INFORMATION TO USERS

The most advanced technology has been used to photograph and reproduce this manuscript from the microfilm master. UMI films the original text directly from the copy submitted. Thus, some dissertation copies are in typewriter face, while others may be from a computer printer.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyrighted material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each oversize page is available as one exposure on a standard 35 mm slide or as a 17" × 23" black and white photographic print for an additional charge.

Photographs included in the original manuscript have been reproduced xerographically in this copy. 35 mm slides or 6" × 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.
Processes of speaker normalization in vowel perception

Johnson, Keith Allan, Ph.D.
The Ohio State University, 1988

Copyright ©1988 by Johnson, Keith Allan. All rights reserved.
PROCESS OF SPEAKER NORMALIZATION IN VOWEL PERCEPTION

DISSERTATION

Presented In Partial Fulfillment of the Requirements for
the Degree Doctor of Philosophy in the Graduate
School of the Ohio State University

By

Keith Allan Johnson, B.A., M.A.

* * * * *

The Ohio State University
1988

Dissertation Committee:
Mary E. Beckman
Ilse Lehiste
Robert Allen Fox
Neal F. Johnson

Approved by
Mary Beckman
Adviser
Department of Linguistics
Copyright by
Keith Allan Johnson
1988
ACKNOWLEDGEMENTS

My parents have been with me in a very real sense throughout my graduate career. They have seen me through some very hard times with not only care and concern for my well being but also with respect for me as a person. Though this love they have given me myself. This is a gift which cannot be compared to any other. And, it cannot be discharged by a simple acknowledgement.

Debbie Stollenwerk has been my confidant and help. She has been with me, encouraging and supporting me. She believed in me and challenged me to believe in myself.

Wayne Cowart has been a source of encouragement and help for a long time. His interest in my work has always been enthusiastic and uplifting.

I want to also say thank you to the 'colorful ones' in the Ohio State Linguistics Department; those who through their belief in the dignity of people give themselves to others. It would be a drab world without the color infused by such conviction and action.

Finally, I acknowledge the debt that I owe to the members of my committee. I had the unusual good fortune to have a committee of friends. Each in his or her own way provided a substantial contribution to my work and development.
My advisor Mary Beckman made the experience of writing this dissertation a time of learning. She could have let me get by with less, but through her hard work and integrity she taught me something of hard work and integrity. I should also say that her direction has been seasoned with a goodly portion of compassion and patience. Thank you, Mary.

Ilse Lehiste gave me a foundation in phonetics and in the practice of experimental investigation. She has been my outspoken supporter in public and my unyielding critic in private. For both I am most grateful.

Rob Fox was my first teacher in phonetics. The fact that I have written a dissertation which is so closely aligned with his work shows the depth of his influence on me. The design of the experiments reported here and a large part of my understanding of speech perception theory have benefitted from his input. And there is one other intangible something; I always came out of your office, Rob, feeling as if I could conquer the world.

Neal Johnson has had the openness, wisdom and personal magnetism to draw together psycholinguists at Ohio State in a weekly seminar meeting for as many years as I have been at the university. The creative power of this meeting has much to do with my decision to pursue psycholinguistic issues and has influenced my thinking along these lines immeasurably.
VITA

1980 ............................................... ......................... B.A., Abilene Christian University, Abilene, Texas.
1985 ........................................................................... M.A., Department of Linguistics, Ohio State University.
1986 ............................................................................ Visiting Lecturer, Department of Linguistics, University of Illinois.

PUBLICATIONS


FIELDS OF STUDY

Major Field: Linguistics
Speech Perception: Rob Fox.
Phonetics: Ilse Lehiste, Mary Beckman, Rob Fox.
Historical Linguistics: Brian Joseph, David Stampe, Ilse Lehiste.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS ................................................................. ii
VITA ................................................................................................ iv
LIST OF TABLES ........................................................................ vii
LIST OF FIGURES ......................................................................... viii
CHAPTER PAGE

I. INTRODUCTION ................................................................. 1

Processes in the perception of speech ........................................ 1
The acoustic description of vowels ........................................... 4
Speaker variation in vowels sounds ........................................... 6
Range normalization ............................................................... 7
Perceptual normalization ......................................................... 10
Dynamic factors in vowel perception ....................................... 12
Auditory processing ............................................................... 14

II. INTONATIONAL CONTEXT AND F0 NORMALIZATION ............. 20

Experiment 1 ................................................................. 27
Method ............................................................... 28
Subjects ....................................................... 28
Materials ................................................. 28
Procedure .................................................. 32
Results ..................................................... 33
Discussion ................................................... 34
Experiment 2 ................................................................. 39
Method ............................................................... 39
Subjects ....................................................... 39
Materials ................................................. 40
Procedure .................................................. 41
Results ..................................................... 41
Discussion ................................................... 42
Conclusion ...................................................... 46
III. HIGHER FORMANT NORMALIZATION ......................................... 47

Experiment 3 ................................................................. 50
  Method .................................................. 51
  Subjects ........................................... 51
  Materials ......................................... 51
  Procedure ......................................... 53
  Results ............................................... 54
  Discussion ........................................... 55

Experiment 4 ................................................................. 57
  Method .................................................. 58
  Subjects ........................................... 58
  Materials ......................................... 58
  Procedure ......................................... 61
  Results ............................................... 63
  Discussion ........................................... 66
  Conclusion ............................................. 68

IV. CONCLUSION ................................................................. 70

Summary ................................................................. 70
Implications .............................................................. 72
  The ‘centre of gravity’ effect ......................... 72
  Levels of processing ................................. 74
  Perception and production ......................... 75
  Sensory processes and cognition ................. 76

BIBLIOGRAPHY .............................................................. 78
# LIST OF TABLES

<table>
<thead>
<tr>
<th>TABLE</th>
<th>PAGE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formant values of the test tokens used in experiment 1</td>
<td>29</td>
</tr>
<tr>
<td>2. Formant values for the tokens in experiment 3</td>
<td>52</td>
</tr>
<tr>
<td>3. F2 values for the 'hid'-'head' continuum in experiment 4</td>
<td>61</td>
</tr>
<tr>
<td>4. Crossover points for F1 and F2 in the 'hood'-'HUD' continuum in 4</td>
<td>67</td>
</tr>
<tr>
<td>FIGURE</td>
<td>PAGE</td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
</tr>
<tr>
<td>1. Vowel formant averages for nine vowels of American English. Average male, female and child values are plotted separately for each vowel. Data from Peterson and Barney (1952)</td>
<td>9</td>
</tr>
<tr>
<td>2. Normalized vowel formant averages for nine vowels of American English. Range normalized values of the data in figure 1</td>
<td>9</td>
</tr>
<tr>
<td>3. The Bark scale of auditory frequency response (from Zwicker, 1961)</td>
<td>15</td>
</tr>
<tr>
<td>4. Bark transformation of the Peterson and Barney (1952) vowel formant data for nine American English vowels</td>
<td>16</td>
</tr>
<tr>
<td>5. Pitch traces for three natural productions of the phrase, 'Is this HUD?'</td>
<td>30</td>
</tr>
<tr>
<td>6. F0 contours of the synthesized intonational contexts in experiment 1</td>
<td>31</td>
</tr>
<tr>
<td>7. Results of experiment 1 showing F0 level and context condition</td>
<td>33</td>
</tr>
<tr>
<td>8. Identification functions of the three F0 levels and two context conditions in experiment 1 as a function of F1/F0 ratio</td>
<td>36</td>
</tr>
<tr>
<td>9. Identification functions of the three F0 levels and two context conditions in experiment 1 as a function of token number</td>
<td>38</td>
</tr>
<tr>
<td>10. F0 contours of the synthesized intonational contexts in experiment 2</td>
<td>40</td>
</tr>
<tr>
<td>11. Results of experiment 2 showing F0 level and context condition</td>
<td>41</td>
</tr>
</tbody>
</table>
12. Identification functions of the three F0 levels and two context conditions in experiment 2 as a function of token number ....................................................... 43

13. Identification functions of the three F0 levels and two context conditions in experiment 2 as a function of F1/F0 ratio ....................................................... 44

14. Results of experiment 3. Identification data are classified by vowel continuum (hid-head vs. hood-HUD) and higher formant condition (high vs. low) ........................................................ 54

15. Results of experiment 3. ‘hood’-'HUD’ continuum by token .. 56

16. Results of experiment 3. ‘hid’-'head’ continuum by token ..... 56

17. Two spectra of token 1 in the ‘hood’-'HUD’ continua which illustrate the result of narrowing the bandwidth of F3. The amplitude of F3 is approximately 4 dB greater with F3 bandwidth of 30 Hz (a) as compared with F3 bandwidth of 110 Hz (b) ....................................................... 60

18. Identification functions for the ‘hid’-'head’ continua of four subjects in experiment 4 who mislabelled the continuum. These data were not included in the ANOVA ............... 62

19. Results of experiment 4. Identification data are classified by vowel continuum (hid-head vs hood-HUD) and higher formant condition (high vs. low) ........................................................ 64

20. Results of experiment 4. Identification functions for the ‘hood’-'HUD’ continua ....................................................... 65

21. Results of experiment 4. Identification functions for the ‘hid’-'head’ continua ....................................................... 65
Chapter I
Introduction

This dissertation is about the perceptual processes involved in speaker normalization in vowel perception. From this rather narrow investigation I hope to be able to expand on the larger issue of how to characterize speech perception processes in general.

Processes in the perception of speech.

Speech perception is a cognitive activity. It involves the identification of speech events as lexical items. The processes which access lexical items in long term memory are necessarily cognitive because lexical items are cognitive categories. However, speech perception is also a sensory activity. It involves the response of the auditory system to speech events. Speech is delivered to the central nervous system via the auditory system, and so speech perception depends upon the auditory sensory system.

A natural concern in speech perception research has to do with identifying those aspects of speech perception which are the result of cognitive level processing and those aspects which result from auditory level processing. By the nature of the task speech perception processes cannot be either cognitive or auditory, rather both cognitive and auditory levels of processing are involved.
It is important to be able to identify the issues in speech perception theory within a view of the task of speech perception which considers both cognitive and auditory level processing. For instance, the debate over the 'specialness' of speech (see Liberman and Mattingly, 1985 for a review) is not a debate about whether speech perception involves cognitive processes or auditory processes, but rather it is a debate about the uniqueness of the cognitive processes involved in speech perception; that the type of cognitive decoding performed during speech perception is unique to speech because of a biological link between speech production and perception. It is important, however, to see what the issue is in relation to an overall understanding of the process of speech perception. The fact that speech perception involves both cognitive and auditory levels of processing leads us toward a more basic and (as yet) unsolved problem. Before debating the nature of the cognitive processes involved in perceiving speech it is essential to be able to identify them.

The emphasis on auditory effects in speech perception (Cutting, 1978; Schouten, 1980; Pastore, 1981) must also be viewed from a comprehensive perspective. Some speech perception phenomena can be explained by reference to characteristics of the auditory system, but this does not demonstrate that these phenomena do not also involve cognitive processing. As I will show in chapter 2 an auditory explanation does not rule out the possibility of a parallel cognitive level process. The fact that an auditory explanation is available for a speech perception phenomenon simply raises the possibility that the cognitive part of the speech perception process may be somewhat less complicated than might have been previously assumed (perhaps a little more like template matching). The goal of
studying speech perception must be the description of all of the processes used to perform speech perception; both auditory level processes and cognitive level processes.

The study of speech perception provides an interesting opportunity to observe how the human cognitive system interacts with and is shaped by the world of experience. Because most other cognitive systems also involve the adaptation of humans to their environment (at some level) it is reasonable to suggest that the study of the processes of speech perception is relevant to the study of other cognitive systems.

In the sections to follow I will focus on the details surrounding the experiments reported here. Some of this background material may be familiar to the reader and so as a guide to reading I will briefly describe the remaining sections in this chapter. The next section, entitled ‘the acoustic description of vowels’, is a short summary of the acoustic theory of speech production and may be skipped by readers already familiar with that theory. Following that section is a section entitled ‘speaker variation in vowel sounds’ which points out the acoustic consequences of differing vocal tract sizes. The section on ‘range normalization’ describes a vowel normalization algorithm and uses the Peterson and Barney (1952) vowel formant data in example calculations of normalized values. The section after that (‘perceptual normalization’) discusses some of the speech perception literature which seems to indicate that hearers perform a type of cognitive level vowel normalization similar to the range normalization algorithm. In the next section I describe some ‘dynamic factors in vowel perception’ which are important stimulus properties used in vowel perception. The final section of the chapter discusses some features of the auditory system which
may be relevant to vowel normalization. An hypothesis about auditory vowel normalization is presented in this final section.

The acoustic description of vowels.

The acoustic properties of vowels are determined by both a sound source and the vocal tract (Fant, 1960, Chiba and Kajiyama, 1941). In voiced vowels the vibrating vocal folds are the source of acoustic energy. The acoustic parameters used to describe the voice source include the rate of vibration of the vocal folds (the fundamental frequency) and the spectrum of the voice source, which includes harmonic components at integral multiples of the fundamental. For whispered vowels the source is the hissing noise of air rushing past the edges of the vocal folds in the open glottis. This noise source is very different from voicing in that it has no fundamental frequency and a relatively smooth spectrum. The acoustic properties of vowels which are determined by the vocal tract have to do with the modification of the source spectrum by the vocal tract’s acoustic filtering properties. The resonant frequencies of the vocal tract are called formants and are normally taken to be the most important acoustic properties of vowels. The formants result from the length and shape of the vocal tract. It is important for an understanding of vowel perception to note that the formants are not necessarily directly observable acoustic energy in vowels but rather can only be seen indirectly as they shape the spectrum of the sound source. In voiced vowels the formants are realized as amplitude modifications on harmonics in the voicing spectrum. Harmonics near the formants are enhanced while other harmonics are damped. Thus, vowel formants may not coincide with any harmonics (in voiced vowels). The
resonances of the vocal tract must be inferred by their influence upon the voicing spectrum (see Carlson, et al. 1975, Assman and Nearey, 1986 and Darwin and Gardner, 1985 for discussion of the role of harmonics in perceiving formants).

Changing the shape and length of the vocal tract changes its resonant characteristics. Thus, acoustically, vowels differ from each other by having different formants. The lowest formant (F1) is inversely related to the degree of mouth cavity openness - being higher in frequency for open vowels such as [a] in 'hot' and [æ] in 'hat' and being lower in frequency for closed vowels such as [i] in 'heat' and [u] in 'hoot'. The second formant (F2) in vowels is related to the location of constriction within the vocal tract - being lower in frequency for vowels with a constriction toward the back of the mouth such as [a] in 'hot' and [u] in 'hoot' and being higher in frequency for vowels with a constriction toward the front of the mouth such as [æ] in 'hat' and [i] in 'heat'. Vowels which are produced with rounded lips have formants which are lower than they would be with the same mouth configuration and no lip rounding. This situation is approximated for the vowels [a] in 'cot' and [ɔ] in 'caught' in some dialects of American English. One reason that lip rounding results in generally lower formants is that the length of the vocal tract is effectively increased by rounding. Just as the pitch produced by a trombone is lowered by increasing the length of the tube through which the buzz of the player's lips must travel, so also the resonances of the vocal tract are lowered as the length of the tract increases.
Speaker variation in vowel sounds.

The connection between vocal tract length and vocal tract resonance is one source of the relatively large differences in the formant values of vowels produced by different speakers. Speakers with different length vocal tracts will necessarily produce vowels which have different formants. This is especially true for children imitating adult productions of words. The child's vocal tract is so much shorter than an adult's that there is no way that the imitation of an adult's production of a word can be an acoustic replication.

So, from the very beginnings of language acquisition the identification of one speaker's production of a vowel with another speaker's production of that vowel cannot be based on their being acoustically identical, but rather some more abstract representation is necessary. Presumably, this abstract level of representation can be characterized by the descriptive apparatus commonly used in linguistics. It is not my goal in this dissertation to define the abstract representation of vowels; rather I am concerned with the spectral normalization processes which seem to be involved in the identification of vowels.

Functionally, these processes must be able to selectively disregard the acoustic variation found in vowels produced by different speakers while maintaining sufficient separation between vowels to be successful in identifying them. Very simply, we could say that spectral normalization processes involve the elimination or reduction of speaker variation (variation from speaker to speaker producing the same vowel) while maintaining phonetic contrast.
Range normalization.

'Vowel normalization' can be used to refer to computational techniques in speech recognition systems or to hypothetical perceptual process in vowel perception. In the remainder of this chapter I will use the term in the first sense, and will call the perceptual process 'perceptual vowel normalization'. In chapters 2 and 3 I will use 'vowel normalization' exclusively to refer to a perceptual process (or processes).

As an example of one type of vowel normalization algorithm consider the proposal of Gerstman (1968). In this approach (called range normalization) each vowel formant value is transformed into a value which is scaled to the range between maximum and minimum values for that formant and that speaker. Instead of basing vowel identification on absolute formant values this approach converts the formant values into relative values. These relative (or range normalized) values indicate where a vowel's formants fall in the range of observed values for each formant for a particular speaker. The rescaled formant value $s$ is found by the use of the following formula:

$$s = \frac{x - l}{r}$$

where:

- $x$ is the observed formant measurement in Hz.,
- $l$ is the lowest observed value for that formant in Hz., and
- $r$ is the range of observed values for that formant in Hz.

The formant range $r$ is found by formula (2).

$$r = h - l$$

where:

- $h$ is the highest observed value for that formant in Hz., and
- $l$ is the lowest observed value for that formant in Hz. (as in formula 1).

For each speaker and each formant there are different values of $h$, $l$, and $r$. 
Using the average formant measurements of Peterson and Barney (1952) as raw data for the purposes of demonstrating range normalization we can calculate the normalized values of F2 for [ɛ] in the following way. The range of male F2 averages in Peterson and Barney’s data is 1450 Hz (2290-840 (F2 of [i] minus F2 of [ɔ])). The absolute value of F2 of [ɛ] is 1840 Hz. This is 1000 Hz up from the bottom of the range (1840-840). The relative position of F2 for [ɛ] then is about .69 of the F2 range (1000/1450).

The F2 range in the female data is 1870 Hz. Average F2 of [ɛ] is 2330 which is 1410 Hz above the bottom of the female range. The relative (or normalized) position of F2 for female [ɛ] is thus .75 (1410/1870). This is much more similar to the male normalized value for F2 of [ɛ] than is the absolute F2 value of female [ɛ] which is higher than F2 of male [i]. Figures 1 and 2 demonstrate the difference between observed and normalized vowel formant values. Figure 1 plots the vowel formant data reported by Peterson and Barney (1952) in the F1-F2 space. Figure 2 presents normalized values for these same tokens. As is obvious from the figures the normalization procedure reduces the variability among speakers while preserving the variation among different vowel qualities.

The normalization procedure illustrated here achieves the same type of results as the hypothetical perceptual processes which were discussed earlier. Speaker-to-speaker differences are reduced while vowel-to-vowel contrast is preserved. The question is, ‘How do hearers perform perceptual vowel normalization?’
Figure 1
Vowel formant averages for nine vowels of American English. Average male, female and child values are plotted separately for each vowel. Data from Peterson and Barney (1952).

Figure 2
Normalized vowel formant averages for nine vowels of American English. Range normalized values of the data in figure 1.
Perceptual normalization.

In order for a range normalization algorithm such as that in Gerstman (1968) to work there must be some indication of the acoustic range of a speaker's vowel sounds. Gerstman recommended providing a computerized recognition system with three vowels produced by a speaker ([i], [a], and [u]) before attempting to perform recognition on other utterances. Clearly this does not accurately model what human listeners do since hearers have no difficulty in identifying the vowels of words which have been produced by different speakers even when the words have been excized out of their original context (Verbrugge, et al., 1976; Strange, et al., 1976); hearers do not require input of actual endpoint values for a given speaker in order to perceive the speaker's vowels accurately. There is, however, some perceptual evidence which is consistent with the existence of a range-normalization process in human vowel perception.

Ladefoged and Broadbent (1957) found that vowel identification results for the same test vowels were different when the range of vowel formants in a precursor phrase were changed. Ladefoged and Broadbent suggested that the precursor phrase caused a perceptual shift in the hearers' internal 'vowel space'. This interpretation of these data relies implicitly on a notion of range normalization. Asserting that the hearer's internal vowel space is shifted is essentially the same as saying that the expected ranges for vowel formants were shifted.

Ainsworth (1975) also found that when test vowels were presented after the presentation of 'corner' vowels ([i], [a], and [u]) whose formant values demarcated different formant ranges the perceptual identification of
the test vowels changed. These results again indicate some sort of range normalization.

Van Bergem et al. (1988), in a replication of the Ladefoged and Broadbent study, found that a precursor phrase influenced the perception of vowel quality while a following phrase did not. It is interesting to note that vowel contrast effects (e.g. Crowder, 1981) are not directional in this way. Rather, a shift in identification due to vowel contrast in auditory memory occurs both when the contrasting vowel precedes and when it follows the vowel being identified. These data support the original interpretation of the data indicating a cognitive level range normalization process.

If a range normalization process does occur in the perception of vowel sounds it is interesting to wonder to what extent it might be triggered by inherent, internal properties of vowels such as fundamental frequency or higher formant structure. Conceivably, such factors could influence the hearer's internal vowel space in a way similar to the perceptual adjustment hypothesized to result from precursor phrases and vowels. High F0 could be taken as an indication that the speaker has a short vocal tract and therefore formant ranges which would be somewhat high and large. Consider for example, a vowel with F1 ambiguous between values appropriate for male [ɔ] and [ʌ] (say 550 Hz). If the vowel has F0 typical of male speakers then the range expected for F1 is that of the male averages in Peterson and Barney (1952) and the range normalized value of F1 is .61 (ambiguous between [ɔ] and [ʌ]). On the other hand, if the expected F1 range is that of females (if the syllable's F0 is in the range typical for female speakers) then the range normalized value is .44 which is much
more like [@]. If the expected range for a formant can be influenced by vowel internal factors such as F0 then we expect to find perceptual vowel normalization effects such as those found in Ladefoged and Broadbent (1952) and Ainsworth (1975) even in the absence of precursor phrases or vowels. Such perceptual vowel normalization effects have been found for both F0 and higher formants (the literature will be discussed in the introductions to chapters 2 and 3).

So, the cognitive processes involved in perceptual normalization may use any kind of speaker information, whether it is the vowel qualities in a precursor phrase or vowel internal properties such as F0 and higher formants. The speaker differences in formant values which result from differences in vocal tract size may be perceptually compensated for by a normalization process which interprets the formant values based on an estimate of the speaker's vocal tract length (the articulatory correlate of expected vowel space) which itself can be based on both vowel external factors such as precursor speech and vowel internal factors such as F0 and higher formants.

**Dynamic factors in vowel perception.**

In describing the process of vowel perception, however, we need to keep in mind the other properties of naturally produced vowels which can be useful for vowel identification. In addition to different formant values among vowels there are also duration, amplitude, pitch and formant movement differences. It has been found that, all other things being equal, the different vowels of American English have differences in 'inherent' duration (Peterson and Lehiste, 1960), amplitude (Lehiste and Peterson,
1959), fundamental frequency (Lehiste and Peterson, 1961b) and formant movement patterns (Lehiste and Peterson, 1961a). (See Lehiste, 1970 for a review of the evidence for cross-language patterns of inherent vowel differences.) Low vowels tend to be longer, louder and lower in pitch than the high vowels. The difference between tense and lax vowels (such as the vowels in 'heed' and 'hid') is primarily one of duration, although their formant movement patterns are also different (see DiBenedetto, 1987; and Huang, 1986).

The relative importance of these additional types of information can be gauged by the results of a very interesting study conducted by Hillenbrand and McMahon (1987). They used the individual formant measurements collected by Peterson and Barney (1952) as synthesis parameters for steady-state vowels which were presented to subjects for identification. They found that when the only vowel identity information available to listeners was the formant values and fundamental frequency these vowels were correctly identified only 75% of the time. Peterson and Barney included in their data base only values from syllables which were unambiguously identified. So, extrapolating from the Hillenbrand and McMahon study, we may conclude that duration, amplitude and dynamic information disambiguated 25% of the syllables found in Peterson and Barney's corpus. Considering that Peterson and Barney's measurements were of careful productions of stressed vowels in the context of [h__d] (which is relatively neutral in terms of coarticulation) this is a remarkable finding and points out the need for research into the role of formant movement patterns in vowel perception. Lehiste and Meltzer (1973) similarly found a 25% error rate in the perception of steady-state vowels
synthesized using the averages for men, women and children's vowels in the Peterson and Barney results. This convergence of results confirms the relative importance of dynamic information in vowel identification.

It is clear that hearers in a natural situation do not depend solely on isolated formant values, but also use dynamic vowel information such as vowel duration (Verbrugge and Shankweiler, 1977; Verbrugge and Isenberg, 1978) and formant movement patterns (Strange, et al., 1977; Jenkins, et al., 1983; and Parker and Diehl, 1984) in vowel perception. Thus, the process of vowel perception does not have to rely on formant values alone. Studies such as those of Syrdal and Gopal (1986), Hillenbrand and Gayvert (1987) and Miller (1987) which use vowel formant and fundamental frequency values as input to vowel identification algorithms leave out stimulus information which is evidently used by hearers in making vowel identification judgments. Of course, in computerized speech recognition tasks it is interesting to know whether vowel formant information is enough for 90-95% correct identification because it is useful to be able to reduce the amount of stimulus information handled in a computer application, but as a model of human perceptual processing such an approach is too simple.

Auditory Processing.

In addition to the information to be found in formant movement patterns and intrinsic duration, amplitude and pitch, it is probably also the case that the auditory transformation of vowels results in a representation which has some 'built in' normalization. Through the use of masking techniques the frequency response of the peripheral auditory system has been estimated (Zwicker, 1961; Scharf, 1970). The relationship between the
acoustic unit of frequency (Hz) and the psychophysically determined auditory unit of timbre (Bark) is expressed by formula 3 and is illustrated in figure 3.

\[
3 \text{ Bark } = 26.81 \frac{f}{(1.96 + f)} - 0.53 \quad \text{where } f \text{ is frequency in kHz}
\]

The nonlinearity of the ear's response to spectral energy may reduce the speaker-to-speaker differences among vowels in two ways. First, variation in the F2 region is reduced due to Weber's law - the ratio between the initial stimulus and the just noticeable difference is invariant. As F2 increases a greater F2 change is required for hearers to notice the difference between two vowels. The nonlinearity of the Bark scale affects F1 values primarily by compressing slightly values in the region from 600-1000 Hz. The Peterson and Barney (1952) averages are presented in figure 4 in their Bark transformed values.
A comparison of this figure with figures 1 and 2 indicates that Bark transformation makes average values for a vowel for the three different speaker types closer to each other, but not as close as would be accomplished using range normalization. Clearly, much speaker variation remains in the Bark transformed values.

One additional auditory effect which may be involved in reducing speaker-to-speaker variation has been called the 'centre of gravity' effect (Chistovich et al., 1979). This perceptual phenomenon is hypothesized to result from an integration of energy in the auditory spectrum over a range
of 3 - 3.5 Bark. An integration of spectral energy can be thought of as a sort of smoothing of the spectrum. Traunmüller (1981) hypothesized that the role of F0 in the perception of vowels arises from the integration of the first harmonic (which is always a prominent peak in the spectrum) and the harmonics around the first formant. The lowest harmonics and the harmonics around F1 are integrated into one 'centre of gravity', thus, as F0 increases and the high amplitude lowest harmonics move closer to the F1 region the perceived F1 is drawn down toward the first harmonic even if the actual resonance of the vocal tract remains fixed. Data reported by Darwin and Gardner (1985) and Assman and Nearey (1987) support this interpretation of auditory F0 normalization. In both of these studies the amplitude of harmonics in the F1 region was manipulated. Increasing the amplitude of a harmonic above F1 resulted in an increase in perceived F1 and increasing the amplitude of a harmonic below F1 resulted in a decrease in perceived F1. The connection between these studies and Traunmüller's explanation of F0 normalization is that as F0 increases the higher amplitude first harmonic approaches the F1 and so the perceived F1 decreases. This linking of F1 and F0 is limited to F0 between 150 Hz and 350 Hz (Traunmüller, 1981; Bladon, et al., 1984). Traunmüller's (1981) experiments suggest that the F0 normalization effect found in the perception literature may actually result from peripheral auditory processing of this sort rather than from the kind of range normalization process suggested earlier.

Traunmüller (1982) suggested that the centre of gravity effect may be speech specific or at least non-auditory in nature. The results which prompted this suggestion came from an experiment in which subjects were asked to rate the similarity of simple tones to noise excited two formant
vowels (or single formant tokens and two formant tokens). The tokens presented were 'speech-unlike' and subjects were not instructed to use vowel category labels in making their judgments. In the Chistovich et al. matching experiments (in which the test tokens were identified by vowel category labels) when F2 minus F1 > 3.5 Bark subjects tended to adjust a single formant token so that the formant approximated F2 in the two formant standard. Traunmüller (1982) found that subjects judged simple tones (or single formant tokens) which had a frequency equal to F2 in a two formant standard to be quite different from the two formant vowel when F2 minus F1 > 3 Bark. Subjects judged the simple tone tokens (and the single formant tokens) to be more similar to the two formant standard when the frequency of the test token matched the F1 of the standard. This is not really evidence of a lack of three Bark integration when the test items are nonspeech (in fact the figures presented in this study seem to support the centre of gravity effect), rather these data indicate a greater perceptual saliency for F2 in speech as opposed to nonspeech. Thus far, there seems to be no reason to conclude that the centre of gravity effect is speech specific. Rather, it seems to result from general properties of the auditory periphery (though perhaps not very well understood properties).

This brings us then to the central question of this dissertation. How should speaker normalization be characterized? Is it an auditory phenomenon which is entirely determined by the psychophysical properties of the peripheral auditory system, or is it necessary to propose the existence of a higher level cognitive process which uses speaker information to interpret vowel sounds in a rough range normalization process? If we were to include in our description of vowel stimuli all of the properties of vowels
which are a part of the auditory system's analysis would there be any work left for a cognitive normalization process?

In this dissertation I will report on a set of experiments which address this issue. In chapter 2 I hypothesize the existence of two types of F0 normalization: the auditory normalization which was proposed by Traunmüller (1981) and a speaker-based range normalization process based on F0. The experiments of chapter 2 are an attempt to separate the two types of normalization and by doing so to demonstrate the existence of a speaker-based, cognitive level process of F0 normalization.

The experiments in chapter 3 test the hypothesis that higher formant normalization results from the centre of gravity effect by comparing the effect of changing higher formants in vowels with high F2 to the effect of changing higher formants in vowels with low F2.

Finally, chapter 4 is a summary and further discussion of the theoretical implications of these experiments.
Peterson and Barney (1952), in a descriptive study of the acoustics of vowel production, found great variability in the formant values of vowels produced by men, women and children. They also found that average F0 was correlated with formant values. Since the initial study of Peterson and Barney, a number of researchers have proposed algorithms to predict vowel identities from the acoustic measures reported for the vowels by using F0 as a factor for normalizing the observed formant values (Hillenbrand and Gayvert, 1987; Miller, J.D. 1987; and Syrdal and Gopal, 1986). The correlation between F0 and relative formant values makes for reduced interspeaker variation for the computationally normalized values within the corpus and labeling results which are nearly as accurate as the listening data reported by Peterson and Barney. Of course the fact that a relationship between F0 and formant values exists and that it is useful in computer simulations of vowel identification does not establish that there are perceptual normalization processes which include F0. However, there are a number of studies which seem to indicate that just such a perceptual normalization of vowel formants based on F0 takes place.

Miller (1953) found that by doubling the fundamental frequency of vowel tokens (from 144 Hz to 288 Hz) the perceptual boundaries between
vowel categories along synthetic continua could be shifted. Among back vowels he found that there was an 80 Hz shift in the [o] - [A] boundary for F1 (the F1 had to be 80 Hz higher in the high F0 condition to be labeled as [A]). Among front vowels the [ɛ] - [ɛ] boundary showed the greatest amount of displacement (30 Hz along the F1 dimension). All of the shifts in identification reported by Miller (1953) are consistent with the existence of a perceptual process of normalization based on F0. Fujisaki and Kawashima (1968) similarly found that an increase in F0 resulted in an increase in the number of [o] responses in an [o]-[a] continuum for Japanese listeners. Also, in an [u]-[e] continuum they found a boundary shift toward [e] (ie. increased number of [u] responses) as a result of increased F0, although it was not as large as the effect found in the purely back vowel series. Again, these results indicate that a perceptual normalization takes place.

In a study of subjects' estimates of vowel differences Slawson (1968) found that doubling F0 (from 135 Hz to 270 Hz) resulted in responses indicative of a perceptual shift in a vowel perception task. Subjects estimated the differences between vowel pairs in a scaling paradigm. From these difference estimates it is possible to compute a perceptual vowel space. The effect of F0 shift was a shift in the locations of greatest perceived differences which was consistent with the hypothesis of perceptual normalization based on F0. This effect was greater for the compact vowels ([ɔ],[a], [ɔ]) than for the noncompact vowels in the study. Ainsworth (1971,1975) studied the effect of F0 on vowel perception by calculating (using a least squared error method) the point in the F1-F2 space which best represented listeners' identification responses. Using five different levels of F0 (75,120,170,260,365 Hz) he found that there was an effect of F0 upon
vowel perception and that it seemed to affect perceived F1 more than perceived F2 (5% shift of perceived F1 versus 2% shift of perceived F2).

A common thread running through these studies is that the perceptual process called F0 normalization seems to affect the low frequency components in the spectra of vowels (say below 1000 Hz) more than it does high frequency components. Thus, vowels with low F2 are more likely to shift in identification when F0 is changed than are vowels with high F2, and F0 normalization seems to affect F1 more than F2.

The research of Traunmüller (1981) supports this generalization and provides an auditorily based interpretation of the phenomenon. He found that in the identification of single formant vowels the perception of openness seemed to depend upon a relation between F0 and F1 such that as F0 increased perceived F1 decreased. This relationship was constant below F0 of 350 Hz. Above 350 Hz fewer categories of vowel height could be perceived and absolute formant value seemed to be more important to categorization of vowels. In a second experiment he demonstrated that shifting F0 and F1 together along a modified tonality scale in vowels which were synthesized with five formants the perceived openness was not greatly changed, while shifting F1 alone (in a third experiment) resulted in category shifts. Finally, he varied F0 alone and found results indicating an F0 normalization process similar to the findings of Miller (1953) and Fujisaki and Kawashima (1968) and Slawson (1968), although the shift in vowel identification function was much greater in magnitude. Unlike previous researchers, however, Traunmüller explains the F0 normalization process as auditory in nature rather than as a cognitive level perceptual process.
His description of the auditory process of F0 normalization is essentially that of Chistovich et al. (1979). Chistovich et al. found that when the interval between two formants is less than 3 to 3.5 Bark the single formant vowel which is judged by hearers to match the quality of the two formant standard has F* (the single formant) such that F1 < F* < F2. However, when the interval between F1 and F2 exceeds 3 Bark F* is no longer positioned between the two formants of the standard but rather is adjusted to match the F2 of the standard. Chistovich et al. hypothesized that this pattern of response is the result of a feature of peripheral auditory processing. They propose that vowel perception involves two auditory integration processes. These integration processes can be conceptualized as a kind of smoothing or filtering of the frequency spectrum, and in fact, in computational applications a Bark transformation of the signal is often achieved by running the signal through a bank of filters which are equally spaced in Bark.

The first stage is integration over 1 Bark intervals on the basilar membrane. This level of frequency integration has been extensively studied using masking techniques and is well established.

The second stage, on the other hand, is novel to Chistovich et al. They propose that there is a second level of processing at which another frequency integration occurs. This second stage of spectral smoothing is hypothesized to extend over a range of 3 to 3.5 Bark. Traunmüller (1981) hypothesized that the role of F0 in the perception of vowels arises from the integration of the first harmonic (which is always a prominent peak in the spectrum) and the harmonics around the first formant. The lowest harmonics and the harmonics around F1 are integrated into one 'centre of
gravity', thus, as F0 increases and the high amplitude lowest harmonics move closer to the F1 region the perceived F1 is drawn down toward the first harmonic even if the actual resonance of the vocal tract remains fixed. He further hypothesizes that the auditory pattern which results from F0 and F1 is matched to a feature template, but that the entire vowel spectrum does not have to match a vowel template. Sussman (1986) also argues for the existence of low level auditory phenomena which are sensitive to ratios of different stimulus properties (like F1/F0).

The hypothesis to be tested in this chapter is that the F0 normalization effect reported in the vowel perception literature is actually the result of a combination of two different perceptual processes. These two normalization processes I will call: (1) auditory F0 normalization and (2) speaker-based F0 normalization. Slawson (1968) proposed a similar explanation of speaker normalization in which he hypothesized the existence of two stages of normalization (1) auditory sensation and (2) phonetic classification. Although he didn’t pursue this proposal very far it does seem that Slawson’s proposal is similar in spirit to my own. In this view auditory F0 normalization is hypothesized to operate as a low level auditory effect perhaps in a manner similar to that suggested by Traunmüller(1981). As a low level effect it is not open to later modification (i.e. cannot be undone by later processing, but rather provides input to later perceptual processes). Auditory F0 normalization as envisioned here would also not depend on speaker identity, sentential context or any other information that implies the application of knowledge about the linguistic system or speaker population. Rather, this effect is the result of the nature of auditory spectral processing in the human auditory system, and in particular is the
result of how the spectrum of incoming complex sounds is transformed by the peripheral auditory system.

The type of F0 normalization which I am calling speaker-based normalization, on the other hand, is a cognitive perceptual process. As it is envisioned here, speaker-based F0 normalization involves a perceptual rescaling of formant values in a vowel spectrum based on an estimate of the speaker's vocal tract length. This estimate of vocal tract size is derived (partly at least) from an awareness of the correlation of F0 and formant values noted above. Obviously, by its very nature we would expect speaker-based F0 normalization to depend on extrinsic factors (speaker voice quality, sentential context, etc.) in a way that could not be expected for auditory F0 normalization. It is this difference between the two types of F0 normalization which allows for a test of this hypothesis about F0 normalization.

Why should we suppose that the F0 normalization data reviewed above reflect the simultaneous operation of two separate types of perceptual process? We can assume that at least some auditory normalization takes place because of the reduction of perceptual distances in the F2 range produced by Bark transformation (see chapter 1), so the real question here is, 'why should we suppose that speaker-based normalization was involved in these studies?'

When a subject is presented with syllables in isolation for identification, and the syllables vary in F0 there are two possible interpretations of the F0 variation open to the subject. He may attribute the F0 differences among the tokens to speaker variation, or he may interpret
these differences as variation in the range of pitch used by the same speaker.

One way that pitch range variation is used in American English (and many other languages) is when speaking up (to speak in order to be heard at a distance). When speaking up the range of F0 for the utterance is expanded. The expansion of pitch range due to speaking up occurs mainly as an increase in the pitch of peaks in the intonational contour rather than a decrease in the pitch of intonational valleys (Liberman and Pierrehumbert, 1984). Intonational peaks show greater variability under changes of pitch range than do intonational valleys. Another way that pitch range variation is used is to signal degree of emphasis. In emphatic speech the location of increased pitch range variation is normally the prominent words of an utterance. Emphasis is achieved by increasing pitch range, amplitude and duration of stressed syllables within accented words. The main difference between pitch range expansion in speaking up and in producing emphasis is that the expansion of the pitch range is global in nature when speaking up, while it is local in nature in producing emphasis. In other words, the range is expanded primarily for the emphatic word in producing emphasis, while the entire utterance is affected by pitch range expansion in speaking up.

One aspect of the typical F0 normalization experiment which indicates that hearers are probably not interpreting the F0 variation as pitch range variation but rather as speaker variation is that in these experiments it is normal for only the F0 level to be varied from one condition to the next. Thus, the only difference between the high F0 condition and the low F0 condition is the F0 level of the tokens. The duration, voice quality and every other parameter of the synthesized tokens is left the same. In natural
speech, however, it is normal to expect durational and amplitude changes to correlate with F0 changes due to increased emphasis and at least some voice quality change due to speaking up. Since this type of correlation does not occur in the tokens of the typical F0 normalization experiment it is likely that subjects interpret the F0 variation in the tokens as being speaker variation. So, one assumption of the experiments to be reported below is that the F0 normalization of items presented in isolation is a combination of both an auditory effect and a cognitive level effect.

The goal of the experiments in this chapter was to separate these two effects.

Experiment 1.

In this experiment I attempted to reduce the likelihood that the speaker-based F0 normalization process would apply by presenting tokens in an intonational context such that the F0 variation among the experimental items could be associated with linguistic variation (intonational difference) rather than with speaker variation. If there is a speaker-based F0 normalization process we would expect its effects in the identification of the continuum to be reduced by the intonational context.

The experiment involves the presentation of vowel continua from \[\alpha\] to \[\lambda\] which have formant values ambiguous between male and female speakers. The tokens are steady state vowels in the consonantal environment [hVd]. Three different continua were synthesized using three different levels of F0. Pilot results for \[\alpha\]-[\lambda] vowel series replicated the findings of Fujisaki and Kawashima (1968) and Miller (1953).
In addition to F0 level as an independent variable the intonational context of the test items was manipulated. The tokens were presented in isolation and also inserted into sentences which provided an intonational context appropriate for the F0 of the token. It was hoped that these intonational contexts would lead subjects to conclude that all of the utterances were produced by the same speaker.

Method.

Subjects.

Ten subjects (2 males and 8 females) participated in the experiment. Subjects were native English-speaking undergraduate students at Ohio State University, who reported normal hearing and no history of speech or hearing trouble. Data from each subject were collected in a single one hour session. Subjects received $3.50 for their participation in the experiment and were recruited by an advertisement in the University paper, and in introductory Linguistics courses.

Materials.

Three vowel continua from [hɔd] to [hʌd] were synthesized using a version of the Klatt (1980) software formant synthesizer. The formant values in all three continua were identical, and were chosen so as to be ambiguous between average values for male and female speakers reported by Peterson and Barney (1952). Table 1 lists the first three formants for each of the seven items in the continua. Bandwidths of the first three formants were 110, 75, 110. The fourth and fifth formants were identical for all of the tokens. Their values, respectively, were 3500 Hz and 4200 Hz, both with bandwidth of 300 Hz.
Table 1
Formant values of the test tokens used in Experiment 1

<table>
<thead>
<tr>
<th>Token #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>474</td>
<td>491</td>
<td>509</td>
<td>526</td>
<td>543</td>
<td>561</td>
<td>578</td>
</tr>
<tr>
<td>F2</td>
<td>1111</td>
<td>1124</td>
<td>1137</td>
<td>1150</td>
<td>1163</td>
<td>1176</td>
<td>1189</td>
</tr>
<tr>
<td>F3</td>
<td>2416</td>
<td>2424</td>
<td>2432</td>
<td>2440</td>
<td>2448</td>
<td>2456</td>
<td>2464</td>
</tr>
</tbody>
</table>

The voice source of the synthesizer was given a 28 dB per 3000 Hz tilt to improve the naturalness of tokens synthesized with high F0. Total duration of the [hVd] syllables was 325 ms. The [h] was 50 ms long. The final transition was 30 ms and there was a closure interval of 50 ms at the end of the syllable. There was no release spike or burst after the closure interval.

The three continua differed in terms of the fundamental frequency of the tokens. The low F0 series had a steady-state pitch of 100 Hz through the first 120 ms of the voiced portion of the token and rose linearly for 100 ms to 120 Hz where it remained during the [d] closure. A mid F0 series had the same time function but the F0 values were 200 Hz and 240 Hz. A high F0 series likewise was identical except that the F0 values were 300 Hz and 360 Hz.

For the intonational phrase condition a carrier phrase ('Is this ___ ') was synthesized for each of the three F0 continua. Figure 5 shows pitch traces of three different natural productions (speaker KJ) of the frame question 'Is this HUD?' with nuclear stress on 'this'. These contours were used as models for the intonational context stimuli. The intonation contours can be transcribed using the intonational transcription system of Pierrehumbert (1980) as H%L*HH% for the lowest F0 level of the test word, and L*HH% at two different pitch ranges for the two highest F0 levels of the
test word. In this transcription system a starred tone indicates a pitch accent (in this case the accent for the nuclear stress on 'this'), the percent symbol indicates that the tone is a 'boundary tone' and the unmarked H or L is used to indicate a 'phrase accent' - a tone which occurs after a nuclear accent and fills up the time until the phrase boundary. So the utterances in Figure 5 (and the synthesized phrases that they serve as models for) are in two cases composed of a low nuclear accent followed by a high phrase accent and a high boundary tone, and in the low F0 case is composed of a high boundary tone followed by a low nuclear accent then a high phrase accent and a high boundary tone. The difference between the mid and high F0 phrases is not the intonational transcription but rather the pitch range of the utterance. In the high F0 case the utterance spans a greater range of F0 than in the mid F0 case, as if the speaker is speaking up.

Note that the last word (the abbreviation of Housing and Urban Development) has about the same pitch contour at three different values of F0. These intonation contours were chosen because they provide a natural

\[ \text{Is this HUD?} \]

**Figure 5**

Pitch traces for three natural productions of the phrase, 'Is this HUD?'
context in which the F0 variation of the test tokens could be interpreted as intonational differences among tokens rather than as speaker differences. Also, the fact that the same basic F0 contour occurs on the test words at each of the different F0 levels allows for this aspect of the tokens to be left constant across the tokens in the experiment. The intonational patterns of the synthesized carrier phrases are presented in figure 6. The main difference between the F0 traces in figure 5 and figure 6 is that the pitch range spanned by the utterances in the natural productions is not as great as that covered by the synthesized tokens.

Each context was synthesized with two criteria in mind: its naturalness and its comparability to the other contexts. All of the precursor phrases are identical in terms of vowel formants and segment durations; only the F0

![Figure 6](image)

**Figure 6**

F0 contours of the synthesized intonational contexts in experiment 1
pattern is different. As with the vowel formants of the test tokens the
formant values for \([\varepsilon]\) in the carrier phrase were chosen so as to be
ambiguous between values typical of males and females \((F1 = 456, F2 =
1740, F3 = 2796)\). The source characteristics of the phrase were identical to
those of the test tokens so that when the tokens were inserted into the
phrases no perceived change of speaker occurred. The number of
observations in each combination of variables was 12. The total number of
observations per subject was 504 (7 tokens \(*\) 3 F0 levels \(*\) 2 contexts \(*\) 12
repetitions).

Procedure.

The experimental sessions were conducted by the ERS experiment
running program (Johnson, 1985) on a New England Digital Able 60
computer. Subjects were seated in an anechoic chamber and the tokens
were played through a 4.5 kHz low pass filter and Sennheisser HD 420
headphones at a comfortable listening level. Subjects indicated their
response by pressing buttons labeled 'hood' and 'HUD'. The test items were
randomized separately for each subject. Vowel identification and reaction
time data were collected by the computer and subjects were given reaction
time feedback via a computer for each item presented. Only the
identification data were analyzed.

Subjects were instructed to identify the test tokens as either 'hood' or
'HUD' and to disregard the pitch differences among the tokens. They were
told to expect to hear both words at all three pitch levels.

Both of the independent variables in this experiment (fundamental
frequency of test items, and presence vs. absence of a context phrase) were
treated as within subjects variables. Thus, all subjects responded to all
combinations of the variables. The context variable was presented in blocks (all of the items in one level of the condition were presented and then all of the items in the other level of the condition were presented). The order of blocks was counterbalanced among subjects. The dependent variable was the number of ‘hood’ responses.

Results.

Figure 7 presents the results for this experiment (there were no noticeable counter-trends among subgroups of subjects to warrant an analysis of individual subjects separately). The ordinate of the figure is the

![Figure 7](image)

Results of experiment 1 showing F0 level and context condition.
percent ‘hood’ responses. The abscissa classifies the responses into F0 level and context. A 2X3 repeated measures ANOVA was performed on the number of ‘hood’ responses in each combination of conditions for each subject. The figure presents these data as a percentage of ‘hood’ responses for ease of interpretation. As is clear from the figure there was a strong effect of F0 normalization (F(2,18) = 115.77, p < .001). Average percentage of ‘hood’ responses for high, mid and low F0 conditions were respectively 83.4%, 62.7% and 15.5%. Also, as indicated by the figure there was no effect for intonational context (F(1,9) = 0.12). The average percent ‘hood’ responses for items presented in context was 54.5% while the average percent ‘hood’ responses for items presented in isolation was 53.3%. The interaction between F0 level and context was also not significant (F(2,18) = 0.37). At the high F0 level the averages for items in context and in isolation were 82.6% and 84.2%. At the mid F0 level these values were 63.8% and 61.7%, while at low F0 the average response to items in context was 17% while in isolation 13.9% of these items were classified as ‘hood’.

Discussion.

It was hypothesized that the perceptual process of F0 normalization is made up of two component processes, an auditory level transformation and a higher level speaker-based process which is triggered when hearers interpret items as having been produced by different speakers. Further, it was hypothesized that presenting items in an appropriate intonational context would reduce the overall shift in vowel identification by reducing hearers’ tendency to interpret items with different F0 as having been produced by different speakers. The data indicate that there was no
The difference between presenting items in context versus in isolation. Rather, the shift in perceived vowel identity which was found when items were presented in isolation was also found when these same items were presented in an intonational context. One possible explanation for this finding is that there is no speaker-based component in the F0 normalization process. Another possibility is that the design of the tokens used in this experiment prevented any separation of auditory and speaker-based F0 normalization processes.

There are two ways that the design of the tokens could have been inappropriate. First, the continuum was designed with a small change in F1 from the 'hood' token to the 'HUD' token in order to enhance the shift in identification attributable to an F0 normalization effect. In figure 8 average percent 'hood' responses for each token are presented as a function of the F1/F0 ratio. Because of the large range of F0 values (200 Hz) and the small range of F1 values (104 Hz) used in this experiment the range of the F1/F0 ratio is quite large. If the auditory F0 normalization process depends on the F1/F0 ratio (as Traunmüller’s hypothesis suggests) then it may be the case that the range of F1/F0 ratios spanned by the tokens in the experiment was so large that even if subjects heard the intonational phrase items as having been produced by the same speaker the perceived F1 for the high F0 series was always low enough to be identified as 'hood' and the perceived F1 of the low series was always high enough to require the 'HUD' response.

Nearey (to appear) has suggested that the relationship between F0 and perceived F1 in isolated syllable identification is such that perceived F1 changes 10-20% for a 100% change in F0. In this experiment F0 was increased 200%, thus we might expect an increase in perceived F1 of 20-
40% or 95 to 190 Hz. Since the F1 of the tokens only increased 104 Hz from token 1 to token 7 and since the F0 change across continua was 200 Hz it would not be unreasonable to suppose that the elimination of the speaker-based normalization process in the context condition (by making the utterances appear to have been spoken by the same speaker) might not be noticed due to a large shift in identification caused by auditory F0 normalization based on F1/F0 ratio.

One issue that is raised by this finding has to do with the production of vowels in natural contexts. If F1 and F0 are auditorily linked as is suggested by this and other experiments having to do with F0 normalization to what
extent is it necessary to adjust F1 as F0 changes? There is some evidence that speakers do make adjustments of F1 at different levels of F0 and that perhaps some sort of 'harmonic efficiency criterion' is used by speakers (see Bladon, et al., 1984). Before reaching any conclusions as to the relevance of the results of this experiment to this question we must consider two facts. First, it is not clear that the hearers in this experiment actually identified the intonational context items as having been produced by the same speaker and in fact the results of experiment 2 seem to suggest that they did not make such an identification. Second, the range of F1 variation in these tokens was small. The difference between the average F1 of [ə] and the average F1 of [ɪ] for men in Peterson and Barney (1952) was 200 Hz and the difference for women was 190 Hz. In the tokens used in this experiment this difference was only 104 Hz.

The second problem in the design of the tokens has to do with the intonational contexts. In natural speech an increase in pitch range due to speaking more emphatically is normally accompanied by durational changes, amplitude changes, and even glottal wave shape changes as well. The different intonational contexts which were used in this experiment were not accompanied by any durational or other segmental level adjustments. This may have lead hearers to interpret the different pitch ranges in intonational context as having been produced by different speakers. If this was the case then the hypothesis was not tested.

One trend in the data seems to indicate the existence of a speaker-based normalization process. Figure 9 presents the average percent 'hood' responses for the different F0 levels and context factors as functions of the tokens in the 'hood' - 'HUD' continuum. Notice that the items with high F0
are classified as 'hood' about equally often for each token except token number 7. In this case the token presented in context was classified as 'hood' less often than the token presented in isolation. A similar trend can be seen for the low F0 items in the 'hood' end of the continuum (tokens 1-3) where the F0 normalization effect for items in context is somewhat less than it is for those same items presented in isolation. As was mentioned in the results section this interaction was not significant, so we are not here talking about statistically reliable results of the experiment, but the fact that the differences between the two conditions (context and isolation) occur where they do raises the possibility that there is a context effect to be found but that the design of this experiment obscured it. To test this possibility
another experiment was conducted in which the F0 range of the tokens was reduced in an attempt to reduce the floor and ceiling effects found in these tokens. The intonational contexts were also changed to be more easily interpretable as having been produced by the same speaker.

Experiment 2.

The design of experiment 2 is exactly that of experiment 1 with two changes. First, the F0 levels of the test tokens were changed so as to be less extreme than those used in experiment 1. Second, the intonational contexts were given F0 contours which were more readily interpretable as having been produced by the same speaker by reducing the differences between the F0 contexts.

Method.

Subjects.

10 subjects (3 males and 7 females) participated in this experiment. They were all native speakers of English enrolled at Ohio State University. None reported any history of speech or hearing difficulties. They were recruited through an advertisement in the school paper and in introductory Linguistics classes and were paid $3.50 for their participation. Data from two subjects were eliminated from analysis because the subjects reported (in post-experiment interviews) using F0 level explicitly as a cue for vowel identification judgements in both the isolation condition and the context condition.
Materials.

Three different ‘hood’ - ‘HUD’ continua were synthesized. They had the same formant values, formant bandwidths and durational properties as those used in experiment 1. The high F0 series had F0 of 250 Hz for 170 ms. followed by a rise to 300 Hz over the last 100 ms. of the token. The mid F0 series had F0 of 200 Hz followed by a rise to 240 Hz. The low F0 series had F0 of 150 Hz and a rise to 180 Hz. The intonational contexts into which these tokens were placed for the ‘context’ condition had the same segmental properties as the phrases used in experiment 1. They all had the intonational transcription L* H H% with the L* on ‘this’ in the phrase, ‘Is this ____?’ The F0 contours of the phrases are displayed in figure 10.

![Figure 10](image-url)

**Figure 10**

F0 contours of the synthesized intonational contexts in experiment 2.
Procedure.

The procedure of experiment 2 was almost identical to that of experiment 1. In the 'context' condition subjects were told that they would be hearing different instances of the same person asking the question 'Is this ___?' In this way it was hoped that subjects would be biased to hear all of the items as being produced by the same speaker. As before, subjects were instructed to disregard pitch variation in the tokens and to simply base their responses on vowel quality.

Results.

The results of experiment 2 are presented in figure 11. Once again there was a very strong F0 normalization effect ($F(2,14) = 303.6, p < .0001$).

![Figure 11](image_url)

Figure 11
Results of experiment 2 showing F0 level and context condition
87.9% of the high F0 items were classified as 'hood', while 53.9% of the mid F0 items and only 11.9% of the low F0 items were identified as 'hood'. In this experiment as opposed to experiment 1 there was also an interaction between F0 and context (F(2,14) = 8.54, p < .01). In a post-hoc Newman-Keuls ranked comparison test it was found that the means of the context and isolation conditions were significantly different only in the low F0 condition. The low F0 items presented in context were identified as 'hood' 15% of the time while only 7% of the low F0 items presented in isolation were identified as 'hood'. The context main effect was not significant (F(1,7) = 0.65, n.s.).

When the data are plotted by token (figure 12) it is clear that the interaction of F0 with the context effect was the result of a shift in the identification response to the low F0 continuum in the direction predicted if speaker-based normalization is prevented from occurring in context but not in isolation. Presented in isolation tokens 1 and 2 of the low F0 continuum were perceived as 'HUD' while in the intonational context there was a greater tendency to hear them as 'hood'. Interestingly, the high F0 continuum was not subject to such a shift in perception.

Discussion

The results of this experiment are consistent with the hypothesis of the existence of both auditory F0 normalization and speaker-based F0 normalization. Although the shift in identification from mid F0 to high F0 was just as large in the intonational context as it was when the items were presented in isolation the shift in identification function for the low versus
mid F0 series was reduced when the low F0 tokens were given an intonational context comparable with that of the mid F0 tokens.

In order to explain this difference in response note in figure 10 that the mid F0 and low F0 context conditions had identical F0 patterns on the first word of the context while the high F0 items had a somewhat higher F0 on this word of the phrase. The fact that the high F0 intonational context contour had a minimum different from that for the low and mid contours may have been enough to cause subjects to interpret the high contexts as having been produced by a different speaker. This follows from the fact that the low end of a speaker's pitch range is more stable than is the high end of the range, and thus the lowest portion of a phrase's F0 contour could
serve as an indication of speaker identity - especially when, as in these tokens, segmental durations do not differ from one instance of the phrase to another. So the lack of any change in identification responses when the high F0 tokens were presented in an intonational context (as compared with the identification of these tokens in isolation) seems to result from the application of a speaker-based normalization process in both types of presentation.

In figure 13 the data of this experiment are presented as functions of F1/F0 ratio. From this display of the data it seems to be possible to find a single crossover point between 'hood' and 'HUD' when these items are presented in isolation. However, when items are presented in context the relationship between F1/F0 ratio and vowel categories is not so simple. In
view of the two component processes hypothesis this result seems to indicate that the auditory explanation of F0 normalization cannot simply rely on the F1/F0 ratio (as suggested by Syrdal and Gopal, 1986, and Sussman, 1986).

In this experiment I have demonstrated the existence of the speaker-based normalization process by creating a situation in which it fails to apply. It should also be possible to conduct an experiment in which the application of the speaker-based process can be observed. One design which might achieve this goal makes use of two different intonational patterns at different F0 levels. The intonational contexts for this experiment would be (1) falling intonation as in a statement (transcribed H*LL%) and (2) rising intonation as in a question (L*HH%). The different pitch ranges in these items would indicate different speakers because: (1) durational cues would not change as they would under different degrees of emphasis, (2) low points in the phrases would be as different as the high points are. The tokens would be designed so that the test word would be at the same F0 at the end of a high falling context and at the end of a low rising context. Since the test words would have the same F0 level there should be no shift in identification due to auditory F0 normalization, but since they would be identified as having been produced by different speakers the speaker-based normalization process should apply. In this way the existence of speaker-based F0 normalization could be demonstrated by creating a situation in which its application can be separated from the nonapplication of the auditory level effect.
Conclusion

The experiments reported in this chapter seem to indicate that the perceptual process of F0 normalization is composed of two separate perceptual processes. At a cognitive level of processing hearers seem to use F0 as an indication of the length of the speaker’s vocal tract and adjust their expectations for the range to be covered by the formants. At an auditory level of processing the lowest, high amplitude harmonics seem to be integrated with the harmonics around F1 such that increasing F0 results in a lower perceived F1.

The significance of the work here is that it demonstrates that both levels of processing must be included in a full description of F0 normalization. I argued in the first chapter that speech perception research must be carried out in the context of an appreciation of both the auditory and cognitive levels of processing. These experiments are striking because they demonstrate the workings of both levels in a single perceptual phenomenon. Where it has commonly been assumed that the shift in identification found in F0 normalization experiments is the result of a single perceptual process (either cognitive or auditory) I have given evidence to indicate that there are actually two processes involved.
Chapter III
Higher Formant Normalization.

While $F_0$ is correlated with $F_1$ and $F_2$ values in a probabilistic sense, the higher formants are correlated with $F_1$ and $F_2$ directly because like the lower formants the higher formants are properties of the vocal tract, rather than of the larynx. So, while it is possible to find people with short vocal tracts who speak with low $F_0$ or conversely people with long vocal tracts who speak with a generally high $F_0$ (Julia Child for instance) it is not possible to find a speaker for whom $F_1$ and $F_2$ are produced by one vocal tract configuration and $F_3$ and $F_4$ by another. This is not to say that the relationship between men's and women's formants can be expressed as a simple scale factor. For instance, if a woman's vocal tract were 90% as long as a particular man's vocal tract we could not simply multiply the man's formant values by 0.9 to calculate the woman's formants for the same vowel sounds. Rather, the relationship between formant values for men and women is complex partly as a result of non-uniform differences in vocal tract size and differing articulatory strategies.

Fant (1966) cites x-ray data which indicate that the ratio of mouth size to pharynx size is different for men and women. He reports that a typical male has a mouth 8.25 cm long (from incisors to back pharynx wall) and a pharynx 9.1 cm long (from soft palate to glottis) while the typical female has a mouth which is 7 cm long and a pharynx which is also 7 cm
long. Since the formant frequencies in speech exhibit cavity affiliations (see Fant, 1960) the fact that the ratio of mouth to pharynx cavities is different for men and women means that the relationship between men’s formants and women’s formants will be complex. Fant (1966, 1975) also notes that the differences between men and women for formants which are not very cavity dependent (F1 of [i] and F1-F3 of round back vowels) are less than might be expected from calculations based on overall vocal tract length. Fant (1966, 1975) suggests that this reduction of the male-female difference may be the result of articulatory strategies by which speakers attempt to reduce the absolute differences in formant values across speakers. Thus, F1 for [i] as produced by a female speaker may be as low as or lower than a typical F1 value for male speakers even though the female speaker’s vocal tract is shorter than the typical male vocal tract. This phenomenon does not seem to be the result of any anatomical features of male and female vocal tracts or cavity affiliation of F1 in [i], but rather seems to be the result of an articulatory strategy in vowel production. As such it may be seen as evidence for a view of vowel production in which the vowel sound is the target and not the vowel gesture.

Bladon et al. (1984) assert that there may be sociological conditions which have to do with the acoustic differences between men’s and women’s vowels. They note that the average ratio of male to female formants is different in different languages. They suggest that this is evidence for some degree of sociolinguistic conditioning of the relative differences (or lack of differences) between the sexes in vowel sound realization. These data simply illustrate the complexity of the problem of relating vowel formant values to vowel identities.
The higher formants are less cavity dependent than the lower formants. For instance, in American English F3 is only lowered considerably in the case of [r] where lip rounding, pharyngeal constriction and retroflex (or bunched alveolar) constriction conspire to lower the formant. Thus, the higher formants could provide hearers with a somewhat stable indication of vocal tract size. If there is a perceptual process of formant normalization which rescales vowel formant values based on some estimate of vocal tract size we would expect this mechanism to make use of the information available in the higher formants. There is some experimental evidence which suggests that the correlation between higher formants and vocal tract size can influence hearers' judgments of vowel quality.

Fujisaki and Kawashima (1968), in addition to their studies of the role of F0 in vowel perception also investigated the role of higher formants. They found that changes in formants 3 through 5 can shift vowel identification boundaries. This effect was greatest in an [u]-[e] continuum and not so large in a continuum from [o] to [a]. They also found that by using a nonperiodic noise source (with which the higher formants had greater amplitude) the effect of shifting the higher formants could be increased. The largest shift of vowel identification boundaries occurred when both higher formants and F0 were manipulated.

However, there is contradictory evidence. Slawson (1968), using a scaling method of testing for vowel perception change, found that variation in higher formants alone did not result in a shift of vowel perception. He concluded that although higher formants correlate well with vocal tract length, they don’t result in large shifts in vowel quality. Karnickaya et al.
(1975) agree. They found that shifting the F3 and F4 in an [o]-[a] continuum or an [e]-[a] continuum did not result in any boundary shifts.

A possible explanation of this divergence of experimental findings is suggested by Carlson et al.'s (1975) results. They found that when F3 is within 3 Bark of F2 the perceived F2 (F2') is influenced by F3. (Perceived F2 is measured by having subjects match a two formant stimulus to a five formant stimulus.) It may be that in the studies of Slawson (1968) and Karnickaya et al. (1975) the distance between F2 and F3 was enough to avoid any influence of F3 on F2, while in the Fujisaki and Kawashima (1968) study the F3 and F2 of the [e] tokens were close enough to result in auditory integration to a single 'centre of gravity'.

Thus, the data reported in the literature seem to indicate that, rather than being used to estimate vocal tract length in a speaker-based, cognitive level normalization process, it may be that the role of higher formants in vowel perception is purely auditory in nature having to do with the integration of spectral information in the speech signal.

If we assume a frequency integration function similar to that proposed by Chistovich et al. (1979) we can predict that higher formant normalization will occur only when the distance between F2 and F3 is less than 3 to 3.5 Bark.

Experiment 3.

This experiment is simply an attempt to test this hypothesis. Two continua were used - [hɔd]-[hʌd], and [hɔd]-[hɔd]. In the [o]-[ʌ] continuum F3-F2 is greater than 3.5 Bark for both men and women in Peterson and
Barney (1952), while in the [ɪ]-[ɛ] continuum F3-F2 is less than 3 Bark for both men and women.

The auditorily based explanation of higher formant normalization predicts that only the front series will have a normalization effect (this due to the shift of F2' (perceived F2) as a function of the frequencies of F2 and F3 when F3 minus F2 < 3.5 bark). If higher formants are used to estimate speakers' vocal tract length in a speaker-based normalization process we expect to find that formant normalization will not be limited to the front series but will be present in both.

Method.

Subjects.

Ten (2 males and 8 females) native English speaking undergraduate students at Ohio State University served as subjects in this experiment. None of the subjects reported any history of speech or hearing disorders. They were recruited through the university paper and through introductory linguistics classes, and received $3.50 for a single one hour session.

Materials.

Four [hVd] continua were synthesized for this experiment. Two of these continua were front vowel continua from 'hid' to 'head' and two were back vowel continua similar to those used in the experiments of Chapter 2 from 'hood' to 'HUD' (the abbreviation of Housing and Urban Development). One front vowel continuum and one back vowel continuum were synthesized with relatively high higher formants while the other front and the other back series had relatively low higher formants (for ease of reference the higher formants in these five formant tokens will be referred
As with the [ə]-[A] continuum of experiment 1 the values chosen for F1 and F2 were ambiguous between the values for men and women in Peterson and Barney's (1952) data.

The formants and bandwidths for all four vowel continua are presented in table 2. The /h/ portion of the tokens was 50 ms. in duration. Each token had a vocalic portion of 220 ms., which included a 30 ms. transition to /d/. The tokens had a total duration of 270 ms. Alveolar transitions for the 'hid'-'head' continuum ended at 300, 1600 and 2600 respectively for the first, second and third formants. Transitions for the

<table>
<thead>
<tr>
<th>Token #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>BW</th>
</tr>
</thead>
<tbody>
<tr>
<td>'hid'-'head'</td>
<td>F1</td>
<td>432</td>
<td>448</td>
<td>464</td>
<td>480</td>
<td>496</td>
<td>512</td>
<td>528</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>2208</td>
<td>2192</td>
<td>2176</td>
<td>2160</td>
<td>2144</td>
<td>2128</td>
<td>2112</td>
</tr>
<tr>
<td>High</td>
<td>F3</td>
<td>2845</td>
<td>same</td>
<td>---</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td>3700</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F5</td>
<td>4550</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>F3</td>
<td>2701</td>
<td>same</td>
<td>---</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td>3300</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F5</td>
<td>3850</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>'hood'-'HUD'</td>
<td>F1</td>
<td>474</td>
<td>491</td>
<td>509</td>
<td>526</td>
<td>543</td>
<td>561</td>
<td>578</td>
</tr>
<tr>
<td></td>
<td>F2</td>
<td>1111</td>
<td>1124</td>
<td>1137</td>
<td>1150</td>
<td>1163</td>
<td>1176</td>
<td>1189</td>
</tr>
<tr>
<td>High</td>
<td>F3</td>
<td>2512</td>
<td>same</td>
<td>---</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td>3700</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F5</td>
<td>4550</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low</td>
<td>F3</td>
<td>2368</td>
<td>same</td>
<td>---</td>
<td>110</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F4</td>
<td>3300</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>F5</td>
<td>3850</td>
<td>same</td>
<td>---</td>
<td>300</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
'hood'-'HUD' series went to 300 (F1), 1600 (F2) and dipped to 2150 then rose to 2570 (F3). All tokens had F0 which was steadystate at 200 Hz for 115 ms. and then fell linearly over 100 ms. to 160 Hz. Subjects responded to fifteen test trials for each token. The total number of observations per subject was 420(7 tokens * 2 higher formant levels * 2 continua * 15 repetitions).

Procedure.

Experimental sessions were conducted by the ERS experiment running program (Johnson, 1985) on a New England Digital Able 60 computer. Subjects were seated in an anechoic chamber and the tokens were played through a 4.5 kHz low pass filter and Sennheiser HD-420 headphones at a comfortable listening level. Subjects indicated their response by pressing buttons on a response box labeled 'hood' and 'HUD' or 'hid' and 'head'. The test items were randomized separately for each subject. Vowel identification response and reaction time data were collected by the computer and subjects were given reaction time feedback via a computer screen for each item presented. Only the identification data were analyzed.

Subjects were instructed to identify the test tokens as either 'hood' or 'HUD', or as either 'hid' or 'head' and to disregard all other differences between the tokens.

The independent variables in the experiment are: (1) the level of the higher formants and (2) front vs. back vowel continuum. Both were treated as within subjects variables. The vowel continuum variable was blocked and the order of presentation was counterbalanced among subjects (half of the subjects heard the front vowel block first while half heard the back vowel
block first). There were 15 practice trials for both vowel series. The dependent variable was the number of 'hood' or 'hid' responses.

Results.

The number of 'hood' or 'hid' responses to each continuum by 10 subjects was analyzed in a 2X2 repeated measures ANOVA. The results for experiment 3 are presented in figure 14.

As the figure indicates there was an interaction between the continuum condition ([ʊ]-[ʌ] vs. [ɛ]-[ɛ]) and the higher formants condition (low vs. high F3-F5) ($F(1,9) = 9.67, p < .02$). For the 'hood'-'HUD' continuum the percent 'hood' responses for low F3-F5 was almost identical to the percent 'hood' responses for high F3-F5 (52% vs. 51.9%). On the other
hand, for low F3-F5 in the 'hid'-'head' continuum there were 47% 'hid'
responses, while when F3-F5 were high only 38% of the tokens were
identified as 'hid'. There was a significant main effect for the higher
formants condition (F(1,9) = 9.68, p < .02). Vowel identification as 'hood'
or 'hid' was 50% with low F3-F5, while it was 45% with high F3-F5. There
was no main effect for vowel continuum (F(1,9) = 4, p = .077).

The difference between the back vowel continuum and the front
vowel continuum is illustrated further in figures 15 and 16. Figure 15 shows
the percent 'hood' responses for each of the tokens in both the high F3-F5
and the low F3-F5 continua. Category boundaries in the 'hood'-'HUD'
continuum for each subject were computed by linear interpolation. There
was no difference between the high F3-F5 condition and the low F3-F5
condition in a correlated observations t-test (t = 0.456, ns). Figure 16 shows
the percent 'hid' responses for each of the tokens in both higher formant
conditions. As is clearly indicated in the figures, there was no effect of
higher formant shift in the back vowel continuum while there was an
increased tendency toward 'hid' responses throughout the 'hid'-'head'
continuum in the high F3-F5 condition.

Discussion.

The results of this experiment seem to support the hypothesis that
higher formant normalization is auditory in nature and depends upon the
proximity of F2 and F3 in vowel spectra. The main effect for higher formants
noted above was entirely the result of the strong difference between high
and low F3-F5 in the 'hid'-'head' continuum. These data give no reason to
Figure 15.
Results of experiment 3. 'hood'-'HUD' continuum by token.

Figure 16.
Results of experiment 3. 'hid'-'head' continuum by token.
suspect that there is a speaker-based process utilizing higher formants in the perception of F2, and every reason to suspect that the influence of higher formants has to do with the role of F3 in the perception of F2 when F2 and F3 are within 3 Bark of each other.

However, two additional considerations motivate a replication of this experiment with some modifications.

First, Fujisaki and Kawashima (1968) found that amplitude of F3 played a role in the higher formant normalization effect. Nearey (to appear) has suggested that back vowels may exhibit less higher formant normalization than front vowels because the amplitude of the higher formants is lower in vowels with low F2. This provides an alternative interpretation of the results of experiment 3. It is possible that the shift of F3-F5 did not influence the perception of the 'hood'-'HUD' continuum because the higher formants were too low in amplitude. Therefore, a replication of the experiment is needed in order to test this interpretation of the results.

Second, the auditory hypothesis which is supposed to account for the shift in perception of the 'hid'-'head' continuum requires only a shift in F3 in order to shift perceived F2. In this experiment F3, F4 and F5 were all shifted. A more convincing test of the hypothesis would involve shifting only F3.

Experiment 4.

As with experiment 3 the hypothesis being tested is that the higher formant normalization effect reported by Fujisaki and Kawashima (1968) is the result of an auditory frequency integration process by which perceived
F2 is influenced by F3 when F2 and F3 are within 3 Bark of each other.

Experiment 4 is a more stringent test of this hypothesis in two ways. First, the amplitude of F3 in the 'hood'-'HUD' continuum has been increased so as to make F3 audible. Thus, the possibility that the lack of a higher formant effect in the back vowel continuum in experiment 3 was due to low F3 amplitude is addressed. Second, only F3 is shifted in this experiment. This provides a more direct test of the F3-F2 auditory integration hypothesis.

Method.

Subjects.
Thirteen (4 males and 9 females) native English speaking undergraduate students at Ohio State University served as subjects in this experiment. None of the subjects reported any history of speech or hearing disorders. They were recruited through introductory linguistics classes and received $3.50 for a single one hour session. As explained below data from nine subjects (2 males and 7 females) were analysed.

Materials.
As in experiment 3 four different [hVd] continua were synthesized. Two continua were from 'hid' to 'head' and two were from 'hood' to 'HUD'.

The 'hood'-'HUD' continua had the same formant values for F1-F3 that were used in experiment 3. F4 and F5 were changed to 3500 and 4200 Hz respectively in all of the 'hood'-'HUD' tokens. The two 'hood'-'HUD' continua were different from each other only by virtue of the fact that F3 was 2512 in the high F3 continuum and was 2368 in the low F3 continuum.

A change in the amplitude of F3 in these tokens as compared with those used in experiment 3 was effected by narrowing the bandwidth of F3
from 110 Hz (as it was in experiment 3) to 30 Hz (the value in the current experiment). Figure 17 illustrates the effect of narrowing the bandwidth of F3 for a vowel sound with formants at 474, 1111, 2368, 3500 and 4200 Hz. The change in amplitude of F3 from wide bandwidth to narrow bandwidth is 4 dB. Flanagan (1957) found that the JND for F2 in [æ] was 3 dB. One other change in the synthesis parameters from those in experiment 3 is that the spectrum of the voice source was given less tilt. Where in experiment 3 it was given a 28 dB per 3500 Hz tilt it was given only 14 dB per 3500 Hz tilt for this experiment.

In order to verify that the F3 in the 'hood'-'HUD' continuum was audible, a preliminary listening test was conducted using two experienced listeners (KJ and MW). The task was an AX discrimination task in which items from the high F3 and low F3 series were compared. F1 and F2 in the pairs were identical while in half of the pairs F3 was different. A' scores for the two listeners were 0.61 and 0.80 respectively, with 0.5 indicating chance performance and 1.0 indicating perfect performance. These scores indicate that the change from high to low F3 in the 'hood'-'HUD' continuum was audible.

The 'hid'-'head' continuum for experiment 4 was quite different from the one in experiment 3. As with the 'hood'-'HUD' continuum the higher formant modification in this experiment involved F3 only. F4 and F5 were kept constant over all conditions. Vowel duration for the tokens was reduced from 220 ms. to 200 ms.
Figure 17
Two spectra of token 1 in the 'hood'-HUD' continua which illustrate the result of narrowing the bandwidth of F3. The amplitude of F3 is approximately 4 dB greater with F3 bandwidth of 30 Hz (a) as compared with F3 bandwidth of 110 Hz (b).
In order to test the hypothesis that F3 and F2 are auditorily coupled in front vowels it was decided to synthesize the 'hid'-'head' continuum with the same F1 for all tokens. In an attempt to increase the naturalness of these syllables I chose to use an F1 which had a slight rise over the last 45% of the vowel (see Lehiste and Peterson, 1961a concerning offglides in lax vowels). The F1 for all tokens was steady at 520 Hz for the first 110 ms and then rose to 600 Hz over the next 70 ms of the vowel and then fell to 300 Hz over the last 30 ms of the vowel in the [d] transition.

The range of F2 values in the continuum was larger and lower than that used in experiment 3. The values are listed in Table 3. Because the natural cue for openness in vowels is the value of F1 (or the F1/F0 ratio, Traunmüller, 1981) these tokens were quite ambiguous.

Table 3.
F2 values for the 'hid'-'head' continuum in experiment 4.

<table>
<thead>
<tr>
<th>Token #</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>2160</td>
<td>2118</td>
<td>2077</td>
<td>2035</td>
<td>1993</td>
<td>1952</td>
<td>1910</td>
</tr>
</tbody>
</table>

Procedure.

The procedure was identical to that of experiment 3 with the following exception. Of the first eight subjects in the experiment, four identified the 'head' end of the 'hid'-'head' continuum as 'hid' and the 'hid' end as 'head'. The ambiguity of this continuum made it necessary to change the procedure slightly. A demonstration tape was constructed in which the endpoints of both the high and low F3 continua were identified and compared with each other. This tape was played to the subjects before the
'hid'-'head' practice items. This procedure is justified because it does not bear on the hypothesis being tested in this experiment. The experiment is not designed to discover 'cues' for [r] and [s], rather it was designed to assess the role of F3 in the perception of an F2 continuum. Those subjects who identified the continuum 'backwards' seem to have based their identifications on vowel contrast rather than on the F2 of the tokens. For each presentation they seem to have been concerned with whether the token was different from the preceding one. If they noticed any difference they identified it as belonging in the opposite vowel category.

Figure 18 shows identification functions for the four subjects who misidentified the continuum. Comparing this figure to figure 21 it is clear that these subjects perceived the continuum in a more categorical fashion.
than did the subjects who identified the continuum in a way which is consistent with the F2 information in the tokens. This is evidence that these subjects adopted a strategy which did not include using F2 as an indication of vowel quality. Since the hypothesis being tested in this experiment was that perceived vowel quality would be influenced by a shift of F3 and since these subjects don’t seem to be basing their judgments of vowel identity on F2 their identification responses do not constitute a test of the hypothesis. Thus, the data from the four subjects who mislabelled the continuum were not included in the results reported below.

The use of a demonstration tape simply biased subjects to identify the continuum in such a way that their responses could be interpreted. Clearly, this continuum (and the 'hood'-'HUD' continuum with its higher amplitude F3) is not made up of typical naturally produced syllables. The limits on naturalness introduced by the synthesis parameters chosen were necessary in order to provide a test of the hypothesis. Note, however, that these syllables were clearly speech and were natural enough to trigger any automatic speech specific perceptual processes (Liberman and Mattingly, 1985).

Results.

The responses of nine subjects were analysed in a 2X2 repeated measures ANOVA. The dependent measure was the number of 'hood' or 'hid' responses in each combination of conditions and the independent variables were vowel continuum ('hid'-'head' vs. 'hood'-'HUD') and F3 level (high vs. low).
The results are presented graphically in figure 19. As in experiment 3 there was a significant interaction between the continuum condition and the F3 condition \((F(1,8) = 8.57, p < .02)\). In the high F3 condition the percent 'hood' responses to the 'hood'-'HUD' continuum was 53% while in the low F3 condition it was 55%. The percent 'hid' responses to the 'hid'-'head' continuum was 46% with high F3 and 39% with low F3. There was a marginally significant main effect for continuum \((F(1,8) = 5.93, p < .05)\) which reflects a tendency for the 'hid'-'head' boundary to be closer to the 'hid' end of the continuum than the 'hood'-'HUD' boundary was to the 'hood' end of that continuum.
Figure 20
Results of experiment 4.
Identification functions for the 'hood'-'HUD' continua.

Figure 21
Results of experiment 4.
Identification functions for the 'hid'-'head' continua.
Figures 20 and 21 illustrate further the differences between the front and back vowel continua. As in experiment 3 there was a tendency for the front vowel continua to shift throughout the series in the different F3 conditions. These figures also reflect the relative ambiguity of the front vowel continuum. The 'hood'-'HUD' series was perceived in a much more categorical fashion.

Although the overall number of 'hood' responses to the 'hood'-'HUD' continuum was not different under the two levels of the F3 condition there appears to have been a slight shift in the category boundary. Category boundaries were computed by linear interpolation for each subject in the two F3 conditions for the 'hood'-'HUD' continuum. For this analysis the data from the four subjects who had misidentified the 'hid'-'head' continuum were also included. The observed boundary shift in the 'hood'-'HUD' continuum was significant in a correlated observations t-test (t = 2.345, p < .05).

Discussion.

These results seem to parallel those of experiment 3 and offer a verification of the hypothesis that higher formant normalization results from the auditory integration of F2 and F3 when F2 and F3 are within 3 Bark of each other.

There was a shift in the perception of the 'hid'-'head' continuum under the different F3 conditions as was predicted by the auditory hypothesis. And, as predicted there was an interaction of the continuum and F3 level factors and the shift which was observed in the front vowel continuum did not seem to be present in the back vowel continuum.
In a post-hoc test of boundary shift in the ‘hood’-‘HUD’ continuum it was found that there was a boundary shift due to the F3 change. This shift is of a much smaller magnitude than is the shift in the ‘hid’-‘head’ continuum (a 2% change in the number of ‘hood’ responses versus a 7% change in the number of ‘hid’ responses). The absolute magnitude of shift should not be compared, however, because these two continua may span different size perceptual regions.

It is also not clear how the shift in the ‘hood’-‘HUD’ boundary can be attributed to a cognitive speaker-based process involving range normalization. From the range normalization point of view F3 would be used by the hearer as an indication of vocal tract length and thus the hearer’s expectation for vowel formant ranges would be adjusted. If F3 is low then this would be taken as an indication of a long vocal tract and thus of generally lower formants. Thus, the formants in the ‘hood’-‘HUD’ continuum would not have to be as high to be identified as ‘HUD’.

Therefore, the speaker-based process predicts that for low F3 there will be more ‘HUD’ responses and that the boundary for the continuum will be closer to the ‘hood’ end than is the boundary for the high F3 continuum. The boundary shift observed here is in the opposite direction. The average 50% crossover in the low F3 series was 4.544 and in the high F3 series it was 4.198. These values are expressed in terms of token number along the

<table>
<thead>
<tr>
<th></th>
<th>High F3</th>
<th>Low F3</th>
</tr>
</thead>
<tbody>
<tr>
<td>F1</td>
<td>529.4</td>
<td>535.3</td>
</tr>
<tr>
<td>F2</td>
<td>1152.6</td>
<td>1157.1</td>
</tr>
</tbody>
</table>
continuum because the continuum was both an F1 and an F2 continuum. The F1 and F2 locations of the boundaries are given in table 4.

So, neither the centre of gravity hypothesis nor a cognitive range normalization hypothesis can account for the shift in the back vowel continuum. It is necessary to hypothesize that some other perceptual process has resulted in this shift in identification. Perhaps something along the lines of ‘profile analysis’ (Green, et al., 1984) should be considered.

This experiment has resulted in data which are consistent with the auditory explanation of higher formant normalization and has not provided any evidence of a cognitive level speaker-based process which would use higher formants to estimate speakers’ vocal tract size and so guide the interpretation of vowel formants.

Conclusion.

It seems that the ‘higher formant normalization’ effect reported in the literature may be the result of a process in auditory perception and nothing more. In the experiments reported in this chapter it was found that when F3 and F2 were within 3 Bark of each other there was a shift in identification when F3 was shifted. An explanation of the shift in identification in terms of cognitive formant range adjustment was further counterindicated by the small boundary shift in the ‘hood’–‘HUD’ continuum found in experiment 4 which was in the wrong direction to be explained by the hypothetical speaker-based process.

So, unlike the results reported in chapter 2, these results might be interpreted as being in line with a theory of speech perception in which the cognitive level is limited to simple pattern matching after the speech signal.
has undergone auditory processing. Note, however, that it has been reported (Slawson, 1968; Fujisaki and Kawashima, 1968) that in continua for which no higher formant normalization effect was observed there seemed to be an increased perceptual normalization (over the shift in identification observed when F0 alone was varied) when the higher formants covaried with F0. It may be that this type of covariant context is a triggering environment for a cognitive process of speaker-based normalization which utilizes higher formant values in an estimate of the speaker's vocal tract size.
Chapter IV
Conclusion

In this chapter I will summarize the results of the experiments in chapters 2 and 3 and discuss their implications for theories of speech perception and cognition.

Summary.

The experiments reported in chapter 2 were designed to test the hypothesis that the F0 normalization effect reported by Miller (1953), Slawson, (1968) Fujisaki and Kawashima (1968) and others is the result of two separate perceptual processes: (1) an auditory level process of spectral smoothing and (2) a cognitive level process of speaker normalization. It was found that when test words (in a continuum between 'hood' and 'HUD') were presented in an intonational context which biased subjects toward identifying the different tokens as having been produced by the same speaker the shift in identification due to F0 normalization was less than when the test words were presented in isolation. This was taken as evidence that the F0 normalization effect which is found when items are presented in isolation involves two separate perceptual processes simultaneously. These data are noteworthy because they indicate that both low level auditory processes and higher level cognitive processes are involved in the same phenomenon. Any attempt to strictly separate the domains of application
of auditory and cognitive processes to specific types of speech perception phenomena seems to run counter to these data.

The experiments reported in chapter 3 were designed to test the hypothesis that the higher formant normalization effect reported by Fujisaki and Kawashima (1968) and Slawson (1968) is the result of an auditory integration of F2 and F3 in vowels which have F2 within 3 Bark of F3. This hypothesis was tested by varying F3 in two different vowel continua: (1) 'hid'-'head' and (2) 'hood'-'HUD'. In the 'hid'-'head' continuum F2 and F3 are within 3 Bark of each other, while in the 'hood'-'HUD' continuum they are not. There was a shift in the identification function in the front vowel continuum which was consistent with the hypothesis, while (as predicted) there was no such shift in the back vowel continuum. This result is striking because it provides evidence for the auditory basis of a speech perception effect without using non-speech analogs of speech stimuli.

When non-speech analogs are used to demonstrate the existence of an auditory process the interpretation of experimental results is open to question because it is not clear that the perceptual processes which are activated in the 'identification' of non-speech tokens are the same as those that are used in speech perception. By avoiding the use of non-speech tokens in this test of the hypothesis, I was able to avoid this problem of interpretation. Of course, there are other problems in interpreting the results of these experiments, the most basic of which (the nature of the 'centre of gravity' effect) will be discussed below.
Implications.

In this section I will discuss the ways in which I think this research fits into current speech perception theorizing and how the results relate to the study of cognition.

The 'centre of gravity' effect.

The interpretation of these experimental results relies heavily upon the 'centre of gravity' (COG) effect, and to the degree that these results are coherent when interpreted in this way the COG effect is validated. However, I have not attempted to test the effect directly. In particular it is important to be able to define the COG effect.

I have assumed (as have others) that the COG effect is a part of auditory processing and is not cognitive in nature. It is not like other auditory effects, however, because the auditory representation of sound is not limited to its output. Hearers' sensitivity to spectral differences is greater than would be predicted if all they had to go on was the output of the COG effect (which smooths out spectral detail). Thus, the fact that vowels produced by a man and a woman may be identified as being the 'same' (partly at least because of the smoothing performed by the COG effect) does not mean that hearers are unable to tell that the two utterances are different. Speaker recognition is not eliminated by the COG effect. These considerations seem to indicate that auditory perception involves at least two separate levels of representation - roughly corresponding with the abilities of discrimination and identification. The processes involved in the transformation of the acoustic signal into a pattern of electrical activity in the auditory nerve along with memory limitations for these representations
provide the limits of discrimination. The auditory processing involved in the COG effect seems to be more relevant to the ability to recognize similarities among different items (by eliminating detail in the representation).

Schwartz (1987) describes a model of the peripheral auditory system in which there is a level of representation which corresponds to the pattern of activity in the auditory nerve and a second, coexistent level which corresponds to the smoothed spectrum which results from 3 Bark integration. Some perceptual activities make use of the lower, more fine grained level (psychophysical discrimination tasks, speaker recognition) while others depend upon the smoothed spectrum (vowel identification). There is no reason to believe that the lower level auditory representation is replaced by the smoothed spectrum. Rather, it seems more likely (as Schwartz suggests) that they exist in parallel.

So, the COG effect is not a part of the auditory periphery. But, it does not seem right, either, to classify it as 'cognitive'. I would prefer to reserve 'cognitive' to refer to those processes which involve the application of knowledge to the interpretation of new (or other) knowledge. Intuitively, one might say that cognitive processes result from one's experience with the world, but this does not uniquely specify cognitive processes because certain aspects of sensory perception are set by real world experience. So, the defining feature of cognitive processes is that they involve the utilization of knowledge in performing their function.

Given an understanding of cognitive processes which is something similar to this, it is then not reasonable to classify the COG effect as a cognitive process. It does not encode or make use of any information about speakers or speech patterns, but rather is blind in its application. As it is
currently understood, it cannot be attributed to anything in the experience of listening to speech or to making associations between different aspects of speech (say F0 and typical formant levels).

The experiments here seem to support the hypothesis that there is a COG effect and they require no substantial changes in the hypothesized nature of the effect.

Levels of processing.

The approach that I have taken in this research is one in which we attempt to identify and describe the characteristics of different levels of processing. Some of the most productive work along these lines has been pursued in the area of memory and speech perception (Pisoni, 1971; Crowder, 1981). It is interesting that the recent work in levels of memory for speech and my own work on levels of speech processes converge. Recently Fox (1985) has argued that vowel contrast phenomena result from contrast at different levels of memory (with vowel contrast occurring both before and after speaker normalization). He also argues for an intermediate level of memory which is not peripheral auditory memory and also not a cognitive level of phonetic memory. It may be that this level of memory corresponds to the output of a spectral smoothing process like the COG effect. The convergence of results is striking.

Sawusch and his colleagues (Sawusch, 1977; Sawusch and Jusczyk, 1981; Sawusch and Nusbaum, 1983; and Sawusch et al., 1980) have argued for an explanation of both vowel contrast and consonant contrast phenomena which involves different levels of memory. Like the research reported in this dissertation these researchers have found results which
suggest that there are both cognitive and auditory effects in a single class of speech perception phenomena.

The data on levels of memory and processes have implications for parallel distributed processing (PDP) models of speech perception (Elman and McClelland, 1986). Work such as that reported here and other empirical studies into the levels of processing and memory utilized in speech perception should provide a foundation for the design of neural models. PDP models do not provide data about human speech perception processes or abilities. They may be useful (as is any concrete theory) in suggesting directions for future research, but as explicit instantiations of speech perception theory they are useful only if they are constrained by the empirical findings of studies such as those reported here. In particular, the use of hidden layers of nodes (which through a training algorithm may substantially improve the performance of the model) must be avoided unless there is theoretical justification for the level and its characteristics (connections and their weights).

Perception and production.

In chapter 2 I concluded that there is a process of F0 normalization which involves adjusting the expected range for formant values (an internal vowel space) based on F0. This process involves the application of knowledge about the typical relationship between F0 and formant values in vowel perception. The experiment offers no evidence about the origin of the process. I was only concerned to determine whether it was cognitive or not.
If the speaker-based process of F0 normalization exists it is clearly a process of speech perception which involves a link between perception and production. The most straightforward way to describe the process is as a cognitive process in which the hearer's knowledge about typical correlations of F0 and formants is used in the act of vowel identification. I prefer to think of this as an association developed through experience, but the data here do not rule out the possibility of a biologically based module which encodes such a link between production and perception (Liberman and Mattingly, 1985).

What is important about the results and conclusions of chapter II is that they indicate that the cognitive level processing which takes place in speech perception is not simple pattern matching. Evidently, speech perception involves cognitive processes which interpret the speech signal in light of the hearer’s knowledge of speech patterns and regularities.

Sensory processes and cognition.

The relationship between sensory processing and cognitive processing which is indicated by these experiments has important ramifications for the study of cognition. It has been important in this work to keep in mind that the central nervous system does not have access to the speech wave but only to the pattern of activity in the auditory nerve(s). The results which I have presented in this dissertation (especially chapter 3) indicate that by disregarding the activity of the sensory system we will incorrectly define the task to be accomplished by cognitive processes. I believe that this is also true of other areas in which we wish to study cognition. For example, it is common to study syntax as if the only syntactic
data available to hearers is a string of words on the page. Perhaps it would be productive to consider the role of prosody in parsing and the influence of prosody upon the organization of syntactic structures. In view of the fact that spoken language is the foundation for written language it may be that readers use information that they have about the prosodic structure of language to process text.

As another example of the importance of stimulus properties for the study of cognition consider the process of making inferences during sentence processing. Geis and Zwicky (1971) point out that hearers will come to conclusions (invited inferences) which are not logically warranted by the utterance which invites the inference. For example, ‘If you mow the lawn, I will pay you five dollars.’ invites the (logically incorrect) inference, ‘If you don’t mow the lawn, I won’t pay you five dollars.’ This points out the difference between ‘natural’ logic (which results from real world experience and/or social contract) and ‘formal’ logic (which is a more technically oriented system of logic).

The importance of my research for the study of cognition is that I have demonstrated that in order to accurately describe cognitive processes it is necessary to describe the processes which make the outside world available to the mind.
Ainsworth, W.A. 1971. Perception of synthesized isolated vowels and
words as a function of fundamental frequency. JASA. 49:1323-1324.

in Fant and Tatham (Eds.) Auditory analysis and perception of

Assman, P.F. and T.M. Nearey. 1987. Perception of front vowels:
The role of harmonics in the first formant region. JASA. 81:520-534.

Perceptual normalization of the vowels of a man and a child.

Bladon, R.A.W. 1982. Arguments against formants in the auditory
representation of speech. In Carlson, R. and Granström, (Eds.) The
representation of speech in the peripheral auditory system.
Amsterdam: Elsevier Biomedical.

theory of speaker normalization. Language and Communication. 4:59-69.

Carlson, R. Fant, G. and Granstrom, B. 1975. Two-formant models, pitch
and vowel perception. In Fant and Tatham (Eds.) Auditory analysis
and perception of speech. London: Academic.

Chiba, Ts. and Kajiyama, M. 1941. The vowel: Its nature and structure.
Tokyo-Kaiseikan: Tokyo.

Chistovich, L.A., Sheikin, R.L. and Lublinskaja, V.V. 1979. ‘Centres of
Gravity’ and spectral peaks as the determinants of vowel quality.
in Lindblom and Ohman (eds.) Frontiers of speech communication

Crowder, R.G. 1981. The role of auditory memory in speech perception and
discrimination. In Myers, T., Laver, J., and Anderson, J. (eds.). The

Cutting, J.E. 1978. There may be nothing peculiar to perceiving in a
speech mode. In J. Requin (Ed.) Attention and performance. VII.

78


Hillenbrand, J. and Gayvert, R.T. 1987. Speaker-independent vowel classification based on fundamental frequency and formant frequency. JASA. 113th meeting.

_______. and McMahon, B.J. to appear. The role of static spectral properties in vowel identification.


Nearey, T.M. To appear. Static, dynamic and relational properties in vowel perception. JASA.


