UNIVERSITY OF CAMBRIDGE

CNBH, Physiology Department, Cambridge University
www.mrc-cbu.cam.ac.uk/cnbh

Centre for the Neural Basis of Hearing

The perception of scale in whispered vowels P36

David R. R. Smith and Roy D. Patterson

david.smith@mrc-cbu.cam.ac.uk, roy.patterson@mrc-cbu.cam.ac.uk

1

INTRODUCTION

When human listeners are given sequences of voiced vowels which differ systematically in the simulated vocal-tract length (VTL) of the speaker, listeners are capable of discriminating VTL differences of 6-10%, over a range of glottal-pulse rate (GPR) and VTL values much greater than that encountered in normal speech (Smith *et al*., 2005).

- The purpose of this study was to extend this research to the case of *whispered* vowels.

- Is discrimination performance for whispered vowels worse or better than for voiced vowels?

2

RATIONALE

Figure 1 (Panel 4) shows the Fourier spectra of the same vowel /a/ when voiced and when whispered.

The continuous spectrum of the broadband noise source in whispered speech provides a better definition of the spectral envelope and thus the transfer function of the vocal tract (Tartter and Braun, 1994).

Given that speaker size estimates should benefit from more information about the vocal tract, this implies that discrimination performance should *improve* for whispered vowels compared to voiced vowels.

However, if voicing is important for good performance, i.e. because glottal pulses are important in the construction of a stable representation of the resonance pattern (cf. Patterson *et al*., 1995), then discrimination performance should be *worse* for whispered vowels compared to voiced vowels.

3

RESULTS & CONCLUSIONS

The results from the change in VTL discrimination experiments show that, of the five places in the GPR-VTL plane tested, four showed no significant difference between the voiced and whispered vowel conditions (Figs 2-3, Panels 5-6). There was a significant difference only at the high GPR-short VTL position [320 Hz, 9.4 cm], where VTL discrimination performance with whispered vowels was significantly worse compared to performance for the voiced vowels.

- Listeners can discriminate changes in simulated VTL when these changes are carried by whispered speech.
- Discrimination performance for whispered vowels is comparable to performance for voiced vowels, except for at high GPR-short VTLs.

- Implication is that size information can be extracted from whispered vowels to inform perceptual decisions.

4

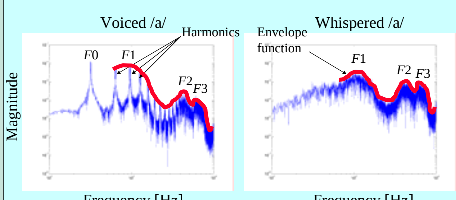

FIGURE 1. The Fourier spectra of the same vowel /a/ when voiced (left panel) and when whispered (right panel). The spectral envelope function is shown by the bold line which is defined by the relative level of the frequencies of the vowel. The formants (*F*1-*F*3) are defined as the peaks of this overlying function. The energy of the formants is carried by the harmonics of the fundamental frequency *F*0 for the voiced and by all frequencies for the whispered.

5

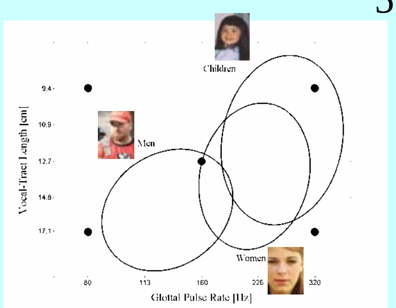

FIGURE 2. The JND for simulated-VTL discrimination was measured using a 2AFC paradigm with the method of constant stimuli at five different points in the GPR-VTL plane (cf. solid circles). The ellipses show estimates of the normal range of GPR and VTL values in speech for men, women and children (derived from an analysis of Peterson and Barney, 1952).

6

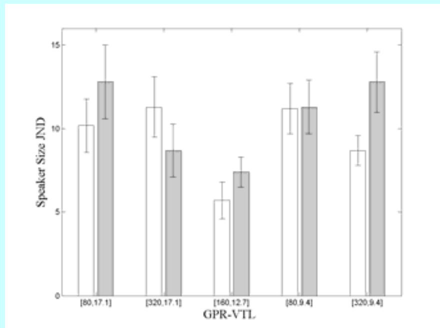

FIGURE 3. Speaker size JNDs for the voiced (light bars) and whispered vowel conditions (dark bars), across the five experimental points in the GPR-VTL plane (cf. Fig. 2). Error bars represent ± 1 standard error of the mean. There were five listeners. The JND is defined as the difference between the values associated with 50 per cent correct (equality match d'=0) and 76 per cent correct (d'=1 in this 2AFC task), relative to the perceived match of the standard, expressed as a percentage. Cumulative Gaussians were fitted to the psychometric function to find the 50 and 76 per cent points.

7

METHODS - I

CANONICAL VOWELS Vowels (/a/, /e/, /i/, /o/, /u/) were extracted from a natural /hVd/ speech sequence spoken by an adult male (RP) – *haad, hayed, heed, hoed, who'd*. Sounds were digitized with 16-bit quantification and a sampling rate of 44.1 kHz. All vowels were 400 ms.

SCALE MANIPULATION Vowels were manipulated to have a range of GPRs and simulated VTLs using STRAIGHT (Kawahara and Irino, 2005). STRAIGHT produces a pitch-independent spectral envelope that accurately tracks the motion of the vocal tract through an utterance. Once STRAIGHT has segregated a vowel into its GPR contour and a sequence of spectral-envelope frames, the vowel can be resynthesized with the spectral-envelope dimension (frequency) expanded or contracted, and the GPR dimension (time) expanded or contracted, and the operations are largely independent.

8

METHODS - II

VOICED and WHISPERED VOWELS Whispered vowels were created directly from the voiced-vowel exemplars, thus ensuring that spectral envelopes were matched across the whispered and voiced-vowel conditions. This was done within STRAIGHT: the voiced-vowel spectral envelope for each vowel type and manipulation level in GPR and VTL was isolated and then re-excited with white noise instead of glottal pulses. The spectral tilt of the whispered vowels was adjusted by lifting it by 6-dB per octave to emphasize the higher frequency components.

DISCRIMINATION EXPT We used a two-interval 2AFC paradigm with the method of constant stimuli. Six-point psychometric functions were measured with 25 trials per point. A trial consisted of two temporal intervals where each interval was composed of a sequence of 4 of the 5 vowels (chosen randomly without replacement), following one of four pitch contours (rising, dropping, up-down, down-up), with different start pitches and where the intensity of the vowels in each interval was roved relative to the other interval (over a 10-dB range). Listener task was to chose the interval in which the vowels were spoken by the smaller speaker. No feedback was given.

9

REFERENCES

Patterson, R.D., Allerhand, M.H., and Giguere, C. (1995). *J. Acoust. Soc. Am.* **98**, 1890-1894.

Peterson, G.E., and Barney, H.L. (1952). *J. Acoust. Soc. Am.* **24**, 175-184.

Kawahara, H., and Irino, T. (2005). In *Speech separation by humans and machines*, P. Divenyi (Ed.), Kluer Academic. Massachusetts, 167-180.

Smith, D.R.R., Patterson, R.D., Turner, R., Kawahara, H., and Irino, T. (2005). *J. Acoust. Soc. Am.* **117**, 305-318.

Tartter, V.C., and Braun, D. (1994). *J. Acoust. Soc. Am.* **96**, 2101-2107.

ACKNOWLEDGEMENTS

Research supported by the MRC (G99003639 and G9901257) and the German Volkswagen Foundation (VWF 1/79 783).