Published in final edited form as:

J Acoust Soc Am. 2009 February ; 125(2): 1114–1124. doi:10.1121/1.3050321.

The interaction of vocal characteristics and audibility in the recognition of concurrent syllables^a)

Martin D. Vestergaard

Centre for the Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

Nicholas R. C. Fyson

Centre for the Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

Bristol Centre for Complexity Science, Department of Engineering Mathematics, University of Bristol, Queen's Building, University Walk, Bristol BS8 1TR, United Kingdom

Roy D. Patterson

Centre for the Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, Downing Street, Cambridge CB2 3EG, United Kingdom

Abstract

In concurrent-speech recognition, performance is enhanced when either the glottal pulse rate (GPR) or the vocal tract length (VTL) of the target speaker differs from that of the distracter, but relatively little is known about the trading relationship between the two variables, or how they interact with other cues such as signal-to-noise ratio (SNR). This paper presents a study in which listeners were asked to identify a target syllable in the presence of a distracter syllable, with carefully matched temporal envelopes. The syllables varied in GPR and VTL over a large range, and they were presented at different SNRs. The results showed that performance is particularly sensitive to the combination of GPR and VTL when the SNR is 0 dB. Equal-performance contours showed that when there are no other cues, a two-semitone difference in GPR produced the same advantage in performance as a 20% difference in VTL. This corresponds to a trading relationship between GPR and VTL of 1.6. The results illustrate that the auditory system can use any combination of differences in GPR, VTL, and SNR to segregate competing speech signals.

I. INTRODUCTION

In multispeaker environments, listeners need to attend selectively to a target speaker in order to segregate their speech from distracting speech sounds uttered by other speakers. This paper is concerned with two speaker-specific acoustic cues that listeners use to segregate concurrent speech: glottal pulse rate (GPR) and vocal tract length (VTL). GPR and VTL are prominent markers of a speaker's size and sex in adults (Peterson and Barney, 1952; Fitch and Giedd, 1999; Lee *et al.*, 1999). Brungart (2001) reported an intriguing concurrent-speech study in which the target and distracting speakers were either the same sex or different sexes. He showed that listeners were better at identifying concurrent speech when the competing voice was from the opposite sex, but it was not possible to determine whether it was the GPR or the VTL difference that was more valuable. Darwin *et al.* (2003) reported

^a)Portions of this work were presented at the 153rd meeting of the Acoustical Society of America, Salt Lake City, UT, 2007.

separate effects of GPR and VTL on target recognition, and showed that there is additional improvement when the target and distracter differ in both vocal characteristics.

In these and other studies, where the speech stimuli are sentences, it is not possible to match the temporal envelopes of the competing speech sounds. As a result, listeners can monitor the concurrent speech for clean segments of target speech, and string the sequence of temporal “glimpses” together to improve performance (Cooke, 2006). To the extent that temporal glimpsing is successful, it reduces the listener's dependence on speaker differences, which in turn reduces the sensitivity of the experiment to the effects of speaker differences. The purpose of the current study was to enhance sensitivity to the value of speaker differences with a concurrent-syllable paradigm, in which the temporal envelope of the speech from the target is matched to the envelope of the distracter in an effort to reduce the potential for temporal glimpsing. Moreover, the domain of voices was extended to include combinations where the GPR was either unusually large or unusually small relative to the VTL, and vice versa.

A. Acoustic cues for segregating concurrent speech

The most significant cue for speech segregation is almost undoubtedly audibility, which is determined by the signal-to-noise ratio (SNR). When the speech material is sentences and they are matched for overall SNR, there are, nevertheless, momentary fluctuations in SNR that allow the listener to hear the target clearly. Miller and Licklider (1950) reported that listeners are capable of detecting segments of the target speech during relatively short minima of a competing temporally fluctuating background noise. Cooke (2006) used a missing-data technique to model the effect of temporal glimpsing, and concluded that it can account for the intelligibility of speech in a wide range of energetic masking conditions.

Vocal characteristics such as GPR and VTL also provide cues that support segregation of competing speech signals. GPR is heard as voice pitch, and a number of studies have demonstrated that performance on a concurrent-speech task increases with the pitch difference between the voices up to about four semitones (STs) (e.g., Chalikia and Bregman, 1993; Qin and Oxenham, 2005; Assmann and Summerfield, 1990, 1994; Culling and Darwin, 1993). The experiments reported in this paper focus on the effects of VTL differences on concurrent syllable recognition, and how VTL differences interact with differences in GPR. The syllable envelopes were matched to minimize temporal glimpsing, and the relative level of the target and distracter was varied across the range where they interact strongly, in order to determine whether SNR affects the form of the interaction between GPR and VTL in segregation.

B. The role of vocal characteristics in speech segregation

The role of pitch in concurrent speech has been investigated in many psychophysical studies, and in many cases the pitch is specified in terms of the fundamental frequency (F0) of the harmonic series associated with the GPR. Chalikia and Bregman (1993) used concurrent vowels to show that a difference in F0 leads to better recognition of both vowels. Furthermore, they showed that a difference in F0 contour can lead to better recognition in situations where the harmonicity of the constituents is reduced. Assmann and Summerfield (1994) used concurrent vowels to show that small departures, from otherwise constant F0 tracks, can improve vowel recognition, especially when the F0 difference is small. Qin and Oxenham (2005) used concurrent vowels to show that performance reached its maximum when the difference in F0 was about four STs. They also found that when the spectral envelope was smeared with a channel vocoder, an F0 difference no longer improved vowel recognition. Summerfield and Assmann (1991) argued that the advantage of an F0 difference derives from the difference in pitch *per se* and not from the difference in spectral

sampling of the formant frequencies, or glottal pulse asynchrony. In a series of related experiments, de Cheveigné and co-workers developed a harmonic cancellation model tuned to the periodicity of the distracter (de Cheveigné *et al.*, 1997b, 1997a; de Cheveigné, 1997, 1993). They showed that the advantage of an F0 difference in double-vowel recognition depends primarily on the harmonicity of the distracter. Culling and Darwin (1993) showed that when F0 tracks of concurrent speech are ambiguous, listeners can use the formant movements of competing diphthongs to disambiguate concurrent speech. They showed that listeners were only able to judge whether the F0 tracks of the speakers crossed or bounced off each other when the constituents had different patterns of formant movement.

When the F0 difference is small or the pitch is otherwise ill-defined, listeners have to use other acoustic cues to segregate concurrent speech. Brungart (2001) used noise and speakers of different sex as distracters in a concurrent-speech experiment. The distracting speech came from (a) the same speaker, (b) a different speaker of the same sex, or (c) a speaker of the opposite sex. Performance was measured with the coordinate response measure (CRM) that consists of sentences in the following form: “Ready /call-sign/ go to /color/ /number/ now,” where the call-sign is a name like Charlie or Ringo (Bolia *et al.*, 2000; Moore, 1981). Brungart (2001) found that the psychometric functions for noise and speech distracters had different shapes. A clear performance advantage was observed when the distracter was a different speaker from the target, and the biggest advantage arose when the distracter was of a different sex. For all of the speech distracters, the worst performance was at 0 dB SNR; at negative SNR, performance recovered somewhat. For noise maskers, performance was better overall, and no recovery was found for negative SNR. The results were interpreted in terms of informational masking, and they suggest that voices are more distracting than noise even when the noise is modulated. However, the release from masking at negative SNR only occurred for identification of the color coordinate in the target sentence. Brungart (2001) speculated that this might be because the numbers appear last in the CRM sentences and so might not overlap in time as much as the color coordinates. Thus, more temporal glimpsing was possible with the numbers than with the colors. This might explain the morphological difference between the psychometric functions for color and number. Glimpsing could also explain the relative inefficiency of the noise masker in that study. In Brungart *et al.* (2001), similar results were found for multiple distracting speakers except for the recovery phenomenon, indicating that it probably has to do with the extent to which words overlap in CRM.

Darwin *et al.* (2003) investigated the effects of F0 and VTL in a study on concurrent speech using the CRM corpus. They used a pitch-synchronous overlap-add technique as implemented in PRAAT to separately manipulate F0 and VTL in a set of sentences and produced a range of speakers with natural combinations of F0 and VTL. For an F0 difference of 12 STs, at 0 dB SNR, they reported an increase in speech recognition of 28%, most of which (~20%) was already apparent at an F0 difference of four STs (see Fig. 1 in Darwin *et al.*, 2003). They also found that individual differences in intonation can help identify speech of similar F0, corroborating the findings of Assmann and Summerfield (1994) mentioned above. For a 38% change in VTL, Darwin *et al.* (2003) reported an increase in recognition of ~20% at 0 dB SNR (see Fig. 6 in Darwin *et al.*, 2003). The largest performance increase was found for a combined difference in GPR and VTL, and they concluded that F0 and VTL interact in a synergistic manner. However, a large asymmetry was reported with regard to the effect of VTL. When the VTL of the target was smaller than the VTL of the distracter, the effect was much larger than when the VTL of the target was larger than the VTL of the distracter (by the same relative amount). As in the study of Brungart (2001), they made no attempt to control glimpsing.

In two related studies involving a lexical decision task, Rivenez *et al.* (2006, 2007) showed that differences in both F0 and VTL between two competing voices presented dichotically facilitated the use of priming cues in an unattended contralateral signal. In both their studies, an advantage was observed in terms of faster response time to the target stimuli, and the results were interpreted to lend support to the notion that early perceptual separation of the competing voices is a necessary prerequisite for lexical processing of the unattended voice.

The results described above support the hypothesis, originally proposed by Ladefoged and Broadbent (1957), that listeners construct a model of the target and distracting speakers, and that they use speaker-specific acoustic cues such as VTL and GPR as part of the model. Smith and Patterson (2005) showed that listeners can judge the relative size/age, and the sex of a speaker based on their vowels even when the GPR and VTL values were well beyond the range of normal speech. Collins (2000) showed that female listeners can make accurate judgments about the weights of male speakers based solely on their voices. To recognize a speech sound, the listener needs to normalize for phonetically irrelevant speaker-dependent acoustic variability (Nearey, 1989) such as that associated with variation in VTL and GPR. For unmasked speech, this is a trivial task for most listeners, and it leads to remarkable robustness in speech perception to speaker differences (Smith *et al.*, 2005). Smith *et al.* (2005) pointed out that the robustness of speech recognition is unlikely to be entirely due to learning, since speakers with highly unusual combinations of GPR and VTL are understood almost as well as those with the most common combinations of GPR and VTL. This is compatible with the suggestion that the auditory system operates early preprocessing stages that detect and normalize for GPR and VTL (Irino and Patterson, 2002). The preprocessing is applied to all sounds and is hypothesized to operate irrespective of attention and task relevance. In the current paper, it is argued that listeners use GPR and VTL in their speaker models and that this facilitates the segregation of competing voices. We hypothesize that the reason why it is possible to attend selectively to a particular speaker in a multispeaker environment is that the processing of VTL and GPR cues is automatic and occurs at an early point in the hierarchy of speech processing.

In natural speech, speakers vary GPR by changing the tension of the vocal folds, and they use GPR to convey prosody information within a range determined largely by the anatomical constitution of the laryngeal structures (Titze, 1989; Fant, 1970). By contrast, it is only possible to change the VTL by a small amount, either by pursing the lips or by lowering or raising the larynx, which require training, and both of which produce an audible change to the quality of the voice. The relative stability of the VTL cue suggests that VTL is likely to be at least as important for tracking a target speaker as GPR.

The purpose of the current study was to investigate the relative contribution of GPR and VTL in the recognition of concurrent speech, while carefully controlling other potential cues. The aims were (1) to quantify the effects of VTL and GPR, and (2) to model the relationship between them.

II. METHOD

The participants were required to identify syllables spoken by a target speaker in the presence of a distracting speaker. Performance was measured as a function of the difference between the target and distracting speakers along three dimensions: GPR, VTL, and SNR. In order to prevent the listeners from taking advantage of temporal glimpses, the temporal envelopes of the target and distracter syllables were carefully matched, as described below.

A. Listeners

Six native English speaking adults participated in the study (four males and two females). Their average age was 21 years (19–22 years), and no subject had any history of audiological disorder. An audiogram was recorded at the standard audiometric frequencies to ensure that the participants had normal hearing. The experiments were done after informed consent was obtained from the participants. The experimental protocol was approved by the Cambridge Psychology Research Ethics Committee.

B. Stimuli

The experiments were based on the syllable corpus previously described by Ives *et al.* (2005) and von Kriegstein *et al.* (2006). It consists of 180 spoken syllables, divided into consonant-vowel (CV) and vowel-consonant (VC) pairs. There were 18 consonants, 6 of each of 3 *categories* (plosives, sonorants,¹ and fricatives), and each of the consonants was paired with 1 of the 5 vowels spoken in both CV and VC combinations. The syllables were analyzed and resynthesized with a vocoder (Kawahara and Irino, 2004) to simulate speakers with different combinations of GPR and VTL. Since all the voices were synthesized from a recording of a single speaker (Patterson), the only cues available for perceptual separation were the GPR and VTL differences introduced by the vocoder. Throughout the experiment the target voice was presented at 60 dB SPL, while the RMS level of the distracting voice varied to achieve an SNR of +6, 0, or –6 dB.

1. Vocal characteristics—The voice of the target (the reference voice) remained constant throughout the experiments, and its characteristics were chosen with reference to typical male and female voices. Peterson and Barney (1952) reported that the average GPRs of men and women are 132 and 223 Hz, respectively, and Fitch and Giedd (1999) reported that the average VTLs of men and women are 155 and 139 mm, respectively. The geometric means of these values were used to simulate an androgynous target speaker with a GPR of 172 Hz and a VTL of 147 mm. The VTL of the original speaker was estimated to be 165 mm, and this value was used as a reference to rescale the original recordings in order to produce the voices used in the experiment. The combinations of GPR and VTL chosen for the distracter are shown by the dots in Fig. 1, which form an elliptical spoke pattern radiating out across the GPR-VTL plane from the reference voice. The ellipse that joins the ends of the spokes had a radius of 26% (4 STs= $2^{4/12}$) along the GPR axis and 41% (6 STs= $2^{6/12}$) along the VTL axis. The VTL dimension is proportionately longer because the just noticeable difference (JND) for VTL is at least 1.5 times the JND for GPR (Ives *et al.*, 2005; Ritsma and Hoekstra, 1974). There were seven points along each spoke, spaced logarithmically in this logGPR-logVTL plane in order to sample the region near the target speaker with greater resolution. Spokes 1 and 5 were tilted by 12.4° to form a line joining the average man with the average woman (Turner *et al.*, 2009; Peterson and Barney, 1952; Fitch and Giedd, 1999). The tilt ensured that there was always variation in both GPR and VTL between the target and distracter voices, which reduces the chance of the listener focusing on one of the dimensions to the exclusion of the other (Walters *et al.*, 2008). In all, there were 56 different distracter voices with the vocal characteristics shown in Table I.

An assumption underlying the design is that there is a trading relationship between VTL and GPR in the perceptual separation between the target voice and any distracter voice, and that the perceptual distance between voices can be expressed by the radial scale displacement (RSD) between their points in the logGPR-logVTL plane. The RSD is the geometrical distance between the target and distracting voices,

¹The category *sonorant* here refers to a selection of consonants from the manner classes: nasal, trill, and approximant (sometimes called semivowels) that are common in the English language (e.g., [m], [n], [r], [j], [l], and [w]).

$$\text{RSD}_\chi = \sqrt{\chi^2 (X_{\text{target}} - X_{\text{distracter}})^2 + (Y_{\text{target}} - Y_{\text{distracter}})^2}, \quad (1)$$

where X is log(GPR), Y is log(VTL), and χ is the GPR-VTL trading value that is 1.5 in the design. The RSD values shown in Fig. 1 are for $\chi = 1.5$. The design only requires the trading value to be roughly correct; so long as the voices vary in combination over ranges that go from indistinguishable to readily distinguishable in all directions, then the optimum trading value can be estimated from the recognition data.

2. Envelope control—A combination of techniques was employed to limit temporal glimpsing. First, the perceptual centers (Marcus, 1981) of the syllables were aligned as described by Ives *et al.* (2005). Second, the target and distracter syllables were matched according to their phonetic specification in the following way: (1) the CV order of the target and distracter syllables was required to be the same, and (2) the consonants in a concurrent pair of target and distracter syllables were from the same category. The result of these manipulations is that the temporal envelopes of the target and distracter syllables were closely aligned and similar in shape, as illustrated in Fig. 2. Within the six categories of syllables, pairs of target and distracter syllables were chosen at random with the restriction that the pair did not contain either the same consonant or the same vowel. These restrictions leave 20 potential distracter syllables for each target syllable.

C. Procedure

The study consisted of three parts: (1) pre-experimental training, (2) the main experiment, and (3) an SNR extension. The procedure was the same in all three: The target syllables were presented in triplets; the first syllable (the precursor) was intended to provide the listener with cues to the GPR and VTL of the target speaker; the second and third target syllables were presented with a concurrent distracter syllable. The three intervals of each trial were marked by visual indicator boxes on the graphical user interface (GUI) illustrated in Fig. 3. After the third interval was complete, the box for interval 2 or 3—chosen at random—was illuminated to indicate which of the two target syllables the listener was required to identify. The multisyllable format was intended to promote perception of the stimuli as speech and the sense that there were two speakers, one of whom was the target. The format makes it possible to vary the task difficulty in order to maximize the sensitivity of the recognition scores to the variation in GPR and VTL. In the present implementation, the trial includes *two* concurrent syllable pairs after the precursor, because pilot work showed that the task was too easy with just one concurrent syllable pair, and that the memory load was too great with three or more concurrent syllable pairs.

Listeners indicated their answers by clicking on the orthographical representation of their chosen syllable in the response grid on the screen. The participants were seated in front of the response screen in a double-walled IAC (Winchester, UK) sound-attenuated booth, and the stimuli were presented via AKG K240DF headphones.

1. Pre-experimental training—The ambiguity of English orthography meant that the response grid shown in Fig. 3 required some introduction. Listeners had to learn that the notation for the vowels was like that in the majority of European languages; that is, “a” is [ɑ:], “e” is [ɛ:], “i” is [i:], “o” is [o:] and “u” is [u:]. Moreover, they had to learn to find the response syllables on the grid rapidly and with confidence. In the first training session, target syllables without distracters were presented to the listeners who were instructed to identify the syllable in interval 3. This training comprised 15 runs with visual feedback. Each run was limited to a subset of the syllable database in order to gradually introduce the

stimuli and their orthography. Then followed 380 trials without distracters in which the target syllable was in either interval 2 or interval 3. In addition, the visual feedback was removed to demonstrate that the listeners could operate the GUI and that the complexity of the GUI was not the basis of subsequent errors. In a second training session, the distracters were progressively introduced with the SNR starting at 15 dB and then gradually decreased to -9 dB. In this session, the target syllable was in either interval 2 or interval 3. There were 10 runs of 48 trials in this session with visual feedback.

During training, performance criteria were used to ensure that the listener could perform the task at each SNR before proceeding to the next level. The pass marks were set based on pilot work that had shown that the performance of most listeners rose rapidly to more than 80% correct for unmasked syllables, and more than 30% correct when distracters were included at -9 dB SNR. Consequently, the criterion for the initial training session with targets but no distracters was 80%, and the criterion for the second training session with targets and distracters decreased from 70% to 30% as the SNR decreased from 15 to -9 dB. If a listener did not meet the criterion on a particular run, it was repeated until performance reached criterion. In total, the listeners did, at least, 1050 trials before commencing the main experiment, by which time they were very familiar with the GUI.

2. Main experiment—In the main experiment, recognition performance was measured as a function of GPR, VTL, and SNR for the CV syllables only. There were 56 different distracter voices (cf. Fig. 1 and Table I) and 3 SNRs (-6, 0, and 6 dB). The RSD between the target and distracter voices was varied over trials in a consistent fashion, from large to small and back to large. In this way, the distracter voice cycled through spoke points 7 to 1 and back to 7 so that the task became progressively harder and then easier in an alternating way. This was done to ensure that the listener was never subjected to a long sequence of difficult trials with small RSD values. The conditions at the ends of the RSD dimension were not repeated as the oscillation proceeded, so one complete cycle contained 12 trials. The main experimental variable—the combination of GPR and VTL in the distracter voice—was varied randomly without replacement from the eight values with the current RSD value (one on each spoke). The main experiment consisted of 120 runs of 48 trials (four cycles as explained above). Between runs, the SNR cycled through the three SNR values. The combination of RSD oscillation and controlled spoke randomization meant that when all runs had been completed, all of the RSD values, other than the endpoints, had been sampled 40 times at each SNR, and the endpoints had been sampled 20 times.

3. SNR extension—The final part of the study measured the psychometric function for six distracters at SNR values of -15, -9, 0, +9, and +15 dB using an up-down procedure. This was done in order to demonstrate that the effect of RSD decreases as SNR moves out of the range (-6 to +6 dB) used in the main experiment because performance becomes dominated by energetic masking in one way or another. Above +6 dB, the distracter is becoming ever less audible, and below -6 dB, the target is becoming ever less audible. At the same time, the SNR extension ensures that the main experiment was performed in the region where the paradigm was most sensitive to the interaction of GPR, VTL, and SNR.

Four of the distracters were the outermost points of spokes 1, 3, 5, and 7, marked with circles in Fig. 1. The remaining two distracters were the target voice itself and a noise masker. The noise maskers were created by extracting the temporal envelopes of distracters chosen in the usual way, and then filling the envelopes with speech-shaped broadband noise (Elberling *et al.*, 1989). Randomization was performed so that after 24 runs of 40 trials each, all midpoints in the SNR range had been measured 40 times for each distracter.

III. RESULTS

Two measures of performance were analyzed: (1) target syllable recognition rate, and (2) distracter intrusion rate, where the listener reported the distracter rather than the target syllable. The effects of vocal characteristics and audibility on these target and distracter scores in the main experiment were analyzed with a three-way repeated-measures analysis of variance (ANOVA) [$3(\text{SNRs}) \times 8(\text{spoke numbers}) \times 7(\text{spoke points})$]. In the SNR extension, the effects of vocal characteristics and audibility were analyzed with a two-way repeated-measures ANOVA for the four different voices [$5(\text{SNRs}) \times 4(\text{spoke numbers})$]. Paired comparisons with Sidak correction for multiple comparisons were used to analyze the effects of all six distracters (four different voices, one identical voice, and one noise masker). Greenhouse–Geisser correction for degrees of freedom was used to compensate for lack of sphericity, and partial eta squared values (η_p^2) are quoted below to report the effect sizes.

For target recognition in the main experiment, there were significant main effects of SNR ($F_{2,10}=201.4$, $p<0.001$, $\epsilon=0.64$, $\eta_p^2=0.98$) and spoke point ($F_{6,30}=123.2$, $p<0.001$, $\epsilon=0.47$, $\eta_p^2=0.96$), and an interaction between SNR and spoke point ($F_{12,60}=5.4$, $p=0.004$, $\epsilon=0.33$, $\eta_p^2=0.52$). There was no main effect of spoke number and no interaction between spoke number and either SNR or spoke point. For the distracter intrusions, there was a similar pattern of results: There were the same significant main effects of SNR ($F_{2,10}=50.2$, $p=0.001$, $\epsilon=0.54$, $\eta_p^2=0.91$) and spoke point ($F_{6,30}=168.1$, $p<0.001$, $\epsilon=0.47$, $\eta_p^2=0.96$), and the same interaction between SNR and spoke point ($F_{12,60}=26.2$, $p<0.001$, $\epsilon=0.19$, $\eta_p^2=0.84$). There was no main effect of spoke number and no interaction between spoke number and either SNR or spoke point. Notice that it is spoke *point* that indicates the radial distance between voices, whereas spoke *number* specifies the angular relationship between sets of voices in the VTL–GPR plane. For target recognition in the SNR extension, there was again a significant main effect of SNR ($F_{4,10}=114.5$, $p<0.001$, $\epsilon=0.53$, $\eta_p^2=0.96$). There was no main effect of distracter voice and no interaction between distracter voice and SNR. For distracter intrusions, there was a similar pattern of results: a significant main effect of SNR ($F_{4,10}=7.0$, $p=0.011$, $\epsilon=0.54$, $\eta_p^2=0.58$), no main effect of distracter voice, and no interaction between distracter voice and SNR.

A. Effects of vocal characteristics

The fact that there was no main effect of spoke number, and no interaction of spoke number with the other variables, for either of the performance measures, means that the best estimate of the main effect of RSD is provided by collapsing performance across spoke number. The results are shown in Fig. 4, where average performance is plotted as a function of spoke point. The target recognition scores are shown by dashed, solid, and dotted black lines in the upper part of the figure. Performance is best, as would be expected, for conditions where the target and distracter voices are maximally dissimilar (spoke point 7). The effect of vocal characteristics (i.e., spoke point) is greatest for 0 dB SNR, where the cumulative effect along the spoke-point function is 29%. For the +6 and –6 dB SNRs, the cumulative effect is somewhat smaller, 15% and 22%, respectively. The convergence of the 0 dB function with the –6 dB function as spoke point decreases between points 4 and 3 shows that listeners can use the level difference in the –6 dB condition to prevent performance falling as low as it might when there is little, or no, difference in vocal characteristics.

The distracter intrusion scores are shown by the dashed, solid, and dotted dark gray lines in the lower part of Fig. 4. Distracter intrusions occur most often, as would be expected, when the voices are similar (spoke point 1). When the voices are dissimilar (spoke points 5–7), distracter intrusions are very rare. The intrusion rate for similar voices (point 1) is greatest (34%) when the SNR is 0 dB, and least (12%) when the SNR is +6 dB. The fact that the

intrusion rate is higher at 0 dB SNR than it is at -6 dB SNR (22%) shows that listeners derive a consistent advantage from a loudness difference even when the distracter is louder than the target.

The sum of the target and distracter scores is presented in the upper part of Fig. 4, separately for each SNR, by thin, light gray lines, using the same line type. These lines exhibit no effect of vocal characteristics; the only difference between them is their vertical position, which reflects the effect of SNR. This means that for a given SNR, the listeners identified the syllables in a pair at a roughly constant level, but they were only successful at segregating the voices when there was a perceptible difference in vocal characteristics.

B. Effect of SNR

An extended view of the effect of SNR is presented in Fig. 5, which shows performance as a function of SNR from -15 to +15 dB, for speech distracters with different voices (lines with squares), a speech distracter with identical vocal characteristics (lines with triangles), and a noise masker (lines with diamonds). The speech distracters with different voices were the four speakers farthest from the target voice on spokes 1, 3, 5, and 7. The black lines in the upper part of the figure show average target recognition, which drops from between 78% and 88% at 15 dB SNR to between 17% and 34% at -15 dB SNR; the largest drop is seen for the noise masker. Performance with speech distracters that differ in vocal characteristics from the target is always better than performance when the distracter has the same vocal characteristics, even when the SNR is large. The gradient is steepest for the noise masker between 0 and -9 dB, and steepest for the identical voice between +9 and 0 dB. In these regions performance drops about 3%/dB.

With the exception of the case where the distracter was identical to the target, recognition performance for speech distracters did not vary with the spoke of the distracter speaker, that is, the angle of the spoke in the GPR-VTL plane. Paired comparisons showed that the target scores for all of the different voices were significantly higher than the target scores for the noise masker for the three lowest SNRs, and lower than the target scores for the identical voice for -9 and 0 dB SNRs. The target scores for the noise masker were significantly higher than the target scores for the identical voice for 0 and 9 dB SNRs.

The distracter scores are shown in the lower part of Fig. 5 with gray lines. Listeners made this type of error only if the target and distracter voices were identical. As the SNR decreased from 15 dB, the proportion of distracter intrusions increased to a maximum (33%) at 0 dB SNR where there is no loudness cue to distinguish the target from the distracter. The rate of distracter intrusions then decreased from 33% at 0 dB SNR to 20% at -15 dB SNR. Paired comparisons of the distracter scores for the identical voice showed that the maximal distracter intrusion rate at 0 dB SNR was significantly higher than the distracter intrusion rates at -15, 9, and 15 dB SNRs. These results confirm that the listeners were able to attend selectively to the target when the only difference between the target and distracter voices was loudness, even in adverse listening conditions where the distracter was much louder than the target.

IV. MODELING THE INTERACTION OF VOCAL CHARACTERISTICS AND AUDIBILITY

The fact that there is no main effect of spoke angle, and no interaction between spoke angle and either spoke point or SNR, suggests that it might be possible to produce a relatively simple model of the target recognition performance in this experiment. Such a model would provide an economical summary of the data and enable us to provide performance surfaces to illustrate how speaker differences interact with SNR differences in the performance space.

The model would also allow us to test the assumption made when generating the stimuli that a difference in VTL has to be about 1.5 times a difference in GPR in order to produce the same effect on recognition performance.

The psychometric function that relates target recognition to RSD (along a spoke) should be asymmetric because the RSD variable is limited to positive values, by definition, and the slope of the psychometric function should, therefore, be 0 for an RSD of 0. This means that the traditional cumulative Gaussian would not provide a good fit in this experiment. Accordingly, the psychometric function was modeled with a cumulative gamma function rather than a cumulative Gaussian. In this way, the probability of target recognition as a function of RSD was defined as

$$p(\text{RSD}) = \lambda + (\mu - \lambda) \int_0^{\text{RSD}} \Gamma(x|\alpha, \beta) dx, \quad (2)$$

where Γ is the gamma function and α and β are its shape and scale parameters. The gamma function rises from 0 to 1 over its range, so the function was offset and scaled on the p -axis by λ and μ , where λ is the intercept on the p -axis, and μ is the asymptotic recognition score for large RSD. The shape parameter, α , was restricted to be greater than 2 to ensure that the cumulative gamma function would have a gradient of zero at the p -axis intercept. The other parameters were limited only by their theoretical maximum range.

The psychometric function was fitted to the data using maximum likelihood estimation (MLE), in which the trading value between VTL and GPR, χ , was included as a free parameter whose optimum value was estimated in the process. Specifically, MLE was used to fit the cumulative gamma function to all of the individual sets of data, and between iterations, the value of χ was varied (that is, the relative lengths of the GPR and VTL dimensions were varied) to find the value of χ that was most likely to have produced the observed data. In one case, the procedure was applied separately to the three groups of eight data sets associated with each SNR (+6, 0, and -6 dB), and in a second case, the procedure was applied to the collective set of data from the three SNRs taken together. In this way, a trading value was derived for each SNR as well as an optimum trading value describing the trading relation for all SNRs. The psychometric functions for the collective fit are presented along with the data in Fig. 6; the optimum trading value for this collective fit was 1.9. Since the trading value is different from the value of 1.5 used to generate the stimuli, the points from different spokes occur at different RSD_{1.9} values. The trading values for the three SNRs, fitted individually, were 1.9, 1.6, and 3.2 for SNR values of 6, 0, and -6 dB, respectively.

In order to illustrate the effects of the trading relationship between GPR and VTL, the trading value from the collective fit was used to generate performance surfaces, which are shown in Fig. 7. The surfaces are sets of elliptical, equal-performance contours fitted to the values from the eight spokes associated with each RSD value, separately for the three SNRs (+6, 0, and -6 dB). The dotted lines in the GPR-VTL plane below the surfaces show the combinations of GPR and VTL that defined the distracters. The surfaces show the main effects of SNR, GPR, and VTL on performance reported above. The main effect of SNR is illustrated by the displacement of the surfaces around their outer edges, where the radial distance between the target and distracter is maximum; the average performance for the largest speaker difference was 77%, 64%, and 52% when the SNR was +6, 0, and -6 dB, respectively. The effect of speaker difference is illustrated by the indentation at the center of each surface. In each case, as the radial distance between the target and distracter voice decreases, recognition performance decreases, and in each case, the worst performance occurs when there is minimal difference between the target and distracter in terms of GPR and VTL. The indentation in the surface is deepest for the 0 dB condition where the

loudness cue is smallest. The average performance at the center of the surface drops 16% when the SNR is +6 dB, 29% when the SNR is 0 dB, and 20% when the SNR is -6 dB.

V. DISCUSSION

The main experiment showed that listeners take advantage of VTL differences as well as GPR differences when recognizing competing syllables. The effect is most notable when the SNR is 0 dB and there is no loudness cue to assist in tracking the target speaker. In noisy environments where extra vocal effort is required, normally hearing speakers tend to allow the level of their voice to fall to an SNR of around 0 dB (Lombard, 1911; Plomp, 1977). Hence, the 0 dB SNR condition of the current experiment represents a common condition in multispeaker environments, and the data show the value of the vocal characteristics in speaker segregation.

A. Trading relationship between VTL and GPR

In the current experiment, the trading values for the three SNRs were found to be 1.9, 1.6, and 3.2 for SNR values of 6, 0, and -6 dB, respectively. The values were determined by the minima of the cost functions for the MLE. However, the cost functions were highly asymmetric about their minima. They drop rapidly as χ increases from 0 toward their minimum, but beyond the minimum the cost functions rise slowly. The trough for the 0 dB cost function was the sharpest, which indicates that the trading value of 1.6 is the most reliable. The trough was least sharp for the -6 dB cost function, indicating that the trading value of 3.2 is the least reliable.

The fact that the smallest trading relationship is found for 0 dB SNR is consistent with the finding that the effects of vocal characteristics are largest when there are few other cues. Pitch is a very salient cue, so it may be that, as soon as there are both pitch and loudness cues, it requires a relatively large VTL difference to improve performance further. The trading value for 0 dB SNR, 1.6, means that a two-ST GPR difference provides the same performance advantage as a 20% difference in VTL.

Darwin *et al.* (2003) reported data on the interaction of VTL and F0 in their study on concurrent speech. At 0 dB SNR, for a 38% difference in VTL, the performance advantage was 20%, and for a four-ST difference in F0, the performance advantage was 20%.² This corresponds to a trading relationship of 1.4, just below that observed at 0 dB SNR in the current experiment. They also found that when the VTL and GPR values both shifted toward larger speakers, there was more benefit than when they both shifted toward smaller speakers. Given their results, one might have expected to see that the improvement in recognition performance along spoke 6 was greater than along spoke 2. Darwin *et al.* (2003) also reported that there was synergy between GPR and VTL; a result that could have led to the expectation that performance would increase less along spokes 1, 3, 5, and 7 (where there is mainly variation in one variable) than along spokes 2, 4, 6, and 8 (where there is comparable variation in both variables). However, we do not find any asymmetry in target recognition performance about the target speaker in the GPR-VTL plane. Our results show that there was no statistical effect of spoke number; the only effect of vocal characteristics was an effect of spoke point, i.e., the radial distance from the target speaker. At this point, it is not clear why Darwin *et al.* (2003) found asymmetries that we did not.

²Darwin *et al.* (2003) stated that a nine-ST F0 difference corresponded to the advantage derived from a 38% VTL difference. However, in their Fig. 1, it is apparent that a four-ST difference yielded a 20% performance advantage while a nine-ST difference yielded an advantage of approximately 25%.

Van Dinther and Patterson (2006) varied the equivalent of GPR and VTL in musical instrument sounds and measured the amount of change in the variables (pulse rate and resonance scale) required to discriminate the relative size of instruments from their sustained note sounds. The design allowed them to estimate the trading value between the variables, which was found to be 1.3 with these sounds. This suggests that when listeners compare sounds with different combinations of pulse rate and resonance rate, pulse rate has a somewhat larger effect in judgments of relative instrument size.

B. Determinants of distraction

The present experiments show that the vocal characteristics of a competing speaker have a large effect on the amount of distraction caused by that speaker. This was evident in Fig. 5, which showed the psychometric functions for six different distracters. At 0 dB SNR, the noise masker reduced performance to 58%; but when the distracter was the identical voice, performance dropped further to 40%. This phenomenon (a drop in recognition performance without a change in SNR) is sometimes referred to as “informational masking” (for a recent review, see Watson, 2005). The idea is that the degree of disturbance depends not only on the distracter's ability to limit audibility (energetic masking), but also on its ability to pull attention away from an otherwise audible or partially audible target sound. In other words, informational masking is a property separate from energetic masking—a property which differentiates the effects of equally intense distracters. Note, however, that as the vocal characteristics of the distracting voice become more and more different from those of the target voice in our study, performance rises above that achieved when the target voice is presented in broadband, envelope-matched noise. When the SNR is 0 dB, performance rises from 40% when the target and distracter are the same voice, to 58% when the distracter is a noise masker, and on up to 70% when the distracter is one of the different voices. Hence, the vocal characteristics of the distracter caused a drop in performance (relative to a noise masker of the same level) when the voices were similar, and an increase in performance when the voices were dissimilar.

The difference between masking and distraction is also evident in the distribution of errors, and the interaction of error type with SNR. When the voice of the distracter caused a drop in recognition performance, it was often because the listener reported the distracter syllable rather than the target syllable (the gray lines in Fig. 4). If these errors are scored as correct, performance is observed to be largely independent of speaker characteristics for a given SNR. In other words, the main effect of reducing the RSD between the target and distracter is that listeners are increasingly unsuccessful at segregating the syllable streams of the competing voices. Target-distracter confusions occurred most often for 0 dB SNR as illustrated in Fig. 5. In a recent study on the release of informational masking based on spatial separation of the competing sources, Ihlefeld and Shinn-Cunningham (2008) also reported that most target-distracter confusions occurred for 0 dB SNR.

For conditions where the SNR was not zero, listeners appear to derive some value from loudness cues, even when the SNR is negative. The depth of the indentation in the performance surface (Fig. 7) for data with an SNR of -6 dB is less than for an SNR of 0 dB. A different view of this effect is provided in Fig. 4 in which performance drops off most for an SNR of 0 dB as RSD decreases. The improvement in relative performance at negative values of SNR is similar to, although less pronounced than, the recovery phenomenon reported by Brungart (2001) at negative SNRs mentioned earlier. The data suggest that listeners can use loudness to reject the distracter and focus on the target even when it is softer than the distracter.

The depth of the indentation in the performance surface for data with an SNR of 0 dB confirms the importance of differences in vocal characteristics when there is no loudness

cue. It is in this condition that listeners are observed to derive the most advantage from differences in VTL, GPR, or any combination of the two. A similar result was reported by Ihlefeld and Shinn-Cunningham (2008). They showed that recognition performance was poorer for an SNR of 0 dB than it was for an SNR of -10 dB, and that the largest relative advantage provided by spatial separation occurred when the SNR was 0 dB.

VI. SUMMARY AND CONCLUSION

The experiments in this paper show how two speaker-specific properties of speech (GPR and VTL) assist a listener in segregating competing speech signals. In multispeaker environments, where there are substantial differences between speakers in GPR and VTL, the performance for a particular SNR depends critically on these speaker differences. When they are not available, target recognition is severely reduced as evidenced by the indentations in the performance surfaces in Fig. 7. The results also show that, when there is a large difference between the speaker-specific characteristics of the target and distracter voices, performance is primarily determined by SNR. As speaker-specific differences between the target and distracter are reduced, performance decreases from the level imposed by the SNR by as much as 30%.

There is a strong interaction between the effects of GPR and VTL that takes the form of a relatively simple tradeoff. When the two variables are measured in logarithmic units, and there are no loudness cues to assist in tracking the target speaker, then a change in VTL has to be about 1.6 times a change in GPR to have the same effect on performance. The results are consistent with the notion that audibility is the prime determinant of performance, and that GPR and VTL are particularly effective when there are no loudness cues. The study has also demonstrated that the auditory system can use any combination of these cues to segregate competing speech signals.

Acknowledgments

The research was supported by the UK Medical Research Council (Grant Nos. G0500221 and G9900369) and the European Office of Aerospace Research and Development (Grant No. FA8655-05-1-3043). The authors thank Joan Sussman and two anonymous reviewers for constructive comments on earlier versions of the manuscript.

References

- Assmann PF, Summerfield Q. Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies. *J. Acoust. Soc. Am.* 1990; 88:680–697. [PubMed: 2212292]
- Assmann PF, Summerfield Q. The contribution of waveform interactions to the perception of concurrent vowels. *J. Acoust. Soc. Am.* 1994; 95:471–484. [PubMed: 8120258]
- Bolia RS, Nelson WT, Ericson MA, Simpson BD. A speech corpus for multitalker communications research. *J. Acoust. Soc. Am.* 2000; 107:1065–1066. [PubMed: 10687719]
- Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. *J. Acoust. Soc. Am.* 2001; 109:1101–1109. [PubMed: 11303924]
- Brungart DS, Simpson BD, Ericson MA, Scott KR. Informational and energetic masking effects in the perception of multiple simultaneous talkers. *J. Acoust. Soc. Am.* 2001; 110:2527–2538. [PubMed: 11757942]
- Chalikia MH, Bregman AS. The perceptual segregation of simultaneous vowels with harmonic, shifted, or random components. *Percept. Psychophys.* 1993; 53:125–133. [PubMed: 8433910]
- Collins SA. Men's voices and women's choices. *Anim. Behav.* 2000; 60:773–780. [PubMed: 11124875]
- Cooke M. A glimpsing model of speech perception in noise. *J. Acoust. Soc. Am.* 2006; 119:1562–1573. [PubMed: 16583901]

- Culling JF, Darwin CJ. The role of timbre in the segregation of simultaneous voices with intersecting f_0 contours. *Percept. Psychophys.* 1993; 54:303–309. [PubMed: 8414889]
- Darwin CJ, Brungart DS, Simpson BD. Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *J. Acoust. Soc. Am.* 2003; 114:2913–2922. [PubMed: 14650025]
- de Cheveigné A. Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing. *J. Acoust. Soc. Am.* 1993; 93:3271–3290.
- de Cheveigné A. Concurrent vowel identification. III. A neural model of harmonic interference cancellation. *J. Acoust. Soc. Am.* 1997; 101:2857–2865.
- de Cheveigné A, McAdams S, Marin CMH. Concurrent vowel identification. II. Effects of phase, harmonicity and task. *J. Acoust. Soc. Am.* 1997a; 101:2848–2856.
- de Cheveigné A, Kawahara H, Tsuzaki M, Aikawa K. Concurrent vowel identification. I. Effects of relative amplitude and f_0 difference. *J. Acoust. Soc. Am.* 1997b; 101:2839–2847.
- Elberling C, Ludvigsen C, Lyregaard PE. Dantale: A new Danish speech material. *Scand. Audiol.* 1989; 18:169–176. [PubMed: 2814331]
- Fant, GCM. *Acoustic Theory of Speech Production*. Mouton; The Hague: 1970.
- Fitch WT, Giedd J. Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *J. Acoust. Soc. Am.* 1999; 106:1511–1122. [PubMed: 10489707]
- Ihlefeld A, Shinn-Cunningham BG. Spatial release from energetic and informational masking in a selective speech identification task. *J. Acoust. Soc. Am.* 2008; 123:4369–4379. [PubMed: 18537388]
- Irino T, Patterson RD. Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-mellin transform. *Speech Commun.* 2002; 36:181–203.
- Ives DT, Smith DR, Patterson RD. Discrimination of speaker size from syllable phrases. *J. Acoust. Soc. Am.* 2005; 118:3816–3822. [PubMed: 16419826]
- Kawahara, H.; Irino, T. Underlying principles of a high-quality speech manipulation system straight and its application to speech segregation. In: Divenyi, PL., editor. *Speech Separation by Humans and Machines*. Kluwer Academic; Boston, MA: 2004.
- Ladefoged P, Broadbent DE. Information conveyed by vowels. *J. Acoust. Soc. Am.* 1957; 29:98–104.
- Lee S, Potamianos A, Narayanan S. Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *J. Acoust. Soc. Am.* 1999; 105:1455–1468. [PubMed: 10089598]
- Lombard E. Le signe de l'élévation de la voix (The characteristics of the raised voice). *Ann. Mal. Oreil. Larynx Nez Pharynx.* 1911; 37:101–119.
- Marcus SM. Acoustic determinants of perceptual center (p-center) location. *Percept. Psychophys.* 1981; 30:247–256. [PubMed: 7322800]
- Miller GA, Licklider JCR. The intelligibility of interrupted speech. *J. Acoust. Soc. Am.* 1950; 22:167–173.
- Moore, TJ. Voice communications jamming research. AGARD Conference Proceedings 311: Aural Communication in Aviation; AGARD; Neuilly-Sur-Seine, France. 1981. p. 2:1-2:6.
- Nearey TM. Static, dynamic, and relational properties in vowel perception. *J. Acoust. Soc. Am.* 1989; 85:2088–2113. [PubMed: 2659638]
- Peterson GE, Barney HL. Control methods used in a study of the vowels. *J. Acoust. Soc. Am.* 1952; 24:175–184.
- Plomp R. Acoustical aspects of cocktail parties. *Acustica.* 1977; 38:186–191.
- Qin MK, Oxenham AJ. Effects of envelope-vocoder processing on f_0 discrimination and concurrent-vowel identification. *Ear Hear.* 2005; 26:451–460. [PubMed: 16230895]
- Ritsma, RJ.; Hoekstra, A. Frequency selectivity and the tonal residue. In: Zwicker, E.; Terhardt, E., editors. *Facts and Models in Hearing*. Springer; Berlin: 1974.
- Rivenez M, Darwin CJ, Bourgeon L, Guillaume A. Un-attended speech processing: Effect of vocal-tract length. *J. Acoust. Soc. Am.* 2007; 121:EL90–EL95. [PubMed: 17348552]

- Rivenez M, Darwin CJ, Guillaume A. Processing unattended speech. *J. Acoust. Soc. Am.* 2006; 119:4027–4040. [PubMed: 16838545]
- Smith DR, Patterson RD. The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age. *J. Acoust. Soc. Am.* 2005; 118:3177–3186. [PubMed: 16334696]
- Smith DR, Patterson RD, Turner R, Kawahara H, Irino T. The processing and perception of size information in speech sounds. *J. Acoust. Soc. Am.* 2005; 117:305–318. [PubMed: 15704423]
- Summerfield Q, Assmann PF. Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony. *J. Acoust. Soc. Am.* 1991; 89:1364–1377. [PubMed: 2030224]
- Titze IR. Physiologic and acoustic differences between male and female voices. *J. Acoust. Soc. Am.* 1989; 85:1699–1707. [PubMed: 2708686]
- Turner RE, Walters TC, Monaghan JJM. A statistical formant-pattern model for segregating vowel type and vocal tract length in developmental formant data. *J. Acoust. Soc. Am.* 2009 in press.
- van Dinther R, Patterson RD. The perception of size in musical instruments. *J. Acoust. Soc. Am.* 2006; 120:2158–2176. [PubMed: 17069313]
- von Kriegstein K, Warren JD, Ives DT, Patterson RD, Griffiths TD. Processing the acoustic effect of size in speech sounds. *Neuroimage.* 2006; 32:368–375. [PubMed: 16644240]
- Walters TC, Gomersall P, Turner RE, Patterson RD. Comparison of relative and absolute judgments of speaker size based on vowel sounds. *Proc. Meet. Acoust.* 2008; 1:050003.
- Watson CS. Some comments on informational masking. *Acta. Acust. Acust.* 2005; 91:502–512.

FIG. 1.

The specification of the distracter voices forms an elliptical spoke pattern in the GPR-VTL plane. The ordinate has been compressed by a factor of 1.5 to illustrate the relationship between GPR and VTL used in the experiment. It is this, that makes the ellipses appear circular. The pitch of the distracter voices varied from 137 to 215 Hz, and the VTL varied from 11 to 21 cm. The target voice is in the center of the spoke pattern with a pitch of 172 Hz and a VTL of 15 cm. In the text, points in the plane are described in terms of their radial displacement from the target (a vector distance). For convenience, they are arbitrarily numbered 1–7 where 1 is closest to, and 7 is farthest away from, the reference voice in the center. The $RSD_{1.5}$ values for spoke points 3, 5, and 7 are shown; see Table I for the complete specification. Some of the distracter voices were used in the extended SNR experiment; they are marked with circles. The gray areas show the region of the plane occupied by 95% of male and female speakers in the population reported by Peterson and Barney (1952), and modeled by Turner *et al.* (2009).

FIG. 2.

Examples of temporal-envelope matching pairs of sonorant-vowel syllables (left column), plosive-vowel syllables (central column), and fricative-vowel syllables (right column). The syllables are p -centered to effect the optimum match for arbitrary syllable pairings within a CV group. See the text for details.

	b	d	f	g	h	k	l	m	n	p	r	s	sh	t	v	w	y	z
a	ba	da	fa	ga	ha	ka	la	ma	na	pa	ra	sa	sha	ta	va	wa	ya	za
e	be	de	fe	ge	he	ke	le	me	ne	pe	re	se	she	te	ve	we	ye	ze
i	bi	di	fi	gi	hi	ki	li	mi	ni	pi	ri	si	shi	ti	vi	wi	yi	zi
o	bo	do	fo	go	ho	ko	lo	mo	no	po	ro	so	sho	to	vo	wo	yo	zo
u	bu	du	fu	gu	hu	ku	lu	mu	nu	pu	ru	su	shu	tu	vu	wu	yu	zu
a	ab	ad	af	ag	ah	ak	al	am	an	ap	ar	as	ash	at	av	aw	ay	az
e	eb	ed	ef	eg	eh	ek	el	em	en	ep	er	es	esh	et	ev	ew	ey	ez
i	ib	id	if	ig	ih	ik	il	im	in	ip	ir	is	ish	it	iv	iw	iy	iz
o	ob	od	of	og	oh	ok	ol	om	on	op	or	os	osh	ot	ov	ow	oy	oz
u	ub	ud	uf	ug	uh	uk	ul	um	un	up	ur	us	ush	ut	uv	uw	uy	uz

1

2

3

FIG. 3.

Schematic illustration of the GUI. The response area, in which the listeners indicate their answer by a click with a computer mouse, is shown at the top. Underneath are shown the three visual interval indicators that light up synchronous with the three syllable intervals of each trial. In this example, interval 3 stays lit to indicate that the listeners should respond to the target syllable that was played in interval 3.

FIG. 4.

Recognition scores as a function of spoke point (see Fig. 1). The black lines at the top show target recognition scores, and the dark gray lines in the lower part of the figure show distracter intrusion scores. The sums of the target and distracter scores for each SNR are shown in the lightest gray lines at the top of the figure. They are largely independent of spoke point.

FIG. 5. Psychometric functions for the speech distracters and the noise masker. The black lines at the top show target recognition scores, and the dark gray lines in the lower part of the figure show distracter intrusion scores. Speech distracters for the different voices are shown with squares, the speech distracter with identical vocal characteristics is shown with triangles, and the noise masker is shown with diamonds.

FIG. 6.

Cumulative gamma functions fitted by MLE to the behavioral data associated with the eight stimulus spokes at each of the three SNR values are shown in thick black lines. The GPR-VTL trading parameter was allowed to vary along with the parameters describing the gamma functions. The optimized trading value was 1.9. Performance is plotted as a function of the RSD ($RSD_{1.9}$) in logarithmic units. The mean data for the individual spokes are shown by the gray lines. The dashed lines are for an SNR of 6 dB; the solid lines are for 0 dB SNR and the dotted lines are for -6 dB SNR.

FIG. 7.

Performance surfaces for -6, 0, and 6 dB SNRs from below to above, respectively. The tip of the middle surface for 0 dB SNR extends down almost to the tip of the surface for -6 dB below it. The vocal specifications of the distracters are indicated in the GPR-VTL plane below the surfaces. The surfaces were modeled by MLE of the psychometric functions along the spokes and by interpolating between the spokes; see the text for details.

TABLE I

Vocal characteristics of the distracter voice for the spoke numbers shown in Fig. 1. Spoke point refers to the position on the spokes ascending outwards from the reference voice in the center. The radial scale displacement ($RSD_{1.5}$) is shown for each spoke point; see text for details.

Spoke point		1	2	3	4	5	6	7
Spoke No.	RSD _{1.5}	0.01	0.03	0.06	0.11	0.18	0.25	0.35
1	GPR (Hz)	170.9	168.6	164.7	159.5	153.0	145.5	137.0
	VTL (cm)	14.7	14.8	14.9	15.1	15.3	15.5	15.8
2	GPR (Hz)	171.3	170.0	167.8	164.9	161.1	156.7	151.6
	VTL (cm)	14.8	15.0	15.5	16.2	17.0	18.2	19.7
3	GPR (Hz)	171.9	172.4	173.3	174.5	176.1	178.1	180.4
	VTL (cm)	14.8	15.1	15.6	16.4	17.5	18.8	20.6
4	GPR (Hz)	172.4	174.5	178.0	183.0	189.6	198.1	208.6
	VTL (cm)	14.7	14.9	15.2	15.6	16.2	16.8	17.7
5	GPR (Hz)	172.5	174.9	179.0	184.8	192.7	202.7	215.2
	VTL (cm)	14.7	14.6	14.5	14.3	14.1	13.9	13.6
6	GPR (Hz)	172.1	173.5	175.7	178.8	183.0	188.1	194.5
	VTL (cm)	14.6	14.3	13.9	13.4	12.7	11.9	11.0
7	GPR (Hz)	171.5	171.0	170.1	168.9	167.4	165.6	163.4
	VTL (cm)	14.6	14.3	13.8	13.2	12.4	11.5	10.5
8	GPR (Hz)	171.0	169.0	165.7	161.1	155.5	148.8	141.3
	VTL (cm)	14.6	14.5	14.2	13.8	13.4	12.8	12.2