UNIVERSITY OF ALBERTA

SPEAKER NORMALIZATION
IN THE
PERCEPTION OF ADOLESCENT VOWELS

BY

© SUSAN R. MERINO

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND
RESEARCH IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF MASTER OF SCIENCE

IN

SPEECH PRODUCTION AND PERCEPTION

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA

FALL 1990

ISBN 0-315-64954-2

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Susan R. Merino

TITLE OF THESIS: Speaker Normalization
in the Perception of Adolescent Vowels

DEGREE : Master of Science

YEAR THIS DEGREE GRANTED: 1990

PERMISSION IS HEREBY GRANTED TO THE UNIVERSITY OF ALBERTA LIBRARY TO REPRODUCE SINGLE COPIES OF THIS THESIS AND TO LEND OR SELL SUCH COPIES FOR PRIVATE, SCHOLARLY OR SCIENTIFIC RESEARCH PURPOSES ONLY.

THE AUTHOR RESERVES OTHER PUBLICATION RIGHTS, AND NEITHER THE THESIS NOR EXTENSIVE EXTRACTS FROM IT MAY BE PRINTED OR OTHERWISE REPRODUCED WITHOUT THE AUTHOR'S WRITTEN PERMISSION.

2608 Ontario Street
Vancouver, British Columbia
V5T 2X 9

Date: August 1,1990

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

THE UNDERSIGNED CERTIFY THEY HAVE READ, AND RECOMMEND TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH FOR ACCEPTANCE, A THESIS ENTITLED "SPEAKER NORMALIZATION IN THE PERCEPTION OF ADOLESCENT VOWELS",

SUBMITTED BY SUSAN R. MERINO

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE

IN SPEECH PRODUCTION AND PERCEPTION

DR. T. M. NEAREY

B. Rochet

DR. B. ROCHET

DR. F. B. WILSON

DATE: 1 Aug 1990

Abstract

Theories attempting to explain the lack of invariance of an F1 by F2 representation of vowels brought about by speaker differences often propose different candidates for a speaker normalization factor, including: (intrinsic) F0, (intrinsic) F3, and (extrinsic) formant range information. Human listeners' identification of adolescent vowels presented in two listening conditions was compared to the predictions from a statistical model of speech recognition in order to test hypotheses concerning the importance of intrinsic and extrinsic information in speaker normalization. The vowels from eight speakers, whose mean fundamental frequency correlated poorly with their mean formant frequency (r= .21), served as the stimuli (data) for both the human listener and pattern recognition identification tests. The potenetial intrinsic cue (F0) to speaker identification (formant range information) was not available in these. Nonetheless, the human rates of identification are still comparable to those for "normal" adult vowels, in similar studies, in which the expected correlation between fundamental and formant frequencies exists. While the addition of syllable-intrinsic F3 or F0 slightly improved the statistical classification, extrinsic information (provided by a formant normalization procedure) was even more effective. Evidence for extrinsic formant range normalization is provided by the improvement in the identification rates in a speaker blocked (over a mixed speaker) presentation being correlated with the change in the response profiles between a model that included F1 & F2 and a model that included formant range normalized F1 & F2. Cues to speaker identity useful in the normalization of vowels seem to be available from both intrinsic and

extrinsic sources, and are perhaps used in varying degrees dependent on their value under the particular conditions.

ACKNOWLEDGEMENTS

Thanks to my classmates Murray Munro and Jean Andruski for all their help over the last years. For genuine moral support, heartfelt thanks go to Satomi, Jean, Ming Ming, Lois Browne and of course, my parents. I appreciate the time spent and the comments made by my committee members, Drs. Rochet and Wilson. I am grateful to the Department of Linguistics for support in the form of assistantships and to the Faculty of Graduate Studies and Research for a Travel Grant. Most of all I must thank my supervisor for his patience and encouragement.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Speaker differences (graphically represented as the overlap of vowel categories in a plot of F1 x F2) are a major source of variation in the acoustic characteristics of vowels. Joos (1948) and Chiba and Kajiyama (1941) were among the first to point out that the first two formant frequencies, though important determinants of vowel identity, are not invariant across different speakers. One way that vowel perception theory has attempted to reduce the overlap problem that is introduced by speaker variation, is through an appeal to normalization of the vowels. Not all contemporary theorists embrace normalization explicitly. One alternate explanation appeals to formant ratios. Both a normalization view and a formant ratio view can explain the overlap problem to some degree. The difference between approaches hinges on whether or not a distinct normalization factor is involved in the perceptual process. If such a factor is in fact part of the speaker normalization process, we should be able to discover where it comes from. The candidates for normalizer include factors such as F0, F3, and formant range information. Confusion sets in when both normalization and formant ratio explanations appeal to the same source of information. For example, a ratio theory may also claim F0 or F3 to be important, but not refer to it as a normalization factor. Therefore the distinction in how the information is used must be made clear. This will be called the manner distinction. Historical differences among vowel perception theories will be outlined in a framework of two basic distinctions. The first addresses the question of where the normalization information comes from. I will use the terms "intrinsic" and "extrinsic" (as was done by Ainsworth, 1975; and

Nearey, 1989) in discussing the information-source issue. The second distinction, the manner-of-use issue, will be discussed in terms of Johnson's [in press] terminology : "direct" or "indirect" usas of such information. A description of how some theories differ in terms of these distinctions is presented. While this study looks at the possibility of either intrinsic or extrinsic sources providing speaker normalization information, it assumes an "indirect" *use* of such information. That is, a normalization view rather than a ratio view is held. Finally, the motivation for using adolescent voices to investigate speaker normalization is outlined.

Terminology

Normalization, as opposed to a ratio theory, involves a re-scaling of formant values. This is typically achieved by using a speaker scale factor calculated by reference to some extra-syllabic factor (e.g. formant range information from context vowels). Extra-syllabic information may be thought of as coming from an extrinsic source. Relative formant normalizations such as the constant log interval hypothesis (CLIH) of Nearey (1978) and the linear rescaling normalizations of Gerstman (1968) and Lobanov (1971), fall into the category of extrinsic theories. They may differ in what specific information is required to calculate the scale factor (and in the number of scale factors), but the source of the information is always extrinsic. Evidence for the formant range effects of context on test vowels abounds (most of it is from synthetic carrier sentences and precursor syllables examples): Ladefoged and Broadbent, 1957; Ainsworth, 1975; Nearey, 1978; Assmann, Nearey and Hogan, 1982 (hereafter ANH); Dechovitz, 1977.

Classic ratio theories, on the other hand, do not refer to anything outside the syllable; they are based exclusively on information intrinsic to the vowel. Intrinsic factors include the formant frequencies, the fundamental frequency and any other components of the vowel itself (also mostly tested by synthetic stimuli: Fujisaki and Kawashima, 1968). Support for ratio "normalization" has been fuelled by evidence of the strong relation between F0 and formant frequencies in natural voices. For example, in the intonation range of a single speaker, F1 is seen to weakly follow F0 (Syrdal and Steele, 1985). There is also cross-speaker evidence for the link. An increase in formant frequencies of approximately 30%, accompanying a 100% increase in fundamental frequency, was noted by Ainsworth and confirmed in calculations of natural data, including Peterson and Barney's (1952) study of male, female and child vowels. However, this 30% to 100% relation between formant and fundamental frequencies is not always upheld. Although there have not been many studies on natural voices having a low correlation between fundamental and formants, the Gottfried and Chew (1986) study showed that in vowels intoned by a (male) counter-tenor, the increase in formant frequency corresponding to a 100% increase in pitch was only 5 tc 10%, (much lower than the expected standard 30%). Slawson (1968), with synthetic experiments found that difference judgements in vowel quality due to a 100% rise in F0 were minimized when accompanied by a 12% increase in formant frequencies. Such evidence weakens the argument for the direct use of F0 as a normalizer.

Both intrinsic and extrinsic factors have been empirically shown to be useful sources of information for vowel perception. Ainsworth (1975) found both intrinsic (F0) and extrinsic (formant ranges of precursor vowels) effects to influence the categorization of synthetic stimuli. Nearey (1989), expanding on Ainsworth's design with better-quality synthetic stimuli, also found both intrinsic and extrinsic effects. Both experimenters found greater effects for context formant ranges than for F0, however. Therefore, intrinsic information alone certainly cannot be the entire answer, in view of the good evidence for effects of context (Ainsworth,1975; Dechovitz, 1977). Neither, though, does a purely extrinsic explanation appear to be completely accurate, since isolated vowels of different speakers in random order are quite well identified (ANH, 1982). If speaker context is required for vowel perception, a mixed speaker presentation (of voices from very different formant ranges, e.g. male and female ranges) would cause misidentifications. However, ANH found that a mixed speaker condition still yielded high identification rates of vowels of men and women (94.5%). Nevertheless, they also found that when vowels were presented blocked by speaker, (providing extrinsic formant range information) identification rates were significantly higher than in the mixed speaker condition. A number of studies show similar advantages under blocked speaker conditions. See Nearey (1989, Tables I and II, p. 2089) for a summary.

In addition to the source of information used for normalization, Johnson (in press) distinguishes the manner in which the information is used. It can be used directly to characterize the vowel, as intrinsic ratio theories propose; or indirectly, suggesting a referential perceptual space for vowel interpretation, as extrinsic formant range normalizations suggest. Johnson's direct/indirect distinction incorporates the "pure" theories referred to by Nearey (1989).[1] Ratio theories (Traunmüller, 1981[2]; Syrdal and

[1] Johnson's terms "indirect-extrinsic" and "direct-intrinsic" are equivalent to Nearey's "pure-extrinsic" and "pure- intrinsic", respectively.

[2] Traunmüller's more recent work appears to be less extreme.

Gopal, 1986) contend that formant spacing is the only important factor for vowel identification. Early suggestions of this nature were made by Potter and Steinberg (1950) and, physiologically, are frequently related in terms of tonotopic basilar membrane sensitivity. The proponents belie\ e that the frequency ratio or Bark difference values of F1 to F0, F2 to F1 and F3 to F2 are the defining characteristics of vowels. The manner in which F0 and F3 are treated is crucial to the distinction being made between direct and indirect theories. In a ratio theory the role of F0 is direct since it makes up part of the ratio. The role of F3 would also be direct. In an indirect theory, on the other hand, the role of F0 and/or F3 is only indirect, in the sense that they are used to calculate a normalization factor.

Fujisaki and Kawashima (1968) found both F0 and F3 to be influential in the perception of vowels, and while the design of their experiment strictly tested a ratio theory, they suggested that these intrinsic factors may in fact be used indirectly. An intrinsic factor, in an indirect treatment, acts as a cue to the referential F1 X F2 space for a particular speaker. A theory that proposes the indirect use of intrinsic factors is logically distinct from either a ratio-type (direct-intrinsic) or a formant normalization (indirect-extrinsic) theory, though experiments necessary to differentiate among them empirically may be quite subtle. (See Johnson, in press).

Slawson (1968) posits the possible combination of uses of information. A *direct* use is made of inherent lower formant cues in a "sensory" stage. An *indirect* use is made of F3, as an aid in the interpretation of the lower formant information, in a phonetic or cognitive stage. Listeners made judgements as to the vowel quality differences between pairs of synthetic sounds. Pairs of sounds in a series (presented in an AX experimental design) differed by a formant shift factor ranging from 0.95 to 1.3. One experiment raised the higher formants (F3 and F4) to values comparable to those of a child (based partly on Peterson and Barney, 1952), another experiment doubled the standard F0 from 135 Hz to 270 Hz , and a third experiment raised both the F0 and the higher formants. Slawson found that the greatest magnitude differences in quality were due to shifts in the lower two formants. Smaller differences in vowel quality were functions of both the first two formants and the fundamental. Higher formants also played a very small role. He proposed F3 as a candidate for the cue (with little effect on vowel quality) that is used as an aid in the identification of children's vowels. The 100% increase in F0, from typical male to child values, caused minimal difference judgements when it was accompanied by an approximate 12% rise in formant frequency. Slawson therefore suggests that the identity of the talker (expected formant range) may be determined by the length of the vocal tract. Since higher formants remain fairly stable across vowels for a single talker, F3 is often thought of as a correlate to vocal tract length, in an inverse relation (Peterson and Barney, 1952). The higher formant information could be used to cue the expected range of the lower formants. Slawson suggests that "the small quality differences, due to higher formants, indicate how the greater differences, due to lower formants, should be used to classify the vowel."

Methodology: Stimuli

By virtue of the hypotheses, much of the experimentation testing intrinsic theories has used synthetic stimuli. Using synthetic stimuli allows the experimenter easy control of experimental factors. Work on naturally produced vowels does not permit the same freedom. However, evidence from studies with natural speech is needed to verify the results from synthetic speech experiments. Adult male versus adult female vowels can be used when formant range differences or F0 differences are desired in experimental stimuli. However, in adult data, where F0 and resonance information are strongly correlated, it is difficult to differentiate between their individual contributions.

Natural voices that lack such a correlation are required for the investigation of the differential effects of the two sources. Nearey (1989) mentions several notable voices that meet the "lack of correlation" requirement, those of: Julia Child (high fundamental but low formant frequencies) and Popeye (high formant frequencies and a low fundamental), for example. It was conjectured that a similar situation might be found in a larger sample, namely in an adolescent population. Because rapid changes in the suprarlaryngeal vocal tract and the larynx occur dunng adolescence, a weak correlation of their acoustic products might also be found. An example of the rapid growth that takes place in this age group (as well as gender differences) is illustrated by the amount of change in size of the larynx itself. The male larynx is known to double in size, while the female larynx increases in size by only half[3] (F. B. Wilson, personal communication). If a weak correlation can be shown to exist between the two sources of information, the opportunity to examine differential effects would be provided. The measurement study of vowels from adolescent speakers that was undertaken to this end is presented in Chapter two.

[3] While the length of the vocal folds is generally believed to be a main determinant of fundamental frequency, many other factors may be involved, therefore the size of vocal apparatus and fundamental frequency may not always be directly related (F. B. Wilson, personal communication).

Chapter three describes a perceptual test involving the vowels from those speakers whose voices were found to differ widely in terms of vocal tract resonances yet had similar fundamental frequencies.

Analyses

Algorithmic testing is a viable demonstration of reduction of the physical overlap in production measures. In favor of using natural stimuli for perceptual tests, ANH (1982) developed a technique to compare statistical classification of vowels from measurements of their acoustic characteristics with actual identification by listeners. An index of resolution that such a method provides is an *a posteriori* probability of membership in an intended category, denoted (APPi). "Intended category" is that which was intended by the speaker in production.

The statistical model serves two functions. It first provides a direct index of resolution by which to compare the performance of algorithms. Secondly, its APP predictions can be used to make relatively fine comparisons between the listeners' and the statistical model's categorization behavior. Though the classification modelling technique provides good evidence for which parameters are salient to vowel perception (by means of testing various combinations of parameters and comparing the resultant measures of resolution), one cannot infer that an additional parameter which causes the model to classify vowels with better separation is necessarily a "normalizing" factor. The model does not explicitly test the contribution of the parameters nor how they interact with each other. In order to qualify a factor as playing a normalization role, it has to be shown that its contribution is indirect, acting to rescale F1 and F2. An *a priori* normalization must be performed. Of the relative formant normalization procedures compared in ANH (1982), CLIH (from Nearey 1978) was one of the more effective, and is therefore employed as the procedure for the current test. Intrinsic factors are tested for their contribution to the separation power of an algorithm merely by their presence or absence in an algorithm.

It will be shown that extrinsic information attained in blocked speaker presentation is of perceptual importance. In addition to extrinsic relative formant information, the possibility of intrinsic F0 and F3 (and other sources of "speaker identity" information) being used indirectly will be entertained in the concluding discussion.

Chapter 2

Measurement Study

Recordings

Thirty-two speakers from ages 11 to 15 were recorded reading /hVd/ words from flash cards presented in succession but random order until the entire list containing the 11 Canadian vowels had been run through at least three times. Repetitions of the list for the same speaker were in different orders, and the order changed from speaker to speaker as well. Two tokens of each of the vowels were digitized for 27 of the speakers. The speakers were volunteers from a junior high school in Edmonton. They were told to read the words in their normal speaking voice. Each speaker was seated at a table upon which the microphone (Sony Dynamic MTL-F96) was distanced approximately 45 cm from, and almost level to his/her mouth. The recorder was a portable Sony CR22 Cassette-corder. The recordings were made in a generally quiet room in the school, over the course of several days, during which recording conditions varied.[4] The floor was tile and the walls thin. There were sounds of footsteps from outside the room captured on the recording. The hum of the air conditioning system was particularly noticeable on one of the recording days. Fortunately, 27 (14 girls and 13 boys) of the speakers' recordings were of suitable quality to be analyzed.

Non-invasive physiological measures (height and neck size) were taken at the time of the recording session.[5]

[4] As will be shown below, the subset of stimuli used in the perceptual experiment were highly identifiable to listeners. In addition, high classification rates in self-trained and jackknifed analyses make it likely that valid measures were available for those tokens.

[5] The physiological measures are not discussed in this thesis, but were presented in another work (Merino & Nearey, 1989).

Digitization

The signals were bandpass filtered at 80 to 7800 Hz on a Wavetek Rockland filter (model 852). The CSRE signal processing software for PC (Jamieson et al.,1990) was used to digitize at a sampling rate of 16.6 KHZ and 12 bit resolution. The signals were stored on a Zenith PC.

Analysis:

Selective-range auto-correlation LPC and cepstral analyses were performed as implemented on a Macintosh II computer (Welz, 1989) using software that was the prototype for the raw track component of the CSRE system (Jamieson et al., 1990).

Parameter settings were changed depending on the expected range of formants, either 9 or 11 coefficients were used (analysis set to the " order" of 9 or 11). The order parameter then would determine the number of formants that could be extracted from the range, two coefficients being required for any given formant. See Jamieson et al. (1990) for a further explanation of the parameters. The first three of four formants were thereby tracked throughout the signal. The number of formants that would be tracked by the analysis also depended on the frequency range setting. A focussed range was used in the higher frequency region if the second and third formants seemed to merge or the tracking was not smooth in that region. For example, the lower limit of the frequency range would be raised to 500 Hz and the number of coefficients decreased.

The upper frequency cutoff was pre-selected at 4000 for initial analysis, but was increased to 4500 Hz if the third formant had not been clearly tracked for a particular speaker.

A typical setting of the parameters for a signal analysis might consist of the following:

LPC window duration: 15 ms.

Hop duration: 5 ms.

Cepstral window duration: 30 ms.

Analysis Order: 11 coefficients.

Frequency range: 0-4000 Hz.

Cepstral limit: 75- 400.

Silence amplitude threshold: 5 dB

Zero crossing threshold: 1500.

The tracking files were displayed in a format with the frequency values printed every 5 ms. From special graphic displays showing unit accuracy to 1 Hertz below the 1000 Hz range, and to 10 Hz above 1000 Hz, measures of the pitch and the first three formants were made "early" and "late" in the vowel. See the appendix for a graphic formant tracking example. The criteria for the taking of measurements were as follows: The EARLY (initial) VOWEL MEASURE was taken as close to the beginning of the signal as possible where (1) the pitch appeared stable, and (2) the overall amplitude was near its maximum for the syllable. The LATE (final) VOWEL MEASURE was taken where (1) the overall amplitude of the signal was within 3 dB of peak amplitude, and (2) the first and second formants did not exhibit rapid movement due to consonantal transition. The third consideration in determining the frame for the late vowel measure was that (3) the frame preceded, by approximately 40 ms, the characteristic, rapid decrease in amplitude due to closure for the final /d/. The DURATION of the vowel was taken as the time between the initial reading of 10 dB below peak amplitude and the next 10 dB below peak reading. The duration, determined almost exclusively by the reliable "amplitude-at-closure" strategy (above), did not always coincide with the duration between the formant frequency measurement frames. The early and late points were used for the measurement of the first three formants as well as the fundamental frequency, so that, ideally, all three formant frequency values and the fundamental frequency were taken from the same analysis frame. If, however, the chosen frame did not provide a reliable reading for one of the formants or the fundamental, the closest time frame was used. Figure 2.1 shows the F1 x F2 vowel space (early measures only) for the twenty-seven speakers.

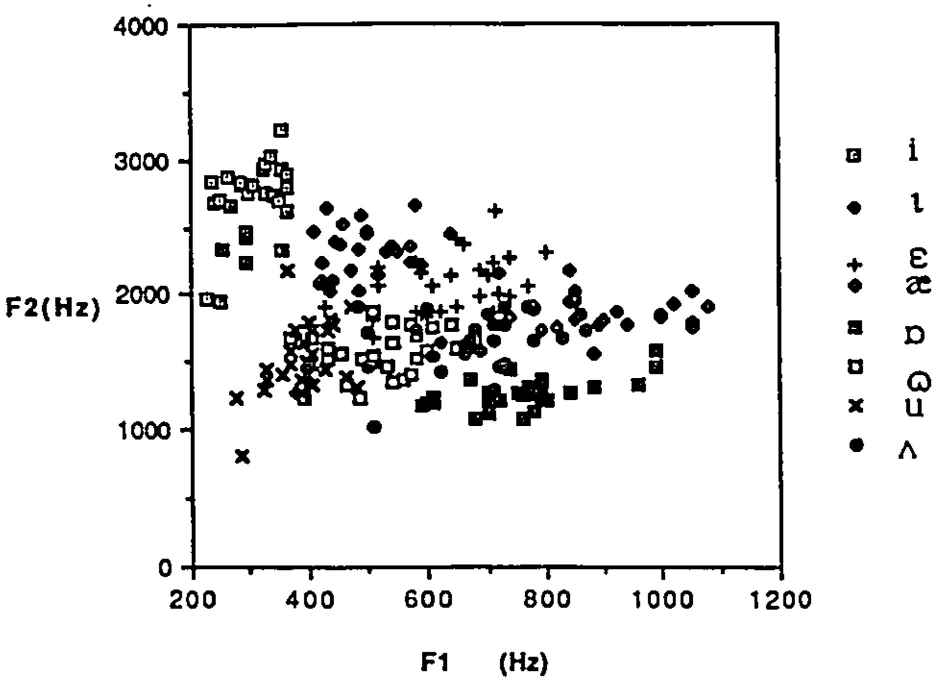


Figure 2.1 The F1 x F2 vowel space of twenty seven speakers. Eight vowels are plotted for each speaker. The vowels without American approximates, /e/ and /o/ are not shown.

Descriptive Comparison

The frequency values from the early measures of the three formants were averaged into a geometric mean for each of the twenty seven speakers. The fundamental frequency values across all vowels for a speaker were collapsed into a mean F0 value. Only the initial measures are used in these averagings for comparison with the Peterson and Barney 1952 study, in which there was one measure taken in mid-vowel. With two measures per speaker, one to represent vocal tract resonance ($\overline{F}$) and the other source frequency ($\overline{F0}$), $\overline{F}$ was plotted against $\overline{F0}$ to determine the amount of correlation between the two potential "size estimating" measures. The Peterson ans Barney data was likewise treated to the same averaging applications, and the two groups were compared for amount of correlation between F0 and formants. The two American vowels /ɑ/ and /ɔ/ were averaged to produce an approximate to the Canadian /ɒ/. The Canadian vowels /e/, and/o/ were not included in the comparison, as the Peterson and Barney data have no equivalents to thəm. Neither was /ɚ/ included.

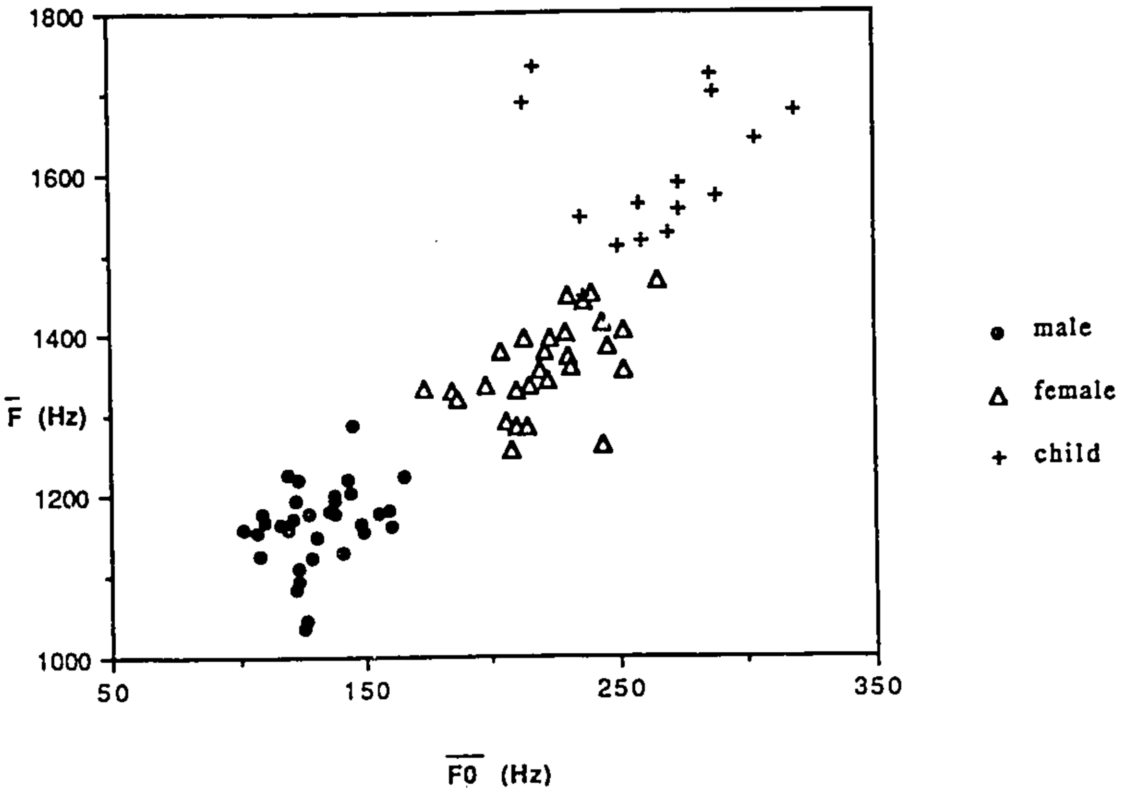


Figure 2.2: Geometric mean of formant frequencies by geometric mean of fundamental frequency, for Peterson and Barney (1952) data (adult male, adult female and child vowels).

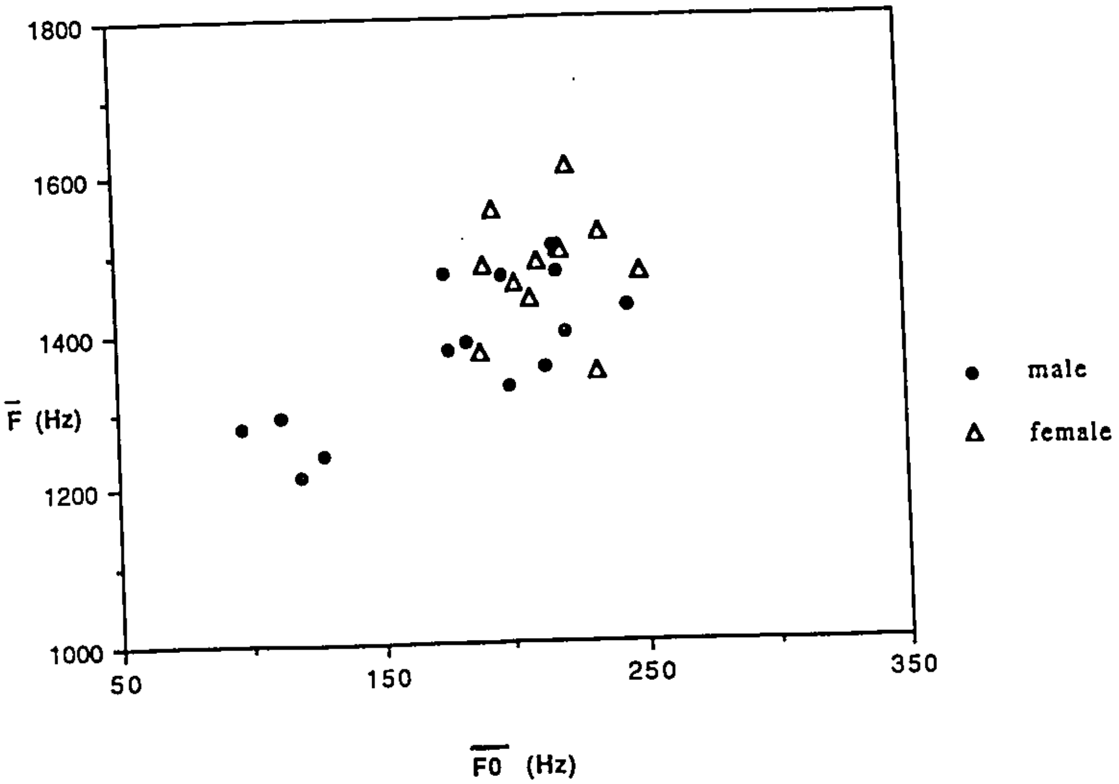


Figure 2.3: Geometric mean of formant frequencies by geometric mean of fundamental frequency, for Hillcrest data (adolescent male and adolescent female vowels).

Table 2.1:Correlation of $\overline{F0}$ with $\overline{F}$
(based on Peterson and Barney 1952 data)

	male	female	child	all
$\overline{F1}$	.36	.50	.29	.84
$\overline{F2}$	.16	.29	.15	.85
$\overline{F3}$	.04	.38	-.09	.84
$\overline{F}$	31	.53	.16	.88

Table 2.2: correlation of $\overline{F0}$ with $\overline{F}$
(Adolescent data)

	male	female	hi-F0	all
$\overline{F1}$	.72	-.01	.12	.58
$\overline{F2}$	.64	.25	.29	.67
$\overline{F3}$	.63	.05	.11	.62
$\overline{F}$	.76	.09	.21	.70

The Peterson and Barney data showed higher correlation (r=.88) as a group than the group of Hillcrest speakers taken as a whole (r=.70). The correlation was radically reduced when the sub-group of higher pitch voices was isolated (r=.21), i.e. , the speakers that had adult male-like voices were in an obviously separated group from the others, as seen in figure 2.3. When the male-like pitched voices were removed, the remaining voices showed an average pitch of approximately 200 Hz. A comparison of the Peterson and Barney female speakers with the Hillcrest high-F0 group shows the average fundamentals of the two groups to be similar. The Hillcrest high-F0 group's vowels are most similar to the Peterson and Barney adult female vowels in terms of fundamental frequency, but in terms of formant frequencies they show a much greater variation than those of the Peterson and Barney adult females.

There is a cluster of voices at a female-like pitch (around 200 Hz), these speakers' vowels will be looked at in greater detail as they will serve as the stimuli for the following perceptual experiment. Before the perceptual experiment is run, the "end-vowel" measures are added to the data corpus.

Another screening of the data found some measurement error in the formant frequency values. Corrections were made prior to the perceptual experiment and pattern recognition analyses. Correlations were checked for any change caused by the new values, and it was found that the $\overline{F}$ was not greatly affected by those few errors.

Chapter 3

Perceptual Experiment

A perceptual experiment was designed to test how well listeners could identify the vowels of adolescent speakers. As in ANH (1982), speaker formant range information is presumed to be available when a speaker's vowels are presented in a blocked order. Further, if the information is used in a manner like that predicted by the formant normalization hypothesis, the vowels should be correctly identified at a higher rate in the blocked than the mixed presentation order, since "exposure to several of a speaker's vowels [facilitates] identification of subsequent vowels" (ANH, 1982, p. 979). The results of ANH 1982 showed a significant improvement in the blocked condition, 95.9% correct, over the mixed condition, 94.5%,(t= 2.36; d.f.= 34; $p < .02$).

In the present experiment the speaker dependent factors are reduced to formant information almost totally. The vowels from nine speakers, four boys and five girls, were selected on the basis of similar values of the calculated geometric mean of the fundamental frequency ($\overline{Fo}$). The nine speakers' mean $\overline{Fo}$ was 221 Hz, with a 5.88 Hz standard deviation.

The geometric mean of formant values ($\overline{F}$) on the other hand, varied greatly (standard deviation 108.74). Speaker information is presented in Table 3.1. A Pearson correlation between $\overline{Fo}$ and $\overline{F}$ for the selected speakers actually gave a negative r value (- 0.245). The negative correlation is opposite to what would be expected in an adult population.

Table 3.1: Chosen speaker information

Speaker	Sex	Age	$\overline{F}$	$\overline{F0}$
1	male	12	1870	218
2	male	12	1688	213
3	male	13	1814	218
4	male	13	1650	222
5	female	12	1959	223
6	female	12	1853	219[6]
7	female	11	1814	220
8	female	14	1802	221
9	female	14	1659	233
mean $\overline{F0}$				221 (s.d.=5.88)
mean $\overline{F}$			1782	(s.d.=108.74)

Other differences in the voices may exist that have not been measured in this study, but the "direct" intrinsic scaling factor of pitch cue has effectively been removed, as much as possible by the selection of voices. The test hypothesis weighs the extrinsic scale factor of formant normalizer more heavily as the favorable information to aid perception. If the vowels are better identified in a blocked order, despited the pitch cue to expect similar formant ranges, then it can be said that the extrinsic information is useful in aiding vowel identification in its own right and not merely a

[6] This speaker's vowels were deleted from the corpus before analysis of the data because of experimental error in labelling. One vowel was missing and another was over-represented. To avoid complications in the scoring the subject's data was deleted completely, leaving eight speakers' vowels, a dataset of 80 vowels.

supplementary cue that backs up the intrinsic pitch cue. A pure (direct) intrinsic hypothesis would predict that since the vowel inherent cues do not correlate, vowel identification should be poor in both conditions. Speaker blocking, which is advantageous according to a formant normalization hypothesis, would not be considered so by a pure intrinsic hypothesis.

Method:

Three repetitions of each vowel were heard in a mixed order presentation and again in a speaker-blocked presentation. Half of the listeners heard the mixed order first followed by the blocked, the other half heard the blocked followed by the mixed order. All listeners heard both orders. Each of the two sessions was exclusively one presentation order. The test sessions were separated by a minimum break of two hours; most listeners came on another day for their second session, and for some the sessions were separated by several weeks. Listeners were not permitted to do the sessions consecutively, in order to avoid a drop in attention, and at the same time to avoid any learning effects (presuming the familiarity with the voices would have been lost after some time.)

Listeners were seated at a Zenith 286 PC terminal, heard the stimuli at a comfortable listening level over tight-fitting headphones, and used the keyboard to respond. The response to each presented vowel was made by hitting the appropriately labelled key on the keyboard: /i,ɩ,e,ɛ,æ,ɒ,ʌ,o,ɷ,u/. A stimulus vowel was heard only once, and the response time was determined by the listener. When the response was made, there was a 250 ms wait period before the next stimulus vowel was presented. Labels for the forced-choice vowel categories were in phonetic transcription and accompanied by a sample /hVd/ word typed above the phonetic symbol as a reminder. The responses for the three trials (per session) were concatenated, so that there was one response table per subject per presentation order.

Subjects

Twelve listeners participated voluntarily. All were familiar with phonetic transcription, having taken at the very least two courses in phonetics. One listener was an unoergraduate linguistics major, three were professors and eight were graduate students in the Department of Linguistics at the University of Alberta. The three professors were American, two of whom had been residents of the Edmonton area for more than ten years. The third professor was a visiting professor who had lived in the area just under a year. He expressed discomfort with one vowel category, finding it unnatural, and somewhere between his Mid-Atlantic dialect's two low back vowels:/ɑ/ and /ɔ /.

Results

Percent error was calculated for each of the listening conditions: mixed and blocked, and the mean error rates compared. Listeners in the mixed condition got 92.42 % correct identification, while in the blocked condition they improved to 94.19 % correct. Although the mean percent correct was higher in the blocked than the mixed condition, a t-test of paired scores was not significant. A systematic labelling error by one of the listeners was noted, and rather than discard his response data, an additional t-test was run with the revised mean score.

Table 3.2(a): Listeners' percent correct identification

Listener	Mixed(x)	Blocked(y)	Difference
bd	78.33	90	11.67
ea	94.16	94.17	0.01
ja	96.67	95.83	-0.84
kd	95.4	97	1.6
kt	95.4	97.5	2.1
mm	97.08	99.17	2.09
rb	93.3	93.75	-0.45
rd	90.83	92.5	1.67
rt	93.75	95	1.25
ss	87.9	83.3	-4.6
td	93.3	98.3	5
tn	92.9	93.75	0.85
mean	92.42	94.12	

Table 3.2(b): Listeners' percent correct identification.
Correction made for labelling error

Listener	Mixed(x)	Blocked(y)	Difference
bd	95	90	-5
ea	94.16	94.17	0.01
ja	96.67	95.83	-0.84
kd	95.4	97	1.6
kt	95.4	97.5	2.1
mm	97.08	99.17	2.09
rb	93.3	93.75	-0.45
rd	90.83	92.5	1.67
rt	93.75	95	1.25
ss	87.9	83.3	-4.6
td	93.3	98.3	5
tn	92.9	93.75	0.85
mean	93.81	94.12	

Table 3.3 Listeners' mean identification scores

Listening condition	Percent correct	Standard deviation
Random(table 3.2a)	92.42	5.1
Random(table 3.2b)	93.81	2.5
Blocked	94.19	4.3

The corrected version (see Table 3.3) while reducing the difference in means also lowered the standard deviation for the mixed condition percent correct scores. A significant difference was still not found between the listening conditions. A randomization test (the most powerful non-parametric test) was run on the paired scores. It is calculated as follows:

dif = score(i,c)- score (i,c'),

With *c,i,c'* tested over all possible permutations of assignments of mixed/blocked treatments ($2^{12} = 4096$) the result was not significant ($p<0.082$).

Most of the listeners (eight) did better in the blocked than the mixed condition. The trend toward higher scores seen here may have become a significant difference if the sample population had been larger (d.f.= 11). The high identification scores in the mixed condition at least rule out the most extreme relative formant range hypothesis, which predicts much lower scores in a speaker-mixed than blocked presentation.

The slight improvement in the blocked over the mixed orders, consistent with findings from previous studies, suggests that extrinsic properties in the vowels aid identification. To strengthen the evidence for the influence of extrinsic factors, further comparative analyses are needed. The data analytic methods used in ANH (1982) and Nearey and Assmann (1986), hereafter NA, were emplcyed for that purpose.

The particular vowel errors made are of interest. Pooled responses (pooled over all listeners in the mixed speaker condition) were plotted in a confusion matrix for each speaker. (That confusion matrix of mixed order responses is later compared to the automatic recognizer in Chapter 4). Figure 3.1 is an example of a confusion matrix from which errors in categorization are easily seen as any response off the diagonal. The presented vowel token is shown on the vertical and the response vowel categories are on the horizontal. Correct responses fall on the diagonal (and the particular vowel substitutions are easily spotted as anything off the diagonal). The height of the bar represents the pooled listeners' number of responses, where the highest bar is 100% of the responses. When a vowel was not correctly identified 100% of the time, the height of the bar on the diagonal is reduced and other bars appear in that row under the columns of the incorrect vowel responses.

Figure 3.1 shows the pooled responses to the first 10 tokens (the vowels of speaker 1). White bars represent the mixed condition and shadow bars are the blocked condition responses.

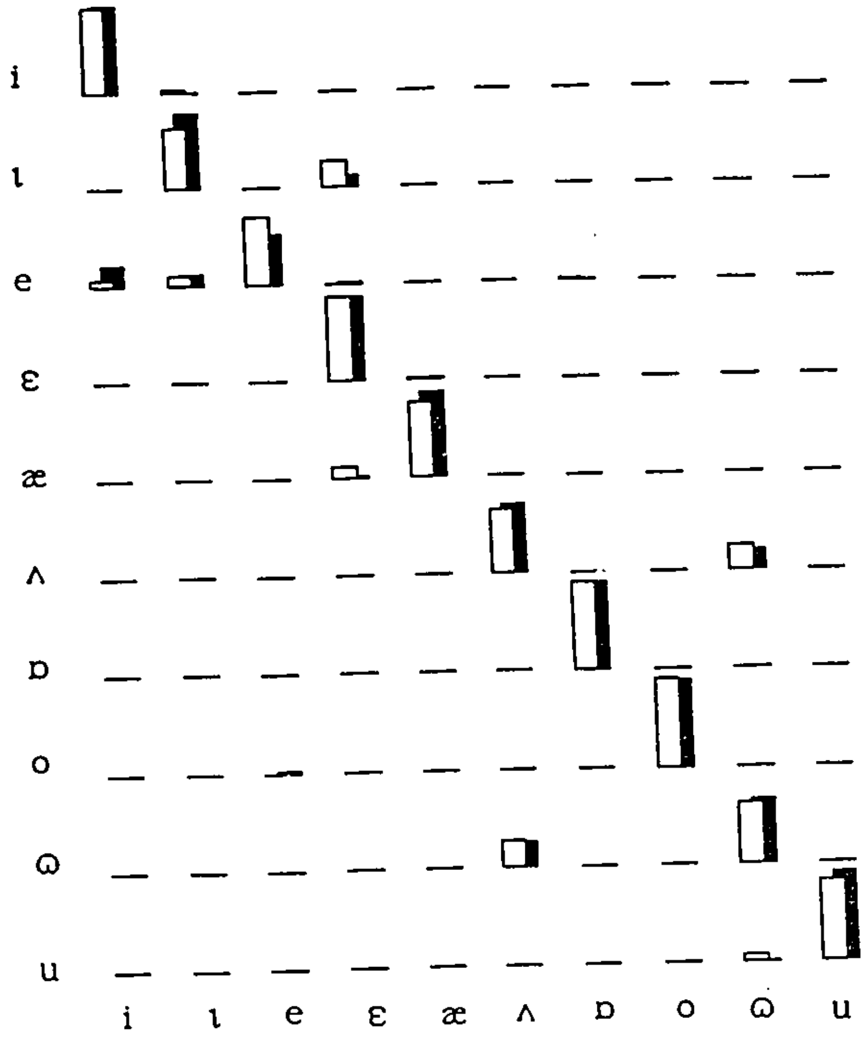


Figure 3.1 Vowel response confusion matrices for two speaker conditions. The presented vowel is indicated on the vertical and the perceived vowel on the horizontal. Maximum height of a bar is 100% identification. The white bars represent the responses to vowels presented in mixed speaker order and the shadow bars represent responses to vowels presented in speaker-blocked order.

It can be seen, for example, that /e/ is better identified in the mixed condition, but /ɩ/ was better identified in the blocked condition. Discounting the /ʌ–u/ replacement by listener 1 (explained earlier as a transcription error) the following confusions were made more than 15% of the time in the mixed speaker condition:

Speaker 1:	/ɪ/ as /e/,
	/æ/, /ɪ/ as /ɛ/
	/ɑ/ as /ʌ/
Speaker 2:	/ʌ/ as /ɑ/
	/æ/ as /ʌ/
	/ɒ/ as /ʌ/, /ɑ/
Speaker 4:	/e/ as /ɪ/
	/æ/ as /ʌ/
Speaker 6:	/ɪ/ as /ɛ/
Speaker 7:	/e/ as /ɪ/, /u/, /i/

Some of the errors are understandable in that the vowels are phonetically similar, (e.g., the confusion amongst the front vowels, for example, seen for speaker 7 and especially for speaker 1). The central vowels in /ɑ/ and /ʌ/ seem to present a problem for listeners, especially those vowels from speakers 1 and 2. Speakers 1 and 2 were problematic in general. There were scattered errors in all their vowels, but none as prominent as the /ɑ–ʌ/ confusion[7]. Speakers 1 and 2 may have a tendency to over-neutralize lax[8] vowels. Speaker 1 also caused listeners to confuse /ɛ/ for /æ/, and /ɪ/. Speaker 2 caused problems primarily with back vowels, and all involving the neutral vowel /ʌ/.

Perhaps many of the substitution errors can be explained in terms of the proximity of the mistaken token's formant frequency values to the mean

[7] Assmann (1979, p.76), of the vowels commonly confused, /ɑ–ʌ / and /ʌ–ɒ /do not show improvement in the blocked condition

[8] Lax occurs as a comparison to tense; vowels that are the lax counterparts of tense ones are shorter, lower, and more central (Ladefoged, 1982).

frequency values of the category it was mistaken for. Plotted F1 x F2 means for the vowel categories are shown in Figure 3.2. The means of the category are based on the eight speakers' values. The problematic vowels are plotted in the F1 x F2 space in Figure 3.3. A comparison can be made of the location of category means and the location of the mistaken vowel (see Figure 3.3).

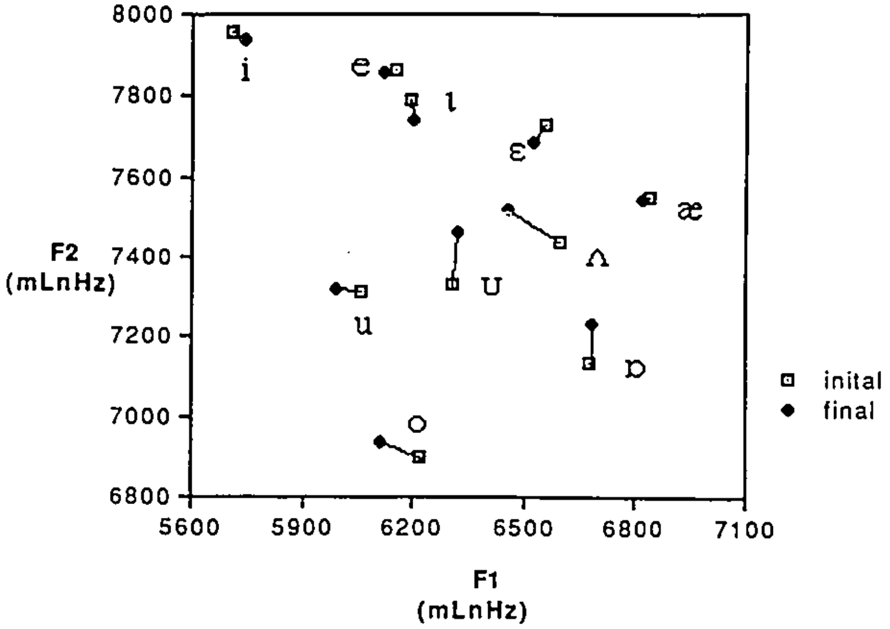


Figure 3.2: Category means of initial and final F1 and F2 values averaged across eight speakers.

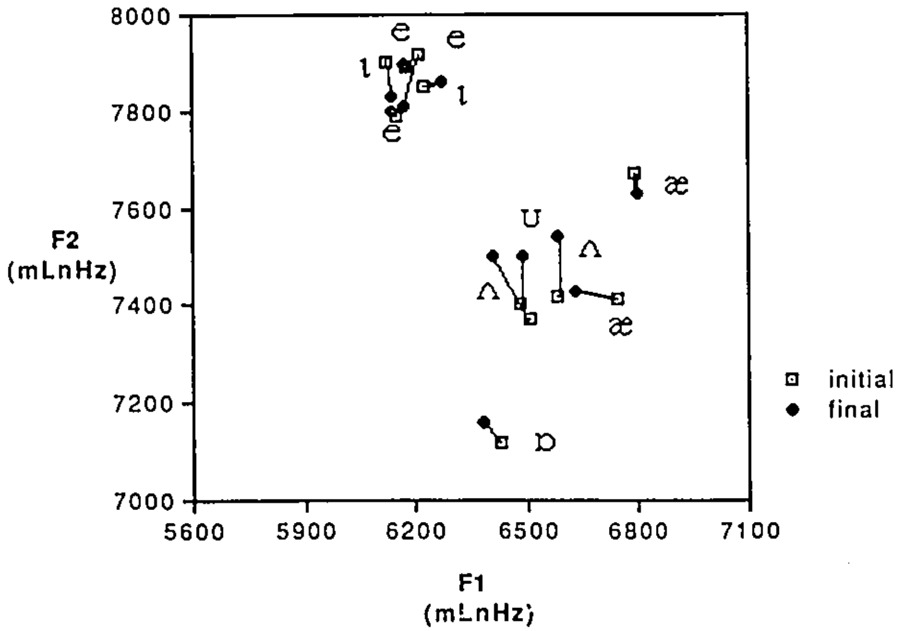


Figure 3.3 Initial and final F1 and F2 values for misidentified vowels. Labels are for the intended vowel category.

If one were to overlay the plots it would be seen that the erroneously identified vowels are often closer to a category mean other than their own category. For example, an intended /ɩ/ (in Figure 3.3) appears to be closer to the /e/ than the /ɩ/ category (in Figure 3.2). Although not all information relevant to categorization is shown in these plots, they nonetheless lend some insight into the problem.

Chapter 4

Data Analytic Vowel Recognition

The purpose of this chapter is to compare listeners' performance with a series of pattern recognition models in a manner similar to that of ANH (1982) and NA (1986). The pattern recognition model is a "'normal *a posteriori* probability model', implemented via linear discriminant analysis" (NA, 1986). In ANH, several distinct pattern recognition model hypotheses were tested. The predictions of the models were compared to listeners' correct identifications in various listening conditions. If a model's performance approximates the listeners' behavior then the parameters that the model uses are assumed to be salient to vowel perception. The algorithm the model uses is likened to the perceptual process of the listener.

From ANH (1982) it was found that the model classified vowels better when given both end points of a formant. Formant trajectory information, "inherent spectral change," was required for the model to be able to classify vowels as well as listeners did. The model also did better (as expected, from Nearey, 1978) when the frequency values were log-transformed. Very good identification rates were attained from a data set including: G0 (averaged), G1, and G2 (initial and final) ("G" refers to a natural log-transformed frequency value). The normalized version, a variation of the constant log interval hypothesis (CLIH) discussed in Nearey 1978, had even higher identification rates. The normalizing procedure of the CLIH is one in which the mean of each formant (across all vowels for a speaker) is subtracted from the formant value of each token for that speaker.

The Pattern Recognition Model

In the present experiment the model is used in each of three ways: (1) in a self-training mode; (2) in a jackknife self-training (partial cross-validation) mode; and (3) in a pure cross-validation mode. In a self-training mode the data to be tested is used to determine category means and the covariance matrix, then the probability of membership in the categories is calculated from that matrix for each token. The Jackknife method (also known as the U method) calculates the membership probabilities slightly differently. In effect, it reduces the optimistic bias otherwise inherent in self-trained classifications (Gray and Schucany, 1972). In a pure cross-validation mode, category means from a matrix calculated on a previous training set are used to test the new data set. Such an a priori test run is important, because it shows that the algorithm does as well on an independent set of vowels.

With the exception of the jackknifed analyses mentioned below, all of the classification procedures and *a posteriori* probability (APP) estimates were carried out within a maximum likelihood Bayesian classification framework. Mean vectors for each vowel category and a single pooled (over all vowels) covariance matrix were estimated from a training set. *A posteriori* probabilities, assuming multivariate Gaussian distribution were calculated for each element of the test set. For the simple self-training mode, the test set was equivalent to the training set. Each test token is classified as belonging to the category for which its APP score is highest. In the case of the jackknifed estimates, classification and *a posteriori* probability estimates were performed using the jackknife option of the BMDP linear discriminant function analysis (LDFA) program. Although in some cases classification via

LDFA takes place in a reduced parameter space, in all the current examples, the number of variates preserved was equal to the total number of variates in the original Bayesian classification analysis. In such cases, LDFA and Bayesian classification analysis are equivalent.

The NA (1986) data set (100 adult vowels, 10 vowels from each of 5 male and 5 female speakers) was always used as the training set for cross-validation runs of the classifier.

Level one comparison:

Statistical Resolution

A subset of the models investigated by ANH (1982) was chosen for evaluation on the present data (see ANH, Table III, p. 982). The measure of resolution adopted here is that used by ANH, the average APP for intended category, which will be denoted by avgAPP_i . The overall percent correct classification scores, based on the calculated APP_i, are also reported. Self-trained runs of the model identified the 80 Hillcrest vowels with an accuracy rate of 91 to 92.5%[9] . The best results, 92.5%, are yielded when only the first two formants, but both the initial and final measures, were given. Performance decreased to 91.25% when either the averaged G0 or initial G3 measures were added to the data set. The initial G3 value is used rather than the average of the initial and final, but it should not differ much from what an averaged value would be since F3 does not change very much from head to tail position. The APP scores reflect a finer measure of fit than the

[9] The self-trained classification is a method of screening the data for measurement error. Since the recording conditions were less than optimal, doubt as to the accuracy of the formant measures may have arisen. The high rate of identification by the pattern recognition model reduces any such doubt.

overall percent correct classification, because an APP score is a continuous variable. As shown in Table 4.1 below, there is actually an increase in performance (from 0.8656 to 0.8694 and 0.8754, respectively) when a "normalizing" [or rather an additional intrinsic factor] measure (G0, or G3) is additionally provided. Moreover, G3$_i$ is slightly more helpful than G0. ("G" before a number indicates the natural log of the formant frequency value). What appears as an improvement by the addition of the extra factor in the maximum likelihood classification is not corroborated in the less biased Jackknife method of classification,[10] where it appears that the addition of either of the extra variables lowers classification performance.

Table 4.1:Self-trained data and listeners classification scores

	Maximum Likelihood		Jackknife method	
Parameters	% Correct	AvgAPPi	% Correct	AvgAPPi
G1i,G1f, G2i,G2f	92.5	0.8656	85	0.8070
G1i,G1f, G2i,G2f, G3i	91.25	0.8754	83.7	0.8015
G1i,G1f, G2i,G2f, G0av	91.25	0.8694	82.5	0.7907
Listeners	% Correct			
Mixed	93.81			
Blocked	94.19			

The combination of this study's data set and that of NA (1986) make up the combined data set. It includes the same five parameters as above for 180 vowels: the 100 NA (1986) adult vowels plus the 80 "Hillcrest"

[10] The Jackknife method was used in NA 1986. It is a less positively biased classification method so its results are therefore more reliable or honest than those from other methods(i.e. the R-method of resubstitution). The test is a cross-validation of sorts (Hand,1981).

adolescent vowels. It is seen that more "power of separation" is attributed to G0 (avgAPPi =0.8158) than to G3 (avgAPPi =0.7678) in the combined data set. The improvement that seems to be attributed to the inclusion of G0 may just be a reflection of its utility for the adult vowels. In a training run of the NA (1986) data set alone, G0 was seen to improve avgAPPi from 0.91 to 0.95, although it may be an oversimplification to separate the effects of each data set in that way. The centroids are custom-calculated on the means for the categories over all the vowels in the training set, and the classification of adolescent data shows a slight improvement itself with the inclusion of G0 (0.8656 to 0.8694 avgAPPi).

Table 4.2: Combined data set (self-trained)

Parameters	% Correct	AvgAPPi
G1i,G1f,G2i,G2f	83.89	0.7678
G1i,G1f,G2i,G2f,G3i	85	0.7883
G1i,G1f,G2i,G2f,G0av	88.33	0.8158

The most unbiased test of the Hillcrest data is by cross-validation with the independent training set. The training set used is the data set from NA (1986). From the avgAPPi scores in Table 4.1 and 4.2 it is evident that G0 or G3 aid identification by the algorithm. Now, from cross-validated measures of resolution (see Table 4.3) it is credible that G0 (avgAPPi = 0.607) is more helpful than G3 (avgAPPi = 0.588)

Table 4 3: Cross-validated classification scores

Parameters	% Correct	AvgAPPi
G1i,G1f,G2i,G2f	57.5	(0.557)
G1i,G1f,G2i,G2f, G3i	62.5	(0.588)
G1i,G1f,G2i,G2f,G0	63.8	(0.607)

It is not known how G0 is used to aid separation of the vowels. The effects of either a direct or indirect use of the intrinsic parameter are not distinguished here. The effects of a third possibility, of a mere difference in intrinsic pitch [11] (Lehiste,1970), which could also be responsible for the improvement in separation, are not distinguishable from these classification rates either. If it were the case that intrinsic pitch differences help distinguish vowels that overlap in two-dimensional space, there would be no need for a normalization explanation. G1 and G2 would not have to be rescaled, by the use of G0. A simple three dimensional solution would be all that was required. On the other hand, since most neighboring vowels in the G1 by G2 space do not differ appreciably in G0, compared to inter-speaker differences, a normalization explanation seems warranted.

In ANH (1982, p. 983), it was found that, for gated vowels, an improvement in identification by the statistical model involving normalized formant measures resulted that was roughly parallel to the improvement of listeners' identification rates in blocked (versus mixed) speaker condition. It was decided to conduct a similar set of tests on the current data. To calculate the normalized formant values for a token, the mean G1 value for

[11] If intrinsic pitch differences exist between vowel categories there may be an apparent advantage to using F0 in the discriminant analysis. The analysis will show if there was an improvement in classification with the addition of F0, but it will not distinguish how it is that F0 helped.

the particular speaker across all of his vowels is subtracted from each of his vowels' G1 values and his mean G2 value is subtracted from each of his vowels' G2 values. The calculation, for each speaker, is represented in the following equation:

$ng_{xy}= g_{xy}-(\Sigma g_{xi})/10$, where

x= formant number 1 or 2

y= i or f, where,

i=initial , f = final

Results

The self-trained normalized data was identified 96.25% correctly. In a cross-validated run the normalized data was 72.5% correctly identified, which is better than even the best unnormalized group of parameters in cross-validation mode (G1,G2,and G0), but still not anywhere near the accuracy of identification which listeners' demonstrate.

Table 4.4: Classification rates of normalized and blocked speaker data

Normalized data	% correct	AvgAPPi
Self trained	96.25	0.930
Jackknife	95	0.905
Cross-validated	72.5	0.711
Pooled listeners		
Blocked condition	94.19	

The jackknife classification results are comparable to the listeners' rates of identification in the blocked condition (see Table 4.4).

Table 4.5: Summary table:
Percent correct and (AvgAPPi scores)

Model*	Self-trained	Jackknife	NA (1986) data	Combined dataset	x-validated (from NA)
I	92.5(0.866)	85(0.807)	97(0.91)	83(0.768)	57.5(0.557)
II	91.25(0.875)	83.7(0.802)	97(0.92)	85(0.788)	62.5(0.588)
III	91.25(0.869)	82.5(0.791)	98(0.95)	88(0.816)	63.8(0.607)
IV	96.25(0.930)	95(0.905)	99(0.98)		72.5(0.711)

Listeners	%Correct
Mixed	93.81%
Blocked	94.19%

* Model types:

model I = first two formants	[G1,G2]
model II= I + F3	[G1,G2,G3]
model III= II + F0	[G1,G2,G0]
model IV= "CLIH2"(Nearey 1978)	[nG1nG2]

Level Two comparison:
Degree of Correspondence between STATISTICAL RESOLUTION and IDENTIFICATION BY LISTENERS

Whether the information used by the algorithm bears any resemblance to that used by listeners can only be demonstrated by a further comparison. The analysis comparing predictions and observations used in NA (1986) is closely followed. It is expected that if the measures are closely related to the perceptual parameters the APPi scores will be closely correlated to the identification profiles. Product-moment correlations are calculated for observed listener and predicted response profiles for individual tokens. In ANH (1982) the amount of association between the perceptual and APPi values was measured, and it was assumed that if the measures were closely related to the perceptual parameters, then the APPi scores would be closely correlated with the correct identification rates by listener. Although in the ANH study the correlation was only of the correct responses "*a posteriori* probability of group membership for the intended category" (Spearman rank correlation), in the NA paper the comparison was of the entire profile of responses (Pearson correlation). In ANH (for gated vowels) there was a tendency for higher correlations between listener identification data (LID) from the random condition and the APPi analyses on unnormalized measures, while the blocked condition LID data tended to be more highly correlated with the APPi analyses on normalized measures (see Table V in ANH, 1982, p. 984). In this study the attempt was made to provide a condition in which the effects of normalization would be apparent. Following ANH, a range-normalizing condition was simulated by presenting vowels blocked by speaker. In that study it was found that the normalized models' APPi profiles paralleled the listeners' blocked speaker response pattern better than did unnormalized models' profiles.

Confusion matrices are a graphical index of the correspondence between two profiles. Observed and Predicted profiles are plotted in a confusion matrix in figure 4.1. Listener identifications are averaged over all listeners into one "observed" profile for each listening condition (vowel tokens are not averaged as was done in NA, 1986). A prediction profile is calculated from the APPi scores of each algorithm.

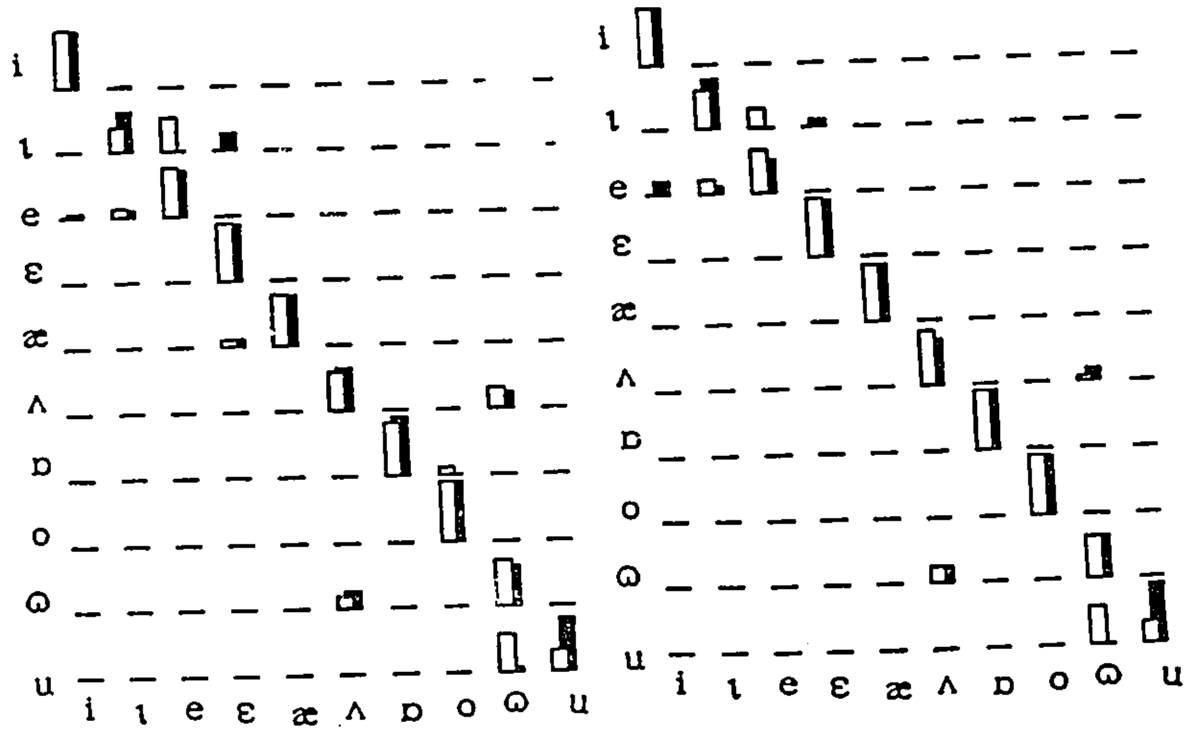


Figure 4.1: Confusion matrices comparing APPi predicted scores (front white bars) with listener condition profiles (shadow dark bars). Matrices on the left are G1G2G3 against random condition. Matrices on the right are nG1nG2 against blocked listener condition. Shown only are the first 10 vowel tokens, the first 10 rows of eighty-row profiles.

Following NA (1986), a raw correlation test (Pearson product-moment) is used as an index of the visual "goodness of fit" (NA, 1986, appendix, p.1306). It is calculated over the 80 token, 800 cell matrices. The model types are determined by the measurements and/or treatment of them prior to analysis. The raw correlation test compares the following four models:

Model I : G1i, G1f, G2i, G2f
Model II : G1i, G1f, G2i, G2f +G0
Model III : G1i, G1f, G2i, G2f +G3
Model IV : nG1i, nG1f, nG2i nG2f

A raw correlation test was run on the normalized and unnormalized models against the listening conditions. The APPi scores of the discriminant analysis are compared with the total response profile of the listeners. The profile shows more than just the binary right or wrong classification, (height of the bars on the diagonal of the confusion matrix) It is the complete information about the rate at which a token is categorized as a member of each vowel category. Following the terminology of ANH (1982) and NA (1986), APP scores are the model's classification *predictions* and listeners' response profiles are *observations.* APP scores are like the responses in that there are APP scores available for every category for each token. "Complete information about the identification of a token is available only from a response profile" (NA, 1986, p.1301). APP scores are easily compared with response profiles since, like response profiles, they have scores for each response category. They are directly compared, at the cell to cell level.

The assumption is made, from the findings of ANH (1982), that the normalized model (Model IV) corresponds to blocked listeners, and the unnormalized minimal two formant model (Model I) corresponds to a random condition (see confusion matrices Fig 4.1). Based on the first level comparison, comparing percent correct and the AvgAPPi scores from the discriminant analysis the choices for matching are not unsupported, but the level of comparison is quite superficial.

The r value of the raw correlation is a better indication of correspondence than just comparing the percent correct identification scores of a prediction and an observation. If both the prediction and the observation were wrong, but the mistakes were entirely of a different sort, the difference would not be reflected at that level of comparison. It is reflected, however, in a correlation test.

Which model best correlates with the random condition? Which model best correlates with the blocked condition? The unnormalized models' APP scores were slightly better correlated with listeners' random responses (0.954 and 0.948) than they were with the listeners' blocked responses (0.951 and 0.944), see Table 4.6 below.

Table 4.6: Raw Correlation tests (p<.001)

	Random profile	Blocked profile
Self-trained		
I) G1i,G1f,G2i,G2f	0.954	0.951
II) G1i,G1f,G2i,G2f,G3	0.955	0.953
III) G1i,G1f,G2i,G2f,G0	0.948	0.944
IV) nG1i,nG1f,nG2i,nG2f	0.974	0.978
Cross-validated		
Xv -I	0.607	0.604
Xv -II	0.637	0.633
Xv -III	0.628	0.626
Xv -IV	0.732	0.732

The normalized APP profile (Model IV) was better correlated with the blocked response profile (0.978) than with the mixed response profile (0.974); the correlation results support the choice of representative models for each of the experimental conditions. The blocked-speaker condition is best represented by a formant range normalization model.

While the blocked condition is slightly more highly correlated with the normalized model than the random condition, the random condition profile is also most highly correlated with the normalized model (Model IV) of all the models tested. That is, both conditions are best correlated with the normalized model. This might be an indication that listeners are not only able to normalize when the speakers are blocked (formant range information) but also in a mixed speaker condition. On the other hand, the high correlation of the random condition with the normalized model may be due partly to the larger diagonal elements. Both the normalized model and the listeners classified at high rates of accuracy, 96% and 94% respectively. The correlation is expected to be high for that reason alone.

Level Three comparison:

Changes in the response profiles

The correlation tests are a way of quantitatively describing similar patterns. In addition to scanning the similarity of patterns between the listeners' two conditions and the two APP predictions (G1G2 vs. nG1nG2) as represented in the confusion matrix bar plots, a difference correlation describes similar patterns of change in a pair of profiles. Was a change that occurred from the random to the blocked condition observations also seen in the unnormalized to the normalized APP predictions? From inspection of the confusion matrices: Fourteen vowels exhibited a marked change (of at least 10%) from listening condition A to listening condition B (where A is the mixed condition and B is the blocked condition). Eleven of those cells were common to the set of cells that showed a change from prediction matrix A to prediction matrix B (where A is the unnormalized model and B is the normalized model). Many more vowels (61) had changed response patterns in the algorithmic APPs, since there was much more room for improvement, and the magnitude of change was usually larger than that seen across the listener profiles for the same reason. The direction of change is also consistent in the two comparisons. Nine of the twelve were categorized more correctly, while two were more poorly classified. The direction of change, whether it was better or worse, is not important, only that whatever direction of change occurred between the listener profiles is the same direction of change reflected in the APP profiles. The criteria for a "marked change" was different for observations and predictions: for observations, a marked change was at least 10 percent, while for predictions it meant at least 8 percent. The higher criteria for listeners' profiles was used in order to avoid reporting the behavior of a single listener. Since the responses from one out of twelve listeners constitutes slightly more than 8 percent of the total responses, 10 percent was chosen as a conservative criterion. The magnitude of changes across cells in this comparison is measured by the difference correlation test.

The difference correlation compares the *change* in a profile across two listening conditions with the change in a profile across two models' predictions. It is a token by token count, that looks at every cell in the 800 cell confusion matrix. The differences for the cells in a pair of matrices are correlated with those difference values for the corresponding cells in another pair of matrices. One pair of matrices is the observed listeners responses, the other is the pair of APP predictions for models correlated (by the raw correlation above) to those listening conditions. "If the model reflects listeners' performance in a detailed way, [it is expected] that changes in listeners' response profiles for the same token in different conditions would be reflected in changes in APP scores in the corresponding conditions" (NA, 1986, p. 1302).

The change (i.e. improvement or worsening) from one condition to the next is compared to the change across prediction profiles. Model II is used as the unnormalized model in this test rather than the minimal Model I, as it was slightly better correlated with the random condition (r = 0.955 as compared to r = 0.954). Moreover, the difference in the correlations is greater when the cross-validated profiles were used: Model II and Model I have r scores with the random condition of 0.637 and 0.607, respectively. The listeners in the blocked condition did better than the listeners in the random condition and Model IV did better than Model II, but was the change across listening conditions similar to the change from Model I to Model IV? To test the significance of the relation between the change across a pair of profiles (listeners) and the change across another pair profiles (APP), correlation analysis was performed between the differences in observed and predicted scores for all pairs of conditions.

This correlation is again a cell by cell analysis. But each cell in a profile is paired this time with the corresponding cell in another profile. So there are four response profiles to consider. Two are the listeners' profiles and two are the predicted responses. Each prediction is matched with the observed profile that it best correlated with (level 2). Because the direction of the change is of the essence, this test is concerned with improvement for a token. For a high correlation to occur the cells that show improvement across listening profiles should also show improvement across the observed profiles and the cells that worsen across the listening profiles should also worsen across the observed profiles. The difference correlation is represented by the following equations:

$Y_{v,t} = LID_{av,t,} - LID_{bv,t}$ and $X_{v,t} = APP_{av,t} - APP_{bv,t}$,where,

v =vowel category index,

t = token index,

$APP_{av,t}$ = *a posteriori* probability for Model IV,

$LID_{av,t}$ = listeners' i.d. rate in blocked listening condition,

$APP_{bv,t}$ = *a posteriori* probability for Model II,

$LID_{bv,t}$ = listeners' i.d. rate in random listening condition

The difference correlation value calculated for $Y_{v,t}$ and $X_{v,t} = 0.275$

The difference correlation between blocked-to-random observed and Model IV to Model II predicted gave an r-value 0.275 ($p < .05$)[12]. The value seems low, but in fact is quite good considering the only place where there was room for change was if there had been any error to begin with for a token, and the correct identification scores were very high. Randomization (correlation) tests provide a *nonparametric* method of testing the significance of the relationship between predictions and observations. The complete set of predicted and observed responses can be observed as four confusion matrices for each speaker.

Summary

In summary, three levels of analysis with a statistical model verify what was hinted but not proved significant with only the human listeners' percent correct scores. That is, the statistical model shows that the improvement in the blocked condition is due to the availability of formant range information. The first level comparison showed that the addition of either G0 or G3 (in a cross-validated run) is helpful in separating categories, but not as effective as a CLIH normalization of the formant frequency values. The second level comparison showed that the response profile from the speaker-blocked condition is better correlated with the normalized model's prediction profile than unnormalized models' prediction. Likewise, the response profile from the mixed speaker condition is better correlated with the unnormalized (than the normalized) models' predictions. Moreover, it

[12] For an idea of the range of correlation encountered in such analyses, consider the Spearman rank correlation coefficients from difference correlations tested in ANH (1982): [0.106. to 0.322] and [-0.039 to 0.260] (Full isolated vowels Table VI, ANH1982, p. 984).

has been demonstrated that the improvement in classification from a speaker-mixed to a speaker-blocked condition is paralleled by a change in predictions of a model given unnormalized formant frequency information and the predictions of the same model given CLIH-normalized frequency information. Although the results of the perceptual test (in the preceding chapter) did not show a significant improvement in the blocked (over the mixed) condition, the comparisons above seem to suggest that there is indeed formant range normalization taking place in the blocked speaker condition, and that the lack of a significant difference in listeners' results (over the two conditions) may have been due to a ceiling effect.

Chapter 5

Conclusion and Discussion

The most salient characteristics of vowel quality (F1 and F2) show variation graphically as overlap of vowel categories. Adolescent voices show acoustic variation in formant frequencies where there is yet little gender distinction of fundamental frequency, leaving the most obvious normalizing factor, F0, unavailable for listeners' use. Identification of adolescent vowels by listeners is not hindered, however, by the lack of correlation between the two independent sources. The perceptual test in Chapter 2 shows that listeners identified the vowels in a mixed speaker condition and a blocked speaker condition with very high rates of identification. At a first approximation, the rates of identification did not differ significantly for the two listening conditions, even though there was a slight improvement in the expected direction. This unexpected negative result is countered by the positive finding in Chapter 3, that changes in response profiles (from mixed to blocked) are correlated in a manner consistent with CLIH normalization.

A problem remains with respect to the cross-validated classification rates of the pattern recognition model. It was shown that the addition of either F0 or F3 information to the F1 and F2 values improved classification rates. Self-trained on only intrinsic information, the model classifies vowels very well (92%). When extrinsic information is added (in the form of normalized formant frequencies) the automatic classification is improved (96%) and even exceeds the identification rate of human listeners (94%). This, together with the results from the difference correlation tests, indicates that the extrinsic information is important for vowel perception, though not exclusively so. When the model was trained on the independent vowel data set of 100 adult vowels, the recognition rates of adolescent vowels was noticeably reduced. While F0 was quite helpful in the model's identification of adult vcwels, it was less so for adolescent vowels. F3 was more helpful than F0 in the identification of self-trained adolescent vowels. The problem that remains is the failure of the cross-validation classification of adolescent vowels with the adult data (rates 57 % to 72%) when the same cross-validating data set has been used to classify other adult vowels more successfully, i.e., the adult vowels of Andruski (1990) were classified 81.4% correctly and vowels from ANH (1982) were classified at much higher rates than even 80 %. Coarticulation effects may be the cause of the discrepancy. The vowels had been produced in a /hVd/ context and although that is usually considered a "neutral" context, the late vowel measures may yet have been influenced by the final /d/. Self-training however, produced successful classifications very close to the listeners rates of classification. It would seem then that the model lacks the ability to generalize from isolated vowel measures to vowel measures from a consonantal context. The model will have to be given more information about the effects of consonantal context before it will be able to cross-validate vowels from different environments. Andruski found the classification of context vowels (57.8%) to be considerably worse than that of isolated vowels (81.4%)[13]. The 81.4% with the Andruski vowels in consonantal context is comparable to the 74 % with the Hillcrest vowels. Therefore it seems that consonantal context effects may be the cause of the poor cross-validation of the Hillcrest vowels.

[13] The consonantal context in Andruski (1990) was /bVb/.

A future study including a comparison of isolated vowels and vowels in context for adolescent speakers is needed.

Human rates of identification under conditions of speaker blocking are best approached by a normalization of formant frequencies. It is shown that a good model of listeners' categorization behavior is calculated by CLIH-type normalization. Normalization formula CLIH2 uses independent speaker factors on each of the two lowest formant frequencies. It categorized natural vowel measures with an accuracy of 95% (jackknife method), better than any unnormalized model of the same data. An identification rate of 95% is also very close to the listeners' identification rate (94.2%). The biggest puzzle is how listeners were able to do as well as they did in the mixed speaker condition (93.8%).

The salient parameters of perception of natural vowels have been determined. To determine how they are used specifically, a more sensitive algorithm modelling listeners' vowel categorization has to be developed. Such an algorithm might include parameters for speaker characteristics to specify speaker gender and/or identity. For now though, this investigation suggests that the intrinsic factors F0 and F3 lend themselves to setting a perceptual vowel space in which the more variable F1 and F2 can be judged for vowel quality. The indirect use of intrinsic information should be incorporated as a step in the process of perception which permits the adjustment of parameter weightings according to speaker identity.

If the speaker normalizing scale factor is indeed a consequence of a speaker identification, that speaker identity information has to be available to the listeners even in a single syllable. The speaker characteristics responsible for this accurate mapping of formant space, that contribute so generously to the scale factor, may be the same ones that are used secondarily to cue speaker sex. Formant frequency differences were enough to allow listeners to accurately perceive speaker gender in F0-neutralized speech (Coleman, 1973, 1976). Formant frequency differences are found in preadolescent vowels of boys and girls that are much greater than the F0 differences. Gender differences in intonation patterns (Tielen,1989), spectral tilt, and even glottal waveform shape (Carrell,1984; Klatt & Klatt, 1989) exist, none of which were addressed in the present study. Johnson's and Carrell's studies have shown that speaker identification can be attributed to cues such as these. The speaker scale factor may not necessarily be determined solely on a gender basis. It certainly may be more specific than that. It would take a series of further tests to convincingly demonstrate the link between speaker identification and vowel perception. To begin with it would be necessary to procure general size and gender percepts from listeners as they identify vowels. Then those *impressions of speaker* would have to be shown to influence the categorization of the vowels. Even if listeners do formulate an impression of speaker size and/or gender, it would be difficult to know where to begin quantifying such impressions, for I suspect an identity perception is more complex than binary gender (male speaker/female speaker) and size (big speaker/small speaker) alternatives.

Bibliography

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

Andruski, J. E. (1990). Perceptual Information for Vowel Identification. Master's Thesis, University of Alberta, Edmonton, Alberta.

Assmann, P. F. (1979). The Role of Context in in Vowel Perception. Master's Thesis, University of Alberta, Edmonton, Alberta.

Assmann, P. F., Nearey, T. M., & Hogan, J. T. (1982). Vowel identification: Orthographic, perceptual and acoustic aspects. Journal of the Acoustical Society of America, 71, 975-985.

Carrell, T. D. (1984). Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification. Doctoral Dissertation, Indiana University, University Microfilms International.

Chiba, T., &. Kajiyama, M. (1941). The Vowel: Its Nature and Structure. Tokyo: Tokyo Publishing.

Coleman, R. O. (1973). Speaker identification in the absence of inter-subject differences in glottal source characteristics. Journal of the Acoustical Society of America, 53, 1741-1743.

Coleman, R. O. (1976). A comparison of the contribution on two voice quality characteristics to the perception of maleness and femaleness in the voice. Journal of Speech and Hearing Research, 19, 168-180.

Dechovitz, D. (1977). Information conveyed by vowels. Haskins Laboratory Status Report of Speech Research, SR-51/52, 213-219.

Edgington, E. (1980). Randomization Tests. New York: Dekker.

Fant, G. (1966). A note on the vocal tract size factors and nonuniform F-pattern scalings. Royal Institute of Technology Speech Transmission Laboratory Quarterly Progress and Status Reports, 66(4), 22-30.

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

Gerstman, L. (1968). Classification of self-normalized vowels. IEEE Transactions on Audio and Electroacoustics, AU-16, 78-80.

Gottfried, T. L., & Chew, S. (1986). Intelligibility of vowels sung by a counter-tenor. Journal of the Acoustical Society of America, 79, 124-130.

Gray, H. L., & Schucany, W. R. (1972). The Generalized Jackknife Statistic. New York: Dekker.

Hand, D. H. (1981). Discrimination and Classification . Chichester: Wiley.

Jamieson, D. J., Ramji, K. V., Nearey, T. M., & Baxter, T. A. (1990). CSRE 3.0 Canadian Speech Research Environment Users Manual. London, Ontario: Communication Lab, Department of Communicative Disorders.

Joos, M. (1948). Acoustic phonetics. Language, Suppl. 24,1-136.

Johnson, K. (in press). F0 normalization and adjustment to talker. Journal of the Acoustical Society of America.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. Journal of the Acoustical Society of America, 87, 820-857.

Ladefoged, P. A (1982). A Course in Phonetics . (2nd ed.). New York: Harcourt Brace Jovanovich.

Ladefoged, P., & Broadbent, D. (1957). Information conveyed by vowels. Journal of the Acoustical Society of America, 29, 98-104.

Lehiste, I. (1970). Suprasegmentals. Cambridge, MA.: MIT Press.

Lobanov, B. (1971). Classification of Russian vowels spoken by different speakers. Journal of the Acoustical Society of America, 49, 606-608.

Merino, S. R., & Nearey, T. M. (1989). Fundamental and formant frequencies in an adolescent population. Journal of the Acoustical Society of America, 85(S1), S54.

Miller, J. D. (1989). Auditory-perceptual interpretation of the vowel. Journal of the Acoustical Society of America, 5, 2114-2134.

Nearey. T. M. (1978). Phonetic Feature Systems for Vowels . Bloomington, IN.: Indiana University Linguistics Club.

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

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

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

Potter, R., & Steinberg, J. (1950). Toward the specification of speech. Journal of the Acoustical Society of America, 22, 807-820.

Slawson, A. W. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. Journal of the Acoustical Society of America, 43, 87-101.

Syrdal, A. K., & Steele, S. A. (1985). Vowel F1 as a function of speaker fundamental frequency. Journal of the Acoustical Society of America, 78(Suppl 1), S56.

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

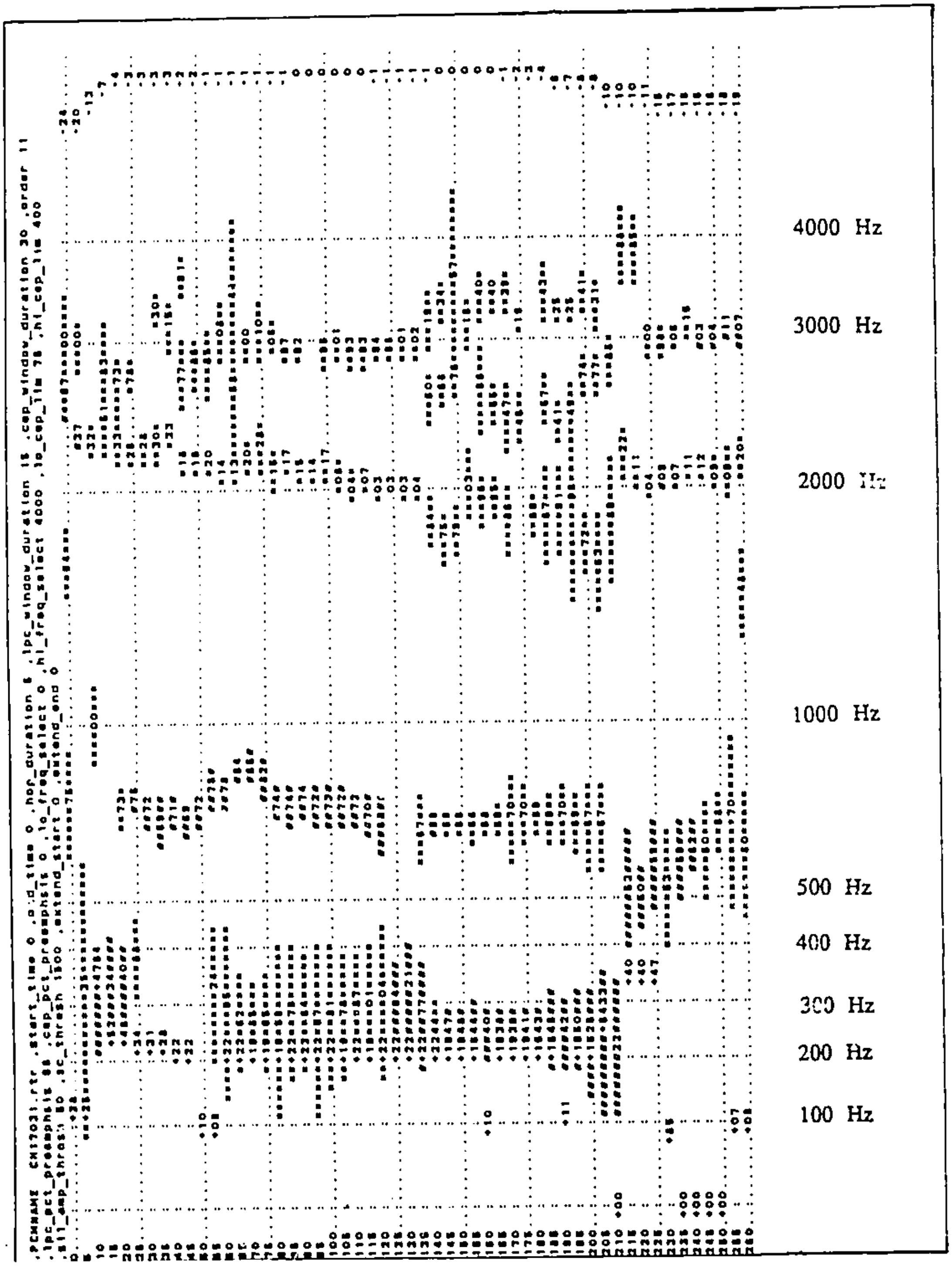
Tielen, M. T. (1990). Fundamental frequency characteristics of middle aged men and women. Proceedings: Institute of Phonetic Sciences,University of Amsterdam, 13, 45-58.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. Journal of the Acoustical Society of America, 69, 1465-1474.

Traunmüller, H. (1988). Paralinguistic variation and invariance in the characteristic frequencies of vowels. Phonetica, 45, 1-29.

Appendix

The following is a sample graphic display of the signal analysis from which the acoustic measurements were taken. Time values (ms) are shown in the bottommost line of numbers. Pitch values (Hz) are preceded by a plus (+) sign. The frequency values (Hz) of formant candidates are displayed with amplitude and bandwidth information as follows: a pound (#) indicates a high amplitude and an equal (=) sign indicates a middle level amplitude; bandwidth is indicated by the approximate location (in the frequency scale) of amplitude characters.



,PCMNAME CH17031.rtr ,start_time 0 ,end_time 0 ,hop_duration 5 ,lpc_window_duration 15 ,cep_window_duration 30 ,order 11
,lpc_pct_preemphsis 88 ,cep_pct_preemphsis 0 ,lo_freq_select 0 ,hi_freq_select 4000 ,lo_cep_lim 75 ,hi_cep_lim 400
,sil_amp_thresh 50 ,sc_thresh 1500 ,extend_start 0 ,extend_end 0