Published in final edited form as:

J Acoust Soc Am. 2009 December ; 126(6): 2860–2863. doi:10.1121/1.3257582.

Effects of voicing in the recognition of concurrent syllables (L)a

Martin D. Vestergaard and 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

Abstract

This letter reports a study designed to measure the benefits of voicing in the recognition of concurrent syllables. The target and distracter syllables were either voiced or whispered, producing four combinations of vocal contrast. Results show that listeners use voicing whenever it is present either to detect a target syllable or to reject a 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.

I. INTRODUCTION

It has been shown that segregation of competing voices is facilitated if the target and distracter voices differ either in glottal pulse rate (GPR) or vocal tract length (VTL) (e.g., Darwin *et al.*, 2003). Moreover, Vestergaard *et al.* (2009) measured the interaction of GPR and VTL in concurrent syllable recognition using a paradigm that controlled temporal glimpsing and the idiosyncrasies of individual voices. They showed that, when the signal-to-noise ratio (SNR) is 0 dB, a two-semitone (ST) difference in GPR produced the same performance advantage as a 20% difference in VTL. This letter reports an extension of Vestergaard *et al.*, 2009, designed to measure the benefits of voicing to the recognition of concurrent syllables.

It has been proposed that the benefit of a pitch difference in the recognition of concurrent speech is that it helps the listener reject sounds that fit the harmonic structure of the distracting voice (cancellation theory) rather than assists the listener in selecting sounds that fit the harmonic structure of the target voice (enhancement theory). In a series of double-vowel experiments, de Cheveigné and colleagues (de Cheveigné *et al.*, 1997b, 1997a; de Cheveigné, 1997, 1993) developed a harmonic cancellation model tuned to the periodicity of the distracter. They showed that the advantage of a pitch difference depends primarily on the harmonicity of the distracter. In order to evaluate the feasibility of the cancellation theory for connected syllables, the current study employed the paradigm described by Vestergaard *et al.* (2009) for voiced and whispered syllables. Whispered speech signals do not provide any acoustic cues to the harmonicity that characterizes speech in voiced phonation (Abercrombie, 1967).

Whispered phonation is produced by allowing turbulent air to flow through partially open glottal folds. Turbulence reduces the gradient of airflow velocities and results in a noise-like excitation of the vocal tract resonances (the formants). In naturally produced speech, whispered syllables are elongated (Schwartz, 1967) and air consumption is dramatically

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

increased (Schwartz, 1968a). Whispered speech is typically 15–20 dB softer than voiced speech (Traunmüller and Eriksson, 2000) and has a spectral tilt of approximately +6 dB/octave (Schwartz, 1970). These features lead to a reduction in the perceptibility of whispered speech while it remains relatively robust. Indeed, whispered speech can convey much of the information that voiced speech can convey. Tartter (1991) found that the intelligibility of whispered vowels was 82%; only 10% lower than for voiced vowels. The recognition of whispered consonants was 64%; much above chance for the 18 consonants in their experiment (Tartter, 1989). Listeners are also able to identify speaker sex given isolated whispered vowels (Schwartz and Rine, 1968) or even when presented with isolated voiceless fricatives (Schwartz, 1968b). Lass *et al.* (1976) reported that the recognition of speaker sex dropped from 96% correct for voiced speech to 75% correct for whispered speech. Moreover, Tartter and Braun (1994) showed that listeners can accurately distinguish “frowned,” neutral, and “happy” speech in both voiced and whispered phonations. Thus, while the purpose of whispering is to reduce audibility, whispered speech remains highly functional in other respect. When instructed to adjust the frequency of a pure tone so that it matched the perceived pitch of a whispered vowel [sic], listeners matched the frequency of the pure tone with the frequency of the second formant of the vowel (Thomas, 1969). Thus, in the absence of temporarily defined pitch, listeners revert to an extracted spectral pitch measure as described by Schneider *et al.* (2005). Together, these studies suggest that we can investigate the functional role of harmonicity on the segregation of concurrent syllables by removing the temporal regularity of voiced speech samples and applying a spectral lift, thus simulating whispered speech.

In the current investigation, audibility was varied by testing at different SNRs. Because whispered speech has relatively more high-frequency energy than voiced speech, it has the potential to be a more efficient masker and a more robust target than voiced speech when presented at the same RMS level. To control such effects, syllable recognition was analyzed as a function of audibility as defined by the speech intelligibility index (ANSI, 1997).

The purpose of the current study was to determine the importance of voicing in the recognition of concurrent speech. It was hypothesized that performance in a syllable recognition task would improve whenever the auditory system can make use of voicing, either to detect a target or reject a competing distracter. Moreover, if the mechanism of cancellation were more effective than the corresponding enhancement mechanism, listeners would be more successful at using voicing to reject a distracter than to detect a target.

II. METHOD

The participants were required to identify syllables spoken by a target voice in the presence of a distracting voice. The target and distracter voices were voiced and whispered in all combinations, and the SNR was varied over a wide range.

The stimuli were taken from the syllable corpus previously described by Ives *et al.* (2005). It contains 180 syllables, divided into 90 consonant-vowel (CV) and vowel-consonant (VC) pairs. Six consonants from each of the three categories (plosives, sonorants,¹ and fricatives) were paired with one of five vowels. 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/octave. This procedure removes pitch from the voiced part of the syllables (in consonants as well as in

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 ([m], [n], [r], [j], [l], [w]).

vowels) and produces an effective simulation of whispered speech. The target voice simulated a tall male speaker (VTL: 182 mm; GPR: 157 Hz when voiced) and the distracting voice sounded like a female of normal height (VTL: 139 mm; GPR: 203 Hz when voiced). The difference in VTL is several just noticeable differences for the discrimination of resonance scale for syllables (Ives *et al.*, 2005), so even when pitch was removed from both the target and distracter voices, it was still easy to hear the difference between the two voices (Vestergaard *et al.*, 2005; Vestergaard, 2007). The target and distracter syllable pairs were presented in triplets to promote the perception of connected speech, and the syllables within a pair were matched according to their phonetic specification to reduce temporal glimpsing (Vestergaard *et al.*, 2009). The listeners responded by clicking on a syllable matrix on a computer screen, and they were given training on the use of the interface before commencing the main experiment. They were seated in an IAC (Winchester, United Kingdom) double-walled, sound-attenuated booth, and the stimuli were presented bilaterally via AKG K240DF headphones at 60-dB sound pressure level.

There were eight listeners (19–21 years old; three male) who all provided informed consent, and the experimental protocol was approved by the Cambridge Psychology Research Ethics Committee (CPREC). Audiograms were recorded at octave frequencies between 500 and 4000 Hz bilaterally, to ensure that the listeners had normal hearing.

Recognition performance was measured for the target voice in a $2 \times 2 \times 7$ factorial design (i.e., 28 conditions). The target voice was either voiced or whispered, and the distracter was either voiced or whispered. Each pair of target and distracter voices was measured at seven SNRs [−15, −9, −3, 0, 3, 9, 15 dB]. The trials were blocked in runs of 40 trials, within which the voice combination and the SNR were constant. Between runs the condition was randomly chosen from the full set of the 28 conditions. Each condition was repeated three times. To increase the sensitivity of the experiment to the variation in voicing, the task was made slightly more difficult by playing the target syllable in either interval 2 or 3. A visual indicator marked the interval to which the listener should respond (for details of the rationale for this paradigm, see Vestergaard *et al.*, 2009).

III. RESULTS

Three different scores were computed: percent correct syllable recognition (the primary task), percent correct consonant recognition, and percent correct vowel recognition. The average values for the three scores are shown as a function of SNR for the four vocal conditions in the top row of Fig. 1. The results show the expected effect of SNR and some notable effects of voicing. To control for the predictable effect of audibility caused by the difference in spectrum between voiced and whispered speech, the following analysis was run: For each trial, an audibility index [the speech intelligibility index (SII) (ANSI, 1997)] was calculated by deriving the spectrum levels for the target and distracter syllables. The distracters' spectrum levels were used as masker when estimating the audibility of the targets. An important function for English nonsense syllables was used according to the ANSI standard. A transfer function (Sherbecoe and Studebaker, 1990) was then fitted to the data. This analysis allows for a data-driven audibility-controlled prediction of recognition performance. The results of this transform are shown at the bottom row of Fig. 1. They show the effects of voicing once audibility has been taken into account.

The scores from the experiment and the scores from the prediction described above were converted to rationalized arcsine units (RAUs) (Studebaker, 1985; Thornton and Raffin, 1978). The effects of voicing on performance were analyzed by assessing the departure of the observed scores from the corresponding predictions. A three-way repeated-measures analysis of variance [$2(\text{target}) \times 2(\text{distracter}) \times 7(\text{SNR})$] was run on the prediction mismatch

units (observed RAU scores–predicted RAU scores) for the three different scores. Greenhouse–Geisser correction was used to compensate for lack of sphericity, and paired comparisons with Sidak correction were used to analyze effects within condition. For the syllable scores (A panels in Fig. 1), recognition of voiced target syllables was above the predicted value, and recognition of whispered-target syllables was below the predicted value ($F_{1,7}=39.9$, $p=0.001$, $\eta_p^2=0.84$). Prediction accuracy increased with increasing SNR ($F_{6,42}=8.4$, $p=0.002$, $\epsilon=0.40$, $\eta_p^2=0.55$), and this trend was more pronounced for whispered targets and less pronounced for voiced targets ($F_{6,42}=3.59$, $p=0.018$, $\epsilon=0.68$, $\eta_p^2=0.33$). Recognition performance was below the predicted value for whispered distracters and above the predicted value for voiced distracters ($F_{1,7}=59.6$, $p<0.001$, $\eta_p^2=0.89$), and this effect was entirely driven by the whispered-target condition ($F_{1,7}=11.8$, $p=0.011$, $\eta_p^2=0.63$). The results of this analysis are illustrated in Fig. 2, which also shows the direction of each effect.

There are two pronounced kinks in the results for the voiced targets at 0–3 dB in Fig. 1, panel A1 (and B1). However, paired comparisons of the prediction mismatch data showed that none of the corresponding troughs in Fig. 2, panel A1 (and B1), was significantly different from the neighboring points, so the kinks would appear to be just statistical fluctuations.

The pattern of consonant scores and vowel scores is similar to the pattern of syllable scores, except the interaction between target and distracter is only significant for vowel recognition ($F_{1,7}=20.8$, $p=0.003$, $\eta_p^2=0.75$; panel C2 in Fig. 2) and not for consonant recognition (panel B2 in Fig. 2). Furthermore, for the vowel data, the effect of SNR interacted not only with the voicing of the target ($F_{6,42}=5.83$, $p=0.012$, $\epsilon=0.36$, $\eta_p^2=0.46$) but also with the distracter ($F_{6,42}=4.72$, $p=0.011$, $\epsilon=0.51$, $\eta_p^2=0.40$). Moreover, target and distracter voicing and SNR showed a three-way interaction ($F_{6,42}=3.30$, $p=0.030$, $\epsilon=0.6$, $\eta_p^2=0.32$) indicating that the effect of SNR on the prediction accuracy of vowel recognition was entirely driven by the whispered-target/whispered-distracter condition (see panel C1 in Fig. 2).

IV. DISCUSSION

Overall, voiced syllables were better recognized than whispered syllables, and whispered distracters led to lower recognition performance than voiced distracters. It would appear that the listeners could use voicing to reject a distracter as well as to detect a target. However, the effect of voicing for distracters was only present when the target was whispered. Similarly, recognition of voiced syllables was well predicted by the audibility model, whereas for whispered-target syllables, recognition was lower than predicted, and more so at low SNRs. The fact that whispered syllables were less intelligible than voiced syllables corroborates previous studies on the perception of whispered speech (Tartter, 1991,1989).

The harmonic cancellation model suggests that the benefit of voicing to recognition should be greater for the distracter than the target. Thus, it was expected that there would be an asymmetry in which listeners suffer more from removal of voicing in the distracter than in the target. The interaction of target and distracter in the vowel recognition data of Fig. 2, panel C2, supports this prediction: The difference between the bars with solid-line hatching shows the drop in recognition performance associated with the removal of voicing in the target syllable, and the difference between the gray bars shows the drop in recognition performance associated with the removal of voicing in the distracter syllable. Since the difference between the bars with solid-line hatching is smaller than the difference between the gray bars, it could be argued that the results are compatible with the cancellation model. However, it is also the case that the effect of removing voicing in the distracter is only pronounced when the target itself is whispered. For voiced target syllables, there is no significant effect of removing voicing in the distracter for any of the scores, possibly

because there was already a cue to the target voice provided by its resonance scale. Vestergaard *et al.* (2009) previously showed that when there is a sizable difference in VTL between the competing voices, then the benefit of additional cues is diminished. Overall, while the data do not provide strong evidence for the cancellation hypothesis, they do not rule it out.

V. CONCLUSION

The main conclusion is that listeners use voicing whenever it is present, either to detect the target speech or to reject the distracter. This study also illustrates the importance of controlling for the effects of audibility in experiments with voiced and whispered speech. Direct interpretation of the recognition scores in the top panels of Fig. 1 would lead to an overestimation of the robustness of whispered speech. When the effects of audibility are included, as in the bottom panels of Fig. 1, the effect of voicing is observed to be considerably smaller than when performance is presented as a function of SNR. To wit, 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, especially below an audibility index of 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.

Acknowledgments

The research was supported by the United Kingdom Medical Research Council under Grant Nos. G0500221 and G9900369. We would like to thank James Tanner and Sami Abu-Wardeh for help with collecting the data, and Nick Fyson for assistance in producing the programs that ran the experiments.

References

- Abercrombie, D. Elements of General Phonetics. Edinburgh University Press; Edinburgh: 1967.
- ANSI. S3.5. Methods for the Calculation of the Speech Intelligibility Index. American National Standards Institute; New York: 1997.
- 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, 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.
- 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.
- 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.
- Lass NJ, Hughes KR, Bowyer MD, Waters LT, Bourne VT. Speaker sex identification from voiced, whispered, and filtered isolated vowels. *J. Acoust. Soc. Am.* 1976; 59:675–678. [PubMed: 1254794]

- Schneider P, Sluming V, Roberts N, Scherg M, Goebel R, Specht HJ, Dosch HG, Bleeck S, Stippich C, Rupp A. Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nat. Neurosci.* 2005; 8:1241–1247. [PubMed: 16116442]
- Schwartz MF. Syllable duration in oral and whispered reading. *J. Acoust. Soc. Am.* 1967; 41:1367–1369. [PubMed: 6074795]
- Schwartz MF. Air consumption, per syllable, in oral and whispered speech. *J. Acoust. Soc. Am.* 1968a; 43:1448–1449. [PubMed: 5659499]
- Schwartz MF. Identification of speaker sex from isolated, voiceless fricatives. *J. Acoust. Soc. Am.* 1968b; 43:1178–1179. [PubMed: 5648107]
- Schwartz MF. Power spectral density measurements of oral and whispered speech. *J. Speech Hear. Res.* 1970; 13:445–446. [PubMed: 5421464]
- Schwartz MF, Rine HE. Identification of speaker sex from isolated, whispered vowels. *J. Acoust. Soc. Am.* 1968; 44:1736–1737. [PubMed: 5702036]
- Sherbecoe RL, Studebaker GA. Regression equations for the transfer functions of ANSI s3.5–1969. *J. Acoust. Soc. Am.* 1990; 88:2482–2483. [PubMed: 2269742]
- Studebaker G. A “rationalized” arcsine transform. *J. Speech Hear. Res.* 1985; 28:455–462. [PubMed: 4046587]
- Tartter VC. What's in a whisper? *J. Acoust. Soc. Am.* 1989; 86:1678–1683. [PubMed: 2808917]
- Tartter VC. Identifiability of vowels and speakers from whispered syllables. *Percept. Psychophys.* 1991; 49:365–372. [PubMed: 2030934]
- Tartter VC, Braun D. Hearing smiles and frowns in normal and whisper registers. *J. Acoust. Soc. Am.* 1994; 96:2101–2107. [PubMed: 7963024]
- Thomas IB. Perceived pitch of whispered vowels. *J. Acoust. Soc. Am.* 1969; 46:468–470. [PubMed: 5804119]
- Thornton AR, Raffin MJM. Speech discrimination scores modeled as a binomial variable. *J. Speech Hear. Res.* 1978; 21:507–518. [PubMed: 713519]
- Traunmüller H, Eriksson A. Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.* 2000; 107:3438–3451. [PubMed: 10875388]
- Vestergaard MD. The effect of voicing, pitch and vocal tract length on the recognition of concurrent speech. *J. Acoust. Soc. Am.* 2007; 121:3200.
- Vestergaard MD, Fyson NRC, Patterson RD. The interaction of vocal characteristics and audibility in the recognition of concurrent syllables. *J. Acoust. Soc. Am.* 2009; 125:1114–1124. [PubMed: 19206886]
- Vestergaard, MD.; Ives, DT.; Patterson, RD. Identification of voiced syllables masked by whispered syllables and vice versa. *Proceedings of the British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness; Cardiff, Wales.* 2005. p. 58

FIG. 1. Recognition scores as a function of (1) SNR (top panels) and (2) SII (bottom panels). The left panels (A#) show syllable recognition, the middle panels (B#) show consonant recognition, and the right panels (C#) show vowel recognition. The black lines show performance for voiced target syllables and the dark gray lines show performance for whispered-target syllables. The solid lines are for voiced distracters and the dashed lines are for whispered distracters. In the bottom panels, the thick light gray curve shows predicted recognition according to the transformation by Sherbecoe and Studebaker (1990). See text for details.

FIG. 2.

The prediction mismatch for (A#) syllable, (B#) consonant, and (C#) vowel recognition scores in RAU. The top panels show (1) the interactions of vocal characteristics with SNR, and the bottom panels show (2) the interaction between target voicing and distracter voicing. The black lines and hatching are for the voiced target syllables and gray is for the whispered-target syllables. Solid lines and hatching are for voiced-distracter syllables and dashed lines and hatching are for whispered-distracter syllables.