THE ACOUSTIC CHARACTERISTICS OF WHISPERED PLOSIVES AND THEIR RELIABILITY FOR THE PERCEPTION OF 'VOICING'

Sandra P. Whiteside (1) & Kevin L. Baker (2)
(1) Speech Science, University of Sheffield, Sheffield S10 2TA, U.K.
(2) Department of Human Communication, De Montfort University, Leicester LE7 9SU, UK.

ABSTRACT

This study examines a range of acoustic characteristics of whispered plosives in initial and final position in minimal word pairs produced by two speakers. Perception tests are carried out to see whether listeners can make judgements about whether the whispered stimuli represent 'voiced' or 'voiceless' tokens. Whether the acoustic characteristics of the whispered stimuli are reliable for the perception of 'voicing' is discussed.

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.

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.

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.

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 subjects with normal hearing served as subjects for the perceptual part of the study. All listeners native speakers of British English with an age range of 20 to 34 years.

Stimuli

The speech samples consisted of 55 CVC whispered words in the frame 'a 'CVC'. This frame was used to produce even stress. They were recorded once

both by an adult female speaker (F) and an adult male speaker (M). The stimuli 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 once). The minimal word pairs represented bilabial (B), alveolar (A) and velar (V) places of articulation.

Table 1. Whispered minimal word-pairs

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

The 60 words were repeated once and randomised into a list for recording.

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.

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 second 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 used a computer mouse to play back the stimuli and make their forced choices about the stimuli they were presented with. Each subject repeated the experiment so that they made judgements of both the male and female speech stimuli.

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 (DAT) 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 (dB) 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 (ms) 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 (dB) between the peak of the burst and the following vowel. This was done using a similar method as for i) and ii); iv) The closure duration (ms) of the initial plosive measured from the end of the preceding schwa (/ə/ to its release; v) The release phase (ms) of the initial plosive, measured from the point of the plosive's release to the onset of F1 in a wide band FFT spectrogram and vi) The overall energy (SPL dB) of the CVC using the same method as measures i) to iii). The statistical results of these analyses can be found in table 4 below.

For the word-final stimuli the following measurements taken were: i) the frequency (Hz) of the first formant (F1) offset preceding the closure for the word-final plosive, using an FFT wideband spectrogram and a graphical interface which allows the measurement of formant frequency values; ii) The duration (ms) of the noise-excited vowel preceding the closure, given that for post-vocalic plosives 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 (/ə 'pæt/) the duration of /æ/ would be taken immediately following the plosion of /p/ until the acoustic closure for /t/; iii) The duration (ms) of the acoustic closure following the vowel and preceding the final release of the plosive; iv) The energy (SPL dB) of the release burst of the final plosive using the method described above and v) The overall energy (SPL dB)of the CVC as described above. The statistical results of these analyses can be found in table 5 below.

RESULTS AND DISCUSSION

Perception Tests

Table 2 provides a summary of perception test 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 table 2 we can see that the mean correct perception scores range from 41% to 100%. This represents an overall mean of 77% and 96% for the word initial and word final stimuli respectively. 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%). 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].

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 significantly above the chance level of 50% expected if identification of the consonants was based on 'voicing' which of course is absent in our stimuli (see table 3).

Tables 4 and 5 shows the t-scores and their significance levels for the acoustic parameters of the 'voiced' and 'voiceless' pairs of word initial and word final stimuli

Table 2. Summary Table of Perception Scores (%)

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

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

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

All values are $p \leq 0.0001$.

respectively. We can see from these scores that of the 36 t-scores for the word initial stimuli, only 10 (5 bilabial, 3 alveolar and 2 velar) are statistically significant, whereas for the word final stimuli 16 out of the 30 scores are statistically significant. The latter findings lend some support to the better perception scores for the word final stimuli. However, although the acoustic data for the word initial stimuli show less statistical significance, the χ^2 values given in table 3 are statistically significant. This suggests that there was enough information in the whispered plosives for listeners to make accurate judgements about 'voicing' in the absence of laryngeal voicing.

From table 4, we can see that significant differences between the 'voiced' and 'voiceless' tokens were variable and patchy. For example, highly significant differences were found for both speakers in the 'release phase' of the bilabial stimuli. However, for the 'burst amplitude' parameter, no significant differences were found between the 'voiced' and 'voiceless' stimuli . The findings suggest that different acoustic cues in the whispered stimuli may be operational for different tokens and different speaker characteristics in the perception of 'voicing' in plosive-initial position.

Table 5 also shows variation in the levels of statistical significance for all the stimuli, however it also shows that the vowel duration and duration closure parameters show significant differences for all the 'voiced' and 'voiceless' stimuli. This suggests that these parameters may be playing a key role in the perception of voicing. Given that the vowel and closure duration preceding the final plosive are available to the listeners for a longer period of time, it is probable that these acoustic characteristics are serving as a robust cues in the perception of voicing .

However one must also bear in mind that there may be other acoustic cues operating in the perception of voicing for the word-initial stimuli that we have not considered in this study. Further research is planned in this area.

Table 4. T scores for Word Initial Acoustic Parameters .

Ac. Param.	Bilabial	Alveolar	Velar
Burst Amp.			
M	0.227	-1.532	-0.348
F	0.708	-2.013	-0.686
Peak to Peak Dur.			
M	3.885*	0.157	0.903
F	-1.796	-3.131*	1.191
Amp. Diff.			
M	0.722	1.719	-1.113
F	3.059*	1.364	0.985
Closure Dur.			
M	1.444	3.258*	-1.135
F	-0.61	-2.011	-5.63**
Rel. Phase Dur.			
M	17.77**	0.333	9.66**
F	13.51**	5.44**	2.031
Overall Energy			
M	-1.363	-2.607	1.574
F	-3.721*	1.114	0.459

* significant at $p \leq 0.05$,
** significant at $p \leq 0.01$

Table 5. T scores for Word Final Acoustic Measures.

Ac. Param.	Bilabial	Alveolar	Velar
F1 offset			
M	-1.395	-2.619	-5.17**
F	-2.978*	-2.671	-3.13*
Vowel Dur.			
M	5.1**	4.07*	3.983*
F	4.433*	4.362*	6.066**
Energy of Rel.			
M	-0.723	0.843	-0.657
F	1.026	2.479	1.646
Dur. of Closure			
M	17.39**	4.172*	7.665**
F	7.621**	6.022**	6.491**
Overall Energy			
M	1.62	-0.627	0.065
F	3.566*	-0.75	1.156

* significant at $p \leq 0.05$,
** significant at $p \leq 0.01$

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*, vol. 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*, vol. 27, 245-251.

[3] Tartter, V. C. (1991). Identifiability of vowels and speakers from whispered syllables. *Perception and Psychophysics*, vol. 49, 365-372.

[4] Strange, W.(1989) Evolving theories of vowel perception, *Journal of the Acoustical Society of America*, vol. 85 2081-2087.

[5] Tartter, V. C. (1989). What's in a whisper? *Journal of the Acoustical Society of America*, vol. 86, 1678-1683.

[6] Dannenbring, G. L. (1980). Perceptual discrimination of whispered phoneme pairs. *Perceptual and Motor Skills*, vol. 51, 979-985.

[7] Munro, M. J. (1990). Perception of 'voicing' in whispered stops, *Phonetica*, vol. 47, 173-181.