come from a particular location. On the other hand, although interaural time difference (ITD) is the most powerful cue for determining the direction of a complex sound (Culling, Summerfield, & Marshall, 1994; Wightman & Kistler, 1992), it is remarkably ineffective at helping listeners to group together the simultaneous frequency components that make up a particular sound source (Culling & Summerfield, 1995; Hukin & Darwin, 1995b). We propose a resolution to the paradox that distinguishes between grouping mechanisms responsible for the formation of auditory objects (which make very little use of ITD) and the determination of the subjective location of a grouped auditory object, which may be based on the pooled ITDs of the grouped frequency components. We show in the first experiment that listeners can attend across time to one of two spoken sentences distinguished by small differences in ITD. By contrast, in the second and third experiments, we show that listeners do not use such continuity of ITD to determine which individual frequency components should form part of a sentence.

Although people can attend to one of two voices, or other sound sources, that both come from a single loudspeaker, auditory attention can also be directed readily to a particular spatial location (Spence & Driver, 1994). When attention is directed spatially in this way, what is it that is being attended? Our subjective experience suggests that we attend to auditory objects (individual sound sources), and the theoretical framework proposed by Bregman (1990) adds that these auditory objects or streams have been formed (at least in part) by preattentive grouping mechanisms based on such common properties as harmonicity and common onset time. The subjective location of an auditory object could then be determined on the basis of the location cues of its component frequencies (Hill & Darwin, 1996; Trahiotis & Stern, 1989; Woods & Colburn, 1992).

An alternative, more reductionist view could, with an eye to the physiological basis of sound localization, propose that attention is directed to those frequency components that have each come from a particular direction. Because for complex sounds a difference in the time of arrival of sound at the two ears (ITD) is the most salient cue to azimuth (Wightman & Kistler, 1992), a simple model of auditory spatial attention can be based on Jeffress's (1948) physiologically valid (Yin & Chan, 1990) cross-correlation model in which frequency-specific fibers from either ear excite coincidence detectors after a specific interaural delay. A coincidence detector fires when a spike arrives at the same time from each ear. Attention could be directed to those coincidence detectors that share, across frequency, the same interaural delay.

The recent discovery (Culling & Summerfield, 1995) and verification (Hukin & Darwin, 1995b) of the surprising weakness of ITD as a perceptual grouping cue for simultaneous sounds potentially allows these two different views to be compared. Listeners are unable either to use a difference in ITD to pair together into vowels four individual simultaneous formant-like noise bands (Culling & Summerfield, 1995) or to segregate a single, resolved harmonic from the phonetic percept of a voiced vowel (Hukin & Darwin, 1995b). But although ITD is a weak cue for perceptually grouping simultaneous sounds, it may be more effective at grouping sounds sequentially (Darwin & Hukin, 1997). In

C. J. Darwin and R. W. Hukin, Laboratory of Experimental Psychology, School of Biological Sciences, University of Sussex, Brighton, Sussex, England.

The research was supported by Medical Research Council Grant G9505738N. The pitch-synchronous overlap-add (PSOLA) procedure used in Experiment 1 was adapted for use as a Macintosh Programmers' Workshop tool by Paul Russell. Nick Hill, Stuart Leech, Brian Moore, and Quentin Summerfield made many helpful comments on an earlier version of this article.

Correspondence concerning this article should be addressed to C. J. Darwin, Laboratory of Experimental Psychology, School of Biological Sciences, University of Sussex, Brighton, Sussex BN1 9QG, England. Electronic mail may be sent to cjd@biols.susx.ac.uk.

the first experiment, we asked whether a difference in ITD can be used to determine across time which words form a sentence. In the second and third experiments, we pursued whether the positive result from the first experiment was due to the use of ITD at the level of individual frequency components or of grouped objects.

In the experiments described here, speech was used as a convenient example of a complex auditory object. The theoretical position leading to the present experiments is that speech and other sounds share some of the mechanisms that allow listeners to partition the auditory input according to whether individual components are more likely to have come from one or another sound source. It has been claimed that speech sounds are not subject to such grouping mechanisms (Remez, Rubin, Berns, Pardo, & Lang, 1994) on the grounds that speech that is reduced to three sine waves can still be understood. However, acknowledging that there are undoubtedly constraints on particular types of sound (such as speech) that are unique to them ("schemata" in Bregman's, 1990, terminology) and that can help listeners allocate sound components appropriately does not preclude the contribution of more general auditory mechanisms to the perceptual separation of speech either from other speech sounds or from nonspeech (Darwin, 1981, 1991). In addition, the claim that speech does not use these more general mechanisms does not provide an adequate framework for explaining changes in the intelligibility of speech masked by other sounds or of changes in the identity of speech syllables when simple auditory cues such as fundamental frequency (Fo) are manipulated (Assmann & Summerfield, 1990; Bird & Darwin, 1998; Culling & Darwin, 1993; Darwin, 1981, 1984, 1997; Darwin & Carlyon, 1995). It is interesting in this context to note the recent finding that listeners have extreme difficulty in identifying a mixture of two sine-wave-speech sentences (Barker & Cooke, 1999). The lack of harmonic structure in these stimuli removes a powerful low-level cue for their perceptual segregation.

Experiment 1

There are a number of simple cues that could serve to group together sequentially the sounds from a common source, such as the speech of a particular talker. Spatial location is one, but others such as continuity of fundamental frequency (Fo) have also been proposed. Over a short duration, continuity of individual harmonics or of formant frequencies can be useful, but they lack generality across pauses and voiceless stops and fricatives.

Experimental evidence for the use of spatial location to define a particular sound source or talker across time comes from a variety of sources. Listeners presented with three pairs of synchronous dichotic digits recalled all three digits presented to one ear before those presented to the other, provided the rate of presentation was faster than about 1.5 s/pair (Broadbent, 1953). Speech that is alternated (at about 4 Hz) between the ears loses intelligibility (Cherry & Taylor, 1954), which is substantially restored if noise is added to the silent ear (Schubert & Parker, 1956). A similar effect occurs in music: A melodic line is destroyed if the notes alternate between the ears but is partially restored if a constant-frequency drone tone is added synchronously to the silent ear (Deutsch, 1979). The implication of Deutsch's experiments is that spatial location may be less clear when there are multiple simultaneous tonal sources present, and hence sequential segregation by spatial location may be less effective.

There is also experimental evidence from a number of paradigms for the use of pitch continuity in defining a complex sound source across time. If an Fo contour that alternates between two values is imposed on a smoothly changing, repeating, formant pattern, then after a few alternations of Fo the sound subjectively breaks up into two talkers on different Fos, with a consequent change of the phonetic percept from semivowels to stop consonants—cued by the implied silence of one talker during the other's turn (Darwin & Bethell-Fox, 1977). Simpler stimuli (sequences of four 100-ms single-formant sounds) will segregate into separate streams on the basis of Fo differences (Bregman, Liao, & Levitan, 1990). Another example used shadowing of natural speech rather than the perception of repeating synthetic formant patterns. Listeners were asked to shadow the passage played to one ear, ignoring a different passage read by the same talker presented to the other ear. Listeners who were successful at continuously shadowing the target passage showed intrusion errors from the opposite ear when the intonation was suddenly switched between the two passages, even though the switching led to syntactic and semantic discontinuities in the text (Darwin, 1975). Similarly, Brokx and Nooteboom (1982) explained the improvement in intelligibility of sentences given different Fo contours against a competing passage of continuous speech as in part due to listeners' using continuity of Fo to track a particular utterance across time. Finally, although there is some evidence that vowel-length effects in consonant perception depend on Fo continuity (Green, Stevens, & Kuhl, 1994), spectral effects, such as continuity of individual harmonics, are now also known to be involved (Lotto, Kluender, & Green, 1996). It is possible that such spectral effects also modify other apparent demonstrations of Fo continuity.

The aim of the present experiment was to assess the relative effectiveness of differences in ITD and in Fo between two sentences in allowing listeners to track a particular sound source over time. Our choice of paradigm in this experiment was guided by the need to emphasize the role of ITD or Fo continuity in defining a sound source across time rather than in helping to detect individual auditory elements or to group them simultaneously.

The experiment had two carrier sentences and two target words embedded in the carriers. The same two carriers and the same two target words were used throughout the experiment. On each trial, the listeners were presented with the two sentences simultaneously. Their task was to attend to a particular carrier sentence (the same one throughout the experiment) and to indicate which of the target words was part of the attended sentence. The two carrier sentences could have the same or different Fos and the same or different ITDs. The two target words had the same Fos and

ITDs as the carrier sentences but not necessarily in the same combination. Consequently, on some trials one of the target words had both the same ITD and Fo as the attended carrier; on others one target word shared the same Fo and the other shared the same ITD, and vice versa. The target words began and ended with stop consonants, and the stop closures were made silent to minimize cues to source continuity other than ITD and Fo.

Method

Participants. The 14 participants were native speakers of British English between the ages of 21 and 52; all had pure-tone thresholds within the normal range at octave frequencies between 250 Hz and 4 kHz.

Stimuli. Two sentences, "Could you please write the word *bird* down now" and "You will also hear the sound *dog* this time," were spoken with a nearly flat intonation contour at around 125 Hz by a native speaker of British English (C.J. Darwin) and recorded in a soundproof booth onto digital audio tape. The sentences were digitized at 22050 Hz. The duration of the target word "dog" was lengthened and that of the target word "bird" shortened by adding or removing pitch periods from their centers to make them similar in duration. About 20 ms of silence was added to the beginning of the "Could you please . . ." sentence to align the target word onsets across the two sentences. The target words started about 1.24 s from the onset of the carrier sentences.

The two sentences were resynthesized on a monotone by means of a pitch-synchronous overlap-add (PSOLA) algorithm (Moulines & Charpentier, 1990) at Fos of 100, 106, 112.3, and 125 Hz, corresponding to approximately 0, 1, 2, and 4 semitones above 100 Hz. This range of Fo differences is sufficient to produce substantial segregation both in speech identification tasks (Assmann & Summerfield, 1990; Culling & Darwin, 1993; Scheffers, 1983) and in across-frequency integration of ITDs (Hill & Darwin, 1996).

To maintain alignment of target-word onsets, we made small adjustments to the silent closure interval before the target word. These adjustments compensated for the fact that PSOLA resynthesis rounds durations to whole numbers of pitch periods.

The target words "dog" and "bird" were then digitally switched around at stop-closure silences between various combinations of files to create a new set of files in which the target word did not have the same Fo as its carrier sentence. The durations of the acoustic segments in the 100-Hz versions of the two target words are given in Table 1. The target words and their immediate context had been chosen to minimize coarticulation across the stop closures (overall, listeners made 51% "bird" responses, so there was no bias toward "bird," the target word originally spoken in the attended sentence).

Procedure. Each listener was tested individually in a sound-attenuated booth. They were told that they would always hear the same two carrier sentences, which might come from the same or different positions. They should attend to the sentence "Could you please write the word *X* down now" and to press the *d* or *b* key if it contained the target word "dog" or "bird," respectively. On each trial the listener heard both carrier sentences and both target words.

Table 1
Durations (in Milliseconds) of Main Acoustic Segments in 100-Hz Versions of Target Words in Experiment 1

Word	Silence	Burst	Vocalic	Voice bar	Silence
Bird	83	13	210	41	78
Dog	32	15	200	66	54

Pairs of files, prepared as described above, were digitally mixed at presentation with ITDs of 0, ±45.3, ±90.7, and ±181.4 μs corresponding to 0, ±1, ±2, and ±4 samples at 22050 Hz. The term *±1 sample* indicates that one of the sentences led in one ear by 1 sample, and the other sentence led in the other ear by 1 sample. The ITDs were paired symmetrically so that if one sentence and target word had an ITD of +2 samples, the other had an ITD of −2 samples. The sentences when mixed at each headphone (Sennheiser 414, Wedemark, Germany) gave an average level of 68 dB (SPL) through a flat-plate coupler.

One carrier sentence and one target word always had an Fo of 100 Hz; the other carrier sentence and the other target had an Fo that was either the same or 1, 2, or 4 semitones higher. The attended carrier sentence was thus separated from the other sentence by seven different intervals (−4, −2, −1, 0, 1, 2, or 4 semitones).

For the trials on which the ITD was zero, these seven conditions were combined with two conditions in which the target word that had the same Fo as the attended sentence was either "dog" or "bird," resulting in a total of 14 conditions (2 of which were identical, with zero ITD and zero difference in Fo).

For the trials on which the ITD was not zero, there were three values of ITD combined with Fo difference (seven values), whether the target with the same ITD was "dog" or "bird" (two values), whether the attended sentence had a positive or a negative ITD (two values), or whether the target word with the same ITD as the carrier sentence also had the same Fo as the carrier sentence or not (two values). This combination resulted in a total of 168 conditions (some identical), which were presented five times each; each listener was presented with a different pseudorandom order.

Figure 1 illustrates the condition in which the ITDs were ±45 μs (the attended carrier is toward the left side), the Fos were 100 and 106 Hz (the attended carrier sentence had the lower Fo), and the "dog" target had the same ITD as the attended sentence but a different Fo.

Results

The data analysis was based on the number of "correct" target words reported by each listener (out of a maximum of five) for each different stimulus. For stimuli that had an ITD of zero, the correct target was defined as that with the same Fo as the attended carrier sentence. For stimuli that had an ITD not equal to zero, the correct target was defined as that with the same ITD as the attended carrier sentence. The latter data were subjected to an analysis of variance (ANOVA) with the following factors: ITD (±45, ±91, ±181 μs), Fo difference between the attended carrier sentence and the distractor (ΔFo = −4, −2, −1, 0, +1, +2, +4 semitones), correct target ("dog," "bird"), correct target's Fo relation to attended carrier (same, different), and side of attended sentence (left, right). The reported significance levels had the Greenhouse–Geisser correction for sphericity applied by means of SuperANOVA (Abacus Concepts, Berkeley, CA).

Continuity of Fo. When the two carrier sentences and target words have the same zero ITD, Fo is the only cue to which target word belongs with the attended carrier. Overall, across all trials on which there was an Fo difference, listeners chose the target word with the same Fo as the carrier on 57.4% (*SEM* = 1.4%) of trials, which is slightly, though significantly, above chance, $t(13) = 4.36$, $p < .001$.

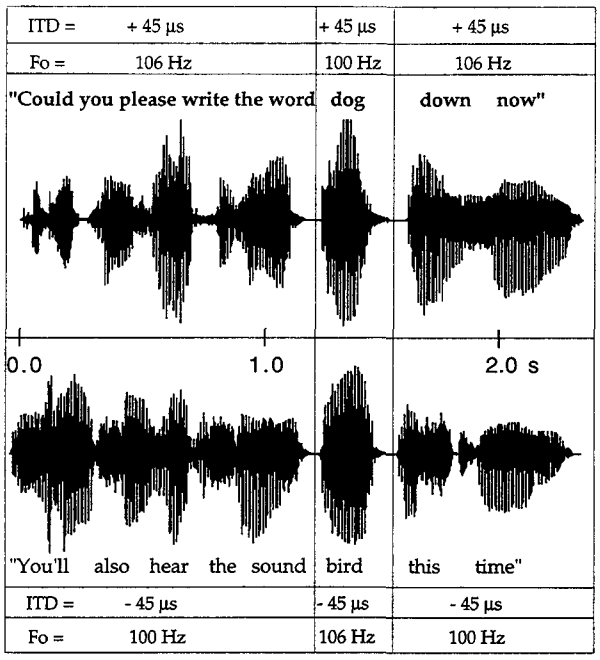

Figure 1. Example stimulus from Experiment 1. Two sentences were presented with interaural time differences (ITDs) of ±45 µs and fundamental frequencies (Fos) of 100 and 106 Hz (with the attended carrier sentence—in bold—having the higher Fo). The "dog" target had the same ITD as the attended sentence but a different Fo.

There was no reliable variation in this figure with absolute difference in Fo across the 14 listeners, $F(2, 26) = 1.4$. These data (see Figure 2A) show that listeners only weakly used continuity of Fo to identify the target word in the monotonous sentences used here.

Continuity of ITD when carrier Fos same. When the carriers and targets all had the same Fo but differed in ITD, listeners tended to report the target that had the same ITD as the attended carrier sentence. The percentage of trials on which listeners reported the target word that had the same ITD as the attended sentence is shown in Figure 2B. When ΔFo was zero, listeners were substantially above chance at all three ITDs (79%, 91%, and 94% correct for ITDs of ±45, ±91, and ±181µs, respectively); the increase with ITD was significant, $F(2, 26) = 23.2, p < .0001$.

Continuity of ITD when carrier Fos different. Listeners continued to report the target word that had the same ITD as the attended sentence when there were also differences in Fo present. Their performance increased slightly compared with a ΔFo of zero for the smallest ITD. As shown in Figure 2B, both the above-chance performance at an ITD of ±45 µs and the subsequent increase in performance with ITD, $F(2,$

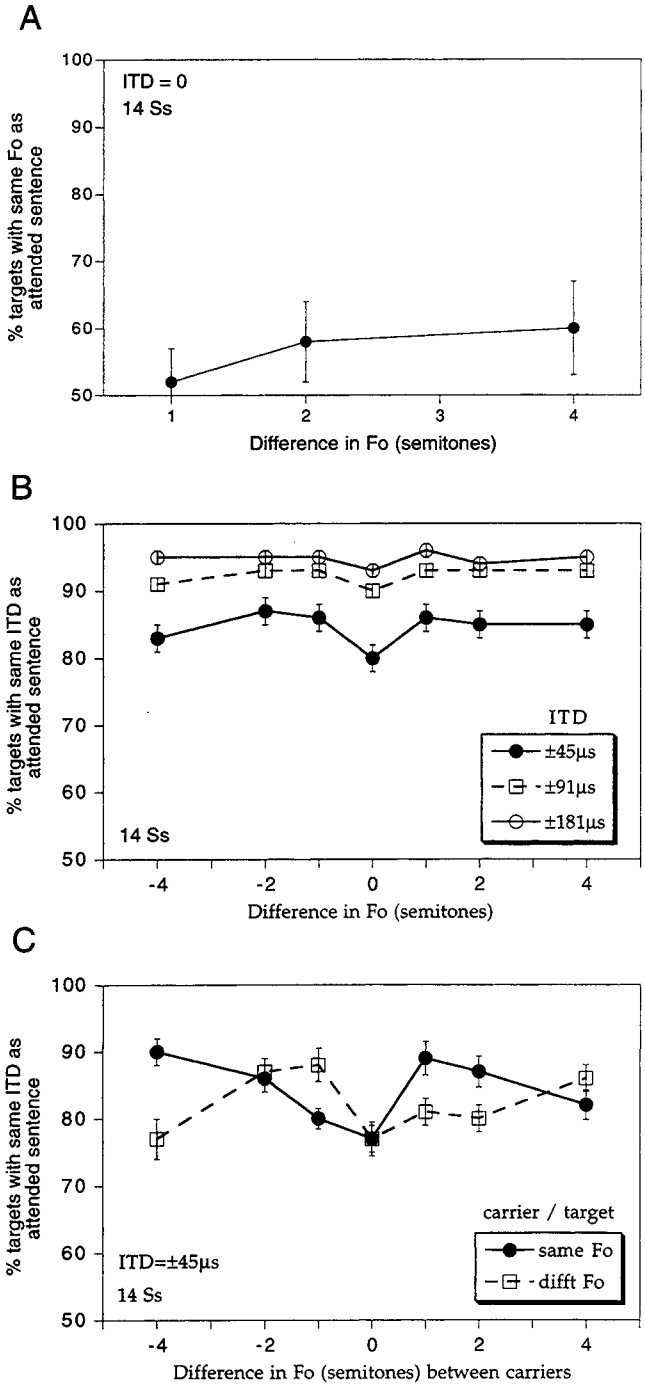

Figure 2 (opposite). A: Percentage of reported target words (±1 *SEM*) in Experiment 1 having the same fundamental frequency (Fo) as the attended sentence as a function of the difference in Fo between the two sentences on trials on which the interaural time difference (ITD) of each sentence was zero. Chance performance on this task was 50%. B: Percentage of reported target words (±1 *SEM*) in Experiment 1 having the same ITD as the attended sentence as a function of the difference in Fos between the two sentences. The parameter was the difference in ITDs between the two sentences. Chance performance on this task was 50%. C: Percentage of reported target words (±1 *SEM*) having the same ITD as the attended sentence as a function of the difference in Fos between the two sentences. The parameter was whether the correct target word (that shared the attended sentence's ITD) had the same Fo as the attended sentence or a different (difft) one. Ss = participants.

24) = 70.8, $p < .0001$, persist as the difference in Fo increases from 0 to 4 semitones.

Note that Figure 2B includes data from trials in which the target word that shared an ITD with the attended carrier sentence had either the same Fo as that sentence or a different Fo. This variable generally had rather little effect, confirming the slight role of Fo in this experiment. However, this variable did show an interaction with ΔFo, $F(6, 72) = 6.5, p < .005$. This interaction was more pronounced for an ITD of ±45 μs (because of ceiling effects at the larger ITDs) and is shown in Figure 2C. Because we adopted the convention that positive and negative values of ΔFo refer respectively to whether the attended sentence was higher or lower in Fo than the unattended sentence, the interaction shows that listeners preferred the target on the higher Fo for ΔFo of 1 and 2 semitones but the opposite at 4 semitones. The reason for this effect is not clear. Had listeners strongly used continuity of Fo to define the target word, the *same* Fo points in Figure 2C would have been consistently higher than the *different* Fo points. They clearly were not.

The overall ANOVA also showed some other, weakly significant interactions. The relative number of correct "dog" and "bird" responses varied weakly with the difference in Fo, $F(6, 78) = 2.6, p < .05$, and with ear of presentation, $F(1, 13) = 6.7, p < .05$, and these three variables gave a further weak four-way interaction with ITD, $F(2, 26) = 3.5, p < .05$. In addition, there was an interaction between ear and whether the target word had the same Fo as the carrier sentence, $F(1, 13) = 6.4, p < .05$; the left ear was more sensitive to the pitch manipulation than was the right ear. None of these interactions prejudiced the main conclusions drawn from the experiment.

Discussion

Listeners used differences in ITD much more effectively than they did differences in Fo to track a particular speaker over time. When the two carrier sentences did not differ in ITD, listeners showed only a weak preference to report the target word that had the same Fo as the carrier sentence. The largest Fo difference used in the experiment—4 semitones—gave only about 60% correct by Fo (against 50% chance). This rather surprising result may not extend to sentences that have natural intonation contours, unlike the monotones used here, or to Fo differences larger than 4 semitones, or to longer duration target words.

In contrast to the weak effect found here for a difference in Fo, an ITD difference between the two carrier sentences of only ±45 μs was sufficient to give a large and highly significant preference for the target word sharing the same ITD as the attended carrier sentence. Increasing the difference in ITD between the two carrier sentences further increased the preference. A time difference of ±45 μs corresponds to an angular separation between sources of about 10°. Our finding that this amount of separation produces above-chance tracking of a sound source across time is compatible with early experiments on selective attention. In one experiment (Spieth, Curtis, & Webster, 1954), listeners had to respond to one of two messages, each consisting of a call sign, a source identifier, and a question ("Oboe, this is Able 2, where in Box 5 is the triangle?") spoken over loudspeakers by two different voices. An angular separation of 10° or 20° increased the number of correctly named sources in the message containing the listener's call sign. The correct source (e.g., Able 2) was identified about 76% correctly with no spatial separation and about 92% correctly with 10–20° separation. Because there was a very limited number of call signs and source identifiers, it is likely that some of this improvement arose from listeners' using spatial cues to identify which source identifier followed the correct call sign. Such an effect is likely to have been smaller than in the experiment reported here, because different (male) voices were used as the two talkers, whereas a single voice was used here, and the key words were not synchronized.

Teder and Näätänen (1994) proposed a relatively narrow angular focus of auditory spatial attention on the basis of an experiment in which they used event-related potentials (ERPs). They measured ERPs to tones that came randomly from loudspeakers in different azimuthal positions as a function of whether the listener was attending to one or the other of two passages coming simultaneously from two of the loudspeakers that were separated by about 60°. They found that although the ERP N1 peak to tones that came from these two loudspeakers did show clear changes as a result of which speech message was being attended, peaks to tones coming from loudspeakers only 3° away from them showed a much reduced effect of which passage was being attended. They interpreted these results as indicating that auditory spatial attention has a narrow but graded focus.

It is surprising that a difference in ITD is almost as effective at allowing listeners to track a particular sound source when both sound sources are synthesized on the same Fo as when they are synthesized on different Fos. Adding together two sentences with the same Fo will produce a single set of harmonics at each ear; the amplitude and phase of each frequency in this set will be the vector sum of the components from the two constituent sentences. If a harmonic at a particular frequency from one sentence is instantaneously substantially more intense than the harmonic with the same frequency from the other sentence, its phase and amplitude will dominate the sum and so will have a broadly appropriate ITD. But if the amplitudes are similar, the resultant will in general have a very different phase and amplitude from the two constituents, leading to an ITD and an interaural level difference (ILD) that are inappropriate for either sound source.

For the speech materials used in this experiment, the amplitudes of individual harmonics in the two sentences were generally different, primarily because of instantaneous differences in formant frequencies and in the level of voiced excitation. For example, for the vowels of the target words "dog" and "bird" used here, first formant values were around 400 and 520 Hz, and second formant values were around 800 and 1400 Hz, respectively. If the harmonics near the formant peaks of one word are sufficiently different in level from the same frequency harmonics from the other

word, then there will be sufficient information in their ITDs to allow the listener, in principle, to hear an auditory object to the appropriate side. When the two target words were listened to simultaneously in isolation with the range of ITDs used in this experiment, the percept was clearly of two distinct words coming from different locations. But this observation raises another problem: If simultaneous grouping by ITD is weak, as previous experiments have demonstrated, how do the two auditory objects become separated? We return to this problem in the General Discussion section.

A further question concerns how reverberation might influence the effectiveness of ITDs and Fo in this paradigm. Plomp (1976) measured the speech reception threshold of connected discourse by masking one talker's speech by that of another. He found that increasing reverberation reduced the advantage of a spatial separation between the talkers: An angular separation of 135° reduced the threshold by 6 dB in anechoic conditions and by 2 dB with the reverberation time T_{60} of 1.4 s (T_{60} is the time for an impulsive sound to drop in level by 60 dB). Using a computer simulation of a reverberant room, Culling et al. (1994) measured the effectiveness of differences in Fo and also in simulated azimuth to reduce the threshold level for identifying a target vowel masked by another steady-state vowel-like sound. A reverberation time of 0.5 s was sufficient to remove the 8-dB advantage given by an angular separation of 120° under anechoic conditions. But the same reverberation time did not reduce the 16-dB advantage produced by giving the target an Fo that was a semitone higher. This resilience of a difference in Fo to reverberation was, however, abolished by giving the Fo of the target and of the masker a 5-Hz, ±2 semitone modulation. In light of the greater resilience of a steady difference in Fo to reverberation than a simulated difference in azimuth, it is perhaps surprising that in our experiment listeners used a difference in ITD more effectively than a steady difference in Fo to track a sound source across time. It was interesting to see whether this advantage persisted when the effects of reverberation were simulated.

Although it is dangerous to compare directly the effectiveness of one dimension, such as ITD, with another, such as Fo, when there is no independent way of equating the size of the manipulation in each dimension, we can contrast their relative effectiveness in two different types of grouping. Experiment 1 has shown that an ITD difference between two sentences of less than 100 μs provides a very effective cue for tracking a sound source over time; it is much more effective than continuity of Fo when the sentences are monotonous and differ by 4 semitones. By contrast, we know from other experiments that a difference in Fo (or harmonicity) of only a few semitones provides good simultaneous perceptual grouping (Assmann & Summerfield, 1990; Darwin & Gardner, 1986; Scheffers, 1979, 1983; Summerfield, 1992), whereas a large difference in ITD (of over ±600 μs) is ineffective at simultaneous perceptual segregation (Culling & Summerfield, 1995; Hukin & Darwin, 1995b). In the next experiment, we pursued this difference in the effectiveness of ITD in simultaneous and sequential grouping.

Experiment 2

In Experiment 1 the listener attended to a particular carrier sentence, which subjectively originated from a particular lateral position. Because all the frequency components that made up the carrier sentence were given the same ITD, the simplest mechanism to explain the results of the experiment is to suppose that listeners attended to those components that shared a common ITD. In Jeffress's (1948) model, this could readily be accomplished by grouping together the outputs of a column of coincidence detectors' responding to a common ITD. But this explanation would predict that we should also be able to perform simultaneous grouping by common ITD (see the left panel of Figure 3), which we know is not the case.

An alternative explanation (see the right panel of Figure 3) is to suppose that attention in Experiment 1 is directed to a particular subjective spatial direction but that the auditory object heard as coming from that direction may contain components that do not necessarily share the same ITD. According to this scheme, the ITD of each individual frequency component is calculated; in parallel with this operation, individual frequency components are grouped together by other grouping cues such as harmonicity and onset time (and also perhaps by phonetic criteria). The location of these groups can then be established from the ITDs of their component frequencies and attention directed to an auditory object in a particular direction. This scheme is similar to one proposed by Woods and Colburn (1992) and is compatible with experiments that have shown that (a) the lateral position of a complex sound can be determined by a weighted averaging of ITDs across its frequency components (Jeffress, 1972; Shackleton, Meddis, & Hewitt, 1992; Trahiotis & Stern, 1989) and (b) other, monaural grouping

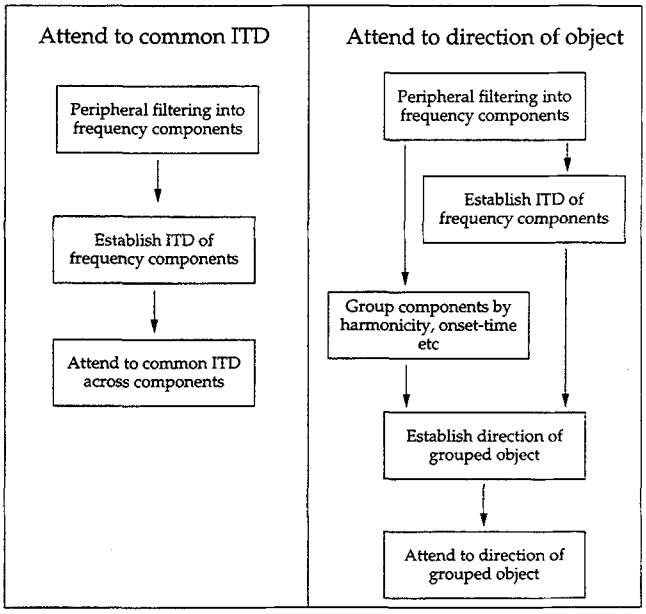

Figure 3. Two theoretical frameworks for interpreting the results of Experiment 1. ITD = interaural time difference.

cues (harmonicity and onset time) determine across which frequency components the weightings of ITD are made (Hill & Darwin, 1996).

If this alternative explanation has value, then we should be able to contrast the extremely effective tracking by common ITD of a target within a carrier sentence that occurred in Experiment 1 (where all the components of the source share the same ITD) with a situation in which, although the components of a target sound source do not all share the same ITD, the whole sound is nevertheless heard as coming from the same direction as the carrier sentence. If the target is constructed to give a different percept depending on whether the part with a different ITD is included or not, then we can distinguish between the two explanations. If listeners are really tracking a particular ITD, then they should perceive the target *excluding* the part that has a different ITD. If listeners are tracking the location of auditory objects, however, they should perceive the target *including* the part that has a different ITD. We tested this prediction in Experiments 2 and 3.

An ILD is not generally a naturally useful cue for the localization of low-frequency sounds. Low-frequency sounds diffract around the head, producing only a small ILD. An exception arises for sounds that are very close to one ear, when the inverse-square law produces level differences that are independent of frequency. Artificial, large interaural intensity differences (such as those generated by playing a sound to only one headphone) provide both a strong lateralization cue, with the sound heard extremely lateralized to one ear, and a stronger simultaneous grouping cue than are provided by large differences in ITD (Culling & Summerfield, 1995; Hukin & Darwin, 1995b, 1995c). In the following experiments we included conditions in which sounds are played to only one headphone ("infinite" ILD). On the basis of previous experiments, we expected an infinite ILD to provide more segregation than a large ITD and consequently for a tone given a different ILD from the rest of the vowel to be more excluded by tracking as a separate auditory object.

Method

An appropriate paradigm for testing these ideas is to use the /ɪ/–/ɛ/ phoneme categorization task that we and others have used previously to investigate simultaneous grouping (Darwin, 1984; Darwin & Gardner, 1986; Darwin & Sutherland, 1984; Hukin & Darwin, 1995a; Roberts & Moore, 1991). Listeners are asked to label vowels that differ in their first-formant (F1) frequency as /ɪ/ or /ɛ/, and their F1 phoneme boundary is established. Physical removal of a harmonic that is just higher in frequency than F1 leads to a more /ɪ/-like percept, with a consequent shift in the phoneme boundary to a higher (nominal) F1 frequency. Conversely, a physical increase in the level of the same harmonic gives a lower F1 boundary. Such boundary shifts can be used to detect perceptual, rather than physical, segregation of the harmonic from the vowel by manipulations that maintain its physical presence. Differences in onset time, harmonicity, and ILD between the harmonic and the rest of the vowel have given upward shifts in the phoneme boundary, thus providing evidence for perceptual segregation from the vowel. Large differences in ITD, however, are not able to segregate the harmonic unless accompanied by other cues to perceptual segregation (Darwin & Hukin, 1997, 1998; Hukin & Darwin, 1995b).

In Experiment 2 we presented one harmonic of a vowel with an ITD different from that of the rest of the vowel. We expected to find that this difference in ITD alone is insufficient to segregate the harmonic from the vowel, as measured by a shift in phoneme boundary. We also asked whether this lack of simultaneous segregation by ITD of a harmonic from a vowel persists when the vowel is presented in a sentence context. If listeners in Experiment 1 were tracking a common ITD, as a function of time, then they should be able to segregate the harmonic from the vowel on the basis of the common ITD, just as they were able to determine which target word was appropriate in Experiment 1. Segregation by ITD should, on the basis of this hypothesis, be increased substantially by putting the vowel in a sentence context with the same ITD as the vowel. However, if listeners in Experiment 1 were tracking a location rather than a common ITD, then the sentence context will not increase segregation by ITD of the harmonic from the vowel.

Because there is already evidence that an ILD, rather than an ITD, does produce some simultaneous segregation, ILD is also included in the experiment as a comparison. We expected to find some segregation due to ILD for conditions in which the vowel was presented alone and an increase in this segregation as a result of placing the vowel in a sentence context. Specifically, if listeners are tracking the location of sound sources, we expected on the basis of previous experiments to find that the sentence context will be more effective at excluding a harmonic with an infinite ILD from the vowel percept than one with a large ITD.

Stimuli. Formant-synthesizer (Klatt, 1980) parameters from a previous experiment (Darwin, McKeown, & Kirby, 1989) for the carrier sentence "Hello, you'll hear the sound [bit] now" were edited to produce a monotone sentence (Fo = 150 Hz, duration = 2.33 s). The parameters were based on a linear predictive coding analysis of a natural sentence (speaker C.J. Darwin). The parameters for the original target words ("bit" and "bet") were edited to produce a continuum of steady-state vowels differing only in F1, which was heard as moving from /ɪ/ to /ɛ/ as F1 increased in frequency. Care was taken to ensure that the target vowel fitted naturally into the carrier sentence by adjusting the steady-state formant frequencies and giving the vowel a natural amplitude envelope. The original continuum had eight members whose F1 ranged from 480 Hz to 620 Hz in 20-Hz steps. The target vowel started 1.46 s into the sentence, after a 70-ms silence; it lasted 160 ms and had F2–F4 set to 1800, 2600 and 3400 Hz, respectively.

The 600-Hz (fourth) harmonic of each of the eight target vowels was extracted with a finite impulse response filter ($n = 301$), and a new continuum in which the 600-Hz component was absent (no-600) was created by subtracting these waveforms (shifted by 150 samples to counter the lag of the filter) from their original vowel. The filtering was not applied to the carrier sentence. The 600-Hz waveforms created by the filtering were also used in some of the stimulus conditions described below.

The experiment had two groups of conditions: one (vowel in sentence) in which the target vowel occurred in the carrier sentence, and one (vowel alone) in which it was presented alone. Within the first group there were eight conditions; in the second, nine. Each condition consisted of a continuum of eight vowel sounds, each derived from the corresponding sound from the original F1 continuum. Four of the conditions were as follows:

1. *Same ILD 0 dB:* The original carrier sentence and target vowel were presented to the left ear, with the 600-Hz component of the target vowel at its original level.

2. *Different ILD 0 dB:* Same as ILD 0 dB but with the no-600 wave played to the left ear and the 600-Hz component of the target vowel played to the right ear at its original level.

3. *Same ITD 0 dB:* Same as ILD 0 dB, but both the no-600 and the 600-Hz waves were presented to both ears, with an ITD of +635 μs (leading on the left ear) applied to both waves.

4. *Different ITD 0 dB:* Same as ITD 0 dB but with the no-600 wave given an ITD of +635 μs and the 600-Hz wave given an ITD of −635 μs.

Four more (6 dB) conditions corresponded to the above four 0-dB conditions, but the level of the 600-Hz component of the vowel was increased by 6 dB. These 6-dB conditions were included to allow a greater effect of the perceptual removal of the 600-Hz component. In previous experiments in which this paradigm was used, the perceptual removal of a +6-dB component has been easier to detect than the removal of an unchanged one. The different ILD 6-dB condition with the sentence carrier is shown in Figure 4. Finally, there was a no-600 condition in the vowel-alone group of conditions that was the same as ILD 0 dB but with the 600-Hz component of the vowel filtered out.

Procedure. The 14 participants from Experiment 1 first completed the vowel-alone group of nine conditions as a separate experiment; on a separate day they completed the vowel-in-sentence group of nine conditions. They were told that they would hear (a carrier sentence with) a vowel in their left ear, which could be either /ɪ/ as in *pit* or /ɛ/ as in *pet* and that they might also hear a tone in their right ear, which they were to ignore. They signaled their response on each trial using the *i* and *e* keys on the Macintosh keyboard. Each sound followed 500 ms after the response to the previous one. Listeners could repeat the previous sound by pressing the "escape" key.

All listeners were native speakers of British English with normal pure-tone thresholds over the range of frequencies of interest in this experiment. The sounds were presented at an overall gain such that the 600-Hz component of the 0-dB vowel with F1 at 600 Hz had a level of 60 dB (SPL).

Results

Phoneme boundaries were estimated (by a least squares fit of a rescaled tanh function) from the number of *i*-key responses to the 10 repetitions of the eight stimuli, differing in F1, in each condition for each listener. The calculated boundaries were all checked by eye. The boundaries of 4 of the listeners in the no-600 Hz condition were too high to be reliably estimated with the range of F1 values that we used. The boundaries for these 4 participants were conservatively placed at 640 Hz for this condition. The average F1 phoneme boundaries across listeners are shown in Figure 5.

Physical changes to 600 Hz. Physical changes to the 600-Hz component had the expected effect. Compared with the 0-dB same ILD and same ITD conditions, removing the 600-Hz component substantially increased (by at least 60 Hz) the frequency of the phoneme boundary. Conversely, increasing the gain of the 600-Hz component by 6 dB decreased the phoneme boundary by about 30 Hz. These results validate the basic paradigm as being sensitive to changes in the relative level of the 600-Hz component.

ILD changes. For the vowel-alone conditions, putting the 600-Hz component on the opposite ear significantly reduced the effect of increasing the level of the 600-Hz component, $F(1, 13) = 19.5, p < .001$, confirming previous results that an infinite ILD produces some segregation of a harmonic from a vowel. Although there was strong segregation by ILD for the 6-dB condition, in these data there was no evidence of segregation by ILD in the 0-dB condition. In a previous experiment (Hukin & Darwin, 1995b) in which we used an infinite ILD but slightly different vowel stimuli, we also found a greater shift for the 6-dB condition than for the 0-dB condition, but the shift at 0 dB was more substantial than that found here.

Placing the vowel in a sentence context that has the same ILD as the body of the vowel has a substantial effect: The 600-Hz component is significantly more segregated from the vowel when it is put in the opposite ear than when it is in the same ear as the rest of the vowel, $F(1, 13) = 23.3, p < .0005$. The boundary increases very substantially in both the 0-dB and 6-dB conditions, $F(1, 13) = 55.6, p < .0001$, with a larger shift in the 6-dB condition, $F(1, 13) = 44.8, p < .0001$. Both the 0-dB and the 6-dB boundaries are comparable to the (albeit conservatively estimated) no-600 bound-

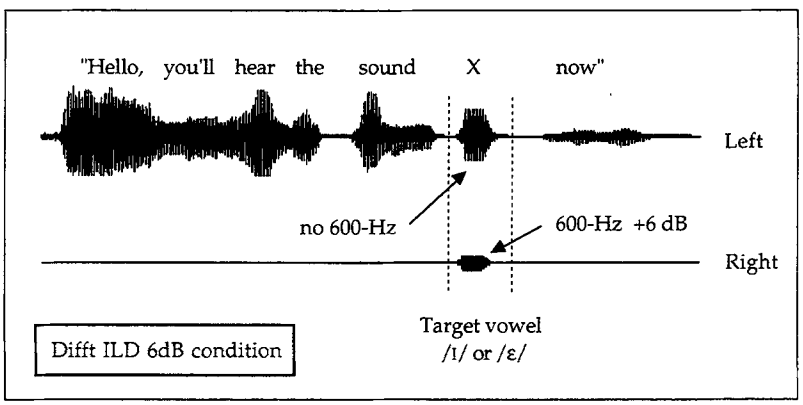

Figure 4. Example stimulus from Experiment 2 (vowel in sentence, different interaural level difference [difft ILD] 6 dB). The synthetic carrier sentence was presented to the left ear on a fundamental frequency of 150 Hz. The target vowel was also presented to the left ear but without its 600-Hz component. The 600-Hz component was given an additional gain of 6 dB and was presented to the right ear.

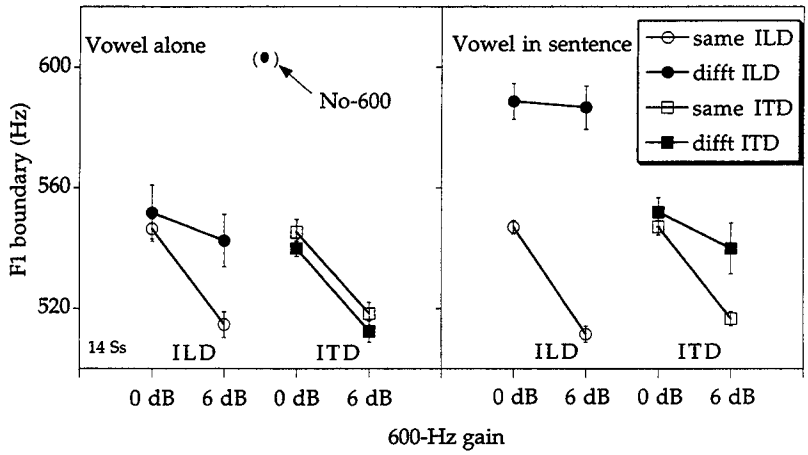

Figure 5. /ɪ/ − /ɛ/ phoneme boundaries (±1 *SEM*) from Experiment 2. The vowel and sentence were always heard on the left side by virtue either of an interaural time difference (ITD) of +635 μs or an infinite interaural level difference (ILD). The 600-Hz component of the vowel had either the same or the opposite sign of ITD or ILD and could be boosted by 6 dB. Across conditions with identical levels, higher first-formant boundaries implied more perceptual segregation of the 600-Hz harmonic from the vowel. difft = different; Ss = participants.

ary where the 600-Hz component was physically removed. These results show that placing the vowel in a sentence context can be a very effective way to increase segregation.

ITD changes. For the vowel-alone conditions, there was no effect of giving the 600-Hz component a different ITD from the rest of the vowel. This result confirms previous findings of the weakness of ITD when it is the only cue for perceptual segregation. Adding the carrier sentence does give evidence of some segregation when the 600-Hz component has a different ITD from the carrier and the rest of the vowel. It reduces the effect of the 6-dB additional gain, $F(1, 13) = 11.4, p < .005$, by elevating the phoneme boundary at 6 dB but not at 0 dB. Nevertheless, the phoneme-boundary shifts produced by the sentence context with ITDs are much smaller than those produced with ILDs.

Differences between ILD and ITD changes. The different pattern of results found between the vowel-alone and the vowel-in-sentence conditions is reflected in a significant three-way interaction between vowel alone–vowel in sentence, ITD–ILD, and same–different, $F(1, 13) = 5.0, p < .05$. This interaction was also present for the more natural 0-dB conditions, $F(1, 13) = 5.8, p < .05$. The interaction confirms that putting the vowel in a sentence context gives a greater increase in segregation for a difference in ILD than for a difference in ITD.

The third experiment was very similar to Experiment 2, so we discuss the results of both experiments together. Whereas in Experiment 2 there was no uncertainty as to which ear the target sentence or the isolated target vowel would come, in Experiment 3 the side to which the vowel base of the carrier sentence was played was randomly varied. In Experiment 1 the attended carrier sentence occurred randomly on the left or the right side, so Experiment 3 was generally more like Experiment 1 by having a variable side of presentation. More specifically, there is evidence that reliable cues can direct attention endogenously to a particular side (Spence & Driver, 1994). Participants may find it easier to use ITD or ILD to segregate a harmonic from a vowel when they know the side of auditory space to which the vowel will be presented. Any such effect is likely to be greater in the vowel alone than in the sentence context, because the part of the sentence before the target vowel is longer than the time it takes participants to orient either exogenous or endogenous attention.

Experiment 3

Method

Experiment 3 was identical to Experiment 2 except that the target sentence and vowel base were played either both to the left or both to the right side at random from trial to trial. All but 1 of the 14 participants involved in Experiment 2 took part in this experiment. Their instructions were similar to those in the previous experiment except that they were told that they would hear (a carrier sentence with) a vowel that could be presented randomly toward either their left or right ear.

Results

Phoneme boundaries were estimated as before from the number of *i*-key responses to the five repetitions of each stimulus, differing in F1, in each condition for each listener. Again, the boundaries of 4 of the listeners in the no-600 Hz condition were too high to be reliably estimated with the range of F1 values that we used and were conservatively placed at 640 Hz. The average F1 boundaries across listeners are shown in Figure 6. The effect of side was not significant; thus all results are shown averaged across side of presentation. The results of Experiment 3 are almost identical to those of Experiment 2.

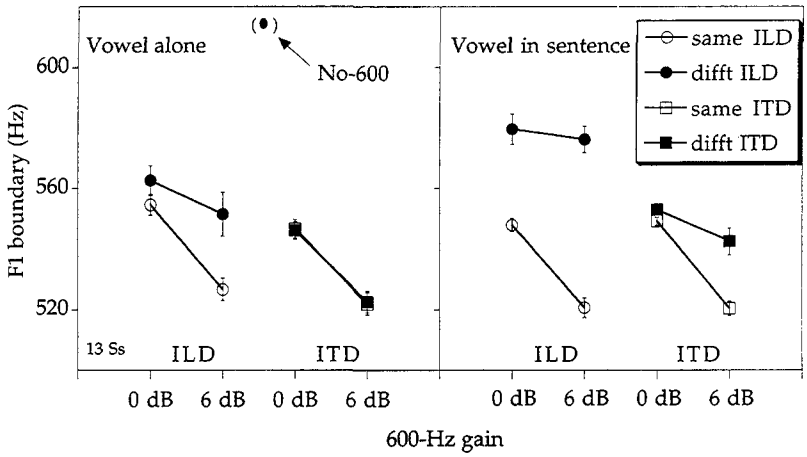

Figure 6. /ɪ/ − /ɛ/ phoneme boundaries (±1 *SEM*) from Experiment 3. The experiment was similar to Experiment 2 except that across trials, the interaural time differences (ITDs) and interaural level differences (ILDs) randomly varied in sign so that listeners heard the vowel and the carrier sentence on either the left or the right side. difft = different; Ss = participants.

Physical changes to 600-Hz. As in Experiment 2, removing the 600-Hz component leads to a substantial increase in the frequency of the phoneme boundary over the 0-dB ILD and ITD conditions. Increasing the level of the 600-Hz component by 6 dB decreases the phoneme boundary by about 30 Hz.

ILD changes. For the vowel-alone conditions, putting the 600-Hz component on the opposite ear significantly reduced the effect of increasing the level of the 600-Hz component, $F(1, 12) = 11.5$, $p < .01$. However, in the vowel-in-sentence conditions, the effect of putting the 600-Hz component in the opposite ear was substantially greater, $F(1, 12) = 33.5$, $p < .0001$. Here the boundary shifted very substantially in both the 0-dB, $F(1, 12) = 35.2$, $p < .0001$, and 6-dB conditions, with a larger shift in the 6-dB condition, $F(1, 12) = 99.5, p < .0001$.

ITD changes. For the vowel-alone conditions, there was no effect of giving the 600-Hz component a different ITD from the rest of the vowel. Adding the carrier sentence reduced the effect of the 6-dB additional gain, $F(1, 12) = 17.7, p < .005$.

Differences between ILD and ITD changes. The different pattern of results found between the vowel-alone and the vowel-in-sentence conditions is reflected in a significant three-way interaction between vowel alone–vowel in sentence, ITD–ILD, and same–different, $F(1, 12) = 8.1$, $p < .05$. This interaction was also present for the more natural 0-dB conditions, $F(1, 12) = 13.8, p < .005$.

General Discussion

Overall, the results of Experiments 2 and 3 support the idea that when listeners attend to a sound whose direction is determined by ITDs, they do this on the basis of the subjective direction of the whole auditory object rather than by attending only to those frequency components that share a common interaural time difference.

Experiments 2 and 3 have confirmed our previous findings that a difference of ITD alone is not effective at segregating a harmonic from a vowel: For the vowel-alone conditions, phoneme boundaries did not change between the same ITD and different ITD conditions. Putting the vowel in a sentence context with the same ITD as the main body of the vowel does produce some segregation of the harmonic when the ITD is different, but this segregation is largely limited to the case in which the harmonic has an increased level (different ITD 6 dB). There is thus only a small increase in the segregation provided by a large difference in ITD (±635 μs) between a harmonic and the rest of a vowel when the vowel is embedded in a sentence with the same ITD.

These results contrast markedly with the results of Experiment 1, in which listeners were about 80% correct in saying which of two target words differing in ITD by only ±45 μs belonged in the attended sentence. That such embedding of the vowel in a target sentence could in principle be effective at increasing segregation was shown by the very substantial increase in such segregation produced by playing the sounds to one ear only (ILD condition).

It is difficult to see how the marked difference in the results of these two types of experiment could be accounted for by allowing auditory attention to be paid to frequency components that share a common ITD, as in the left-hand panel of Figure 5. If the results of Experiment 1 were due to listeners' being capable of attending to those frequency components that shared a common ITD, we would have expected there to be some segregation by ITD in the vowel-alone condition and strong segregation by ITD when the vowel was embedded in a sentence (making the stimulus conditions closer to those of Experiment 1). Neither of these outcomes occurred.

The results of Experiments 2 and 3 can, however, be interpreted using the theoretical scheme outlined in the

right-hand panel of Figure 5. Because the harmonic is synchronous and harmonically related with the rest of the vowel, it will tend to be grouped with it, with a difference in ITD exerting only a weak segregating influence. The whole vowel (including the 600-Hz harmonic) is then labeled and localized by an across-frequency weighting (Trahiotis & Stern, 1989) or integration (Shackleton et al., 1992) of ITDs.

The same scheme can also handle the results of Experiment 1. When the two target words are on the same Fo, there are presumably enough dynamic cues such as small onset-time and offset-time differences, amplitude trajectories, and perhaps also phonetic plausibility (Remez et al., 1994) to segregate, at least partially, those harmonics whose level is determined mainly by "dog" from those determined mainly by "bird." When there is a difference in Fo, this is a major cue to segregation (Assmann & Summerfield, 1990; Bird & Darwin, 1998; Culling & Darwin, 1993; Scheffers, 1983). The two auditory objects formed by the two groups of segregated harmonic frequencies can then be localized. The stability of the lateralized percept in the face of ITDs that change with the relative levels of a particular harmonic in the two sentences and targets may be helped by the well-known sluggishness of the binaural system: Listeners are insensitive to rapid changes in ITD over time (Grantham, 1986; Grantham & Wightman, 1978; Kollmeier & Gilkey, 1990). More speculatively, listeners may also be able to allocate some of the energy of a single harmonic to one sound source and the rest to another on the basis of the available dynamic information (Darwin, 1995; Warren, Bashford, Healey, & Brubaker, 1994).

Experiment 3 gave very similar results to Experiment 2 even though participants did not know to which side the isolated vowel or the sentence would be presented. For the sentence condition, the lack of any substantially greater segregation in Experiment 2 than in Experiment 3 is not surprising because participants have ample time to direct attention to the sentence before the target word arrives. It is more surprising in the isolated vowel condition. Further work is needed to clarify the relation between the endogenous and exogenous shifts of attention discussed by Spence and Driver (1994) and the direction of attention to complex simultaneous sounds.

There is an apparent inconsistency between our results from Experiments 2 and 3 concerning the segregation in the ILD 0-dB condition. Although it was clear both from earlier experiments and from results of the 6-dB condition in the vowel-alone conditions that segregation was greater for ILD than for ITD, this was not the case for the 0-dB conditions. Why then does the sentence context have a larger effect on the ILD 0-dB than on the ITD 0-dB conditions, when neither of them shows appreciable segregation when presented alone without a sentence?

The answer may lie in different types of segregation. In Experiments 2 and 3 we measured the segregation of a harmonic by the change it produced in vowel quality. One could also measure segregation by asking listeners whether they could hear out a harmonic as a separate sound source (Moore, Peters, & Glasberg, 1985) or by a change in other properties such as pitch (Darwin & Ciocca, 1992) or localization (Hill & Darwin, 1996). There are clear quantitative differences between segregation measured in these different ways (Darwin & Carlyon, 1995; Hukin & Darwin, 1995a); it is generally easier to segregate part of a complex sound so that it can be heard out as a separate source than it is to remove it from the calculation of pitch or vowel quality (a form of duplex perception). It is possible that in the ILD 0-dB condition, listeners were able to hear out the 600-Hz component as a separate sound source even though they still included it in the calculation of vowel quality. Although we did not question listeners, our own observations suggest that this is very likely to be the case. Its segregation from the vowel could then have been enhanced by being placed in a sentence context.

In summary, the experiments reported here have shown the following:

1. Listeners can use a small (± 45 μs) difference in ITD between two sentences to say which of two target words was part of an attended sentence but were substantially less able to use differences in Fo—a difference of 4 semitones produced performance that was only slightly above chance.

2. By contrast, a large difference in ITD is not sufficient to exclude a harmonic from a vowel percept when the vowel is in a carrier sentence with the same ITD as the main part of the vowel. The carrier sentence does, however, have a large effect on a harmonic that differs (infinitely) in ILD from the vowel and the carrier sentence.

These results can be explained by assuming that auditory attention is directed toward objects in subjective locations rather than toward those frequency components that share a particular ITD. Such an assumption may allow work on the purely auditory aspects of attention (with which this article has been concerned) to interface with recent work on cross-modal attention (Driver & Spence, 1994; Spence & Driver, 1996, 1997), including the remarkable finding of a strong effect of subjective direction induced by the ventriloquism effect (Bertelson & Radeau, 1981) on listeners' abilities to separate simultaneous voices (Driver, 1996).

References

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

Barker, J., & Cooke, M. (1999). Is the sine-wave speech cocktail party worth attending? *Speech Communication, 27,* 159–174.

Bertelson, P., & Radeau, M. (1981). Cross-modal bias and perceptual fusion with auditory–visual spatial discordance. *Perception & Psychophysics, 29,* 578–584.

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

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* Cambridge, MA: Bradford Books/MIT Press.

Bregman, A. S., Liao, C., & Levitan, R. (1990). Auditory grouping based on fundamental frequency and formant peak frequency. *Canadian Journal of Psychology, 44,* 400–413.

Broadbent, D. E. (1953). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology, 47,* 191–196.

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

Cherry, E. C., & Taylor, W. K. (1954). Some further experiments upon the recognition of speech, with one and with two ears. *Journal of the Acoustical Society of America, 26,* 554–559.

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

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

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.

Darwin, C. J. (1975). On the dynamic use of prosody in speech perception. In A. Cohen & S. G. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 178–194). Berlin: Springer-Verlag.

Darwin, C. J. (1981). Perceptual grouping of speech components differing in fundamental frequency and onset-time. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 33A,* 185–208.

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

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

Darwin, C. J. (1995). Perceiving vowels in the presence of another sound: A quantitative test of the "old-plus-new" heuristic. In C. Sorin, J. Mariani, H. Méloni, & J. Schoentgen (Eds.), *Levels in speech communication: Relations and interactions: A tribute to Max Wajskop* (pp. 1–12). Amsterdam: Elsevier.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Science, 1,* 327–333.

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 B. C. J. Moore (Ed.), *The handbook of perception and cognition: Vol. 6. Hearing* (pp. 387–424). London: Academic Press.

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., & 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,* 838–845.

Darwin, C. J., & Hukin, R. W. (1997). Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *Journal of the Acoustical Society of America, 102,* 2316–2324.

Darwin, C. J., & Hukin, R. W. (1998). Perceptual segregation of a harmonic from a vowel by interaural time difference in conjunction with mistuning and onset-asynchrony. *Journal of the Acoustical Society of America, 103,* 1080–1084.

Darwin, C. J., McKeown, J. D., & Kirby, D. (1989). Compensation for transmission channel and speaker effects on vowel quality. *Speech Communication, 8,* 221–234.

Darwin, C. J., & Sutherland, N. S. (1984). Grouping frequency components of vowels: When is a harmonic not a harmonic? *Quarterly Journal of Experimental Psychology: Human Experimental Psychology, 36A,* 193–208.

Deutsch, D. (1979). Binaural integration of melodic patterns. *Perception & Psychophysics, 25,* 399–405.

Driver, J. (1996, May 2). Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature, 381,* 66–68.

Driver, J., & Spence, C. J. (1994). Spatial synergies between auditory and visual attention. In C. Umiltà & M. Moscovich (Eds.), *Attention and performance XV: Conscious and nonconscious information processing* (pp. 311–331). Cambridge, MA: MIT Press.

Grantham, D. W. (1986). Detection and discrimination of simulated motion of auditory targets in the horizontal plane. *Journal of the Acoustical Society of America, 79,* 1939–1949.

Grantham, D. W., & Wightman, F. L. (1978). Detectability of varying interaural temporal differences. *Journal of the Acoustical Society of America, 63,* 511–523.

Green, K. P., Stevens, E. B., & Kuhl, P. K. (1994). Talker continuity and the use of rate information during phonetic perception. *Perception & Psychophysics, 55,* 249–260.

Hill, N. I., & Darwin, C. J. (1996). Lateralization of a perturbed harmonic: Effects of onset asynchrony and mistuning. *Journal of the Acoustical Society of America, 100,* 2352–2364.

Hukin, R. W., & Darwin, C. J. (1995a). Comparison of the effect of onset asynchrony on auditory grouping in pitch matching and vowel identification. *Perception & Psychophysics, 57,* 191–196.

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

Hukin, R. W., & Darwin, C. J. (1995c). Grouping of vowel components by common interaural time differences. *British Journal of Audiology, 29,* 78.

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

Jeffress, L. A. (1972). Binaural signal detection: Vector theory. In J. V. Tobias (Ed.), *Foundations of modern auditory theory* (Vol. 2, pp. 349–368). New York: Academic Press.

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

Kollmeier, B., & Gilkey, R. H. (1990). Binaural forward and backward masking: Evidence for sluggishness in binaural detection. *Journal of the Acoustical Society of America, 87,* 1709–1719.

Lotto, A. J., Kluender, K. R., & Green, K. P. (1996). Spectral discontinuities and the vowel length effect. *Perception & Psychophysics, 58,* 1005–1014.

Moore, B. C. J., Peters, R. W., & Glasberg, B. R. (1985). Thresholds for the detection of inharmonicity in complex tones. *Journal of the Acoustical Society of America, 77,* 1861–1868.

Moulines, E., & Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication, 9,* 453–467.

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.

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,* 129–156.

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. *Journal of the Acoustical Society of America, 89,* 2922–2932.

Scheffers, M. T. (1979). The role of pitch in perceptual separation of simultaneous vowels. *Institute for Perception Research: Annual Progress Report, 14,* 51–54.

Scheffers, M. T. (1983). *Sifting vowels: Auditory pitch analysis and sound segregation.* Unpublished doctoral dissertation, Groningen University, the Netherlands.

Schubert, E. D., & Parker, C. D. (1956). Addition to Cherry's findings on switching speech between the two ears. *Journal of the Acoustical Society of America, 27,* 792–794.

Shackleton, T. M., Meddis, R., & Hewitt, M. J. (1992). Across frequency integration in a model of lateralization. *Journal of the Acoustical Society of America, 91,* 2276–2279.

Spence, C. J., & Driver, J. (1994). Covert spatial orienting in audition: Exogenous and endogenous mechanisms. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 555–574.

Spence, C., & Driver, J. (1996). Audiovisual links in endogenous covert spatial attention. *Journal of Experimental Psychology: Human Perception and Performance, 22,* 1005–1030.

Spence, C., & Driver, J. (1997). Audiovisual links in exogenous covert spatial orienting. *Perception & Psychophysics, 59,* 1–22.

Spieth, W., Curtis, J. F., & Webster, J. C. (1954). Responding to one of two simultaneous messages. *Journal of the Acoustical Society of America, 26,* 391–396.

Summerfield, A. Q. (1992). Roles of harmonicity and coherent frequency modulation in auditory grouping. In M. E. H. Schouten (Ed.), *The auditory processing of speech: From sounds to words* (pp. 157–165). Berlin: Mouton de Gruyter.

Teder, W., & Näätänen, R. (1994). Event-related potentials demonstrate a narrow focus of auditory spatial attention. *Neuroreport, 5,* 709–711.

Trahiotis, C., & Stern, R. M. (1989). Lateralization of bands of noise: Effects of bandwidth and differences of interaural time and phase. *Journal of the Acoustical Society of America, 86,* 1285–1293.

Warren, R. M., Bashford, J. A., Healey, E. W., & Brubaker, B. S. (1994). Auditory induction: reciprocal changes in alternating sounds. *Perception & Psychophysics, 55,* 313–322.

Wightman, F. L., & Kistler, D. J. (1992). The dominant role of low-frequency interaural time differences in sound localization. *Journal of the Acoustical Society of America, 91,* 1648–1661.

Woods, W. A., & Colburn, S. (1992). Test of a model of auditory object formation using intensity and interaural time difference discriminations. *Journal of the Acoustical Society of America, 91,* 2894–2902.

Yin, T. C. T., & Chan, J. C. K. (1990). Interaural time sensitivity in the medial superior olive of the cat. *Journal of Neurophysiology, 64,* 465–488.

Received October 27, 1997
Revision received February 17, 1998
Accepted April 20, 1998 ■