Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers

Christopher J. Darwin[a)]
Experimental Psychology, University of Sussex, Brighton BN1 9QG, United Kingdom

Douglas S. Brungart
Air Force Research Laboratory, Human Effectiveness Directorate, Wright-Patterson AFB, Ohio 45433

Brian D. Simpson[b)]
Veridian, 5200 Springfield Pike, Suite 2000, Dayton, Ohio 45431

(Received 15 June 2002; revised 9 August 2003; accepted 11 August 2003)

Three experiments used the Coordinated Response Measure task to examine the roles that differences in *F*0 and differences in vocal-tract length have on the ability to attend to one of two simultaneous speech signals. The first experiment asked how increases in the natural *F*0 difference between two sentences (originally spoken by the same talker) affected listeners' ability to attend to one of the sentences. The second experiment used differences in vocal-tract length, and the third used both *F*0 and vocal-tract length differences. Differences in *F*0 greater than 2 semitones produced systematic improvements in performance. Differences in vocal-tract length produced systematic improvements in performance when the ratio of lengths was 1.08 or greater, particularly when the shorter vocal tract belonged to the target talker. Neither of these manipulations produced improvements in performance as great as those produced by a different-sex talker. Systematic changes in both *F*0 and vocal-tract length that simulated an incremental shift in gender produced substantially larger improvements in performance than did differences in *F*0 or vocal-tract length alone. In general, shifting one of two utterances spoken by a female voice towards a male voice produces a greater improvement in performance than shifting male towards female. The increase in performance varied with the intonation patterns of individual talkers, being smallest for those talkers who showed most variability in their intonation patterns between different utterances. [DOI: 10.1121/1.1616924]

PACS numbers: 43.71.Bp, 43.71.Es [PFA] Pages: 2913–2922

I. INTRODUCTION

Over the past 50 years, numerous researchers have studied the ability of human listeners to extract information from a target speech signal that is masked by one or more competing talkers [see Ericson and McKinley (1997) and Bronkhorst (2000) for recent reviews of this literature]. This ''cocktail party'' listening task is particularly difficult when the target and masking speech signals are mixed together into a single channel and then presented monaurally or diotically over headphones. In this condition, binaural speech segregation cues, which are typically available in real-world listening situations, are absent and listeners must rely on monaural cues to perform the task. Although there is some evidence that listeners can use differences in the overall levels of the two voices to perform the segregation task (Egan *et al.*, 1954; Brungart, 2001), the most powerful monaural speech segregation cues seem to be related to differences in the vocal characteristics of the competing talkers (Bregman, 1990; Darwin and Hukin, 2000; Brungart, 2001). Brungart, for example, found that phrases spoken by different-sex talkers were substantially easier to segregate than phrases spoken by same-sex talkers, and that phrases spoken by two different same-sex talkers were substantially easier to segregate than two phrases spoken by the same talker.

To this point, however, little is known about the relative contributions that different voice characteristics make to the voice segregation process. Differences in the overall long-term spectra of the competing speech signals do not seem to have much influence on performance: Festen and Plomp (1990) found little difference between the intelligibility of a speech signal masked by speech-shaped noise with the overall spectrum of a same-sex talker and the intelligibility of a speech signal masked by speech-shaped noise with the overall spectrum of a different-sex talker.

Two relatively simple physical characteristics can be manipulated to change the apparent gender of a voice: the fundamental frequency (*F*0) range, and the vocal-tract length. Women's voices are typically a little under an octave higher in *F*0 than men's, and women's formant frequencies are around 16% higher than men's as a result of male vocal tracts being longer (Peterson and Barney, 1952). Voice individuality is lost if formants are shifted by 8% (Kuwabara and Takagi, 1991), but the successful digital transformation of voice gender normally requires manipulation of both *F*0 and vocal-tract length (Atal and Hanauer, 1971).

Of these two parameters, only differences in the *F*0 of the talkers and differences in the prosodic features of the

[a)]Electronic mail: cjd@biols.susx.ac.uk
[b]Currently at Air Force Research Laboratory, Wright-Patterson Air Force Base, Ohio.

competing utterances have been shown to produce some improvement in voice segregation (Brokx and Nooteboom, 1982; Scheffers, 1983; Assmann and Summerfield, 1990; Bird and Darwin, 1998; Darwin and Hukin, 2000). We know of no studies that have systematically examined both *F*0 and vocal-tract length while preserving the small temporal and prosodic variations that normally occur in repeated utterances spoken by the same talker. In this series of experiments, we examine diotic speech segregation when each of two phrases is spoken by the same talker, similar to the ''same-talker'' (TT) condition examined by Brungart (2001). However, here the speech signals were electronically modified to introduce differences in *F*0 (experiment 1), vocal-tract length (experiment 2), or both *F*0 and vocal-tract length (experiment 3) between the two competing phrases. The results provide valuable insights into the roles that *F*0, vocal-tract length, and target and masker gender play in two-talker speech segregation.

II. EXPERIMENT 1: CHANGES IN *F*0 ONLY

A. Method

The stimuli were derived from the publicly available Coordinate Response Measure speech corpus (Bolia *et al.* 2001). This corpus, which has been used in previous multitalker listening experiments (Brungart, 2001; Brungart *et al.*, in press), consists of sentences of the form ''Ready ⟨call sign⟩ go to ⟨color⟩ ⟨number⟩ now'' spoken with all 32 possible combinations of four colors (''red,'' ''blue,'' ''white,'' and ''green'') and eight numbers (1–8). The present experiments used all eight talkers available in the corpus (four male, four female), but only used four of the eight available call signs (''Arrow,'' ''Tiger,'' ''Eagle,'' and ''Baron''). Thus, a total of 1024 different sentences (8 talkers×4 call signs×4 colors×8 numbers) were used to produce the stimuli employed in the experiments

These 1024 sentences were down-sampled from 40 to 20 kHz, and then processed with the PSOLA algorithm (Moulines and Charpentier, 1990), as implemented in Macintosh version 3.9.28 of the Praat software package (Boersma and Weenink, 1996), to produce six new sets of speech files with *F*0 contours shifted by various numbers of semitones. The sentences spoken by female talkers were shifted by −9, −3, −1, 0, 1, or 3 semitones. The sentences spoken by male talkers were shifted by −3, −1, 0, 1, 3, or 9 semitones. The resulting speech quality was excellent for the range of *F*0 shifts that we used in the experiment.

Each stimulus consisted of a diotic mixture of two sentences spoken by the same talker: one sentence (the ''target'') contained the call sign ''Baron'' and a randomly selected color and number; the other sentence (the ''masker'') was randomly selected from all of the sentences in the corpus with a different call sign, color, and number than the target sentence. The *F*0 shifts of the two sentences were selected from the combinations shown in Table I to produce one of eight different absolute *F*0 differences: 0, 1, 2, 3, 4, 6, 9, or 12 semitones. The masking speech was also scaled to set its rms power to one of six different signal-to-noise ratios (SNRs) relative to the rms power of the target speech: −6, −3, 0, +3, +6, or +9 dB. Finally, as in previous CRM experiments, the overall output was randomly roved over a 6-dB range (in 1-dB increments) before a D/A converter (Tucker-Davis Technologies DD-1) was used to present the stimulus to the listener through headphones (Sennheiser HD-540) at a comfortable listening level (60–70 dB SPL).

TABLE I. *F*0 differences (in semitones) tested in the stimuli of experiment 1. Each pair of sentences in a trial was chosen from the same talker. The *F*0 shifts of each talker were selected differently for male and female talkers to place most of the *F*0 shift into the range between male and female speech. Thus, male talkers were generally shifted up more than down, and female talkers were generally shifted down more than up. Note that, on a given trial, the target sentence was equally likely to receive the higher or lower *F*0 shift.

	Male		Female	
Δ*F*0 (Semitones)	Lower *F*0	Higher *F*0	Lower *F*0	Higher *F*0
0	0	0	0	0
1	0	+1	0	−1
2	−1	+1	+1	−1
3	0	+3	0	−3
4	−1	+3	+1	−3
6	−3	+3	+3	−3
9	0	+9	0	−9
12	−3	+9	+3	−9

The experiment was conducted with the listeners seated at the CRT of a control computer in a sound-treated listening room. The listeners' task in each trial was to identify the color and number spoken in the target phrase, which was identified by the presence of the call sign ''Baron.'' Responses were made by using the computer mouse to select the appropriate color–number combination from an array of 32 colored numbers shown on the CRT. Nine paid volunteer subjects (four male, five female) with normal hearing participated in the study. All were native English speakers from the midwestern United States. Each of these listeners completed ten trials for each of the 384 possible stimulus configurations of the experiment (8 talkers×8 *F*0 differences×6 SNRs), for a total of 3840 trials per listener. These trials were completed in blocks of 192 trials, with separate blocks for the male and female talkers in the corpus, and with the trials within each block randomly balanced to have an equal number of trials for each value of each of the three independent variables (talker, *F*0 difference, and SNR). Each block took approximately 10 min to complete and each listener participated in two to three blocks per day over a 2-week period.

B. Results

1. Overall effects of shift in F0

The overall results averaged across all eight talkers are shown in Fig. 1. The left panel of the figure shows the probability of a correct identification of both the color and number in the target sentence as a function of the signal-to-noise ratio and the absolute *F*0 separation between the two sentences in semitones.

The 0-semitone data represent a stimulus condition that is very similar to the ''same talker'' (''TT'') condition presented in Brungart (2001). The only difference is that the

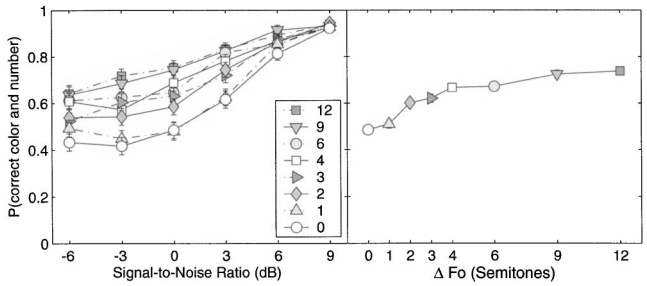

FIG. 1. The left panel shows the probability of a correct identification of the number and color in the target sentence as a function of SNR for eight different separations in semitones of the *F*0 contours between the target and the masker sentence. The right panel shows performance averaged across the lowest four SNRs (−6, −3, 0, and 3 dB) as a function of the separation in *F*0. The data have been averaged over all eight talkers. The error bars represent the 95% confidence intervals of the mean.

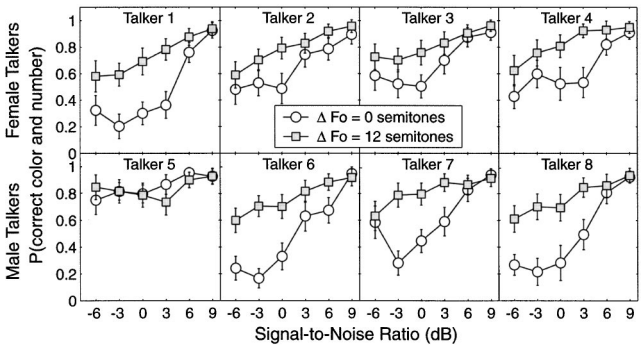

FIG. 2. Data as in the left panel of Fig. 1 but from individual talkers with either zero semitones (open circles) or 12 semitones (shaded squares) of separation in *F*0 between target and masker sentences. The upper row gives the results from the four female talkers (1–4) and the lower row from the four male talkers (5–8).

speech in this experiment has been analyzed and resynthesized by the PSOLA algorithm (without introducing any shift in *F*0). This analysis and synthesis did not appear to have any substantial effects on performance in the CRM task: the present results are nearly identical to those from the previous experiment, with performance declining as SNR decreases from over 90% at +9 dB to around 40% at 0 dB and then leveling out at SNRs less than 0 dB. The results of a *posthoc* pairwise comparison test (Fisher LSD) performed on the arcsine-transformed percentages of correct responses from the individual subjects confirmed that this plateauing was a significant effect: the 0-, −3-, and −6-dB conditions did not differ significantly at the $p<0.05$ level, while the 0, +3, +6, and +9 conditions were all significantly different from one another. The somewhat paradoxical leveling out of performance for SNRs less than 0 dB was found in early experiments on speech intelligibility in the presence of a single competing talker (Egan *et al.*, 1954; Dirks and Bower, 1969) and, as discussed in Brungart (2001), is likely due to a difference in level providing a cue to sound source identity, which compensates for increased energetic masking.

The new result in this figure is the gradual improvement in performance that occurred when target and masker sentences were artificially separated in *F*0 ($F_{1,8}=45.4$, $p<0.0001$, averaged over the lowest four SNRs). This effect is highlighted in the right panel of Fig. 1, which shows the probability of a correct color and number response averaged across the four lowest SNR values (−9 to +3 dB) as a function of the absolute *F*0 separation between the two talkers. (The +6- and +9-dB values were excluded from the average since performance was asymptoting, leaving little room for *F*0 differences to improve performance.) Although the improvement in performance from 0- to 1-semitone separation was only marginally significant ($F_{1,8}=5.9$, $p<0.05$, averaged over the lowest four SNRs), increasing the separation to 2 semitones improved performance by 12 percentage points, and increasing the separation to 12 semitones produced a 24 percentage point improvement in overall performance.

Note that, as the *F*0 separation was increased, the plateau in performance that occurred in the 0-semitone condition at SNRs less than 0 dB gradually disappeared. Statistical support for this observation is provided by the fact that increasing the SNR from −6 to 0 dB produced a significant improvement in performance at *F*0 separations of 3, 4, 6, 9, and 12 semitones (*post hoc* Fisher LSD test on the arcsine-transformed individual subject data, $p<0.05$), but not at *F*0 separations of 0, 1, or 2 semitones. At separations of 9 or 12 semitones, performance declined monotonically from over 90% at +9 dB SNR to around 65% at −6 dB SNR (left panel). A similar result was seen in the earlier Brungart (2001) study, where performance plateaued at negative SNRs in the same-talker (TT) and same-sex (TS) masking conditions, but decreased gradually in the different-sex (TD) masking condition. Thus, the overall shape of the 9- and 12-semitone curves is similar to the TD condition of the earlier study. However, despite the fact that the mean *F*0 values of the male and female talkers in the CRM corpus differ by almost exactly 12 semitones (104 Hz versus 206 Hz), the maximum improvement in performance afforded by a 12-semitone shift in *F*0 (28% at 0 dB SNR) in this experiment was only half as large as the improvement from the TT to the TD condition in the earlier experiment. Thus, it is clear that differences in *F*0 alone cannot account for the difference in performance between the same-talker and different-sex masking conditions of the earlier two-talker study.

The data were also examined to determine what effect the relative values of *F*0 for the target and masker had on performance. There was no consistent difference between conditions with a higher-pitched target and those with a higher-pitched masker. In addition, there was no interaction between relative *F*0 and the sex of the talker.

Nearly all of the incorrect responses in the experiment consisted of the color and/or number that occurred in the masker sentence. Only 1.8% of the incorrect responses were not such intrusions. This result indicates that the performance benefits that occurred with separation in *F*0 were due to improvements in the listeners' ability to allocate the colors and numbers in the stimulus to the appropriate call-signs, and not due to an increase in their ability to correctly perceive the colors and numbers in the stimulus.

2. Differences between talkers

The overall results conceal considerable variation between individual talkers. Figure 2 shows the probability of a correct identification separately for each talker. The data

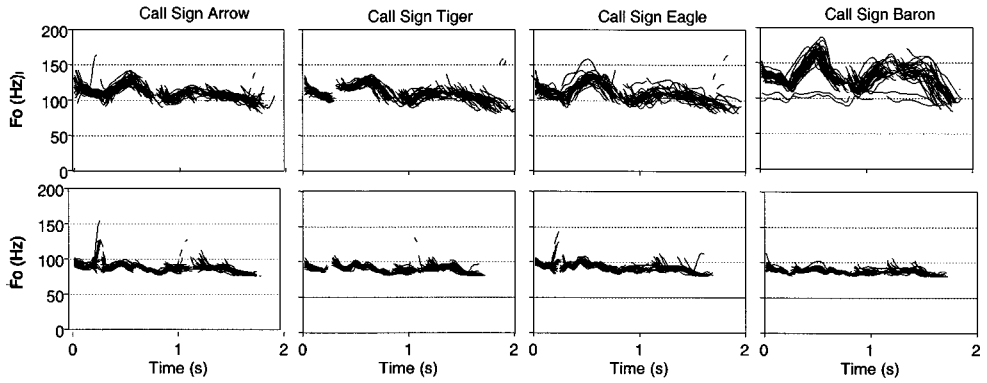

FIG. 3. Plots of the *F*0 contours for two male talkers. The upper row is talker 5, the lower row talker 8. Each plot overlays the *F*0 contours from all 32 sentences that used a particular call sign. The target call sign (''Baron'') is no. 7. The extracted *F*0 contours have been smoothed by low-pass filtering at 10 Hz.

from female talkers are shown in the upper row, with males in the lower row. As an extreme example of the differences between talkers, compare talker 5 (bottom left) with talker 8 (bottom right). For talker 5, an *F*0 separation of 12 semitones between the target and masker phrases did not improve performance over the condition in which there was no *F*0 separation; performance to his utterances is already very high with 0-semitone separation. By contrast talker 8's performance is much lower with 0-semitone separation than is talker 5's, and improves dramatically when the *F*0 separation is increased to 12 semitones.

These striking differences between talkers are at least partly due to their different intonation patterns. The intonation patterns for talkers 5 and 8 are shown in Fig. 3. Talker 8 (bottom row) spoke all his sentences with a rather flat intonation, which varied very little across the different call signs. Consequently, listeners were not able to use either instantaneous differences in *F*0 or differences in the overall contour to help them follow the target rather than the masker sentence. Introducing an artificial overall shift in *F*0 between the target and masker sentences, however, would have allowed listeners to use *F*0 differences in this way. Talker 5, on the other hand, used large excursions of *F*0 in his intonation, which, although similar for sentences with the same call-sign, differed substantially between call-signs: in particular, his intonation for the target call-sign ''Baron'' is distinctively different from the others. Even without any artificial shift in *F*0, this talker's speech provided substantial *F*0 differences between the target and masker sentences to help a listener in tracking the target sentence.

We now provide a quantitative test of whether the improvement with *F*0 separation is greater for talkers who show less initial difference in the original *F*0 between the target and masker sentences. Figure 4 shows a strong inverse correlation ($r^2=0.93$, df$=7$, $p<0.001$) between the improvement between 0 and 12 semitones averaged across SNR for individual talkers, and the average rms instantaneous difference in *F*0 between target and masker sentences for that talker. We estimated this rms difference with Praat by taking 50 random target-masker pairs for each talker, extracting their *F*0 contours (with 0.01-s intervals between points), smoothing these contours with a low-pass 10-Hz filter and then taking the rms difference in Hz for *F*0 at each time slice throughout the sentence pair. Time slices for which either sentence was silent or voiceless did not contribute to the rms average. These averages were themselves averaged across the 50 random sentence pairs to give the data for each talker plotted in Fig. 4.

C. Discussion

1. Effect of F0 separation

The main effect of *F*0 separation is that performance gradually improves as *F*0 separation increases, with little improvement at 1-semitone separation, and maximum improvement at 12 semitones. It is important to bear in mind that these nominal semitone separations in *F*0 have been added to the existing natural variations between utterances of the same talker. This natural variation is of the order of 5% to 10% for seven of the eight talkers and is probably responsible for the fact that overall improvement in the 1-semitone condition was barely detectable. The remaining talker (talker 5) had an unusually large variation in *F*0 between his target and masker sentences, and performance remained essentially unchanged as the separation in *F*0 varied from 0 to 12 semitones.

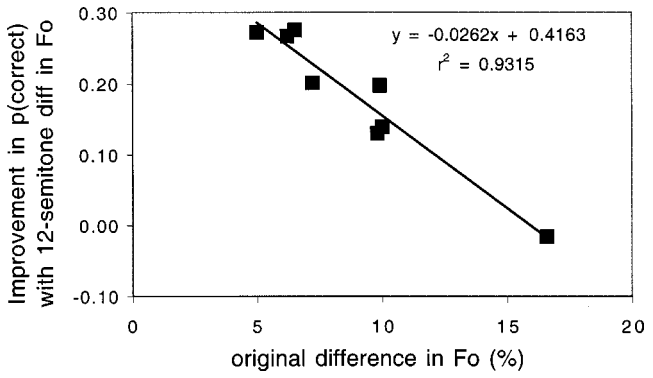

FIG. 4. The abscissa shows the rms % difference in *F*0 between the target voice and the masking voice for a particular talker condition with no alteration to *F*0. The ordinate shows the average change in performance for that talker across all six S/N ratios between an *F*0 shift of zero and 12 semitones. The figure shows a strong inverse relationship between the initial *F*0 difference between the target voice and the masker voice for a particular talker and the improvement in performance obtained by imposing a 12-semitone difference in *F*0 between the two voices.

The overall improvement in performance with *F*0 separation of the target and masking sentences that we have found in this experiment is compatible with previous experiments that have shown that a difference in *F*0 can help listeners segregate a target sentence from competing speech.

Brokx and Nooteboom (1982) asked their listeners (in their first experiment) to recall semantically anomalous sentences of fixed syntactic form (''*The town swims now in a sheep*'') that had been resynthesized (through a formant synthesizer after LPC-based formant tracking) on a monotone *F*0 contour. The sentences were played against a continuous background of speech—a short story read by the same talker—that had also been reduced to a monotone. The signal-to-noise ratio between the sentences and the background speech was either 0, −5, −10, or −15 dB. Content word identification (allowing a single phoneme alteration) of the sentences was measured as a function of the *F*0 difference between the sentences and the background speech. Brokx and Nooteboom found that word identification improved roughly linearly from about 40% to about 60% as the *F*0 difference increased from 0 to 3 semitones, but then identification dropped to about 50% when the separation was 12 semitones.

Two features of our data are different from these results of Brokx and Nooteboom. First, performance in our experiment barely increases overall between the 0- and 1-semitone conditions. Second, in our data the 12-semitone condition gives the best performance; in Brokx and Nooteboom's it did not. Both of these differences have similar explanations, which depend on the natural pitch contours used in our experiment.

In the Brokx and Nooteboom experiment, the *F*0 contours corresponding to a 0-semitone shift were identical, whereas in our experiment they had the variation inherent in natural utterances. As can be seen from the abscissa values in Fig. 4 the target and masker sentences had an average instantaneous difference of between about 5% and 17%, respectively (about 1 and 3 semitones), depending on the talker. These instantaneous differences in *F*0 prevent the fusion, and corresponding decrease in intelligibility (Scheffers, 1983; Assmann and Summerfield, 1989), that occurs with identical *F*0 values. They also provide some differential *F*0-contour information to help a listener to track a particular sound source over time. A small difference in *F*0 will thus have more of an effect on performance with Brokx and Nooteboom's monotone sentences than with ours. The reduction in performance with a one-octave (12-semitone) *F*0 separation in the Brokx and Nooteboom experiment has a similar explanation. The exact octave relationship between the two monotone utterances again causes fusion of the two sounds into a less-intelligible whole. In our experiments the 12-semitone separation was imposed on top of the natural variation in *F*0 and so there was no fusion; moreover, the large *F*0 difference between the sentences probably allowed listeners to track a particular sound source effectively.

Brokx and Nooteboom's (1982) second experiment used stimuli whose *F*0 contours were more similar to ours. Although the content of the utterances was the same as in their first experiment, the speech remained natural rather than being resynthesized by LPC coding, and different *F*0 conditions were obtained by the same talker naturally producing four different types of intonation: (1) normal intonation in the same *F*0 range as was used for the background passage (averaging about 110 Hz); (2) normal intonation at a high pitch (averaging about 160 Hz); (3) intended monotone at about 110 Hz; and (4) intended monotone at about 220 Hz. These four intonation conditions were presented at six different SNRs from 0 to −15 dB. With normal (rather than monotonous) intonation, performance was considerably better when the *F*0 of the target sentences was in a higher range than the background passage (72%) compared with when it was in the same range (55%). However, when the speech was spoken on a monotone, performance improved very little on the higher *F*0.

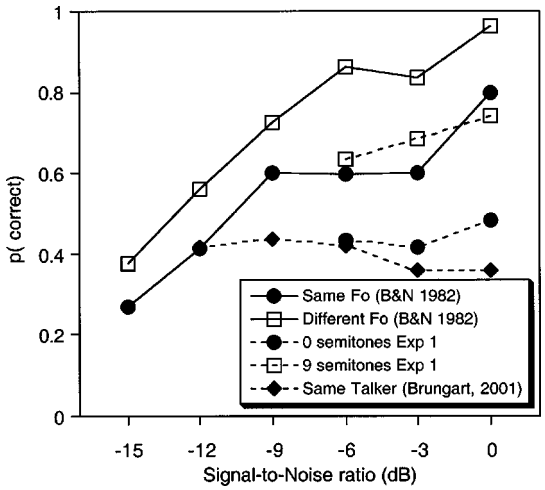

FIG. 5. Comparison of the data from experiment 1 with data from earlier experiments. Data joined by solid lines are from experiment 2 of Brokx and Nooteboom (1982) and show the probability of correct recall of content words from naturally spoken semantically anomalous sentences against a background of speech from the same talker in the same (solid circles) or different (open squares) pitch range. Data joined by dashed lines come from the same talker condition (solid diamonds) of Brungart (2001) and from the 0- (solid circles) or 9-semitone (open squares) conditions of experiment 1.

The *F*0 separation of the first two conditions of Brokx and Nooteboom's second experiment are similar to our 0- and 9-semitone manipulations. Figure 5 compares the results from these conditions at various SNRs. The figure also shows data from the same-talker condition of Brungart (2001) which was very similar to the present 0-semitone condition, but, like the Brokx and Nooteboom experiments, used a wide range of SNRs. Although overall performance in our data is lower (perhaps because of the rhythmic and semantic similarity between our target and mask utterances), the improvement in correct identifications with a difference in *F*0 is similar (c. 30%) in both experiments.

Larger changes in intelligibility of speech with differences in *F*0 have been reported by Bird and Darwin (1998). Their listeners had to recall the shorter of two simultaneous sentences that had been constructed to contain few stop and fricative consonants (e.g., ''*I only moan in the morning*''). The sentences were spoken on a monotone and resynthesized using LPC or PSOLA to have different monotonous *F*0 values. Intelligibility with the lower-quality LPC resynthesis in-

creased from about 20% correct words to 80% as the *F*0 difference increased from 0 to 8 semitones. With the higher-quality PSOLA resynthesis (and a new talker) the change was a little less (from 35% to 75%) from 0 to 10 semitones. The rather larger effect of a difference in *F*0 found in this experiment compared with our present experiment is probably due to fusion of the strictly monotone sentences at 0-semitone separation, together with the greater reliance on *F*0 to track an individual sentence in the absence of the onset and offset cues provided by stop and fricative consonants.

III. EXPERIMENT 2: CHANGES IN VOCAL TRACT LENGTH ONLY

The results of experiment 1 clearly show that artificial separations in the *F*0 values of the talkers can produce substantial performance improvements in a multitalker listening task. However, even a 12-semitone change in *F*0 did not produce as large an improvement in performance as a change in the sex of the masking talker. Thus, it is clear that differences in *F*0 alone cannot account for a listener's ability to segregate different-sex talkers. A second experiment was conducted to examine the effect of another perceptual property that listeners may be able to use to segregate different-sex talkers: the length of the vocal tract.

A. Method

The primary difference between experiments 2 and 1 is that the *F*0 shifts that were examined in the first experiment were replaced by changes in the vocal-tract lengths of the target and masking talkers. As was the case with the *F*0 shifts, these vocal-tract shifts were implemented by processing each of the 1024 sentences in the CRM corpus with the Praat software package. The apparent vocal-tract length of each talker was changed by a factor of vt for each utterance by (1) multiplying *F*0 by vt and duration by $1/vt$ (using PSOLA), (2) resampling at the original sampling frequency multiplied by vt, and then (3) playing the samples at the original sampling frequency. The end effect of this manipulation was to maintain the same duration and *F*0 of the original utterance but to scale the spectral envelope by vt. Scaling the spectral envelope is not identical to a change in vocal-tract length since it scales all those factors that are responsible for the spectral envelope. These factors include for example the spectral envelope characteristics of the voice source (such as spectral tilt) as well as the vocal-tract transfer function. However, the vocal-tract resonances are the main factor responsible for spectral envelope shape, and for simplicity we will refer to the manipulation as a change in vocal-tract length.

The range of vocal-tract ratios that we used is based on the average formant-frequency ratio between female and male voices reported by Peterson and Barney (1952) (of around 16%). Each utterance was processed with the following values of vt: 1.16, 1.08, 1.04, 1.02, 1.0, 0.98, 0.96, 0.92, 0.84. Pairs of utterances from the same original talker were then selected on each trial to produce one of the nine relative values of vt shown in Table II. These relative values were used in an experiment that was otherwise identical to experiment 1. Lengthening the vocal tract leads to a more male-sounding voice (though without the usual lower pitch), shortening it to a more female-sounding voice.

TABLE II. Vocal-tract length differences tested in the stimuli of experiment 2. Each pair of sentences in a trial was chosen from the same talker. The scaling values were selected differently for male and female talkers to place most of the scaled vocal tracts into the range between typical male and female talkers. Thus, male talkers were generally scaled down more than up, and female talkers were generally scaled up more than down. Note that, on a given trial, the target talker was equally likely to receive the higher or lower vt scaling.

	Male		Female	
vt Ratio	Longer VT	Shorter VT	Longer VT	Shorter VT
1.0	1.0	1.0	1.0	1.0
1.02	1.0	0.98	1.02	1.0
1.04	1.02	0.98	1.02	0.98
1.08	1.04	0.96	1.04	0.96
1.13	1.04	0.92	1.08	0.96
1.16B	1.08	0.92	1.08	0.92
1.16A	1.0	0.84	1.16	1.0
1.38	1.16	0.84	1.16	0.84

Eight of the nine paid volunteer listeners who were used in the first experiment had also participated in experiment 2. Four of the listeners were males, and four were females. Each completed ten trials for each of the 384 possible stimulus configurations of the experiment (8 talkers×8 vt differences×6 SNRs), for a total of 3840 trials. These trials were completed in blocks of 192 trials, with the trials within each block randomly balanced to have an equal number of trials for each value of each of the three independent variables (talker, vt difference, and SNR). Each block took approximately 10 min to complete, and each listener participated in two to three blocks per day over a 2-week period.

B. Results

The left panel of Fig. 6 shows the probability of a correct identification of both the color and number of the target sentence as a function of the signal-to-noise ratio and the ratio of vocal-tract length changes. The baseline condition with a vt ratio of 1.0 (open circles in Fig. 6) produced results very similar to those obtained in previous same-talker con-

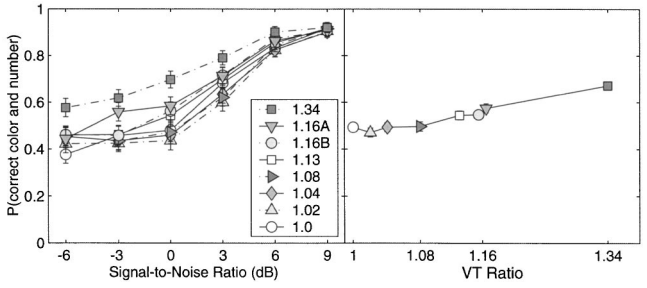

FIG. 6. The left panel shows the probability of a correct identification in experiment 2 of the number and color in the target sentence as a function of SNR for eight different ratios of vocal-tract length between the two sentences. The ratio is always taken so as to be unity or greater. The right panel shows performance averaged across the lowest four SNR values (−6, −3, 0, and 3 dB) as a function of the vt ratio. The data have been averaged over all eight talkers. The error bars represent 95% confidence intervals.

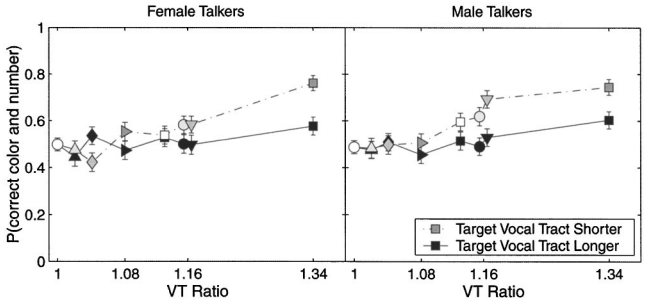

FIG. 7. Data as in the right panel of Fig. 6, except the results have been shown separately for female talkers (left panel) and male talkers (right panel), with separate curves representing trials where the target voice had a shorter vocal tract (open and shaded symbols) or a longer vocal tract (black symbols) than the masking voice.

figurations, with a decrease in performance from +9 to 0 dB SNR and then a plateau in performance at SNR values less than 0 dB.

The right panel of Fig. 6 shows the percentage of correct responses as a function of vt ratio averaged across the lowest four SNR values tested in the experiment (SNR$\leqslant$+3 dB). At 0-dB SNR, detectable improvement only appeared for vt ratios of 1.13 ($p<0.0001$) or greater. Changes of vt ratio of 1.08 or less gave no improvement in performance ($F_{3,21}=1.94$, $p>0.1$). For the largest vt ratio (1.38), however, there was a substantial improvement in performance compared to the baseline condition with equal-length vocal tracts: the percentage of correct responses increased from about 50% correct to 70% correct. This improvement is comparable to that produced by an *F*0 shift of 9 semitones in experiment 1 both in the magnitude of the improvement at 0 dB SNR and in the overall shape of the curve (filled squares in left panel of Fig. 6).

The results so far have not distinguished between trials where the target sentence had the shorter or the longer vocal tract. This distinction is made in Fig. 7, which is similar to the right panel of Fig. 6, but breaks the results down by the relative length of the target vocal tract (shaded and black symbols in each panel) and by the sex of the talker (females in the left panel, males in the right panel). The main result shown in the figure is that performance was consistently higher in trials where the target talker had a shorter vocal tract. Indeed, it appears that manipulating the vt ratio had very little effect on the proportion of correct responses when the masking talker had the shorter vocal tract. It is interesting to note that this effect was equally strong for both male and female talkers. Thus, the performance advantage seen for the shorter vocal tract does not appear to depend on the initial vocal-tract length of the talker. This result suggests that listeners may be biased to focus their attention on the talker with the shorter vocal-tract length when no other cues are available to segregate the two talkers.

As was the case in the first experiment, talker 5 produced the best performance in the baseline condition and showed the least improvement with increased vt ratio. At 0 dB SNR, there was no improvement in performance between the 1.0 and 1.34 vt ratios with talker 5, compared to an average improvement of 30 percentage points for the other seven talkers in the experiment. The large natural variations in *F*0 that occurred in the speech from talker 5 probably provided sufficient segregation such that no additional benefit was gained from changing the vocal-tract length of that talker.

As in experiment 1, the vast majority of incorrect responses contained the color and/or number from the masker sentence. Only 2.7% of the errors were not such intrusions. The improvements that we see with vt separation are thus improvements in the listeners' ability to allocate correctly perceived colors and numbers to the appropriate call-sign sentence.

C. Discussion

This experiment asked whether changes in vocal-tract length between the talkers of the target and masker sentences could improve performance. The main result of this experiment is that although small changes (8% or less) do not improve performance, larger changes do. Across the four lowest SNRs, a change of 13% in vt length increases performance from about 50% to 55% ($p=0.001$) and a 38% change increased it further to 67% ($p<0.0001$).

This result is entirely consistent with previous experiments by Darwin and Hukin (2000). They found that small differences in vocal-tract length ($\leqslant$8%) did not help listeners to allocate a target word to the target carrier sentence, but a difference of $\pm$17% helped substantially. It may be the case that small changes ($\pm$4%) in vocal-tract length are not effective in segregation because listeners do not hear them as a change in the identity of the talker. Kuwabara and Takagi (1991), for example, found that an upward or downward shift of 8% in the first three formant frequencies was sufficient to reduce individual voice recognition to chance. The smaller, bi-directional changes that we used may not be sufficient to change the identity of talker's voices that were familiar to our listeners.

D. Interim discussion

These two experiments have identified two factors which are likely to have contributed to the improvement in performance found by Brungart (2001) for different sex talkers relative to identical talkers. A difference in talker sex produces a difference in overall *F*0 of the target and masker sentences of a little under an octave and a change in vocal-tract length of about 16% (Peterson and Barney, 1952). Artificially manipulating the *F*0 and the vocal-tract length separately also gave substantial increases in performance. However, it is unlikely that these two factors alone are entirely responsible for the performance differences originally found. If we look simply at the data gathered with a SNR of 0 dB, and assume that the effects of *F*0 and of vocal-tract length are additive, then we find that together they provide less improvement than was found with different-sex talkers. Specifically, we converted the probability of a correct response to d' values using published tables (Hacker and Ratcliff, 1979), assuming a choice between four orthogonal alternatives (since almost all errors were the color or number from the masker sentence). Changing the sex of the talker

TABLE III. *F*0 ratio (and equivalent semitone) and vt values used to modify male voices in experiment 3.

Male original	*F*0 ratio	Semi tones	vt
Super-male	0.74	−5.2	1.08
Male	1.00	0.0	1.00
Quarter-female	1.17	2.7	0.96
Half-female	1.35	5.2	0.92
Almost female	1.53	7.4	0.88
Female	1.70	9.2	0.84
Super-female	2.05	12.4	0.76

TABLE IV. *F*0 ratio (and equivalent semitone) and vt values used to modify female voices in experiment 3.

Female original	*F*0 ratio	Semi tones	vt
Super-male	0.49	−12.4	1.24
Male	0.59	−9.2	1.16
Almost male	0.66	−7.2	1.12
Half-male	0.74	−5.2	1.08
Quarter-male	0.85	−2.7	1.04
Female	1.00	0.0	1.00
Super-female	1.35	5.2	0.92

increases d' by about 1.91; a change of 12 semitones in *F*0 increases d' by 0.97, and a change in vocal-tract length of 1.16 increases d' by 0.46. We can combine these two d' measures in one of two ways in order to obtain an expected d' when both cues are present. Any d' scores can be linearly added if the underlying noise distribution is the same for both measures, or orthogonally added (as the square root of the sums of their squares) if the underlying noise distributions are independent. With each of these ways of combination there is a shortfall attributable to other voice characteristics, which is between about 0.48 (assuming linear additivity) and 0.84 (assuming orthogonality). These voice characteristics may also be at least partly responsible for the 0.67 improvement in d' that is found in Brungart's data (2001) for different talkers of the same sex (TS) compared with identical talkers (TT). One of the ways in which talkers differ is in the timing of their speech, but there may also be a super-additive effect of *F*0 and vocal-tract length differences, when they are combined in a natural way. The next experiment examines improvement on the task when both *F*0 and vocal-tract length co-vary in a natural way.

IV. EXPERIMENT 3: CHANGE IN BOTH *F*0 AND VOCAL TRACT LENGTH

The first two experiments examined the effects that isolated changes in *F*0 and vocal-tract length have on a listener's ability to segregate two talkers. However, neither of these manipulations alone was found to produce the level of performance that occurs when the target and masking sentences are spoken by different-sex voices that differ both in *F*0 and in vocal-tract length. Thus, a third experiment was conducted that varied both *F*0 and vt in concert to simulate a smooth transition between a same-sex and different-sex masking voice.

A. Methods

Most of the procedures for the third experiment were the same as for the first two, but the stimuli consisted of speech phrases in which both the vt ratio and the *F*0 frequency were manipulated at the same time. The particular vt and *F*0 values that we used are shown in Tables III and IV. Another difference with the first two experiments is that here on each trial listeners heard one message with the *F*0 and vt unchanged; these unchanged values are referred to as "male" in Table III and "female" in Table IV. From the original male voices we generated "female" voices with *F*0 multiplied by 1.7 and vt by 0.84. These values correspond to the average female/male ratios for the formant and *F*0 data reported by Peterson and Barney (1952). Similar, inverse changes were made to the original female voices to give new "male" voices. Three intermediate voices were then linearly interpolated between each original voice and its changed counterpart (which we have arbitrarily labeled as "quarter-," "half-," and "almost" shifts to the opposite gender in the table), and two more extreme values were linearly extrapolated (the super males and super females of Tables III and IV).

The experimental procedure was similar to the first two experiments. Eight paid volunteer listeners (four male, four female) participated in experiment 3. All but two had previously participated in the first two experiments. Each listener completed ten trials for each of the 336 possible stimulus configurations of the experiment (8 talkers×7 gender differences×6 SNRs), for a total of 3360 trials. These trials were completed in blocks of 168 trials, with each block randomly balanced to have an equal number of trials for each value of each of the three independent variables (talker, gender difference, and SNR). Each block took approximately 9 min to complete and each listener participated in two to three blocks per day over a 2-week period.

B. Results and discussion

The results of the experiment are shown separately for the male and female talkers in the left panels of Figs. 8 and 9. Once again, the results are shown in terms of the percentage of correct identifications of the color and number in the target sentences as a function of the SNR of the stimulus. The results for the unshifted masking voices (open circles in the figures) were once again similar to those from previous experiments, with a rapid decrease in performance from +9 to 0 dB SNR, and a plateau in performance at about 45% correct responses at SNRs less than 0 dB. However, the results for the gender-shifted sentences show a systematic increase in performance up to a substantially higher level of performance than that obtained in either of the first two experiments. At a SNR of 0 dB, overall performance improved approximately 35 percentage points (from 45% to 80%) at the larger gender shifts tested for both the male and female talkers. This brought performance up to a level comparable to that measured with different-sex talkers in the earlier experiment by Brungart (2001).

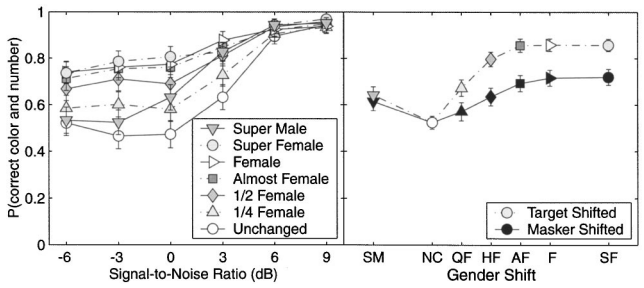

FIG. 8. The left panel shows the probability of a correct identification of the number and color in the target sentence as a function of SNR for the seven different gender shifts introduced into the male voice in experiment 3 (see Table III). The right panel shows performance averaged across the lowest four SNRs (−6, −3, 0, and 3 dB) as a function of the gender shift, with separate curves for trials where the target voice was shifted (open and shaded symbols) and trials where the masking voice was shifted (black symbols). The error bars represent the 95% confidence intervals of the mean.

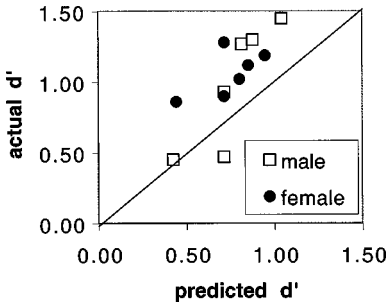

FIG. 10. The actual d' improvement in experiment 3 at 0 dB SNR when one voice changes in both $F0$ and vocal-tract length versus predicted d' from separate changes in either variable in experiments 1 and 2. Published tables (Hacker and Ratcliff, 1979) were used to convert the d' values from the probability of a correct color and number identification in each condition with the assumption that each response represented a choice between four orthogonal alternatives. See text for details.

The right panels of Figs. 8 and 9 show performance averaged across the four lowest SNRs tested in the experiment (−6 to +3 dB) as a function of the gender shift introduced between the two talkers in the stimulus. These data are plotted separately for trials where the target speech was shifted in gender (shaded and open symbols) and where the masker was shifted in gender (black symbols). The data for both the male and female talkers show that performance was generally better when the target was shifted in gender than when the masker was shifted in gender. In part, this may occur because the listener's attention is drawn to the unusual characteristics of the gender-shifted talker. The difference between the target-shifted and masker-shifted configurations was largest when the male target talker was female-shifted, which may reflect the bias in favor of the talker with the shorter vocal-tract length found in experiment 2. The difference between the target-shifted and masker-shifted configurations was substantially smaller for the male-shifted female talkers, which may reflect a conflict between the novelty of the gender-shifted talker and the bias in favor of the talker with the shorter vocal tract. Overall, the results suggest that the performance advantages that occurred with the shorter vocal tract in experiment 2 represent a relatively weak effect that only dominates the results when the vocal-tract length is changed at a fixed value of $F0$.

The results of experiment 3 also reveal other asymmetries between the effects of the gender shifts for the male and female talkers. When the voice characteristics of the shifted voice fell between the normal ranges of male and female voices, there was generally a more rapid improvement in performance with gender change for the female talkers than for the male talkers. This was especially true for the female-$\frac{1}{4}$ male condition, which produced performance about 10 percentage points higher than the corresponding male-$\frac{1}{4}$ female condition (upward-pointing triangles in the right panels of the figures). In both cases, however, nearly 100% of the benefit of separation in gender was achieved in the almost male or almost female condition, which represent the normally occurring differences in vt and $F0$ for a male and female voice (squares in the right panels of the figures). A larger asymmetry between the male and female talkers occurred when the voice characteristics were shifted outside the normally occurring range of vt and $F0$. Specifically, the performance improvement produced by shifting a female voice to a super-female voice (upside-down triangles in the right panel of Fig. 9) was nearly three times as large as the performance improvement produced by shifting a male voice to a super-male voice (upside-down triangles in the right panel of Fig. 8). The reason for this differential improvement is seen more clearly in the left panel of Fig. 8, which shows that performance in the super-male condition falls off substantially more rapidly than performance in any condition except the unshifted condition as the SNR of the stimulus decreased. It appears that the listeners had extreme difficulty segregating a male voice from a super-male voice at low SNRs. At this point we have no explanation for this result.

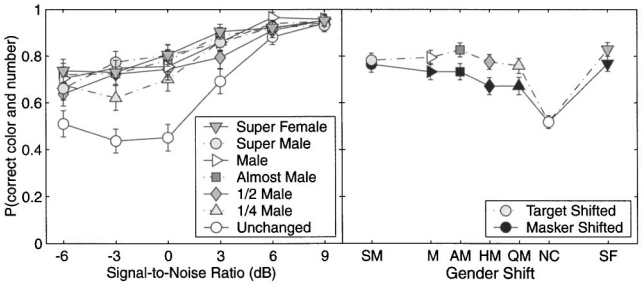

FIG. 9. Identical to Fig. 8 except the data are shown for the female talkers in experiment 3 with the gender shifts outlined in Table IV.

To address the question of whether the effects of a shift in $F0$ and a shift in vocal-tract length are additive, we predicted the improvements in performance in the 0 dB SNR conditions of experiment 3, which varied both $F0$ and vocal-tract length, from the improvements shown for these two parameters separately in experiments 1 and 2. We assumed that the two factors were orthogonal and so the square of the predicted d' was the sum of the squares of the two predicting d's. The results of this analysis in Fig. 10 show that the factors behave super-additively—predicted values for performance when both variables change are, except for 1 point out of 12, lower than the actual values obtained in experiment 3. Listeners therefore gain more benefit from a joint change in

*F*0 and vocal-tract length than is predicted from changes to each of these variables separately.

This super-additive combination of *F*0 and *vt* cues can explain most, but not all, of the improvement in performance that occurs when the target and masking speech signals are spoken by different-sex talkers than when they are spoken by the same-sex talkers. Brungart (2001) found that listeners responded correctly in the CRM task approximately 85% of the time when different-sex competing talkers were presented at a SNR of 0 dB. However, even in the male–super-female and female–super-male configurations of experiment 3, where the *F*0 shift (105%) was slightly larger than the average *F*0 difference between the male and female talkers in the corpus (100%) and the *vt* shift (24%) was twice as large as the average *vt* shift between the male and female talkers in the corpus (12%), the listeners responded correctly only approximately 80% of the time when the SNR was 0 dB. The 5 percentage point performance difference between the best condition in experiment 3 and the ''TD'' condition in the earlier experiment is presumably the result of other variations in the voices of the different talkers in the corpus, such as voice-source characteristics, intonation, and speaking rate.

V. CONCLUSIONS

These three experiments have examined the roles that differences in *F*0 and differences in vocal-tract length have on the segregation of multiple simultaneous speech signals. The imposed *F*0 shifts were additional to the natural differences already present between separate utterances by the same talker. The major results can be summarized as follows:

(1) Differences in *F*0 produce systematic improvements in segregation performance when the difference is 2 semitones or greater, but *F*0 alone cannot improve performance to the level that occurs with different-sex talkers.
(2) Differences in vocal-tract length produce systematic improvements in segregation performance when the ratio of lengths is 1.08 or greater, but differences in vocal-tract length alone cannot improve performance to the level that occurs with different-sex talkers. In general, performance is better when the shorter vocal tract belongs to the target talker.
(3) Systematic changes in both *F*0 and vocal tract that simulate an incremental shift in gender produce substantially larger improvements in performance than differences in *F*0 or *vt* alone. The combined effect is larger than that predicted assuming additivity. In general, shifting one of two utterances from a female voice towards a male voice produces a greater improvement in performance than shifting from male towards female.
(4) The intonation patterns of the individual talkers can play as large a role as *F*0 or *vt* in determining overall segregation performance. The natural variations in intonation in the utterances spoken by talker 5 were so large that no additional improvement was obtained by introducing artificial changes in the *F*0s or vocal-tract lengths of the competing phrases.

ACKNOWLEDGMENTS

The first author was supported by Grant No. GR/M90146 from the UK EPSRC and Grant No. G9801285 from the UK MRC. The second author was supported in part by AFOSR Grant No. LRIR01HE01COR. Peter Assmann and two anonymous reviewers provided helpful comments on the paper.

Assmann, P. F., and Summerfield, A. Q. (**1989**). ''Modelling the perception of concurrent vowels: Vowels with the same fundamental frequency,'' J. Acoust. Soc. Am. **85**, 327–338.

Assmann, P. F., and Summerfield, A. Q. (**1990**). ''Modelling the perception of concurrent vowels: Vowels with different fundamental frequencies,'' J. Acoust. Soc. Am. **88**, 680–697.

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

Bird, J., and Darwin, C. J. (**1998**). ''Effects of a difference in fundamental frequency in separating two sentences,'' in *Psychophysical and Physiological Advances in Hearing*, edited by A. R. Palmer, A. Rees, A. Q. Summerfield, and R. Meddis (Whurr, London), pp. 263–269.

Boersma, P., and Weenink, D. (**1996**). ''Praat, a System for doing Phonetics by Computer, version 3.4,'' Institute of Phonetic Sciences, University of Amsterdam, Vol. 132, pp. 1–182, www.praat.org

Bolia, R. S., Nelson, W. T., Ericson, M. A., and Simpson, B. D. (**2000**). ''A speech corpus for multitalker communications research,'' J. Acoust. Soc. Am. **107**, 1065–1066.

Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (Bradford Books, MIT, Cambridge, MA).

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

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

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

Brungart, D. S., Simpson, B. D., Scott, K. R., and Ericson, M. A., (**2001). ''Informational and energetic masking effects in the perception of multiple simultaneous talkers,'' J. Acoust. Soc. Am. 110**, 2527–2538.

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

Dirks, D., and Bower, D. (**1969**). ''Masking effects of speech competing messages,'' J. Speech Hear. Res. **12**, 229–245.

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

Ericson, M. A., and McKinley, R. L. (**1997**). ''The intelligibility of multiple talkers separated spatially in noise,'' in *Binaural and Spatial Hearing in Real and Virtual Environments*, edited by R. H. Gilkey and T. R. Anderson (Lawrence Erlbaum, Mahwah, NJ), pp. 701–724.

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

Hacker, M. J., and Ratcliff, R. (**1979**). ''A revised table of d' for *M*-alternative forced choice,'' Percept. Psychophys. **26**, 168–170.

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

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

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

Scheffers, M. T. (**1983**). ''Sifting vowels: Auditory pitch analysis and sound segregation,'' Ph.D. dissertation, Groningen University, The Netherlands.