The advantage of spatial and vocal characteristics in the recognition of competing speech

MARTIN D. VESTERGAARD, D. TIMOTHY IVES, AND ROY D. PATTERSON

Centre for the Neural Basis of Hearing, Department of Physiology, Development and Neuroscience, University of Cambridge, United Kingdom

In multi-speaker environments, listeners take advantage of a variety of cues that characterize the target and distracter speakers to improve speech recognition. Spatial cues like interaural time and intensity differences provide binaural unmasking and a better-ear advantage. Vocal characteristics such as pitch and resonance scale help to disambiguate concurrent speech. Temporal misalignment of competing speech signals can improve recognition by virtue of 'listening in the dips'. In this paper, we review a series of experiments on the advantage of spatial and vocal characteristics in the recognition of concurrent speech. Syllable pairs were synthesized to simulate different speakers, and the recognition of syllables that varied in spatial and vocal characteristics was measured. The effect of temporal glimpsing was measured by aligning the temporal envelopes of the competing signals in a controlled way. The results show that spatial and vocal cues compete to provide selectivity of concurrent speech sounds. When they are clearly separated in space, vocal characteristics can only further improve performance marginally. However, when they are temporally and spatially aligned, a substantial advantage can be derived from the vocal characteristics. The paper discusses the interaction of spatial and vocal cues, and the patterns of syllable confusions that listeners make.

INTRODUCTION

The phenomenon that listeners can attend selectively to one speaker in a multi-speaker environment in order to separate their speech from distracting speech sounds uttered by other speakers is sometimes called the cocktail party problem (Cherry, 1953). However, there are several acoustic cues in speech sounds that make normally hearing listeners remarkably good at segregating competing speech. The most significant cue for speech segregation is almost undoubtedly audibility, which is largely determined by signal to noise ratio (SNR). When whole sentences are matched for overall SNR there are 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 glottal pulse rate (GPR) and vocal tract length (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 (ST) (e.g., Chalikia and Bregman, 1993; Qin

Proceedings of ISAAR 2009: Binaural Processing and Spatial Hearing. 2nd International Symposium on Auditory and Audiological Research. August 2009, Elsinore, Denmark. Edited by J. M. Buchholz, T. Dau, J. Christensen-Dalsgaard, and T. Poulsen. ISBN: 87-990013-2-2. The Danavox Jubilee Foundation 2010.

and Oxenham, 2005; Assmann and Summerfield, 1990; Assmann and Summerfield, 1994; Culling and Darwin, 1993). Moreover, interaural time and intensity differences (ITD, IID) provide spatial information about competing sources, which may help listeners segregate target speech from distracting speech (Drennan *et al*., 2003; Culling and Summerfield, 1995).

The role of pitch in concurrent speech has been investigated in many studies. Chalikia and Bregman (1993) 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) showed that small departures from otherwise constant F0 tracks can improve vowel recognition. Qin and Oxenham (2005) show that recognition of concurrent vowels reached a maximum when the difference in F0 was about 4 ST. Summerfield and Assmann (1991) have 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 colleagues (de Cheveigné *et al.*, 1997b; de Cheveigné *et al.*, 1997a; de Cheveigné, 1993) argued that the advantage of an F0 difference depends primarily on the harmonicity of the distracter (*i.e.* harmonic cancellation mechanism).

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. Brungart 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. Darwin *et al.,* (2003) investigated the effects of F0 and VTL in a study on concurrent speech. They reported an increase in speech recognition of 28%, most of which (~20%) was already apparent at an F0 difference of 4 ST. They also found that individual differences in intonation can help identify speech of similar F0, corroborating the findings of Assmann and Summerfield (1994). Moreover, for a 38% change in VTL, Darwin *et al.* reported an increase in recognition of ~20% at 0 dB SPL. 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. These results 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) have shown that listeners can judge the relative size/age, and the sex of a speaker based on their vowels even when the GPR and VTL were well beyond the range of normal speech.

The phonetic identity of vowels is specified by the formant frequencies. They are determined by the filtering of the supralaryngal vocal tract, which consists of the oral and nasal cavities above the larynx. Formant frequencies are largely independent of pitch but they vary with VTL (Lee *et al.*, 1999), so to understand different speakers, the listener needs to normalize for the phonetically irrelevant variation in GPR and VTL. 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 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.

In the current paper, we present a series of experiments intended to measure the interaction of cues specifying auditory size and other cues used for segregating competing speech sounds.

METHOD

Listeners were required to identify syllables spoken by a target voice in the presence of a distracting voice. Performance was measured for target and distracter voices that were voiced and whispered over a large range of different voices, presented at different SNRs and simulated locations. The experimental protocol was approved by the Cambridge Psychology Research Ethics Committee (CPREC).

Subjects

Thirty-eight listeners participated in the study (24 male). Their average age was 21 years (17 – 33 years), and no subject had a history of any audiological disorders. After informed consent was obtained from the participants, an audiogram was recorded at the standard octave frequencies between 500 and 4,000 Hz, bilaterally, to ensure that they had normal hearing. The subjects took part in one or more of the experiments focussing on different aspects of concurrent speech segregation, and they all took part in a pre-experimental measurement intended to provide baseline data for the speech material used in the main experiments.

Procedure

The procedure was the same in all the experiments: syllables were presented in triplets to promote perception of the stimuli as a phrase of connected speech as described by Vestergaard *et al*. (2009). The listeners responded by clicking on an orthographical representation of their answer from a response matrix on a computer screen. They were seated in front of the response screen in an IAC double-walled, sound-attenuated booth, and the stimuli were presented bilaterally via AKG K240DF headphones.

Stimuli

The stimuli were taken from the CNBH syllable corpus previously described by Ives *et al.* (2005). 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 and fricatives), and each of the consonants was paired with one of 5 vowels spoken in both CV and VC combinations. The syllables were analyzed and re-synthesized with the STRAIGHT vocoder (Kawahara and Irino, 2004) to simulate voices with different combinations of GPR and VTL. To simulate whispered speech, the STRAIGHT spectrograms were excited with broadband noise and high-pass filtered at 6 dB/oct. This procedure removes pitch from the voiced part of the syllables and creates an effective simulation of whispered speech.

A combination of techniques was employed to control temporal glimpsing. First, the perceptual centers (Marcus, 1981) of the syllables were aligned as described by Ives *et al*. (2005), and target and distracter syllables were matched according to their. phonetic specification as described by Vestergaard *et al.* (2009). Then, the distracter was offset systematically by values between -400 and 400 ms. The offset describes distracter latency relative to the target syllable; *i.e.* a positive value indicates that the distracter was delayed compared to the temporally matched condition. Within the six types of syllables [2 (CV *vs* VC) × 3 (consonant category)], 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

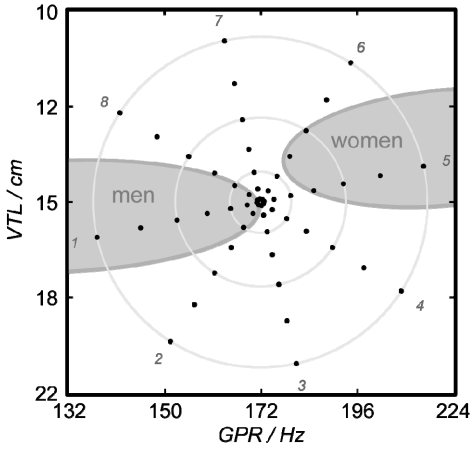

Fig. 1: The vocal characteristics form a spoke pattern across a large range of natural and unnatural combinations of glottal pulse rate (GPR) and vocal tract length (VTL). Figure adapted from Vestergaard *et al.* (2009).

The reference voice used for training had a GPR of 172 Hz and a VTL of 14.7 cm. The combinations of GPR and VTL for the experimental voices are shown by the dots in Fig. 1, which span an ellipse around the reference voice. The ellipse had a radius of 4 ST along the GPR axis and 6 ST 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). In all, there were 57 different voices with the vocal characteristics illustrated in Fig. 1 (see Vestergaard *et al.* (2009) for a table with the exact values).

RESULTS AND DISCUSSION

During the pre-experimental training the listeners took part in a baseline experiment intended to measure the unmasked recognition of the syllables used in the main experiments. The results plotted in Fig. 2 showed that once listeners had been trained on the set of 180 syllables uttered by one voice (voice #0 in Fig. 2), recognition performance remained high for subsequent, novel voices (voices #1-8 in Fig. 2) from the same set. This result is consistent with the hypothesis that the auditory system normalizes out the variability associated with speaker characteristics to enhance robustness.

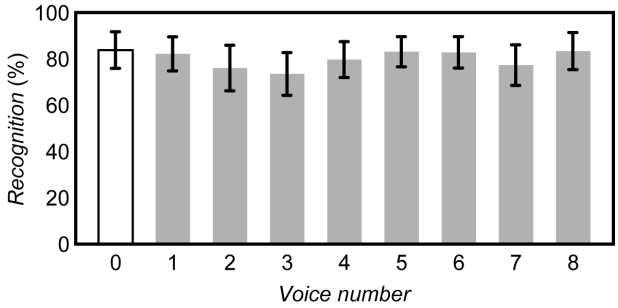

Fig. 2: Baseline recognition for the extreme voices illustrated in Fig. 1. Voice number 0 is the training voice in the centre of the spoke pattern.

The relative contribution of the two speaker-specific properties of speech GPR, and VTL was quantified in Vestergaard *et al.* (2009). In multi-speaker 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 shown in Fig. 3. The results also showed 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. Vestergaard *et al*. modeled this trading relationship and found that when the two variables were measured in logarithmic units, and there are no loudness cues to assist in tracking the target speaker, then a change in VTL had to be about 1.6 times a change in GPR to have the same effect on performance.

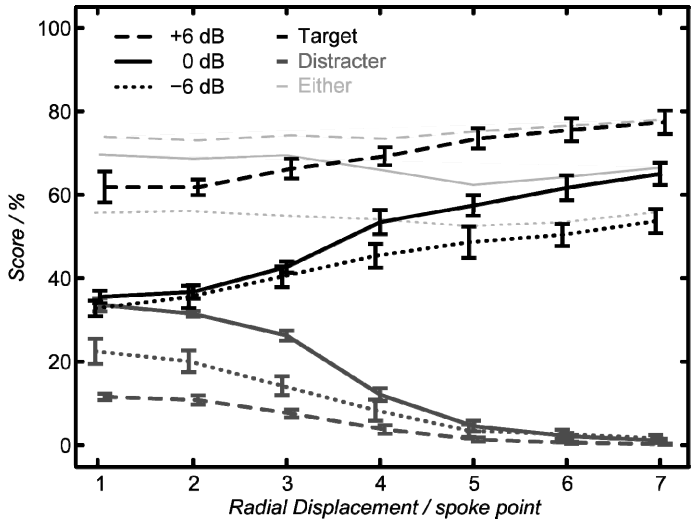

Fig. 3: The interaction of vocal characteristics and audibility on syllable recognition and distracter intrusion in concurrent speech. Figure taken from Vestergaard *et al.* (2009).

To test the effect of voicing and the cancellation theory, which is thought to make rejecting distracter syllables more effective than enhancing target syllables, we generated a small set of four voices from the full set. They had markedly different simulated VTLs and were either voiced or whispered. The difference in VTL ensured that it was easy to hear a vocal difference between them when pitch was removed from both, simulating a whispered voice interfering with another whispered voice. All four combinations of vocal contrast were tested. The results in Fig. 4 show syllable, consonant and vowel recognition scores plotted as a function of audibility (Speech Intelligibility Index (SII), ANSI, 1997). The main conclusion was that listeners can use voicing whenever it is present both to detect the target speech and to reject the distracter. Three of the four vocal conditions contained voicing in either the target, the distracter, or both, and they show comparable results once audibility has been taken into account. By contrast, in the condition in which both target and distracter were whispered, performance drops off progressively with audibility, particularly when the audibility index is below 0.5. In other words, audibility predicts the identification of the target when one of the concurrent syllables is voiced, but it leads to an overestimation of the recognition of whispered syllables when the distracter is a whispered syllable.

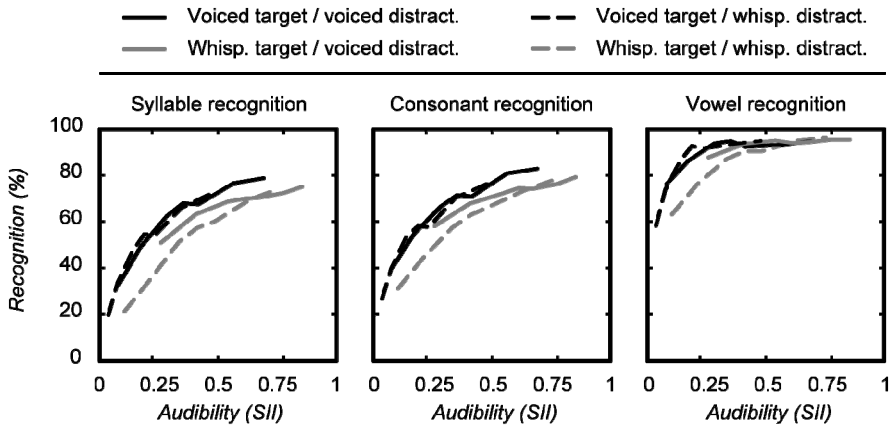

Fig. 4: The effect of voicing compared to whispering on concurrent syllable recognition. Figure adopted from Vestergaard and Patterson (2009).

Ives *et al.* (2009) measured the interaction of vocal characteristics and spatial orientation of concurrent syllables. They used generic head related transfer functions (HRTF) to simulate the location of the competing voices, and manipulated the location of the distracter off to the side under the assumption that listeners normally face the target speaker. They used a subset of the speakers shown in Fig. 1 including four of the extreme voices and found a similar advantage to that reported by Vestergaard *et al.* (2009). Thus, their results show that even in this more ecologically valid listening situation, the listeners were able to derive advantage from the vocal characteristics. However, when a spatial separation between the two voices was introduced, the advantage from differences in vocal characteristics decreased. Their

main results are illustrated in Fig. 5 which shows recognition performance for the five different voices used as a function of simulated spatial separation. When the competing voices were very similar, the listeners derived additional advantage from spatial separation, but when the voices were already very different the additional advantage derived from the spatial cues was dramatically reduced.

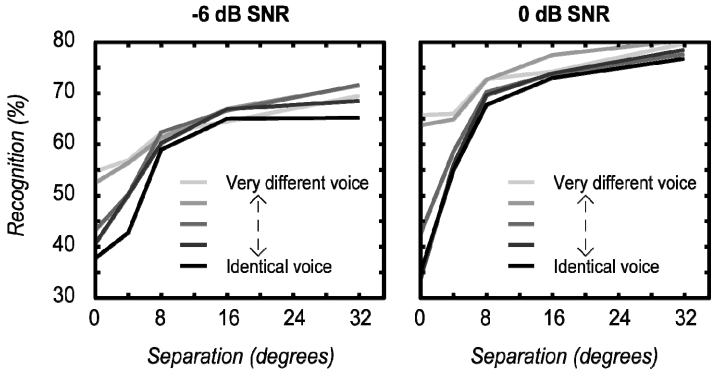

Fig. 5: The interaction of spatial separation and vocal characteristics on syllable recognition in concurrent speech.

In all of the previous experiments the temporal envelopes of the syllables were accurately matched. In order to measure the effect of temporal cues, asynchrony was systematically introduced by shifting the distracter forward or backward by values between –400 ms and +400 ms. The results from this experiment are shown in Fig. 6 which shows syllable recognition as a function of distracter offset, separately for CV and VC syllables for all voices (left panel), and separately for similar and dissimilar voices (right panel). Square symbols are used for CVs and circles for VCs. Overall, recognition performance recovered more and somewhat steeper when the distracter was played before than when delayed. There was a strong interaction between

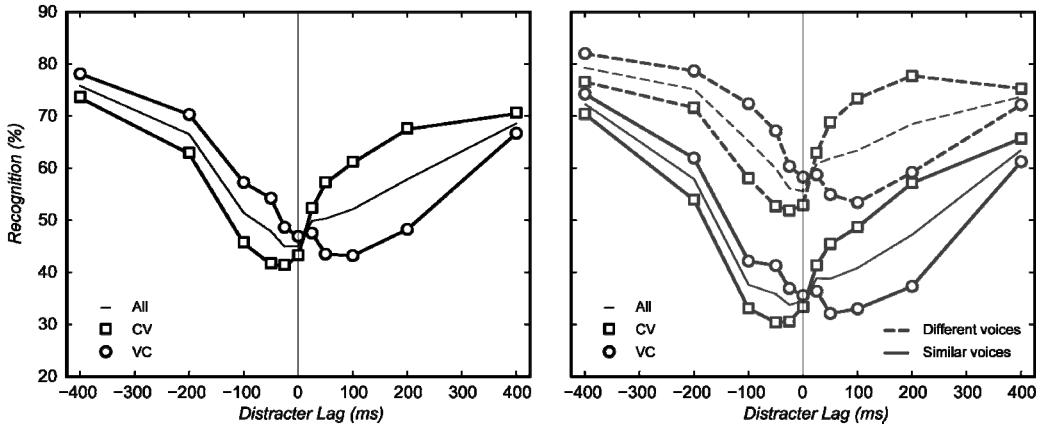

Fig. 6: The interaction of temporal asynchrony and vocal characteristics on syllable recognition in concurrent speech. Figure adopted from Vestergaard *et al*. (2011).

consonant position and offset direction as illustrated in the left panel. For negative offset, the recognition of VC syllables recovered first as offset duration increased, whereas for positive distracter offset, performance recovered first for CV syllables. Moreover, the poorest performance was not observed for an offset of zero when the results are plotted separately for CVs and VCs. For CVs it occurs in the distracter latency range between –25 and –50 ms, and for VCs it occurs in the +50 to +100 ms latency range. At these latencies there is energy from the distracter vowel overlapping with the target consonant.

CONCLUSION

Normally hearing listeners can readily normalize for the acoustic variability associated with different voices. However, they can also use this acoustic variability to disambiguate concurrent speech. The experiments showed how two speaker-specific properties of speech [glottal pulse rate (GPR), and vocal tract length (VTL)] assist a listener in segregating competing speech signals. When speaker-specific differences between the target and distracter were reduced, recognition decreases dramatically. Moreover, listeners can use voicing either to detect the target speech or to reject the distracter. When the predictable effects of audibility were taken into account, limited evidence remained for the harmonic cancellation mechanism thought to make rejecting distracter syllables more effective than enhancing target syllables. Spatially separating target and distracter also enhanced recognition performance. A clear advantage was observed for speaker separations of as little as 4 degrees, and the increase in performance leveled out at around 16 degrees separation. However, the relative advantage from spatial separation was reduced when the concurrent voices were dissimilar, and/or when there were loudness cues to assist the listeners. Furthermore, temporal asynchrony can provide unmasking of the consonants in partially concurrent syllables leading to an increase in recognition performance. The results are consistent with the notion that audibility is the prime determinant of performance, and that vocal and spatial cues are particularly effective when there are no loudness cues. The study also demonstrated that the auditory system can use any combination of these cues to segregate competing speech signals.

AKNOWLEDGEMENT

Research supported by the UK-MRC [G0500221, G9900369]. We would like to thank James Tanner, Sami Abu-Wardeh, Charles Barter, Hamish Findlay, Andy Taylor, Beng Beng Ong and Kristopher Knott for help with collecting the data, and Nick Fyson for assistance in producing the programs that ran the experiments.

REFERENCES

ANSI (**1997**). *S3.5. Methods for the calculation of the speech intelligibility index* (American National Standards Institute, New York).

Assmann, P. F., and Summerfield, Q. (**1990**). “Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies,” J. Acoust. Soc. Am. **88**, 680-697.

Assmann, P. F., and Summerfield, Q. (**1994**). “The contribution of waveform interactions to the perception of concurrent vowels,” J. Acoust. Soc. Am. **95**, 471-484.

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

Chalikia, M. H., and Bregman, A. S. (**1993**). "The perceptual segregation of simultaneous vowels with harmonic, shifted, or random components," Percept. Psychophys. **53**, 125-133.

Cherry, E. C. (**1953**). "Some experiments on the recognition of speech, with one and 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.

Culling, J. F., and Darwin, C. J. (**1993**). "The role of timbre in the segregation of simultaneous voices with intersecting f0 contours," Percept. Psychophys. **54**, 303-309.

Culling, J. F., and Summerfield, Q. (**1995**). "Perceptual separation of concurrent speech sounds: Absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. **98**, 785-797.

Darwin, C. J., Brungart, D. S., and Simpson, B. D. (**2003**). "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," J. Acoust. Soc. Am. **114**, 2913-2922.

de Cheveigné, A. (**1993**). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. **93**, 3271-3290.

de Cheveigné, A., McAdams, S., and Marin, C. M. H. (**1997a**). "Concurrent vowel identification. Ii. Effects of phase, harmonicity and task," J. Acoust. Soc. Am. **101**, 2848-2856.

de Cheveigné, A., Kawahara, H., Tsuzaki, M., and Aikawa, K. (**1997b**). "Concurrent vowel identification. I. Effects of relative amplitude and f0 difference," J. Acoust. Soc. Am. **101**, 2839-2847.

Drennan, W. R., Gatehouse, S., and Lever, C. (**2003**). "Perceptual segregation of competing speech sounds: The role of spatial location," J. Acoust. Soc. Am. **114**, 2178-2189.

Fant, G. C. M. (**1970**). *Acoustic theory of speech production* (Mouton, The Hague).

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

Ives, D. T., Vestergaard, M. D., and Patterson, R. D. (**2009**). "Location and acoustic scale cues in concurrent speech recognition," J. Acoust. Soc. Am. *submitted*.

Kawahara, H., and Irino, T. (**2004**). "Underlying principles of a high-quality speech manipulation system straight and its application to speech segregation," in *Speech separation by humans and machines*, edited by P. L. Divenyi (Kluwer Academic, Boston MA).

Ladefoged, P., and Broadbent, D. E. (**1957**). "Information conveyed by vowels," J. Acoust. Soc. Am. **29**, 98-104.

Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455-1468.

Marcus, S. M. (**1981**). "Acoustic determinants of perceptual center (p-center) location," Percept. Psychophys. **30**, 247-256.

Miller, G. A., and Licklider, J. C. R. (**1950**). "The intelligibility of interrupted speech," J. Acoust. Soc. Am. **22**, 167-173.

Qin, M. K., and Oxenham, A. J. (**2005**). "Effects of envelope-vocoder processing on f0 discrimination and concurrent-vowel identification," Ear Hear. **26**, 451-460.

Ritsma, R. J., and Hoekstra, A. (**1974**). "Frequency selectivity and the tonal residue," in *Facts and models in hearing*, edited by E. Zwicker and E. Terhardt (Springer, Berlin).

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.

Summerfield, Q., and Assmann, P. F. (**1991**). "Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony," J. Acoust. Soc. Am. **89**, 1364-1377.

Titze, I. R. (**1989**). "Physiologic and acoustic differences between male and female voices," J. Acoust. Soc. Am. **85**, 1699-1707.

Vestergaard, M. D., and Patterson, R. D. (**2009**). "Effects of voicing in the recognition of concurrent syllables," J. Acoust. Soc. Am. **126**, 2860-2863.

Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2009**). "The interaction of vocal characteristics and audibility in the recognition of concurrent syllables," J. Acoust. Soc. Am. **125**, 1114-1124.

Vestergaard, M. D., Fyson, N. R. C., and Patterson, R. D. (**2011**). "The mutual roles of temporal glimpsing and vocal characteristics in cocktail-party listening," J. Acoust. Soc. Am. *in press*.