Proceedings of the Institute of Acoustics

THE PERCEPTION OF 'VOICING' IN WHISPERED STOP CONSONANTS

Kevin L. Baker (1) & Sandra P. Whiteside (2)

(1) Department of Human Communication, De Montfort University, Leicester LE7 9SU, U.K.
(2) Department of Speech Science University of Sheffield, Sheffield S10 2TA, U.K.

1. INTRODUCTION

From the few studies that have been carried out into whispered speech, it is apparent that listeners have little difficulty in perceiving vowels. Kallail and Emmanuel [1, 2] presented lengthened and isolated vowels to their subjects and reported that between 63 and 65% were correctly identified when whispered compared to 80% when spoken normally. Tartter [3] improved on this study by presenting whispered CVC syllables following observations by Strange [4] that in normal speech, formant transitions are important for vowel identification. Tartter [3] found a better than 80% identification rate for 10 vowels whispered by 6 speakers compared to over 90% for the normally voiced vowels.

Research into whispered consonants is relatively scarce. Tartter [5] presented whispered consonant- [a] syllables to 6 listeners and found that the overall identification score was 64% with a 72% accuracy for identifying 'voicing'. However, given that her data included other consonants in addition to stops it is difficult to ascertain the levels of accuracy for the stop consonants alone. Dannenbring [6] investigated 12 subjects' ability to discriminate between whispered consonants in CV syllables, where the vowel was either /i/, /a/ or /u/. Dannenbring's results show that listeners were able to make discriminations with confidence but does not provide correct identification scores which makes his results difficult to compare with other studies. Munro [7] presented 32 whispered tokens of /p/ and /b/ in four vowel contexts to 8 listeners where he found an overall mean correct identification score of 63%. Although he showed that the whispered /b/ tended to have a steeper rise slope than the whispered /p/, these showed no relationship to the pattern of identifications. He concludes that it is dangerous to make 'inferences about perceptual mechanisms on the basis of production data alone' (p.180).

The present study is intended as an preliminary investigation into whether 'voicing' contrasts in whispered stop consonants (or plosives) in words can be identified in the absence of both the laryngeal voice source and meaningful contextual information (e.g. a meaningful sentential framework. Vowel context was not controlled for. Instead, the focus of this study was placed upon the place of articulation of the whispered stop consonants in initial and final position in minimal word pairs. Listeners were asked to make judgements about whether whispered stop consonants in word initial and word final position were 'voiced' or 'voiceless'. Subsequently acoustic measurements were taken for the word initial and word final stimuli.

2. METHOD

We presented whispered stimuli to subjects who were given a forced choice for their identification. The forced choice was between the presented stop consonant and its 'voiced' or 'voiceless' counterpart. For example, if the whispered token A PAT was presented to the subject, then the word pair A PAT and A BAT was visually presented as the choice for identification.

2.1 Subjects

The authors served as speakers for the recorded speech samples. Both speakers are native speakers of British English and are in their late twenties. Five female and five male volunteers served as subjects for the perceptual part of the study. All listeners reported normal hearing and were also native speakers of British English. Their ages ranged from 20 to 34 years.

2.2 Stimuli

The speech samples consisted of 55 CVC whispered words in the frame 'a CVC'. They were recorded once both by an adult female speaker (F) and an adult male speaker (M). The 55 words are shown in table 1 and form 30 minimal word pairs for stop consonants in word initial position and 30 minimal word pairs for stop consonants in word final position (5 of the words are used in more than one pairing). The minimal word pairs represented bilabial, alveolar and velar places of articulation.

	Bilabial		Alveolar		Velar	
Word Initial	a pat	a bat	a tip	a dip	a cod	a god
	a peat	a beat	a tab	a dab	a cold	a gold
	a pack	a back	a tuck	a duck	a cape	a gape
	a pay	a bay	a tart	a dart	a coat	a goat
	a pig	a big	a toe	a doe	a cap	a gap
Word Final	a lap	a lab	a pat	a pad	a lack	a lag
	a tap	a tab	a lit	a lid	a tack	a tag
	a swap	a swab	a fat	a fad	a back	a bag
	a cop	a cob	a sort	a sword	a tuck	a tug
	a cap	a cab	a lout	a loud	a rack	a rag

Table 1 Whispered word-pairs

The 60 words were repeated once and then put into a randomly ordered list. A random selection of 30 words was taken from the original list to form a 'dummy' list. Finally, the two lists were mixed for presentation in the following order: 10 dummies, 30 test words, 5 dummies, 30 test words, 5 dummies, 30 test words, 5 dummies, 30 test words, 5 dummies. Judgments for the dummy list were not used in the analysis in order to ignore 'listing' effects and lack of attention.

2.3 Recording

Each whispered word was recorded while the speaker was seated in a sound proof chamber. The whispered speech was recorded digitally using an Apple Macintosh Classic II computer via a microphone connected to a Farallon MacRecorder™. The sampling rate was set at 22kHz (8 bit). The MacRecorder digitizer filtered the analogue sound with a cut off of 11 kHz.

2.4 Perception Tests

Subjects were seated in the sound proof chamber with a loudspeaker and a computer 'mouse'. Outside the chamber the Apple Macintosh was placed in view of the subject through a window in the chamber, and connected to the mouse. A Hypercard™ (Apple Computer Inc., 1990) program written by the first author, was used to play the speech samples from the stimuli list, present the appropriate word pair on the computer screen, and to record the judgements made by the subjects. The subjects were told to click the mouse on the computer screen in order to hear a word through the loudspeaker. They were asked to click as many times as they wished until they could identify the whispered word from the word pair shown on the computer screen and then to record their choice by clicking the mouse on the judged word. Each subject repeated the experiment so that they made judgements of both the male and female speech stimuli.

2.5 Acoustic Analysis

Possible acoustic cues to the perception of 'voicing' were investigated for the whispered stop consonants. These acoustic cues were examined using a KAY Computerised Speech Lab (CSL) Model 4300. The whispered speech stimuli were transferred from the Apple Macintosh computer onto digital audio tape and then transferred on to the KAY CSL using a sampling rate of 10 kHz. The methods of analysis used for each of the measurements are outlined below.

For the word-initial stimuli the following measurements were taken: i) The amplitude of the plosive burst using the graphical results of an algorithm which computes an energy envelope in dB SPL from the speech pressure waveform of the whispered speech sample; ii) The interval between the peak amplitude of the plosive burst and the peak amplitude of the following noise-excited vowel from the computed energy envelope (dB SPL) using the graphical interface provided by the CSL; iii) The amplitude difference between the peak of the burst and the following vowel. This again was done using a similar method as for i) and ii). The results of these analyses can be found in figure 2 below.

For the word-final stimuli the measurements taken were: i) the frequency of the first formant (F1) offset preceding the closure for the word-final stop consonant, using an FFT wideband spectrogram and a graphical interface which allows the measurement of formant frequency values; ii)The duration of the noise-excited vowel preceding the closure, given that for post-vocalic stop consonants one of the acoustic cues of voicing is the duration of the preceding vowel, where a shorter vowel duration cues voicelessness [8]. This was done using the FFT spectrograms and the graphical interface. The duration of the vowel was taken from the point immediately following the plosive burst of the preceding plosive until the acoustic closure for the final plosive. So for example, for the stimulus A PAT (/' 'pœt/) the duration of /œ/ would be taken immediately following the plosion of /p/ until the acoustic closure for /t/. The results of these analyses can be found in figure 3 below.

3. RESULTS AND DISCUSSION

3.1 Perception Tests

Graphic representations of the results of the perception tests are shown in figure 1 below. Table 2 provides a summary of these results. χ_2 tests were carried out on the identification scores with the assumption that the expected identification of the consonants would be at chance level (i.e. 50%). These results are given in Table 3.

From figure 1 and summary table 2 we can see that the word initial stop consonants were correctly identified with levels of accuracy ranging from 0% to 100% giving an overall mean identification score of 77%. However, if we look at the results in more detail we find a variation in the identification results for each place of articulation, For example the 'voiceless' alveolar stimuli are identified for both the male and female stimuli with most accuracy (mean of 98.5%). In addition, the 'voiced' bilabial stimuli for the male speaker are identified with the least level of accuracy (41%) followed by the 'voiced' velar stimuli of the female speaker (42%). However, any conclusions about gender differences can only be tentative with this data coming from two speakers. What is evident from table 2 is that the 'voiced' word-initial stimuli are identified with lower levels of accuracy compared with their 'voiceless' counterparts, a finding also made by Tartter [5].

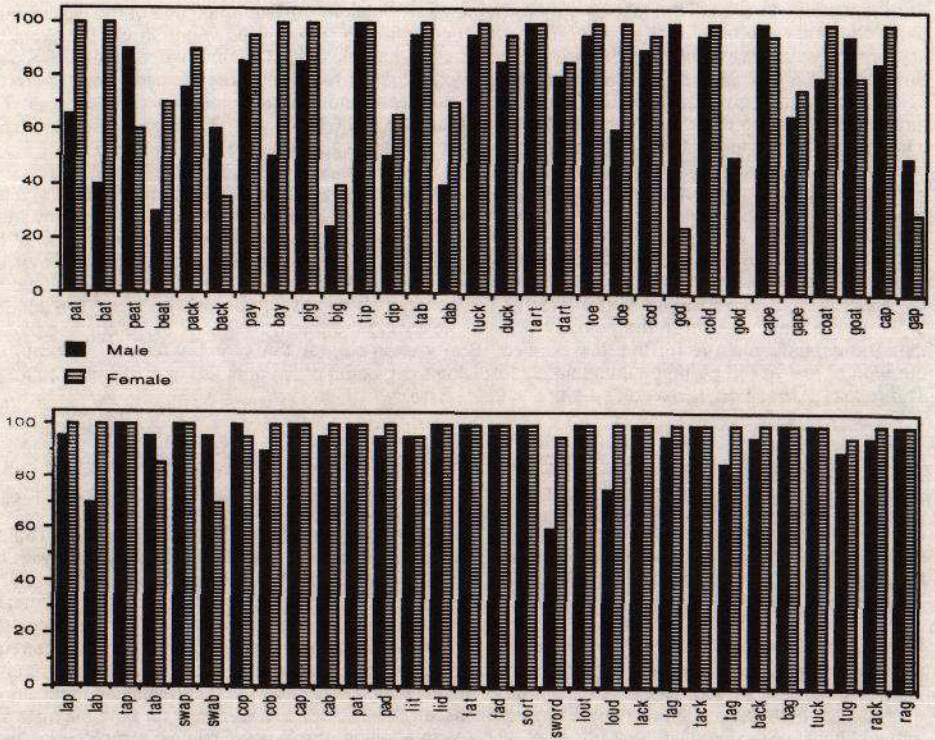

Figure 1: Perception scores for word intitial and word final consonants (bilabial, alveolar, velar), male and female.

For the word final whispered stop consonants the number of stimuli correctly identified ranged from 60% to 100% with an overall mean identification score of 96%, much higher than the word initial scores. These findings suggest that the listeners had little trouble identifying whispered stop consonants in word final position. The correct identification of the word initial and word final whispered consonants are signifcantly above the chance level of 50% expected if identification of the consonants was based on 'voicing' which of course is absent in our stimuli.

	Bilabial		Alveolar		Velar	
	-v /p/	+v /b/	-v /t/	+v /d/	-v /k/	+v /g/
(M%, F%) Mean (%) word intial	(80, 89) 84.5	(41, 69) 55	(97, 100) 98.5	(63, 83) 73	(90, 98) 94	(72, 42) 57
(M%, F%) Mean (%) word final	(99, 99) 99	(89, 91) 90	(99, 99) 99	(86, 99) 92.5	(98, 100) 99	(94, 99) 96.5

Table 2: Summary table of perception scores (% correct).

	Bilabial		Alveolar		Velar	
	Male	Female	Male	Female	Male	Female
Word intial	122.5	289	268	377.5	260	332
Word final	402	423	395.5	481	428	490.5

All values are $p \leq 0.0001$

Table 3: χ^2 values for perception scores assuming chance level of 50%.

3.2 Acoustic Analysis

Graphical representations of the acoustic analyses are given in figures 2 and 3. Results of a set of two-tailed t-tests on the acoustic measurements of the 'voiced' and 'voiceless' word initial stimuli and word final stimuli are given in tables 6 and 8 respectively.

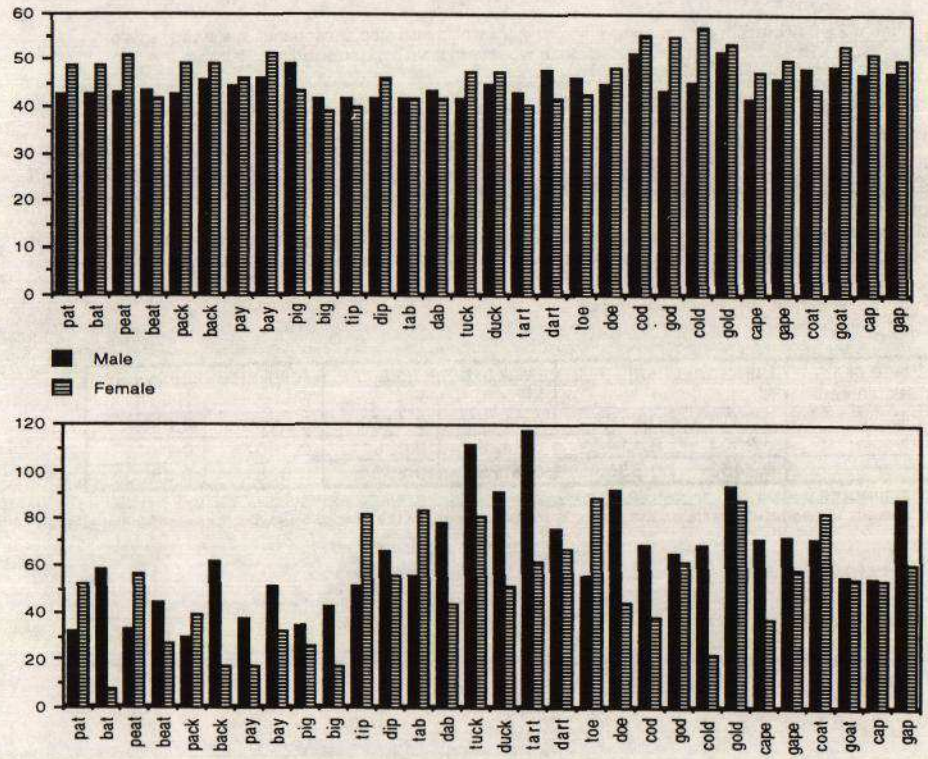

Figure 2 i) & ii): Acoustic measurements for word-initial whispered stop consonants, male and female.
i) Top: Amplitudes of plosive burst (dB SPL). ii): Peak (of plosive) to peak (vowel) duration (ms).

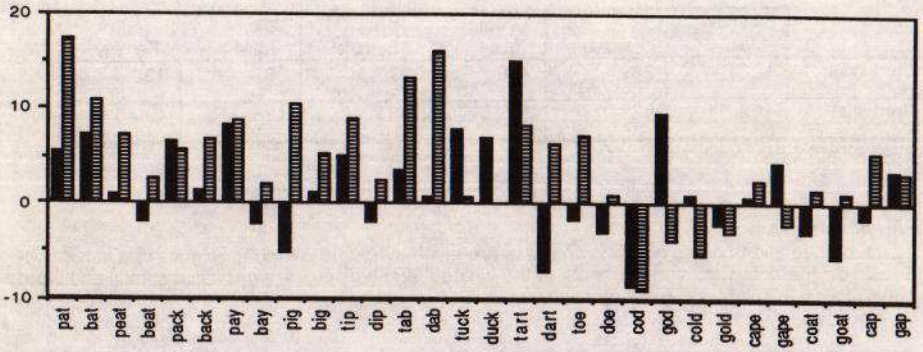

Figure 2 iii): Acoustic measurements for word-initial whispered stop consonants, male and female.
iii): Amplitude difference (vowel - plosive burst, dB SPL).

From table 4 we can see that for the 'burst amplitude' no significant differences were found between the 'voiced' and 'voiceless' stimuli. For the 'peak to peak duration' measurements only the bilabial stimuli of the male speaker and the alveolar stimuli of the female speaker showed significant differences between the 'voiced' and 'voiceless' stimuli. Finally, for the 'amplitude differences' only the female speaker's bilabial stimuli showed significant differences. The non-consitent nature of these analyses lies at odds with the significantly above chance levels of correct identification for these stimuli shown in tables 2 and 3.

Place of articulation	Burst amplitude		Peak to peak duration		Amplitude difference	
	M	F	M	F	M	F
Bilabial	0.8318	0.5183	0.0178*	0.1469	0.5103	0.0377*
Alveolar	0.2004	0.1144	0.8825	0.0351*	0.1607	0.2443
Velar	0.7453	0.5305	0.4175	0.2994	0.3281	0.3806

* significant $p<0.05$

Table 4: Results of two-tailed t-tests on the acoustic measurements of the voiced and voiceless word-initial stimuli.

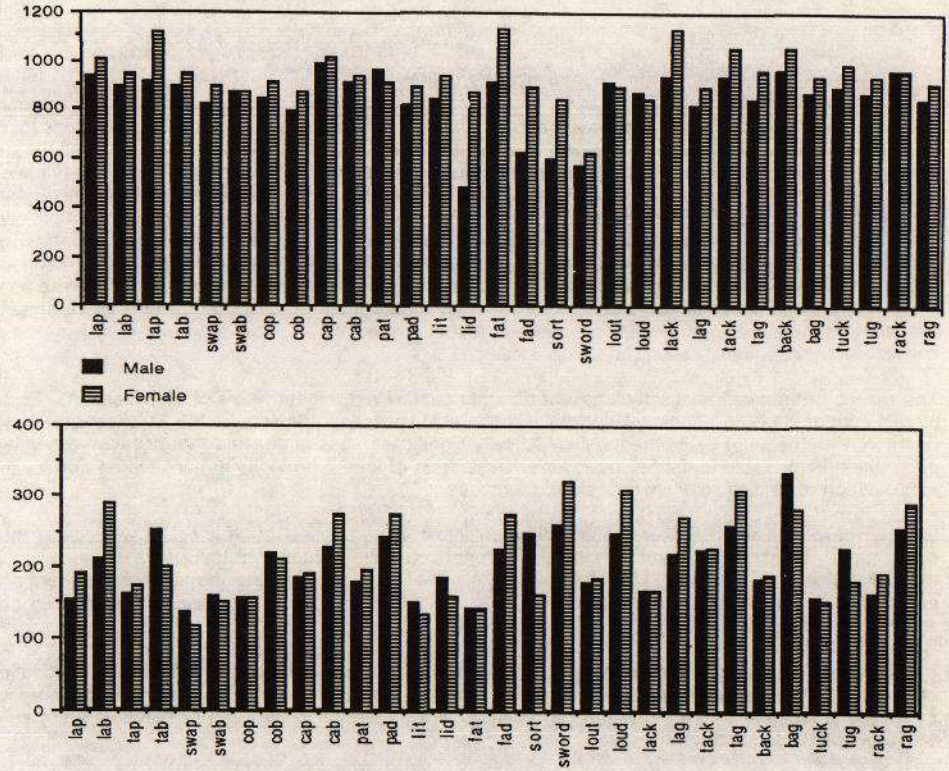

Figure 3: Acoustic measurements for word-final whispered stop consonants, male and female.
i) top: F1 offset (Hz). ii) bottom: Vowel duration (ms).

From table 5 we can see that for 'vowel duration' significant differences were found between all the 'voiced' and 'voiceless' stimuli. However, for 'F1 offset' significant differences were only found for the male bilabial and female bilabial and female alveolar stimuli. If we relate these findings to the perception scores it would appear that vowel duration is the more robust of the two cues analysed in this study. In fact the overall mean values of vowel duration between the 'voiced' and 'voiceless' stimuli are 245.2 ms and 173.2 ms giving a difference of 72 ms and a ratio of 1.42 in favour of the 'voiced' stimuli.

Place of articulation	F1 offset M	F	Vowel duration M	F
Bilabial	0.0408*	0.0352*	0.007**	0.0114*
Alveolar	0.0558	0.0066**	0.0152*	0.012*
Velar	0.0589	0.2355	0.0164*	0.0037**

** significant $p<0.01$
* significant $p<0.05$

Table 5: Results of two-tailed t-tests on the acoustic measurements of the voiced and voiceless word-final stimuli.

4. CONCLUSIONS

The findings of this preliminary experiment appear to agree to some extent, with previous findings [5, 6, 7] that listeners are able to identify 'voicing' in whispered stop consonants in initial position, without any sentential context. However, in this study it was found that although identification was above chance levels, some of the data suggested variability in the correct identification of bilabial word-initial consonants. In addition, these showed no relationship to most of the acoustic measurements made. This supports Munro [7] who states that production data needs to be interpreted with caution when considering perceptual mechanisms. It is suggested here that this 'caution' needs to be exercised particularly when we are dealing with acoustic measurements which are brief and transient in nature as in the syllable initial consonants. In addition, the insignificant findings of the acoustic measurements indicate that the acoustic cues were not reliably present in the data and may therefore explain some of the variability in the identification scores.

The results of the word final stimuli gave a different picture with higher levels of identification. These results appeared to show some relationship with the acoustic measurements for the F1 offset and vowel duration, with the latter being the most statistically significant. It is suggested that because the vowel information is available to the listener over a long period of time it provides a more robust cue for the identification of 'voicing' for word final stop consonants.

Clearly there is a need for further investigation into the perception of 'voicing' in whispered stop consonants. This should include for example: i) Controlled vowel context which appears to play a role in the perception of 'voicing' [6, 7]; ii) Multiple repetitions of data; iii) A greater number of speakers and listeners with similar linguistic backgrounds; iv) Speech data with stop consonants in stressed and unstressed contexts.

In this study we have concentrated on relating acoustic measurements in relation to production contrasts. Further research and experimentation is planned in which we intend to incorporate the above suggestions together with a more in-depth analysis of the correlation between possible acoustic cues and perceptual identification.

5. REFERENCES

[1] Kallail, K. L. and Emmanuel, F. W. (1984a). An acoustic comparison of isolated whispered and phonated vowel samples produced by adult male subjects, *Journal of Phonetics, 12*, 175-186.
[2] Kallail, K. L. and Emmanuel, F. W. (1984b). Formant frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects, *Journal of Speech and Hearing Research, 27*, 245-251.
[3] Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception and Psychophysics, 49*, 365-372.
[4] Strange, W.(1989) Evolving theories of vowel perception, *Journal of the Acoustical Society of America, 85*, 2081-2087.
[5] Tartter, V. C. (1989). What's in a whisper? *Journal of the Acoustical Society of America, 86*, 1678-1683.
[6] Dannenbring, G. L. (1980). Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills, 51*, 979-985.
[7] Munro, M. J. (1990). Perception of 'voicing' in whispered stops, *Phonetica, 47*, 173-181.
[8] Kent, R. D. and Read, C. (1992). *The Acoustic Analysis of Speech*, Whurr Publishers.