Quality perception of vowels with simulated /CVC/ formant trajectories

Shinichi Tokuma

Department of English Literature and Language, Sagami Women's University, 2-1-1, Bunkyo, Sagami-hara, Kanagawa, 228-8533, Japan

Received 8 April 1998; received in revised form 24 December 1999; accepted 6 January 2000

Abstract

This study investigates the perceived vowel quality change caused by formant undershoot, where vowels in /CVC/ environments are compared with steady-state vowels. In the perceptual experiment of this study, 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 results of the study show that, in vowel quality perception, the performance of subjects was affected by the formant trajectory range of the /CVC/ 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. This demonstrated phenomenon is compatible with the results of existing vowel perception studies, although the models proposed by these studies do not account for this phenomenon. © 2000 Elsevier Science B.V. All rights reserved.

Keywords: Formant undershoot; Grid-matching experiment; Vowel quality perception; Formant trajectory range

1. Introduction

In natural speech, vowels in consonantal contexts are modified from isolated form by their neighbouring consonants. In particular, their formant frequencies may not achieve their target or citation form values, a phenomenon called acoustic vowel formant undershoot.

Despite this acoustic variability, listeners are able to identify vowels in consonantal contexts accurately, and one explanation for this is to suppose that listeners, with the assistance of the acoustic attributes available in the acoustically underdetermined vowel, are able to reconstruct the target information which has been ‘lost’ in the course of speech production due to articulatory undershoot.

This hypothesis was first proposed by Lindblom and Studdert-Kennedy (1967). In their study, to interpret the phonological boundary shifts which occurred to vowels in /#V#/ and /jVj/ or /wVw/, they proposed perceptual compensation mechanism where listeners’ perceptual processes compensate acoustic formant undershoot: listeners use the direction and amount of movement in coarticulated formant transitions to calculate the formant target values that the speaker intended to reach but failed to achieve due to such factors as inertia of the articulators.

This hypothesis is now a standard explanation for coarticulated vowel perception, although it has met with criticisms from its recent theoretical alternatives. For example, Nearey and Assmann
(1986) and Andruski and Nearey (1992) proposed the vowel inherent spectral change model, where a vowel has its inherent spectral change found even in isolated forms, and it persists in coarticulated vowels as a major perceptual cue for the identification. Strange and her colleagues (e.g., Strange, 1989) suggested the dynamic specification model of vowel perception, which relies not on any particular spectral cross-section with some ‘target’ formant frequencies, but on the dynamic information available in the time-varying signals corresponding to vowel segments. Some other works examined $F_1$ temporal averaging process in vowel identification. Huang (1985) claimed that some time-average values of $F_1$ frequencies for the duration of the vowel including the transition can contribute to the perception of vowels, and Di-Benedetto (1989) extended this model and proposed “the weighted average time formant theory”, an $F_1$ averaging theory whose weighting function provides more importance to the first half of the $F_1$ /CVC/ trajectory.

Thus the perceptual invariance of coarticulated vowels has been a central problem for speech perception, but there is an interesting aspect to this phenomenon. Do vowels that are acoustically realised with formant undershoot give a different phonetic impression to listeners than those produced in isolation? Do listeners refer solely to the formant trajectory peak of the /CVC/ syllables in vowel quality evaluation, or do they rely on the information obtained from dynamic formant movement?

Another issue related to this is; if listeners utilise such dynamic information to evaluate the quality of vowels showing formant undershoot, do they perceive the vowel quality in the same manner as suggested by previous studies on phonological vowel perception? For example, in quality perception process, do listeners perceptually refer to the hypothetical target values by extrapolating the formant tracks in the CV/VC transition, as Lindblom and Studdert-Kennedy (1967) proposed in phonological vowel perception? Or do they follow a completely novel strategy?

The phonetic processes of vowel quality evaluation have received little previous attention. An experiment by Pols et al. (1984) investigated the perceived quality of a single-direction $F_2$ change: their stimuli had a steady $F_1$ trajectory and their $F_2$ change was single-directional. However, they failed to deliver clear results from an inadequate amount of data and were unable to verify whether the phenomenon reported on was caused by the underlying human perceptual mechanism or the experimental methods. Similarly, Akagi et al. (1994), although they carried out matching tests between /CVC/ and /#V#/., do not disclose the complete results, and their data and their implication cannot be evaluated.

Hence this study aims to: (1) discover by an appropriate experiment how listeners evaluate the quality of vowels in /CVC/ showing formant undershoot; and (2) investigate how and to what extent the phonetic processes of vowel quality evaluation contribute to the phonological identification of vowels, by comparing the results with the predictions from the proposed models of phonological vowel identification.

Finally, it should be emphasised that ordinary categorical tests of vowel perception (like labelling tests) do not allow for the investigation of vowel quality perception, since they would only supply listeners with pre-defined labels (phonemes) to attach to stimuli, which could mean that they need not make a quality judgement on a token which falls perceptually between two phonological prototypes. This view is supported by Pols et al. (1984), who claim, “… one of the drawbacks [of previous perceptual experiments] is that the listener could only specify his percept in terms of a limited number of pre-defined categories …” and therefore they introduced a matching paradigm to “elicit more finely grained judgements” (p. 372). Ladefoged (1967) in his research on vowel quality shares the same viewpoint. He claims, “When a phonetician hears a vowel he is usually capable of allocating it to one of a number of ‘general human categories of a sound’,” (p. 53) and then he argues that this process of describing a vowel sound cannot be considered in terms of phonemes or diaphones. After criticising the method to describe vowel quality in terms of key-words (like ‘vowel [s], pep’) since a vowel in a keyword does not necessarily have the same phonetic quality, he finally proposes, “It is obvious that phonemic
classifications do not provide a satisfactory basis for establishing phonetic categories” (p. 55). Thus, to investigate perceived vowel quality, we have introduced an interactive matching scheme using a two-dimensional grid-display (a revised version of Nord, 1986), which is described in the following section.

2. Experimental

2.1. Materials

Two types of tokens, the reference /CVC/ token and the test /#V#/ token, were created.

(a) test tokens: short vowels in isolation, whose formant frequencies could be modified by listeners in an interactive mode.

(b) reference tokens: /CVC/ syllable, where consonants were /b/ or /d/ and the vowels were /ɛ, æ, ɔ, u/, whose formant frequencies were fixed.

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, in order to investigate the effect of the F2 trajectory direction on perception. (F2 trajectory is concave in /dVd/ while it is convex in /bVb/.)

For vowels in /CVC/, of all RP short monophthongs, the four vowels /ɛ, æ, ɔ, u/ were selected, because in an acoustic (F1 versus F2) vowel diagram they are well separated from each other. /I/ was not selected since its inclusion would change the pattern of F2 trajectory in /dVd/; it would make both trajectories concave. The actual formula followed Nearey (1989), according to which the formant trajectory was synthesised as follows:

If one defines t as relative time within /CVC/ (i.e. t = 0 at the release of initial consonant), then formant frequency \( F_m(t) \), when t is between 0 and the midpoint of /CVC/, is

\[
F_m(t) = F_v + (F_i - F_v)[(t - T_v)b/T_v^b] \text{ (Hz)},
\]

where

\( F_i \): consonantal loci (Hz),

for /bVb/, \( (F_1, F_2) = (150,700) \),

for /dVd/, \( (F_1, F_2) = (150,2000) \),

\( T_v \): durational midpoint of a stimulus (ms),

\( F_v \): frequencies in the steady part of nuclei (Hz).

The total duration of this /CVC/ and the formant frequencies of /CVC/ syllable nuclei were obtained from an acoustic analysis that investigated the F1 and F2 frequencies of the RP short monophthongs produced by one male RP speaker in /CVC/ (C = /b,d,g/) uttered in isolation. Ten repetitions of each token were taken per condition, and its mean value was utilised. The formant values are shown in Table 1.

The mean duration of /CVC/ in the results of the acoustic experiment was 121 ms, and therefore, the duration of the synthetic /CVC/ was set to 120 ms (i.e. the durational midpoint in Formula (1), \( T_v \), was 60 ms). Vowels in these /CVC/ syllables had only two formants, F1 and F2, and they were synthesised using the JSRU parallel formant synthesiser by Holmes (1982) implemented in the Speech Filing System running on a Sun SPARC workstation. In the synthesis, \( F_0 \) declined linearly from 130 to 100 Hz, and every 10 ms point was interpolated linearly. Voicing of the token started at 10 ms after the initial point, reaching the full amplitude at 20 ms, and it started to decline 20 ms before the final point, ceasing completely 10 ms before the end. Between these 20 ms turning points, the formant ampli-

<table>
<thead>
<tr>
<th>Vowel</th>
<th>F1 (Hz)</th>
<th>F2 (Hz)</th>
</tr>
</thead>
<tbody>
<tr>
<td>/ɛVb/</td>
<td>542/1806</td>
<td>760/1621</td>
</tr>
<tr>
<td>/æVd/</td>
<td>524/1848</td>
<td>733/1619</td>
</tr>
<tr>
<td>/ɔVb/</td>
<td>545/1009</td>
<td>491/1122</td>
</tr>
<tr>
<td>/uVd/</td>
<td>565/1135</td>
<td>437/1359</td>
</tr>
</tbody>
</table>

*The values indicated are “F1”/“F2”. All values in Hz.*
tude was kept constant. \( F_1 \) and \( F_2 \) intensities were fixed to 50 dB during the voicing excitation. 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, using Kay Sonagraph Model 5500.

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 Formula (1). 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.

The other type of stimuli were the test /#V#!/, which had only two formants, \( F_1 \) and \( F_2 \), like the /CVC/ tokens, although these formants were static. Their \( F_0 \) pattern was identical: linear declination from 130 to 100 Hz, and with a duration of 220 ms. In the initial 10 ms the voicing energy increased gradually and in the final 10 ms it declined to avoid providing a clipped auditory impression.

2.2. Subjects

Fifteen native speakers of South-East British English participated in this experiment. They were undergraduate students in Linguistics or in Speech Science/Speech Communication at 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 experiment was held in the teaching laboratory of Wolfson House, Department of Phonetics, University College London. 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. At most three subjects were tested at one time.

Experimental instruction was given to the subjects in the following way: they were given a written explanation of the task \(^1\) required in the experiment. After reading it, they were given an oral instruction presented by the author, using one PC terminal in the laboratory, simulating the task that they were to do in the main experiment. Enquiries about the task of the experiment and the way to operate a PC and its mouse were answered during this oral presentation. The subjects then had one trial matching session and any further questions were answered. In addition, they were instructed to have a break between each session if they became tired.

The procedure of this experiment was as follows: in the laboratory, individual subjects were asked to sit in front of a PC terminal. Its screen displayed a schematic grid (6 × 6 blocks) with a cursor on one of these blocks, showing the “rela-

\(^1\) The written explanation of the task given to subjects is as follows: “Find the closest pair of vowel quality” – You can see a 6 × 6 grid on the PC screen. Clicking each block by a mouse, listen to a pair of /CVC/ and /#V#!/ on that block. The first token of each pair is the same in all block, but the second /#V#!/ changes its vowel quality block by block. Your task is to find out a block, where the vowel quality of the first token and second token is closest by playing around the grid”. A few subjects found this explanation was not clear enough, but finally they, having adequate background knowledge, understood the phonetic nature of the task after the following oral instruction, where it was emphasised that the subjects had to listen to the phonetic quality of stimulus vowels, not their phonological identity.
tive” position of the test token. In fact the grid was an acoustic vowel diagram, with $F_1$ step = $F_2$ step = 0.5 Bark (obtained by the formula $^2$ in (Bladon and Lindblom, 1981)), but the subjects were not informed of this. When they moved the cursor into a neighbouring block, so there was an increase/decrease of either $F_1$ or $F_2$ frequency of the test /#V#/ token by one step, while the reference token remained unchanged. The cursor position on the grid was moved by clicking with the mouse, so that the subjects could “hop” from one block in a grid to another in a remote position without passing through intermediate blocks, thus saving the time and effort of the subjects when traversing the grid.

Each time a block of the grid was clicked with the mouse, a pair of a test token and a reference token with an inter-stimulus gap of 300 ms was replayed. These two tokens were played in the order of ‘reference’-‘test’. The procedure is shown in Fig. 1.

The grid was arranged to make the grid values of $F_1$ and $F_2$ all different from the $F_1$ and $F_2$ trajectory peak values of reference /CVC/. The allocation of the direction of $F_1$ and $F_2$ on the two axes was randomised.

Subjects were instructed to tune into the vowel quality of both tokens and to discover in the grid a block where these tokens sounded closer than in any other block, i.e. to find a closest pair of /CVC/ and /#V#/ in vowel quality. When they found one, they pushed the space key on the keyboard and $F_1$ and $F_2$ values of the selected /#V#/ token and the other parameters were stored in an ASCII data file, and the next stimulus pair and the next grid were presented. Since the consonants of the reference /CVC/ were /d_d/ or /b_b/ and the vowels were /e, æ, ø, u/, the number of the reference token types was 2 (consonantal frames) $\times$ 4 (vowel types) = 8 (reference token types). Each reference token type (i.e. each /bæb/ /bæb/ ... /dud/) had six repetitions, therefore producing $8 \times 6 = 48$ matching sessions. The order of presentation of these matching sessions was randomised.

$^2$ The formula by Bladon and Lindblom is: Up to 3000 Hz, the critical band number $z$ (bark) for a given frequency $f$ (Hz) is provided by $z = 7 \times \ln(f/650) + [(f/650)^2 + 1]^{1/2}$ (Bark).
Chapter 4). Its results confirmed that the grid-matching scheme can be managed by the subjects.

2.4. Results and discussion

2.4.1. Adequacy of the obtained results
After examining the results of individual subjects, it was discovered that the data supplied by four subjects (out of 15) seemed to lack credibility for the following reasons:

1. all the responses of one subject fell within the very centre 2 × 2 blocks in the grid, implying that the results of the subject were likely to follow from a non-auditory strategy;
2. the other three subjects produced \( F_1 \) and \( F_2 \) responses which exceeded 3 × 3 steps in the 6 × 6 grid over six repetitions in more than six out of eight /CVC/ token types. This suggests that they made an arbitrary response to the presented /CVC/-/#V#/ stimuli and their results lack consistency.

These four subjects were therefore excluded from the analysis. The analysis of the results of the remaining 11 subjects is discussed in the following sections.

2.4.2. Interpretation of results
To study the general tendency of the matching process, first the shift index ([given matched for-
mant value] – [its corresponding /CVC/ trajectory peak]) was obtained on a Bark scale for each measurement. This shift index represents a formant frequency of /#V#/ chosen for the reference /CVC/ token with respect to its dynamic formant peak frequency. Then, the mean shift index was calculated across 11 subjects for each formant and for each vowel and consonant type. The mean shift index represents a mean formant value of /#V#/ that was chosen for a /CVC/ token with respect to its dynamic formant peak frequency. The results are displayed in Table 3.

Table 3 indicates that the strategy of subjects seems to be complex. The mean shift index fluctuates across all \( F_1 \) and \( F_2 \); the peak \( F_2 \)
frequencies of /dZd/ (1135 Hz) and /buB/ (1122 Hz) are close but the mean shift index of /dZd/ is far greater. This suggests that in the perceptual processing of dynamic formant trajectories, subjects did not pursue a straightforward matching strategy, such as perceptual overshoot or temporal averaging, since the former predicts that all the mean matched formant values would be outside the /CVC/ trajectory (resulting in the positive mean shift index of all F1 and /bVb/ F2, and all negative mean indices of /dVd/ F2), while the latter predicts that all the mean matched formant values would be within the /CVC/ trajectory (resulting in the negative mean shift index of all F1 and /bVb/ F2, and all positive mean indices of /dVd/ F2).

The results in an experiment of Tokuma (1995) suggest 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 investigated for both F1 and F2.

Tables 4 and 5 show the relations between trajectory ranges and the mean shift index to the

| Table 4 |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Trajectory range of F1 /CVC/ and its mean shift index, written as “mean index” ( = mean of all [matched formant value of /#V#/] – [its reference/CVC/ trajectory peak]) (all values in Bark) |
| /bVb/ F1 (F1 range → increase) | /dVd/ F1 (F1 range → increase) |
| Range | u | ε | D | w |
| Mean index | 3.28 | 3.70 | 3.73 | 5.37 |
| Mean index | -0.54 | -0.61 | 0.13 | 0.16 |

| Table 5 |
|-----------------|-----------------|-----------------|-----------------|-----------------|
| Trajectory range of F2 /CVC/ and its mean shift index, written as “mean index” ( = mean of all [matched formant value of /#V#/] – [its reference/CVC/ trajectory peak]) (all values in Bark) |
| /bVb/ F2 (F2 range → increase) | /dVd/ F2 (F2 range → increase) |
| Range | D | u | w | ε |
| Mean index | 2.02 | 2.65 | 4.97 | 5.67 |
| Mean index | -0.02 | -0.34 | -0.15 | 0.54 |

| Range | ε | w | u | D |
| Mean index | -0.52 | -1.39 | -2.51 | -3.62 |
| Mean index | 0.32 | -0.17 | -0.48 | -0.80 |
Fig. 2. Schematic figure of $F_1$ frequency chosen for /bVb/ overlaid on trajectories (a) and $F_1$ frequency chosen for /bVb/ with respect to peak frequencies (b).

Fig. 3. Schematic figure of $F_1$ frequency chosen for /dVd/ overlaid on trajectories (a) and $F_1$ frequency chosen for /dVd/ with respect to peak frequencies (b).

Fig. 4. Schematic figure of $F_2$ frequency chosen for /bVb/ overlaid on trajectories (a) and $F_2$ frequency chosen for /bVb/ with respect to peak frequencies (b).

Fig. 5. Schematic figure of $F_2$ frequency chosen for /dVd/ overlaid on trajectories (a) and $F_2$ frequency chosen for /dVd/ with respect to peak frequencies (b).
trajectory peak. Table 4 is for $F_1$ results, and Table 5 for $F_2$ results. In each formant/consonantal environment, the order of the vowels is arranged so that the /CVC/ trajectory range increases from left to right.

These mean differences are also plotted in Figs. 2–5, together with error bars of one standard deviation, which were obtained by accumulating all the individual frequency differences between the matched value and the /CVC/ trajectory peak according to each token type. Figs. 2–5 also have schematic spectrograms which display the formant frequencies chosen for a /CVC/ token (shown as approximated error bars) overlaid on the /CVC/ formant trajectory. The frequency axis of those schematic spectrograms is a relatively close, but not exact, approximation to the actual Bark scale.

Tables 4 and 5, together with Figs. 2–5, show an interesting relation between the trajectory range and the matched frequency: 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 $F_2$ 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). This is illustrated in Fig. 6.

Finally, there are potential issues to be addressed concerning the effect of the trajectory range: first, in Figs. 2 and 3, the order of the trajectory range reverses with regard to that of the indices from /u/ to /i/ in /bVb/ $F_1$ and /dVd/ $F_1$; and second, from /bVb/ to /bVb/, shift indices drop while trajectory range increases, as shown in Fig. 4.

The first irregularity might be explained in terms of natural variability since the shift index differences between /bVb/ and /bVb/ and /dVd/ and /dVd/ are not large. The second irregularity could be attributed to the peculiarity of /bVb/. In /bVb/ $F_2$, the $F_2$ choices of test /#V#/ are 800, 875, 955, 1041, 1131 and 1227 Hz. In the experiment, sub-

2.4.3. Statistical analysis on subject homogeneity

Since different subjects may have reacted in different ways to the experiment, repeated measures ANOVA were carried out to examine whether these 11 subjects were statistically homogenous, with factors of [subject] (11-levels), [consonant] (2-levels), [vowel] (4-levels) and [trial]
With regard to $F_1$, the analysis showed that [subject] as a main factor was significant ($F(10, 150) = 3.01, p < 0.01$) and the interaction [subject] * [vowel] was also significant ($F(30, 150) = 5.32, p < 0.01$), although the interaction [subject] * [consonant] was not significant ($F(10, 150) = 1.25, p > 0.01$). The analysis of $F_2$ matching confirmed a difference between subjects: subject as a main factor was significant ($F(10, 150) = 3.06, p < 0.01$), and the interactions [subject] * [vowel] ($F(30, 150) = 1.99, p < 0.01$) and [subject] * [consonant] ($F(10, 150) = 3.06, p < 0.01$) were both significant.

Thus the 11 subjects were shown not to form a homogeneous group, making it necessary to create subgroups, and therefore the following procedure was undertaken. First, cluster analysis was made on the mean shift index across six trials, for each subject, consonantal context and vowel type. Then a subject with a distinctive response pattern was eliminated according to the results of the cluster analysis. This elimination process was repeated until the $F$-ratio of the repeated measures ANOVA was less than that of the 1% significance level, for [subject] as a main factor and its interaction by [consonant] and [vowel]. This process was carried out for each formant type across all vowel types, and eventually created two subject groups for $F_1$ and two groups for $F_2$. The two subject groups for $F_1$ are henceforth called Groups A and B and the two subject groups for $F_2$ are called Groups X and Y. There are eight subjects in Group A and three in Group B, while there are nine subjects in Group X and two in Group Y.

---

**Group A**

![Chart of Group A](chart_a.png)

**Group B**

![Chart of Group B](chart_b.png)

---

**Fig. 7.** $F_1$ frequency chosen for /bVb/ with respect to peak frequencies.

**Fig. 8.** $F_1$ frequency chosen for /dVd/ with respect to peak frequencies.
Subsequently the mean shift index was calculated across subjects in each subject group for each formant and for each vowel and consonant type. These mean differences are plotted in Figs. 7–10, together with error bars of one standard deviation, which were obtained by accumulating all the individual frequency differences between the matched value and the /CVC/ trajectory peak according to each token type. Figs. 7 and 8 are for F1 results, Groups A and B, and Figs. 9 and 10 for F2 results, Groups X and Y. As in Figs. 2–5, in each formant/consonantal environment, the order of the vowels is arranged so that the /CVC/ trajectory range increases from left to right.

In Figs. 7 and 8, the results of Groups A and B show the same type of sensitivity to the formant trajectory range observed above although the results of Repeated Measures ANOVAs suggested that the two groups do not form a homogenous group. The same is true for Groups X and Y, as shown in Figs. 9 and 10. In Fig. 9, one could notice a slight difference between results of Groups X and Y, where Group Y shows the matched mean of /bUb/ higher than that of Group X, but the results of Group Y do not weaken the observation made on the sensitivity to the formant trajectory range since they show that the matched formant frequency of test /#V#/ tokens moves from within the trajectory range to outside the trajectory range as the formant trajectory range of the reference /CVC/ tokens grows larger. Therefore, these results demonstrate that all the subjects in this experiment, regardless of a subgroup to which they belong, were influenced in the same way by the effect of the trajectory range.
2.4.4. Concluding remarks on the experiment

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 the formant trajectory range of /CVC/ affected the matching strategy of subjects: when the formant trajectory range 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. The next section examines to what extent this effect of the formant trajectory range found in the experiment is supported or contradicted by the previous studies on phonological or phonetic perception of vowels with formant undershoot.

3. General discussion

3.1. Significance of the results in the context of previous studies

First we examine to what extent the effect of the formant trajectory range is supported or contradicted by previous studies on phonological or phonetic perception of vowels with undershoot.

There have been a number of studies of phonological vowel perception to which we could relate the trajectory range effect. In the labelling test of Lindblom and Studdert-Kennedy (1967), used in tests of the perceptual target compensation, the stimuli /jVj/ had a concave F2 trajectory with trajectory ranges of 464–724 Hz. These values are similar to those used in the main experiment of this study, where the concave F2 trajectory of /dæd/ had a range of 1.39 Bark (= 381 Hz). Our results showed that the mean matched frequency was below the trajectory peak, with its mean shift index being −0.15 Bark in Group X and −0.23 Bark in Group Y. Hence the results of the main experiment predict a perceptual overshoot in the /jVj/ results of Lindblom and Studdert-Kennedy (1967), and indeed this was the case.

Further support is found in the temporal averaging hypothesis of vowel perception, which claims that the phonological perception of a vowel with a dynamic F1 trajectory is determined by the average F1. This seems to be in contradiction to the findings of this study when the trajectory range is large. However, DiBenedetto (1989), outlining a modified version of the temporal averaging hypothesis of F1, used F1 stimuli with peak values that fell within the range of American English /ɪd–ɪd/ continuum, which means that their F1 range was small, and hence the temporal averaging hypothesis is compatible with the findings of this study.

The dynamic specification model of vowel perception, although it is one of the mainstreams of vowel perception theory, is not appropriate to incorporate in this study since it did not provide concrete algorithms. The results obtained in this study cannot verify the vowel inherent spectral change model, since the experiment did not manipulate the acoustic parameter provided by the vowel inherent spectral change.

Kuhl proposes the perceptual magnet effect, which claims that perceptual distinctions between linguistic sounds close to a prototype are minimised, and that other distinctions are maximised between sounds that are not too close to any prototype, but lie on opposite sides of a boundary between categories (see, for example, (Kuhl, 1983; Kuhl and Iverson, 1995)). The interesting aspect of her proposal is that, to explain how the effects of context are reflected in the prototype representations, Kuhl and Iverson (1995) claim that there are data to suggest “the location of best instances of the category shifts with changes in such variables as the rate of speech” (p. 1146). It must be emphasised that this perceptual magnet effect is attributed to the phonological level of perception and the experiment of this study deals with phonetic judgements. However, if we assume that the matched /#V#/ would cluster around the “prototype” of the vowel of /CVC/ syllable, as predicted from the perceptual magnet effect, the trajectory range effect found in this experiment could trigger the shift of the phonological prototype in the perceptual magnet effect. However this assumption should await further investigation.

Next, significance of the results in the context of studies on phonetic vowel perception is discussed. An experiment by Pols et al. (1984) investigated
the perceived quality of a single-direction $F_2$ change. Their stimulus structure was considerably different from that of this study, in that their stimuli had a steady $F_1$ trajectory and their $F_2$ change was single-directional. Their results show that a small formant transition gave a matched frequency beyond the endpoint of the changing formant while in our results the smaller trajectory ranges caused matching between the trajectory peak of /CVC/ and the onset/offset. For compatibility, therefore, we must assume that either the direction of a trajectory change or the absence of a dynamic $F_1$ trajectory explains why no trajectory range effect was found.

Akagi et al. (1994) carried out a matching test, and one of the 8 stimuli types which they used is similar to the /CVC/ stimuli used in the experiment. The $F_1$ of the stimuli, with its peak frequency of 546 Hz (5.13 Bark), is flanked by 30 ms transitions terminating 281 Hz (2.73 Bark), while all the other formants, $F_2$–$F_5$, do not have a transition. Since the stimuli had a quasi-parabola $F_1$ with a trajectory range of 2.4 Bark, the results of the main experiment predict a perceptual undershoot. Although Akagi et al. (1994) do not disclose the complete result, they claim that the result of the matching tests of the eight stimuli groups generally shows the temporal averaging of $F_1$. In this regard, their results are partially compatible with ours.

Some claim that the matching performance of listeners is influenced by whether they match vowel ‘phonemeness’ (phonological) or vowel timbre (phonetic/psychoacoustic), and that in the latter case the assimilation effect occurs, where “the categorical boundary of stimulus series shifts in a direction opposite to the context stimulus” (Shigeno 1991, p. 103). The terms assimilation or contrast in psychophysics normally apply to aspects of identification performance, and it is not certain whether we could extend ideas on these effects to multiple boundaries between response categories in matching, since the response in matching is differently defined. However, if they are operating, they will lead to a perceptual boundary shifting over trials, and this may lead to increased variability of matching in case of assimilation. This was indeed the case, as seen in somewhat large standard deviations of matched formant values, shown as error bars in Figs. 2–5.

Overall, it was found that our results are compatible with the findings of the previous studies on vowel perception. However, it should not be overlooked that the models proposed by these studies do not explain how and why this effect of formant trajectory range occurs in vowel perception process. Perceptual overshoot model predicts that all the mean matched formant values would be outside the /CVC/ trajectory, while temporal averaging model predicts that the mean matched formant values would be constantly within the /CVC/ trajectory range, and neither is the case here. Other previous phonological/phonetic models of vowel perception, for example, the dynamic specification model or the vowel inherent spectral change model, cannot explain this trajectory range effect, nor can the perceptual magnet effect or the assimilation/contrast effect. The next section discusses two possible interpretations of this perceptual phenomenon.

3.2. Interpretation of results within overall speech perception models

Two possible interpretations are discussed here of the results obtained in this experiment to hypothesise how and where the trajectory range effect arises in the process of speech perception.

The first interpretation is that the trajectory range effect is attributed to the level of psychoacoustic perception. One could hypothesise that there are psychoacoustic compensation processes for formant undershoot that could be the origin of phonological compensation processes: phonological processes are dependent upon the ‘bottom-up’ influence of obligatory psychoacoustic processing, since speech sounds are subject to processing by the auditory system as are any non-speech sounds: they are involved in the same obligatory auditory processes.

This interpretation adopts a bottom-up approach: it is assumed that when the trajectory range is small, the output from the underlying psychoacoustic process of vowel quality evaluation provides a vowel quality judgement of the /CVC/ according to some kind of sampled or
averaged value, which is a cause of the observed phonological temporal averaging process; and that when the trajectory range is large, the bottom-up output gives a judgement of vowel quality in /CVC/ by selecting a frequency outside the /CVC/ trajectory range, which is realised as phonological target compensation at the phonological level.

However, this assumption is difficult to justify because it is yet to be investigated why the underlying psychoacoustic/phonetic process might show such an odd sensitivity to the dynamic trajectory range: specifically, it does not explain why, in the underlying psychoacoustic level, temporal averaging or phonological target compensation would operate in some formant patterns but not in others.

The second interpretation is that in the phonetic/psychoacoustic experiment of this study, the proposed phonological processes (such as perceptual target compensation, temporal averaging or dynamic specification) influenced the listeners’ process of matching vowel quality. These higher-level processes may be affecting listeners’ psychoacoustic judgements of vowel quality as a ‘top-down’ influence from the phonological level to the psychoacoustic level, by shifting the matched formant values.

This is based on a top-down approach: this interpretation assumes that although the psychoacoustic processes may play some role in vowel perception, they provide only primitive perceived features of acoustic signals, and the perception of the quality of a vowel is strongly influenced by the intervention from higher phonological perception processes. Note that in this (top-down) interpretation the output of phonological vowel perception processes mainly influences the psychoacoustic judgements of vowel quality, while in the previous (bottom-up) interpretation the direction of influence is reversed: the output of phonetic/psychoacoustic perception process unilaterally affects the higher phonological process.

This interpretation is partially supported by the result of Akagi et al. (1994). They claim that the perceptual overshoot is observed only in the results of the identification test of vowels, not in the matching test, and that consequently the overshoot results from a higher-order central processing mechanism. However, their claim is weak since they failed to deliver clear results from their matching experiment. Also, this interpretation does not specify how phonological perception processes affect the phonetic judgements, and this lack of specific perceptual mechanism remains to be addressed.

Overall, neither of the two interpretations above adequately describes how and where the trajectory range effect arises in human speech perception process. Further empirical exploration of this trajectory range effect is to be pursued in future work to suggest its satisfactory explanation.

4. Conclusion

The results of this study showed that, in vowel quality perception, the formant trajectory range of the /CVC/ stimuli affected the performance of subjects. When the formant trajectory was small, they selected a value between the edge and the peak frequencies, while they selected a value outside the trajectory range when it was large. This demonstrated phenomenon is compatible with the results of existing vowel perception studies, although the models proposed by these studies do not account for this phenomenon.

Acknowledgements

I wish to express my gratitude for the helpful comments made by Dr. Mark Huckvale and Dr. Andy Faulkner of University College London, and for the useful feedback given to the poster presented in Eurospeech 97 (Tokuma, 1997), which is a preliminary version of this paper. I also appreciate detailed statistical support from Professor Burton Rosner of Oxford University. The whole experimental process carried out on a terminal screen was programmed in C-language by Dr. Mark Huckvale.
References


