Effectiveness of spatial cues, prosody, and talker characteristics in selective attention

C. J. Darwin[a] and R. W. Hukin
Experimental Psychology, University of Sussex, Brighton BN1 9QG, United Kingdom

(Received 26 May 1999; accepted for publication 22 October 1999)

The three experiments reported here compare the effectiveness of natural prosodic and vocal-tract size cues at overcoming spatial cues in selective attention. Listeners heard two simultaneous sentences and decided which of two simultaneous target words came from the attended sentence. Experiment 1 used sentences that had natural differences in pitch and in level caused by a change in the location of the main sentence stress. The sentences' pitch contours were moved apart or together in order to separate out effects due to pitch and those due to other prosodic factors such as intensity. Both pitch and the other prosodic factors had an influence on which target word was reported, but the effects were not strong enough to override the spatial difference produced by an interaural time difference of ± 91 μs. In experiment 2, a large ($\pm 15\%$) difference in apparent vocal-tract size between the speakers of the two sentences had an additional and strong effect, which, in conjunction with the original prosodic differences overrode an interaural time difference of ± 181 μs. Experiment 3 showed that vocal-tract size differences of $\pm 4\%$ or less had no detectable effect. Overall, the results show that prosodic and vocal-tract size cues can override spatial cues in determining which target word belongs in an attended sentence. © *2000 Acoustical Society of America.* [S0001-4966(00)01302-3]

PACS numbers: 43.66.Pn, 43.71.Es [RVS]

INTRODUCTION

This paper addresses the general problem of how listeners attend to a particular sound source over time. Cherry's original paper on the ''cocktail-party effect'' (Cherry, 1953) and subsequent work on auditory selective attention by Broadbent (1953) and others (Spieth *et al.*, 1954) emphasized the spatial nature of auditory attention. A number of recent papers have studied the consequences for audition of our ability to direct attention, either intentionally or automatically, to a particular spatial direction (Spence and Driver, 1994; Teder and Näätänen, 1994; Mondor and Zattore, 1995; Quinlan and Bailey, 1995). Auditory attention, though, cannot be purely concerned with selecting between spatially distinct sound sources, since attention is possible, though more difficult, between sources that are not spatially separated. Indeed, there is a variety of demonstrations of listeners' ability to attend selectively to particular frequency regions (Scharf *et al.*, 1987; Schlauch and Hafter, 1991; Hubner and Hafter, 1995). For a complex and highly constrained sound source such as speech, the simple strategy of attending to a particular frequency is of doubtful value. There is a variety of more complex, nonspatial perceptual properties of a speech signal that could be used to maintain attention across time such as its pitch contour and, more abstractly, individual characteristics of the talker (Broadbent, 1952) or of the transmission channel (Egan *et al.*, 1954). More recently, on the basis of EEG data, Woods and colleagues (1984) have argued that selective attention to speech (either silent or with shadowing) is directed not just to a particular spatial location but rather to the ensemble of a particular location and talker. This paper looks at the effectiveness of two properties that help to define a particular talker across time: prosodic continuity and vocal-tract size.

In an earlier paper (Darwin and Hukin, 1999), which was primarily concerned with the use of spatial cues such as interaural time difference (ITD) in auditory grouping and selective attention, we showed that listeners could use differences in ITD of a few tens of microseconds (μs) to decide which of two synchronized target words came from one of two simultaneous sentences. The sentences were spoken in a monotone by the same talker, and were resynthesized to be accurately on a constant fundamental frequency (F_0). Somewhat surprisingly, we found that listeners made very little use of F_0 continuity in doing this task. Differences in F_0 between the two sentences of up to four sentences had very little influence on listeners' preferences. When the two sentences had the same ITD, listeners showed little preference for the target word on the same F_0 as the attended sentence; when the target words were cross-spliced they continued to report the target word that shared spatial location with the attended sentence, tolerating an F_0 jump during the target word. Such ineffectiveness of F_0 continuity was surprising since previous work using rather different paradigms had shown that continuity of prosody or just of F_0 could be useful in assigning speech across time to the same or to different talkers (Darwin, 1975; Darwin and Bethell-Fox, 1977). Speech separation algorithms had similarly found continuity of F_0 to be useful (Parsons, 1976; Weintraub, 1987).

The first experiment in the present paper uses a similar task to that used in the earlier paper to investigate the effectiveness of natural (rather than monotonous) prosody, and uses resynthesized speech to compare the effectiveness of F_0 and level changes produced by varying sentence stress. The experiment titrates these changes against differences in lateral position cued by ITD.

[a]Electronic mail: c.j.darwin@biols.susx.ac.uk

I. EXPERIMENT 1

This experiment first examines how effective the prosodic cues present in naturally spoken sentences are at helping listeners to track a particular utterance in the presence of a second utterance from the same talker. The two sentences could also differ in ITD, and the experiment compares the effectiveness of these two types of cue.

The basic design of the experiment is similar to the first experiment reported in our previous paper: the listener hears the same two carrier sentences on each trial, is asked to attend to one of them, and to report which of two target words occurs in the attended sentence. The same two targets words are present on each trial, so the experiment is strongly weighted towards measuring how effective the different cues are at enabling listeners to track a particular sound source over time, rather then how well listeners can detect or recognize a word.

Unlike our previous experiments, in which the sentences had a monotonous F_0 contour, this experiment uses sentences that are spoken with natural intonation. The sentence stress is placed either towards the beginning of the sentence or towards the end, so that on each trial the two constituent sentences have globally different intonation contours. This change in intonation contour affects the way the pitch, amplitude, and durations of the individual words vary across the sentence. Listeners might be using any or all of these cues to help them to identify which word was spoken as part of the attended sentence. Our paradigm makes it unlikely that rhythmic differences are contributing to listeners' performance, since the two target words start at the same time and have similar durations.

In order to assess whether a difference in fundamental frequency (over and above any difference in intensity) is contributing to the ability of subjects to follow a particular sentence, we manipulate the overall level of the pitch contour. In one manipulation we change the overall pitch level for the whole of each sentence in order to bring the average F_0s in the two target words together, and in the other, we move them further apart. If listeners are using the F_0 contour to help them to track a particular sentence, we could expect them to do this better when the F_0 contours are moved apart, and worse when they are moved together. In order to maintain good speech quality while changing the pitch contour of the speech, we use the pitch-synchronous overlap–add (PSOLA) method of resynthesis. PSOLA is a waveform-based resynthesis method that allows independent manipulation of the pitch and duration of speech while maintaining good speech quality. The method was introduced by Moulines and Charpentier (1990) and its perceptual effect on simple speech-like sounds has recently been evaluated (Kortekaas and Kohlrausch, 1997, 1999).

In this experiment, we also vary the lateral position of the two sentences by giving the two sentences of a pair different ITDs—the dominant lateralization cue for complex sounds (Wightman and Kistler, 1992). In our previous experiment, using sentences spoken and resynthesized on a monotonous F_0 contour, listeners were able to use a small difference in ITD to track one sentence over time. Listeners continued to report the target word with the same ITD as the carrier sentence when the target words were cross-sliced so that there was a discontinuity (of up to four semitones) in the monotonous pitch contour. Experiment 1 thus pitted the effectiveness of prosodic cues against differences in ITD by cross-splicing target words between pairs of sentences, so that one target word had the appropriate prosody but the wrong ITD and *vice versa*.[1]

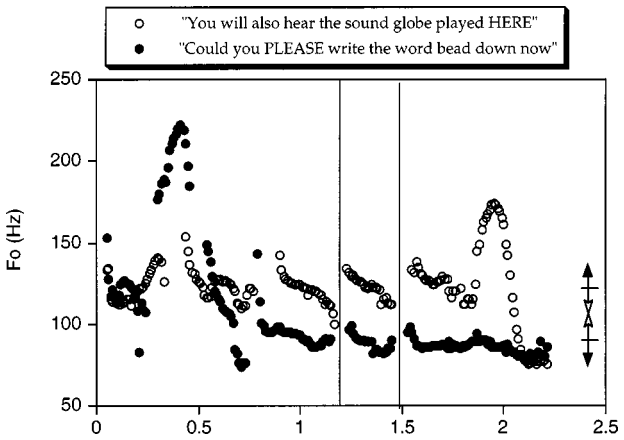

FIG. 1. F_0 values used for PSOLA resynthesis for one of the four original pairs of sentences used in experiment 1. The vertical lines at around 1.5 s delimit the target words ''bead'' and ''globe.'' The open-headed arrows show the change in F_0 to the whole sentence for the together condition, where the F_0 contours for the target words overlap; the filled arrows show the change in F_0 to the whole sentence for the apart condition.

A. Stimuli

Two sentences with each of two target words (''Could you please write the word bead/globe down now'' and ''You'll also hear the sound bead/globe played here'') were recorded by a male, native speaker of British English (C.J.D.). Each sentence was spoken in two versions, one with the main sentence stress early in the sentence (on ''please'' and ''also''), and once with the stress late in the sentence (on ''now'' and ''here''). These eight sentences were recorded on DAT tape and subsequently digitized at 22 050 Hz. Each sentence was between 2.20 and 2.41 s long. Sentences were paired so that a pair contained both carrier sentences, both target words, and both sentence stress positions (early, late). These constraints gave four different pairings. Within each pair, minor durational adjustments were made in order to (i) equate the durations of the two target words by adding/removing individual pitch periods from their centers, and (ii) align the onset of the two target words by adding a short period of silence to the beginning of one sentence in a pair. The target words started between 1.2 and 1.3 s from the start of their respective carrier sentences.

The F_0 contour of each sentence was obtained automatically using ESPS/WAVES+ software (Sensimetrics, Cambridge, MA) and checked for accuracy against the waveform. These original contours for two of the sentences are shown in Fig. 1. Three different resyntheses were then made for each sentence pair using a WAVES+ resynthesis tool (Möhler and Dogil, 1995) based on the PSOLA method (Moulines and Charpentier, 1990) which allowed the F_0 contours of the sentences to be increased or decreased in frequency. These three F_0 conditions were: *original*, in which the F_0 values

TABLE I. Allocation of prosodic and spatial cues across the different experimental conditions. (The vocal tract manipulation is only made in experiments 2 and 3.)

	F_0 Together	F_0 Original	F_0 Apart
Normal	+ITD $0F_0$ +Intensity	+ITD $+F_0$ +Intensity (+Vocal tract)	+ITD $++F_0$ +Intensity
Swapped	−ITD $0F_0$ +Intensity	−ITD $+F_0$ +Intensity (+Vocal tract)	−ITD $++F_0$ +Intensity

were unchanged; *together*, in which the two sentences' F_0 contours were both shifted in order to make the values of F_0 during the two target words similar (F_0 contours of the sentence whose target word had a higher F_0 were shifted down 12% and the F_0 contours of the other sentence were shifted up 15%—the symmetry was introduced to maintain quality of resynthesis); *apart*, in which the two sentences' F_0 contours were shifted the opposite way in order to make their values of F_0 during the two target words more different (the sentence that had the higher-F_0 target was raised by 15% Hz, the one with the lower- F_0 target was lowered by 12%).

Two different splicing conditions were generated from these resynthesized sentence pairs. A *normal* condition which retained the sentences as described in the previous paragraph, and a *swapped* condition in which the target words were swapped between the two sentences of a pair. Notice that this cross-splicing was done *after* the F_0 contours of the sentence pair had been altered but *before* sentences were given different ITDs (which was done by the computer program that presented the sounds to listeners). Thus, in the normal condition prosodic cues act in the same direction as the spatial cue to reinforce the choice of a particular target word, whereas in the swapped condition, the prosodic cues oppose the spatial cue. Table I summarizes the allocation of the two prosodic cues (intensity, F_0) and one spatial cue (ITD) in the different conditions of experiment 1.

B. Procedure

The 13 listeners were native speakers of British English aged between 21 and 52 who had pure-tone thresholds within the normal range at octave frequencies between 125 Hz and 8 kHz. They had all taken part in the experiment reported in the earlier paper, which had used a similar paradigm and broadly similar stimuli.

Listeners were tested individually. They were told that they would always hear the same two carrier sentences which might come from the same or different positions. Six listeners were asked to attend to one of the sentences and seven to the other sentence. They were asked to press the ''b'' or ''g'' key according to whether the attended sentence contained the target word ''bead'' or ''globe,'' respectively. On each trial the listener heard both carrier sentences and both target words.

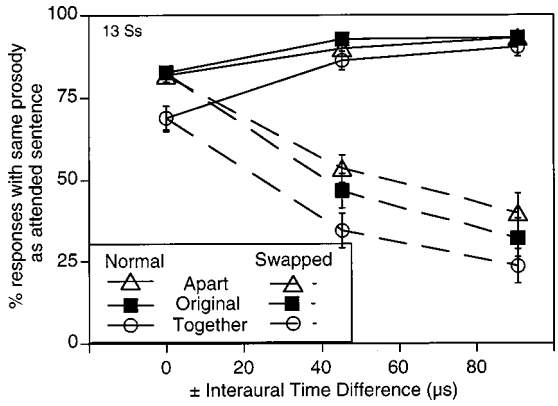

FIG. 2. Percentage of trials on which listeners reported the target word that had been originally spoken in the attended sentence (chance performance is 50%). Two sentences were presented at the same time; their interaural time differences could be different and the F_0 contours of both original sentences could be manipulated so as to move the F_0's of the two simultaneous target words either together or further apart from their original values. In the normal condition, one of the target words has both the same prosody and the same ITD difference as the attended sentence. In the swapped condition, the target that has the same prosody as the attended sentence has the opposite ITD.

Pairs of files, prepared as described above, were digitally mixed at presentation with ITDs of 0, ±45.3, ±90.7 μs (both sentences were therefore played to both ears). These ITDs correspond to 0, ±1, ±2 samples at 22 050 Hz (the terminology ''±1 sample'' indicates that one of the sentences led in one ear by one sample, and the other sentence led in the other ear by one 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. Which sentence of a pair had the positive and which the negative ITD was randomly varied from trial to trial (with the 0 ITD conditions doubled).

Each listener thus heard each of 144 trial types (4 sentence pairs×2 cross-splicing conditions×3 F_0-contour types×6 ITD conditions) 5 times to give a total of 720 trials. Stimuli were presented through a Digidesign Protools interface attached to a Power Mac 7100 which also controlled the experiment. The output of the Protools interface was connected to Tucker-Davis PA4 attenuators which were used to set the overall level for the experiment. Subjects listened over Sennheiser HD414 headphones in a double-walled IAC booth. The sentences, when mixed at each headphone, gave an average level of 68 dB SPL through a flat-plate coupler.

C. Results and discussion

Figure 2 shows the percentage of trials on which listeners reported the target word that had been originally spoken in the attended sentence (and so had the appropriate prosody for the attended sentence) as a function of ITD. Overall, the obvious difference between the normal conditions (solid lines) in which the prosodic cues and ITD work together, and the swapped conditions (dashed lines) where ITD opposes the prosodic cues, shows the effect of ITD in opposing the prosodic cues. However, we first examine the power of the

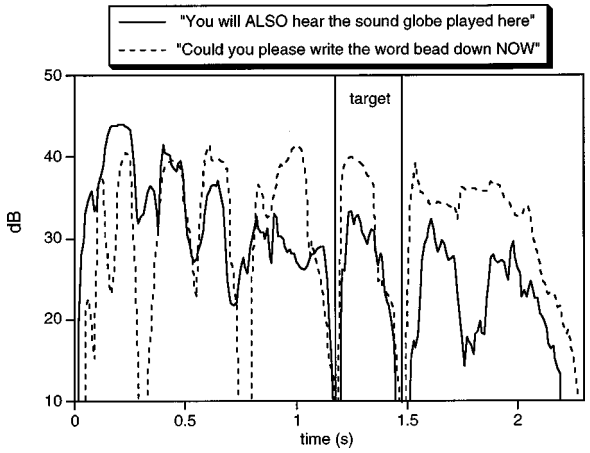

FIG. 3. Amplitude contours of the two sentences of a pair from experiment 1. In this example the target from the sentence with early sentence stress is about 10 dB less intense than the target from the sentence with late sentence stress.

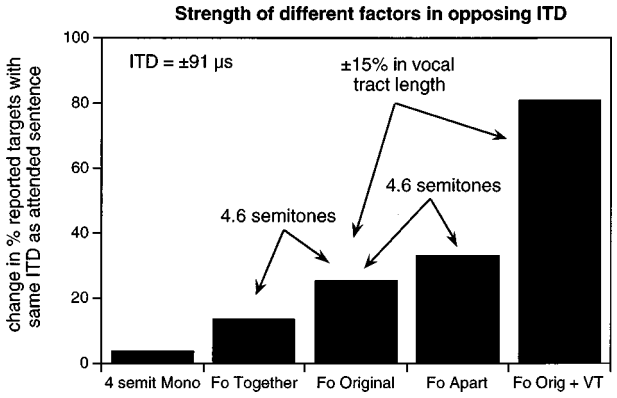

FIG. 4. The strength of different prosodic factors in opposing ITD is illustrated by plotting how the percent of reported targets that have the same ITD as the attended sentence changes when ITD is opposed (swapped condition), rather than reinforced (normal condition), by the prosodic cues listed on the abscissa. A large percent change indicates an effective prosodic cue. All the date in this figure are from the same 13 subjects. The left-hand bar comes from experiment 1 of Darwin and Hukin (1999). The next three bars are from experiment 1 of the present paper; the right-hand bar is from experiment 2.

prosodic cues alone by looking at the results from the conditions where the ITD was zero (and so the distinction between normal and swapped does not exist).

1. ITD=0

When both sentences are presented with an ITD of zero, listeners prefer the target with the same prosody as the attended sentence substantially above chance (83%) for the original condition (with the prosody unchanged). This level of performance is made possible by the prosodic differences between the two sentences. Specifically, the F_0 contour of the attended sentence makes some contribution to listeners' ability to track the attended sentence; changing the pitch contours of both sentences, so that the two target words have similar F_0's (together condition), reduces performance by 14% ($t_{12}=3.9, p<0.001$). This change (produced by a 27% change in the frequency of F_0) is comparable to the 11% improvement in performance produced by a four-semitone (26% change in frequency) monotonous difference in F_0 in our previous experiment (Experiment 1, Darwin and Hukin, 1999). It shows that, in the absence of other cues, listeners can use the continuity of a natural F_0 contour to track a sentence across time, but, like the continuity of a monotonous F_0 contour, it is not a particularly strong effect. Strengthening the prosodic cues by moving the F_0 contours apart does not further increase listeners' preference.

In the together condition, where there is a little difference between the F_0 contours of the target words, other prosodic cues maintain listeners' preferences well above chance at 69% ($t_{12}=4.8, p<0.002$). Listeners are probably using intensity differences between the target words to achieve this level of performance. Figure 3 shows the intensity contours of the two sentences of one pair. In the region of the target word, the sentence with the late stress is around 10 dB more intense than the sentence with the early stress.

2. Normal ITD≠0

When the original sentences are also separated by a difference in ITD, preference for the target word sharing the same prosody as the attended sentence increases from 83% at 0 μs to 93% at ± 91 μs. This change is significant ($t_{12}=5.1, p<0.002$). Changes to the F_0 contour do not significantly alter this already very high level of performance.

3. Swapped ITD≠0

Opposing prosodic cues with a difference in ITD clearly reduces the number of trials on which listeners report the target word that has the same prosodic cues as the attended sentence $F_{1,12}=72.9, p<0.0001$. This reduction is greater for the larger difference in ITD, $F_{1,12}=49.6, p<0.0001$. An ITD of ± 45 μs is sufficient to counter the prosodic cues present in this experiment (performance is 50% for the swapped original condition) and with an ITD of ± 91 μs listeners consistently report the target that has the same ITD as the attended sentence rather than the target that has the correct prosody. These levels of performance are lower than in the normal original, zero ITD condition by 36% for ± 45 μs and 51% for ± 91 μs.

Strengthening the prosodic cues by moving the F_0 contours apart increases the chance of listeners reporting the target word with the same prosody as the attended sentence; conversely, weakening the prosodic cues by moving the F_0 contours together decreases it: the total increase in reports of the target with the same prosody between the together and apart conditions is about 18%. The main effect of the F_0 manipulation (together, original, apart) is significant in the swapped conditions when ITD is not zero ($F_{1,12}=44.7, p<0.0001$).

We can now compare the effectiveness in opposing ITD of the natural intonation contours from this experiment, with the monotone manipulations from our earlier experiment. Figure 4 shows how effective different prosodic factors are in opposing an ITD of ± 91 μs. The score shows how much listeners' preference for the target that had the same ITD as the attended sentence was reduced when the ITD cue was opposed rather than reinforced by the prosodic cue(s) shown

on the abscissa. A small difference shows that listeners tend to follow ITD; a large difference shows that they tend to follow the other cues. The left-hand bar illustrates data from our previous experiment, and shows the difference between a condition where a four-semitone difference in monotone F_0 is acting with or against an ITD of ± 91 μs. There is very little (3%) change, reflecting the weakness of the monotonous F_0 cue. The next three bars illustrate data from the together, original, and apart conditions of the present experiment. Although these prosodic manipulations are insufficient to override the ITD difference, they exercise a greater influence than the monotonous F_0. The together condition is significantly more influenced by prosody than the four-semitone monotone condition, $t_{11}=2.47$, $p<0.02$, and the others progressively more different. A 4.6 semitone difference of F_0 between the together and normal conditions produces a 12% change, and a further 4.6 semitone difference between the normal and apart conditions gives a further 8% change.

In summary, first this experiment has extended, to sentences with natural intonation, our previous result that a difference in ITD between two sentences substantially helps listeners decide which of two simultaneous target words belong to an attended sentence. A difference of ITD of ± 91 μs (equivalent to about $\pm 12°$ separation in azimuth) gives a change of about 50% when opposed to the natural prosodic cues.

Second, the experiment has shown that normal prosodic changes also help listeners: the prosodic cues naturally present here give a performance that is 33% above chance. Specifically removing the F_0 contribution to the prosodic cues (leaving substantial amplitude differences), reduces this figure by 14%, to 19% above chance.

Third, the experiment (in conjunction with experiment 1 of our previous paper) has shown that, for comparable semitone differences, a natural F_0 contour is slightly more effective at maintaining a listener's attention in the face of an opposing ITD difference than a monotonous contour.

II. EXPERIMENT 2

Both the first experiment in our previous paper and the present experiment 1 used two simultaneous sentences spoken by the same talker. It is substantially more natural to have two different talkers speaking at the same time. Globally shifting formant frequencies is an effective way of altering the individuality of a voice (without changing voice pitch or fundamental frequency) and corresponds to an alteration in vocal-tract length. Individuals with longer vocal tracts have lower formant frequencies; those with shorter vocal tracts have higher formant frequencies. Men and women differ on average by about 17% in vocal-tract length (Peterson and Barney, 1952), but an upward or downward shift of 8% in formant frequencies is sufficient to reduce the recognition of individual voices to chance (Kuwabara and Takagi, 1991). Although it is technically possible to resynthesize speech so that only formant frequencies (the resonant frequencies of the vocal tract) are changed, it was simpler (and allowed better speech quality) for us to change the whole spectral envelope, so that not only formant frequencies but also voice source characteristics were changed. When this is done, it is likely that formant frequency changes, rather than changes to the individual characteristics of the voice source, are responsible for most of the change in individuality (Zhu and Kasuya, 1998). It is important to stress, however, that the manipulation that we used leaves the fundamental frequency of the voice unchanged: the same harmonic frequencies are present as in the original voice, but the spectral envelope that defines the amplitudes of those harmonics is transposed up or down in frequency.

In this experiment, we produced two apparently different talkers from the original sentences used in experiment 1 by globally changing the spectral envelope (including formant frequencies) by $\pm 15\%$. For convenience, we will refer to the change as producing a voice from a longer or a shorter vocal tract.

A. Stimuli

The original sentences from experiment 1 were modified to produce two different talkers, with the same F_0 and durations as the originals but with a different spectral envelope and consequently a different apparent vocal-tract size. To produce the longer vocal-tract voice, the PSOLA algorithm that had been used in experiment 1 was used together with the program DSP DESIGNER (Zola Technologies) to: (i) raise F_0 and globally reduce the duration by 15%, (ii) resample the file at a 15% higher sampling frequency and then set the playback rate back to the original value (22 050 Hz). The end results of these manipulations was to produce sentences that had the same durations and F_0s as the originals, but which had the spectral envelope (including formant resonances and voice source properties) lowered by 15%. To produce the shorter vocal-tract voice the opposite manipulations were made. The resulting voices, although still plausibly natural, were very clearly different individuals, neither of whom sounded like the original talker.

The sentences were paired as in experiment 1 with the additional constraint that each pair contained one long vocal-tract sentence and one short vocal-tract sentence. As in experiment 2, target words could be swapped between the sentences of a pair before the sentences were allocated an ITD. In the swapped condition, the target word that had the same ITD as the attended sentence had a prosody and vocal-tract size that was appropriate for the unattended sentence.

B. Procedure

The procedure was similar to that from experiment 1 except that ITDs of 0, ± 45.3, ± 90.7, and ± 181.4 μs were used. The longest ITD value was introduced after the first four subjects had been run. Again, the ITD manipulation was made at the time of presentation. The same 13 listeners from experiment 1 took part.

C. Results and discussion

Figure 5 shows the percentage of trials on which listeners reported the target word that had originally belonged to the attended sentence (and so shared its prosody and vocal-tract size). The data from this experiment are plotted with

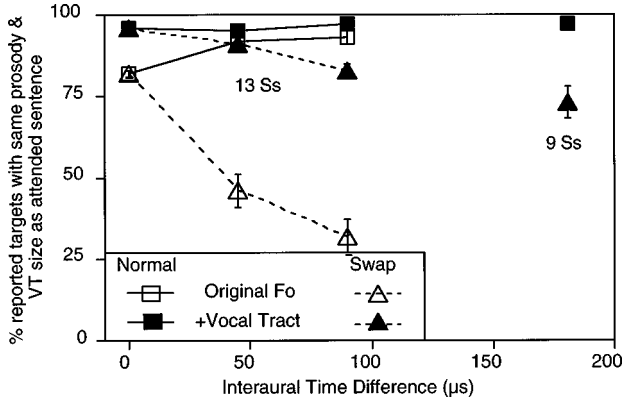

FIG. 5. Percentage of trials on which listeners reported the target word that had been originally spoken in the attended sentence (chance performance is 50%). The solid squares show the results from experiment 3 in which two sentences were presented at the same time, one with each of two different vocal-tract lengths. In the swapped conditions, the original target word has the opposite ITD, and the other target word has the same ITD as the attended sentence. Open triangles replot comparable data from experiment 1 without the difference in vocal-tract length. Only nine listeners contributed data at the longest ITD.

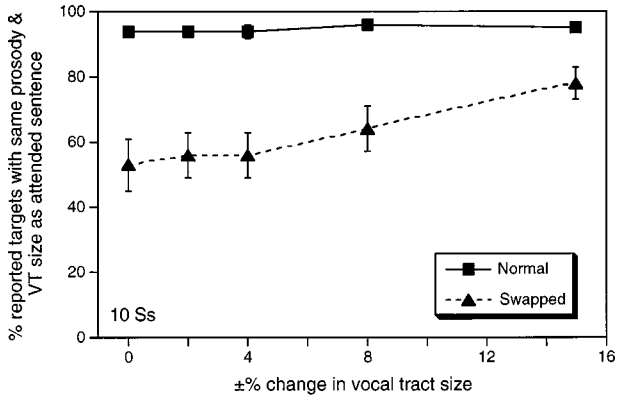

FIG. 6. Percentage of trials in experiment 3 on which listeners reported the target word that had been originally spoken in the attended sentence. In the normal conditions, one of the target words has the same prosody, vocal-tract size, and ±91 μs ITD difference as the attended sentence. In the swapped condition the target with the same prosody and vocal-tract size as the attended sentence has the opposite ITD.

solid symbols; data from the same listeners from experiment 1 (without the vocal-tract change) are plotted as open symbols.

Introducing a difference of ±15% in vocal-tract size clearly has a large effect; even without a difference in ITD, listeners' preferences are almost 100% when the two voices differ in vocal-tract size. In addition, when an increasing ITD acts against prosodic and vocal-tract size cues, listeners have a strong tendency to stay with the vocal tract rather than the spatial position. Although there is some significant progressive reduction as ITD increases in the swapped condition in the number of reported targets derived from the attended sentence, the reduction is small ($F_{2,16}=7.9$, $p<0.02$ for nine listeners across three ITDs, $F_{1,12}=10.3, p<0.01$ for 13 listeners on the two smaller ITDs). Similarly, introducing an opposing ITD of ±91 μs only reduces the number of these targets reported from 98% to 86%. This contrasts with a reduction in experiment 1, which lacked the vocal-tract cue, from 93% to 32%. With a larger opposing ITD of ±181 μs (corresponding to a spatial separation of about ±50°D) listeners still prefer the original target word (that has the same vocal-tract size and the same prosody as the attended sentence) on 73% of trials.

This experiment has shown that a substantial difference in vocal-tract size between two talkers produces a powerful cue for selective attention. As illustrated in Fig. 4, the effects that we have found with our vocal-tract manipulation have been substantially larger than those found for our prosodic manipulations. However, our vocal-tract manipulation is larger, compared with naturally occurring values, than is our prosodic manipulation as shown in the data of Peterson and Barney (1952). Their data shows an average[2] female/male formant ration of 1.17, but an average female/male ratio for F_0 of 1.7 (nine semitones). Consequently, our experiment 2 uses an F_0 range (9.2 semitones between together and apart) that is comparable to the average male/female difference, but a vocal tract difference (30% between the short and the long manipulations) that is at the top of the range of scaling values. Experiment 3 examines the effect of smaller vocal-tract manipulations.

III. EXPERIMENT 3

This experiment asked how changes to the difference in vocal-tract size between two talkers influence the listener's selection of which word belongs in the attended sentence. It is similar in design to the first two experiments but uses a range of vocal-tract size differences with a single prosodic condition which are opposed by a single ITD.

A. Method and procedure

The original prosody condition from experiment 2 at an ITD of ±91 μs was used together with four more conditions with changes of vocal-tract size of ±2, ±4, ±8, and ±15% produced by the same PSOLA and DSP DESIGNER algorithms as in experiment 2. As in the previous experiments, for the normal conditions a particular target word had the same ITD, prosody, and vocal-tract size as one of the sentences, whereas in the swapped condition, if a target word had the same ITD as one sentence, it had the same prosody and vocal-tract size as the other sentence.

Each subject heard each of 80 trial types (4 sentence pairs×5 vocal-tract sizes×2 ITD conditions×normal/swapped) 6 times to give a total of 240 trials. The ten listeners had all taken part in experiment 2.

B. Results and discussion

The overall results for the experiment are shown in Fig. 6 as the percent of target words that shared the same prosody and, where appropriate, vocal-tract length with the attended sentence.

When prosodic, vocal tract, and ITD differences all operate in the same direction, listeners perform the task almost perfectly with scores of over 90% in all conditions. When an ITD of ±91 μs opposes the prosodic and vocal-tract cues, listeners report fewer of the targets that have the same

prosody and vocal-tract size. With no difference in vocal-tract size, the prosodic cues in the experiment are sufficient to give about 54% of the reported target words, which is somewhat larger than the corresponding condition in the previous experiment. With the largest vocal-tract difference augmenting the prosodic cues, this figure increases to about 78%. However, most of the change occurs for the two larger vocal-tract sizes. There is no significant change in performance across the 0%, ±2%, and ±4% vocal-tract length differences. Differences from the 0% condition only begin to appear with the ±8% change in vocal tract with a 9% increase in reported targets that share the same vocal tract as the attended sentence $t_9 = 3.2$, $p < 0.01$. This size difference (16.6% between the short and the long) is comparable to that between male and female talkers, and so is ecologically comparable to an F_0 differences of around nine semitones which gave about a 20% increase in reported targets in experiment 2. This increase is somewhat larger than that produced by the ±8% vocal-tract change, indicating that the apparently dominant effects of a change in vocal tract in the previous experiment are due to the unnaturally large vocal-tract size changes used.

IV. GENERAL DISCUSSION

The experiments described in this paper have presented some new data on the relative roles of spatial cues, prosodic cues, and cues to an individual voice in auditory attention. The experiments have shown that natural prosodic variations are more effective than a monotonous pitch at overriding spatial cues in determining which of two possible target words belongs with a particular attended sentence. Part of their effectiveness is due to amplitude differences between target words in different prosodic contexts, and part is due purely to F_0 differences. The experiments have also shown that a difference in vocal-tract size that is comparable to or larger than the average male/female difference can also override spatial cues.

The task that we have used, though having the virtue of simplicity and requiring a minimum of preparatory signal processing, does not separate out the immediate allocation of attention from a more leisurely assessment of the whole stimulus. Listeners were under no time pressure to make their response and the target word occurred relatively late in the sentence. Since there were only two response alternatives, they could also have brought some variety to a repetitive task by listening to the ''nonattended'' sentence, though because of the symmetry of the stimulus pairings, such a strategy would not have influenced the overall results. However, our own impression in doing the task is that when there is a substantial change in vocal-tract size in the attended sentence, the word with the appropriate vocal-tract size in the unattended sentence ''pops out'' at you. We plan to pursue the questions raised in these experiments by trying to obtain some measure of the allocation of spatial attention in the temporal vicinity of this intrusion and also by testing whether it still occurs under a task such as close shadowing.

Although speech separation algorithms have exploited spatial differences and F_0 differences between competing messages, we know of no attempt to exploit vocal-tract length differences. Although vocal-tract length differences would be of limited value in initially separating simultaneous sounds, they could, along with spatial and prosodic cues, help in solving the problem of source continuity.

ACKNOWLEDGMENTS

This research was supported by Grant No. G9801285 from the UK MRC to the first author.

[1]This experiment was an extension and replication of a pilot experiment carried out by Catherine Brown as a student project.

[2]With a range across vowels from 1.02 to 1.27 reflecting the fact that males have a relatively larger pharynx (Fant, 1964).

Broadbent, D. E. (**1952**). ''Failures of attention in selective listening,'' J. Exp. Psychol. **44**, 428–433.

Broadbent, D. E. (**1953**). ''The role of auditory localization in attention and memory span,'' J. Exp. Psychol. **47**, 191–196.

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

Darwin, C. J. (**1975**). ''On the dynamic use of prosody in speech perception,'' in *Structure and Process in Speech Perception*, edited by A. Cohen and S. G. Nooteboom (Springer, Berlin), pp. 178–194.

Darwin, C. J., and Bethell-Fox, C. E. (**1977**). ''Pitch continuity and speech source attribution,'' J. Exp. Psychol.: Hum. Percept. Perform. **3**, 665–672.

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

Egan, J. P., Carterette, E. C., and Thwing, E. J. (**1954**). ''Some factors affecting multichannel listening,'' J. Acoust. Soc. Am. **26**, 774–782.

Fant, G. (**1964**). ''A note on vocal tract size factors and nonuniform *F*-pattern scalings,'' Speech Transmission Laboratory, Stockholm STL-QPSR **4**, pp. 22–30.

Hubner, R., and Hafter, E. R. (**1995**). ''Cueing mechanisms in auditory signal-detection,'' Percept. Psychophys. **57**, 197–202.

Kortekaas, R. W., and Kohlrausch, A. (**1999**). ''Psychoacoustical evaluation of PSOLA. II. Double-formant stimuli and the role of vocal perturbation,'' J. Acoust. Soc. Am. **105**, 522–535.

Kortekaas, R. W. L., and Kohlrausch, A. (**1997**). ''Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-formant stimuli,'' J. Acoust. Soc. Am. **101**, 2202–2213.

Kuwabara, H., and Takagi, T. (**1991**). ''Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method,'' Speech Commun. **10**, 491–495.

Möhler, G., and Dogil, G. (**1995**). ''Test environment for the two level model of Germanic prominence,'' Eurospeech 1995; Madrid, pp. 1019–1022.

Mondor, T. A., and Zatorre, R. J. (**1995**). ''Shifting and focusing auditory spatial attention,'' J. Exp. Psychol: Hum. Percept. Perform. **21**, 387–409.

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

Parsons, T. W. (**1976**). ''Separation of speech from interfering speech by means of harmonic selection,'' J. Acoust. Soc. Am. **60**, 656–60.

Peterson, G. H., and Barney, H. L. (**1952**). ''Control methods used in a study of the vowels,'' J. Acoust. Soc. Am. **24**, 175–184.

Quinlan, P. T., and Bailey, P. J. (**1995**). ''An examination of attentional control in the auditory modality—further evidence for auditory orienting,'' Percept. Psychophys. **57**, 614–628.

Scharf, B., Quigley, S., Aoki, C., Peachey, N., and Reeves, A. (**1987**). ''Focused auditory attention and frequency selectivity,'' Percept. Psychophys. **42**, 215–223.

Schlauch, R. S., and Hafter, E. R. (**1991**). ''Listening bandwidths and frequency uncertainty in pure-tone signal-detection,'' J. Acoust. Soc. Am. **90**, 1332–1339.

Spence, C. J., and Driver, J. (**1994**). ''Covert spatial ordering in audition: Exogenous and endogenous mechanisms,'' J. Exp. Psychol.: Hum. Percept. Perform. **20**, 555–574.

Spieth, W., Curtis, J. F., and Webster, J. C. (**1954**). ''Responding to one of

two simultaneous messages,'' J. Acoust. Soc. Am. **26**, 391–396.
Teder, W., and Näätänen, R. (**1994**). ''Event-related potentials demonstrate a narrow focus of auditory spatial attention,'' NeuroReport **5**, 709–711.
Weintraub, M. (**1987**). ''Sound separation and auditory perceptual organization,'' in *The Psychophysics of Speech Perception*, edited by M. E. H. Schouten (Martinus Nijhoff, Dordrecht), pp. 125–134.
Wightman, F. L., and Kistler, D. J. (**1992**). ''The dominant role of low-frequency interaural time differences in sound localization,'' J. Acoust. Soc. Am. **91**, 1648–1661.
Woods, D. L., Hillyard, S. A., and Hansen, J. C. (**1984**). ''Event-related brain potentials reveal similar attentional mechanisms during selective listening and shadowing,'' J. Exp. Psychol.: Hum. Percept. Perform. **10**, 761–777.
Zhu, W. Z., and Kasuya, H. (**1998**). ''Perceptual contributions of static and dynamic features of vocal tract characteristics to talker individuality,'' IEICE Trans. Fundam. Electron. Commun. Comput. Sci. **E81A**, 268–274.