Location and acoustic scale cues in concurrent speech recognition[a)]

D. Timothy Ives and 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 and Laboratoire de Psychologie de la Perception, CNRS, Universite Paris Descartes, DEC, Ecole Normale Supérieure, 29 rue d'Ulm, 75005 Paris, France

Doris J. Kistler
Department of Psychological and Brain Sciences, University of Louisville, Louisville, Kentucky 40292

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

(Received 25 June 2009; revised 25 January 2010; accepted 9 March 2010)

Location and acoustic scale cues have both been shown to have an effect on the recognition of speech in multi-speaker environments. This study examines the interaction of these variables. Subjects were presented with concurrent triplets of syllables from a target voice and a distracting voice, and asked to recognize a specific target syllable. The task was made more or less difficult by changing (a) the location of the distracting speaker, (b) the scale difference between the two speakers, and/or (c) the relative level of the two speakers. Scale differences were produced by changing the vocal tract length and glottal pulse rate during syllable synthesis: 32 acoustic scale differences were used. Location cues were produced by convolving head-related transfer functions with the stimulus. The angle between the target speaker and the distracter was 0°, 4°, 8°, 16°, or 32° on the 0° horizontal plane. The relative level of the target to the distracter was 0 or −6 dB. The results show that location and scale difference interact, and the interaction is greatest when one of these cues is small. Increasing either the acoustic scale or the angle between target and distracter speakers quickly elevates performance to ceiling levels. © *2010 Acoustical Society of America.*
[DOI: 10.1121/1.3377051]

PACS number(s): 43.71.Bp, 43.71.An, 43.66.Ba [JES] Pages: 3729–3737

I. INTRODUCTION

Recently, Vestergaard *et al.* (2009) have shown how vocal tract length (VTL) and glottal pulse rate (GPR) (i.e., the rate at which the vocal folds vibrate in voiced speech) interact in concurrent speech recognition. They measured the recognition of a single target speaker in the presence of a distracting speaker while systematically varying the VTL and GPR of the distracter. Not surprisingly, they found that recognition of the target improves as the VTL and/or the GPR of the distracter become progressively more different from those of the target. More importantly, they showed how the VTL and GPR dimensions could be equated, over a large range of values, to produce many different speakers, all of whom caused equal levels of distraction. Specifically, they demonstrated that there is a simple trading relationship between the logarithm of GPR ratio and the logarithm of VTL ratio with a value of around 1.6; that is, a two-semitone (or 12%) difference in the GPR of the distracter produced a similar increase in target recognition to a 20% difference in the VTL of the distracter. In terms of just noticeable differences (JNDs), this equates to about 6 JNDs for GPR and about 4 JNDs for VTL. This trading relationship holds over a wide range of values, which suggests that the internal representation of these sounds maintains independent dimensions for GPR and VTL. These dimensions of acoustic scale may be processed separately to normalize vowels produced with any combination of GPR and VTL. Irino and Patterson (2002) have demonstrated how such normalization might be achieved, and they argue that the by-products of the processing could be used as tracking variables for perceptual judgments about sources. Candidate brain regions for scale processing have been identified by von Kriegstein *et al.* (2007) in bilateral superior temporal gyrus (STG) for general sounds, and left posterior STG for speech sounds. Vestergaard *et al.* (2009) hypothesized that it is VTL and GPR normalization that make human speech recognition so robust to variation in speaker characteristics (e.g., Smith *et al.*, 2005; Ives *et al.*, 2005).

The normalization mechanisms proposed by Irino and Patterson (2002) operate like transforms on the representation of sound observed in the auditory nerve; they are assumed to be applied to all incoming sounds at an early stage in auditory processing. The processes that extract spatial cues from binaural sounds are also applied to all incoming

[a)]Portions of this work were presented in "The interaction of location with acoustic scale in concurrent speech recognition," at the 157th Meeting of the Acoustical Society of America, Portland, OR.

sounds at an early stage in auditory processing, and it has long been known that spatial cues improve speech recognition in multi-speaker environments. For example, Licklider (1948) showed that phase differences between a speech target and a noise masker could be used to improve the intelligibility of the speech. Similarly, Hirsh (1950) and Cherry (1953) both showed that the spatial separation of a target speaker from other masking speakers improved recognition of the target speaker. The increase in intelligibility is normally attributed to either the "better-ear" advantage, i.e., the fact that there is a greater signal to noise ratio (SNR) at one of the ears (Shaw *et al.*, 1947; Hawley *et al.*, 1999), or to binaural unmasking, i.e., the decorrelation of noise and a target signal using phase differences between the sounds at the two ears (Hirsh, 1948; Durlach, 1963).

Darwin and Hukin (2000) looked at the interaction of speaker characteristics and location cues. They showed that VTL information in speech can override the effect of interaural time difference (ITD) for concurrent sentences. In their study, listeners heard two simultaneous sentences and had to decide which of two simultaneous target words came from the attended sentence. Target word parameters such as ITD and VTL were changed compared to those of the carrier sentence. Darwin and Hukin showed that an ITD of 181 μs applied to the target word could be overridden by a VTL difference of 15%. Darwin and Hukin also showed that the prosody of the sentence could be used to selectively attend to a particular talker. Their study did not, however, investigate how GPR (as a component of acoustic scale) interacts with VTL and location cues.

In this study, the work of Vestergaard *et al.* (2009) is extended to include the spatial dimension; specifically, the location of the distracting voice is varied in combination with the vocal characteristics to demonstrate that any combination of differences in these three dimensions can be used to reduce the effect of the distracter, and to estimate the relative importance of the location cue with respect to the vocal cues.

II. METHOD

Listeners were required to identify syllables spoken by one voice (the target) presented concurrently with syllables from a second voice (the distracter). Recognition performance was measured as a function of three parameters of the distracter voice: namely, the acoustic scale of the distracter (as specified by the particular combination of VTL and GPR that defined the speaker), the location of the distracter, and the level of the distracter (relative to the target). On each trial of the experiment, the listener was presented with three stimulus intervals. A syllable from the target voice was present in each of the three intervals and a syllable from the distracting voice was present in the second and third intervals. The syllable from the target voice in the first interval was intended to cue the listener to the vocal characteristics of the target voice. The vocal characteristics of the target and distracter voices were held constant within a trial. The listener was required to identify the syllable spoken by the target voice in either the second or the third interval. The interval that the listener was required to respond to was chosen at random, and the listener was advised of the selection after hearing all three intervals.

A. Stimuli

The syllables were natural speech taken from the database of Ives *et al.* (2005). The database consisted of 180 unique syllables which were divided into 6 groups: three consonant-vowel (CV) groups and three vowel-consonant (VC) groups. Within the CV and VC categories, the groups were distinguished by consonant category: sonorants (son), stops (stp), and fricatives (fri). Thus, the six groups were CV sonorants (CVson), CV stops (CVstp), CV fricatives (CVfri), VC sonorants (VCson), VC stops (VCstp), and VC fricatives (VCfri). Each group contained 30 syllables generated by pairing five vowels with six consonants. The syllables had their perceptual centers (P-centers) aligned by inserting silence before and after the speech signal. As a result, when any combination of the syllables was played in a sequence, they would be perceived to proceed at a regular pace; irregular sequences produce an unwanted distraction. The P-center for each syllable was determined using procedures described by Marcus (1981) and Scott (1993); the specific implementation of the P-center correction is detailed in Ives *et al.* (2005). The total length of each syllable including the silence was 683, ms and there was no additional silence inserted between the intervals of a trial (i.e., the duration of each trial was 3×683 ms$=2.049$ s).

The pairs of target and distracter syllables presented in the second and third intervals of a trial were always selected from the same syllable group; that is, they had the same order of consonant and vowel (CV or VC), and the consonants came from the same category (sonorant, stop, or fricative). This procedure ensures that the two syllables have similar temporal envelopes, which, in turn, minimizes temporal glimpsing (Cooke, 2006), as described in Vestergaard *et al.* (2009). The distracter syllable was further constrained to have a consonant and a vowel that were different from those of the target syllable in the same interval. These constraints leave 20 potential distracter syllables for any given target syllable. The result is a target voice and a distracter voice, both of which are represented by a sequence of syllables, which both have considerable variability in terms of their acoustic properties (e.g., GPR, vowel formant frequencies, and consonant spectrum), and which are distinguished primarily by their fixed VTL. It is argued (Ives *et al.*, 2005) that VTL is extracted simultaneously with vowel type, and thus, at a stage of auditory processing beyond that where resolved harmonics or individual formant frequencies might be extracted.

The presentation level of the target speech was 60 dB throughout the experiment. The level of the distracter speech was adjusted relative to the target speech according to the experimental condition. Two SNRs were used (0 and -6 dB), giving a distracter level of either 54 or 60 dB.

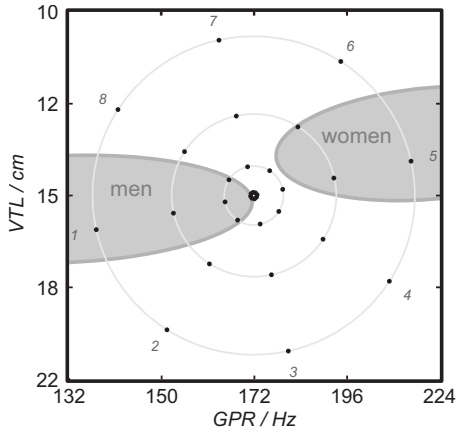

FIG. 1. The distribution of the 33 distracter voices on the VTL/GPR plane. Eight spokes radiate out from a central target voice (not shown for clarity). The spokes are numbered one through eight in a counter clockwise manner. Each spoke contains four points numbered from point 1 nearest the center to point 4 at the outer end. The central target voice has a VTL of 147 mm and a GPR of 171.7 Hz. An additional distracter voice is located at the same position as the target voice. The gray shaded areas show the range of VTL and GPR values, which encompass 95% of male and female speakers in the population. This was modeled by Turner *et al.* (2009), based on measurements from Peterson and Barney (1952)

1. Vocal characteristics

The GPR and VTL of a speaker largely determine the perceived size and sex of the speaker (Smith and Patterson, 2005; Walters *et al.*, 2008). All of the syllables in the database were analyzed and resynthesized with the vocoder STRAIGHT (Kawahara and Irino, 2004) to produce a complete set of the syllables for a target voice and 32 distracter voices with different combinations of VTL and GPR. Figure 1 shows the combinations of GPR and VTL that defined the 33 voices; they are arranged in an elliptical spoke pattern radiating out from the target voice in the center. The range of VTL values is from 10.5 to 20.6 cm, and the range of GPR values is from 137 to 213 Hz; the complete set of 33 voices is specified in Table I. Taller people like adult men tend to have longer vocal tracts and speak with lower GPRs than women and children (Fitch and Giedd, 1999; Peterson and Barney, 1952), so manipulating the vocal characteristics of the voices creates an effective simulation of speakers of different sex and size. The ellipses underneath the spoke pattern show the distribution of VTL and GPR combinations for male and female speakers in the normal population as reported by Peterson and Barney (1952), and modeled by Turner *et al.* (2009). The distracter voices are arranged on eight spokes radiating out from the target voice at the center of the spoke pattern; there are four speakers on each spoke. The target voice had a VTL of 147 mm and a GPR of 171.7 Hz. These values are the geometric means of the average GPRs and VTLs of men and women, respectively [see Vestergaard *et al.* (2009) for further details]. The VTL dimension is proportionally longer than the GPR dimension because the JND for VTL (Ives *et al.*, 2005) is more than 1.5 times the JND for GPR (Smith *et al.*, 2005). The configuration of voices in the GPR-VTL plane was originally devised by Vestergaard *et al.* (2009), who showed that the voices at a given radial distance from the target voice in the center produce the same amount of disruption of target syllable identification. This shows that there is a trading relationship between VTL and GPR in the perceptual separation between the target voice and any distracter voice. The perceptual distance between voices can be expressed by the radial scale displacement (RSD) between their points in the log(GPR)-log(VTL) plane. The RSD is the geometrical distance between the target and distracter voices

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

where X is log(GPR), Y is log(VTL), and χ is the GPR-VTL trading value, which is 1.5 in this experiment. The RSD values shown in Fig. 1 are for χ=1.5.

TABLE I. The set of VTL and GPR values of the voices used in the main experiment. There are four points on each of eight spokes, which radiate out from the target voice at the center point. The target voice (which is also an additional distracter voice) is located at the center.

Spoke	Point	VTL (cm)	GPR (Hz)
1	1	14.7	170.9
1	2	14.9	164.7
1	3	15.3	153.0
1	4	15.8	137.0
2	1	14.8	171.3
2	2	15.5	167.8
2	3	17.0	161.1
2	4	19.7	151.6
3	1	14.8	171.9
3	2	15.6	173.3
3	3	17.5	176.1
3	4	20.6	180.4
4	1	14.7	172.4
4	2	15.2	178.0
4	3	16.2	189.6
4	4	17.7	208.6
5	1	14.7	172.5
5	2	14.5	179.0
5	3	14.1	192.7
5	4	13.6	215.2
6	1	14.6	172.1
6	2	13.9	175.7
6	3	12.7	183.0
6	4	11.0	194.5
7	1	14.6	171.5
7	2	13.8	170.1
7	3	12.4	167.4
7	4	10.5	163.4
8	1	14.6	171.0
8	2	14.2	165.7
8	3	13.4	155.5
8	4	12.2	141.3
	Center	14.7	171.7

2. Location information

The speech stimuli were convolved with head related transfer functions (HRTFs) to simulate location information (Wightman and Kistler, 1989a, 2005). HRTFs contain information about how an incoming sound wave is affected by the head and pinna as a function of its angle, relative to the

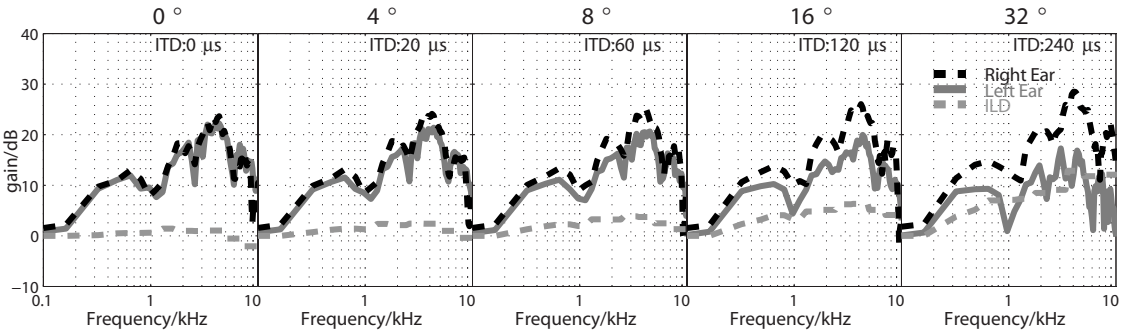

FIG. 2. The set of five HRTFs. The location of the target speaker is kept fixed at 0° azimuth and 0° elevation. The distracting speaker is located at one of five positions (0°, 4°, 8°, 16°, or 32° azimuth, and 0° elevation). The transfer functions are shown for right and left ears (dashed black and solid gray lines, respectively) for each angle. The dashed gray line shows the ILD between the right and left ears (the distracting speaker is moved to the right of the listener so a positive value of the ILD represents a larger value at the right ear relative to the left ear). The ITD is shown for each distracter angle at the top of each subplot.

listener's head. This makes it possible to simulate the spatial cues from a sound source and recreate a sound field as if the stimulus were produced at the simulated location (Wightman and Kistler, 1989b). The original HRTFs were measured with a miniature microphone placed in an earplug in the occluded ear canal (Wightman and Kistler, 2005). The HRTF set consisted of measurements from 613 positions measured in an anechoic chamber. Measurements were taken at 10° intervals for both azimuth (ranging from +180° to 170°) and elevation (ranging from +80° to 80°). In this study, we used a small subset of the original HRTFs to simulate the five angles of 0°, 4°, 8°, 16°, and 32° azimuth (to the right hand side of the listener), all on the horizontal plane (0° elevation). For angles which are not multiples of 10° (4°, 8°, 16°, and 32°), the HRTFs are calculated by interpolating from the nearest two locations. The interpolation was performed in the time-domain on the minimum phase impulse response (Wightman and Kistler, 1999). The HRTFs were recorded from a listener who did not participate in the current experiment. Wenzel *et al.* (1993) and Møller *et al.* (1996) have both shown that non-individualized HRTFs compared well with individualized HRTFs when the locations are restricted to the horizontal plane (0° elevation) and to the front of the listener.

The distracter angles were limited to a relatively small range (0°–32°), close to that of the target angle (0°), to balance the effectiveness of location with that of the acoustic scale cues. In anechoic conditions, segregation by perceived location is trivial for large separation angles.

Figure 2 shows the frequency responses of the five HRTFs with each panel showing the response for a particular angle (0°, 4°, 8°, 16°, or 32°). The gain applied to the signal is shown on the ordinate; the abscissa shows frequency from 0.1 to 10 kHz. The dashed black lines show the responses at the right ear and the solid gray lines, the left ear. The dashed gray lines show the interaural level differences (ILDs) between the right and left ears. The ILD is shown for ten discrete bands over the frequency range. On each panel the ITD is shown at the top; it is expressed as the time lag of the sound arriving at the left ear relative to the right ear. The graphs show that as a sound moves off center to the right, the overall level increases at the right ear, and the increase is greater at higher frequencies.

B. Procedure

Listeners were required to identify the syllables spoken by a target voice, when presented with concurrent syllables spoken by a distracter voice. On each trial, three intervals of speech were presented; a syllable from the target speaker was present in all three intervals, and syllables from a distracting speaker were present in the second and third intervals. The interval selected for identification of the target was randomly selected from the last two intervals, and the listener was advised of the selection after hearing all three intervals. The reason for varying the target interval was to preclude the listener attending to only one of the latter two intervals. The listeners chose their responses from a syllable matrix like that shown in Fig. 3, using a computer mouse.

1. Training

Prior to the main experiment, listeners undertook extensive training, to teach them the orthography of the syllable set, and to familiarize them with the response matrix. The orthography of vowels is ambiguous in English and so the listeners have to learn that the pronunciation was /a/ as in "bar," /e/ as in "bay," /i/ as in "bee," /o/ as in "toe," and /u/ as in "zoo." The listeners respond by clicking on the graphical representation of the syllable which corresponds to the acoustic signal they heard. At the start of training, the task was made easy by restricting the response to a small subset of the syllables in the matrix, then, gradually over blocks the size of the response set was increased until the whole database of 180 syllables was included. The first part of the training (*Training 1*) consisted of 15 runs of 10–20 trials with visual feedback. Each trial had three intervals as in the main experiment with syllables from the target voice on its own, i.e., the distracter was absent. The syllables were presented at 0°, directly in front of the listener. The task of the listener was to identify the syllable in the third interval. Listeners progressed onto subsequent runs of Training 1 once they reached a criterion level of performance; the performance criteria decreased from 80% to 70% as set size increased. Following Training 1, listeners undertook a baseline test with 18 runs each containing 20 trials without visual feedback.

The second part of the training (*Training 2*) consisted of

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

FIG. 3. The response matrix containing 180 syllables. The listeners use a similar matrix onto which they respond using a computer mouse. The syllables are divided into six groups: three CV groups and three VC groups. Within the CV and VC categories, the groups were distinguished by consonant category: sonorants (m, n, l, r, w, and y), stops (b, d, g, p, t, and k), and fricatives (s, f, v, z, sh, and h). Thus, the six groups were CV sonorants, CV stops, CV fricatives, VC sonorants, VC stops, and VC fricatives. Each group contained 30 syllables generated by pairing five vowels (a, e, i, o, and u) with six consonants. The uppermost row and the leftmost column are for information only and are not response options to the listener. The leftmost column also contains "cue tokens" for each vowel (e.g., "fa" and "la" for vowel "a").

18 runs of 20 trials. The purpose of Training 2 was to gradually introduce the distracter to the listener. This was achieved by initially setting the three parameters of the distracter (namely, vocal specification, location, and level) such that minimal distraction would occur and the target would be readily identified. As the listener progressed through the 18 runs of Training 2, the vocal characteristic difference, location difference, and level difference between the target and distracter were reduced, with the result that target identification became progressively more difficult. Listeners were required to identify target syllables from either interval two or interval three, and no visual feedback was given. Listeners were permitted a maximum of three attempts to achieve a criterion level of performance and progress to the next run. The level of performance varied from 70% for conditions in which the distracter differences were greatest, down to 40% in the most difficult conditions. Nine listeners undertook the training from which eight progressed successfully to the main experiment.

2. Experiment

Upon successful completion of the training, listeners progressed to the main experiment. The purpose of the main experiment was to measure recognition performance for target syllables in the presence of distracting syllables, as a function of three parameters of the distracter, namely, its vocal characteristics, location, and level.

For practical reasons, the experiment was split into two halves with the odd- and even-numbered spokes presented in different halves. Three listeners completed the experiment with the odd-numbered spokes, three completed the experiment with the even-numbered spokes, and two listeners completed both experiments. For the two listeners who completed both the odd and even spokes, one completed the odd-spokes version first and the other completed the even-spokes version first. In both halves of the experiment, recognition performance was measured for seventeen distracter voices at five locations; the 17 voices were the four points on all four spokes (either odd- or even-numbered) together with the reference voice in the center. The target voice was always the reference voice, and its location of 0° remained fixed throughout the experiment. The difference in vocal characteristics between the target and distracter voices was varied in a structured manner to prevent listeners from having a sustained period of difficult trials. The difference went from large to small and back to large in an alternating way (Vestergaard *et al.*, 2009). Each run consisted of 32 trials at each of the five locations, for a total of 160 trials. The 32 trials comprised four presentations of the reference voice (4), two presentations each of points 1, 2, and 3 on each of the spokes ($2\times3\times4=24$), and one presentation of point 4 on each of the spokes ($1\times4=4$). Each run took about 20 min and, together with breaks, listeners typically performed four runs in a 2-h session. The main experiment required five sessions. Within a run, the SNR remained constant; it was 0 dB for ten runs and −6 dB for ten runs.

C. Listeners

Eight listeners (four male) were paid an hourly wage to participate in the experiment. They were all students at the University of Cambridge, between 20 and 23 years of age. They had normal hearing (i.e., thresholds within 15 dB of audiometric threshold at 0.5, 1, 2, and 4 kHz). The experimental protocol was approved by the Cambridge Psychology Research Ethics Committee.

III. RESULTS

The performance measure was target recognition rate. The values were collapsed over spoke number as an analysis of variance (ANOVA) for each listener showed there was no significant effect of spoke number for any of the listeners. This was anticipated from the findings of Vestergaard *et al.* (2009), which showed that the functional advantage of a change in vocal specification was the same in all directions about the target voice, once the relative strength of the GPR and VTL dimensions have been balanced. The results for distracter location in the current experiment showed that per-

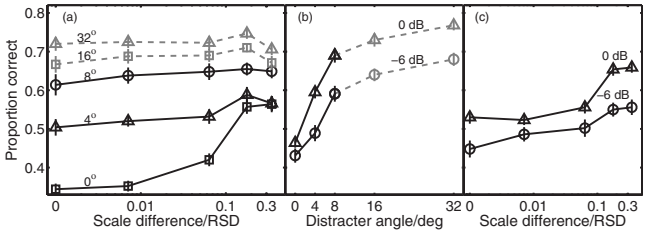

FIG. 4. The three interactions of (a) acoustic scale difference × distracter angle, (b) distracter angle × SNR, and (c) acoustic scale difference × SNR (collapsed over the three distracter angles of 0°, 4°, and 8°). In all three panels the solid lines show data that were included in the ANOVA and dashed lines show data that were excluded from the ANOVA.

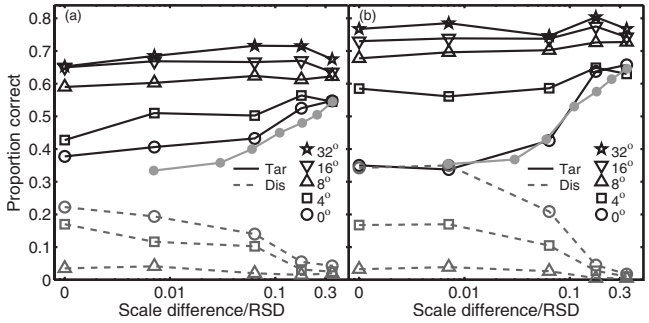

FIG. 5. The combined effect of distracter angle, acoustic scale difference, and SNR on target recognition: (a) SNR is −6 dB and (b) SNR is 0 dB. Target recognition scores are shown by the solid black lines, and distracter confusions are shown by dashed gray lines. The distracter confusions for 16° and 32° have been omitted for clarity; the actual results for these conditions were close to 0%. The solid gray lines with solid circle markers show results from Vestergaard *et al.* (2009); these compare best with the 0° conditions in the current study.

formance rose to ceiling levels as distracter angle increased to 16° (see Fig. 4). In order to increase the statistical power of the ANOVA with regard to the interaction of the vocal characteristics with location and level, the ceiling conditions, 16° and 32°, were excluded from the analysis. Accordingly, the effects of the distracter on target recognition rate were analyzed with a three-way repeated measures ANOVA [3(distracter angles) × 5(acoustic scale differences) × 2(SNRs)]. The Greenhouse–Geisser method was used to correct the degrees of freedom associated with a reduction in sphericity. The ANOVA shows there are three significant main effects and four interactions, all of which are listed in Table II. Partial eta squared (η_p^2) values are included to show the relative sizes of the effects. All three of the main effects are highly significant with η_p^2 values in excess of 0.81; however, it is the statistical interactions that describe the interaction of the main variables, so it is the statistical interactions that are the focus of the analysis.

A. Interaction of factors

All three of the two-way interactions were significant: acoustic scale difference × distracter angle, distracter angle × SNR, and acoustic scale difference × SNR. The three-way interaction was also significant, that is, acoustic scale difference × distracter angle × SNR. The acoustic scale difference is measured in units of RSD using a chi value of 1.5 (see Sec. II A 1). Figure 4(a) shows the interaction of acoustic scale difference with distracter angle; the effect of scale difference is large for small distracter angles, but the effect diminishes rapidly with increasing spatial separation. Chance performance in these experiments was 0.6% correct (1/180 syllables). Figure 4(a) shows all five distracter-angle locations even though only three of the angles (0°, 4°, and 8°) were included in the ANOVA. The data for 16° and 32° are shown as dashed gray lines to distinguish them from the data for 0°, 4°, and 8° shown by the solid black lines. Figure 4(b) shows the interaction of distracter angle and SNR; the advantage obtained from spatial separation is greater for 0-dB SNR than for −6-dB SNR. Again the results for distracter angles 16° and 32° are shown as dashed gray lines. Figure 4(c) shows the interaction of scale difference and SNR collapsed over the three distracter angles of 0°, 4°, and 8°. This interaction appears to be somewhat more complicated. The improvement in recognition rate from an SNR of −6 to 0 dB is greater for the large scale differences (~0.2 and ~0.3) than for two of the smaller scale differences (~0.01 and ~0.1); however, the improvement at RSD=0 is also greater than it is for the intermediate values of scale difference (~0.01 and ~0.1).

The three-way interaction of distracter angle, acoustic scale difference, and SNR is shown in Figs. 5(a) and 5(b) for SNRs of −6 and 0 dB, respectively. In Figs. 5(a) and 5(b), the solid black lines show the proportion of trials in which the target was identified correctly; five solid black lines are shown, one for each distracter angle. The solid gray lines show the results from the study of Vestergaard *et al.* (2009). They measured performance for seven different values of acoustic scale difference, and the results are comparable to the 0° conditions for both −6 (Fig. 5(a)) and 0 dB (Fig. 5(b)). The dashed gray lines show the target-distracter intrusions, i.e., the proportion of trials in which the listeners mistakenly reported the distracter syllable instead of the target syllable; three dashed gray lines are shown, one each for distracter angles of 0°, 4°, and 8°. The distracter confusions for 16° and 32° have been omitted for clarity; the actual results for these conditions were close to 0%.

TABLE II. The significant main effects and interactions of factors for target recognition.

Distracter angle	$F_{(2,14)}=122.7$, $p<0.001$, $\varepsilon=0.71$, $\eta_p^2=0.95$
Acoustic scale difference	$F_{(4,28)}=31.7$, $p<0.001$, $\varepsilon=0.58$, $\eta_p^2=0.82$
SNR	$F_{(1,7)}=29.6$, $p<0.001$, $\varepsilon=1$, $\eta_p^2=0.81$
Acoustic scale difference × distracter angle	$F_{(8,56)}=7.9$, $p=0.003$, $\varepsilon=0.30$, $\eta_p^2=0.53$
Distracter angle × SNR	$F_{(2,14)}=11.9$, $p=0.005$, $\varepsilon=0.67$, $\eta_p^2=0.63$
Acoustic scale difference × SNR	$F_{(4,28)}=5.4$, $p=0.007$, $\varepsilon=0.74$, $\eta_p^2=0.43$
Acoustic scale difference × distracter angle × SNR	$F_{(8,56)}=4.0$, $p=0.028$, $\varepsilon=0.33$, $\eta_p^2=0.36$

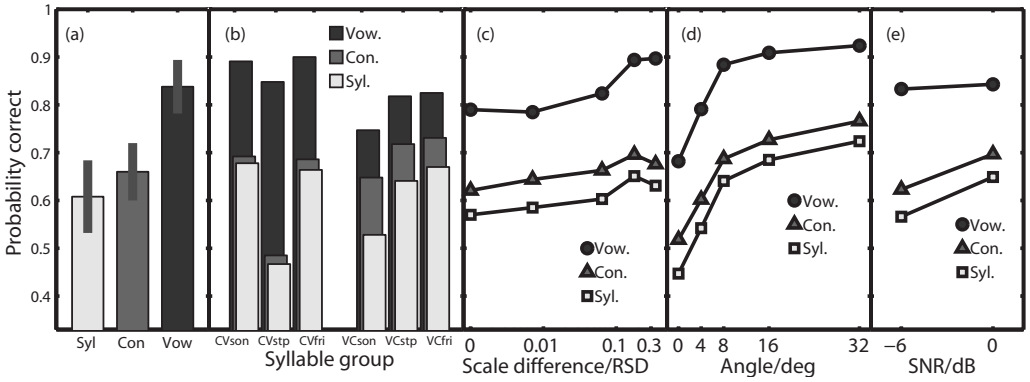

FIG. 6. Performance measures for the correct recognition of the consonant, the vowel, or both vowel and consonant (complete syllable): (a) partial scores averaged across the whole experiment; the vertical bars indicate 95% confidence intervals; (b) partial scores for each of the six syllable groups; (c) partial scores as a function of acoustic scale difference; (d) partial scores as a function of distracter angle; and (e) partial scores for each of the SNR values.

Figure 5 shows that recognition performance increases as either distracter angle or scale difference increases, and performance reaches ceiling levels for distracter angles of 16° and 32°. Figures 5(a) and 5(b) also show that the overall improvement in target recognition, due to increasing angle between target and distracter, is much larger in (b) the 0-dB SNR condition than in (a) the −6-dB SNR condition. There are two reasons for this: first, performance for small scale differences with a 0° angle and a 0-dB SNR is actually lower than that for a −6-dB SNR; second, performance for the larger angles (i.e., 8°, 16°, and 32°) with a 0-dB SNR is greater than with a −6-dB SNR, as would be expected from the increase in audibility. These two effects make the range of performance for the 0-dB condition much larger than for the −6 dB condition, which leads to the interaction. The depression of performance for small scale differences with a 0° angle and a 0-dB SNR is due to the relatively high number of distracter intrusions. Indeed, for the two smallest scale differences (0 and 0.0071), the target recognition level is the same as the distracter intrusion level, because there are no cues whatsoever to distinguish the target from the distracter. The rate of distracter intrusions rapidly decreases as differences between the target and distracter increase, be they differences in angle, acoustic scale, or level. Brungart (2001) also found that performance improved as SNR decreased in the range of 0 to −6 dB.

The solid gray lines show results from Vestergaard *et al.* (2009) with diotic stimuli. For the 0-dB condition, the results from the current study are indistinguishable from those of Vestergaard *et al.* For the −6-dB condition, the results from the current study are slightly above those of the Vestergaard study. Comparison of the 0 and −6-dB conditions shows that when the target and distracter voices are very similar, performance in the 0-dB condition is actually poorer than in the −6-dB condition, despite the fact that SNR is greater in the 0-dB condition. This is due to the increase in the number of distracter intrusions when there are no differences in vocal characteristics. The Vestergaard study shows the increase in the number of distracter confusions, but not the corresponding decrease in performance.

B. Partial scoring

In Sec. II, the measure of performance was percent correct *syllable* identification, i.e., the listener was required to identify both the vowel and consonant in each syllable correctly. It is also informative to analyze the results for the consonants and vowels separately. Figure 6(a) shows performance for syllables, consonants, and vowels averaged across all of the conditions in the experiment. Vowel recognition is significantly better than consonant recognition or syllable recognition and consonant recognition is not significantly greater than syllable recognition. This may simply reflect that fact that there are fewer vowels (5) than consonants (18) in the syllable database, or it may be that consonants mask consonants better than vowels mask vowels. Figure 6(b) shows performance for the six, individual syllable types. The color coding of the bars is the same as in Fig. 6(a), and the syllable group is shown along the abscissa. Although vowel recognition is better than consonant or syllable recognition for all syllable types, there is, nevertheless, a striking difference between the pattern of results for the CV syllables and the VC syllables. For the CV syllables, performance appears to be driven almost exclusively by consonant recognition; consonant and syllable recognition are very similar in all three cases. In contrast, consonant recognition is consistently greater than syllable recognition for the VC syllables, and vowel recognition is correspondingly lower in the VC syllables.

Figure 6(c) shows how performance varies as a function of acoustic scale difference, separately for vowels, consonants, and syllables. The pattern reflects the overall pattern of performance in Fig. 6(a); however, Fig. 6(c) shows that there is an increase in vowel recognition for large scale differences. Performance is presented as a function of distracter angle in Fig. 6(d), and it shows that the effect of angle is largely independent of phoneme category. Finally, Fig. 6(e) shows the effect of SNR on consonant and vowel recognition. Vowel recognition is the same for the two SNR values used in this study (−6 dB and 0 dB), which probably means that vowel recognition performance was close to ceiling values.

IV. DISCUSSION

The study of Vestergaard *et al.* (2009) measured the effect of acoustic scale differences on concurrent speech recognition. Their highly controlled stimuli reduced the linguistic and acoustic cues that were available to listeners. The use of syllables minimized contextual language cues, and the matching of temporal envelopes minimized glimpsing cues. Minimizing the language and glimpsing cues, in turn, increased sensitivity to the acoustic scale cues. Vestergaard *et al.* (2009) showed that the auditory system can use the acoustic scale cues, GPR and VTL to improve concurrent speech recognition. The stimuli of Vestergaard *et al.* (2009) were diotic. The current study extended the paradigm to include spatial information as it would occur in a free-field anechoic environment, to determine whether listeners would still be able to extract the acoustic scale information from a more complex stimulus. Figure 4(a) shows that in the main results: recognition performance increases with scale differences for distracter angles of 0°, 4°, and 8°. For larger distracter angles (16° and 32°), there is a ceiling effect which limits performance and masks any effects of acoustic scale that might otherwise be observed. The fact that acoustic scale information can be extracted from speech stimuli independent of speaker separation supports the hypothesis that vocal characteristics are important in multi-speaker environments.

The effect of acoustic scale difference is not as great as that of distracter angle. Figure 5(a) shows that, for −6-dB SNR, when there are no scale cues, recognition improves from 37% for a distracter angle of 0° to 65% for a distracter angle of 32°. When there are scale cues, the improvement is smaller rising from about 55% for a distracter angle of 0° and a scale difference of 0.34 to about 70% for a distracter angle of 32° and a scale difference of 0.34. For the 0-dB SNR (Fig. 5(b)), when there are no scale cues, recognition rises from about 35% for a distracter angle of 0° to 77% for a distracter angle of 32°. When there is scale difference of 0.34, performance rises from about 63% for a distracter angle of 0° to about 80% for a distracter angle of 32°. The largest advantage due to acoustic scale cues arises when the distracter and target are in the same location and have the same level (i.e., Figure 5(b), compare a scale difference of 0 at 0° with a scale difference of 0.34 at 0°). In this case performance rises from 35 to about 63% correct.

The effect of acoustic scale difference is largely obscured by ceiling performance in this experiment for distracter angles larger than 8°. This shows that either the benefit from the spatial cues is so great that any potential improvement from the scale cues is overwhelmed, or that for larger distracter angles the scale cues cannot be extracted from the signal in an efficient manner. The individual subject data show that the two listeners with the worst overall performance continue to benefit from acoustic scale cues at the larger distracter angles (16° and 32°). This suggests that acoustic scale cues can be used with larger separation angles if recognition performance is not already near ceiling.

Separate analysis of the consonant and vowel data shows that syllable recognition is driven mainly by consonant recognition (Fig. 6(a)), which may be partly due to there being fewer vowels (5) than consonants (18). At the same time, there is a clear difference between consonant and vowel recognition for CV and VC syllables (Fig. 6(b)). CV syllable recognition is less dependent on vowel identification than VC recognition. This difference is probably due to the vowels in VC syllable pairs overlapping more than for vowels in CV syllable pairs.

It is clear from the results (Figs. 4(a) and 5) that the spatial separation of the target and distracter greatly improves performance. This spatial release from masking arises due to the differences in either the arrival time and/or the level of the sounds at the two ears. An improvement due to level differences (ILDs) is known as the "better-ear" advantage and an improvement due to arrival time differences (ITDs) at the two ears is known as "binaural unmasking." These differences are shown in Fig. 2; the ITDs are 0 μs, 20 μs, 60 μs, 120 μs, and 240 μs, for angles of 0°, 4°, 8°, 16°, and 32°, respectively; and the maximum ILDs are 1.6, 2.5, 4, 6.5, and 13.2 dB for angles of 0°, 4°, 8°, 16°, and 32°, respectively. Assuming that acoustic scale cues are beneficial for distracter angles of up to 8°, then the combined effect of an ITD of at least 60 μs and an ILD of at least 4 dB is sufficient to produce ceiling performance and mask any potential advantage from scale cues for most listeners. This is a relatively small change in the signal. However, the experiment simulates anechoic conditions wherein location cues are unrealistically clean due to the absence of reflections. In a reverberant environment, which is what is typical for multi-speaker situations, we would anticipate that location cues derived from binaural unmasking or level differences would be greatly reduced (Hartmann *et al.*, 2005) and that acoustic scale cues would be of greater relative benefit. Although binaural unmasking and ILD cues are greatly reduced in reverberant conditions, listeners can take advantage of a perceptual segregation of speakers, which occurs for larger spatial separations, to improve recognition performance. This perceptual segregation is trivial in anechoic conditions, and as such, it was anticipated that there would be no benefit from acoustic scale cues because recognition would already be at ceiling levels. In reverberant environments, recognition performance would decrease (Lavandier and Culling, 2007) and listeners might well benefit more from acoustic scale cues.

V. CONCLUSIONS

It was argued by Vestergaard *et al.* (2009) that there exists a trading relationship between log(VTL ratio) and log(GRP ratio) of about 1.6. Therefore, a change in GPR of two semi-tones would produce a similar effect to a 20% change in VTL. In the current study, the trading relationship was set to a value of 1.5 [this is the same as the initial trading relationship of the stimuli used in the study of Vestergaard *et al.* (2009)] and there was found to be no effect of spoke angle, i.e., performance was no better for one particular spoke than another. This shows that, for the current experiment, there also exists a trading relationship between the logarithms of VTL ratio and GPR ratio of about 1.5.

The aim of the current study was to extend the work of Vestergaard *et al.* (2009) and show that acoustic scale information could be extracted in conjunction with location from speech sounds. The results show that, in dichotic anechoic conditions, differences in the acoustic scale of two concurrent speakers can be used to improve recognition performance. Recognition improved over a range of distracter locations (0°, 4°, and 8°), where performance was below the ceiling set by spatial unmasking. In reverberant conditions, where the signal is contaminated by reflections, spatial unmasking is reduced, and we would expect recognition performance to decrease (Lavandier and Culling, 2007). In such conditions, listeners might well exhibit greater benefits from acoustic scale cues.

ACKNOWLEDGMENTS

Research supported by the U.K. Medical Research Council (Grant Nos. G0500221 and G9900369). The authors would like to thank Kristopher Knott and Beng Beng Ong for their assistance in running the experiments.

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

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

Cooke, M. (**2006**). "A glimpsing model of speech perception in noise," J. Acoust. Soc. Am. **119**, 1562–1573.

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.

Durlach, N. I. (**1963**). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. **35**, 1206–1218.

Fitch, W. T., and Giedd, J. (**1999**). "Morphology and development of the human vocal tract: A study using magnetic resonance imaging," J. Acoust. Soc. Am. **106**, 1511–1522.

Hartmann, W., Rakerd, B., and Koller, A. (**2005**). "Binaural coherence in rooms," Acta. Acust. Acust. **91**, 451–462.

Hawley, M. L., Litovsky, R. Y., and Colburn, H. S. (**1999**). "Speech intelligibility and localization in a multi-source environment," J. Acoust. Soc. Am. **105**, 3436–3448.

Hirsh, I. J. (**1948**). "The influence of interaural phase on interaural summation and inhibition," J. Acoust. Soc. Am. **20**(4), 536–544.

Hirsh, I. J. (**1950**). "The relation between localization and intelligibility," J. Acoust. Soc. Am. **22**, 196–200.

Irino, T., and Patterson, R. D. (**2002**). "Segregating information about the size and shape of the vocal tract using a time-domain auditory model: The stabilised wavelet-Mellin transform," Speech Commun. **36**, 181–203.

Ives, D. T., Smith, D. R. R., and Patterson, R. D. (**2005**). "Discrimination of speaker size from syllable phrases," J. Acoust. Soc. Am. **118**, 3816–3822.

Kawahara, H., and Irino, T. (**2004**). "Underlying principles of a high-quality speech manipulation system STRAIGHT and its application to speech segregation," in *Speech Segregation by Humans and Machines*, edited by P. Divenyi (Kluwer Academic, Boston), pp. 167–180.

Lavandier, M., and Culling, J. F. (**2007**). "Speech segregation in rooms: Effects of reverberation on both target and interferer," J. Acoust. Soc. Am. **122**, 1713–1723.

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

Marcus, S. M. (**1981**). "Acoustic determinants of perceptual centre (Pcentre) location," Percept. Psychophys. **30**, 247–256.

Møller, H., Sorensen, M. F., Jensen, C. B., and Hammershøi, D. (**1996**). "Binaural technique: Do we need individual recordings?," J. Audio Eng. Soc. **44**, 451–469.

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

Scott, S. K. (**1993**). "P-centres in speech: An acoustic analysis," Ph.D. thesis, University College, London.

Shaw, W. A., Newman, E. B., and Hirsh, I. J. (**1947**). "The difference between monaural and binaural thresholds," J. Acoust. Soc. Am. **19**, 734.

Smith, D. R., and Patterson, R. D. (**2005**). "The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex and age," J. Acoust. Soc. Am. **118**, 3177–3186.

Smith, D. R. R., Patterson, R. D., Turner, R., Kawahara, H., and Irino, T. (**2005**). "The processing and perception of size information in speech sounds," J. Acoust. Soc. Am. **117**, 305–318.

Turner, R. E., Walters, T. C., Monaghan, J. J. M., and Patterson, R. D. (**2009**). "A statistical, formant-pattern model for segregating vowel type and vocal-tract length in developmental formant data," J. Acoust. Soc. Am. **125**, 2374–2386.

Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2009**). "The interaction of vocal tract length and glottal pulse rate in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**, 1114–1124.

von Kriegstein, K., Smith, D. D. R., Patterson, R. D., Ives, D. T., and Griffiths, T. (**2007**). "Neural representation of auditory size in the human voice and in sounds from other resonant sources," Curr. Biol. **17**, 1123–1128.

Walters, T. C., Gomersall, P., Turner, R. E., and Patterson, R. D. (**2008**). "Comparison of relative and absolute judgments of speaker size based on vowel sounds," Proc. Meet. Acoust. **1**, 050003.

Wenzel, E. M., Arruda, M., Kistler, D. J., and Wightman, F. L. (**1993**). "Localisation using nonindividualized head-related transfer functions," J. Acoust. Soc. Am. **94**, 111–123.

Wightman, F. L., and Kistler, D. J. (**1989a**). "Headphone stimulation of free-field listening. I: Stimulus synthesis," J. Acoust. Soc. Am. **85**, 858–867.

Wightman, F. L., and Kistler, D. J. (**1989b**). "Headphone stimulation of free-field listening. II: Psychophysical validation," J. Acoust. Soc. Am. **85**, 868–878.

Wightman, F. L., and Kistler, D. J. (**1999**). "Resolution of front-back ambiguity in spatial hearing by listener and source movement," J. Acoust. Soc. Am. **105**, 2841–2853.

Wightman, F. L., and Kistler, D. J. (**2005**). "Measurement and validation of human HRTFs for use in hearing research," Acta Acust. **91**, 429–439.