EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

QUALITY OF A VOWEL WITH FORMANT UNDERSHOOT: A PRELIMINARY PERCEPTUAL STUDY

Shinichi TOKUMA
Phonetics Laboratory, Sophia University
7-1, Kioi-cho, Chiyoda
Tokyo 102 Japan
e-mail: s-tokuma@hoffman.cc.sophia.ac.jp

ABSTRACT

In this study vowels in /CVC/ environments are compared with steady state vowels to investigate the perceived vowel quality change caused by undershoot. This study uses a perceptual task, whereby listeners match constant /CVC/ stimuli of /bVb/ or /dVd/ to variable /#V#/ stimuli, using a schematic grid on a PC screen. The grid represents an acoustic vowel diagram, and the subjects change the F1/F2 frequencies of /#V#/ by moving a mouse. The main results of the study show that while subjects referred to the trajectory peak of the /CVC/ stimuli in vowel quality perception, their performance was also affected by the formant trajectory range of the stimuli. When the formant trajectory range was small, they selected a value between the edge and peak frequencies, while they selected a value outside the trajectory range when it was large.

10.21437/Eurospeech.1997-571

1. INTRODUCTION

This study focuses on the perceived quality of vowels with formant undershoot. Although the phonological perception of vowels showing formant undershoot has been frequently investigated, the phonetic vowel quality change caused by the formant undershoot has received little previous attention. Since it is impossible to investigate all the acoustic parameters involved in the dynamics of vowel formants, this study concentrates on how the coarticulation of a vowel with neighbouring segments shows different vowel qualities, using a synthetic /CVC/ whose vowel formant values are found from acoustic analysis. The two formants of the /CVC/ are dynamic. Furthermore, an interactive matching experiment scheme equipped with a grid-display, a modified version of the experiment performed by Nord (1986), is introduced in this study.

2. EXPERIMENT

2.1 Materials

First the reference material was synthesised. This reference material consisted of a vowel with a dynamic trajectory simulating the /CVC/ pattern according to the formula devised by Nearey (1989). The consonants /b/ and /d/ were selected for consonants in /CVC/, and for vowels, of all RP short monophthongs, the four vowels /ɛ, æ, ɒ, ʊ/ were selected. The total duration of this /CVC/ (120 ms) and the formant frequencies of /CVC/ syllable nuclei were obtained from an acoustic analysis. The formant values are shown in Table 1 below.

	ɛ	æ	ɒ	ʊ
b-b	542/1806	760/1619	545/1009	491/1122
d-d	526/1848	733/1621	565/1135	437/1359

Table 1: Input peak values obtained from the acoustic study. The values indicated are; F1/F2. All values in Hz.

Vowels in these /CVC/ syllables had only two formants, F1 and F2, and they were synthesised using the parallel formant JSRU synthesiser by Holmes (1985) implemented in the Speech Filing System running on a Sun SPARC workstation. In the synthesis, F0 declined linearly from 130 Hz to 100 Hz. To ensure the reliability of the experiment, the actual formant frequencies of the output /CVC/ syllables were confirmed by obtaining a spectral section of the durational midpoint and measuring the formant centre frequencies. To address a potential criticism that the two-formant stimuli may not be easily identifiable or natural, three native speakers of South-East British English with phonetic training were asked to judge all types of the synthesised /CVC/ tokens by listening to them through headphones twice. The subjects all agreed that the stimuli all had acceptable quality of synthesised speech although two of them remarked on the unnaturalness of the /d/ in /dVd/ tokens synthesised according to Nearey's formula. Subsequently, the /dVd/ tokens in this experiment were modified by the addition of an initial intense burst and a final voiceless release, which improved the naturalness of the /d/ segments. This modification was accepted positively in the second informal survey on the stimulus quality.

Also the test /#V#/ materials were created using the JSRU synthesiser: short monophthongs in isolation, whose formant frequencies could be modified by subjects in an interactive mode.

2.2 Subjects

15 native speakers of South East British English participated in this experiment. They were undergraduate students of BA in Linguistics, or BSc in Speech Science/Speech Communication, of University College

London. They had no history of hearing problems. They had taken several phonetics/phonology courses and at least one course involving phonetic ear-training sessions, and that fact assured that they brought adequate background knowledge to the task in this experiment. For attendance over the whole experimental period, each subject was paid four pounds. They were not informed of the nature and aim of this experiment before its end.

2.3 Procedure

The procedure of this experiment was as follows: individual subjects were asked to sit in front of a PC terminal. Its screen displayed a schematic grid (6 x 6 blocks) with a cursor on one of these blocks, showing the "relative" position of the test token. In fact the grid was an acoustic vowel diagram, with F1/F2 = 0.5 Bark step, but the subjects were not informed of this. Sitting in front of a PC terminal showing a 6 x 6 grid, the subjects were required to match the vowel quality of the reference /CVC/ and the test /#V#/, as the latter changed its F1/F2 according to the cursor position on the grid, which was moved by clicking with a mouse. Each time the cursor was moved, or when a space key was pressed, a pair of a test token and a reference token having an interval of 300 ms between them was replayed through a speaker. These two tokens were played in the order of 'reference'-'test'. The allocation of the direction of F1/F2 on the two axes was randomised. There was no particular block whose F1/F2 values exactly corresponded to the peak F1/F2 values of /CVC/ as before. The grid cell with values closest to the F1/F2 peak values of the reference /CVC/ was randomly assigned to one of the central 4 x 4 cells.

There is a potential criticism of this interactive grid-matching scheme: whether subjects are really able to cope with this task and can really tune into the vowel quality and make judgements finer than phonological categories. Two pilot experiments of Tokuma (1995) and Tokuma (1996) examined its validity. Tokuma (1995) and Chapter 4 of Tokuma (1996) report on the results of the experiments. The results showed that while subjects referred to the trajectory peak of the /CVC/ stimuli in vowel quality perception, their matching process of F2 was affected by the low F1 frequency values, and also that the F1 trajectory of the reference /CVC/ stimuli may affect the F1 matching of the test /#V#/.

The whole experimental process carried out on a terminal screen was programmed in C-language by Mark Huckvale, University College London. The experiment was held in the teaching laboratory of Wolfson House, Department of Phonetics, University College London to accommodate more than one subject at each time slot. The laboratory was kept quiet by removing sources of background noise as much as possible, and subjects listened to the stimuli through covered-ear headphones. None of them reported that their attention had been compromised by background noise.

2.4 Results and Discussion

First of all, it was found that each of the 15 subjects finished the 48 sessions within 45 minutes. All subjects claimed that the experimental task was manageable. It was found that one subject had all the responses fall within the centre 4 x 4 blocks, and that implies that the subject probably adopted the strategy of selecting the visual centre-4 blocks on the screen without involving auditory judgement. Therefore the subject was excluded from the analysis. It was also shown that the responses of other three subjects produced a larger F1 x F2 range than 3 x 3 in 6 or 7 tokens types, which suggests a quasi-random response by these subjects (under the assumption that all consistent responses should target one particular block, with one step up/down as an error range). They were also therefore excluded.

Then, to examine the homogeneity of the subjects, Repeated Measures ANOVA was carried out for each formant number (i.e. F1 / F2), with factors of factors of [subject] (11-levels), [consonant] (2-levels), [vowel] (4-levels) and [trial] (6-levels). The significance level was set to 1 %., and the analysis was made separately on each formant. With regard to F1, the analysis showed that [subject] as a main factor was significant ($F(10,150)= 3.01$, $p<.01$). The analysis of F2 matching also confirmed a difference between subjects: subject as a main factor was significant ($F(10,150)= 3.06$, $p<.01$). Therefore the process of grouping subjects was carried out for each formant type across all vowel types, and eventually created two subject groups for F1 and two groups for F2. The two subject groups for F1 are henceforth called Group A (8 subjects) and Group B (3 subjects) and the two subject groups for F2 are called Group X (9 subjects) and Group Y (2 subjects). Details of this process are discussed in Chapter 5 of Tokuma (1996).

The result of a pilot matching experiment between /CVC/ and /#V#/ in Tokuma (1995) and Tokuma (1996) suggests that the F1 trajectory range of the reference /CVC/ may affect the F1 matching of the test /#V#/. Hence the relations between the formant trajectory range of the reference /CVC/ and the matched frequency of the test /#V#/ were also investigated for both F1 and F2. Tables 2 and 3 show the relations between trajectory ranges and the mean shift index to the trajectory peak (written as MI). Table 2 is for F1 results, Groups A and B, and Table 3 for F2 results, Groups X and Y. In each formant / consonantal environment, the order of the vowels is arranged so that the /CVC/ trajectory range increases from left to right. Also the mean shift indices of a group with the larger subject number (i.e. Group A for F1 and Group X for F2) are plotted in Figures 1 and 2, together with error bars of one standard deviation. All numbers are in Bark.

F1 Group A

	bʊb	bɛb	bɒb	bæb
range	3.28	3.70	3.73	5.37
MI	-.46	-.48	.09	.11
	dʊd	dɛd	dɒd	dæd
range	2.80	3.57	3.89	5.18
MI	-.31	-.46	.06	.39

F1 Group B

	bʊb	bɛb	bɒb	bæb
range	3.28	3.70	3.73	5.37
MI	-.75	-.97	.25	.31
	dʊd	dɛd	dɒd	dæd
range	2.80	3.57	3.89	5.18
MI	-.25	-.61	.19	.64

Table 2: Trajectory range of F1 /CVC/ and its mean shift index.

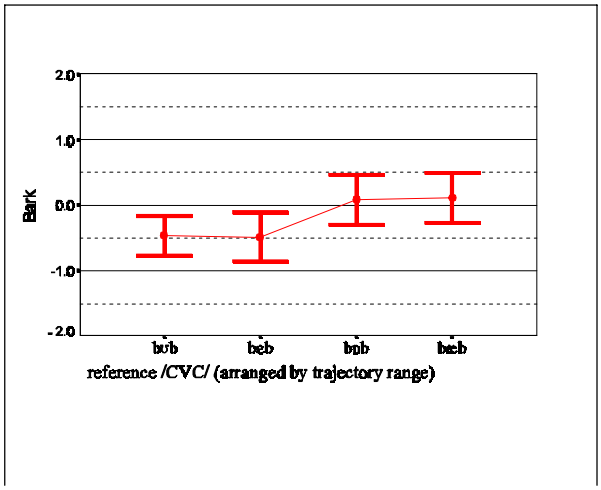
F2 Group X

	bɒb	bʊb	bæb	bɛb
range	2.02	2.65	4.97	5.67
MI	.00	-.45	-.13	.54
	dɛd	dæd	dʊd	dɒd
range	-0.52	-1.39	-2.51	-3.62
MI	.36	-.15	-.31	-.66

F2 Group Y

	bɒb	bʊb	bæb	bɛb
range	2.02	2.65	4.97	5.67
MI	-.07	-.07	-.20	.54
	dɛd	dæd	dʊd	dɒd
range	-0.52	-1.39	-2.51	-3.62
MI	.23	-.23	-.95	-1.19

Table 3: Trajectory range of F2 /CVC/ and its mean shift index.



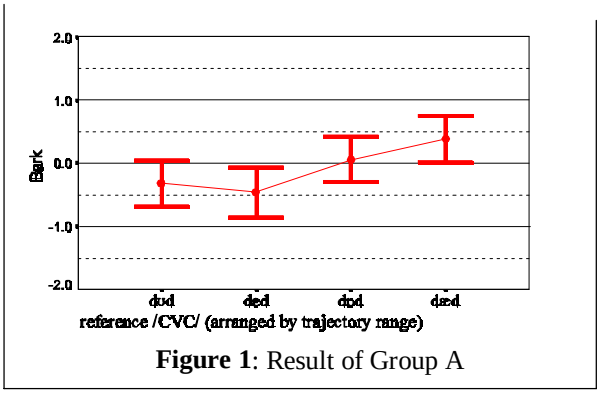


Figure 1: Result of Group A

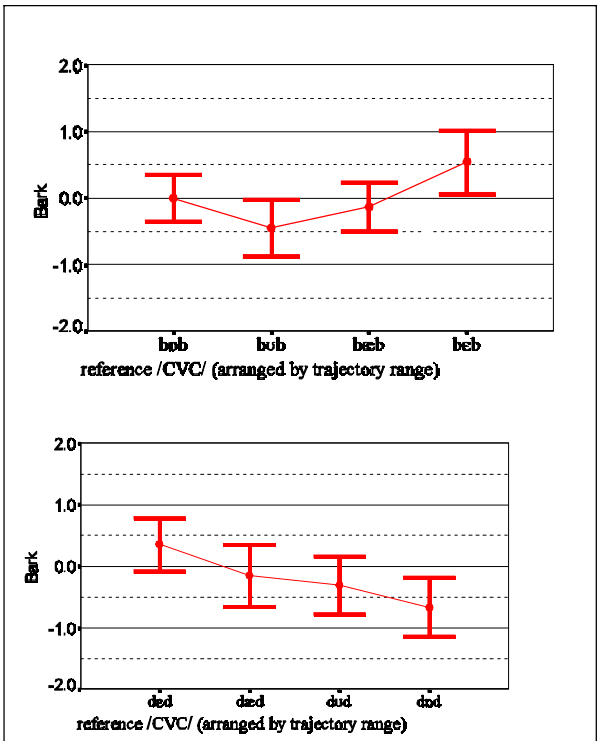


Figure 2: Result of Group X

Tables 2 and 3, together with the Figures 1 and 2, show an interesting relation between the trajectory range and the matched frequency across two groups of subjects: as the trajectory range of a formant in a reference /CVC/ increases, the mean matched formant frequency shifts from within the trajectory range to outside the trajectory. Note that the F2 trajectory in /dVd/ is concave. In other words, when the formant trajectory range is small, subjects select a value somewhere between the /CVC/ edge and peak frequencies to represent its vowel quality, and when the formant trajectory range is large, they select a value beyond the trajectory range (i.e. a value higher than the peak if the trajectory is convex, and a value lower than the peak if it is concave).

2.5 Discussion of the trajectory range effect

This section deals with the potential issues to be addressed concerning the effect of the trajectory range.

First, the order of the trajectory range reverses with regard to that of the indices from /ʊ/ to /ɛ/ in /bVb/ F1 and /dVd/ F1; and second, in Group X, from /bɒb/ to /bʊb/, shift indices drop while trajectory range increases.

The first irregularity might be explained in terms of natural variability since the shift index differences between /bʊb/ and /bɛb/ and /dʊd/ and /dɛd/ are not large. The second irregularity could be attributed to the peculiarity of /bɒb/. In /bɒb/ F2, the F2 choices of test /#V#/ are; 800, 875, 955, 1041, 1131 and 1227 Hz. Figure 3 shows the histograms of responses for F1 results of Groups A and B, for F2 results of Group X.

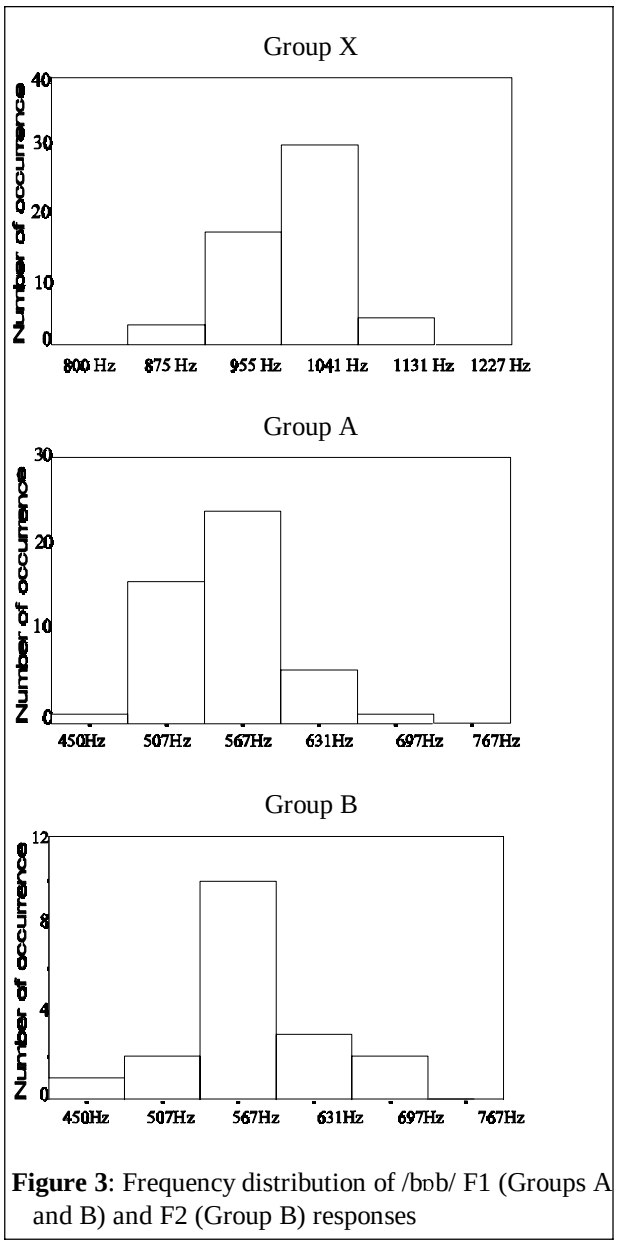


Figure 3: Frequency distribution of /bɒb/ F1 (Groups A and B) and F2 (Group B) responses

In the experiment, subjects in Group X selected F2=1041 Hz most frequently while in Groups A and B, the most frequent F1 choice was 567 Hz. If the interval between two formants is considered, one possible explanation for the peculiarity of /bɒb/ can be presented: since the Bark interval between 567 Hz and 1041 Hz is only 3.23 Bark, the two formants are very close. Furthermore, the test tokens of (F1,F2) = (567,955), which has 2.73 Bark F1-F2 distance, and (F1,F2) = (567,855), which has 2.23 Bark F1-F2 distance, show a slightly unnatural vowel quality since their two formants are closer than those of any natural vowels. This reason might account for /bɒb/ F2 matching to a higher frequency than is expected by its trajectory range.

3. CONCLUSION

Overall, this experiment was designed to investigate the strategy that listeners use to evaluate the quality of a vowel with dynamic formant trajectories. The results show that while subjects referred to the trajectory peak of /CVC/, the formant trajectory range of /CVC/ also affected their matching strategy: when the formant trajectory was small, subjects selected a value somewhere between the /CVC/ trajectory end frequency and peak frequency to represent its vowel quality, and when the formant trajectory range was large, they selected a value beyond the trajectory range: a value higher than the peak if the trajectory was convex, and a value lower than the peak if it was concave.

ACKNOWLEDGEMENT

I cordially appreciate the helpful comments made by Mark Huckvale of University College London, and also the detailed statistical analysis design proposed by Prof Burton Rosner of Oxford University.

REFERENCES

Holmes, J.N. (1985) “A parallel synthesizer for machine voice output.” In *Computer Speech Processing*. Eds. by F. Fallside & W.A. Woods, Prentice Hall International, page 163.

Nearey, T.M. (1989) “Static, dynamic and relational factors in vowel perception.” *Journal of the Acoustical Society of America*, **85**, 2088-2113.

Nord, L. (1986) “Acoustic studies of vowel reduction in Swedish.” *Speech Transmission Laboratory Quarterly Progress and Status Report*, **4**, 19-36.

Tokuma, S. (1995) “The problem of the missed target: how undershoot affects vowel quality.” *Journal of the Phonetic Society of Japan*, **208**, 45-55.

Tokuma, S. (1996) *The perceived quality of vowels showing formant undershoot*. PhD dissertation, University of London.