UNIVERSITY OF ALBERTA

PERCEPTUAL INFORMATION FOR VOWEL IDENTIFICATION

BY

JEAN E. ANDRUSKI

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH

IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE

OF MASTER OF SCIENCE

IN

SPEECH PRODUCTION AND PERCEPTION

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA

FALL, 1990

ISBN 0-315-64981-X

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Jean E. Andruski

TITLE OF THESIS: Perceptual Information for Vowel Identification

DEGREE: M. Sc.

YEAR THIS DEGREE GRANTED: 1990

Jean Andruski
(Student's signature)

11439-48 Avenue
Edmonton, Alberta
T6H 0C9
(Student's permanent address)

Date: June 8/1990

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled PERCEPTUAL INFORMATION FOR VOWEL IDENTIFICATION, submitted by Jean E. Andruski in partial fulfilment of the requirements for the degree of Master of Science in Speech Production and Perception.

Dr. T.M. Nearey (Supervisor)

Dr. B.L. Derwing

Dr. B. Rochet

June 8, 1990
Date

ABSTRACT

It has been proposed that the endpoints of vowels in consonant context contain cues to vowel identity that are not available in isolated vowels, and that these supplementary, coarticulatory cues are perceptually superior to cues which the vowels themselves provide regarding their identity. However, a vowel-*inherent* cue, namely spectral change, measured from the initial to the final portion of the vowel itself, is also shown to persist in the endpoints of [bVb] syllables. Perceptual tests using 'silent center' stimuli, in which listeners hear only brief initial and final portions of the syllable, with the center section reduced to silence, show that listeners' identification patterns and error rates for isolated vowels and [bVb] syllables are very similar. This suggests that the same type of information is used to identify both types of stimuli. Predictions of listeners' identification patterns using a statistical model of vowel identification, which uses only vowel-inherent spectral change cues, are also shown to be significantly correlated with listeners' responses. These findings indicate that, although coarticulatory information may play some minor role in the perception of English vowels in consonant context, speakers of Western Canadian English rely largely on vowel-inherent spectral change to identify both isolated vowels and vowels in [bVb] syllables, at least when these are presented in silent center form.

ACKNOWLEDGEMENTS

My family deserves first, and very special mention for their unfailingly solid support and encouragement, and for the confidence they've always shown in my ability to successfully switch careers. They are all very much loved and appreciated.

I'd also like to express very special thanks to my supervisor, Dr. Terry Nearey, for the many hours he spent helping with experimental designs and data analysis, for the vast amounts of programming he has done, and for being such an excellent source of constructive comments, questions and criticism at every stage of this research. He and my other committee members, Dr. Bruce Derwing and Dr. Bernard Rochet, have provided much in the way of encouragement and inspiration, not just with regard to this thesis, but throughout my stay in the Department of Linguistics.

The CSRE gang – especially Terry Baxter and Tom Welz – have made important contributions as well, and I thank them for their willingness to write special routines to control and analyze various aspects of these experiments.

My fellow graduate students, notably Grace, Lucille, Ming-Ming, Radecka, Satomi, Susan and Yuhuan, have also been a wonderful source of support. Elyse Abraham and Murray Munro have been particularly helpful in a host of different ways, and I'd like to thank both for everything, including their editorial help with earlier drafts.

Finally, I would like to thank the Department for providing financial support in the form of numerous research and teaching assistantships, and the Province of Alberta, for similar support in the form of Graduate Scholarships.

TABLE OF CONTENTS

Chapter

Chapter

Chapter

LIST OF TABLES

Table

LIST OF FIGURES

Figure

CHAPTER 1

Literature Review

Although formant structure is traditionally recognized as the primary acoustic cue to vowel identity, the frequencies of formants for a given vowel vary with speaker identity, consonant context and speaking rate, so that the 'raw' or untransformed formant values do not provide invariant cues to vowel identity. In this thesis, perception of isolated Canadian English vowels is compared with perception of the same vowels in [bVb] syllables, and the formant cues to vowel identity used by listeners in each case are examined.

Magnitude of formant variation introduced by various sources

At the present time, there is no consensus regarding the degree to which perception of vowels in consonant context makes use of the same cues as perception of isolated vowels. In comparison with other sources of formant variation, such as vowel and speaker identity, frequency variations introduced by consonant context tend to be smaller, yet of great enough magnitude to be perceptible. In his summary of sources of variation in vowel formant structure, Nearey (1989) points out that the largest source of variation in formant frequency is vowel identity itself. For example, for female speakers in the Peterson and Barney data (1952, as cited by Nearey 1989, p. 2090) the average F1 spacing between vowels, ranges from 38% to 44% of baseline F1 values. As an example of F2 spacing, Nearey cites [æ] and [ʌ] as differing by about 46% in F2.[1] Speaker identity is the next largest source of variation, with formant values for a

[1] Calculating % change "from some baseline values, or more precisely as

$$\%\text{change} = 100\,[(x/V_{ref}) - 1]$$

where x is the modified value and V_{ref} is the baseline or reference value." (Nearey, 1989, p. 2090) single vowel varying by an average of 30% when spoken by children versus adult males (Peterson and Barney, 1952, as cited in Nearey 1989, p. 2090). As Nearey points out, consonant context tends to have a smaller effect on vowel formant frequencies than vowel or speaker identity, at least when the consonants are obstruents. In Stevens and House (1963), full duration stressed vowels spoken by three speakers in fourteen consonantal environments showed average variations of 4% in F1 and 10% in F2 (1963, p.114-115).[2] Nearey also cites slightly larger variations reported by Lindblom (1963) for eight vowels spoken by a single speaker in three consonant frames with varying stress and timing patterns. Durations for these vowels ranged from 80 to 300 ms (Lindblom 1963, p. 1774) and actual formant values varied from predicted steady-state targets by an average of 10% for F1 and 17% for F2.[3]

Given that formant spacing between different English vowels is on the order of 40% to 50% of baseline F1 and F2 values, and given that deviations from baseline values for vowels in consonant context are in the range of 4% to 17%, are the same formant cues to vowel identity present in both isolated vowels and vowels in consonant context, and if so, to what extent do listeners use them in each case?

Consonants as a source of information for vowel identity

Strange, Verbrugge, Shankweiler and Edman (1976) suggested that, since the formant structure of isolated vowels is sometimes considerably different from the structure of the same vowels in consonant context, listeners may use different acoustic cues to identify vowels in context versus isolation. Their suggestion was prompted by the results of several perceptual experiments in which listeners' identification rates for

[2] Percentage change calculated by this author using Nearey's formula, see footnote 1.
[3] Averages calculated by this author from Nearey's Table III, 1989, p. 2091.

isolated vowels were inferior to rates for vowels in consonant context (Fairbanks and Grubb (1961); Fujimara and Ochiai (1963)).

Strange *et al.* suggested two ways in which consonant context might contribute to more accurate vowel identification. Since earlier experiments did not systematically control for speaker identity, listeners might be using cues introduced by consonant context to assist in normalizing changes in vocal tract size and shape when speaker identity changes. The presence of such consonant-based cues might allow listeners to adjust more accurately for speaker-dependent variations in vowel formant structure, resulting in fewer errors for vowels in consonant context than for isolated vowels. If this were the case, the formant cues used to identify vowels both in and out of consonant context could be the same. The consonants might simply provide more accurate information on how to normalize those cues for speaker variation.

The second possibility suggested by Strange *et al.* is that vowel information may be distributed throughout the syllable in which a vowel is embedded, so that vowel identity cannot be well-specified "by formant frequencies at any particular cross section in time, but rather is carried in the dynamic configuration of the whole syllable" (1976, p. 214). A corollary of this view is that vowels in isolation are "under-specified" in comparison with vowels in CVC context. If this should prove to be true, the information used to identify vowels in consonant context would be quite different from that used to identify isolated vowels, since the former, but not the latter, contain 'supplementary' information in the CV and VC transitions.

In their experiment, Strange *et al.* compared identification rates for vowels in [pVp] context with rates for the same vowels spoken in isolation by a group of fifteen male, female and child speakers. Listeners heard the tokens in either "segregated talker" condition, in which a single talker produced all tokens for the listening test, or in "mixed talker" condition, in which talkers were presented in random order.

Talker Arrangement	Error Rate	
	Isolated Vowels	[pVp] Syllables
Segregated	31.2%	9.5%
Mixed	42.6%	17.0%

Table 1-1: Summary of error rates for Strange, Verbrugge, Shankweiler and Edman (1976), Experiment 1.

Changes in speaker identity contributed significantly to identification errors, but the largest difference in identification rates was attributed to the presence or absence of consonant context. Error rates are summarized in Table 1-1. An acoustic analysis of the vowel tokens showed that, when a single slice from the most steady state portion of each vowel was considered, the isolated vowels were somewhat better distinguished by their F1-F2 structure than the vowels from [pVp] context. Hence, if accurate perception were contingent on finding a single, well-differentiated cross section from the vowels, the isolated vowels should have been better identified than the [pVp] vowels. Since vowel duration also provides a cue to vowel identity (Peterson and Lehiste, 1960; House, 1961; Ainsworth, 1972; Mermelstein, 1977), Strange *et al.* compared durations in the two sets of stimuli, and concluded that differences in duration of the vowels in context as compared with isolation were insufficient to explain the confusion patterns and differences in error rates. Based on these two 'vowel-inherent' factors then (i.e., duration, plus formant values in a single cross section of the vowel) the isolated vowels appeared to be acoustically as well specified, and perhaps better specified than the [pVp] vowels. Strange *et al.* suggested that the [pVp] syllables must therefore provide some additional cues to vowel identity, since they were better identified. In a second experiment using several consonant contexts, listeners again identified vowels in context more accurately than isolated vowels, even though consonant identity was unpredictable from trial to trial.

Since identification rates for the segregated talker condition improved significantly when vowels were heard in consonant context, Strange *et al.* concluded that coarticulation of vowels with consonants improves vowel identification not by facilitating normalization of talker vocal tract differences, but rather, by providing a source of vowel information unavailable in isolated vowels. Specifically, they suggested that vowels are not well-specified by any single, temporal cross section, including a cross-section taken through the 'steady state' of an isolated vowel. Instead, they concluded that vowels are specified dynamically, and it is the coarticulatory information contained in consonant transitions that provides the dynamic cues required by listeners for accurate vowel identification.

Dialect and response-task as factors in vowel identification

Subsequent experiments cast considerable doubt on the conclusion that isolated vowels are less well specified than vowels in consonant context by showing that listeners can identify natural isolated vowels at error rates comparable to rates for vowels in context when factors such as dialect and orthographic interference are controlled, as summarized in Table 1-2. Macchi (1980) suggested that variations in speaker dialect, and the lack of any unique orthographic representation for English vowels might have affected error rates in Strange *et al.* (1976). She matched speakers and listeners for regional dialect, and controlled for differences in response accuracy that may be introduced by the ambiguous orthographic representation of English isolated vowels, by asking listeners to rhyme both isolated vowel and [tVt] syllable responses with English words. Her data showed no significant differences between error rates for isolated vowels and vowels in [tVt] syllables, suggesting that coarticulatory information is not necessary for accurate identification of vowels.

Diehl, McCusker and Chapman (1981) suggested that finding the spelled version of a syllable may be simpler than finding a word which 'contains the same vowel sound' as the test syllable. They tested three different response conditions and found that circling written [bVb] words (a task similar to that used by Strange *et al.*) resulted in a significant advantage for vowels in [bVb] context over isolated vowels. However, no advantage was found for [bVb] syllables when answer sheets consisted of 'spelled' isolated vowels. Similarly, no advantage was found for [bVb] syllables when listeners were asked to "vocally mimic" the test items. Diehl *et al.* concluded that listeners identify context vowels more accurately than isolated vowels in some experiments not because consonant context provides additional acoustic cues, resulting

Experiment	Error Rates (%) by Type of Task				
	Rhyming	Keyword	Spelling	Phonetic Symbols	Repetition (Mimicking)
Macchi:		—	—	—	—
[tVt]'s	5.3				
Iso V's	4.7				
Diehl *et al.*:	—	—			
[bVb]'s			3.4 8.2	3.8	2.3
Iso V's			7.8 8.6	7.1	2.1
Assmann:	—				
[pVp]'s		15.1	4.0	5.1	4.9
Iso V's		17.2	9.0	7.3	4.6

Table 1-2: Summary of error rates from experiments by Macchi (1980), Diehl *et al.* (1981) and Assmann (1979) on differences in task difficulty when identifying isolated vowels as compared to vowels in consonant context. Results for Diehl *et al.*'s spelling task and repetition task are "adjusted" error rates, in which results for 15 "poorly produced" tokens (1981, p. 244) are excluded. Adjusted error rates for the phonetic symbols group are not reported by Diehl *et al.* For spelling task results, left-hand results are for the [bVb] answer sheet, on which responses were spelled [bVb] words. Right-hand results are for the vowel answer sheet, on which responses were "spelled" isolated vowels (e.g. EE, I, A, etc.)

in more accurate perception, but rather because certain response tasks are more difficult for subjects to perform for isolated vowels than for vowels in consonant context.

Earlier experiments by Assmann (1979) yielded essentially the same outcome as the above experiments by Macchi (1980) and Diehl *et al.* (1981). Assmann concluded that subjects have difficulty labeling isolated vowels in certain response tasks, resulting in higher error rates for isolated vowels, and consequent over-estimation of the advantage of consonant context. In three experiments, he compared four different response tasks and found that 'keyword' tasks are more difficult to perform than 'spelling' tasks. In keyword tasks, responses are English words that contain the test vowel. In spelling tasks, reponses are the spelled versions of test syllables. Since no unique spelled equivalents exist for English isolated vowels, vowels in consonant context have an advantage over isolated vowels on spelling tasks. When keyword tasks are used, neither vowel type has an advantage. When listeners were asked to transcribe their responses using IPA symbols, results were similar to those for keyword tasks, in that both vowel types were approximately equally well identified. When subjects were asked to repeat the test vowel, error rates dropped for both isolated vowels and vowels in consonant context. From these results, Assmann suggested that listeners frequently label vowels incorrectly in written responses even though the vowels are correctly perceived.

Like Macchi (1980) and Diehl *et al.* (1981), Assmann (1979) concluded that coarticulatory information is not necessary for accurate identification of vowels. He suggested that listeners may indeed rely on dynamic information for assistance in discriminating spectrally similar vowels. However, whereas Strange *et al.* (1976) proposed that the perceptually important dynamic information is contained in consonant transitions, Assmann hypothesized that the dynamic characteristics of vowels themselves, including duration and diphthongization, may help listeners to disambiguate vowels when speaker identity and consonant context result in separate vowel categories having overlapping formant frequencies.

Vowel-inherent dynamic cues to vowel identity

To test this hypothesis, Assmann used a digital windowing procedure to gate 100 ms sections from 10 isolated vowels. In these shortened vowels, diphthongization is reduced and information regarding relative duration is eliminated. Assmann found that error rates for the artificially shortened vowels were more than twice as great as for the unmodified isolated vowels, and most errors involved confusions between spectrally similar vowel pairs such as [e−ɪ] and [e−ɛ], suggesting that full isolated vowels do contain perceptually useful dynamic information.

Since error rates on the shortened isolated vowels were still very low (13.8% in mixed speaker condition and 9.5% in blocked speaker condition) Assmann also suggested that isolated vowels carry a great deal of redundant information, making them perceptually quite robust, rather than impoverished and in need of perceptual reinforcement. Linear discriminant function analyses based on F0 and the first three formants from the windowed vowels indicated that these shortened segments did contain sufficient information to discriminate between most of the vowels. In Assmann, Nearey and Hogan (1982) a further analysis was performed on data from these vowels. In this paper, a discriminant analysis based on the 'steady state' parameters alone was compared with an analysis that included parameters for both steady state formant values and the dynamic properties of formant slope and vowel duration. The dynamic measures were found to be significantly correlated with listeners' identification rates for the full isolated vowels, and correlations involving the dynamic measures were "uniformly higher than any of the analyses involving steady-state measures alone" (1982, p. 984), suggesting that these 'vowel-inherent' sources of dynamic information are perceptually useful.

These experiments clearly demonstrated that isolated vowels are acoustically well-defined stimuli which do not require consonant context to be perceptually viable. However, Strange and Gottfried (1980) provided evidence that vowels in consonant context can retain an advantage over isolated vowels, even when task variables are controlled. In their experiment, vowels in [kVk] syllables were better identified than isolated vowels on both a rhyming task and a keyword task, as summarized in Table 1-3.

Stimulus Type	Keyword Task	Rhyming Task
[kVk]'s	6.7%	4.5%
IsoV's	27.7%	18.5%

Table 1-3: Summary of error rates for Strange & Gottfried (1980)

Strange and Gottfried suggested that conducting perceptual tests on vowel identification under such highly controlled conditions as those in Macchi's (1980) study produces identification rates that are too good: under these optimum conditions identification rates reach a ceiling which conceals the differences between the two stimulus types, making it impossible to compare their relative perceptual effectiveness. Although Strange and Gottfried conceded that isolated vowels do contain sufficient information for accurate identification, they emphasized that the information in static vowel targets may not be "a *necessary* source of information for accurate vowel identification when other dynamic acoustic sources are available" (1980, p. 1625). This conclusion is in good agreement with the conclusion of Assmann (1979) and

Assmann *et al.* (1982), but the source of the *dynamic* information that is perceptually most important remains in question: is it inherent to the vowel, or is it introduced by coarticulation of the vowel with consonants?

"Eliminating" vowel-inherent cues: silent center stimuli

In 1983, Strange, Jenkins and Johnson developed a technique for electronically modifying vowels in order to address this question further. In their 'silent center syllables', the central, most steady-state portion of vowels from CVC context is separated from the consonant transitions. Since the silent center syllable technique isolates the transitions and the vowel nucleus, it becomes possible to estimate the amount of perceptually useful information that each of these sections contains regarding the identity of the vowel. In effect, Strange, Jenkins and Johnson suggested that the silent center syllable technique allows a single token of a CVC syllable to be digitally modified to present either 'isolated vowel' information or 'coarticulatory' information. Using this technique, they compared the value of the central portion of vowels in twenty [bVb] syllables spoken by an adult male, with the value of two dynamic cues to vowel identity, namely formant transitions and syllable duration.

Two kinds of 'isolated center' stimuli and three kinds of 'silent centers' were produced, as summarized in Table 1-4, below. The 'isolated centers' consisted of only the central portion of the [bVb] syllables, with consonant transitions removed. In the 'fixed isolated centers', each vowel nucleus was trimmed to a constant length of 57 ms. The 'variable isolated centers' were also produced by removing the consonant transitions, but the relative duration of each vowel was preserved. The 'silent centers' consisted of consonant transitions only, with the central, nucleus portion of the vowel reduced to silence. In the 'full silent centers' the nucleus was replaced by a silence of the same length as the original center portion, so that the duration of each original syllable was preserved. The 'shortened' and 'lengthened' silent centers were comparable to the 'full silent centers', but with uniformly short (57 ms) or long (163 ms) silent intervals respectively. Listeners also heard the initial and final transitions alone.

For the 'variable' isolated centers and the full silent centers, both of which preserved durational cues, error rates were not significantly different from error rates for the unmodified control syllables. Listener performance on the 'fixed' isolated centers and the lengthened silent centers was significantly worse than for the unmodified control syllables. Neither of these two stimulus types contained durational cues. However, the shortened silent centers, which also lacked durational cues, did not fare significantly worse than stimuli which retained durational information.

Stimulus Type	Duration Cues	Transition Cues	Steady State Cues	Error Rate
Control	Yes	Yes	Yes	1%
Isolated Centers:				
Fixed	No	No	Yes	13%
Variable	Yes	No	Yes	8%
Silent Centers:				
Full	Yes	Yes	No	6%
Shortened	No	Yes	No	7%
Lengthened	No	Yes	No	21%
Transitions:				
Initials	No	Yes	No	47%
Finals	No	Yes	No	46%

Table 1-4: Summary of stimuli used in Strange, Jenkins & Johnson (1983), Experiment I.

In order to assess the value of durational information, the authors made separate comparisons of the silent centers and the isolated centers. They concluded that, when transition information is not present, as in the isolated centers, listeners use durational information to disambiguate between spectrally similar vowels. However, since there were no significant differences in error rates among the three types of silent center syllables, they concluded that duration is not a primary source of information for differentiating vowels when dynamic information is available in the form of consonant transitions. As an explanation for the higher error rates for the lengthened silent centers, they suggested that these stimuli may have begun to lose their integrity as syllables, resulting in an "assymetry of perceptual results" (1983, p. 700).

Error rates for the initial and final transitions presented alone were dramatically worse than for all other stimulus types. The authors contended that these results were also in support of their hypothesis that "dynamic sources of information are sufficient for highly accurate identification of coarticulated vowels" since they expected that "sufficient information for vowel identification was not 'contained within' either of the [transitional] components taken by itself" (1983, p. 699-700). In another sense, though, these results appear to contradict that hypothesis: if dynamic information from the transitions is both supplementary to and more effective than information in the vowel nucleus, we should anticipate that vowels which are bounded on one side by a consonant, as for example in the words "bay" or "Abe", should be better identified than isolated vowels. In other words, initial and final consonants may make some independent contributions to the shape of vowel formants. The assumption that initial and final consonant transitions do make such independent contributions to vowel formant trajectories is supported by the results of several modelling experiments, by Broad and Fertig (1970) and Broad and Clermont (1987). In these experiments, the authors showed that formant frequency contours for vowels in CVC syllables can be predicted by additive models which use a vowel target, plus *independent* initial and final transition functions. In other words, the interaction between initial and final consonants was small enough to be ignored in predictions of the formant contours for CVC syllables. To carry this reasoning further, if silent center syllables act as 'pseudo [bVb] syllables', then stimuli which contain only the initial and final transitions might reasonably be expected to act as 'pseudo [bV] and [Vb] syllables' respectively, in which case identification rates for the initial and final transition stimuli should be better than rates for the centers. As noted above, however, this is not the case: initial and final transitions must occur together, as in the silent center syllables, for error rates to approach those of the centers and the control syllables.

Preservation of vowel-inherent dynamic information in silent center stimuli

A second possible explanation for these results can be proposed based on the results of Assmann (1979) and Assmann *et al.* (1982). If listeners attend to dynamic formant information in vowels, they might make use of initial and final formant targets (as opposed to a single, central target), or some perceptual estimate of vowel-inherent spectral change from the onset to the offset of the vowel, in order to identify vowels in silent center syllables. Assuming that silent center syllables retain some vowel-inherent formant information, listeners could estimate initial and final targets, or the amount and direction of formant movement, from the combined initial and final consonant transitions. Under this hypothesis, the poor performance of the initial and final transitions when they occur alone can be explained by the fact that no estimate of dual targets or vowel-inherent formant movement can be made from either side without the other. Similarly, the higher error rate on the fixed centers, as compared to the variable centers, might result in part from the loss of information regarding vowel-inherent targets, or vowel-inherent spectral change, when the centers are shortened. In summary, the same results would be expected in Strange *et al.* (1983) if the primary source of dynamic spectral information were the vowel itself, rather than coarticulatory information contained in the combined initial and final transitions.

Target extrapolation as a cue to vowel identity in silent center stimuli

A third interpretation of the nature of the dynamic information is possible, namely that the consonant transitions 'point to' static vowel targets. Lindblom (1963) and Lindblom and Studdert-Kennedy (1967) suggested that, when vowel formants fail to reach their expected values in consonant context, listeners make use of formant trajectories in the CV and VC transitions to estimate the intended vowel targets. Strange *et al.* (1983) tentatively rule out this hypothesis since it predicts that, when syllable duration is artificially changed, listeners should alter their estimates of vowel target values and thus make more identification errors. Although lengthening the silent interval resulted in higher error rates in their experiment, shortening it had virtually no effect.

In a 1986 experiment, Verbrugge and Rakerd (also Rakerd and Verbrugge, 1987) attempted to contrast two of these hypotheses, namely that listeners use information in the transition regions to compute missing vowel targets versus the hypothesis that these regions "convey vowel information that is complementary to, and distinct from, target information" (1986, p. 40). To do so, they used 'hybrid' silent center syllables. Like Strange *et al.*'s silent center syllables, Verbrugge and Rakerd's hybrid silent centers contain only information from the initial CV and final VC transitions of [bVb] syllables. Unlike Strange *et al.*'s silent centers however, the initial and final portions of hybrid stimuli come from tokens spoken by speakers of the opposite sex.

In their experiment, listeners heard either unmodified control syllables, silent center syllables, hybrid silent centers, or the initial or final transitions alone. Both types of silent center stimuli preserved durational cues. Error rates are summarized in Table 1-5. All pairwise differences among syllable types were statistically significant except for the difference between silent center and hybrid silent center stimuli. The same results for statistically significant pairwise differences were produced in ar experiment by Jenkins and Strange (1987), which used hybrid [dVd] stimuli.[4] Unlike the silent center experiment by Strange *et al.* (1983), error rates for silent center stimuli in both of these experiments were significantly higher than for control syllables, suggesting that some perceptually important information was lost in the silent centers.

Stimulus Type	Error Rate
Control	8.8%
Silent Center	23.1%
Hybrid Silent Center	27.4%
Initial Transition	56.4%
Final Transition	73.8%

Table 1-5: Summary of error rates for [bVb] stimuli from Verbrugge and Rakerd (1986)

The authors of both papers concluded that listeners most likely do not use the information in transitions to estimate vowel targets, since formant frequencies in the two halves of hybrid syllables are quite dissimilar and, therefore, 'point to' quite different targets. However, this neglects the possibility that listeners may normalize each half of a hybrid syllable independently, with the potential result that the normalized sections point to the same target. These experiments therefore do not provide

[4] However, since Jenkins and Strange did not include a Final Transition condition, this category was not included in their comparisons.

conclusive evidence against a target extraction hypothesis. Nevertheless, the results of both experiments support the hypothesis that transition regions of CVC syllables contain information regarding vowel identity, and suggest that the information is quite robust, in that listeners can use it even when speaker identity changes from the initial to the final portion of the syllable.

Perceptual value of steady state regions for vowel identification

Murphy, Shea and Aslin (1989) also used silent center syllables to investigate sources of vowel information in [bVb] syllables. They compared listeners' use of steady state vowel information with their use of coarticulatory information from formant transitions in synthetic syllables. They tested the ability of children between three and four years of age to distinguish between synthetic tokens of [bæd] and [bʌd] in which either 0%, 10%, 35%, 65% or 90% of the steady state formants were reduced to silence (1989, pp. 376-77). Since the 90% silent center tokens were identified above chance levels (approximately 65% correct) they concluded that their listeners did not use the steady state information remaining in the stimuli to identify these vowels, but rather "extracted the differences in the second and third formant transitions from the full-vowel syllables and recognized the corresponding pattern of formant transitions when the steady-state formant information was eliminated from the silent center syllables" (1989, p. 378). The authors base this conclusion on the fact that the 90% silent center tokens retained only one pitch period (8.3 ms) of vowel steady state in each transition, in combination with their intuition that "it seems unlikely that the auditory system can extract *steady-state formant information* from a single pitch period" (1989, p. 378, italics added).

Their conclusion presumes, however, that a steady state target is the primary item of information, and perhaps the only information which listeners extract from a vowel's formant pattern for the purpose of vowel identification. Since a single period does not constitute a steady state, the 90% silent center tokens retain no perceptually viable 'vowel-inherent' information on vowel identity. This may prove to be true, but other results argue against accepting this conclusion without further evidence. For example, Hyde (1971) found that vowels could be identified from segments only slightly longer than a single glottal pulse, and production data shows that many natural vowel tokens (including isolated vowel tokens) do not contain any steady state (e.g. Potter, Kopp and Green, 1947; Shearme and Holmes, 1962). If listeners extract relevant information from moving formants, then the fact that the 90% silent center vowels contain only one period of steady state may be less important than the fact that these syllables also contain a longer span which approaches the steady state frequency. The question of whether or not listeners favour steady state regions for vowel perception will arise again with regard to research on perception of diphthongs.

So far then, it appears that listeners do extract information regarding vowel identity from the transition regions of CVC syllables, but it is not clear whether this information is introduced by coarticulation of the vowel with consonants, and is therefore different from information used to identify isolated vowels, or whether it is inherent to the vowel itself and therefore, at least potentially, the same as information used to identify isolated vowels.

Identification of isolated vowels in silent center form

If durational cues, coarticulatory information, and the central, most steady state portion of the vowel provide all of the information which listeners use to determine vowel identity, then error rates should be very high for stimuli which contain none of these three cues. Under such an extreme hypothesis, the endpoints of isolated vowels, presented in fixed durations, would be expected to provide no perceptually useful information regarding vowel identity. A 'vowel-inherent dynamic information' hypothesis, on the other hand, predicts that, since stimuli of this nature contain information regarding vowel-inherent spectral change, error rates should approach those for control stimuli.

In 1986, Nearey and Assmann examined isolated vowels using a technique analogous to Strange *et al.*'s silent centers (1983). These 'silent center isolated vowels' allowed the perceptual value of the endpoints of isolated vowels to be evaluated. Nearey and Assmann gated tokens of ten naturally-produced Canadian English vowels spoken by two male and two female speakers using a digital windowing procedure, to produce two 30-ms portions of each vowel, centered at 24% and 64% of the vowel's total duration. The initial portion (the 'nucleus'), and the final portion (the 'offglide') were presented to listeners in three different arrangements. In the natural order condition the nucleus was followed by the offglide and separated from it by 10 ms of silence. In the repeated nucleus condition, the nucleus was presented twice, with an intervening 10 ms of silence; and in the reverse order condition, the offglide was presented first, followed by 10 ms of silence, then the nucleus. Error rates are summarized in Table 1-6.

Error rates for the repeated nucleus and reverse order stimuli were significantly higher than for both the unmodified control syllables and the natural order stimuli. The authors concluded that "these results provide clear evidence that sufficient information is retained for reliable vowel identification in two 30-ms sections taken from the *nucleus* and *offglide* sections of vowels" (1986, p. 1289). Thus, a hypothesis which proposes that there are only three perceptually relevant sources of vowel information: a central target, consonant-imposed formant transitions, and durational cues, must be rejected.

Stimulus Condition	Error Rate
Control	12.5%
Natural Order	14.4%
Repeated Nucleus	31.0%
Reverse Order	37.5%

Table 1-6: Summary of error rates for windowed isolated vowels in Nearey and Assmann (1986)

Nearey and Assmann also suggested that, given the large increase in errors in the repeated nucleus condition, the offglide component must make a significant independent contribution to perception and that, since error rates rise significantly in the reverse order condition, the temporal order of the two sections must be critical. In combination, these findings suggest that information regarding initial and final formant targets or regarding formant movement may, indeed, be perceptually important. In order to further test the hypothesis that listeners' correct identification of the silent center isolated vowels is based on perception of vowel-inherent spectral change, Nearey and Assmann did an acoustic analysis of the stimuli, and used a statistical pattern recognition model to compare patterns found in the acoustic measurements with listeners' identification patterns for the different types of stimuli.

Spectral change in isolated vowels

An evaluation of the acoustic results using t-tests and a Bonferroni multiple comparisons procedure indicated that "the 'nominal monophthongs' [ɪ], [ɛ] and [æ] as well as the 'phonetic diphthongs' [e] and [o] showed significant movement in either

F1, F2 or both" (1986, p. 1299). The average size of the significant formant changes was "considerably larger than difference limens for formant frequency of steady-state vowels given by Flanagan (1955) as about 3% to 5% of the formant frequency" (1986, p. 1300). The acoustic results, therefore, showed that some Canadian English vowels do exhibit perceptually detectable formant movement.

Listeners' use of vowel-inherent spectral change as a cue to vowel identity

Results from the pattern recognition modelling tended to support the hypothesis that vowel-inherent spectral change is perceptually important. Nearey and Assmann used a "normal *a posteriori* probability model" (Nearey and Hogan, 1986) implemented by linear discriminant analysis to develop response profiles for the test vowels from log-transformed frequencies of F0, F1 and F2. In order to evaluate the vowel-inherent spectral change hypothesis, two values each of F1 and F2 were used, taken from the initial and final portions of the vowel, and thus encompassing change in F1 and F2 across the vowel. The response profiles generated by the model on the basis of these formant values provide probability estimates for each vowel token, indicating the likelihood that the token will be correctly identified, as well as a likelihood value for its misidentification as a member of each other vowel category. The model's response profiles were compared with listeners' response profiles for all four stimulus conditions (control, natural order, repeated nucleus and reverse order). The nature of the response profiles allowed the authors to evaluate the model as a predictor of listeners' perceptual behaviour both by examining correct identification rates and by comparing the types of errors made. Product-moment correlations reflecting the "goodness of fit" of the model's overall predictions, as well as separate correlations reflecting the goodness of fit of predictions for individual tokens, were significant for all conditions. Nearey and Assmann also examined how well *changes* in listeners' identification patterns from one stimulus condition to another were correlated with changes in the model's response profiles. Significant positive correlations were found for five of the six comparisons[5].

Difficulties in predicting listeners' identification patterns

The Nearey-Assmann model was, however, somewhat unsuccessful in predicting a general tendency on the part of listeners to misidentify long vowels as short vowels. Attempts to incorporate duration into the model's predictions resulted in over-prediction of short vowel responses (1986, p. 1303). The model also failed to predict accurately the types of errors listeners made in the reverse order condition. More precisely, correct identification rates and error patterns were well-predicted for two 'reverse order' vowels, namely [ɪ] and [e]. The formants of these two vowels are in a similar frequency range, but they move in opposite directions, so that a 'backwards' [ɪ] resembles [e], and vice versa. In the reverse order condition, both listeners and the model tended to mistake [e] for [ɪ] and [ɪ] for [e]. However, the model was not as successful in predicting listeners' error patterns for vowels such as [o], for which the reverse order formant movement is not characteristic of any Canadian English vowel. In addition, the model tended to over-estimate the number of errors made in the reverse order condition, suggesting that it may give "too much weight to frequency change" (1986, p. 1303). This affirms that other cues are perceptually important and must be taken into consideration in order to accurately describe listeners' identification of vowels. Certainly, from the long history of vowel perception research it is evident that formant movement is not essential for identification of all vowels, since listeners can identify some vowels from steady state formants, providing of course that these fall within an appropriate range for the given vowel (e.g. Ladefoged and Broadbent, 1957;

[5] The correlation between the control and normal order conditions was actually negative, but the authors note that since identification rates in these conditions were quite high, this correlation was based on relatively small changes in the response profiles.

Fry *et al.*, 1962). However, other experiments show that certain vowels are difficult to identify from steady state formants (e.g. synthesized steady-state vowels in Diehl *et al.*, 1981 and in Ryalls and Lieberman, 1982) and it seems likely from the experimental results considered above that formant movement can provide a perceptually useful cue to vowel identity.

Cues to perception of diphthongs

Another body of literature which supports the notion that formant movement can be of perceptual importance for vowel recognition is the literature related to perception of diphthongs. Not surprisingly, there is good agreement in this literature that vowel-inherent formant movement is necessary for diphthong perception. However, the exact nature of the cues to which listeners attend in order to identify diphthongs remains under debate. In their paper, Nearey and Assmann evaluated the adequacy of three hypotheses of diphthong perception as explanations of listeners' identification of the silent center vowels used in their experiment. They found that all three hypotheses were "adequate to characterize the main tendencies in listeners' responses in the perceptual experiment" (1986, p. 1305), even though the vowels they examined were not phonemic diphthongs.

It seems reasonable to assume, as did Nearey and Assmann, that the findings of experiments on perception of formant movement in diphthongs may also have implications for perception of formant movement in other vowels. Briefly, then, the three hypotheses considered by Nearey and Assmann were (i) the dual target hypothesis, under which listeners attend to the actual frequencies of the relatively steady state endpoints of diphthongs, and ignore the transition region between them; (ii) the target plus slope hypothesis, which holds that the starting frequencies of the diphthong formants and the rate, or slope of frequency change are perceptually important; and (iii) the target plus direction hypothesis, which states that the initial target and the direction of formant movement are important, but rate, or slope of movement and the exact frequency of the final target are not. Experiments on diphthong perception provide varying amounts of support for each of these hypotheses.

Wise (1964) and Bladon (1985) interpret their work as supporting a dual target hypothesis. In their experiments, listeners were able to identify diphthongs when presented with the endpoints only. This was true whether the transition region was replaced with silence or eliminated entirely, so that the two endpoints were temporally adjacent. These results might also be interpretable as supportive of a target plus direction hypothesis, since the direction of formant movement may be determined from the two endpoints.

Gay (1970) and Bond (1982) found that listeners could identify diphthongs which did not have any initial or final steady state target. Gay interpreted these results as indicating that transitional rate of change (or slope) is the most important perceptual cue. However, the results of experiments by Wise (1964), Bond (1982) and Bladon (1985) present evidence that goes against a slope hypothesis. In these experiments, slope was varied by maintaining the same initial and final target values, and changing the duration of the transitions from as little as 0 ms (Wise and Bladon) to as much as 160 ms (Bond). Correct identification rates reached between 75% and 100% both when the slope was very steep, as in the 0 ms transitions, and when it was very shallow, as in the 160 ms transitions. Bladon (1985) argues strongly against Gay's 'rate of change' hypothesis. As he suggests, any hypothesis which completely ignores the contribution of formant frequencies seems unlikely to account well for listeners' identification of vowels. Gay, however, does not rule out the contribution of actual formant frequencies, but instead relegates them to a position of secondary importance.

Based on her finding that identification rates were high either when the transition duration was relatively long (approximately 70 to 160 ms) or when the steady state target components were relatively long (approximately 20 ms or more), Bond concludes that "the perceptual requirement for identifying a synthetic token as a diphthong may be variable" (1982, p. 203). Given that identification rates were equally good when glide durations were long as when steady states were long, it would seem that listeners do not have a strong perceptual preference for the steady state portions of vowel formants.

Interpretation difficulties in diphthong-perception experiments

In general, the results of these experiments tend to support either a dual target or a target plus direction hypothesis. However, there are obvious difficulties in defining a target in perceptual terms, since listeners appear not to require steady-state targets for diphthong perception. In addition to this problem, difficulties in interpretation tend to arise due to the small number of stimuli used, and restrictions on stimulus types: none of these experiments includes a sufficient variety of stimuli to test all three of the hypotheses proposed above. Further difficulties arise from the complexity of the problem under investigation. In experiments such as these it would seem that the most reliable method of interpretation is through the use of well-defined computer models of perception, such as those used by Assman, Nearey and Hogan (1982), and Nearey and Assmann (1986). Without such well-defined models, the process of predicting listeners' responses to stimuli which vary along several dimensions becomes largely guesswork, and therefore unverifiable; yet unless dependable predictions can be made, hypotheses cannot be subjected to rigorous tests.

F1 onset frequency and timing of F1 maximum as cues to vowel identity

Another, somewhat different approach to the question of what role formant movement plays in vowel perception is taken by Di Benedetto (1989*a* and *b*). This approach was anticipated to some degree by Lehiste and Peterson (1961), by Strange, Jenkins and Johnson (1983), and by Strange (1989). In a production study of six vowels, Di Benedetto examined the temporal and spectral properties of F1 in two neutral contexts, [Vd] and [hVd], and in twelve consonant contexts, in an effort to find "some different or additional cues to a single [i.e. invariant] value of F1 that could supply the listener with the information necessary to identify the vowel" (1989*a*, p. 61). Di Benedetto concludes that certain vowels are better discriminated by the onset value of F1 and the length of time from F1 onset to maximum F1 frequency, than by their F1-F2 structure.

Her interpretation of the results is problematic for several reasons, three of which place her arguments in serious jeopardy. First of all, her conclusions are based primarily on results for only two of the six vowels she examined. For the other four vowels, results were largely inconclusive. Secondly, conclusions regarding these two 'good' vowels ([ɪ] and [ɛ]) rely on differences in timing of the maximum F1 frequency. For [ɪ], choosing the time when F1 reached its maximum frequency appears to have been unproblematic, since the example trajectories for [ɪ] (1989*a*, Fig. 8, p. 61) show clear upward movement throughout their entire course. For [ɛ], however, choosing a time for the F1 maximum appears to have been very subjective. The F1 trajectories which Di Benedetto provides as examples of [ɛ] are relatively flat, showing overall rises of less than about 30 Hz in three out of four cases, with no obvious peak. Although the author places the F1 maximum for [ɛ] at between 30 and 40 ms, the sample trajectories (1989*a*, Fig. 9, p. 62) would appear to allow much later placement of the F1 maximum (making [ɪ] and [ɛ] more alike, rather than more distinct) and in fact to cite any time point as an F1 "maximum" seems to be endowing the [ɛ] trajectories with a property to which they have no claim. Finally, average identification rates for [ɪ] and [ɛ] from linear discriminant analyses improved by only 1% with the inclusion of the relevant temporal and spectral information, from an average of 99% correct using F1-F2 values, to 100% correct using either timing and F1 maximum, or onset frequency and F1 maximum.

In a companion paper, Di Benedetto conducted several perceptual experiments to further test the conclusions reached in the production study. Her stimuli were synthetic tokens of [dVd] syllables, in which a range of frequency values were used for the F1 onsets, steady states, and offsets. In the first set of stimuli, F1 had relatively long onset transitions (70 ms) and shorter offset transitions (30 ms). The second set were mirror images of the first, with short onsets and long offsets. The vowels in each set had a steady state of 15 ms, and in both sets the 70 ms transitions terminated at lower frequencies than the 30 ms transitions. Although the stimuli were designed to test the effect of onset frequency and position of the F1 maximum on vowel identification, the resulting patterns are also good representations of vowels that show upward movement across the course of F1 (for example, [ɪ] in the previous study) versus vowels that show downward movement across the course of F1 (for example, [e]). Since Di Benedetto's theory of F1 onset frequency and position of F1 maximum has already been shown to be problematic, the experimental results will be discussed here only in terms of their ability to support or refute a vowel-inherent spectral change theory of vowel identification.

Those stimuli which had a long onset and short offset, resulting in a rising pattern for F1, were heard by all four English subjects as either [i] or [ɪ] until the F1 steady state reached a frequency of approximately 450 Hz.[6] Two of the four subjects continued to identify these stimuli as [i]-[ɪ] at higher frequencies. Since a rising F1 is characteristic of [ɪ], these results appear to fit well with a vowel-inherent spectral change theory. The other two subjects identified the remaining 'high F1' stimuli as [e]-[ɛ]. The steady state value of F1 in these tokens was between 470 and 500 Hz, and was therefore well above the average value of F1 for a male speaker. The average value of F1 for [ɪ] for a male (Peterson and Barney, 1952, as cited in Borden and Harris, 1984, p. 107) is 390 Hz. In comparison, the average F1 for [ɛ] is 530 Hz. Thus, the F1 values of these stimuli are approaching values for [ɛ].

The 'falling F1' tokens were consistently identified as [e]-[ɛ] when the F1 steady state was above 430 Hz,and consistently as [i]-[ɪ] when F1 was below 370 Hz. Since Di Benedetto does not distinguish between [i]-[ɪ] or [e]-[ɛ] responses, these results are somewhat difficult to interpret. At best, they again appear to provide mild support for a vowel-inherent spectral change theory, in that a falling F1 is more often heard as [e]-[ɛ] than as [i]-[ɪ]. These results do tend to support the further conclusion that the appropriate direction of formant movement will not necessarily 'bring a vowel in' to a given category if the formant frequencies are outside the expected range.

In another experiment in this same paper, Di Benedetto examines listeners' identification of four synthetic [dVd] stimuli which had an abrupt F1 onset (60 Hz rise in 10 ms), and F1 offsets that varied in duration from 30 to 85 ms. Total duration of each of these vowels was 120 ms, so that those vowels with a short duration offset had

[6] Di Benedetto groups together [i]-[I] responses, as well as [e]-[ɛ] responses, thus no further detail can be reported.

a long steady state, and vice versa. Listeners heard all of these stimuli as [e], except for the token with the longest steady state, which was heard as a vowel somewhere between [e] and [ɛ] 40% of the time and as [e] the rest of the time (1989b, pp. 73-74). Each of the three stimuli which were consistently identified as [e] showed a falling pattern in F1, characteristic of the vowel [e]. In the remaining stimulus, the long steady state may account for its identification as a cross between [e] and [ɛ], since a steady state F1 appears to be characteristic of [ɛ] in this dialect, as discussed above. Listeners may have been dissuaded from hearing the long steady state vowel as [ɛ] more often because of a very high F2. In fact, the F1-F2 space is undersampled in all of the perceptual experiments in Di Benedetto, 1989b, which makes interpretation of the results more difficult. At 2800 Hz, the frequency of F2 is higher than expected for [ɛ] when spoken by an adult male. The average F2 for [ɛ] for a male speaker is 1840 Hz, (Peterson and Barney, 1952, as cited in Borden and Harris, 1984, p. 107). Two other stimuli were also tested in this experiment. The first of these had an F1 onset which rose 150 Hz over 90 ms, and an offset that dropped 100 Hz over 15 ms. This stimulus was always heard as [ɪ]. As discussed above, [ɪ] tends to show a rising F1, so again, this result tends to support a vowel-inherent spectral change hypothesis. The onset and offset of the last stimulus were symmetrical about a 20 ms steady state. This token was consistently identified as [ɪ] by two subjects, and occasionally identified as [e] by the third. Since it contained identical rising and falling components, its consistent identification as [ɪ] by two out of three subjects may indicate that the onset portion, in which F1 was rising, has a somewhat greater perceptual weight than the offset portion, which in this case was falling.

Overview

To summarize, listeners are able to identify vowels in CVC syllables when they hear only the initial and final transitions from these syllables. It has been proposed that CV and VC transitions contain 'additional' information regarding vowel identity, i.e. information not available in isolated vowels. It has also been proposed that this information is perceptually superior to information which vowels themselves provide regarding their identity. However, other experiments have shown that isolated vowels can be as well identified as vowels in consonant context, and also that isolated vowels can be identified, like vowels in CVC syllables, from brief sections taken from their endpoints. The possibility remains, then, that the information which listeners extract from the endpoints of both isolated vowels and vowels in consonant context is the same kind of information, perhaps providing listeners with details of formant movement that is inherent to the vowel. Listeners have been shown to attend to formant movement in the perception of diphthongs, although details regarding the nature of the cues they attend to have not been agreed upon. Finally, experiments done in support of an 'F1 onset and F1 maximum' theory of dynamic perceptual information for vowel identification have not produced convincing evidence in support of that theory, but have produced results which suggest that the direction of F1 movement may help listeners distinguish between vowels with similar F1 frequencies.

Purpose and design of this study

To this point, no acoustic analyses have been performed to investigate the possibility that vowel-inherent spectral change persists in the trajectories of CVC syllables. Similarly, no silent center experiments have been done which directly compare listeners' identification of vowels in consonant context with their identification of isolated vowels. If vowel-inherent spectral change does persist in consonant context, and if listeners use it as a cue to vowel identity both in silent center CVC and in silent center isolated vowel stimuli, then a direct comparison of perceptual results for these two types of stimuli should reveal similarities in listeners' identification rates and identification patterns. More specific evidence regarding the nature of the formant cues used by listeners in each case may also be provided by the Nearey-Assmann model. If listeners consistently rely on vowel-inherent spectral change as a cue to vowel identity in silent center stimuli, then the Nearey-Assmann model should predict listeners' identification patterns relatively well for both CVC syllable and isolated vowel stimuli. However, if listeners use different cues to vowel identity when vowels appear in consonant context, error rates, listeners' identification patterns, and the Nearey-Assmann model's predictions should show differences when compared with error patterns, identification rates, and predictions for isolated vowels.

The above approach will be used in this thesis to investigate listeners' perception of vowels in one silent center consonant context, namely [bVb] context. Listeners' perception of silent center isolated vowels will also be tested and compared with results for the [bVb] stimuli, in order to estimate the degree to which listeners use the same cues to identify the vowels in both cases. The first step, investigating perseverence of vowel-inherent spectral change cues in [bVb] context, is presented in the next chapter.

CHAPTER 2

Acoustic Evidence for Vowel-Inherent Spectral Change in [bVb] Syllables (Experiment 1)

It is reasonable to ask whether vowel-inherent spectral change persists in consonant context, before proceeding to compare listeners' identification of vowels in isolation with their identification of vowels in consonant context. Nearey and Assmann (1986) tentatively suggest, on the basis of formant tracks for vowels in several consonant contexts, that vowel-inherent spectral change may, indeed, remain, at least in some consonant contexts. In this section, acoustic evidence will be examined in order to evaluate whether or not vowel-inherent spectral change persists when vowels are produced in [bVb] context.

Nearey and Assmann define vowel-inherent spectral change as "the relatively slowly varying changes in formant frequencies associated with vowels themselves, even in the absence of consonantal context" (1986, p. 1297). As a parametric representation of vowel-inherent spectral change, they measured change in F1 and F2 from the initial to the final portion of the vowel. The same measure of vowel-inherent spectral change will be used here, and the amount and direction of formant movement found in vowels in [bVb] syllables will be compared with the movement found in isolated vowels by Nearey and Assmann.

Method

Ten vowels, (i,ɪ,e,ɛ,æ,ʌ,ɒ,o,ʊ,u), produced in [bVb] context by three male and three female speakers of Western Canadian English were recorded in random order in a sound-treated room using the left channel of a Sony TC-K55II stereo cassette deck and a Sennheiser MD 42 1N cardioid directional microphone. All six speakers spoke the same dialect of Canadian English as the speakers whose isolated vowels were analyzed by Nearey and Assmann. The recorded [bVb] syllables for four of the speakers (two male and two female) were bandpass filtered through a Rockland Programmable Dual Hi/Lo Filter (Series 1520) set to 80 to 8000 Hz, and then digitized at 16 kHz on a Digital PDP-12A minicomputer, using the University of Alberta Department of Linguistics' Alligator system (Stevenson and Stephens, 1979). [bVb] syllables for the remaining two speakers were also bandpass filtered and digitized at 16 kHz, but tokens for these two speakers were digitized on a Zenith Data Systems 286 personal computer, using software developed for CSRE (the 'Canadian Speech Research Environment' project, see Jamieson *et al.*, 1989).

In order to select formant candidates from the digitized syllables, a 30 ms Hamming window was advanced 4 ms for each analysis frame. Each frame was first order pre-emphasized using a coefficient of 0.98. Autocorrelation LPC analysis, using a range of 0 to 4000 Hz with 10 predictor coefficients, was used to estimate a smoothed spectrum. Formant candidates from the analysis were displayed graphically on the computer screen in a form similar to a wideband spectrogram. Formant measurements were made from the graphic display by clicking the computer mouse on the chosen formant candidate. Initial and final measurements of both F1 and F2 were taken for each vowel. The formant frequency, bandwidth and amplitude, as well as the time of each formant measurement point, were automatically recorded in a log file, and the logged values were checked for appropriateness. Formant measurements were taken as early and as late in each vowel as possible, subject to the following three restrictions: 1) initial measurements were taken in the first half of the syllable, but at least 40 ms after the initial burst, 2) final measurements were taken in the second half of the syllable, but before the rapid, final consonant transitions, and 3) all measurements were taken at points where the amplitude and bandwidth were within approximately 5 dB and 60 Hz of the amplitude and bandwidth at the syllable peak. In general, if initial measurements had been taken earlier in the syllable, and final measurements had been taken later, estimates of formant movement would have been somewhat greater than those provided here. For a few vowels in which F1 and F2 were closely adjacent, the above analysis did not provide adequate resolution of the formants. These vowels were re-analyzed with the analysis area restricted approximately to the frequency range of F1 and F2, with fewer predictor coefficients, using Markel and Gray's (1976) method for selective linear prediction.

Analysis and Results

In order to evaluate the degree to which formant movement found by Nearey and Assmann in isolated vowels persists in vowels in the recorded [bVb] syllables, tokens were grouped according to the expected direction and amount ('significant' or 'insignificant') of movement in F1 and F2. Average change (in hertz) in F1 and F2 was calculated for each of the six speakers, across the vowels within each test group. One-tailed t-tests were performed on the subject means of vowel groups in which significant upward or downward movement was expected, and two-tailed tests on the subject means of groups in which no significant movement was expected in either direction. Results are summarized in Table 2-1.

Because of the small number of subjects involved in this measurement study, no firm conclusions can be drawn regarding the hypothesis that vowel-inherent formant movement persists in [bVb] syllables. However, the results do provide tentative support for that hypothesis. All four groups in which significant upward or downward movement is expected show significant movement in the expected direction were taken at points where the amplitude and bandwidth were within approximately 5 dB and 60 Hz of the amplitude and bandwidth at the syllable peak. In general, if initial measurements had been taken earlier in the syllable, and final measurements had been taken later, estimates of formant movement would have been somewhat greater than those provided here. For a few vowels in which F1 and F2 were closely adjacent, the above analysis did not provide adequate resolution of the formants. These vowels were re-analyzed with the analysis area restricted approximately to the frequency range of F1 and F2, with fewer predictor coefficients, using Markel and Gray's (1976) method for selective linear prediction.

Analysis and Results

In order to evaluate the degree to which formant movement found by Nearey and Assmann in isolated vowels persists in vowels in the recorded [bVb] syllables, tokens were grouped according to the expected direction and amount ('significant' or 'insignificant') of movement in F1 and F2. Average change (in hertz) in F1 and F2 was calculated for each of the six speakers, across the vowels within each test group. One-tailed t-tests were performed on the subject means of vowel groups in which significant upward or downward movement was expected, and two-tailed tests on the subject means of groups in which no significant movement was expected in either direction. Results are summarized in Table 2-1.

Because of the small number of subjects involved in this measurement study, no firm conclusions can be drawn regarding the hypothesis that vowel-inherent formant movement persists in [bVb] syllables. However, the results do provide tentative support for that hypothesis. All four groups in which significant upward or downward movement is expected show significant movement in the expected direction

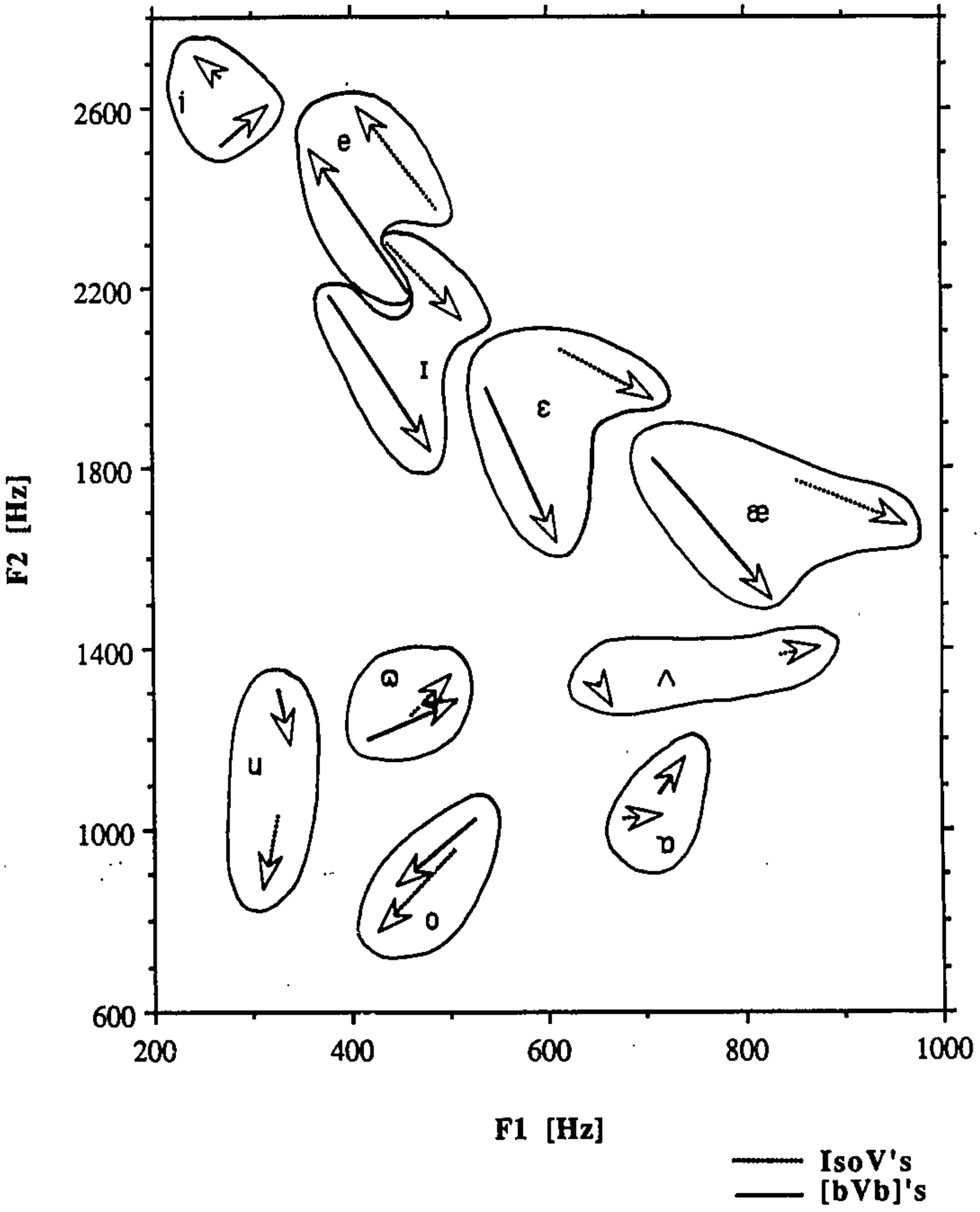
($p \leq 0.05$). The groups in which no significant movement is expected (from prior work on isolated vowels) should fail to reach significance, and this is the case for F2. The F1 group in which no significant movement is expected (again from prior work on isolated vowels) does, however, reach significance at the $p \leq 0.05$ level, indicating that more movement is present than predicted. The sample mean indicates that vowels in this group display a tendency to move upward in F1. This tendency to move upward is not totally unexpected, since several vowels in the group are lax vowels. The amount of F1 movement for individual vowels in this group can be seen in Figure 2-1.

Vowel Group and Expected Movement	Sample Mean	t	d.f.	Probability
F1 Significant downward (e,o)	-90.7	-4.594	5	≤ 0.0029
F1 No significant movement (i,æ,ʌ,ɒ,ɷ,u)	52.5	2.589	5	≤ 0.0489*
F1 Significant upward (ɪ,ɛ)	88.3	3.098	5	≤ 0.0135
F2 Significant downward (ɪ,ɛ,æ,o)	-286.0	-3.802	5	≤ 0.0063
F2 No significant movement (i,ʌ,ɒ,ɷ,u)	22.2	1.045	5	≤ 0.3440*
F2 Significant upward (e)	316.5	6.336	5	≤ 0.0007

Table 2-1: Results of t-tests on direction and amount of formant movement. Vowels are grouped according to the results of acoustic measurements and t-tests done on formant movement in isolated vowels by Nearey and Assmann (1986). Categories marked with asterisks were tested with two-tailed t-tests, all other categories were tested with one-tailed t-tests.

Figure 2-1 provides a graphic representation of the average change from initial to final F1 and F2 for each of the ten vowels examined. In this figure, formant movement in the [bVb] context vowels is compared with formant movement in the equivalent isolated vowels. Isolated vowel values were calculated from tokens spoken by two males and two females from the same group of speakers as analyzed for the

Figure 2-1: Mean Initial and Final F1-F2 Values



[bVb] syllables. Formants in the isolated vowels were also measured in the same manner as those in the [bVb] syllables.

From Figure 2-1, it can be seen that [ɪ,ʌ,ɒ] and [ʊ], four of the six vowels expected to have no significant F1 movement in either direction are, indeed, relatively 'stable' in F1. However, the remaining two vowels in this group, [æ] and, to a lesser degree, [ɷ] show upward F1 movement in [bVb] context. Upward movement was also found in F1 for isolated versions of these two vowels by Nearey and Assmann (1986), but it failed to reach significance. Some upward F1 movement is also apparent in the isolated versions of these vowels for the four subjects considered here. In fact, the isolated version of [æ] shows upward F1 movement of approximately the same magnitude as its [bVb] counterpart. Similar movement can apparently be found in these two vowels in American English: Klatt provides dual F1 targets for synthesis of both [æ] and [ɷ], (1980, p. 986). The Klattalk program for speech synthesis by rule (which forms part of Digital Equipment Corporation's commercial DECTalk text-to-speech system) also includes "schwa offglides" for lax vowels (1987, p. 756). Hence, the *direction* of F1 movement in these two [bVb] context vowels appears consistent with findings on F1 movement for isolated vowels. A larger sampling of speakers might indicate that the *amount* of F1 movement is also comparable in isolation and in [bVb] context.

In general, a visual comparison of the isolated vowels and their [bVb] counterparts in Figure 2-1 suggests that formant movement which is very similar to that found in isolated vowels persists in [bVb] context. Indeed, even this small sample provides a good example of how vowel-inherent spectral change may contribute to the accuracy of vowel perception: vowels which have either a high F1 or high F2 in isolation tend to show lower values for that formant in [bVb] context. This results in noticeable overlap in the F1-F2 space for the vowels [ɪ] and [e], which would be difficult to distinguish based on a single formant slice. However, when the direction of formant movement is taken into consideration, they are easily distinguished.

Conclusion

In summary, this study shows that spectral change in the same direction, and of approximately the same magnitude, as found in isolated vowels does persist in Western Canadian English vowels produced in [bVb] context. A larger study could provide useful verification and expansion of this conclusion by analyzing vowels for a larger group of speakers, by examining vowels in a wider variety of consonant contexts, and by examining vowels produced by speakers of other English dialects.

Since cues to vowel-inherent spectral change do seem to persist in the endpoints of [bVb] syllables, listeners may use them to identify vowels in silent center [bVb] syllables, rather than using dynamic information that is introduced by coarticulation of the vowels with consonants. This possibility will be investigated in the second experiment.

CHAPTER 3

Perception of Silent-Center Isolated and [bVb] Context Vowels (Experiment 2, Part a)

In this experiment, silent center stimuli are used to test two extreme hypotheses of vowel perception: an extreme vowel-inherent view, which states that exactly the same vowel-inherent cues are used to identify vowels in isolation and in consonant context; and an extreme coarticulatory view, which states that coarticulation of vowels with consonants introduces powerful new cues to vowel identity, and when these are available (as in all CVC syllables) vowel-inherent cues are ignored in favour of the coarticulatory cues. It should be noted that these extreme views are not necessarily held by any researcher or group of researchers. The extreme hypotheses do, however, provide a convenient starting point for comparing listeners' identification of vowels in isolation and in consonant context.

Method

In the silent center stimuli used here, as in those described previously, the traditional vowel 'target', in other words the central, most steady-state portion of the vowel, is replaced by silence. In addition, in this experiment durational cues to vowel identity are neutralized. This leaves only those cues to vowel identity which reside in the temporal endpoints of the stimuli. The temporal endpoints of isolated vowels provide information regarding vowel-inherent spectral change. Identification rates and error patterns for silent center isolated vowels can be predicted relatively well based only on this vowel-inherent information (Nearey and Assmann, 1986). The results of Experiment 1 suggest that the endpoints of vowels in at least one type of syllable ([bVb] syllables) also retain information regarding vowel-inherent spectral change. If so, information on vowel-inherent spectral change may act as a useful perceptual cue both in silent center isolated vowels and in silent center [bVb] stimuli. On the other hand, [bVb] syllables contain a second type of information, namely coarticulatory information, which may provide listeners with a separate, possibly more powerful cue to vowel identity, as suggested by Strange and others (Strange *et al.*, 1976; Verbrugge and Rakerd, 1986; Murphy *et al.*, 1989). Since coarticulatory information is distinct in nature from vowel-inherent information, vowels identified on the basis of coarticulatory cues could exhibit differences in identification patterns when compared with vowels identified on the basis of vowel-inherent cues. Hence, if listeners rely on coarticulatory information to identify vowels in silent center [bVb] stimuli, but on vowel-inherent information to identify silent center isolated vowels, the identification rates and error patterns for these two stimulus types may be quite different.

In addition to silent center stimuli, listeners' perception of *hybrid* silent center syllables is tested in this experiment. In hybrid stimuli, speaker identity changes from the initial to the final portion of the syllable. If listeners are able to identify isolated vowels in hybrid silent center form, and at rates equivalent to those for hybrid silent center [bVb] syllables, we may not need to postulate (as Verbrugge and Rakerd did) that the information which listeners use to identify vowels in hybrid [bVb] syllables must be "defined over the syllable as a whole" (1986, p. 56).

Stimuli

Ten vowels (i,ɪ,e,ɛ,æ,ʌ,ɒ,o,ɷ,u), produced in isolation and in [bVb] context by two male and two female speakers of Western Canadian English were used to make the silent center stimuli for this experiment. The vowel [ɑ], which raised error rates in Strange *et al.* (1976) through confusion with [ɒ], was not included in this experiment. All tokens were recorded and digitized on the PDP–12A minicomputer for Experiment 1 above. Markers were added manually to the digitized tokens for use by an automatic windowing program. For the [bVb] syllables, markers were placed at the initial burst release and final labial closure. Markers were added to the isolated vowels at vowel onset and vowel offset. The silent center stimuli were produced as shown in Figure 3-1. The initial portions of the [bVb] syllables were windowed with a 30 ms plateau, followed by a 10 ms down ramp in order to produce the "heads" of the silent center [bVb] stimuli. For the heads, the beginning of the plateau section was aligned with the burst of the initial [b]. A mirror image of this window was used to produce

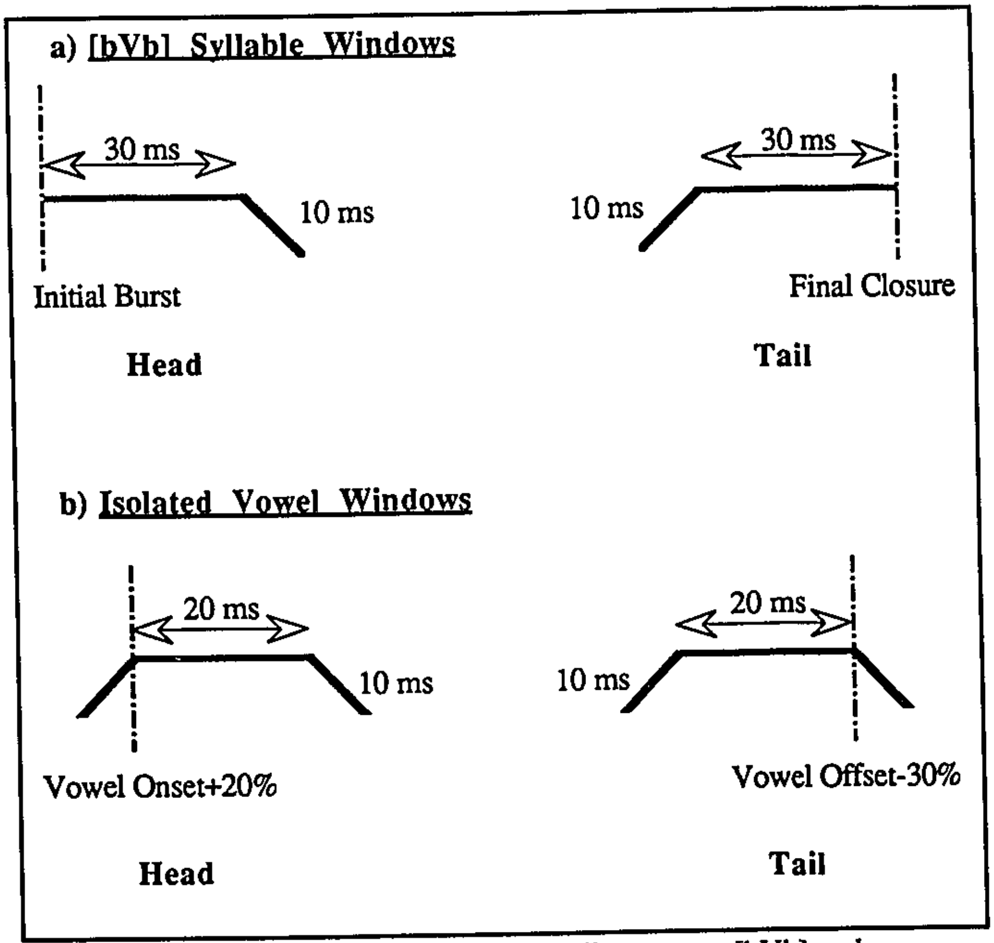


Figure 3-1: Windowing functions for the silent center [bVb] and isolated vowel stimuli

the final, or "tail" sections of the silent center [bVb] stimuli. The plateau for the tails was aligned with the time of labial closure for the final [b] (see Figure 3-1a).

For the isolated vowels, trapezoidal windows were employed. Total window length was again 40 ms, but in this case the windows consisted of a 10 ms up ramp, a 20 ms plateau, and a 10 ms down ramp. Since isolated vowels provide no clear time alignment points for windowing, windows were aligned at fixed proportions of the total syllable duration. The alignment points were selected so as to avoid the relatively low amplitude sections at the beginning and end of the isolated vowels. For isolated vowel heads, therefore, the beginning of the plateau was aligned to a point 20% of the total duration from the onset of the vowel. For the tail sections, the end of the plateau was aligned to a point 30% of the total duration from the offset (Figure 3-1b).

To assemble the silent center stimuli, the resulting 40 ms head and tail sections were concatenated with an intervening 150 ms of silence, yielding an overall stimulus duration of 230 ms. The hybrid silent center stimuli were produced by combining the head of each isolated vowel and [bVb] token with the tail of that vowel for each of the other three speakers. A total of 160 silent center tokens were constructed for both the isolated vowels and the [bVb]'s (4 head speakers X 4 tail speakers X 10 vowels).

Two separate tapes were made for the perceptual test, one for the isolated silent center vowels and one for the silent center [bVb] syllables. Each tape was approximately twenty minutes long. Stimuli were recorded in random order with an interstimulus interval of five seconds. After every fifth stimulus there was a six second pause and a tone.

Response Task

Listeners' responses were recorded on specially-designed answer sheets. Each line of the answer sheets consisted of ten spelled words representing the ten vowels in the study ("heed, hid, hayed, head, had, hud, hawed, hoed, hood, who'd"). Thus, the task involved was a keyword task for both stimulus types. Lines on the answer sheets were grouped into sets of five, to coincide with the grouping of stimuli on the listening tapes. The columns of spelled words were labelled, above each group of five lines, with the phonetic symbol for the appropriate vowel. A sample answer sheet page is provided in the Appendix. For each stimulus, subjects were asked to cross out the word on the appropriate line, that contained the vowel sound they thought was most like the vowel just heard.

Subjects

Subjects for the experiment were volunteers enrolled in an introductory linguistics course at the University of Alberta. All had approximately two weeks of training in the phonetic transcription of Canadian English sounds, and all were native speakers of Canadian English. Separate groups of subjects heard the isolated vowel and [bVb] stimuli. Each stimulus group heard ten practice stimuli from one tape. After the practice, subjects were given time to ask any other questions they had, and were then tested using the other tape. Testing took place in a quiet classroom. The tapes were played at a comfortable listening level, using a Sony Stereo Cassette Deck, Sony Integrated Amplifier, and a Heco Sound Master 15 Speaker. Twenty subjects heard the silent center [bVb] stimuli, and nineteen heard the silent center isolated vowel stimuli. Five subjects were dropped before the analysis since they apparently lost their place on the answer sheets and responded to less than 90% of the stimuli. This left seventeen subjects in each group.

Since identification rates for the first group of subjects were lower than expected (53.4%), the experiment was run a second time. Subjects in the second group were again phonetically trained volunteers who were native speakers of Canadian English, but all had completed at least one linguistics course, and most were either undergraduate majors in linguistics, or graduate students in linguistics. The twelve subjects in this group were tested on both types of stimuli. In order to improve listening conditions, subjects were tested individually in a quiet office. Stimuli were played on a Tandy Cassette Recorder, over Koss headphones. The tapes used were the same as those used for the first subject group, but in order to allow more time for responses, subjects in the second group were told they could stop the tape if necessary. They could not, however, listen to a stimulus more than once. As with the first group, this group of subjects first heard ten practice stimuli from one tape. After the practice they were given time to ask further questions, and were then tested on the other tape. Following a break of at least fifteen minutes, they were tested on the second tape. The order of the test tapes was randomly varied for subjects in this group, with the condition that each tape was played first for an equal number of subjects.

Results and Discussion

Once testing was complete, a check of the recorded stimuli revealed that the heads and tails of three [bVb] tokens were of questionable quality, due to errors in digitization and the placement of markers for windowing. These errors were untraceable, since the PDP-12 computer, on which digitization and windowing was done, had been retired in the intervening period of time. In order to determine whether or not the quality of these three stimuli was sufficiently poor to drop them from the analysis, the range of identification rates for tokens that contained the 'problem' heads and tails was compared with the range for other tokens of the same vowel. The results of this comparison are presented in Table 3-1. None of the differences was considered large enough to justify dropping stimuli from the analysis. The figures presented in the discussion that follows therefore include data for these tokens.

Speaker Number and Syllable	Range for Problem Tokens	Range for Good Tokens
Speaker 1 [bɛb]	5 to 9	5 to 11
Speaker 3 [bib]	6 to 10	6 to 12
Speaker 4 [bʌb]	4 to 12	6 to 12

Table 3-1: Ranges of identification rates, out of a possible 12, for silent center [bVb] tokens which contained a poor quality head or tail, compared with identification rates for tokens of the same vowel which did not contain the poor quality head or tail. Seven tokens of each syllable contained the problem head or tail, and nine tokens did not. All figures are for subject group 2.

Identification rates, averaged over all stimuli, were 53.4% for the first subject group and 68.8% for the second group. In comparison, identification rates in two other silent center experiments, for [bVb]'s (in Strange *et al.*, 1983), and for isolated vowels (in Nearey and Assmann, 1986), were substantially higher, at 86% and 85.6% respectively. Identification rates for Verbrugge and Rakerd's hybrid silent center [bVb] syllables were closer to rates for this experiment, at 72.6% (1986).

Identification rates here may have been affected by a number of factors. Subject training was less extensive in this study than in some similar studies. In both Strange and Gottfried (1980), and Strange (1989), subjects heard five blocks of ten or eleven "familiarization" stimuli. Feedback was given after each block in Strange and Gottfried

(1980, p. 1623) and after the first and fourth blocks in Strange (1989, p. 2138). In Strange (1989), approximately one third of the subject group was dropped because they failed to pass the familiarization criteria. Although subject training procedures are not described in detail by Strange *et al.* (1983), they may have followed a similar paradigm. Verbrugge and Rakerd's subjects heard a "demonstration sequence" of 22 control tokens, then responded, without feedback, to two randomized "practice blocks" containing all 22 of the test stimuli (1986, p. 47). Subjects in this experiment heard one practice block, containing only 10 of the 320 test stimuli, and received no feedback.

Hybrid syllables may also be more difficult to identify than single-speaker syllables. Experiments have consistently shown that listeners perform better on vowel identification tasks when tokens are blocked by speaker, rather than randomized by speaker (Strange *et al.*, 1976; Strange and Gottfried, 1980; Assmann *et al.*, 1982). In this experiment, speaker identity varied randomly not only from token to token, but 'within' tokens as well. In a similar vein, Mullennix and Pisoni (1988), and Johnson (1988) report that listeners' response latencies on word identification tasks increase from single-talker to multiple-talker conditions. Johnson suggests that this "reaction time disadvantage" results from perceptual adjustments that are required when a new voice is encountered (1988, p. 256).

The retention of durational cues to vowel identity in the experiments cited above, and the lack of durational cues here can also be expected to contribute to differences in identification rates. Both Strange *et al.* (1983) and Verbrugge and Rakerd (1986) based their windowing procedure on proportions of individual tokens, so that the duration of each head, tail and silent interval varied according to the duration of the original token. When silent interval durational cues were neutralized by lengthening the silence to a uniform 160 ms in Strange *et al.*, error rates increased from 6% to 13% when speakers were blocked, and from 14% to 24% when speakers were randomized. In comparison, the windowing procedure used for this experiment completely removed durational cues, and this, in combination with a silent interval of 150 ms (approaching the 160 ms duration of the "lengthened" silent interval in Strange *et al.*) very likely contributed to lower identification rates.

Statistical Analysis

An analysis of variance conducted on the results of the first subject group, with three factors 1) head-speaker, 2) tail-speaker, and 3) syllable-type, found significant main effects for head-speaker (f=13.12; p=0.0000) and syllable-type (f=6.10; p=0.0356). Significant interactions were also found for head-speaker by syllable-type, and head-speaker by tail-speaker. Results of an analysis of variance for the second group of subjects were the same as those for the first, with the exception that a significant main effect was found only for head-speaker (f=16.16; p=0.0000). Since ANOVA results were very similar for both subject groups, only results for the fully-crossed subject group will be presented in some portions of the discussion that follows.

Speaker-Related Effects

The head-speaker main effect resulted, in both subject groups, from significantly lower identification rates for speaker 1, a male. Speaker-related effects are also a component of both statistically significant interactions. The presence of speaker-related effects indicates that perceptual information for vowel identification is not fully invariant across speakers in the silent center stimuli. Speaker-related effects in the form of head-speaker by tail-speaker interactions are graphed in Figure 3-2. One might expect, here, that head by tail interactions occur because of differences in 'voice

Figure 3-2a: Head x Tail Effects for Group 1 (p<0.0003)

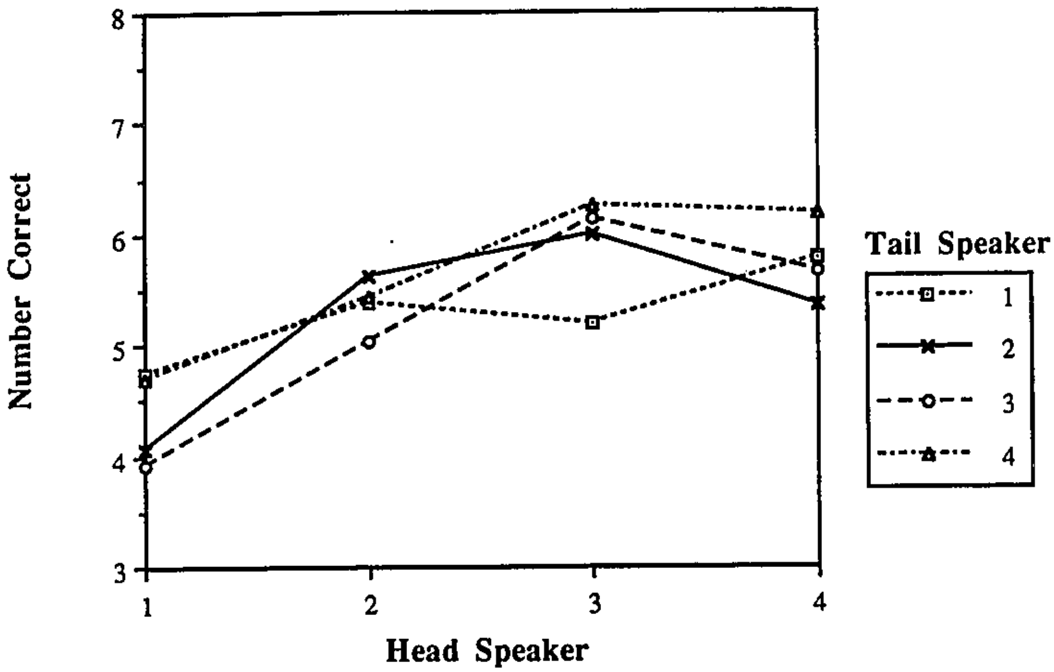
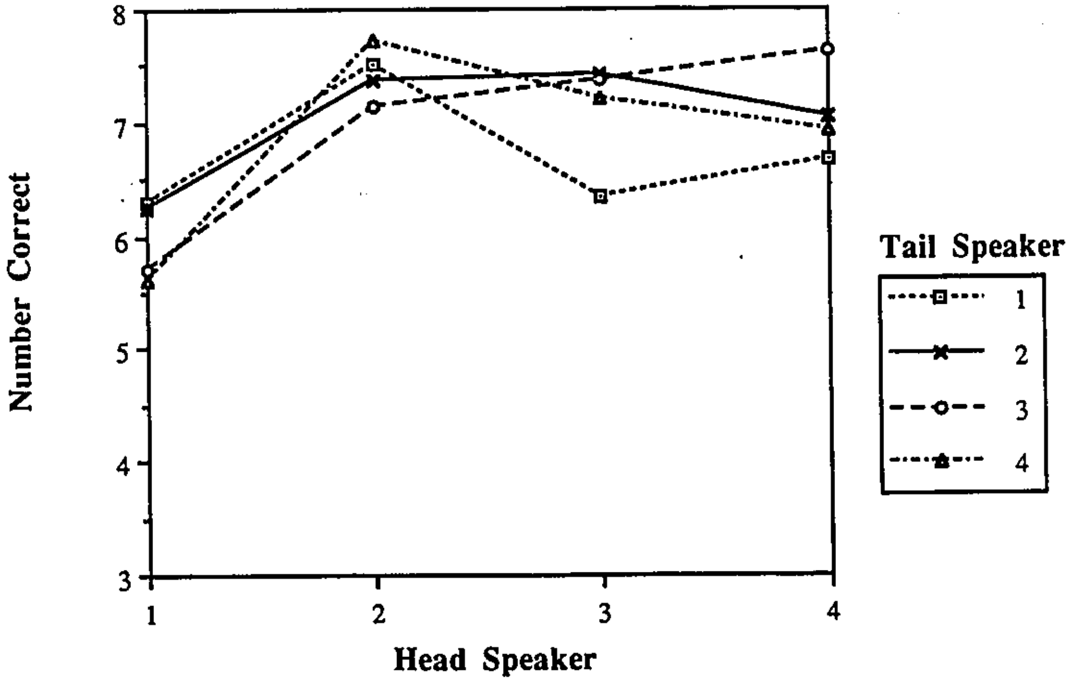


Figure 3-2b: Head x Tail Effects for Group 2 (p<0.0001)



compatibility'. It might be anticipated, for example, that tokens with the same head and tail speaker would be better-identified than tokens involving two different speakers, or that tokens involving speakers of the same sex would be better-identified than tokens which involve a male/female mix. However, there is no clear evidence for such a pattern. Average identification rates for each head and tail combination are provided in Tables 3-2 and 3-3 for the isolated vowel and [bVb] tokens respectively, for subject group 2.

Isolated Vowels		Tail Speaker			
Head Speaker		Male		Female	
		1	2	3	4
Male	1	72.5	71.7	64.2	62.5
	2	78.3	76.7	73.3	75.8
Female	3	67.5	80.8	76.7	74.2
	4	68.3	72.5	75.8	69.2

Table 3-2: Percent correct identification for each head and tail speaker combination for isolated vowels (subject group 2). 'Same speaker' cells are outlined along the diagonal.

[bVb]'s		Tail Speaker			
Head Speaker		Male		Female	
		1	2	3	4
Male	1	53.3	53.3	50.0	50.0
	2	71.7	70.8	69.2	78.3
Female	3	59.2	67.5	70.8	70.0
	4	65.0	68.3	76.7	69.2

Table 3-3: Percent correct identification for each head and tail speaker combination for [bVb] syllables (subject group 2). 'Same speaker' cells are outlined along the diagonal.

For the isolated vowels, the four 'same speaker' combinations (outlined along the diagonal) averaged 74% correct identification, only 1% better than the 73% average for mixed-speaker combinations. Results for the [bVb] tokens are comparable, with 'same speaker' combinations (at an average of 66%) again only 1% better identified than mixed-speaker combinations (at 65%). Indeed, the best speaker combinations for both stimulus types involve head and tail speakers of the opposite sex, namely speakers 3 and 2 for the isolated vowels, and speakers 2 and 4 for the [bVb]'s.

'Speaker compatibility', in the sense of simple voice similarity, does not, then, appear to explain the head by tail interactions. Strange (1989) also reports a speaker-related effect for silent-center stimuli, in which tokens produced by one speaker were less well identified than tokens for several other speakers. Strange concludes that this is "in part because of this speaker's slower articulatory rate during opening and closing gestures" (1989, p. 2148). A visual examination of formant tracks for the speakers involved in this experiment suggests that similar factors may contribute to speaker-related effects here. Figure 3-3 provides a comparison of the first 200 ms of formant tracks for three high F1 vowels ([ɛ,æ] and [ʌ]) in [bVb] context, for the two male speakers. Although fundamental frequencies are similar, and although F1 reaches a similar level in each vowel pair, speaker 1's F1 tends to move upward more slowly, and reach its maximum later than speaker 2's F1. Since only the first 40 ms of each syllable was used in the silent center 'heads', the slower articulatory rate of speaker 1 may be partly responsible for significantly lower identification rates when speaker 1 is head speaker. Differences in articulatory rates may also contribute to other speaker-related effects. However, since no apposite acoustical or statistical analyses have been conducted, this can be offered only tentatively as an explanation of the speaker-related effects in this experiment.

Strange attributes the lower identification rates for the "slow" speaker in her experiment to the incompleteness of coarticulatory information in the windowed heads

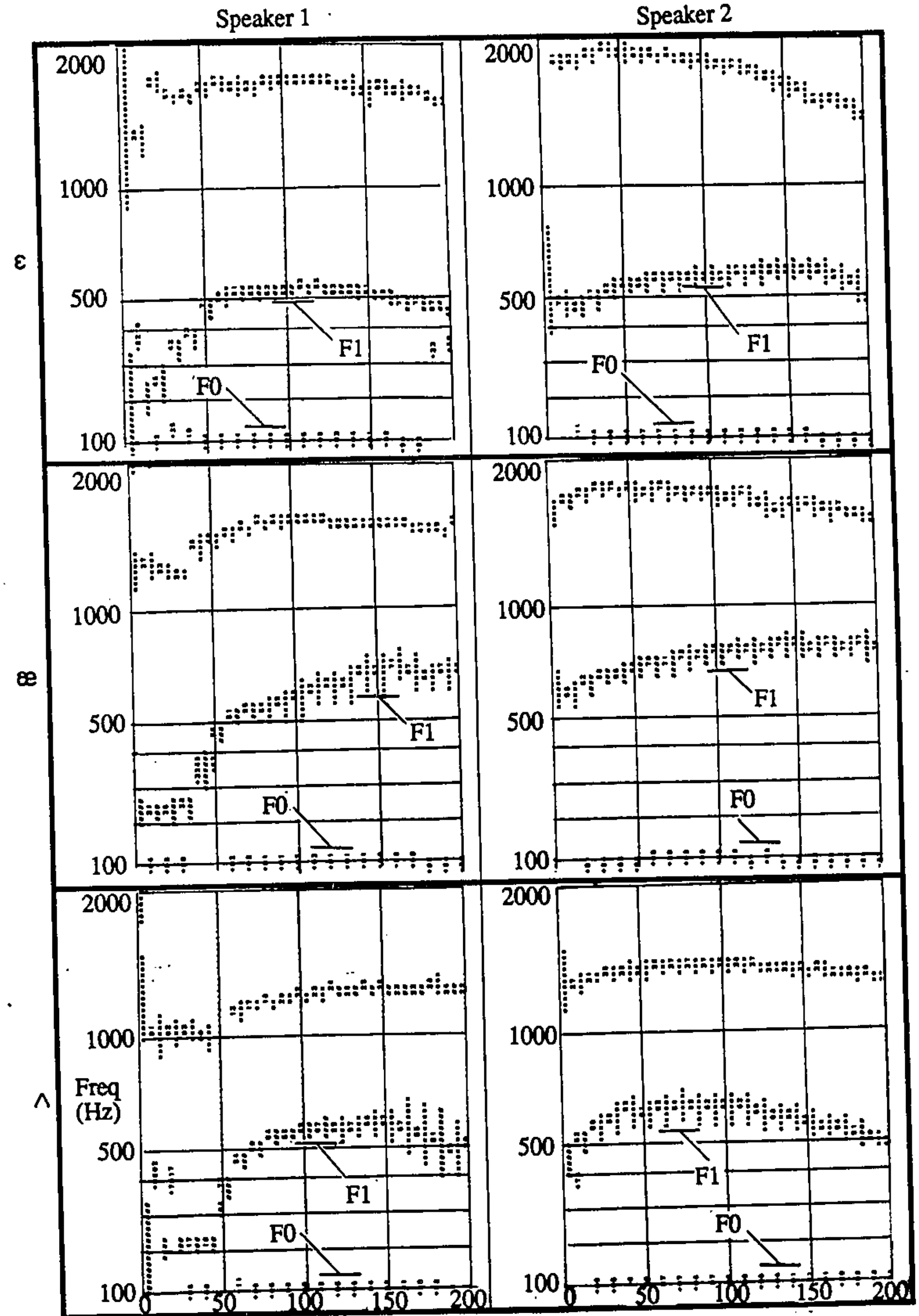


Figure 3-3: Formant tracks for [bɛb], [bæb] and [bʌb] for speakers 1 and 2. Each track begins at 0 ms after the initial burst release, and includes 200 ms of the vowel. The 'heads' of the silent center [bVb] stimuli included the first 40 ms of these syllables.

and tails. However, lower rates would also be expected if listeners rely primarily on vowel-inherent information. Since formants do not reach vowel-inherent levels until relatively late in the syllable, the brief windowed heads and tails may contain insufficient or misleading vowel-inherent cues for some speakers.

Syllable-Type Effects

Head-speaker by syllable-type interactions are graphed in Figure 3-4. Tukey tests of honestly significant differences indicate that, for both subject groups, vowels are significantly better identified when they occur in isolation than when they occur in [bVb] context, when either speaker 1 or speaker 3 is head speaker. The presence of an effect related to syllable-type indicates that, like information across speakers, the information available to listeners across syllable types is not fully invariant. Consequently, an extreme vowel-inherent view, which predicts that exactly the same cues to vowel identity are available in both cases, does not appear to be supported by this data. Similarly, since the syllable-type advantage goes to the isolated vowel stimuli, an extreme coarticulatory hypothesis is unsupported. The coarticulatory hypothesis predicts that listeners will identify silent center vowels better in [bVb] syllables than in isolation, since the endpoints of coarticulated vowels contain additional, more powerful cues to vowel identity.

Discussion

So far, then, neither extreme hypothesis appears to provide a satisfactory explanation of the results. An extreme coarticulationist view is unsupported, since the endpoints of isolated vowels in this experiment appear to provide as much, and sometimes more perceptual information than the endpoints of the [bVb] stimuli. An extreme vowel-inherent view is similarly unsupported, since the endpoints of the [bVb]

Figure 3-4a: Head x Syllable-Type Effects for Group 1 (p<0.0002)

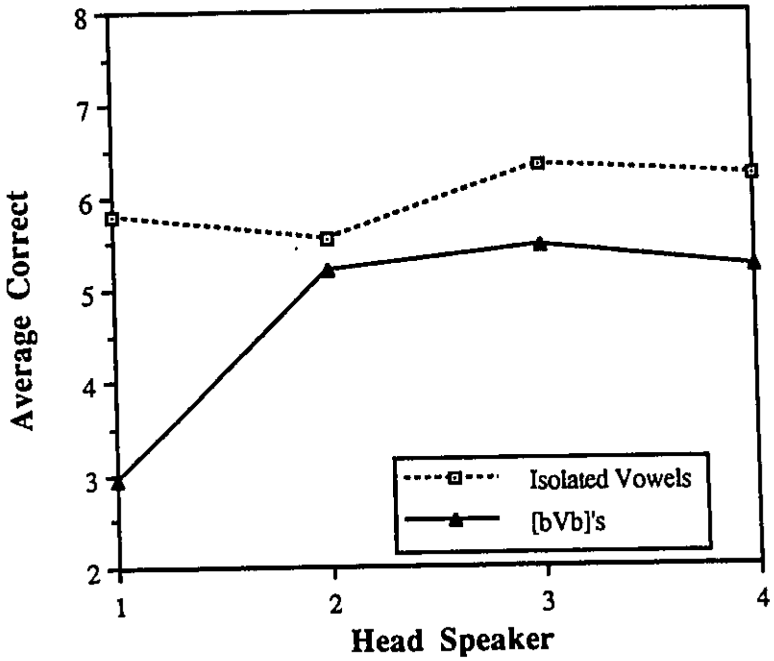
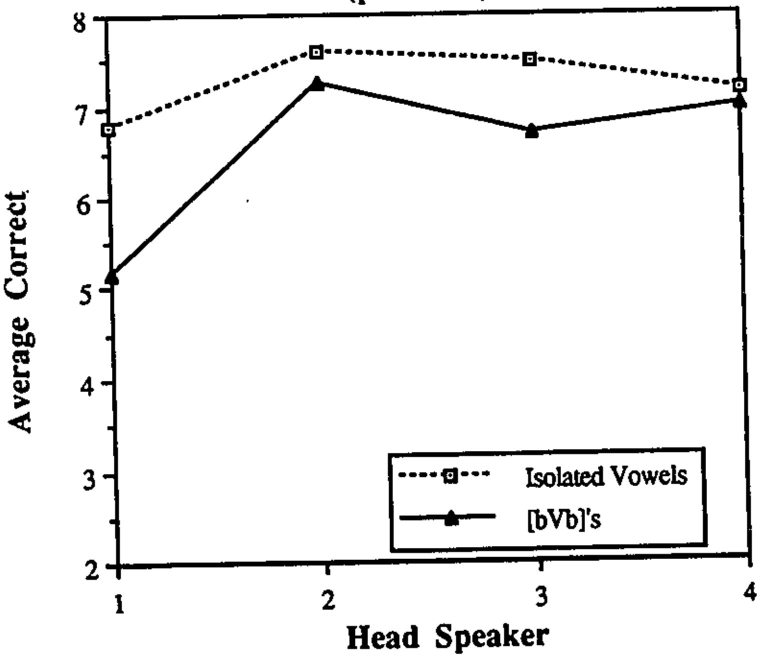


Figure 3-4b: Head x Syllable-Type Effects for Group 2 (p<0.0004)



stimuli do not provide listeners with precisely the same perceptual cues as the endpoints of the isolated vowels.

Before the extreme vowel-inherent view is abandoned, however, it is important to distinguish between cues that differ in terms of their exact values, and cues that differ in terms of how they are used. The conclusion that cues taken from the endpoints of vowels in [bVb] context are not the same as those taken from the endpoints of isolated vowels is neither new, nor very provocative, if the difference being discussed is strictly one of precise values: formant frequencies for a single vowel are known to vary with factors such as speaker identity and consonant context. In other words, the presence of differences in the actual values of perceptual cues need not, in itself, be an indication that listeners are extracting a different type of cue from [bVb] syllables, even if these differences result in some changes in perception. Vowel-inherent cues, including 'steady state targets', are also susceptible to changes in value, and these changes presumably have an effect on vowel identification. If, on the other hand, the cues provided by the isolated vowel and [bVb] stimuli differ in how they are *used* by listeners, then the vowel-inherent view must be abandoned.

It is not clear precisely how these two possibilities can be distinguished, based only on perceptual results. Some indication of the nature of differences in cues for these two stimulus types may, however, be found through a comparison of identification patterns. If cues from both sets of stimuli are used in the *same* manner by listeners, the disparities that arise due to differences in precise values should be relatively minor, and the overall patterns of results should show a strong resemblance. In particular, error patterns may help to distinguish whether the cues are used in different ways: since the same speakers produced both the isolated vowel and [bVb] tokens, characteristics of voice or pronunciation which lead to errors in one stimulus set should also do so in the other. However, if isolated vowel cues are used in a different way from cues for coarticulated vowels, the specific misidentifications which listeners make may also be expected to differ.

Comparison of Identification Patterns for [bVb]'s and Isolated Vowels

For their 1986 experiment, Nearey and Assmann developed a graphical means of comparing confusion matrices, thereby allowing complete response profiles, including both correct identification rates and error patterns, to be contrasted. They also used correlations to measure the "goodness of fit" between pairs of confusion matrices (1986, p. 1301-1302). These same methods will be used in this study to compare identification patterns for the [bVb] and isolated vowel stimuli.

A graphical comparison of confusion matrices for all tokens of the isolated vowel and [bVb] stimuli is provided in Figure 3-5. Listeners' identification patterns for the isolated vowels are represented by the white bars in the foreground, and patterns for the [bVb]'s are represented by the dark bars in the background. Although isolated vowels were significantly better identified than [bVb]'s for certain head-speakers, the overall identification patterns are remarkably similar. In particular, listeners' error patterns for the [bVb]'s correspond very well to error patterns for the isolated vowels. This suggests that perceptual cues are used in a similar manner in both cases.

A descriptive statistical measure of similarity for the two confusion matrices can be provided by a Pearson's product moment correlation. The *r* coefficient over the 100 cells of these two confusion matrices is 0.8878. In order to obtain an index against which this coefficient can be compared, subjects can be randomly assigned to two groups, and the confusion matrix for one half of the subjects can then be correlated

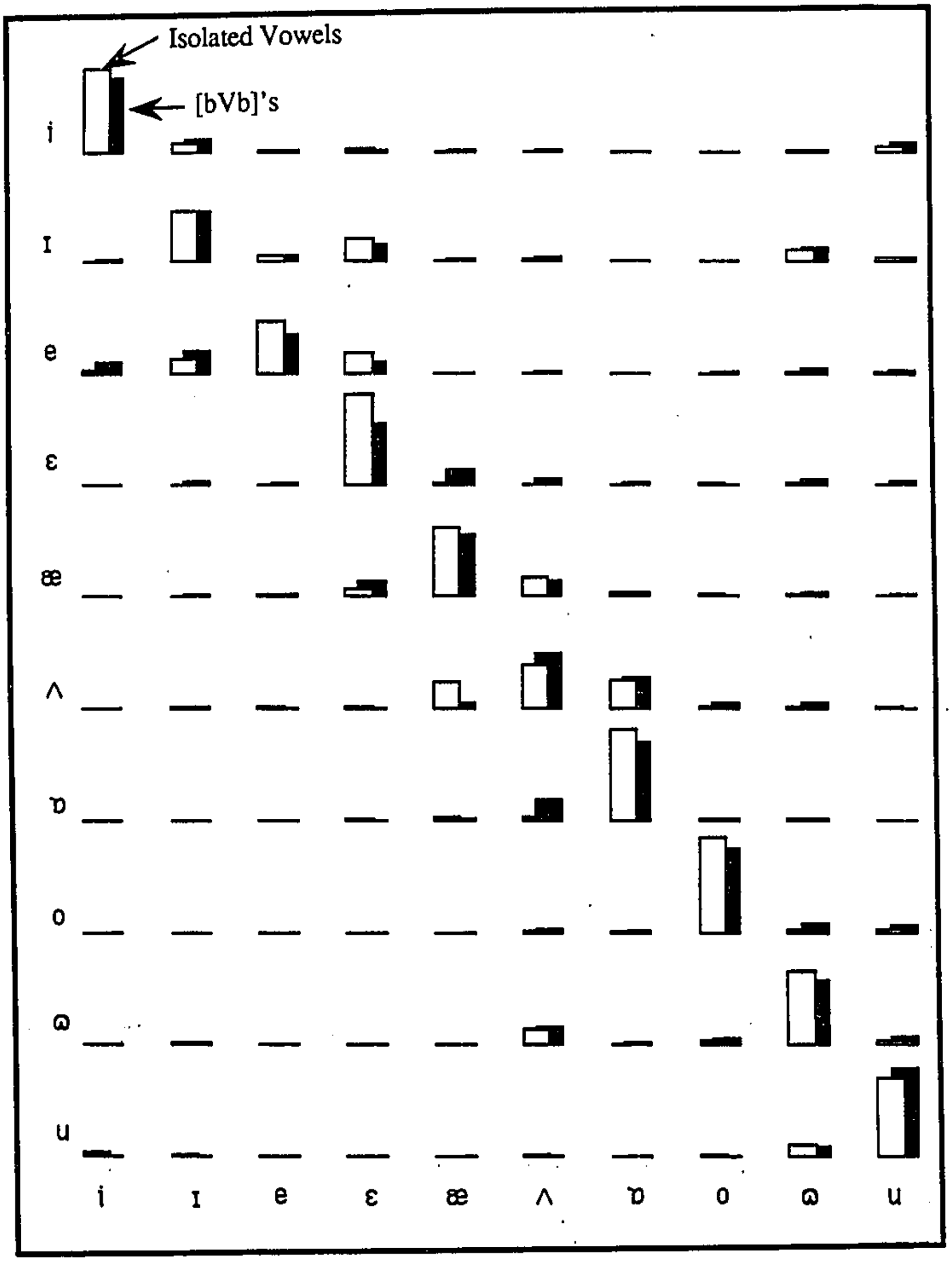


Figure 3-5: Comparison of listeners' identification patterns for the silent center isolated vowel and [bVb] stimuli. Rows represent intended vowel categories, and columns represent actual responses. Number of responses in any given category is indicated by the height of the bars. White bars in the foreground represent listeners' responses to the isolated vowel stimuli; background, black bars represent responses to the [bVb] stimuli.

with the matrix for the remaining subjects, for each stimulus type. Since different groups of subjects presumably use similar perceptual strategies to classify identical stimuli, the correlation between subject responses should approach its upper limit when responses for a single stimulus set are compared across subject groups. This comparison therefore provides a reasonable estimate of the ceiling which the correlation coefficient may reach when exactly the same perceptual cues are present, and subjects tend to use them in the same manner. When this procedure is used, the coefficients are 0.8723 for the isolated vowels, and 0.8190 for the [bVb]'s. Thus, at 0.8878, the coefficient for correlation between the isolated vowel and [bVb] confusion matrices is within the approximate expected range.

Since the majority of responses for the perceptual experiment are correct, and therefore fall along the diagonal, a high correlation is expected even if the two confusion matrices resemble each other only in the pattern of correct responses. A much more rigorous test of the similarity between identification patterns can be made by means of difference correlations. Difference correlations can be performed over *changes* in identification profiles, thus eliminating the inherent correlation between correct responses. A difference correlation analysis was performed over listeners' responses for the isolated vowels and [bVb]'s, by defining

$$X_{v,c,h,t} = \mathrm{ISO}\ (v,c,h_i,t_i) - \mathrm{ISO}\ (v,c,h_i,t_j) \qquad \text{and}$$

$$Y_{v,c,h,t} = \mathrm{BVB}\ (v,c,h_i,t_i) - \mathrm{BVB}\ (v,c,h_i,t_j),$$

where $X_{v,c,h,t}$ is the difference in listeners' classification of isolated vowel token v as a member of vowel category c, when speaker h_i is head speaker, and tail speaker identity changes from t_i to t_j. This difference is calculated for each of the three tail speakers who differ in identity from speaker h_i. $Y_{v,c,h,t}$ is similarly defined for the equivalent difference in identification of the [bVb] tokens. A small positive correlation was found for this comparison (*r*=0.1687). As above, a ceiling for the coefficient can be estimated by performing the same correlation over responses for one half of the subjects against responses for the other half, for each stimulus type. Using this procedure, a coefficient of 0.2249 is obtained for the isolated vowels, and 0.1258 for the [bVb]'s. Again, the coefficient for correlation *between* responses for the isolated vowel and [bVb] stimuli is within the resulting range.

The fact that correlations between identification patterns for isolated vowels and [bVb]'s are as good as those between responses for different groups of subjects who have access to exactly the same perceptual cues, strongly suggests that listeners do use the same kinds of cues to identify the [bVb] and isolated vowel stimuli in this experiment, and that they use them in the same way. Thus, contrary to the conclusions of Strange *et al.* (1983) and Verbrugge and Rakerd (1986), some proportion of the perceptual information which listeners extract from silent center [bVb] syllables, whether these are single-speaker or hybrid stimuli, appears to be inherent to the vowel.

Conclusion

To summarize, the results of this experiment argue against an extreme coarticulationist view, since there is no evidence that vowels in silent center [bVb] syllables are either significantly better-identified, or significantly differently-identified than silent center isolated vowels. If vowel-inherent cues are used in both cases, however, they are generally 'comparable', but not identical. In other words, although listeners may extract and use cues in the same way for both stimulus types, the perceptual results do differ when the actual values of the cues differ. This implies that, in certain cases, the silent center [bVb] stimuli contain somewhat 'degraded' versions of isolated vowel cues, since the [bVb] stimuli are less well identified than the isolated vowel stimuli for certain head speakers.

The results of this experiment do provide support for this version of the vowel-inherent hypothesis. However, to this point no evidence has been offered to show that the similarity in identification patterns results specifically from listeners' attention to vowel-inherent *spectral change.* Since such evidence would also provide more convincing support for the general 'vowel-inherent' view, the question of how well listeners' identification patterns can be predicted by a vowel-inherent spectral change hypothesis will be addressed in the second part of this experiment.

CHAPTER 4

Modelling Listeners' Perception of Silent-Center Isolated Vowels and [bVb] Syllables (Experiment 2, Part b)

Although identification rates and error patterns in Experiment 2a argue relatively convincingly against an extreme coarticulationist view of vowel perception, this does not necessarily imply support *for* the vowel-inherent spectral change hypothesis. The spectral change view can, however, be further appraised through the use of Nearey and Assmann's (1986) pattern-recognition model[7]. Using only measurements of vowel-inherent spectral change, Nearey and Assmann were relatively successful in predicting listeners' identification rates and error patterns for silent center isolated vowels. The analysis presented here represents a stringent cross-validation test of the Nearey-Assmann model: the pattern recognition model constructed by Nearey and Assmann in 1986 is used here, with no further 'tuning', to predict listeners' responses for two new sets of silent center stimuli. Since the model was 'trained' to identify vowels solely on the basis of vowel-inherent spectral change, it can be used to estimate the degree to which listeners rely on these cues, particularly to identify the silent center [bVb] stimuli.

Method

The Nearey-Assmann model is illustrated in Figure 4-1. In order to train the model for the 1986 experiment, five measurements were taken from ten tokens each of

[7] Broad and Clermont's (1987) perceptual model was also used to predict listeners' perception of the vowels tested in Experiment 2. However, a peculiarity of their solution causes the effects of [b] context to become exponentially more prominent as time passes, so that predicted formant values for long vowels (such as the 230 ms stimuli in Experiment 2) are far outside the normal range. Since the results were uninformative with respect to the nature of perceptual cues used for vowel identification, they will not be discussed here.

The Nearey-Assmann Model

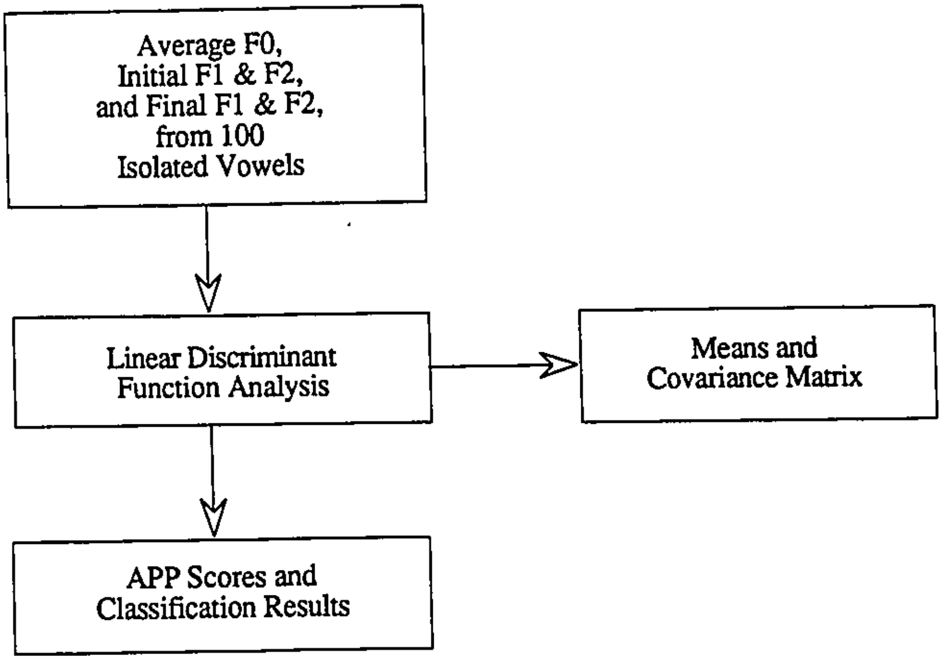


Figure 4-1a: Training on 1986 data

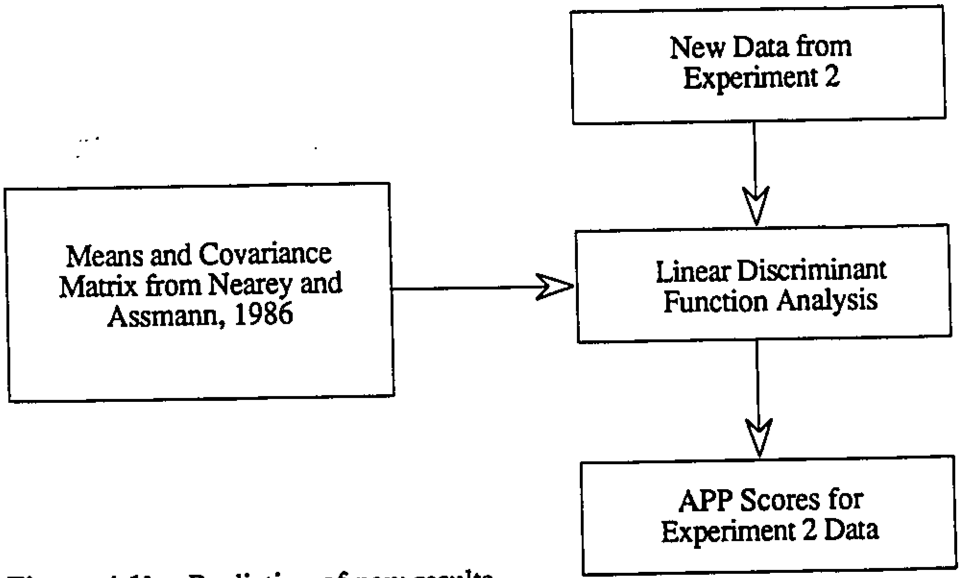


Figure 4-1b: Prediction of new results

the same isolated vowel categories as used in Experiment 2a, above. The five measurements used by the model are an average F0, and initial and final values for both F1 and F2. Log-transformed values of these five measures were subjected to linear discriminant function analysis, from which a-posteriori estimates of group membership were derived. As argued by Nearey and Assmann, the a-posteriori probability, or APP scores can be viewed as fuzzy classification scores, and can serve as predictors of listeners' responses (1986, p. 1301). Specifically, APP scores can be used to predict listeners' confusion matrices, including the proportion of correct responses for each vowel, and the number of incorrect responses assigned to each other vowel category.

Since the formant measurements which the model requires for vowel identification are the same as those used in Experiment 1, the F1 and F2 measurements from Experiment 1 were used for pattern recognition modelling in this experiment. In addition, F0 measurements for each head and tail section were made using the same procedure as described in Experiment 1 for formant measurements. An average F0 was then calculated for each token, and all values were log-transformed for use by the model. The procedure used for predicting listeners' responses is outlined in Figure 4-1b. The log means of the five variables (average F0, plus initial and final F1 and F2), and the pooled covariance matrix from the 1986 training data were used to predict listeners' responses for this experiment.[8]

[8] Several other variations of the model were tested in addition to the 'average F0' version. These included the following: 1) a version in which only an initial value was provided for F0, thus giving the head speaker priority; 2) a version in which formants were 'pre-normalized' by subtracting 30% of the average F0 from all formant values (this model is also discussed by Nearey and Assmann (1986, footnote 5, p. 1307); 3) a version in which formants were pre-normalized by subtracting 30% of the *initial* F0 from all formant values; and 4) a version in which formants were pre-normalized by subtracting 30% of the initial F0 from formants in the head of the syllable, and 30% of the final F0 from formants in the syllable tail. These variations were used in order to test several hypotheses of intrinsic speaker normalization (Nearey, 1989 p. 2090), and were felt to be justified by the presence of hybrid syllables in the perceptual test. Since none produced results that were better than those for the original ('average F0') model, they will not be discussed.

Results

For the isolated vowels, a comparison of the model's predicted correct identification rates and the observed correct identification rates is presented in Table 4-1. Although the model was not 'trained' to identify hybrid stimuli, which can exhibit considerable changes in formant values from head to tail, it nonetheless does extremely well on both the same-speaker stimuli and the hybrid isolated vowels, in fact performing better than the listeners in all cases.

Percent Correct Identification for Isolated Vowels		
Head-Tail Combination	**Observed**	**Predicted**
Same Speaker	73.8	81.4
Different Speakers: Male - Male	75.0	81.4
Male - Female	69.0	73.6
Female - Female	75.0	88.5
Female - Male	72.3	79.1

Table 4-1: Comparison of correct identification rates on isolated vowel stimuli for listeners ('observed' values) and the Nearey-Assmann model ('predicted' values) for different speaker combinations.

A comparison of the predicted and observed confusion matrices for the isolated vowels is provided in Figure 4-2. Listeners' identification patterns are represented by the white bars in the foreground, and the model's predictions are shown in the background, by the dark bars. Generally, the model's accuracy is comparable to that of the listeners. The model is somewhat better than listeners at identifying the vowels [i,ɪ,e] and [ʌ]. In addition, when errors are made, the model's predictions tend to follow the direction of listeners' errors relatively well. However, the proportion of errors assigned to any given vowel category tends to be predicted less successfully.

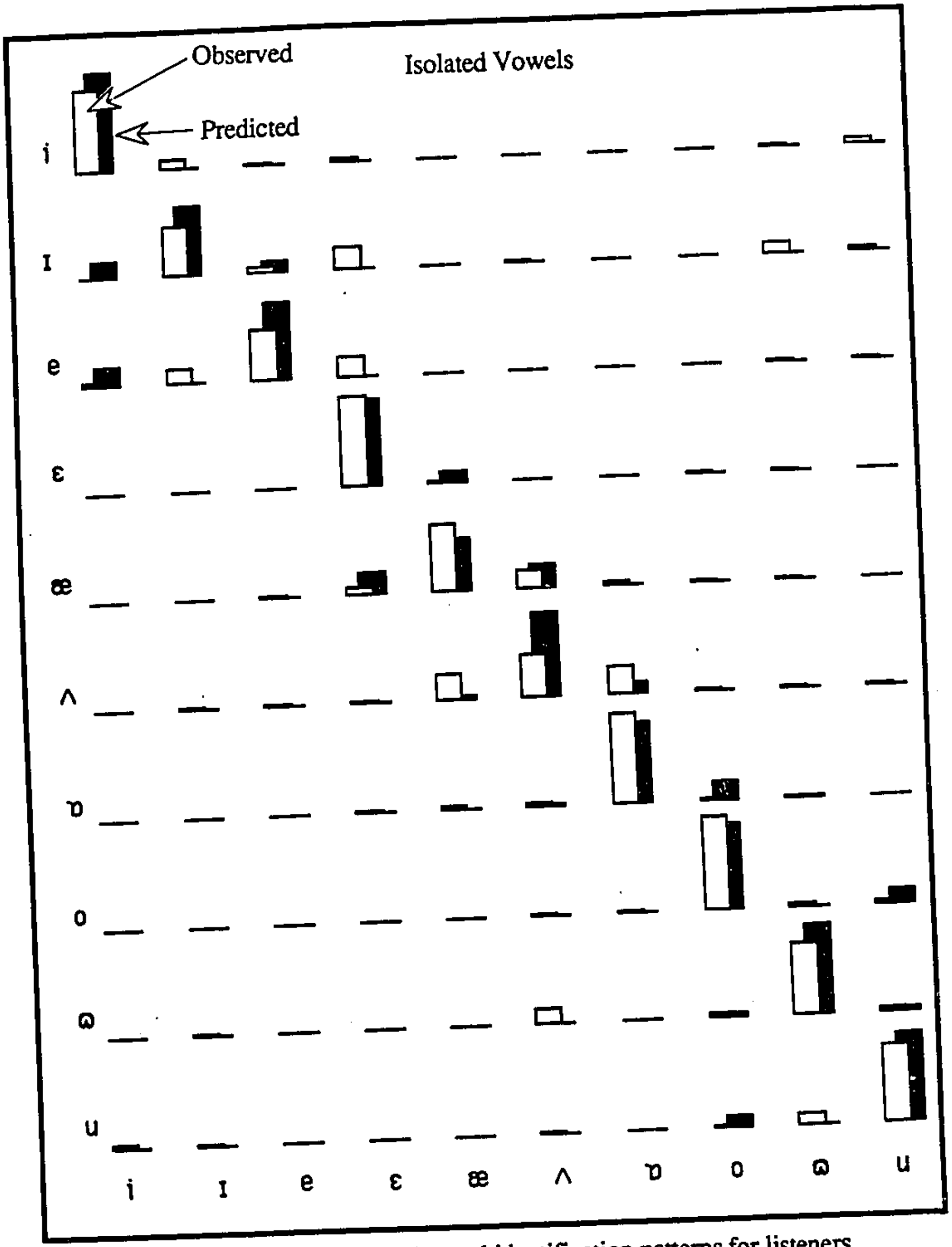


Figure 4-2: Comparison of isolated vowel identification patterns for listeners ("Observed") and the Nearey-Assmann model ("Predicted"). Rows represent intended vowel categories, and columns represent actual responses. Number of responses in any given category is indicated by the height of the bars. White bars in the foreground represent listeners' responses to the isolated vowel stimuli; background, dark bars represent the model's predictions for those stimuli.

Observed and predicted correct identification rates for the [bVb] stimuli are summarized in Table 4-2. Both listeners and the model have more difficulty identifying the [bVb]'s, and in this case the listeners perform better than the model on some speaker combinations. The drop in the model's ability to predict listeners' responses suggests that the [bVb] stimuli may contain some additional information which listeners exploit, but the model does not. Such additional information could be coarticulatory in nature. However, it is also possible that listeners are simply more familiar than the model with the range of variation that is permissible in formant frequencies, especially when vowels occur in consonant context. Since the model is trained on isolated vowels produced by a group of just ten speakers (none of whom were included in the test group for this experiment) a substantial increase in the size of the training group might improve the model's performance on [bVb] context vowels, even without the specific inclusion of training on vowels from consonant context.

Percent Correct Identification for [bVb]'s		
Head-Tail Combination	**Observed**	**Predicted**
Same Speaker	66.0	58.7
Different Speakers: Male - Male	62.6	66.7
Male - Female	61.9	49.4
Female - Female	73.4	64.1
Female - Male	65.0	74.6

Table 4-2: Comparison of correct identification rates on [bVb] stimuli for listeners ('observed' values) and the Nearey-Assmann model ('predicted' values) for different speaker combinations.

Predicted and observed confusion matrices for the [bVb]'s are compared in Figure 4-3. The model is again somewhat better than listeners at identifying the high front vowels, but in this case it is much less successful than listeners on the vowel [ʌ]. To a large degree, the difficulty the model has with the vowel [ʌ] may be due to its

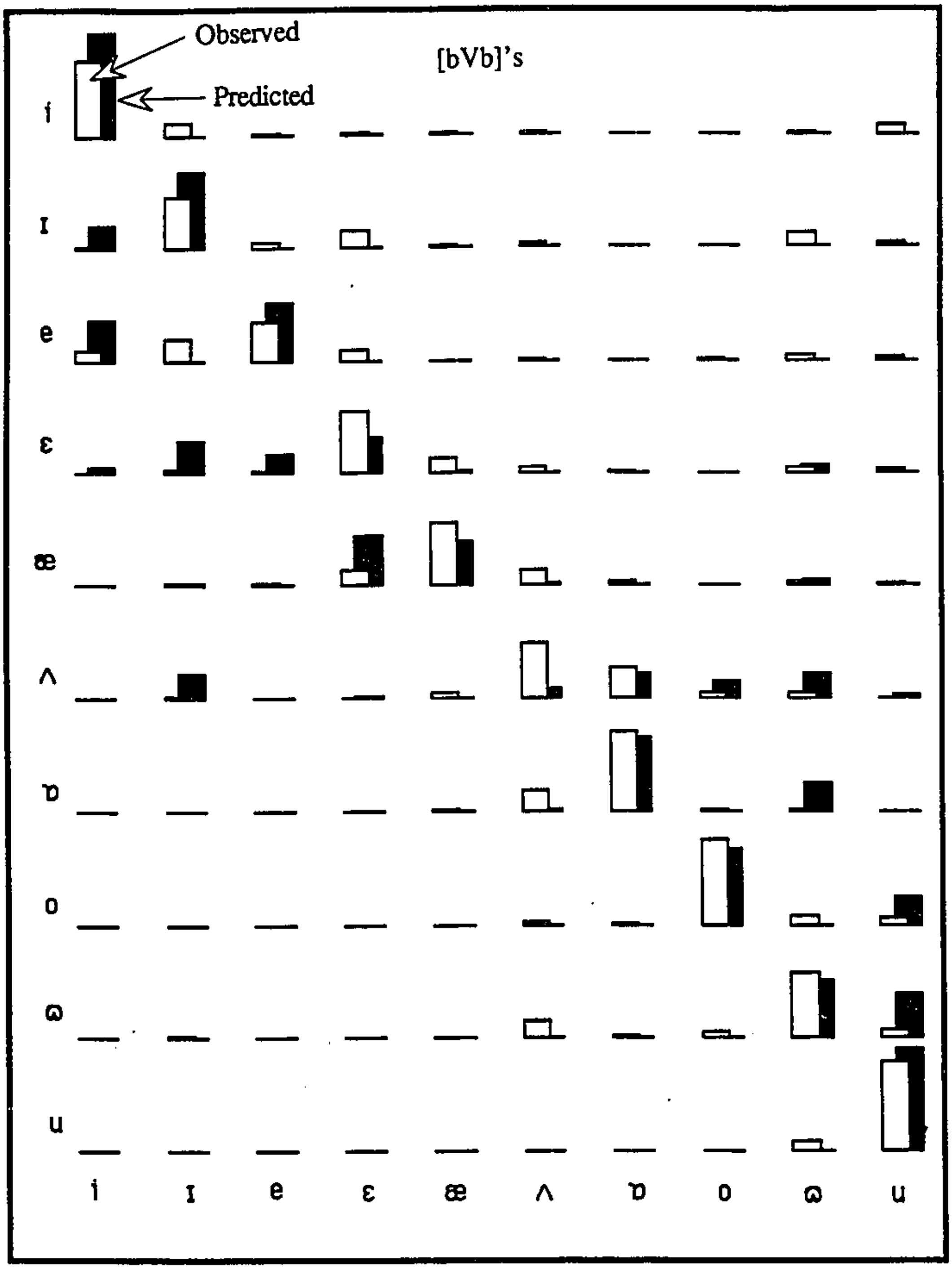


Figure 4-3: Comparison of [bVb] identification patterns for listeners ("Observed") and the Nearey-Assmann model ("Predicted"). Rows represent intended vowel categories, and columns represent actual responses. Number of responses in any given category is indicated by the height of the bars. White bars in the foreground represent listeners' responses to the [bVb] stimuli; background, dark bars represent the model's predictions for those stimuli.

position in the F1-F2 space. As can be seen in Figure 2-1 (p. 35), [ʌ] is 'surrounded' in comparison with other vowels. Thus, a relatively small error on [ʌ] will be relatively likely to result in its misidentification as some other vowel. It would appear, from other experiments on vowel identification, that human listeners also encounter this difficulty. Error rates for [ʌ] are often higher than for other vowels in a given experiment (e.g. Strange *et al.*, 1976; Strange and Gottfried, 1980; Diehl *et al.*, 1981)As with predictions for the isolated vowels, the model's errors tend to follow the direction of listeners' errors, but the proportion of misidentifications assigned to any particular vowel category tends to differ.

An estimate of the similarity between the model's predictions and the observed identification patterns can again be made by means of a difference correlation. In this case, the variables are defined as

$$Y_{v,c} = LID_{iso}\,(v,c) - LID_{bVb}\,(v,c) \qquad \text{and}$$

$$X_{v,c} = APP_{iso}\,(v,c) - APP_{bVb}\,(v,c)$$

where $Y_{v,c}$ is the difference in listeners' identification for the isolated and [bVb] versions of vowel token *v*, as a member of vowel category *c*; and $X_{v,c}$ is the difference in the model's a-posteriori predictions for those tokens.

The coefficient for this correlation is *r*=0.1301. As in Experiment 2a, a ceiling for the correlation coefficient can be estimated by performing the same correlation over responses for one half of the subjects against responses for the other half. In this case, the correlation between groups of listeners is slightly better than the correlation between listeners' responses and the model's predictions, at *r*=0.1484. Although the coefficient is lower when listeners are compared to the model than when one group of listeners is compared to another, randomization tests (Nearey and Assmann, 1986, p. 1306) indicate that the correlation between listeners' responses and the models' predictions is significant at the p=0.0001 level. Once again, this suggests that listeners and the model are using the same type of cues, and using them in the same general manner. Since the model receives only information on vowel-inherent spectral change, listeners would also appear to be using vowel-inherent spectral change as a cue to vowel identity, both for the isolated vowel stimuli and for the vowels in [bVb] context.

Conclusion

The results of this modelling experiment show that listeners' identification of both the silent center isolated vowel and the silent center [bVb] stimuli can be predicted relatively well from initial and final values of F1 and F2, encompassing spectral change across the vowel. This is true even though the stimuli include hybrid syllables, on which the model has never been trained. The model, like listeners, performs less well on the [bVb]'s than on the isolated vowel stimuli. However, whereas the model is consistently more accurate than listeners on the isolated vowels, it is less accurate than listeners in about half the cases on the [bVb] stimuli. This may be because listeners extract additional information from the silent center [bVb]'s, to which the model does not have access. Such additional information could be coarticulatory in nature. Conversely, listeners may identify some vowels in [bVb] context better than the model simply because they have had far more exposure to the range of permissible variation in vowel-inherent cues. A replication of this experiment using a more extensive training set for the model (i.e. a training set which includes isolated vowels produced by a much larger group of speakers) would be of value in distinguishing between these two possibilities.

Although the model is somewhat less successful than listeners in identifying certain [bVb] stimuli, the fact that it is generally successful in predicting listeners' behaviour for the [bVb]'s (as demonstrated by the presence of a significant correlation between the model's predictions and listeners' identification patterns) strongly suggests that listeners and the model are, to an important degree, using the same types of cues to identify the [bVb]'s. Since the model has access only to information on vowel-inherent spectral change, listeners presumably are also using information on vowel-inherent spectral change to identify both the silent center [bVb] and the silent center isolated vowel stimuli.

Chapter 5

General Discussion and Conclusions

It has been proposed that consonant context provides listeners with supplementary cues to vowel identity, and that these coarticulatory cues are perceptually superior to the cues which vowels themselves provide regarding their identity. However, the results of the experiments discussed here suggest that, to an important degree, listeners use the same cues to identify isolated vowels and vowels in [bVb] context, at least when such stimuli are presented in silent center form.

Summary of Earlier Evidence Against a Coarticulatory Hypothesis

Much of the evidence presented to date in support of a coarticulatory hypothesis of vowel perception has been questioned. To recapitulate briefly, the evidence in favour of a coarticulatory hypothesis tends to fall into three categories: 1) identification rates which are better for natural vowels in consonant context than for natural isolated vowels; 2) acoustic analyses which show that, in terms of steady state formant frequencies, isolated vowels are somewhat better-distinguished than vowels in CVC context, even though the CVC context vowels are better identified; and 3) high identification rates for vowels in CVC syllables when vowel-inherent perceptual cues are 'removed'.

As discussed in Chapter 1, it has been shown that identification rates are sometimes higher for vowels in CVC context than for isolated vowels because of dialectal differences between speakers and listeners, and because certain commonly used response tasks (especially those which employ spelled versions of the test syllables) are more difficult for listeners to perform for isolated vowels than for vowels in CVC context (Assmann, 1979; Macchi, 1980; Diehl *et al.*, 1981). Thus, better identification rates for context vowels are not necessarily an indication that such vowels are perceptually better-defined by the presence of coarticulatory information. Instead, such results may indicate that experimental designs require modification.

The value of the second type of evidence, the degree of separation between steady-state formant values, has also been questioned. The smaller degree of separation between steady state targets for CVC context vowels, in combination with higher identification rates for these vowels, brought some researchers to conclude that listeners must use coarticulatory information to identify CVC context vowels, since they are able to distinguish between context vowels more effectively than between isolated vowels. As mentioned, speaker and task factors contribute to higher CVC identification rates in some experiments. Additionally, the results of other experiments suggest that listeners neither require, nor prefer steady state cues for isolated vowel identification (Gay, 1970; Hyde, 1971; Bond, 1982; Nearey and Assmann, 1986). If listeners do not rely on steady state cues to identify isolated vowels, then the degree of acoustic separation between steady state targets does not necessarily indicate the degree of perceptual separation which listeners can achieve between vowels, using exclusively vowel-inherent cues. Thus, a lack of separation between steady state targets in CVC context vowels does not necessarily indicate that listeners must resort to coarticulatory cues for accurate perception.

The third type of evidence has also been shown not to exclusively support a coarticulatory hypothesis. Experiments in which the central portions of vowels are removed have clearly demonstrated that listeners can identify CVC context vowels from dynamic information contained in their endpoints (Strange *et al.*, 1983;

Verbrugge and Rakerd, 1986; Rakerd and Verbrugge, 1987; Jenkins and Strange, 1987; Murphy *et al.*, 1989). However, listeners can also identify isolated vowels from such information (Nearey and Assmann, 1986). This suggests that the perceptual information provided by the endpoints of CVC context vowels need not be coarticulatory in nature. Instead, the endpoints of coarticulated vowels may retain vowel-inherent information, and listeners may rely on the vowel-inherent cues, rather than coarticulatory cues, to identify silent center vowels. This possibility was investigated in the experiments discussed here, and the results support this conclusion.

Evidence Presented Against a Coarticulatory Hypothesis in this Thesis

As shown in Experiment 1, vowel-inherent spectral change comparable to that found in isolated vowels by Nearey and Assmann (1986) persists in vowels in [bVb] context. Hence, the same information is available, and may be used by listeners to identify both silent center isolated vowels and silent center [bVb] vowels. In Experiment 2a it was shown that listeners' identification patterns for isolated and [bVb] context vowels are remarkably similar. This was true even though listeners heard only the endpoints of the vowels, without durational or steady-state cues to vowel identity. For certain head speakers, the isolated vowel stimuli were significantly better identified than the [bVb] stimuli. Thus, contrary to Strange *et al.*'s early (1976) conclusion that isolated vowels are underspecified in comparison with vowels from consonant context, these results support the conclusion that isolated vowel stimuli are equally as well specified as vowels from [bVb] context, and occasionally better specified, even when only the dynamic information in the endpoints of the vowels is considered. With regard to Verbrugge and Rakerd's (1986) conclusion that listeners must attend to information spread throughout the syllable in order to identify hybrid silent center

[bVb] stimuli, the results presented in this thesis again suggest that sufficient information for accurate identification is available in the vowel alone, even when the central 'target' is removed, durational information is neutralized, and speaker identity changes from the beginning to the end of the syllable. The results of Experiment 2a do not establish that listeners use the same cues to identify the vowels in both types of stimuli, but the resemblance between identification patterns suggests that this may be the case.

Evidence that listeners do use the same cues, to a large extent, to identify [bVb] context and isolated vowels is provided by Experiment 2, Part b. Listener identification patterns are predicted relatively well, both for isolated vowels and vowels from [bVb] context, using strictly vowel-inherent spectral change cues. The fact that the Nearey-Assmann model successfully identifies vowels from [bVb] context, indicates that the same type of spectral change which is found in isolated vowels also provides a cue to the identity of coarticulated vowels. This is clearly the case, since the model identifies the [bVb] stimuli at much greater than chance levels, but relies solely on vowel-inherent spectral change cues from isolated vowels, to do so. The fact that the model's identification rates are approximately as good as listeners', and that the model's predicted identification patterns show a significant resemblance to listeners' actual identification patterns, also suggests very strongly that listeners, like the model, are using vowel-inherent spectral change cues to identify both the isolated vowel and [bVb] stimuli.

Results Which May Favour a Coarticulatory View

Although much of the evidence in favour of a coarticulatory hypothesis has been challenged, several facts remain which are not, at least at the present time, satisfactorily explained by a purely vowel-inherent hypothesis of vowel identification. These include the consistent finding that natural vowels are slightly better identified in consonant context than in isolation, even when dialectal and response-task factors are controlled (e.g. Strange and Gottfried, 1980). The failure of the Nearey-Assmann model (in Experiment 2b, above) to perform as well on the [bVb] stimuli, relative to listeners' performance, as on the isolated vowel stimuli also suggests that listeners may use some additional information, to which the model does not have access, to more accurately identify the [bVb] stimuli. These findings (and possibly others) suggest that listeners may rely on coarticulatory information to some degree for accurate perception of vowels in CVC context.

Other explanations are, however, also possible. Vowels produced in consonant context may exhibit more regular differences in duration than isolated vowels, in which case durational cues, rather than coarticulatory cues, could account for the slight difference in identification rates. It should be noted, though, that Strange *et al.* (1976) did examine differences in duration between isolated vowels and coarticulated vowels, and concluded that such differences were insufficient to explain the confusion patterns and differences in error rates that were present in the data.

The disparity between identification rates for the model versus listeners, on the [bVb]'s as compared to the isolated vowels, may also be explained without reference to coarticulatory information. As discussed in Chapter 4, it is reasonable to assume that listeners are more familiar than the model with the range of variation that is allowed in vowel-inherent cues within any given vowel category, and that they are therefore able to identify the [bVb] stimuli more accurately, without refering to coarticulatory information. Without further evidence to this effect it is, of course, equally reasonable to assume that listeners do use some amount of coarticulatory information to identify the [bVb] stimuli.

Degree of Listeners' Reliance on Coarticulatory Information

The experimental evidence discussed above suggests that coarticulatory information is not the primary source of information for vowel identification, even for vowels in consonant context. Two additional experiments can be mentioned which further suggest that listeners' reliance on coarticulatory information for vowel perception, if it exists at all, is relatively minimal. Using highly simplified synthetic versions of [bVb] stimuli, Andruski and Nearey (in preparation) showed that listeners' identification patterns for [bVb] context vowels which contain only vowel inherent spectral change cues closely resemble identification patterns for "full" [bVb] syllables (i.e. syllables which contain both vowel-inherent and coarticulatory cues). Identification patterns for the synthetic 'vowel-inherent' stimuli also closely resembled those for synthetic silent center [bVb] syllables. This again suggests that listeners rely on essentially the same information in all three cases. Since only vowel-inherent cues were common to all three sets of synthetic stimuli, the findings imply that listeners tend to rely largely on vowel-inherent cues. Nearey (in preparation) also provides evidence which suggests that coarticulatory information contributes relatively little to listeners' perception of vowels in [bVb] context. Through the use of several optimized logistic models of vowel perception, Nearey showed that listeners' identification patterns for silent center [bVb] stimuli can be largely accounted for by refering only to vowel-inherent spectral change cues. The addition of coarticulatory cues, in the form of information about the extent and direction of initial and final transitions, resulted in only a very small improvement in the model's ability to predict listeners' identification patterns.

Perception of Vowels in Natural Speech

The question remains whether the cues which listeners use to identify silent center vowels under experimental conditions are the same, or even similar, to those which they use to identify unmodified vowels in natural speech. Unmodified control syllables are consistently slightly better, if not always significantly better identified than silent center stimuli (Strange *et al.*, 1983; Nearey and Assmann, 1986; Verbrugge and Rakerd, 1987; Strange, 1989). This does not, in itself, imply that listeners use different cues or strategies to identify vowels in unmodified speech. However, listeners may alter their perceptual strategies in reponse to unusual stimuli, so that the strategies used to identify silent center vowels do not reflect strategies which are generally used in speech perception. Hence, there is something of a Catch-22 for this type of research: it is not possible to distinguish between the opposing explanations considered here, purely on the basis of *correct* identification rates. However, lowering the identification rates often requires the introduction of some type of artificiality in the stimuli, or in the experimental conditions (for example, high levels of noise), and this may, inevitably, be reflected by some degree of artificiality in listeners' response strategies.

Summary and Directions for Further Research

In large measure, speakers of Western Canadian English use the same, vowel-inherent cues, for perception of isolated vowels and vowels in [bVb] context. This conclusion can be extended only tentatively, since it is based on comparisons of isolated vowels with vowels in just one consonant context, for one dialect of English, and since a relatively small number of speakers and hearers were involved in the experiments. A much larger experiment, incorporating a variety of consonant contexts, and speakers and hearers of other English dialects, would allow more general conclusions to be drawn regarding perception of English vowels in consonant context.

Although the above conclusion cannot be extended without risk, it can be made with some confidence within the given bounds. The use of the Nearey-Assmann model allows firm predictions of perceptual results to be made under a purely vowel-inherent hypothesis of perception, for the specific stimuli used in this experiment. Comparisons of these predictions with actual listener performance show that the two are significantly correlated. The conclusion that listeners rely primarily on vowel-inherent cues could be seriously challenged if evidence were presented showing that equally good predictions can be made on the basis of purely coarticulatory cues. However, since the nature of coarticulatory information has not been precisely defined, and since current discussions often assume that coarticulatory information is what remains once durational and steady state cues to vowel identity are removed, no attempt has been made here to conduct this test. If, in the future, a generally-accepted acoustic definition of coarticulatory cues can be provided by supporters of the coarticulatory hypothesis, perceptual modeling will provide an excellent means of testing both the adequacy of that definition, and listeners' reliance on the cues incorporated in the definition, in comparison with their reliance on vowel-inherent cues.

Since the cues proposed by both of these hypotheses of vowel perception are relatively complex, and since the perceptual results are difficult to interpret without the assistance of well-defined models, it will likely be difficult to resolve the opposing points of view without the aid of perceptual models. Further development and use of models to provide precise, objective predictions therefore seems likely to provide the most promising avenue for comparing the merits of competing perceptual hypotheses.

Bibliography

Ainsworth, W. A. (1972). Duration as a cue in the recognition of synthetic vowels. Journal of the Acoustical Society of America, 51, 648-651.

Ainsworth, W. A. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant, & M. Tatham (Ed.), Auditory analysis and perception of speech (pp. 103-113). London: Academic Press.

Andruski, J.E., & Nearey, T.M. (in preparation). Modelling listeners' perception of vowels in synthetic [bVb] syllables.

Assmann, P. (1979). The role of context in vowel perception. Unpublished MSc thesis, University of Alberta, Edmonton, Alberta.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: orthographic, perceptual, and acoustic aspects. Journal of the Acoustical Society of America, 71(4), 975-989.

Benedetto, M. G. D. (1989a). Vowel representation: some observations on temporal and spectral properties of the first formant frequency. Journal of the Acoustical Society of America, 86(1), 55-66.

Benedetto, M. G. D. (1989b). Frequency and time variations of the first formant: properties relevant to the perception of vowel height. Journal of the Acoustical Society of America, 86(1), 67-77.

Bladon, A. (1985). Diphthongs: a case study of dynamic auditory processing. Speech Communication, 4, 145-154.

Bond, Z. S. (1982). Experiments with synthetic diphthongs. Journal of Phonetics, 10, 259-264.

Borden, G. J., & Harris, K. S. (1984). Speech science primer (2nd ed.). Baltimore: Williams & Wilkins.

Broad, D. J., & Clermont, F. (1987). A methodology for modeling vowel formant contours in CVC context. Journal of the Acoustical Society of America, 81, 155-165.

Broad, D. J., & Fertig, R. H. (1970). Formant-frequency trajectories in selected CVC-syllable nuclei. Journal of the Acoustical Society of America, 47(6), 1572-1582.

Dechovitz, D. (1977). Information conveyed by vowels: a confirmation. Haskins Laboratories: Status Report on Speech Research, SR-51/52, 213-219.

Diehl, R. L., McCusker, S. B., & Chapman, L. A. (1981). On the identifiability of synthesized steady-state isolated vowels in isolation and in consonantal context. Journal of the Acoustical Society of America, 68, 1626-1635.

Disner, S. F. (1980). Evaluation of vowel normalization procedures. Journal of the Acoustical Society of America, 67(1), 253-261.

Fairbanks, G., & Grubb, P. (1961). A psychophysical investigation of vowel formants. Journal of Speech and Hearing Research, 4, 203-219.

Flanagan, J. (1955). Difference limen for vowel formant frequency. Journal of the Acoustical Society of America, 27, 613-617.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. Language and Speech, 5, 171.

Fujimura, O., & Ochiai, K. (1963). Vowel identification and phonetic contexts. Journal of the Acoustical Society of America, 35, 1889(A).

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. IEEE Transactions on Audio and Electroacoustics, AU-16(1), 73-77.

Gay, T. (1970). A perceptual study of American English diphthongs. Language and Speech, 13, 65-88.

Gerstman, L. J. (1968). Classification of self-normalized vowels. IEEE Transactions on audio and electroacoustics, AU-16(1), 78-80.

House, A. S. (1961). On vowel duration in English. Journal of the Acoustical Society of America, 33, 1174-1178.

Hyde, S. R. (1971). Perception of very brief sounds. Second International Congress on Applied Linguistics London: Cambridge University Press.

Jamieson, D. G., Nearey, T. M., & Ramji, K. (1989). CSRE: a speech research environment. Canadian Acoustics / Acoustique canadienne, 17(4), 23-25.

Jenkins, J. J., & Strange, W. (1987). Identification of 'hybrid' vowels in sentence context. Journal of the Acoustical Society of America, 82(S1), S82.

Johnson, K. (1988). F0 normalization and adjusting to talker. Research on Speech Perception Progress Report, 14, 237-258.

Kent, R. D., & Forner, L. L. (1979). Developmental study of vowel formant frequencies in an imitation task. Journal of the Acoustical Society of America, 65(1), 208-217.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. Journal of the Acoustical Society of America, 67, 971-995.

Klatt, D. (1987).Review of text-to-speech conversion for English. Journal of the Acoustical Society of America, 82(3), 737-792.

Ladefoged, P., & Broadbent, D. E. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29(1), 94-104.

Lehiste, I., & Peterson, G. E. (1961). Transitions, glides, and diphthongs. Journal of the Acoustical Society of America, 33(3), 268-277.

Lindblom, B. (1963). Spectrographic study of vowel reduction. Journal of the Acoustical Society of America, 35, 1773-1781.

Lindblom, B., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. Journal of the Acoustical Society of America, 42, 830-843.

Macchi, M. J. (1980). Identification of vowels spoken in isolation versus vowels spoken in consonantal context. Journal of the Acoustical Society of America, 68, 1636-1642.

Markel, J., & Gray, A. (1976). Linear prediction of speech. Berlin: Springer Verlag.

Mermelstein, P. (1977). On the relationship between vowel and consonant identification when cued by the same acoustic information. Haskins Laboratories: Status Report on Speech Research, SR-51/52, 201-212.

Miller, R. L. (1953). Auditory tests with synthetic vowels. Journal of the Acoustical Society of America, 25(1), 114-121.

Mullennix, J. W., & Pisoni, D. B. (1988). Detailing the nature of talker normalization in speech perception. Research on Speech Perception Progress Report, 14, 289-305.

Murphy, W. D., Shea, S. L., & Aslin, R. N. (1989). Identification of vowels in 'vowelless' syllables by 3 year olds. Perception and Psychophysics, 46(4), 375-383.

Nearey, T.M. (in preparation). A case study in linear logistic analysis of speech perception data: Assessing the role of fundamental frequency and formant transition information in hybrid syllables.

Nearey, T. M. (1989). Static, dynamic and relational properties in vowel perception. Journal of the Acoustical Society of America, 85(5), 2088-2113.

Nearey, T., & Assmann, P. (1986). Modeling the role of inherent spectral change in vowel identification. Journal of the Acoustical Society of America, 80, 1297-1308.

Nearey, T. M., & Hogan, J. T. (1986). Phonological contrast in experimental phonetics: relating distributions of production data to perceptual categorization curves. In J. Ohala, & J. Jaeger (Ed.), Experimental phonology Orlando, Fl.: Academic Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. Journal of the Acoustical Society of America, 24(2), 175-184.

Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. Journal of the Acoustical Society of America, 32, 693-703.

Pols, L. C. W., van der Kamp, L. J. T., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46(2,2), 458-467.

Potter, R. K., Kopp, G. A., & Green, H. C. (1947). Visible speech . New York: Van Nostrand.

Rakerd, B., & Verbrugge, R. R. (1987). Evidence that the dynamic information for vowels is talker independent in form. Journal of Memory and Language, 26, 558-563.

Ryalls, J. H., & Lieberman, P. (1982). Fundamental frequency and vowel perception. Journal of the Acoustical Society of America, 72(5), 1631-1634.

Shearme, J. N., & Holmes, J. N. (1962). An experimental study of the classification of sounds in continuous speech according to their distribution in the formant 1 - formant 2 plane. Proceedings of the 4th International Congress of Phonetic Sciences (pp. 234-240). Hague: Mouton.

Stephenson, D., & Stevens, R. (1979). The Alligator reference manual (Unpublished manuscript). University of Alberta, Department of Linguistics.

Stevens, K. N., & House, A. S. (1963). Perturbation of vowel articulations by consonantal context: an acoustical study. Journal of Speech and Hearing Research, 6(2), 111-128.

Strange, W. (1989). Dynamic specification of coarticulated vowels spoken in sentence context. Journal of the Acoustical Society of America, 85(5), 2135-2153.

Strange, W., & Gottfried, T. (1980). Task variables in the study of vowel perception. Journal of the Acoustical Society of America, 68, 1622-1625.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. Journal of the Acoustical Society of America, 74, 695-705.

Strange, W., Verbrugge, R. R., Shankweiler, D., & Edman, T. (1976). Consonant environment specifies vowel identity. Journal of the Acoustical Society of America, 60(1), 213-224.

Syrdal, A. K., & Gopal, H. S. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. Journal of the Acoustical Society of America, 79(4), 1086-1100.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. Journal of the Acoustical Society of America, 69(5), 1465-1475.

Verbrugge, R. T., & Rakerd, B. (1986). Evidence of talker-independent information for vowels. Language and Speech, 29(1), 39-57.

Verbrugge, R. R., Strange, W., Shankweiler, D. P., & Edman, T. R. (1976). What information enables a listener to map a talker's vowel space? Haskins Laboratories: Status Report on Speech Research, SR-45/46, 63-94.

Wise, C. M. (1964). Acoustic structure of English diphthongs and semi-vowels vis-a-vis their phonemic symbolization. Proceedings of the 5th International Congress of Phonetic Science (pp. 589-593). New York: S. Karger.

Appendix

Sample Answer Sheet for Experiment 2, Part a

	i	ɪ	e	ɛ	æ	ʌ	ɔ	o	ʊ	u
1	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
2	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
3	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
4	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
5	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd

	i	ɪ	e	ɛ	æ	ʌ	ɔ	o	ʊ	u
1	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
2	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
3	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
4	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
5	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd

	i	ɪ	e	ɛ	æ	ʌ	ɔ	o	ʊ	u
1	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
2	[illegible]	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
3	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
4	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
5	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd

	i	ɪ	e	ɛ	æ	ʌ	ɔ	o	ʊ	u
1	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
2	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
3	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
4	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd
5	heed	hid	hayed	head	had	hud	hawed	hoed	hood	who'd