National Library of Canada
Bibliothèque nationale du Canada

Ottawa, Canada
K1A 0N4

17

TC –

0-315-22883-0

CANADIAN THESES ON MICROFICHE SERVICE – SERVICE DES THÈSES CANADIENNES SUR MICROFICHE

PERMISSION TO MICROFILM – AUTORISATION DE MICROFILMER

• Please print or type – Écrire en lettres moulées ou dactylographier

AUTHOR – AUTEUR

Full Name of Author – Nom complet de l'auteur

PETER FRANK ASSMANN

Date of Birth – Date de naissance 04 - 08 - 1953	Canadian Citizen – Citoyen canadien ☒ Yes / Oui ☐ No / Non
Country of Birth – Lieu de naissance Netherlands	Permanent Address – Résidence fixe MRC Institute of Hearing Res. University of Nottingham Nottingham England NG7 2RD

THESIS – THÈSE

Title of Thesis – Titre de la thèse

THE Role of harmonics' and formants in the perception of vowel quality.

Degree for which thesis was presented Grade pour lequel cette thèse fut présentée Ph.D	Year this degree conferred Année d'obtention de ce grade 1985
University – Université University of Alberta	Name of Supervisor – Nom du directeur de thèse Dr. T. Nearey

AUTHORIZATION – AUTORISATION

Permission is hereby granted to the NATIONAL LIBRARY OF CANADA to **microfilm** this thesis and to **lend** or **sell copies of the film.**

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.

L'autorisation est, par la présente, accordée à la BIBLIOTHÈQUE NATIONALE DU CANADA de **microfilmer** cette thèse et de **prêter** ou de **vendre des exemplaires du film.**

L'auteur se réserve les autres droits de publication; ni la thèse ni de longs extraits de celle-ci ne doivent être imprimés ou autrement reproduits sans l'autorisation écrite de l'auteur.

ATTACH FORM TO THESIS – VEUILLEZ JOINDRE CE FORMULAIRE À LA THÈSE

Signature Peter Assm	Date August 8 1985

Canada

THE UNIVERSITY OF ALBERTA

THE ROLE OF HARMONICS AND FORMANTS IN THE PERCEPTION OF VOWEL QUALITY

BY

© PETER F. ASSMANN

A THESIS

SUBMITTED TO THE FACULTY OF GRADUATE STUDIES AND RESEARCH IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF:

DOCTOR OF PHILOSOPHY

IN

SPEECH PRODUCTION AND PERCEPTION

DEPARTMENT OF LINGUISTICS
EDMONTON, ALBERTA, CANADA

Fall, 1985

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR Peter F. Assmann

TITLE OF THESIS The role of harmonics and formants
in the perception of vowel quality

DEGREE FOR WHICH THESIS WAS PRESENTED
Doctor of Philosophy

YEAR THIS DEGREE GRANTED 1985

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.

(SIGNED) Peter As

PERMANENT ADDRESS:
MRC Institute of Hearing
Research, University of
Nottingham, Nottingham
ENGLAND NG7 2RD

DATED August 1 1985

THE UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled

The role of harmonics and formants in the perception of vowel quality .... submitted by

Peter F. Assmann .... in partial fulfilment of the requirements for the degree of

Doctor of Philosophy .... in

Speech Production and Perception ....

Terrance M. Nearey...

Supervisor

A Rozsypal..........

B Rochet..........

John T. Hogan...

Steven Greenberg.......

External Examiner

Date...June 28, 1985.......

To the memory of my father, Oswald J. Assmann and to the memory of my father-in-law, Morris E. Calman

ABSTRACT

This study investigates several aspects of the formant hypothesis of vowel perception, which states that listeners identify vowels by estimating the frequencies of the lowest 2 or 3 formants. Matching and identification experiments were carried out to investigate the role of individual harmonic components in the perception of height in front vowels. It is shown that vowel height is not determined solely by the frequency of the most prominent harmonic, but depends on a weighting of harmonics near the first formant peak. Synthetic vowel stimuli were constructed with a band of 2-5 harmonics of equal amplitude substituted for F1. When listeners were asked to find the best match from an F1 continuum, they selected F1 values near the highest frequency harmonic in the band. Attenuation of the highest frequency harmonic resulted in <u>lower</u> matched F1 values, while attenuation of the lowest harmonic had little effect, except in the 2 harmonics condition where <u>higher</u> F1 matches were observed. Matching experiments with a pair of harmonics in the F1 region showed that the effects of attenuation are asymmetrical: the higher frequency harmonic appears to carry greater weight than the lower component. The effects of attenuation were largely independent of changes in harmonic rank (harmonics 1 to 4) and fundamental frequency (125 and 250 Hz). The data were modelled fairly well by a local centre of gravity measure which computes the weighted mean of the 2 most prominent harmonics in the F1 region of the preemphasized spectrum; very similar results were obtained using LPC estimates of F1. In an identification experiment, synthetic front vowels along an F1 continuum were presented in various filter conditions which removed harmonics above or below the 2 most prominent. The identification functions were not substantially altered by these manipulations, indicating that the 2 most prominent harmonics play an important role in vowel height. The second part of the study examined the perception of back vowels with close spacing of F1 and F2. It has been proposed that 2 formants in close proximity are not separately resolved, and that back vowel quality depends on the centre of gravity of the formant cluster. It is shown in a matching experiment that formant frequencies and formant amplitudes do not trade off in the manner suggested by the formant centre of gravity hypothesis. Formant amplitude was shown to have a small effect on the identification of synthetic vowels along an F1-F2 continuum. This effect was reliably present in a high pitch condition (f_0=250 Hz) but not in a low pitch condition (f_0=125 Hz). Contrary to predictions of the formant centre of gravity hypothesis, formant amplitude effects were not restricted to vowels with close spacing of F1 and F2; in fact, the largest effects were present in stimuli with the widest separations of F1 and F2. It is suggested that listeners may be able to

infer the frequencies of closely spaced formants in back vowels from other cues such as relative formant amplitudes, overall spectral balance or the slope of the spectrum above the F1-F2 cluster.

ACKNOWLEDGEMENTS

I am grateful to the faculty members and graduate students of the Department of Linguistics who provided encouragement and support during the preparation of this dissertation.

My thesis supervisor, Dr. T. Nearey, deserves my very special vote of thanks: our discussions were a constant source of inspiration for me. His guidance and help with computer programming, data analysis and planning of experiments is gratefully acknowledged.

I would like to thank the external examiner of the thesis committee, Dr. S. Greenberg, for many interesting comments and suggestions, and the other members of the committee, Drs. A. Rozsypal, B. Rochet and J. Hogan for their careful reading of the dissertation and for their helpful suggestions and comments.

Special thanks to those who volunteered as listeners in the experiments. I truly appreciate the many hours that were sacrificed in the phonetics laboratory. My colleagues Maureen Dow and Tracey Derwing provided generous assistance in the editing and revision of the text; Shaunie Shammass and Murray Munro also provided assistance in many ways.

My greatest debt is to my wife Albi, whose patience, support and encouragement helped me more than words will ever express.

Part of this research was supported by a Doctoral Fellowship from the Social Sciences and Humanities Research Council.

TABLE OF CONTENTS

PAGE

CHAPTER ONE. INTRODUCTION

CHAPTER TWO. AUDITORY CODING OF VOWEL SOUNDS

CHAPTER THREE. A PERCEPTUAL STUDY OF FRONT VOWELS

CHAPTER FOUR. A PERCEPTUAL STUDY OF BACK VOWELS

CHAPTER FIVE. SUMMARY AND CONCLUSIONS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

1.1 Scope of the present study

Speech perception is an intricate pattern recognition process, which in normal communication takes place against a background of competing signals and noise. The perceptual system is able to partition the complex acoustic structure of the speech waveform to extract phonetic information (necessary for speech comprehension) and non-phonetic information (e.g., individual speaker characteristics and properties of the environment, such as the distance between the speaker and the hearer, and the presence of noise or simultaneous speech). This research focusses on phonetic quality in vowel sounds and the acoustic properties which enable listeners to distinguish among them.

The acoustic properties of natural speech are "dynamic", changing constantly in formant pattern and fundamental frequency (Potter & Steinberg, 1950). These changes alter the short term amplitude spectrum in a manner similar to simultaneous frequency and amplitude modulation of sinusoidal components. In vowels these changes take place at a sufficiently slow rate that a steady state description of their acoustic properties is possible. Vowel segments represent regions of relative stability in the speech stream.

Speakers can readily produce sustained vowels in isolation, and compare and contrast them in terms of steady state properties. This capacity is employed in phonetic transcription and introductory phonetics training (Jones, 1956; Ladefoged, 1975). Speech synthesis studies in several languages have demonstrated that steady state approximations to monophthongal vowel sounds are readily labelled by listeners in terms of the categories of their native language (e.g., Delattre, Liberman, Cooper & Gerstman, 1952; Carlson, Granstrom & Fant, 1970).

There are some indications, however, that a characterization of vowel quality in terms of static 'target' positions may not be fully adequate. For example, isolated vowels (including some ordinarily classified as monophthongs) may exhibit dynamic properties (change in the formant pattern over time) as a result of diphthongization (Joos, 1948). Additional information is provided by duration differences among the vowels (e.g., Bennett, 1968; Ainsworth, 1972). When vowel duration and diphthongization differences were neutralized in naturally spoken isolated English vowels (by excising a portion of the waveform), a significant increase in the number of identification errors was found. However, the majority of the vowels were still identified correctly (Assmann, Nearey & Hogan, 1982).

It was found that two brief sections from the initial and final portions of the waveform, separated by a 10 ms silent interval, provided sufficient information for accurate identification of isolated vowels (Nearey & Assmann, 1984). These results indicate that a simple representation of vowel quality in terms of steady state properties may be adequate for isolated vowels.

The present study investigates the perception of isolated oral vowel sounds, presented individually or in pairs to listeners in the absence of background noise. Issues such as the perceptual effects of other sources of acoustic variability, including speaker or voice quality differences, the dynamic changes which take place over time, and the influence of segmental or suprasegmental context and speaking rate will not be addressed here. Although a full account of vowel perception must deal with these additional factors, it is clear that the study of isolated vowel sounds can reveal basic perceptual processes underlying normal speech understanding.

1.2 Phonetic classification of vowel quality

The traditional classification of vowel quality has employed a set of phonetic features by which vowel sounds may be compared and contrasted. Feature systems have provided important insights in vowel phonology and have helped to reveal the patterning and symmetry underlying phonological processes and sound change. Experimental studies of vowel similarity judgments have indicated a close relationship between phonetic features and the perceptual space derived from multidimensional scaling analysis (e.g., Hanson, 1967; Singh, 1974; Terbeek, 1977; Fox, 1983).

It is generally acknowledged that the features of vowel 'height' and 'advancement' constitute the principal dimensions for classifying the English vowel phonemes[1]. Although these features were originally defined in terms of the vertical and horizontal placement of the tongue durihg vowel articulation, this interpretation has been found to be unsatisfactory (Russell, 1928; Ladefoged, DeClerk, Lindau & Papçun, 1972; Lieberman, 1976; Nearey, 1977). There is, however, a close correspondence with the frequencies of the two lowest formants (Joos, 1948). Vowel height is negatively correlated with the frequency of the first formant: high vowels (such as /i/) have low F1 values, while low vowels (such as /&ae/) have high F1 values. Advancement is correlated with the frequency of the second formant: front vowels (e.g., /e/) have a high F2, while back vowels

[1] Two additional features, 'tenseness' and 'rounding', which are also frequently used to describe vowel contrasts in English, will not be investigated here. There is some question as to whether these features consitute independent dimensions of vowel quality (Nearey, 1980). Neither appears to have acoustic correlates in steady state vowels which are independent of the parameters specifying vowel height and advancement. Some authors have regarded these features as primarily articulatory in nature (Ladefoged, 1975).

(e.g., /o/) have a low F2. The terms height and advancement will be used here to describe aspects of vowel quality associated with F1 and F2.

1.3 Overview

This study is an investigation of the acoustic and auditory factors underlying the perception of phonetic quality differences in vowels, in particular vowel height. A survey of the literature on the auditory coding of vowel sounds is given in chapter two. The chapter begins with a discussion of the acoustics of speech production and the formant representation of speech. It is well established that vowel quality depends primarily on the frequencies of the lowest two or three formants. However, the perceptual grounds for this association have not been extensively studied.

Several recent studies of vowel perception have emphasized the importance of auditory frequency analysis. Attempts have been made to model vowel perception by simulating the transformations which take place in the auditory system. Some of the neurophysiological and psychophysical data on which these models are based will be reviewed and the application of these models in the study of vowel perception will be considered.

In the second part of chapter two, several alternative hypotheses concerning the pattern recognition process in vowel perception are evaluated. The evidence is strongly in favour of the formant hypothesis, which states that listeners are able to identify vowels by estimating the frequency values of the lowest two or three formants. There are, however, some unresolved problems for this hypothesis.

The first problem concerns the recovery of formant locations from the vowel waveform by the auditory system. Studies of auditory frequency resolution have indicated that the frequency analysis performed by the hearing system may be capable of resolving individual harmonics throughout the frequency range of the first formant. Since there is generally no energy physically present at the formant frequencies, it is of interest to determine how formant information is coded by the auditory system on the basis of individual harmonics in the perception of vowel height contrasts.

Several hypotheses about the contribution of individual harmonics to the perception of vowel height are considered in chapter three. These proposals are evaluated by means of a series of experiments employing vowel matching and vowel identification tasks. Strong experimental evidence is presented against the hypothesis that vowel height is determined exclusively by the frequency of the most prominent harmonic in the first formant region. Instead, the results suggest a formant estimation procedure based on a weighting of prominent harmonics in the first formant region. When several alternative heuristic procedures were compared, it was found that the best predictions of the matching data were obtained using procedures which also gave the most accurate estimates of F1 in synthetic vowels.

In chapter four the problem of closely spaced formants is addressed. In back vowels, F2 is often very close to F1, with the result that only a single peak appears in the low frequency region of the amplitude spectrum. According to the formant hypothesis of vowel perception, the auditory system can infer the frequency locations of F1 and F2 even when the formants are in close proximity. If the line spectrum of a back vowel has only a single peak in the F1-F2 region, a simple weighting of prominent harmonics cannot be used to estimate the positions of both formants. However, there may be two distinct peaks in the auditory representation of the vowel, as a result of sharpening or enhancement of spectral contrast. Moreover, other cues such as the presence of a small 'shoulder' in the frequency region of the second formant, formant amplitude relationships, and spectral balance may help to specify the locations of the two formants.

A modification of the formant hypothesis was recently proposed by Chistovich and Lublinskaya (1979) to account for the perception of vowels with closely spaced formants. They proposed that vowel quality depends on the spectral centre of gravity in the F1-F2 region when F1 and F2 are separated by less than 3-3.5 Bark. With larger separations of F1 and F2, vowel quality is determined by the frequencies of the two formants and is independent of their amplitudes. This 'formant centre of gravity' hypothesis predicts an interaction between formant frequency and formant amplitude in vowel stimuli with closely spaced formants.

Chapter 4 presents the results of two experiments designed to evaluate the formant centre of gravity hypothesis. The first of these experiments investigated the effects of formant amplitude on back vowel matching. While changes in formant amplitude did lead to some shifts in matching, the overall pattern of results was not consistent with the formant centre of gravity hypothesis. The presence of a multimodal matching pattern for some stimuli suggested a conflict between formant frequency and formant amplitude cues rather than a tradeoff as suggested by Chistovich and Lublinskaya (1979).

The second experiment used vowel identification responses to further evaluate the formant centre of gravity hypothesis and to assess the importance of a 3-3.5 Bark separation of the formants. Formant amplitude manipulations resulted in small shifts of the identification functions. However, these effects were somewhat variable; they appeared only in a high pitch condition, and did not follow the predictions of a model based on the formant centre of gravity hypothesis. Formant amplitude effects were not restricted to vowels with close spacing of F1 and F2; in fact, stimuli with greater than 3.5 Bark separation of F1 and F2 were affected more by formant amplitude manipulations than those with closely spaced formants.

The experimental evidence presented in chapter four suggests a multi-cue hypothesis of back vowel perception. The locations of the formants appear to be important for the perception of vowel quality, even when F1 and F2 are close together. It appears that formant amplitude can affect phonetic quality in back vowels, but its influence is small relative to that of the formant frequencies.

Since formant amplitudes can be predicted from a knowledge of the formant frequencies (Fant, 1956), the perceptual system may take advantage of these constraints in estimating the frequencies of closely spaced formants. In more general terms, a change in formant frequency produces a redundant pattern of correlated changes throughout the amplitude spectrum. The perceptual system may analyze the entire pattern of these changes, group them in terms of coherence or consistency, and code them as a formant change (Haggard, 1978). In this view, the presence of a local maximum in the spectrum provides only one clue to the frequency of a formant; there are a number of others.

A formant estimation procedure based on a weighting of prominent harmonics can account for the results of the front vowel experiments. Other factors appear to be involved in the perception of back vowels. Further research is needed to identify the spectral cues which determine phonetic quality in back vowels.

CHAPTER TWO

AUDITORY CODING OF VOWELS

2.1 Introduction

It is often suggested that the perception of speech sounds engages specialized decoding mechanisms. The sequence of articulatory movements involved in speech production has complex acoustic consequences, with the result that most phonetic distinctions have multiple acoustic correlates (Repp, 1983). This complexity, coupled with the apparent lack of acoustic invariance across different phonetic contexts, has led some speech researchers to postulate a link between speech perception and speech production. Liberman and Studdert-Kennedy (1978) suggest that listeners may perceive speech sounds in terms of their implicit knowledge of articulatory-acoustic correspondences. According to this view, diverse acoustic properties are grouped together in perception by virtue of their common origin in a single articulatory configuration or gesture.

Vowel sounds are often characterized in terms of their <u>formant patterns</u>. The formant representation is based on the source-filter theory, which characterizes the relationship between the acoustic properties of speech sounds and speech production. A brief review of this theory is given in section 2.2.

It is well established that vowel quality is dependent on the frequencies of the lowest two or three formants. According to the formant hypothesis of vowel perception, vowel identification involves a perceptual estimation of the frequencies of the two or three lowest formants. This hypothesis suggests a close link between speech production and perception, although it does not entail that vowel perception is mediated by a knowledge of articulatory-acoustic correspondences.

An alternative approach to the study of vowel perception is based on the assumption that speech perception does not employ specialized mechanisms, but makes use of the same general auditory processes that are involved in the recognition and discrimination of non-speech sounds (e.g., Schouten, 1980). A basic premise of the present work is that the operation of both speech-specific and general auditory factors should be considered in the investigation of vowel perception. This includes basic auditory phenomena such as masking and higher-level cognitive factors involving listeners' implicit 'knowledge' of the common acoustic properties of physically different productions of the same vowel.

It has been suggested that a simulation of the transformations which take place at lower levels of the auditory system may help to reveal the basic information-bearing elements of vowel sounds. In section 2.3 a review is given of psychophysical and neurophysiological studies of auditory frequency analysis and the application of auditory models to the study of vowel perception is discussed.

Several hypotheses concerning the pattern recognition process in vowel perception are considered in section 2.4. Three main proposals are discussed: a "whole spectrum" template matching process; a statistical model based on the extraction of "spectral shape" factors; and the formant hypothesis. The evidence is in favour of the formant hypothesis, although few studies have directly investigated the perceptual basis for formant frequency estimation.

2.2 Source-filter theory of vowel production

Vowel production can be described as a sequence of coordinated movements. Airflow from the lungs is modulated as it passes through the glottis by the opening and closing of the vocal folds. This action results in a sequence of quasi-periodic pulses whose rate determines the voice fundamental frequency. The glottal air pressure waveform is subsequently modified as it passes through the vocal tract, a non-uniform tube whose configuration depends on the positions of the articulators, particularly the lips and the tongue. The source-filter theory (Fant, 1960; Flanagan, 1972) characterizes vowel sounds in terms of the output of a linear system (the vocal tract) in response to an excitation source (quasi-periodic glottal pulses in the case of voiced sounds, random noise for unvoiced sounds; or a mixture of the two). The effects of radiation at the lips, and the optional presence of a side branch resonator (in nasal sounds) are also taken into account.

The amplitude spectrum of a vowel can be characterized in terms of <u>harmonic fine structure</u> (reflecting the contribution of the glottal source) and <u>formant pattern</u> (reflecting the contribution of the vocal tract filter). Strictly speaking, the term 'formant' refers to a resonance or natural mode of vibration of the vocal tract. However, it is also sometimes used to describe the properties of a local maximum in the spectrum envelope of a vowel.

The frequencies of the formants do not necessarily coincide with harmonics, a consequence of the independence of source and vocal tract transfer function (Fant, 1960). Formant frequencies can be estimated by determining the locations of peaks in the spectrum envelope. The spectrum envelope is extracted from the speech signal by means of interpolation procedures based on LPC or cepstral analysis (Rabiner and Schafer, 1978). At high fundamental frequencies, there are fewer harmonics in each of the formant regions; vowel formant structure may be poorly specified, with a corresponding increase in errors of formant mea ment. In natural speech, however, additional energy may be present in the formant regions as a result of a changing fundamental frequency contour or the presence of aperiodic glottal source components.

Where necessary, a distinction will be made between nominal formant frequency (the actual resonance frequency, or the value specified for a formant in producing a synthetic vowel) and estimated formant frequency, the measured value determined by any one of a number of speech analysis procedures.

The formant frequencies of vowel sounds characterize their essential physical structure in a fairly accurate and compact manner, and provide a reliable method for describing differences between vowels (Peterson, 1951; Peterson and Barney, 1952). Experimental and theoretical studies of vowel production have indicated a number of relationships among the acoustic parameters. The spectrum envelope of a vowel can be calculated once the formant frequencies, formant bandwidths and glottal source characteristics are known (Fant, 1956). The relative amplitudes of the formants are thus predictable.

These built-in correlations among the acoustic parameters account for a large proportion of the redundancy in vowel spectra. It has been suggested (e.g., Stevens &

House, 1961) that the relationships between formant frequencies and amplitudes make it inappropriate to manipulate these parameters independently in perceptual experiments with synthetic speech. The relationship between formant frequencies and formant amplitudes in natural vowels has been incorporated in the design of cascade formant synthesizers: formant amplitudes cannot be independently manipulated. However, the possibility that formant amplitude may have an independent effect on vowel perception cannot be ruled out *a priori*. It will be argued below that the perception of synthetic speech-like sounds which are not physically realizable by a human vocal tract may provide important insights into natural speech perception.

2.3 Auditory frequency analysis and vowel perception

The most complete specification of the acoustic structure of vowel sounds is in terms of the time-varying sound pressure waveform. However, it is generally agreed that the principal acoustic differences among vowels are reflected most directly in their short-time frequency spectra, using a time window ranging from 3 to 30 ms. Although speech is an inherently dynamic process, the articulators move fairly slowly, especially during vowel production, and thus the formant pattern is relatively stable over this time interval.

A number of procedures have been proposed for the frequency analysis of vowels; the traditional and most widespread method in phonetics is the sonagraph or sound spectrograph (Koenig, Dunn and Lacy, 1946). The design features of the spectrograph reflect certain aspects of the frequency analyzing properties of the auditory system. For example, the spectrograph performs a short time spectral analysis, allowing for the measurement of fundamental frequency and formant information believed to be important in vowel perception.

The assumption of an auditory basis for spectrographic analysis of speech has been made explicit by Studdert-Kennedy (1976: p.245):

> "We may imagine, therefore, an early stage of the auditory display, soon after cochlear analysis, as the neural correlate of a spectrogram."

In order to determine how realistic this assumption is, it will be necessary to make a brief excursion into psychophysics and neurophysiology.

The spectrograph provides a display of the short term amplitude spectrum over time, discarding phase information. The amplitude dimension is represented in terms of a gray scale with fairly limited resolution. In contrast with the narrow dynamic range of the spectrograph, human speech perception employs a range exceeding 100 dB (Pollack and Pickett, 1958). The spectrograph uses a constant absolute bandwidth filter of continuously varying centre frequency to scan the signal. Other standard speech analysis procedures based on the Fast Fourier Transform (FFT analysis) and linear predictive coding of speech (LPC analysis) also employ constant bandwidth analysis. Studies of the frequency analyzing ability of the ear have shown that auditory frequency resolution is not constant but increases in proportion to frequency (for a review, see Plomp, 1976; Flanagan, 1972).

Either a narrow band filter (45 Hz bandwidth) or a wide band filter (300 Hz bandwidth) is commonly used in spectrographic analysis. The narrowband filter provides a detailed frequency analysis, displaying the individual harmonics as a series of horizontal lines as a function of time. When the wideband filter is applied to low pitched voices, these horizontal lines are "merged" to form dark bars indicating the presence of formants.

The narrowband filter, which provides better frequency resolution, will have a longer rise and decay time (and hence poorer temporal resolution) than the wideband filter. It has been suggested that the narrowband filter may not be capable of accurate measurements of rapidly changing acoustic properties such as bursts and formant transitions associated with consonants (Searle, Jakobson and Rayment, 1979). Limited temporal resolution is probably less important for the analysis of vowel sounds, which are characterized by a spectral pattern which is fairly stable within the time window of the wideband filter generally used in formant measurement.

A classic problem in spectrographic analysis is the selection of an analysis bandwidth which is not affected by differences in fundamental frequency and formant ranges among speakers. Spectrographic displays of a vowel sound produced by men, women and children show radically different characteristics. For example, a wideband filter (300 Hz bandwidth) will "merge" the individual harmonics of an adult male voice so that a formant shows up as a single horizontal dark band; while a child's voice may show a "splitting" of the band into individual narrow lines corresponding to the harmonics.

Klatt (1982b) addressed this problem and concluded that (1) broader bandwidths are not justifiable on the basis of evidence from studies of auditory frequency resolution; (2) broader bandwidths will frequently merge two formants in close proximity, as in some back vowels. In subsequent sections we will review psychophysical and neurophysiological evidence which indicates that the auditory system is capable of resolving the lower harmonics of vowels over a wide range of fundamental frequencies. It will be shown that the auditory system does not employ a a constant bandwidth analysis as in the spectrograph; a better characterization of auditory frequency analysis is provided by a "constant-Q" filterbank which provides a more detailed frequency analysis in the low frequency region and better temporal resolution at high frequencies (Youngberg and Ball, 1978; Searle, Jakobson and Rayment, 1979).

2.3.1 Psychophysical studies of auditory frequency resolution

One of the first systematic studies of auditory frequency resolution was conducted by Helmholtz (1863) who proposed that each of the sinusoidal components of a complex harmonic tone has a distinct pitch corresponding to its frequency. He conducted a series of experiments on the ability of listeners to distinguish the component tones of simultaneously sounded tuning forks and musical instruments. His observations suggested that the ear carries out a frequency analysis allowing the listener to identify a number of distinct pitches; this capacity is limited by the presence of other components, and it is relatively insensitive to changes in phase.

Recent psychophysical research has shown that the limitations of frequency analysis as revealed by masking studies, and the limits on the ability to identify individual harmonic components are closely related, and can be characterized in terms of the concept of critical bandwidth, a psychophysical measure of auditory frequency resolution (Scharf, 1970; Plomp, 1976). A number of studies have investigated the sensitivity of the ear to changes in phase relationships, and some of these have indicated that the timbre of complex sounds (including vowels) may be affected by such changes. In subsequent sections we will consider these findings and their implications for the perception of vowel quality.

2.3.2 Phase sensitivity

Changes in the short term phase spectrum appear to have relatively minor effects on the intelligibility of speech sounds. Presentation of speech in a reverberant sound field (which modifies both the phase and amplitude spectrum) has little effect on intelligibility. However, a number of studies have demonstrated that phase changes in complex tones may have an effect on timbre (Zwicker, 1952; Licklider, 1957; Schroeder, 1959).

Plomp and Steeneken (1969) obtained dissimilarity judgements for steady state vowels. Judgements of vowel timbre were affected much more by changes in the amplitude spectrum than they were by changes in the phase configuration. The maximum effect on vowel timbre due to phase changes was similar in magnitude to the distance between two vowels with similar formant patterns. They speculated that phase effects might be large enough to change one vowel into another. However, this conclusion is probably not justified if phase effects alter only non-phonetic aspects of timbre such as roughness, as indicated by other studies.

Carlson, Granstrom and Klatt (1979) reported a harsh quality when the initial phases of steady state vowels were randomized. They found, however, that listeners rated the "psychoacoustic" or overall dissimilarity in pairs of vowels differing in phase as much larger than their "phonetic" dissimilarity - the latter based on perceived changes in vowel quality. The effects of deleting individual harmonic components from synthetic vowels was studied by Kakusho, Nakashima, Yanagida and Mizoguchi (1983). The removal of harmonics frequently resulted in changes in vowel timbre; however, these effects were not affected by changes in the phase spectrum (constant versus random phase).

Taken together, these studies indicate that phase is relatively unimportant in the perception of vowel quality. There is evidence, however, that the phase spectrum has an important influence on the naturalness and quality of synthetic speech (Schroeder & Mehrgardt, 1982; Atal & Remde, 1982).

2.3.3 Discrimination of components in complex tones

Psychophysical studies have shown that the auditory system can resolve individual harmonics of complex tones in the low frequency region. Plomp (1964) and Plomp and Mimpen

(1966) presented listeners with complex tones composed of 12 harmonically-related components of equal sensation level. Listeners were asked to determine which of a pair of alternating probe tones was present in the complex. One of the tones coincided with a harmonic, the other occurred between two harmonics. Listeners were able to identify 5 to 8 components reliably; performance declined slightly as the fundamental frequency was increased from 125 to 1000 Hz. Similar results were obtained for inharmonic complexes. The frequency resolution of individual components was approximately 15 to 20 percent, which is roughly consistent with other psychophysical measures of frequency selectivity discussed in the next section.

2.3.4 Masking, critical bandwidth and the 'auditory filter'

The phenomenon of masking (the elevation of signal threshold by the presence of noise or a competing signal) has been extensively studied to measure the frequency analyzing ability of the ear. One of the first systematic studies was conducted by Fletcher (1940) who presented a tone to listeners in the presence of narrowband noise maskers of variable bandwidth, centred at the frequency of the probe tone. He found that the amount of masking increased as a function of noise bandwidth, up to a certain critical value. Increases in noise bandwidth beyond this value did not lead to a further elevation of tone threshold.

Fletcher proposed that the peripheral auditory system may be modelled as a bank of bandpass filters with continuously overlapping centre frequencies. When a tone is presented in noise, only components occurring within a narrow frequency region around the tone (which he called the critical bandwidth) will have a masking effect. Fletcher hypothesized that noise power inside the critical band is equal to the power of the tone at its detection threshold. The bandwidth of the filter can then be estimated as the ratio of tone-to-noise power.

A number of subsequent studies have shown that the behaviour of listeners in a variety of response paradigms changes abruptly when a critical frequency separation (e.g., between probe tone and masker) is exceeded. Consistent estimates of the critical bandwidth have been obtained from studies of two-tone masking, loudness summation, phase sensitivity, discrimination of harmonics, and several other sources (Scharf, 1970). This degree of consistency has led to the widespread acceptance of critical bandwidth as a measure of auditory frequency selectivity.

However, direct methods of determining the critical bandwidth have produced estimates which are approximately 2.5 times larger than the bandwidths estimated by Fletcher's method (Hawkins & Stevens, 1950; Scharf, 1970). Estimates based on Fletcher's method are now referred to as critical ratios.

Fletcher made the simplifying assumption of a rectangular filter shape. This assumption has been shown to be incorrect: changes in threshold may appear even when the frequency separation of the signal and masker exceeds the critical bandwidth. This suggests a filter with sloping skirts. Several attempts have been made to infer the characteristics of the 'auditory filter' from masking data. One of the difficulties in estimating filter properties is "off-frequency listening" (Moore, 1982). If a narrowband masker is introduced at a lower frequency than the tone, for example, the filter centred on the tone will not provide an optimal signal-to-noise ratio; shifting to a filter with a higher centre frequency may attenuate much of the masking noise without significantly affecting the tone.

Patterson (1976) developed a method for determining the shape of the auditory filter which limits the possible effects of off-frequency listening. He measured tone threshold in the presence of wideband noise as a function of the width of a notch in the band, which was arithmetically centred on the tone. The notch had very steep edges, with each edge equidistant from the tone. As notch width decreased, tone threshold increased monotonically. Patterson estimated the shape of the auditory filter, under the following assumptions:

(1) symmetric filter shape, on a linear frequency scale[1];

[1] Patterson and Nimmo-Smith (1980) have shown evidence for some slight asymmetries in filter shape (steeper high frequency slope; reversed in the masking patterns) at higher

(2) the filter is centred on the tone;
(3) the power of the tone at threshold is proportional to the power of the noise passed by the filter.

If the tone power at threshold is proportional to the area under the auditory filter spanned by the noise, then the derivative of the function relating tone threshold to notch width can be used to estimate the transfer function of the assumed filter. Filter shapes derived by this method have relatively broad, rounded tops and steep skirts. The 3-dB bandwidths of the filters were approximately 13 percent of their centre frequency, slightly smaller than estimates of critical bandwidth. The equivalent rectangular bandwidths were somewhat larger, fairly close to the critical bandwidth function, except below 500 Hz where the bandwidth continued to decline. A number of subsequent studies (e.g., Moore & Glasberg, 1981; Patterson, Nimmo-Smith, Weber & Milroy, 1982) have confirmed these basic findings. Patterson et al. (1982) modelled their empirically derived filter shapes in terms of a rounded exponential function.

Another method for determining auditory filter shape from masking data was developed by Houtgast (1974a, 1977) using rippled (comb-filtered) noise maskers. The estimated filter shapes were similar to those obtained by Patterson (1976) using the notched noise method, with fairly wide

[1](cont'd) sound levels

passbands, rounded tops and steep skirts. Although Houtgast's method resulted in somewhat larger bandwidths, the derived filter shapes were fairly similar and could be characterized in terms of a single linear filtering process (Glasberg, Moore & Nimmo-Smith, 1984).

2.3.5 Non-simultaneous Masking

Houtgast (1972, 1974a) reported a change in the pattern of masking results when the test tone and masker were presented non-simultaneously (e.g. using pulsation threshold or forward masking procedures). These differences were attributed to lateral suppression, a sharpening or enhancement of frequency selectivity which is typically found in non-simultaneous masking but not in simultaneous (direct) masking. Houtgast proposed that suppression effects are not observed in direct masking because suppression acts on both the signal and masker; both are attenuated, maintaining the sa signal-to-noise ratio. In non-simultaneous masking the suppression is assumed to end abruptly at the end of the masking stimulus, while the masking effect continues. Estimates of filter shape in simultaneous and non-simultaneous masking (Houtgast, 1974a, 1977; Glasberg, Moore & Nimmo-Smith, 1984) revealed sharper filter shapes with much narrower bandwidths and steeper slopes. A negative trough appeared on the high frequency side of the derived filter characteristic, suggesting that suppression acts primarily from higher toward lower frequencies (Houtgast, 1974a).

Houtgast (1974b) investigated non-simultaneous masking with steady state vowel maskers, /i,e,a/. He observed that peaks corresponding to each of the formants were preserved in the masking pattern. Spectral contrast in pulsation threshold patterns was higher than observed in a one-third octave bandpass filter representation. Tyler and Lindblom (1982) compared direct masking and pulsation threshold measurements using steady state vowels /i,e,y/ as maskers. Pulsation patterns revealed greater spectral contrast (peak-valley differences) than did direct masking patterns. They suggested that suppression might act to preserve spectral peaks in the auditory representation of vowels.

Further support for the hypothesis that suppression plays a role in spectral contrast enhancement comes from a study by Moore and Glasberg (1983) which compared simultaneous and forward masking patterns for the synthetic vowels /i/ and /æ/. Peaks corresponding to the first three formants were enhanced relative to the physical spectrum in forward masking, and somewhat blurred in simultaneous masking. In some cases a formant peak which showed up as a slight "ripple" in simultaneous masking appeared as a clear peak in forward masking patterns. However, Stelmachowicz, Small and Abbas (1982) have shown that suppression effects using pulsation threshold measurements are complex and in some cases may actually diminish spectral contrast. Filter shapes derived from non-simultaneous masking (Glasberg, Moore & Nimmo-Smith, 1984) differ somewhat depending on the type of masking stimuli employed. The presence of a negative region in the filter weighting function suggests an inhibitory process. At present very little is known about the physiological basis of lateral suppression and its role in the perception of vowels.

2.3.6 Excitation and loudness pattern models

The results of psychophysical studies suggest that the auditory system performs a short term frequency analysis of sound, which may be simulated by a set of continuously overlapping parallel bandpass filters whose bandwidths increase in proportion to their centre frequencies. Bandwidth estimates from direct masking studies range from about 13 to 20 percent of the centre frequency; bandwidth estimates from non-simultaneous masking are somewhat narrower. The filter shapes inferred from masking data are generally consistent across studies, with a relatively flat region in the passband and steep skirts.

The concept of the *psychoacoustic* *excitation* *pattern* has been advanced by Zwicker and his colleagues (Zwicker & Zwicker & Feldtkeller, 1967). Excitation patterns are considered to represent the auditory projection of a sound stimulus at a peripheral level of the auditory system. The shape of the excitation pattern is inferred from masking patterns, using data from narrowband noise masking. According to this model, the pattern of excitation evoked by pure tones differing in frequency is invariant in shape when plotted as a function of critical band rate (in Bark units) rather than linear or log frequency.

Excitation patterns can be converted to "loudness patterns" by means of a power function. Excitation levels are thus converted to "specific loudness" or loudness density (in sone units); integration of loudness density as a function of critical band rate gives the total loudness of a complex tone. The model has been used to successfully predict the performance of listeners in a variety of psychophysical tasks including loudness and pitch judgements (e.g. Terhardt, 1979).

Schroeder, Atal and Hall (1979) developed a computational procedure related to Zwicker's model for the purpose of measuring the degradation of speech quality by noise in speech coding systems. This procedure was first discussed by Schroeder (1977) as a possible alternative to spectrographic analysis, providing an "auditorily oriented data reduction" consistent with psychoacoustic data.

The power spectrum of a speech sound is computed using a time window of approximately 20 ms. Frequency is converted to critical band rate using a fitted approximation to the data provided by Zwicker, Flottorp and Stevens

(1957). The resulting critical band spectral density is convolved with a "masking filter" with a low frequency slope of +25 dB/Bark and a high frequency slope of -10 dB/Bark Zwicker (1963)[1]. Convolution with the masking filter results in a pattern of excitation level as a function of critical band rate. This pattern may be converted to a "loudness pattern" by applying a power function with an exponent of .25. Suprathreshold levels are scaled so that the loudness of a 1 kHz tone at 40 dB above the threshold of hearing is equal to 1 sone; components below threshold are set equal to zero.

One of the assumptions underlying this model is that the threshold of a tone is proportional to the amount of excitation produced by the masker in the frequency region of the tone. Moore (1982) has argued that masking patterns may not provide an accurate projection of the activity level in the region of the signal if off-frequency listening is involved.

Moore and Glasberg (1983) proposed an alternative procedure for calculating excitation patterns based on Patterson's (1976) notched noise masking studies. They proposed a filter characteristic based on the rounded exponential function used by Patterson, Nimmo-Smith, Weber and Milroy (1982) to model masking data. They also provided

[1] The masking filter function of Schroeder, Atal and Hall (1979) was found to have a bandwidth of approximately 1.4 Bark rather than 1 Bark (or one critical band). A corrected version was proposed by Sekey and Hanson (1984).

analytic expressions for calculating equivalent rectangular bandwidth (ERB) of the filter and a frequency-to-ERB-rate conversion similar to the critical band rate or Bark scale. There are two important differences between the Moore and Glasberg method for calculating excitation patterns and the procedures developed by Zwicker and Schroeder et al. The filter bandwidths based on notched noise masking data are narrower than classical estimates of critical bandwidth; these bandwidths are not constant below 500 Hz but but continue to decline with decreasing frequency, reflecting a possible error in the classical critical band function (Fidell, Horonjeff, Teffeteller & Green, 1983). In addition, filter shapes obtained by the notched noise method are less likely to be affected by off-frequency listening (Moore, 1982).

2.3.7 Applications of auditory models to vowels

Can the study of auditory frequency analysis provide a representation of the acoustic stimulus that will help to reveal the information-bearing elements of vowel spectra in a manner that standard methods (e.g. spectrographic analysis) cannot? Several studies have addressed this question, using auditory models to obtain spectral patterns whose characteristics were compared with the perceptual responses of listeners. Carlson and Granstrom (1979) evaluated several alternative spectral representations of vowels for the prediction of perceptual distance judgements. Pairwise comparison of 66 synthetic versions of a vowel similar to /æ/ were made, each differing along acoustic dimensions such as formant frequencies, bandwidths, overall amplitude, phase, spectral tilt, and various types of filtering. They compared six different models. The first two were based on FFT spectra, with summation of spectrum levels over 200 Hz intervals[1] (called the "FFT" model) or over a one Bark interval (called the "BARK" model). The third model applied the filterbank analysis of Schroeder, Atal and Hall ("MASK" model). The fourth ("PHON") obtained loudness patterns in phon units[2] from equal loudness contours, applied to the output of the MASK model. In the fifth model ("SONE") loudness level was converted from phons to sones using the formula proposed by Bladon and Lindblom (1981). The sixth model ("DOMIN") was based on temporal analysis of filterbank output, and will be discussed separately in section 2.3.10.

Plomp (1970) proposed that the relative similarity between two complex tones could be predicted in terms of the

[1] This preprocessing stage, which involves a rectangular smoothing function with a bandwidth larger than the fundamental frequency of the vowels, is not based on psychoacoustic data and would appear to eliminate spectral fine structure, i.e. the peaks corresponding to individual harmonics of a low fundamental frequency. In fact, data from a variety of sources including pitch discrimination and masking have suggested that low frequency harmonics of complex tones can be resolved by the ear, at least for fundamental frequencies in the speech range.

[2] Loudness level (in phon units) of a tone is the intensity, in dB SPL, of a 1 kHz tone which it equals in perceived loudness (Moore, 1982, p.45).

distance D_{ij} between pairs of spectra i and j, each represented by k frequency channels:

$$D_{ij} = (\Sigma_k |A_{ik}-A_{jk}|^n)^{1/n} \quad ,$$

When parameter n=1, the result is a "city block" metric; for n=2 it is Euclidean distance. Euclidean and city block distances were computed between pairs of spectra. Correlations between these predicted distances and perceptual data ('phonetic' and 'psychoacoustic' judgements of perceived distance[1]) were calculated for each model.

Overall, correlations were very high. However, the correlations with "phonetic" judgements were much lower than correlations with "psychoacoustic" judgements. The implications of this finding will be discussed below. Each of the models using the "MASK" filter (models 3, 4, and 5) yielded lower correlations than the models based on the unfiltered amplitude spectrum. Evidently listeners were basing their judgements in part on information lost through filtering, possibly because the filter bandwidths were actually larger than critical bands (Sekey and Hanson, 1984). The assumption of a single dimension of distance, and the assignment of equal weights to each channel may be invalid. Carlson and Granstrom found a large degree of task and stimulus dependence. Certain kinds of manipulations

[1] This distinction is based on a change in instructions to listeners. Phonetic distance: "Rate only changes that tend to influence vowel identity, disregard changes associated with harshness, speaker identity or transmission channel". Psychoacoustic distance: "...take into account any difference between the vowels" Carlson and Granstrom, (1979: p.85)

(e.g., formant frequency changes) affected phonetic distances much more than others (e.g., spectral tilt or overall amplitude). These differences were less evident in the psychoacoustic distances judgements. Changes which affect formant peaks had the largest effects on phonetic distance judgements.

Each of the models proposed by Carlson and Granstrom (1979) was incorporated in an automatic recognition scheme by Blomberg, Carlson, Elenius and Granstrom (1982, 1983). In all cases, the FFT model gave better results than any of the models employing the "masking filter" supporting the conclusion that important vowel information is lost in filtering.

Bladon and Lindblom (1981) also predicted dissimilarity judgements between two- and four-formant vowels using two of the models described by Carlson and Granstrom (1979). The correlation of judged dissimilarity on a five-point scale (expressed in log units) with Euclidean distance between pairs of excitation patterns (MASK model) was fairly high (r=.79). A marked improvement was found using loudness patterns (SONE model) (r=.89), but loudness normalization (dividing each loudness channel by the total loudness) led to a lower correlation (r=.83).

The correlations are generally higher than those reported by Carlson and Granstrom (1979) which may be due to several factors. Listeners were asked to rate the overall dissimilarity (as in the "psychoacoustic" task of Carlson and Granstrom) rather than phonetic changes. Acoustic differences between two- and four-formant vowels are spread over a large region of the spectrum, and these can result in large changes in timbre. If phonetic distances are based on local differences in spectral profile, i.e. listeners assign greater importance to discrepancies in the region of formant peaks (e.g., Carlson, Granstrom & Klatt, 1979; Klatt, 1982a) poorer predictions might be expected with phonetic distance judgements. At the same time, the high correlation with loudness patterns is difficult to evaluate without an appropriate reference condition, such as distance predictions based on the unfiltered amplitude spectra.

Plomp (1970,1976) proposed that the sound pressure levels of one-third octave bandwidth filters may provide a good approximation to the loudness pattern as described by Zwicker. He suggested that one-third octave provides an adequate approximation to critical bandwidth in the range above 500 Hz. Filters below 500 Hz were assigned a fixed bandwidth of about 100 Hz, reflecting the classical critical band function (Zwicker, Flottorp & Stevens, 1957).

Judgements of timbre dissimilarity of complex steady-state tones similar to musical tones, or vowel sounds were highly correlated with distances between one-third octave spectra, represented as distances in a multidimensional space. Klein, Plomp and Pols (1970) also found very high correlations between the perceptual space derived from multidimensional scaling of dissimilarity judgements, and the factor space resulting from principal components analysis of the bandfilter spectra of vowel sounds. Pols, van der Kamp and Plomp (1969) presented similar results using reiterated single pitch periods from natural vowels.

In order to evaluate whether one-third octave spectra provide a perceptually realistic portrayal of vowel spectra, it is necessary to compare model predictions using one-third octave filters with the more detailed models of the loudness pattern which they are intended to approximate.

While each of these auditory models can predict overall timbre dissimilarity fairly successfully, they do not appear to assign enough weight to formant changes in predicting phonetic quality judgments and are too sensitive to factors such as speaker and voice quality and characteristics of the transmission channel.

Chistovich and her colleagues (Chistovich, 1971; Chistovich, Kozhevnikov, Lesogar, Shupljakov, Taljasin & Tjulkov, 1974; Karnickaya, Mushnikov, Slepokurova & Zhukov, 1975) have argued that loudness patterns must be subjected to a subsequent level of processing in the form of a lateral inhibition function (which sharpens the response to spectral peaks) to account for the results of vowel perception studies. For example, the presence of a small "shoulder" on the upper slope of a formant peak may cause a radical change in vowel quality, shifting a back vowel to a front vowel. Karnickaya et al. described a 30 channel lateral inhibition weighting function, divided into 3 equal portions one Bark wide. The central symmetrical, positive portion was flanked by two negative (inhibitory) portions on either side, larger on the low frequency side. This aspect of the model is not strictly motivated by psychoacoustic or neurophysiological studies, and must be regarded as tentative. The effects of lateral inhibition are similar to the calculation of the second derivative of the loudness pattern along the frequency scale, and result in a clear enhancement of spectral peaks (Huggins & Licklider, 1951).

2.3.8 Frequency analysis in hearing: physiological factors

The auditory system performs a frequency analysis of sound. A frequency-to-place transformation of the acoustic stimulus takes place on the basilar membrane. The position of maximum displacement along the basilar membrane varies as a function of frequency. The basal portion of the basilar membrane responds maximally to high frequencies while the apical portion is most sensitive to low frequencies (v. Bekesy, 1960).

At the level of the auditory nerve, information about the sound spectrum is preserved in the distribution of neural activity in the population of auditory nerve fibres.

Individual fibres exhibit behaviour suggestive of narrowband filtering (Evans, 1982). Stimulus energy near the characteristic frequency (CF) of a fibre causes an increase in mean discharge rate over the spontaneous discharge rate; each unit fires at a rate proportional to the energy within its tuning curve (Kiang and Moxon, 1974).

Most cochlear fibres have a limited dynamic range, however, and respond at their maximum rates about 20-40 dB above threshold, suggesting that a rate-place representation may not adequately represent vowel formant structure at high intensities. Sachs and Young (1979) investigated average firing rates of a population of auditory nerve fibres in anesthetized cats, in response to synthesized vowel sounds. Peaks in the rate profiles for /I/, /ɛ/ and /a/ were observed in the vicinity of formant peaks at low intensities (38 dB SPL). At moderate to high levels (58-78 dB SPL), the discharge rate profile showed a loss of spectral contrast (smaller peak-to-valley ratios).

The limited dynamic range of the average rate profiles conflicts with estimates of the dynamic range of speech intelligibility, which is reportedly stable over at least 100 dB in humans (Pollack & Pickett, 1958). There is, however, a small proportion of fibres with low spontaneous firing rates which have a wider dynamic range; these fibres may preserve spectral contrast in vowel sounds even at high intensities.

The temporal pattern of neural discharges provides an alternative source of information about a sound stimulus. Fibres with CFs below about 4 kHz fire at intervals which are synchronized or phase-locked to the stimulating waveform. These phase-locked responses approximate a half-wave rectified version of the stimulating waveform following cochlear filtering (Rose, Brugge, Anderson & Hind, 1967; Rose, Kitzes, Gibson & Hind, 1974).

Young and Sachs (1979) obtained temporal discharge patterns from a large population of auditory nerve fibres in response to synthetic vowel sounds'. They demonstrated that temporal responses to components near formant peaks were larger than responses to other components. Many fibres with CFs near a formant peak were dominated by a single harmonic, possibly because the frequency of each formant coincided with the frequency of a harmonic. Separate peaks were retained for the lowest two or three formants in the response profiles, up to at least 80 dB SPL. The temporal discharge pattern thus had a much wider dynamic range than the average rate representation. One reason for this extended range is the presence of synchrony suppression at high levels, which reduced the amount of phase-locking to frequencies outside the formant regions. At high

'Temporal response patterns were derived from period histograms, which indicate the instantaneous discharge rate over the course of a single period of a vowel. The degree of synchrony to each harmonic component of the vowel was defined as the magnitude of the Fourier Transform of the period histogram, averaged over all fibres within 1/4 octave above and below the frequency of the harmonic.

intensities a large number of fibres were synchronized to the first formant, and a small number to F2 and F3. While average rate suppression resulted in a loss of spectral contrast, synchrony suppression appeared to preserve the formant structure of vowel sounds.

These results have recently been extended by Delgutte and Kiang (1984) to a larger sample of nine steady state, two-formant vowels whose formant frequencies in most cases did not coincide with harmonics of the fundamental frequency. They reported that any harmonic close to a formant peak which exceeds its neighbours by more than about 6 dB dominated many fibres with CFs in its vicinity; synchronized responses to other harmonics were suppressed. However, two components near the peak whose levels differed by less than 6 dB both maintained a high response level.

Sachs, Young and Miller (1982) and Delgutte (1984) demonstrated that the individual harmonics of a vowel stimulus are resolved in temporal response patterns. Delgutte, who adopted a one-sixth octave filterbank analysis of the period histograms, found that the lowest 8 harmonics appeared as separate peaks in the temporal response patterns, corresponding roughly to estimates of psychophysical frequency selectivity.

In summary, the transformation of a vowel sound into the spatial pattern of vibration along the basilar membrane, and subsequently to the firing pattern of auditory nerve fibres, appear to result in a certain loss of information about its spectrum. The results of Young and Sachs indicate that peaks corresponding to the lowest formants or harmonic components near the formant peaks may be represented in a more robust fashion by a temporal code than by a rate/place code at high sound levels.

2.3.9 Timing pattern models

Temporal response patterns of auditory nerve discharges to vowel sounds (Young & Sachs, 1979; Delgutte & Kiang, 1984) are often dominated by components near formant peaks, even for fibres whose CFs are remote in frequency from the peak. In these studies, temporal response magnitudes were determined by spectral analysis of period histograms. Response levels were plotted as a function of fibre CF, resulting in a tonotopic representation of the spectrum. This coding scheme is related to the 'central spectrum' model of Srulovicz and Goldstein (1983). In their model, the central spectrum is determined by the output of a matched filter applied to the interval histogram corresponding to each auditory nerve fibre, tuned to the characteristic frequency of that fibre.

It was noted by Plomp (1964) that listeners are better able to resolve individual harmonics of two-tone than multitone complexes. Moore (1982) discussed a possible temporal mechanism that could account for this. In the multitone complex, analysis channels may be responding to several harmonics simultaneously; no channel will be dominated by a single component. In the two-tone complex, there may be channels remote in frequency that are dominated exclusively by one of the components, as a result of the sloping skirts of the analysis filters and the absence of other signal components in the vicinity. The "edge" components of a multitone complex and possibly the spectral peaks of vowels might be expected to show similar behaviour.

A histogram of the number of channels dominated by a given frequency component, plotted against channel frequency provides an alternative spectral representation which gives special prominence to dominant components of the spectrum, enhances spectral peaks and suppresses energy in spectral valleys (Carlson, Granstrom & Fant, 1970; Carlson, Fant & Granstrom, 1975; Delgutte, 1984). Carlson et al. applied this model to the prediction of second formant matching in two-formant approximations to four-formant vowels. In their model ("DOMIN"), the number of zero-crossings are counted at the output of each channel of a critical band filterbank over a selected time interval; dominant frequencies are identified by counting the number of channels with the same zero crossing frequency within a 75 Hz quantization range. The number of channels are plotted as a function of channel frequency, resulting in a spectral representation which resolves individual harmonics of vowels in the low frequency region, and formant peaks in the high frequency region.

The DOMIN model yields very high correlations with phonetic distance judgements, and has also given excellent results in automatic vowel recognition (Blomberg, Carlson, Elenius & Granstrom, 1983, 1984). Although the application of this model to vowel sounds has not been studied in detail, the preliminary results of Carlson and Granstrom (1979) suggest that it may be worth pursuing. The DOMIN model consistently performed better than any of the auditory models based on the masking filter proposed by Schroeder, Atal and Hall (1979).

2.4 Pattern recognition in vowel perception

Vowel identification may be regarded as a pattern recognition process operating on the spatial and/or temporal distribution pattern of neural activity. The incoming signal is assumed to be scrutinized, compared with stored patterns and assigned a category label corresponding to the pattern which best matches the input. Present research in speech perception is aimed at a functional characterization of the relationship between incoming acoustic signals and the outcome of the decision mechanism, based on a set of hypotheses concerning the nature of the underlying mechanisms.

2.4.1 "Whole spectrum" specification

One of the simplest models of the vowel identification process involves the systematic comparison of the whole spectrum of an incoming signal with each of a set of stored spectral patterns corresponding to different vowel categories. The decision mechanism may depend on a cross-correlation process, or on the best match with a template or matched filter (Licklider, 1952). No intermediate feature extraction stage is postulated as the basis for comparison with stored patterns; hence a "whole spectrum" specification.

This model has been implemented in automatic vowel recognition procedures (e.g., Blomberg, Carlson, Elenius & Granstrom, 1983, 1984; White & Neely, 1976; Suomi, 1984). Attempts to apply whole spectrum-based distance measures were discussed in previous sections, where it was noted that these measures were closely correlated with listeners' judgements of timbre. The main difficulty is that these models do not perform as well in the prediction of phonetic quality judgements.

The whole spectrum approach generally tends to overestimate phonetic distances resulting from changes such as high pass, low pass and notch filtering; changes in spectral tilt; changes in formant amplitudes and bandwidths (Carlson & Granstrom, 1976; Carlson, Granstrom & Klatt, 1979; Klatt, 1982a,b). Changes in the frequencies of spectral peaks are much more important than other kinds of changes for phonetic quality.

Mushnikov and Chistovich (1971) reported that changes in the level of a single harmonic in the F1 region of a synthetic front vowel could induce a shift in phonetic quality from /i/ to /e/. Further evidence was presented by Chistovich (1971, 1985) indicating that changes in a very narrow spectral region can produce large changes in phonetic quality. These results are not consistent with a whole spectrum approach and indicate that certain regions of the spectrum are perceptually more important than others.

An additional problem for a whole spectrum specification concerns the role of fundamental frequency. Spectral patterns based on the auditory models discussed in previous sections contain peaks in the low frequency region corresponding to individual harmonics. Altering the fundamental frequency, which shifts the locations of these peaks, will have a large effect on the calculated distance but only a small effect on phonetic quality. One possible solution is a spectrum interpolation procedure which eliminates the harmonic structure; however, it is known that f_0 can affect vowel quality in some circumstances (Miller, 1953; Carlson, Fant and Granstrom, 1975; Ainsworth, 1975).

2.4.2 Principal components analysis of vowel spectra

An alternative to the whole spectrum hypothesis is suggested in a series of studies employing a statistical procedure for reducing the dimensionality of speech spectra (e.g., Plomp, Pols & van der Geer, 1967; Klein, Plomp & Pols, 1970; Pols, 1977). Principal components analysis (PCA) represents the spectrum in terms of a reduced number of dimensions. These dimensions, or principal components are each a linear combination of levels recorded in the original spectrum. Each new dimension accounts for as much of the residual variance in the spectrum levels as possible.

These studies have demonstrated that the variability in speech spectra can be described by a small number of statistically independent spectral shape factors. These factors have been used successfully as the parameters of a synthesis system producing intelligible speech, and in automatic speech recognition (Pols, 1977). The factor space is closely related to the formant representation. Pols et al. have argued that their analysis procedure is preferable to a formant analysis for the following reasons:

(1) The same representation can be used for both speech and non-speech sounds; there is no need to impute a special decoding mechanism (e.g. a formant extractor) for speech

sounds.

(2) From a practical point of view, formant measurement is less efficient, and runs into difficulties with high fundamental frequencies and when formants are close together or asymmetical. PCA captures the essential information of a formant specification, can be carried out in real time and avoids the problems of "missing" or "spurious" formants.

(3) The factor space derived from PCA is correlated with the perceptual space derived from vowel identification performance by listeners.

While PCA has not been explicitly formulated as a theory of vowel perception, Klein, Plomp and Pols (1970) have suggested that it may provide "a useful model of what a human listener unconsciously does" (p. 1008). If so, it may be of interest to determine whether an optimal statistical data reduction can specify the differences between vowels in a manner which reflects their perceptual attributes. A more detailed study of the correspondence between the factor space and the performance of listeners is needed to answer this question.

The studies reviewed above suggest that a separate representation may be needed for timbre and phonetic quality differences, and it is difficult to see how a perceptual model based on PCA can account for these differences. Carlson, Granstrom and Klatt (1979) have shown that differences between steady state vowels may be judged differently depending on whether listeners are instructed to respond in a phonetic mode or not. Large changes in the spectrum may have only minor effects on phonetic distance if the locations of spectral peaks are not altered. At the same time, very small changes in spectrum envelope can result in enormous differences in vowel quality (Chistovich, 1971). For example, the introduction of a small shoulder on a single formant vowel can shift its quality from a back vowel to a front vowel (Chistovich, Sheikin & Lublinskaya, 1979). Further difficulties for this model may arise when the effects of fundamental frequency, speaker differences, and the presence of simultaneous speech sounds or noise are considered.

2.4.3 Formant hypothesis of vowel perception

The formant hypothesis of vowel perception maintains that listeners identify and judge the phonetic quality of vowels on the basis of the frequency locations of the formants. A number of sources of evidence have been cited above indicating that formants play a special role in the perception of vowels. Formants have received special attention because of their role in speech production and speech synthesis, The first two formants are thought to be sufficient for representing known vowel qualities, except for nasalized or 'r-coloured' vowels (Joos, 1948; Delattre, Liberman, Cooper & Gerstman, 1952; Miller, 1953; Carlson, Granstrom & Fant, 1970). Typical results for synthetic vowels based on measured formant values were presented by Lehiste and Meltzer (1973) who obtained identification scores of about 75 percent, approximately 10 to 15 percent lower than identification of naturally spoken vowels.

Other studies have investigated the position of vowels in the F1-F2 plane in relation to listeners' confusion errors (e.g. Fairbanks & Grubb, 1961; Peterson & Barney, 1952; Tiffany, 1959; Koopmans-van Beinum, 1980; Assmann, Nearey & Hogan, 1982). These studies have shown a close correspondence between formant overlap and confusion errors by listeners. An additional source of evidence comes from multidimensional scaling of vowel similarity judgements (e.g. Terbeek & Harshman, 1971; Singh, 1974; Fox, 1983). These studies have consistently extracted perceptual dimensions corresponding to vowel height (or F1) and advancement (or F2).

Formant parameters have formed the basis for a number of successful automatic vowel recognition algorithms (e.g, Gerstman, 1968). Recognition scores of over 90 percent can be obtained using only F1 and F2; even higher rates are obtained when fundamental frequency, higher formant frequencies and/or speaker normalization parameters are included.

The existence of cross-speaker variability in formant frequencies has led to the proposal that vowel perception is based on invariant relational properties of the formant pattern, shifted along the frequency scale (e.g. Joos, 1948; Potter & Steinberg 1950; Lieberman, 1976; Nearey, 1977). The lack of invariance is mainly due to the overlap in formant frequencies for different vowels resulting from vocal tract size differences. A number of procedures have been developed for mapping invariant relationships among the formants across speakers. It appears that some relatively simple transformations of the formant space will reduce the degree of overlap considerably, and that listeners are able to compensate for speaker differences on the basis of context (e.g., Ladefoged & Broadbent, 1957; Nearey, 1977; Assmann, Nearey & Hogan, 1982).

<u>Perception of the first formant</u>. At present very little is actually known about the auditory encoding of formant frequency. One of the problems faced by the formant hypothesis concerns the estimation of formant location when no energy is physically present at the frequency of the formant[1]. Psychophysical studies have shown that individual harmonics in the low frequency region (spanning the range of the first formant in front vowels) can be resolved by the ear. Kakusho, Hirato, Kato and Kobayashi (1971) have shown

[1] It should be pointed out that in natural speech, changes in fundamental frequency and formant pattern over time may help to specify the location of the formant peak. Even when energy is present at the formant peak, it is still of considerable interest to determine whether vowel quality is determined exclusively by the peak location or whether spectrum levels in adjacent frequency regions are also taken into account.

that listeners can discriminate very small changes in the amplitudes of individual harmonics in the vicinity of a formant; the profile of difference limens (DLs) versus frequency shows a pattern very similar to the inverted spectrum envelope.

These considerations suggest that the frequency of a formant may be inferred from the frequencies and amplitudes of prominent harmonics in the formant region. One possibility is a perceptual weighting of harmonic levels to estimate the location of the a peak in the spectrum envelope (Carlson, Fant & Granstrom, 1975). An alternative suggestion advanced by Mushnikov and Chistovich (1971) is that the most prominent harmonic in the formant region provides the acoustic cue for the perception of formant changes in the low frequency region. The evidence for these two hypotheses will be considered in chapter three, where a series of experiments are described which investigate the role of individual harmonics in the F1 region of front vowels.

Perception of the second formant. Although vowel synthesis studies have shown that two formants provide sufficient information for vowel identification, the preferred frequency location for the second formant is often higher than its value in natural tokens (Fant, 1972; Fant & Risberg, 1963; Carlson, Granstrom & Fant, 1970; Bladon & Fant, 1978). Several empirical formulas were developed for predicting the "effective" second formant, F2', using a weighting of the higher formants (F2, F3 and F4). Carlson, Fant and Granstrom (1975) found that they were also able to predict F2' values fairly well by selecting the highest peak above F1 in the spectral profiles produced by their DOMIN model. One problem was the sensitivity of the model to formant amplitude changes which they did not find in the matching data.

While Carlson et al. (1970, 1975) reported no evidence of bimodal matching, Bladon and Fant (1978) found that F2' matches often alternated between F2 and F3 or F4 in high front vowels. Mushnikov, Slepokurova and Zhukov (1974) found that the proportion of matches to F3 was increased when the amplitude of this formant was raised. This suggests that matching may depend on a peak selection process which determines the frequency of the most intense formant in the region above F1, rather than on a weighted average of the higher formants as proposed by Fant (1959).

There are indications from other studies that a weighted average of F2, F3 and F4 cannot account for all of the perceptual data. For example, Fujimura (1967) found varying both the distance between log F2 and log F3 and their mean value had an effect on vowel identification in the region of /i y u/, with an interaction between the two variables. Further empirical data are needed to assess whether the F2' experiments are the result of a peak selection process or a weighting of higher formants.

One possibility discussed by Stalhammar (1978), Chistovich and Lublinskaya (1979) and Bladon (1983) is that F2' is positioned above F2 of the reference vowel because of an auditory spectral integration process, which leads to a perceptual "merging" of closely spaced formants.

<u>The formant centre of gravity hypothesis in back vowels</u>. The formant centre of gravity hypothesis was originally proposed by Delattre, Liberman, Cooper and Gerstman (1952) who discovered that back vowels could be synthesized using only a single formant. They proposed that when F1 and F2 are close together (as in back vowels) the two formants are combined to form a single "effective" peak whose spectral centre of gravity determines the phonetic quality of the vowel.

Chistovich and Lublinskaya (1979) provided further experimental data in support of the formant centre of gravity hypothesis in back vowels and obtained estimates of the critical separation of two formants necessary for these apparent fusion effects to occur. They proposed that the phonetic quality of back vowels is determined by the centre of gravity of the F1-F2 cluster when F1 and F2 are separated by a frequency distance smaller than 3-3.5 Bark.

The formant centre of gravity hypothesis in back vowels is evaluated in chapter four, and two experiments will be reported which evaluate the contribution of formant frequency and formant amplitude changes in the perception of back vowels.

CHAPTER THREE

A PERCEPTUAL STUDY OF FRONT VOWELS

3.1 Introduction

The present chapter is an experimental investigation of the acoustic and auditory bases for the perception of height differences in front vowels. Previous studies have demonstrated the importance of F1 in the perception of vowel height, yet the auditory basis for this aspect of vowel quality has been little studied. It has been suggested that the frequency of the first formant is inferred by the perceptual system from the relative amplitudes and frequencies of individual harmonics in the F1 region. In this chapter several alternative hypotheses about the role of individual harmonic components in the perception of vowel height will be considered. Several of these hypotheses are evaluated by means of vowel matching and identification experiments.

3.1.1 General approach

In the following experiments we will draw upon the observed relationship between vowel height and the frequency of the first formant to assess the effects of spectrum changes on the perceived height of a vowel. It is assumed that F1 changes have a consistent and reliable effect on vowel height. A shift in formant frequency, however, represents a series of correlated changes throughout the vowel spectrum. In order to explore the fine structure determinants of vowel height, the frequency of the first formant is used as a physical measure by which the contribution of individual harmonics to vowel height can be determined.

In the matching experiments, listeners attend to pairs of synthetic vowel sounds. Each pair is comprised of a matching stimulus and a reference stimulus. The matching stimulus is a member of a continuum of vowels differing in F1. Listeners are asked to find the closest match in vowel height by adjusting the F1 value of the matching stimulus.

The reference stimulus is produced by modifying the amplitude spectrum of one member of the F1 continuum (referred to as the baseline stimulus). Typically, these are modifications of the spectrum which are predicted to shift vowel height according to a given hypothesis A but are deemed irrelevant according to a second hypothesis *B*. A subsequent examination of the spectral characteristics of each reference stimulus and its "best match" is carried out to determine which attributes of the spectrum are important for the perception of vowel height differences. From a comparison of the shared acoustic properties of vowel pairs judged to be closest in height, characteristics of the perceptual *weighting function* underlying vowel height differences can be inferred.

The value of the first formant frequency used to synthesize each stimulus will be refered to as its *nominal F1*. As a result of the spectral modifications carried out on a given reference stimulus, its vowel height may be altered. Its *effective F1* is assumed to be given by the nominal F1 value of the stimulus judged to be its best match[1].

The effects of F2 and higher formants will not be investigated here. There is an inverse correlation of F1 and F2 in front vowels: higher F1 values are generally associated with lower F2 values (see Figure 3.2 for an example from western Canadian English). However, this correlation does not appear to have a large role in

[1] An analogous procedure is often used in pitch matching experiments where the pitch of a complex tone is considered to be given by the frequency of a sinusoid judged to have the same pitch (Moore, 1982). The perceived height of a vowel with modified spectrum is expressed here in terms of the F1 value of a vowel judged to have the same phonetic quality.

perceived vowel height.

Studies with synthetic speech (e.g., Karnickaya, Mushnikov, Slepokurova and Zhukov, 1975; Hose, Langner and Scheich, 1983) have indicated that vowel responses to stimuli on an F1-F2 continuum often occupy regions whose boundaries lie parallel to the F1 or F2 axis. Mushnikov and Chistovich (1971) demonstrated that the distinction between the vowels /i/ and /e/ in Russian could be controlled by F1 changes alone, positioning F2 near the average value for these two vowels. In Experiment 4 it is shown that Canadian English listeners can readily assign synthetic vowel stimuli on an F1 continuum (with a single fixed F2 value) to one of the 5 front vowel classes.

3.1.2 Additive harmonic synthesis

In order to determine the contribution of individual harmonics to the perception of vowel height, the vowel stimuli used in the following experiments were produced by additive harmonic synthesis. This procedure enables independent adjustments of the frequency, amplitude and phase of each component. Such adjustments are not possible using conventional formant synthesis.

For the matching experiments, the amplitude spectra of the baseline stimulus and matching stimuli were computed using the cascade formant synthesis model [illegible] 1960; Flanagan, 1972). The amplitude spectrum of each reference stimulus was generated by modifying the low frequency region (0-1.5 kHz) of the spectrum of the baseline stimulus. The waveform was calculated by summation of digitally produced sinusoids whose amplitudes were specified by the cascade formant synthesis model; all components were added in sine phase. The fundamental frequency was constant at either 125 or 250 Hz in all experiments.

3.2 Some hypotheses concerning the perception of vowel height

According to the formant theory of vowel perception, listeners are able to infer the frequencies of vocal tract resonances from the speech signal. In this section several alternative hypotheses concerning the nature of this process will be considered. One possibility is that the first formant peak in front vowels is only roughly estimated by means of a simple peak selection process which locates the largest (most prominent) harmonic in the formant region. Alternatively, the perceptual system may employ a weighting of two or more harmonics near the formant peak; or a global procedure which takes into account the shape of the entire spectrum.

3.2.1 Frequency of the most prominent harmonic (FMPH hypothesis)

According to Mushnikov and Chistovich (1971, 1973), the frequency of the most prominent harmonic in the first formant region determines vowel height, and serves to distinguish vowels such as /i/ and /e/. Their evidence comes mainly from experiments involving steady-state, synthetic front vowels with all harmonic components in the F1 region replaced by a single harmonic, or a small number of harmonics of equal intensity.

Chistovich (1971) described an experiment in which the first formant region of a synthetic vowel was replaced by a pure tone of variable frequency controlled by the subject. The reference was a synthetic two formant vowel with F2 equal to that of the standard stimulus. Listeners were asked to adjust the frequency of the tone to find the best match in vowel quality. It was found that for <u>low</u> f_0 values (<100 Hz) the tone was matched near F1. At a higher f_0 (where each harmonic in the F1 region presumably occupies a single critical band) the tone was not matched to F1 but to the frequency of the most intense harmonic in the formant region. Chistovich concluded that vowel height is based solely on the frequency of the most prominent harmonic in the F1 region, provided the harmonics are separated by a frequency interval greater than the critical bandwidth.

Carlson, Fant and Granstrom (1975) rejected this conclusion, suggesting that the task employed by Chistovich may have forced listeners to compromise between matching for vowel quality and matching for pitch. Listeners may have positioned the sinusoid at the frequency of a harmonic because of the need for harmonic congruence, even if this resulted in a slight mismatch in vowel quality.

In a subsequent study, Mushnikov and Chistovich (1971) discovered that a increase in the level of a harmonic either immediately above or below 450 Hz (a frequency value near the F1 boundary for the /i-e/ distinction in Russian) shifted the identity of the vowel from an /i/ to an /e/. The level difference between a pair of harmonics (one above and one below 450 Hz) at the phoneme boundary was fairly constant over a range of overall intensities. The relative amplitude of the harmonics at the phoneme boundary did not correspond to equal loudness of the components in isolation.

Mushnikov and Chistovich (1973) constructed stimuli with 1 to 3 equal amplitude harmonics either below or above the 450 Hz cutoff, and a single harmonic on the opposite side. Listeners altered the amplitude of this single harmonic to induce a shift from /i/ to /e/ or vice versa. The number of harmonics on the opposite side did not affect this amplitude setting. The authors proposed that the maximum on the loudness density curve in the F1 region (corresponding to the most prominent harmonic) is critical for the perception of changes in vowel identity.

The maximum on the loudness density curve will generally correspond to the most intense harmonic. At low fundamental frequencies, however, there will be more than one harmonic within a critical band which may displace the peak in the loudness density curve from the frequency value of the most prominent harmonic toward an adjacent component.

Two harmonics near the formant peak may have equal loudness density levels, and hence there will be competing candidates for the most prominent harmonic, each specifying a distinct vowel height. The perceived quality of such a vowel may be expected to be ambiguous[1]. If two harmonics are equal in intensity (as in the synthetic stimuli used by Mushnikov and Chistovich) an examination of the loudness density patterns may help to determine which harmonic is perceptually the more prominent. Several alternative methods for estimating the relative auditory prominence of harmonics will be considered below.

Peterson (1959) suggested that the frequency of the highest amplitude harmonic in the vicinity of a formant might provide a relatively simple and reasonably accurate

[1] An alternative model may be considered in which "prominence" is based on a "noisy" or stochastic representation of the loudness density levels of the individual harmonics. A random fluctuation in the loudness levels of prominent harmonics can be expected to produce shifts in vowel height from trial to trial, and a multimodal pattern of matching may be expected. The size of these shifts will be determined by the harmonic spacing.

measure of the formant frequency. However, this measure will become increasingly less accurate at higher fundamental frequencies, because the density of harmonics is reduced, with a maximum error of $f_0/2$. One problem noted by Peterson is that vowels with very close spacing of F1 and F2 and a high fundamental frequency (e.g., a child's /a/) may contain only a single harmonic or pair of harmonics in the F1-F2 region. This problem will be considered further in chapter four.

If vowel height is determined exclusively by the frequency of the most prominent harmonic, one would expect that changes in F1 will lead to discrete shifts in vowel quality (separated by intervals equal to the fundamental frequency), rather than a continuous, gradual shift which might be expected with more accurate estimation of first formant frequency.

3.2.2 Local centre of gravity procedures

Local centre of gravity measures such as the first spectral moment in the region of a formant have often been used to estimate the frequencies of formant peaks from the amplitude spectrum. The first spectral moment may be defined as

$$F_m = \Sigma_i (F_i \cdot S_i)/\Sigma_i S_i \qquad (3.1)$$

where S_i is the spectrum level for the ith component and F is its frequency. The range of summation is restricted to the region of the spectrum spanning a formant peak.

The calculation of the first moment or weighted mean of spectral components in the vicinity of a formant peak was discussed by Lloyd (1896) and attributed to the earlier work of Hermann (who adopted a weighting of the amplitudes and frequencies of the three most intense harmonics in the formant region) and Pipping (who proposed a weighting based on log frequencies, in accordance with musical scales). More recently, Potter and Steinberg (1950), Suzuki, Kadokawa and Nakata (1963), and Schroeder (1975) suggested spectral moment calculations over selected frequency ranges for estimating formant frequencies. Suzuki et al. proposed that a global spectral moment, computed over a broad frequency range, might be used to separate the F1 range from that of the higher formants. This procedure provided fairly good estimates of F1 and F2, except when F2 was close to F1 (as in back vowels) or with high fundamental frequencies, as in the voices of children.

Potter and Steinberg (1950: 811) considered that centre of gravity procedures could be used to characterize the perception of vowel sounds, suggesting that "the ear deals with something akin to effective pitch centres of loudness of the energy concentrations".

3.2.3 Weighted mean of the two most prominent harmonics (FMPH2 hypothesis)

Carlson, Fant and Granstrom (1975) reported the results of an identification experiment with synthetic vowels in which F1 varied along a continuum from /i/ to /e/. Five versions of this continuum were generated, differing in fundamental frequency (100, 115, 130, 145 and 160 Hz). Carlson et al. found that F1 values at the /i-e/ boundary increased as a function of f_0. These boundary shifts did not correspond to changes in the frequency of the most prominent harmonic in the F1 region; their gradual nature suggested instead a weighting of harmonics in the F1 region. Carlson et al. did not observe an abrupt discontinuity in phonetic quality as F1 passed from the frequency of one harmonic to another which would be predicted by the FMPH hypothesis. Instead, they proposed that height differences in front vowels depend on the weighted mean of the two most prominent harmonics in the first formant region.

The relationship between F1 and harmonic levels in the F1 region is illustrated in Figure 3.1. Line spectra are shown for three vowels with a fundamental frequency of 125 Hz and F1 values of 450, 500 and 550 Hz. It can be seen that all harmonics in the F1 region are affected by a change in formant frequency; those near the peak are affected most.

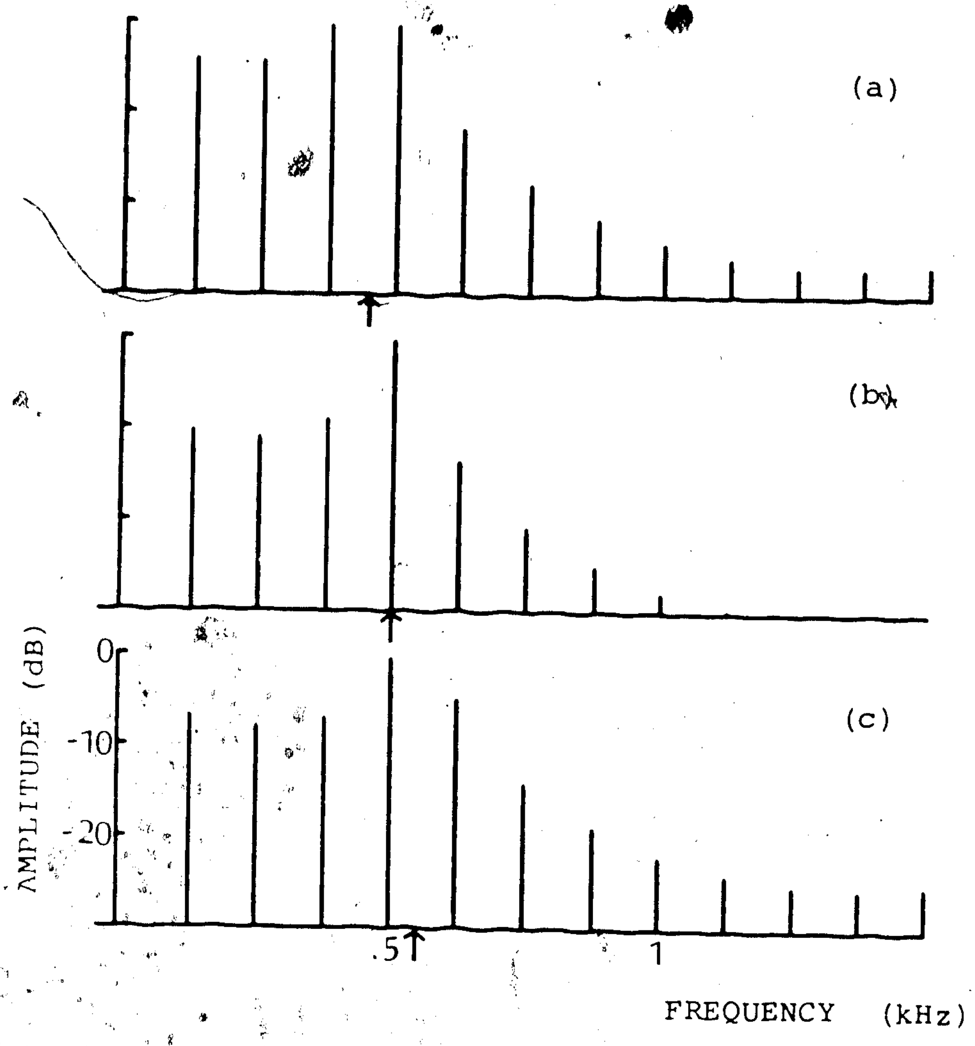


Figure 3.1 Effects of F1 changes on the amplitude spectrum of a front vowel.
f0=125 Hz. F1 indicated by arrows
(a) F1=450 Hz
(b) F1=500 Hz
(c) F1=550 Hz

Note the asymetrical shape of the formant (steeper high frequency slope) in all 3 spectra. This asymmetry in spectrum levels reflects the lowpass nature of the first formant (Fant, 1956).

Carlson, Fant and Granstrpm (1975) proposed that the weighted mean of the two most prominent harmonics in the F1 region could account for the perceptual effects of f_0 in their identification experiment. They suggested the following model:

3.2 $$F_m = (f_{i1} \cdot S_{i1} + f_{i2} \cdot S_{i2})/(S_{i1} + S_{i2})$$

where f_{i1} and f_{i2} are the frequencies of the two most prominent harmonics i_1 and i_2; S_{i1} and S_{i2} are their loudness density levels. Carlson et al. proposed a preemphasis of the spectrum in the region below 600 Hz by 6 dB/octave to approximate loudness equalization. F_m was a continuous function of the fundamental frequency, whereas the frequency of the most prominent harmonic showed abrupt discontinuities with increasing f_0. The former appeared to be more consistent with the perceptual data than the latter.

There may be an additional factor influencing these results. Previous studies have shown that fundamental frequency changes may result in perceptual "speaker normalization" effects (Fujisaki and Kawashima, 1968; Ainsworth, 1975). There is a correlation between fundamental frequency and formant ranges in natural speech measurements (Peterson and Barney, 1952). Listeners in the Carlson et al. experiment may have interpreted a change in fundamental frequency as reflecting a change in the implied speaker, and indirectly, as a change in vocal tract size. Slawson (1968) found that F1-F2 changes of 10-15 percent were needed to compensate for an octave increase in f_0. This perceptual compensation is in rough agreement with the relationships in natural speech data. The extent to which perceptual compensation for changes in perceived speaker identity may have influenced the boundary shifts observed by Carlson et al. is unclear. It is therefore desirable to obtain evidence from other sources as a further test of their hypothesis.

3.2.4 Weighted mean of the three most prominent harmonics (FMPH3 hypothesis)

This proposal was examined by Carlson, Fant and Granstrom (1975) as an alternative to the FMPH2 hypothesis. According to this model, the height of a vowel depends on the weighted mean of the three most prominent harmonics:

$$3.3 \quad F_m = \{(f_{i1} \cdot S_{i1}) + (f_{i2} \cdot S_{i2}) + (f_{i3} \cdot S_{i3})\} / (S_{i1}+S_{i2}+S_{i3})$$

where S_{i1}, S_{i2}, and S_{i3} are the loudness densities of the three most prominent (contiguous) harmonics in the region of the F1 peak, and f_{i1}, f_{i2} and f_{i3} are their frequencies.

Carlson, Fant and Granstrom (1975) reported that this measure was more strongly affected by changes in f_0 than the weighted mean of the two most prominent harmonics; the effects of f_0 were larger than might be expected on the basis of the identification data.

3.2.5 Weighted mean in the entire F1 region

Each of the hypotheses discussed above involves an explicit isolation and ranking of the relative prominence of harmonic components in the F1 range. If it is assumed that all harmonics in the F1 region contribute equally to vowel height, this procedure may be unnecessary. Vowel height may be estimated by the first moment (centroid) of the spectrum over the entire F1 range.

Beddor (1984) and Beddor and Hawkins (1984) used a measure of this type to predict the results of a matching experiment in which listeners adjusted the first formant frequency of a synthetic oral vowel to match the phonetic quality of a corresponding nasal vowel. They computed the centroid of the LPC spectrum (log magnitude versus frequency) in the F1 region of the nasal reference vowels. It was found that listeners frequently selected matches whose F1 values were intermediate between F1 of the reference vowel and its centroid. Centroid calculations appeared to assign too much weight to energy distant from the F1 peak and hence did not yield accurate predictions of the oral-nasal matching data. Nonetheless, the direction of the departure from F1 matches was successfully predicted by the centroid.

3.2.6 LPC based estimates of the first formant

There are a number of alternative procedures for estimating formant frequencies based on models of speech production. One of the most commonly used methods is LPC analysis, which models the speech waveform by means of an all-pole digital filter (Makhoul, 1975; Markel and Gray, 1976).

In LPC analysis, speech sample points are approximated as a linear combination of previous sample points. The predictor coefficients are determined by minimizing the mean square error between observed and predicted samples. These coefficients represent the filter coefficients of the all-pole model which reflects the combined effects of the glottal source, vocal tract and radiation characteristics. The all-pole model provides a close fit to vowel spectra; the degree of interpolation is determined by the number of coefficients. Because the error criterion assigns greater weight to regions of the spectrum whose intensity levels are underestimated by the model, the LPC spectrum provides a closer fit in the region of spectral peaks than in the spectral valleys (Makhoul, 1975; Rabiner and Schafer, 1978).

There are several alternative procedures for determining the filter coefficients, including the autocorrelation method, the covariance method, and the lattice method. The autocorrelation method is frequently used in formant estimation. Formant frequencies can be obtained from LPC analysis either by solving for the roots of the predictor polynomial, or by evaluating the LPC spectrum and using a peak selection algorithm to locate spectral maxima corresponding to formants. In the present analysis the autocorrelation method was adopted, and a peak selection algorithm was applied to estimate the frequency of the lowest peak in the LPC spectrum. If spectral interpolation followed by peak selection is a reasonable model of the perception of vowel height, then it might be expected that LPC based estimates of the F1 peak will closely model the matching data.

In the following matching experiments direct tests of each of these hypotheses were carried out. In addition, procedures derived from each hypothesis were implemented and comparisons drawn between the resulting predictions and the observed data.

3.3 Experiment 1: The effects of harmonic amplitudes on vowel height

3.3.1 Introduction

The FMPH hypothesis of Mushnikov and Chistovich (1971, 1973) maintains that vowel height is determined by the frequency of the most prominent harmonic in the F1 region. Harmonic amplitude changes will have no effect on vowel height, according to this hypothesis, unless such changes alter the frequency location of the most prominent harmonic. As the most prominent harmonic is attenuated, therefore, a single discrete change in vowel height is expected when this harmonic is no longer the most prominent. An alternative possibility is that vowel height depends on a perceptual weighting of two or more harmonics (Carlson, Fant and Granström, 1975). If so, it is expected that attenuation of the most prominent harmonic will lead to a gradual shift in vowel height.

A matching experiment was designed to test these hypotheses using synthetic vowel stimuli similar to those used by Mushnikov and Chistovich (1971, 1973). Listeners judged pairs of synthetic stimuli with respect to vowel height. One member of the pair was a reference stimulus in which a small number of harmonics (2-5) of equal amplitude was substituted for the first formant peak, with all additional components up to 1.5 kHz set to zero amplitude. The other member was a matching stimulus from an F1 continuum spanning the F1 range for the vowels /i I e ɛ æ/ in Canadian English (see Figure 3.2). The reference stimuli differed somewhat in timbre from full spectrum synthetic vowels, but the degree of approximation to natural vowel quality was not diminished. Their phonetic identities were preserved, and listeners had little difficulty labeling the stimuli as one of the Canadian English vowels /i I e ɛ æ/ (see Experiment 4).

The importance of the number of harmonics in the F1 region was investigated, as well as the effects of the amplitude level of the highest frequency harmonic ("upper edge" condition) and the lowest frequency harmonic ("lower edge" condition).

3.3.2 Method

Stimulus materials

All of the stimuli were produced digitally at a sampling rate of 16 kHz using a PDP-12 minicomputer and the OS/8 operating system. The amplitude spectrum for each of the matching stimuli was computed using the cascade formant synthesis model with 6 formants (Fant, 1960). The contributions of each formant, glottal source and radiation characteristics and higher pole correction factor were

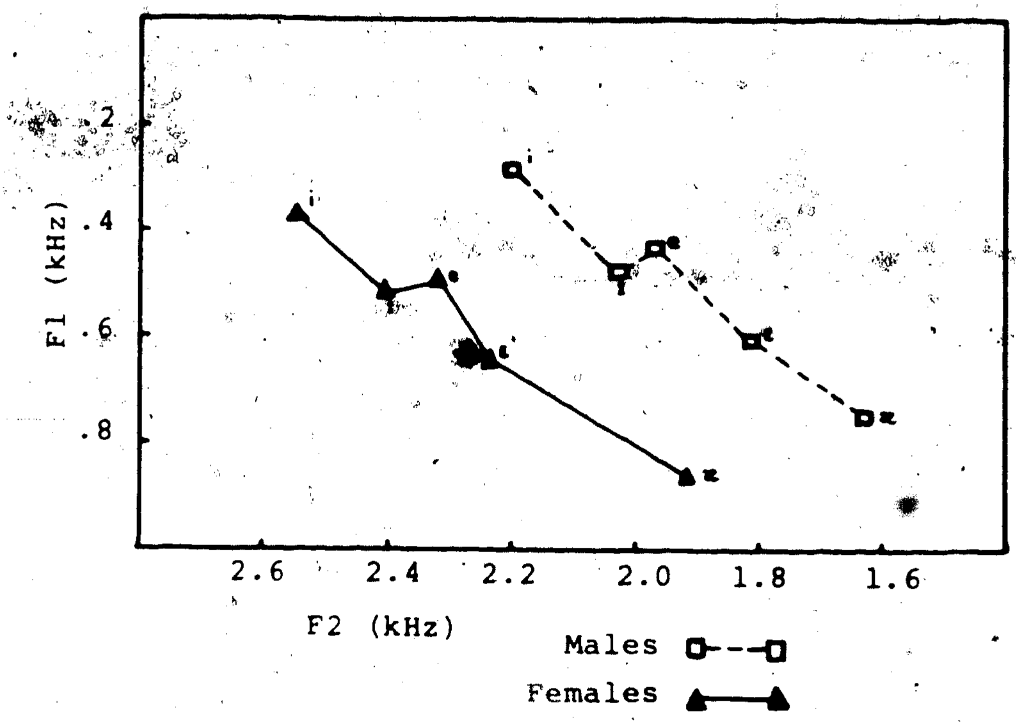


Figure 3.2 Average formant values for the front vowels of western Canadian English speakers

calculated as described by Flanagan (1972: pp.217-218).

The fundamental frequency of each stimulus was 125 Hz (a typical male voice pitch). The 25 matching stimuli formed an F1 continuum ranging from 175 to 850 Hz, in 25 Hz steps. The frequencies and bandwidths of the 6 formants are shown in Table 3.2. The frequencies of higher formants were set close to measured means; formant bandwidths were similar to the values suggested by Klatt (1980).

The baseline stimulus used to produce each of the reference stimuli was drawn from the middle of the F1 continuum (F1=500 Hz). Its amplitude spectrum was modified to generate a series of "upper edge" and "lower edge" reference stimuli.

"Upper edge" stimuli. Eighteen reference stimuli were constructed as follows: All harmonics in the 0 to 1.5 kHz region were removed and replaced with a band of either 3, 4, or 5 contiguous harmonics, the lowest member of which was always 125 Hz. The amplitude level of each component within the band was adjusted to the level of the highest amplitude harmonic in the F1 region of the baseline stimulus. The highest frequency component in the band, which was either the third (375 Hz), fourth (500 Hz) or fifth (625 Hz) harmonic, was attenuated by either 0, 3, 6, 9, 12, or 15 dB, relative to the peak amplitude value. For convenience, we will refer to this harmonic as the *upper edge harmonic* in subsequent discussions. Line spectra for each of the 18

F1 (see text)	B1= 70 Hz
F2= 2125 Hz	B2= 90 Hz
F3= 2750 Hz	B3= 150 Hz
F4= 3625 Hz	B4= 250 Hz
F5= 4500 Hz	B5= 300 Hz
F6= 6000 Hz	B6= 700 Hz

Table 3.1. Formant frequencies and formant bandwidths for the matching and reference stimuli of Experiment 1

reference stimuli (6 amplitudes x 3 component numbers) are shown in Figure 3.3. The frequency region 0-1 kHz is shown in the figure; harmonic levels at higher frequencies were identical in each of the reference stimuli.

"Lower edge" stimuli. The 18 lower edge reference stimuli were produced by replacing the F1 region (0-1.5 kHz) of the baseline stimulus with a band of either 2, 3 or 4 harmonics of equal amplitude, with the *highest* frequency harmonic fixed at 500 Hz[1].

The amplitude of each component within the band was set to the level of the harmonic of greatest amplitude in the F1 region of the baseline stimulus.

[1] A 5 harmonics condition was included in pilot studies, but was replaced with the 2 harmonics condition when it was discovered that stimuli in the former condition were very similar, while stimuli in the latter condition differed in vowel height.

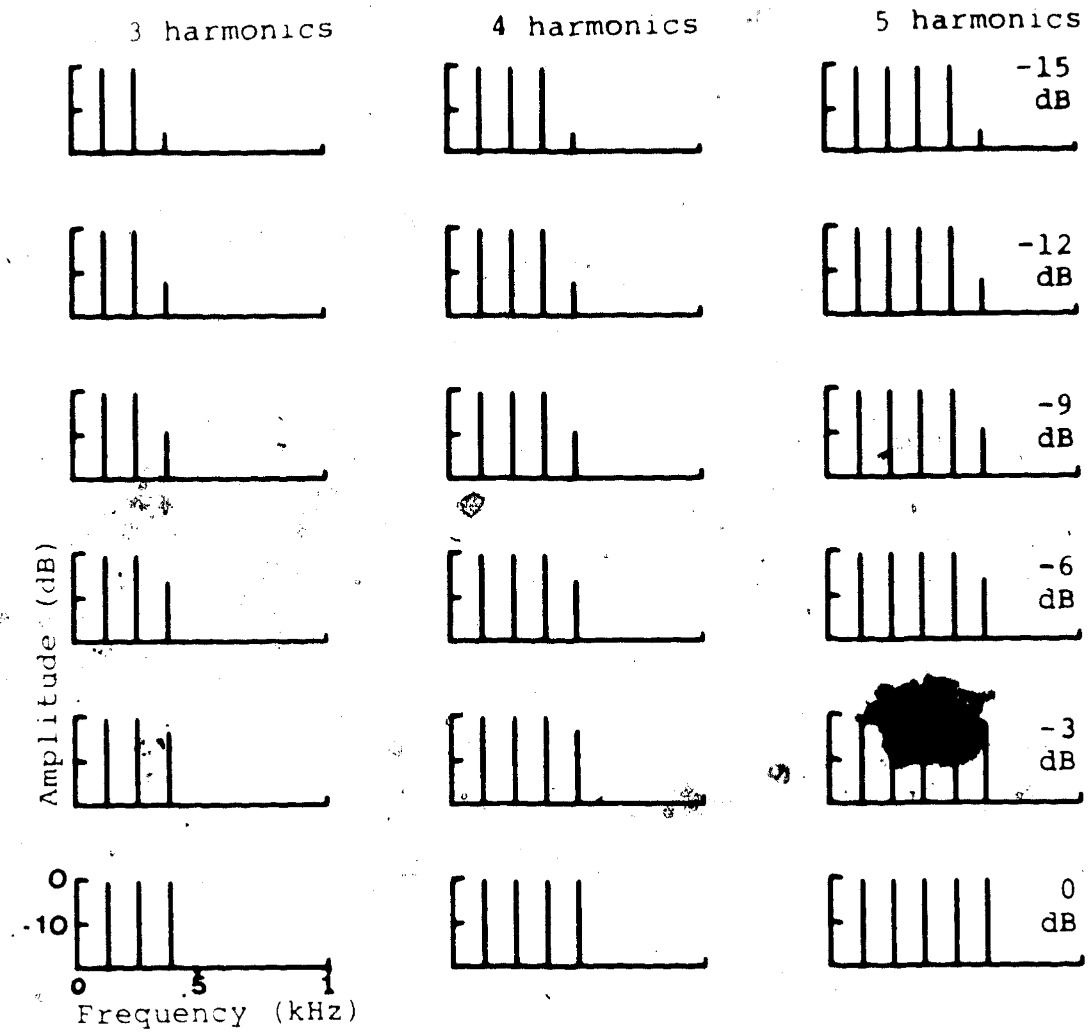


Figure 3.3 Line spectra of the upper edge stimuli used in experiment one

We will refer to the lowest frequency harmonic in each stimulus (with a frequency of 125 Hz in the 4 harmonics condition, 250 Hz in the 3 harmonics condition, and 375 Hz in the 2 harmonics condition) as the lower edge harmonic. This component was attenuated in 6 steps of 3 dB, ranging from 0 to -15 dB, as in the upper edge condition. Line spectra of each stimulus are shown in Figure 3.4.

A single pitch period of each stimulus was generated using additive harmonic synthesis, with all components in sine phase[1]. A total of 32 sinusoidal components (all harmonics up to 4 kHz) was summed. A number of pitch periods were concatenated to produce a 250 ms stimulus whose onset and offset were smoothed using the initial and final half portions of a 30 ms Hamming window.

The set of 36 reference stimuli and 25 matching stimuli were stored in computer disc files. In order to preserve absolute amplitude relations among the reference stimuli, the waveform of each one was scaled by the same scale factor. This procedure was not adopted in the matching stimuli because of the wide range of peak amplitudes (greater than 25 dB). Separate scale factors were therefore used for each of the matching stimuli, so that the peak

[1] Individual pitch periods of spoken vowels rapidly attain maximum amplitude, then decay exponentially over the course of their duration (Rosenberg, 1971). However, studies of phase sensitivity (reviewed in section 2.3.2) have indicated that phonetic quality in vowels is unaffected by changes in phase.

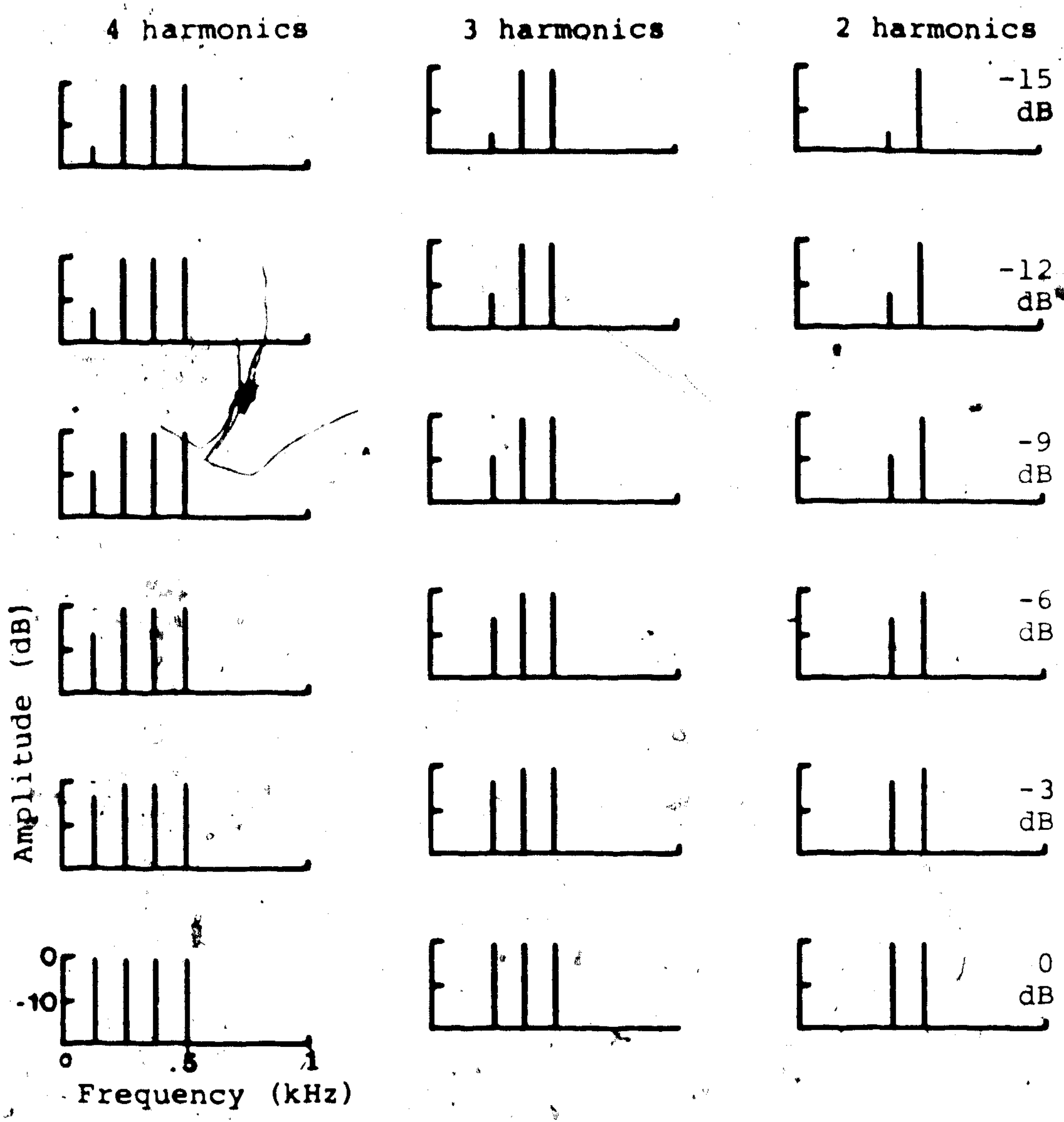


Figure 3.4 Line spectra of the lower edge stimuli used in experiment one

amplitude of the waveform was identical in each one.

Procedure

The stimuli were stored in computer disc files and presented on-line using a 10-bit D/A converter, and transmitted to a sound laboratory via shielded lines. The stimuli were bandpass filtered using a Wavetek Rockland Series 1520 filter (68 Hz to 6800 Hz, 24 dB/octave spectral rolloff) and amplified using a Braun AG Amplifier (Type [illegible]V 250) with modified output impedance. The measured signal-to-noise ratio of the complete presentation system (including filter, amplifier and headphones) was approximately 55 dB.

The signal level was checked by means of a Digital Multimeter (Hewlett-Packard 3469B) at the beginning of each session, and adjusted to a level of 82.5 dB SPL for a 1 kHz sinusoid with the same peak amplitude level as each of the matching stimuli. The absolute intensity setting was established with the aid of a Bruel and Kjær artificial ear (Type 4153) calibrated with a Bruel and Kjær Pistonphone (Type 4230). The signals were presented to listeners in a sound-treated room, and delivered to the right ear with a set of Telephonics TDH49 headphones. The presentation levels ranged from 79 to 85 dB SPL. The response of the entire system was flat (less than 5 dB maximum deviation) in the range 100-4000 Hz, as measured using a Bruel and Kjær

Frequency Analyzer (Type 2120). A block diagram of the stimulus presentation facilities is shown in Figure 3.5.

The stimuli were presented in pairs. The first member of each pair was one of the 36 reference stimuli. Different randomization lists were used to specify the order of items for each replication and for each listener. The second member of the pair, presented after a 3 second intra-pair interval, was a matching stimulus from the continuum whose F1 value was determined by the position of a slide lever adjusted by the subject. There were 2 replications of the 36 reference stimuli in each experimental session, and each listener completed 5 sessions.

The slide lever used in matching controlled a potentiometer which attenuated an input voltage of 5 volts from a Lambda Dual Regulated Power Supply (Model LPD 422 FM). The output level was sampled through an A/D converter and the entire 5 volt range was quantized into 25 equal steps, each specifying a single stimulus on the F1 continuum. A five second inter-pair interval was used.

Listeners were instructed to adjust the position of the slide lever to find the best match in vowel quality. When satisfied with the match they pressed a switch on a response box. The F1 value of the best match was automatically recorded, and the next pair was presented. A pause button which interrupted stimulus presentation was also available, to avoid listener fatigue. Listeners were allowed as many

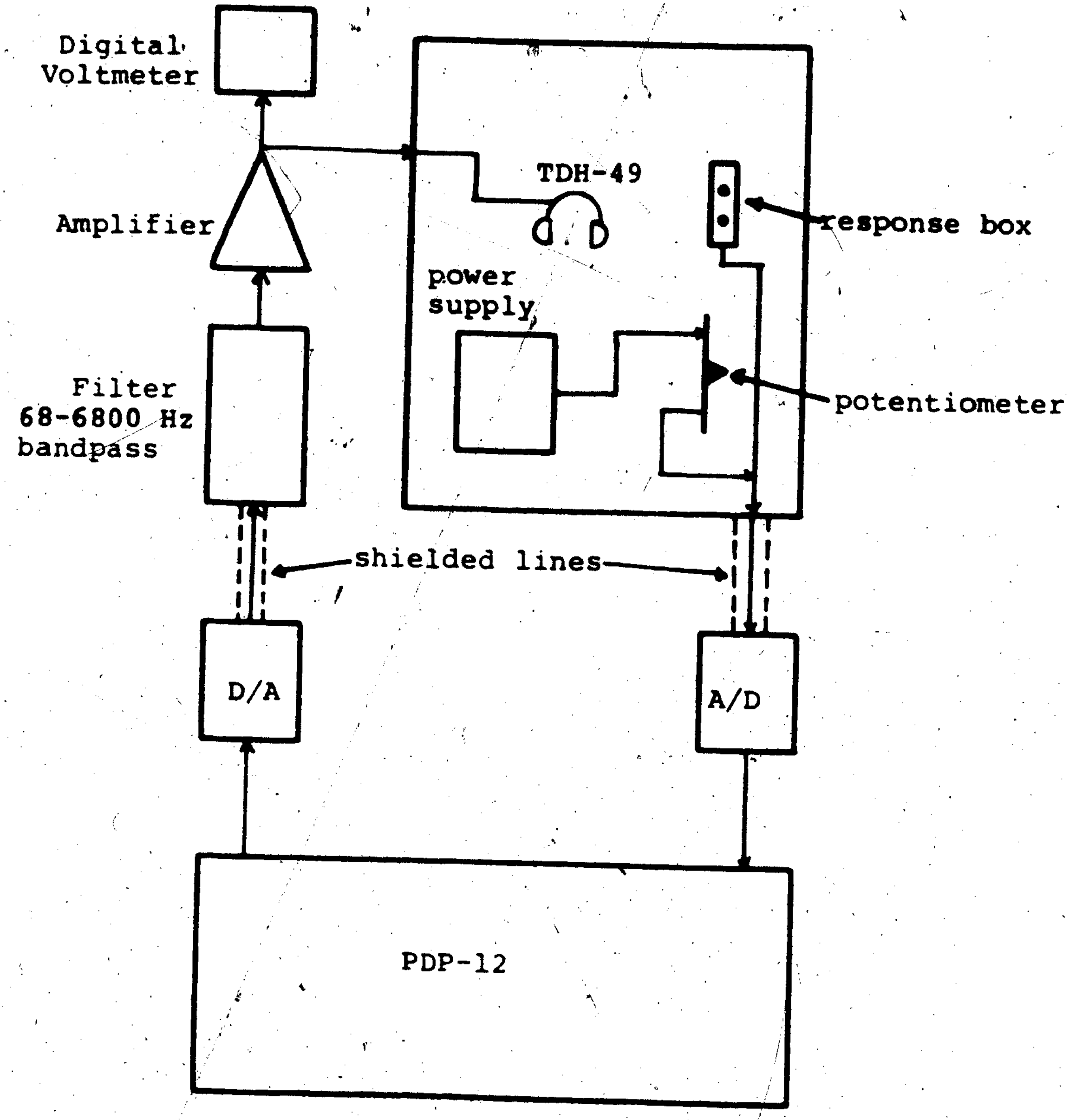


Figure 3.5 Block diagram of stimulus presentation facilities

trials as needed to arrive at a satisfactory match. They were instructed to adjust the position of the lever several times for each reference stimulus, so that the reference was compared with a number of matching stimuli. Experimental sessions lasted approximately 15 to 20 minutes.

Listeners

Six listeners (3 men and 3 women, ranging in age from 25-40 years) completed the task. All but one had received some prior training in phonetics, and had participated in other speech perception experiments involving natural or synthetic stimuli. All had normal hearing (less than 15 dB HL) in the right ear as determined by audiometric testing (Grason-Stadler Model 1703B Recording Audiometer). None of the subjects had any difficulty with the task.

3.3.3 Results and discussion

The experimental results for the upper edge and lower edge conditions differed substantially, and will be discussed separately.

1. **Upper edge stimuli.** Histograms of the matching data (representing the number of times each matching stimulus was selected, for each of the 18 reference stimuli) are shown in Figure 3.6. Each individual histogram represents a total of 60 matches (10 trials from each of the 6 listeners). Matches are generally clustered around a single mode, with

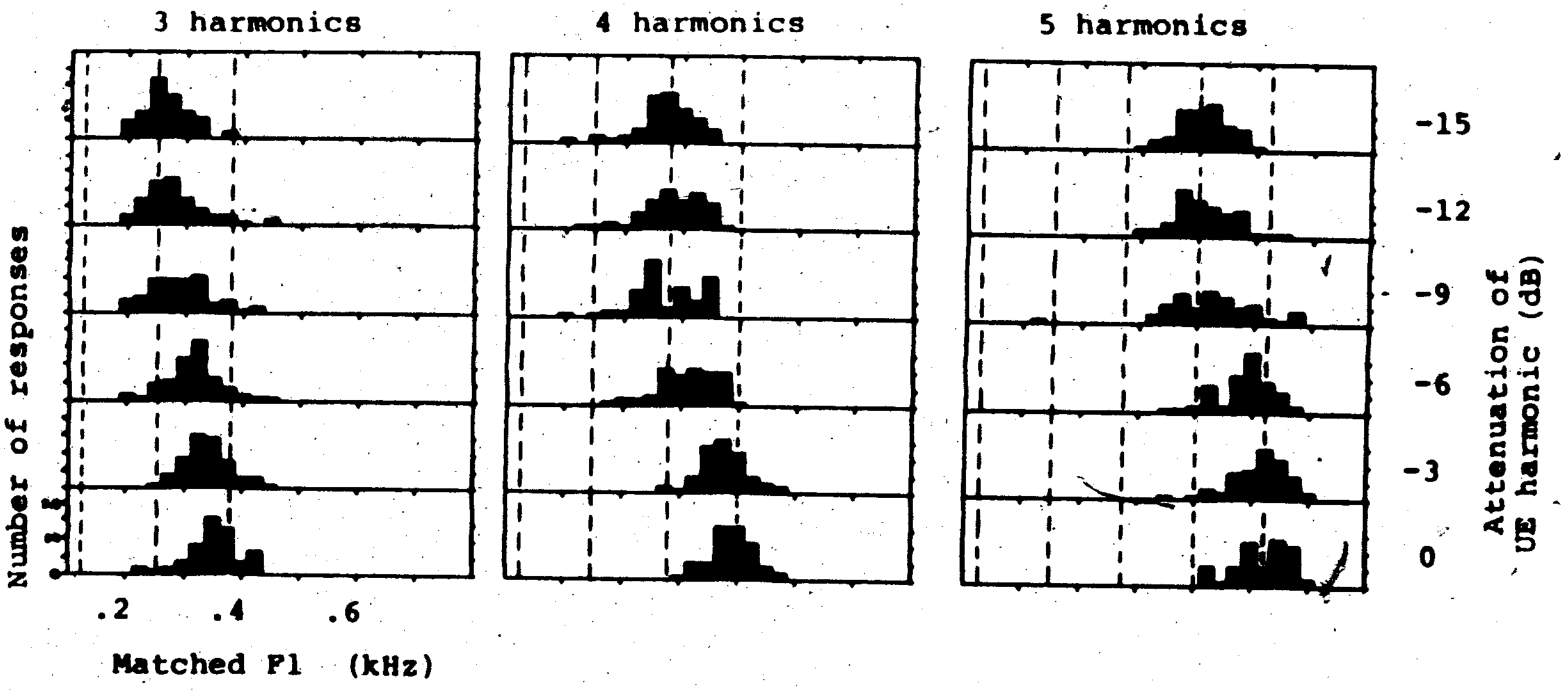


Figure 3.6 Histograms of matching data for the upper edge (UE) condition of experiment one.

Dashed lines indicate the frequencies of the harmonics of the reference stimuli most matches within 100 Hz of the mode. It is apparent that attenuation of the upper edge harmonic resulted in gradual shifts toward lower matched F1 values, down to at least -9 dB.

These gradual changes in matched F1 argue strongly against the hypothesis that vowel height matching depends solely on the frequency value of the most prominent harmonic; rather, they suggest a perceptual weighting of the levels of two or more harmonics.

The matching data from each listener was summarized in terms of the median matched F1 values across the 10 trials. Means and standard deviations (across listeners) of the median F1 matches are shown in Figure 3.7. The 3 curves follow a similar pattern, with a frequency separation close to the spacing between the harmonic components (125 Hz). A gradual decrease in matched F1 is seen as the amplitude of the upper edge harmonic is reduced[1].

The F1 peak of the matching stimuli was positioned near the upper edge harmonic of the reference. Attenuation of the upper edge harmonic led to a gradual decline in mean matched F1, indicating that the upper edge harmonic is important in matching judgements. Figure 3.7 indicates that

[1] Gradual shifts can be observed in the data for each individual listener (see Appendix A), eliminating the possibility that the shifts observed in the average data are the result of single, abrupt shifts in the data from individual listeners which take place at different amplitude levels.

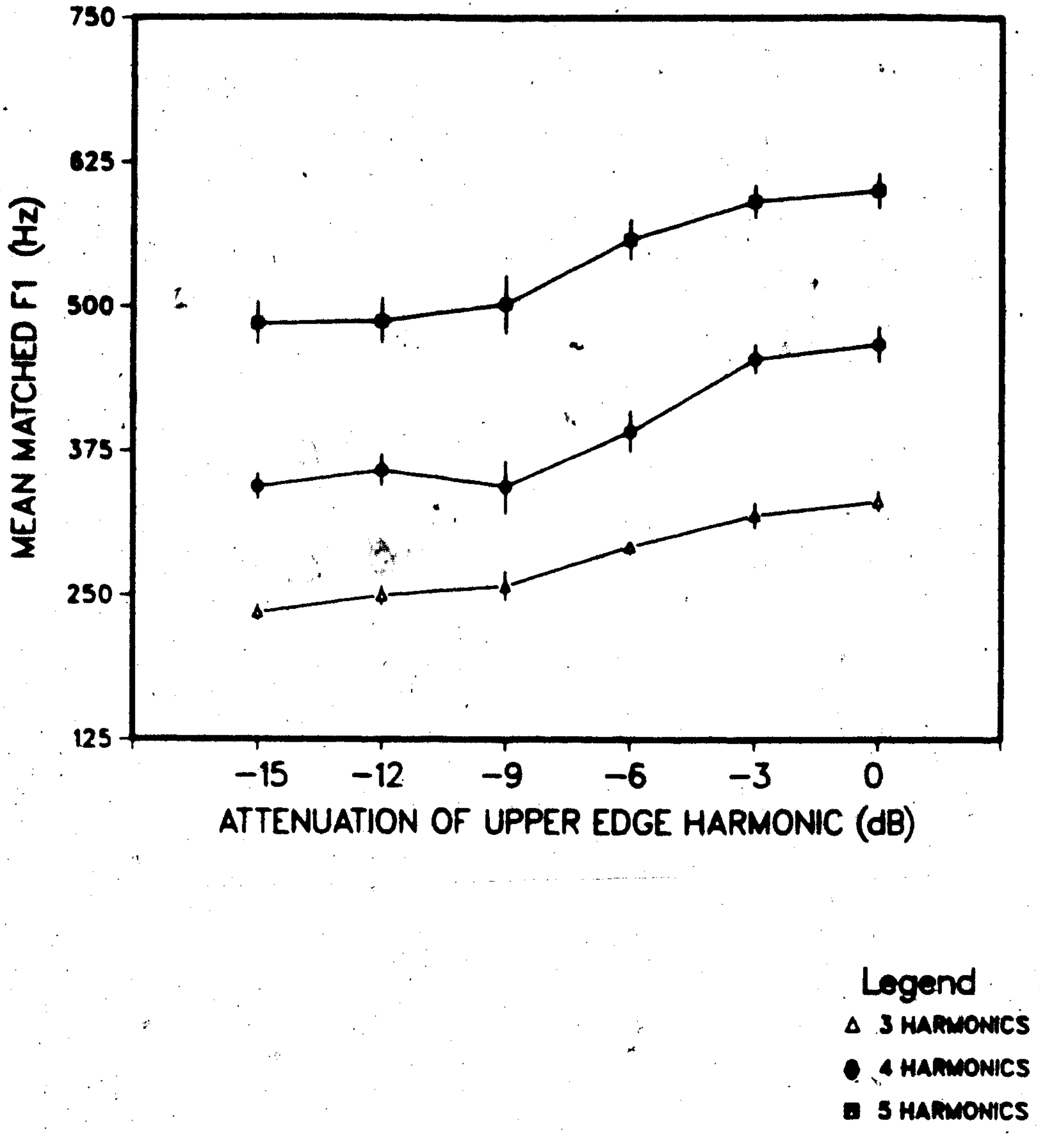


Figure 3.7 Mean matched F1 values for the upper edge condition of experiment one

the mean matched F1 for the 0 dB stimuli occurs just below the frequency value of the upper edge harmonic. Stimuli with 15 dB attenuation were matched with mean F1 values slightly lower than the second highest harmonic.

A repeated measures analysis of variance was performed on the median matched F1 data with the factors AMPLITUDE (level of the upper edge harmonic) and NUMBER OF HARMONICS (3, 4, or 5 components in the F1 region). A significant main effect was found for AMPLITUDE (F(5,25)=115.4; p<.001) and NUMBER OF HARMONICS (F(2,10)=321.9 ;p<.001), but the interaction of these two factors was not significant.

Post hoc tests (Tukey HSD procedure) were applied to compare means within the factor AMPLITUDE (Table 3.2). Significant differences were found between most pairs of means: lower F1 matches were selected with increasing attenuation. If vowel height is determined solely by the frequency of the single most prominent harmonic, one would expect at most one shift in matched F1 with progressive attenuation of the most prominent harmonic. The finding of more than one statistically significant shift with attenuation of the upper edge harmonic indicates that the FMPH hypothesis can be rejected[1].

----

[1] Informal listening to the entire set of reference stimuli suggested instead a fairly continuous change in vowel height with decreasing amplitude and number of harmonics, ranging from 0 dB attenuation in the 5 harmonics condition (an /ɛ/-like vowel) to 15 dB attenuation in the 3 harmonics condition (an /i/-like vowel).

	0	-3	-6	-9	-12	-15 dB
0		11.2	51.9**	97.8**	100.7**	111.1**
-3			40.7**	86.6**	89.4**	99.9**
-6				59.2**	48.7**	46.9**
-9					13.3	2.8
-12						10.5

** p<.01

Table 3.2. Differences in means and significance levels for Tukey HSD tests in the upper edge condition of Experiment 1

The reference stimulus with 0 dB attenuation in the 4 harmonics condition may also be viewed as a 5-harmonic stimulus with complete attenuation of the upper edge harmonic; the 0 dB stimulus in the 3 harmonics condition may be considered as a 4 harmonic stimulus with complete attenuation of the upper edge harmonic. In both cases, the -15 dB stimulus is matched slightly higher than the corresponding stimulus with full attenuation of the harmonic, suggesting that a small difference in vowel height may be detected.

Amplitude spectra are shown in the upper portion of Figure 3.8 for each reference stimulus and its best match (medians from 10 trials x 6 listeners). In the bottom portion of the figure, 6 dB/octave preemphasized spectra are shown for the same stimuli. The most prominent harmonic in

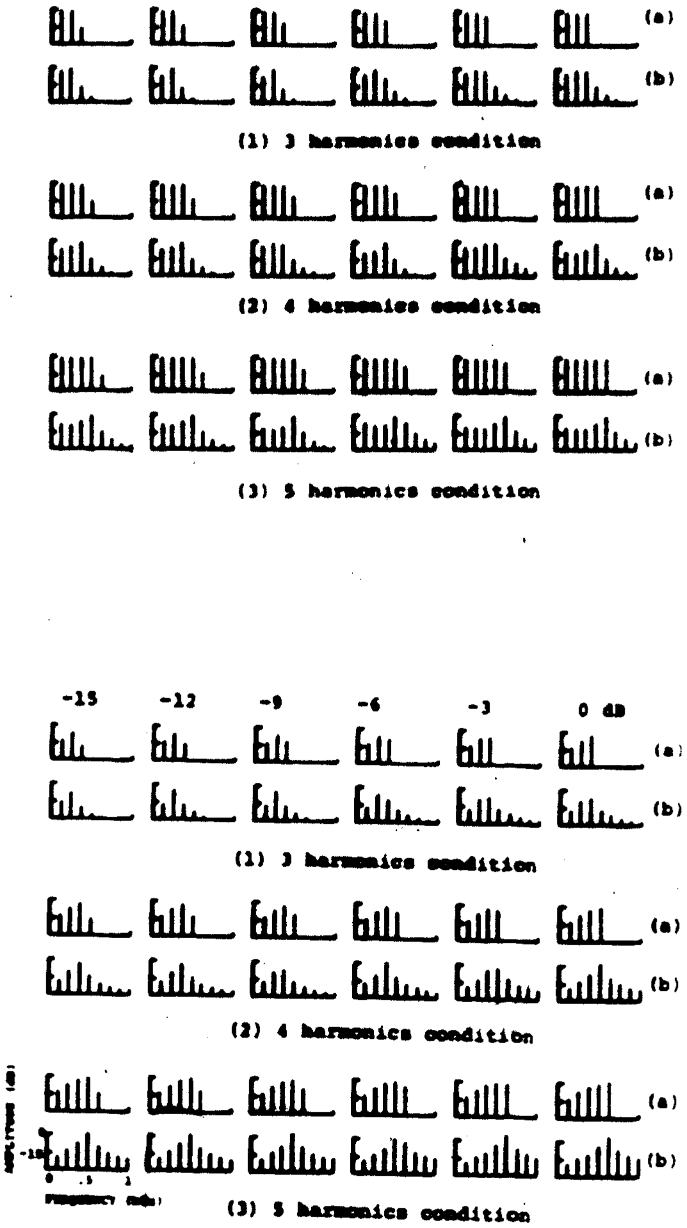


Figure 3.8 Line spectra of the upper edge reference stimuli of experiment one, (a), and "best" matching stimuli, (b).
Top panel: unpreemphasized spectra
Bottom panel: spectra with 6 dB/oct preemphasis

the amplitude spectrum of the matching stimulus coincides with the upper edge harmonic in all of the 0 dB and -3 dB stimuli; and with the adjacent harmonic for reference stimuli with more than 3 dB attenuation. Listeners do not appear to use a global averaging procedure, since this would predict lower F1 matches to compensate for the absence of energy in the F1 region above the upper edge harmonic of the reference stimuli.

The two most prominent harmonics in the preemphasized spectra of the matching stimuli are aligned in nearly all cases with the upper edge and adjacent harmonic of the reference. The preemphasized spectra of "best matched" vowel pairs appear to be more similar in shape than the corresponding unpreemphasized spectra.

2. **Lower edge stimuli.** Histograms of the matching data (10 trials from 6 listeners) are shown in Figure 3.9 for each of the lower edge stimuli. Unlike the upper edge condition, the effects of amplitude are very small, and there do not appear to be systematic changes as a function of amplitude. There is a trend, however, toward higher F1 matches with decreasing amplitude in the 2 harmonics condition.

Median matched F1 values were determined for each listener. Means and standard deviations (across the 6 listeners) are shown in Figure 3.10. A repeated measures analysis of variance was conducted on the median F1 match values, as described previously for the upper edge stimuli.

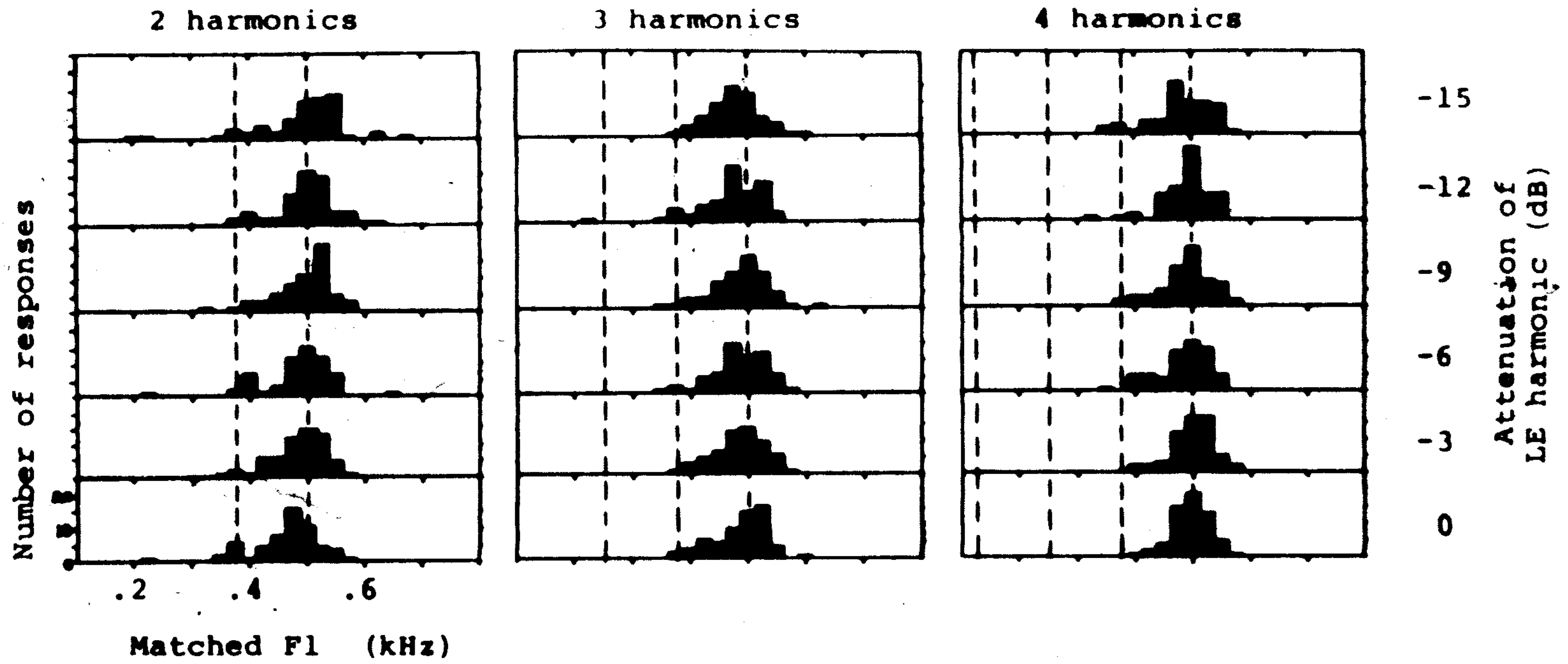


Figure 3.9 **Histograms of matching data for the lower edge (LE) condition of experiment one**
Dashed lines indicate the frequencies of the harmonics of the reference stimuli

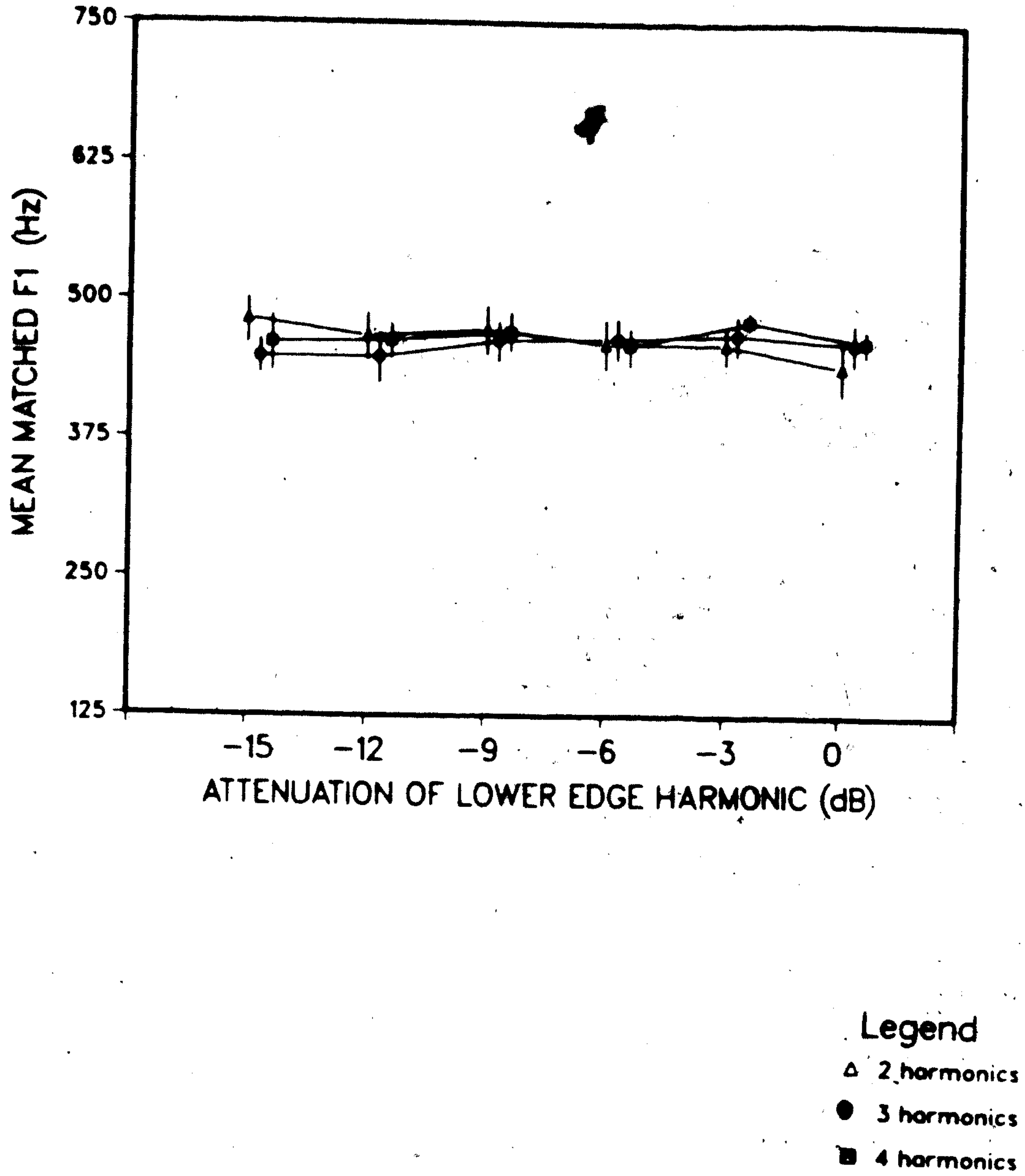


Figure 3.10 **Mean matched F1 values for the lower edge condition of experiment one**

No significant main effects were found for AMPLITUDE or NUMBER OF HARMONICS. However, there was significant interaction of these two factors (F(10,50)=2.62; p<.02).

Tukey HSD tests indicated significant effects of amplitude only in the two harmonics condition. These involved comparisons between the 0 and -9 dB conditions (a difference of 33.3 Hz; p<.05) and 0 and -15 dB conditions (a difference of 41.7 Hz; p<.01). The general trend toward higher F1 matches with decreasing amplitude of the lower edge harmonic is consistent with a change in the centre of gravity or weighted mean of the 2 harmonics: a decrease in the amplitude of the lower edge harmonic raises the apparent vowel height, while a decrease in the amplitude of the upper edge harmonic lowers it. However, the maximum shift for the lower edge condition (with 0 and -15 dB attenuation) was only 42 Hz, less than half of the F1 shift observed for corresponding stimuli in the upper edge condition.

These findings indicate that at least two prominent harmonics can affect vowel height. Attenuation of the upper edge harmonic lowered the mean matched F1, while attenuation of the lower edge harmonic resulted in significant shifts only for the 2 harmonics condition, where significantly higher F1 matches were observed. Shifts in the lower edge condition were smaller than corresponding shifts in the upper edge condition. In both upper and lower edge conditions, the F1 peak in the matching stimulus was often close to the frequency of the upper edge harmonic of the reference.

3.3.7 Predictions of the matching data

In this section several spectral measures for predicting the matching data will be considered. These measures are derived from the hypotheses considered in section 3.2, and will be used to predict the "best match" (along the F1 continuum) for each reference stimulus. The relative merits of different hypotheses can be assessed in terms of the degree of fit of the predictions to the observed data.

Spectral measures X_i and X_j were computed for each reference stimulus i and each matching stimulus j, respectively. These measures were used to calculate a spectral distance D_{ij} between reference stimulus i and matching stimulus j:

$$D_{ij} = |X_i - X_j| \qquad (3.1)$$

In preliminary analyses, a minimum distance criterion was used to determine the predicted match for each reference stimulus. However, this criterion was found to be unsatisfactory because it often happened that several matching stimuli were equidistant from the reference. For this reason, a weighted spectral similarity measure was adopted.

(1) For each matching stimulus j, a weight W_j was calculated as the inverse square of its spectral distance to reference stimulus i:

3.2 $W_j = 1/(X_i - X_j)^2$

where X_i and X_j are the measured values for the ith reference stimuli and jth matching stimulus[1].

(2) The F1 value of the predicted match for the ith reference stimulus, $PF1_i$, was computed as:

3.3 $PF1_i = \Sigma_j^{nj} (F1_j \cdot W_j) / \Sigma_j^{nj} W_j$

where $F1_j$ represents the nominal F1 value (the actual formant value specified for the synthesizer) of the jth matching stimulus, and nj represents the total number of matching stimuli. The RMS distance between observed and predicted F1 values was computed as an index of the degree of fit:

3.4 $RMS = [1/(ni \cdot nj)\Sigma_i^{ni} \Sigma_k^{nk} (OF1_{ik} - PF1_i)^2]^{1/2}$

where $OF1_{ik}$ represents the median matched F1 value for the ith reference stimulus from the kth listener; ni is the total number of reference stimuli and nk is the total number

[1] This formulation takes into account the relative proximity of all matching stimuli to the reference (rather than just the closest match). Matching stimuli which are close to the reference (i.e. with the smallest values of D_{ij}) contribute more to the weighting function than those which are distant from the reference.

of listeners.

We will first consider several local peak estimation procedures, based on prominent harmonics near the first formant peak. Each procedure uses as its starting point the harmonic line spectrum of the vowel, specified by the synthesis procedures described in section 3.3.2. For vowels with fundamental frequencies greater than approximately 100 Hz, the separation of harmonics in the F1 region is larger than the critical bandwidth. Psychophysical data from several sources indicates that these components can be resolved by the ear. It is important, therefore, to consider the harmonic structure of vowels in the perception of F1 differences.

FMPH hypothesis

The results of this experiment have provided direct evidence that vowel height is not determined exclusively by the frequency of the most prominent harmonic in the F1 region as proposed by Mushnikov and Chistovich (1971). However, it is still of interest to see how well predictions based on the frequency of the most prominent harmonic can account for the matching data.

We will define the "F1 region" as the frequency range up to and including 1.5 kHz; the expression "F1 harmonics" will be used to refer to components within this range. This range will always incorporate the F1 peak, but will also include F2 in back vowels (e.g., see Peterson and Barney, 1952). The problem of formant range overlap will be considered in chapter four.

The "most prominent harmonic" was initially defined as the largest component in the F1 region of the amplitude spectrum. For 31 of the 36 reference stimuli, there were 2 or more F1 harmonics with the same peak amplitude. Having noted that the F1 peak was aligned near the upper edge harmonic, it was decided to select the highest frequency component in case of a tie.

Predictions based on the frequency of the most prominent harmonic in the F1 region of the amplitude spectrum are shown in Figure 3.11 (solid lines) along with the observed means (dashed lines). Predicted F1s for the upper edge condition were mostly within 100 Hz of the observed means. The overall RMS distance between observed and predicted F1s was 50.9 Hz.

The predicted values show a single, abrupt decline between between 0 and -3 dB; with more than 3 dB attenuation the predicted F1 values are constant. The difference between the predicted F1 values in the 0 and -3 dB conditions was close to the average separation of the 3 curves; however, the observed shift was a gradual one, suggesting a weighting of 2 or more harmonics.

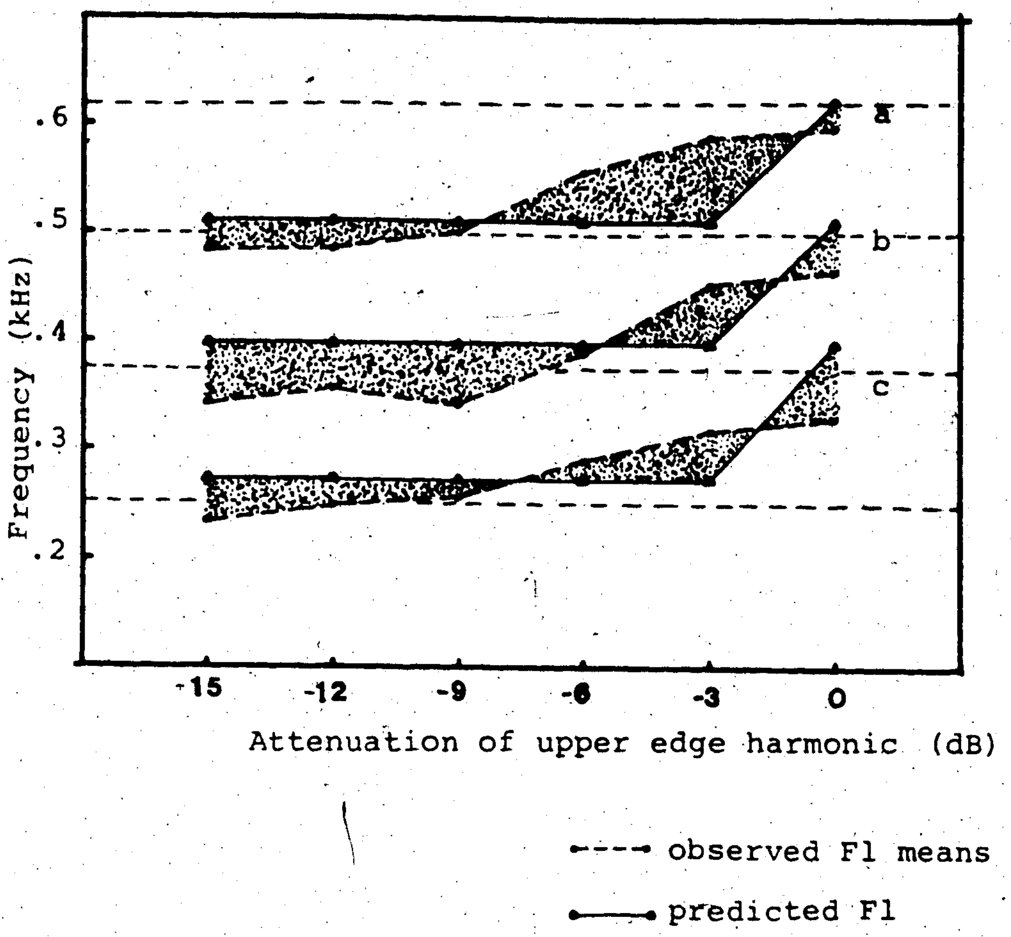


Figure 3.11 Model predictions for the upper edge condition of experiment one based on the frequency of the most prominent harmonic in the F1 region.

(a) 5 harmonics condition
(b) 4 harmonics condition
(c) 3 harmonics condition

Dashed lines indicate harmonic frequencies

Predictions in the lower edge condition were constant; all reference stimuli were predicted to be matched with the same F1 value, close to the upper edge, or fourth harmonic. These predictions resulted in an RMS value of 61.5 Hz. The predicted values were close to observed means, but failed to account for the increase in matched F1 with attenuation of the lower edge harmonic in the 2 harmonics condition.

T-tests were carried out to compare the observed medians with predictions based on the FMPH hypothesis. Significant differences ($p<.05$) were found for 9 of the 18 upper edge reference stimuli and 14 of the 18 lower edge reference stimuli, indicating a significant lack of fit for a substantial proportion of the stimuli.

Excitation and loudness pattern model predictions

Predictions based on the frequency value of the largest amplitude harmonic in the F1 region do not account for certain important features of the matching data. However, Mushnikov and Chistovich's (1971) hypothesis actually states that vowel height differences depend on the frequency of the most prominent harmonic in the low frequency region of the auditory spectrum, or loudness density pattern (Zwicker and Feldtkeller, 1967; Schroeder, Atal and Hall, 1979).

To investigate this hypothesis, excitation patterns were calculated using the model proposed by Sekey and Hanson (1982). A filterbank analysis was performed using a set of critical band filters, evenly spaced at .1 Bark intervals along the scale of critical band rate, from 0 to 10 Bark. Each filter had a bandwidth equal to one Bark, with a filter shape inferred from psychophysical data (Zwicker, 1963). The low frequency skirt had a +25 dB/Bark slope, while the high frequency skirt declined at a rate of -10 dB/Bark (Sekey and Hanson, 1982). Filtering was simulated by weighted summation of power spectra. An equal loudness preemphasis function was applied (Hermansky, Hanson and Wakita, 1985), followed by a .3 power transformation taking into account the relationship between intensity and loudness (Stevens, 1966).

An alternative model proposed by Moore and Glasberg (1983) for calculating excitation patterns was also investigated. Their procedure, which was discussed in section 2.3.6, employs a rounded exponential filter shape, with filter characteristics inferred from notched noise masking data (Patterson, 1976). The bandwidths of the filters are somewhat narrower than critical bands, and contipue to decline below 500 Hz where the critical bandwidth is nearly constant. Filters were separated by a frequency interval of 7.8 Hz.

Excitation patterns for each of the reference stimuli based on the model of Sekey and Hanson (1982) are shown in Figures 3.12 (upper edge condition) and 3.13 (lower edge condition). Each pattern represents the excitation level

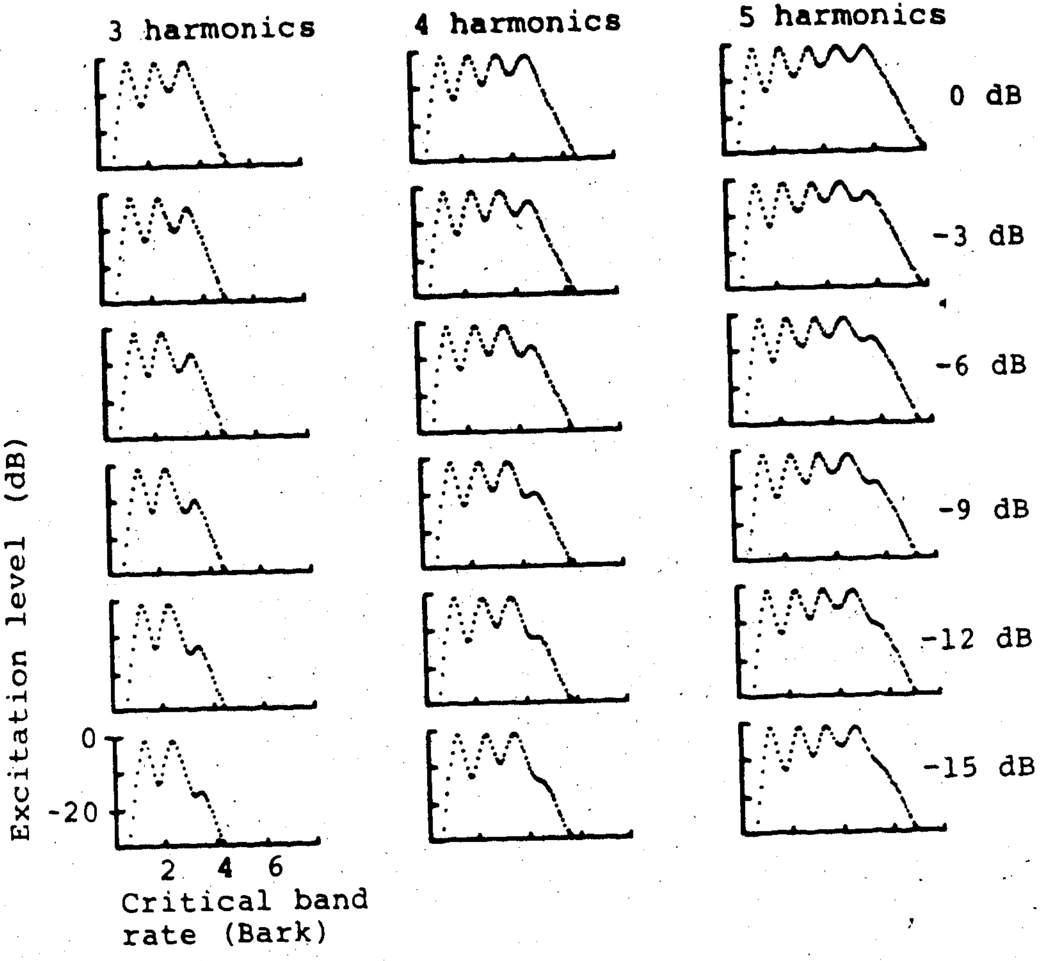


Figure 3.12 Excitation patterns for the upper edge stimuli of experiment one

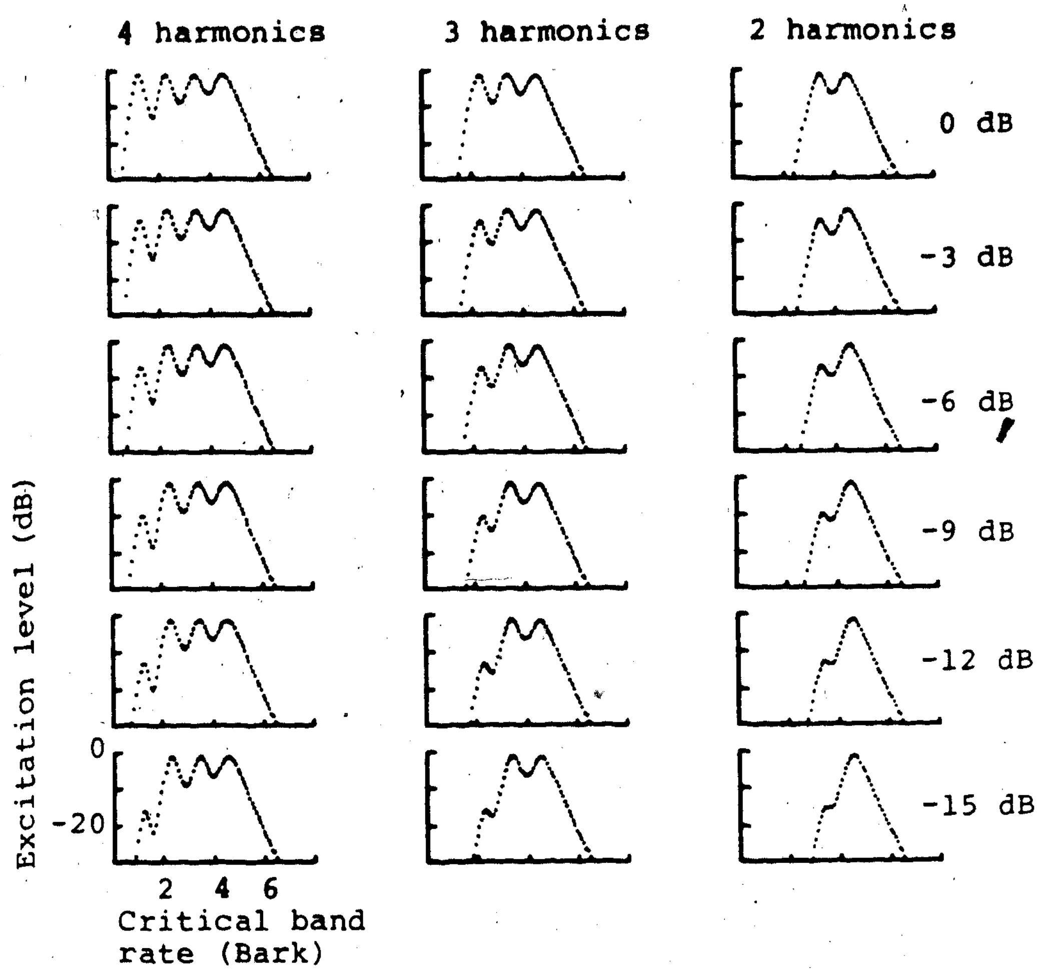


Figure 3.13 Excitation patterns for the lower edge stimuli of experiment one

(in dB) as a function of the critical band rate (in Bark). The low frequency region (0 to 8 Bark) is shown here. Individual harmonics (with the exception of the attenuated harmonic for some of the stimuli) are clearly resolved as peaks in the excitation patterns. Note that while the depth of the valley between each peak decreases with harmonic rank, the level of each peak is fairly similar. There are small differences in the heights of the peaks: the excitation level of the higher frequency member of a pair of harmonics was always greater than the lower.

Predictions based on the frequency of the largest peak in the F1 region of the loudness pattern gave very similar results to predictions based on the frequency of the most prominent harmonic in F1 region of the amplitude spectrum. The loudness pattern model resulted in an RMS distance of 51.9 Hz for the upper edge condition, and 51.8 Hz for the lower edge condition.

Analogous results were obtained using the excitation pattern model of Moore and Glasberg (1983). FMPH predictions based on the excitation patterns were very similar to those based on the amplitude spectrum or loudness pattern.

FMPH2 hypothesis

The model of Carlson, Fant and Granstrom (1975) was applied to the reference and matching stimuli to obtain predictions of the matching data.. Carlson et al. did not describe an explicit procedure for determining the relative prominence of F1 harmonics, but the following procedure will be adopted here:

(1) The most prominent harmonic i_1 was identified by searching the F1 region of the spectrum for the largest component, selecting the higher frequency component in case of tied values:

3.5 $$i_1 = i, \quad A_{i-1} \leq A_i > A_{i+1}$$

(2) The second most prominent harmonic i_2 was identified as the larger of the two adjacent harmonics (selecting the higher frequency component in case of a tie):

3.6 $$\begin{aligned} i_2 = i_1-1, &\quad A_{i1-1} > A_{i1+1} \\ i_1+1, &\quad A_{i1+1} \geq A_{i1-1} \end{aligned}$$

An alternative formulation, which effectively combines (1) and (2), involves the calculation of the sum of all pairs of harmonics in the F1 region; harmonics i_1 and i_2 are selected as the pair with largest sum, $A_{i1} + A_{i2}$. It should be pointed out that both of these procedures involve a harmonic adjacency constraint: only contiguous components are included. This avoids the possibility of selecting one component from the F1 peak region and the other from the low frequency region (associated with the peak in the glottal source spectrum).

Predictions of the upper edge matching data based on the two most prominent harmonics of the amplitude spectrum are shown in Figure 3.14. The predicted F1 values are constant at all attenuation levels except the 0 dB condition where higher values are predicted for all three conditions. The pattern is similar to that of Figure 3.11, predictions based on FMPH, but lead to a higher RMS value of 66.7 Hz. The upper edge harmonic is not included in the weighting formula, except in the 0 dB condition, because it does not qualify as one of the two most prominent under the criteria stated above. Predicted F1 values based on the weighted mean of two harmonics were lower than those based on a single harmonic: the second most prominent harmonic lowers the weighted mean.

Predictions of the lower edge data based on two harmonics led to a smaller RMS value (39.5 Hz) than predictions based on FMPH (61.5 Hz). According to the former model, the predicted F1 values are constant for all attenuation levels in the 3 and 4 harmonic conditions. For the 2 harmonic stimuli, attenuation of the lower edge harmonic results in an increase in predicted F1. With more

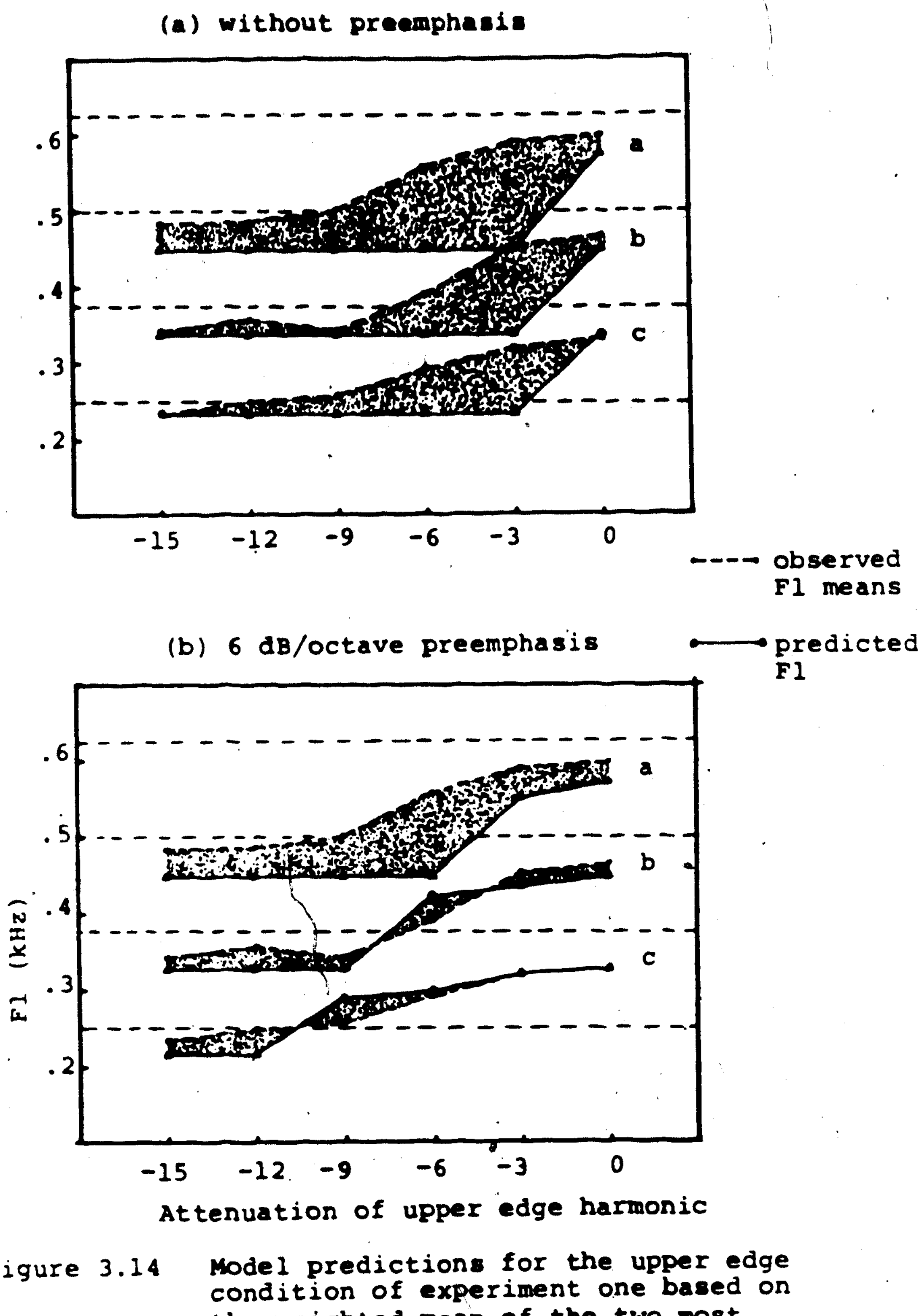
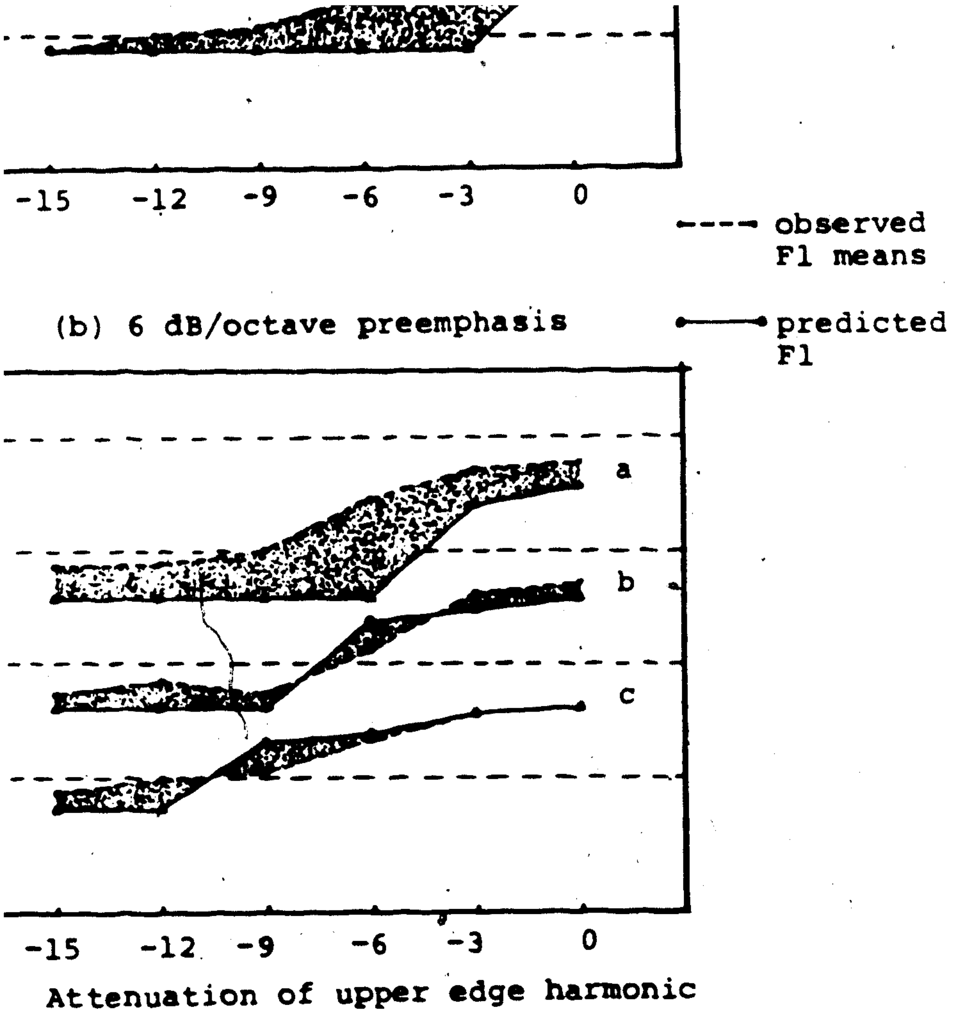


Figure 3.14 Model predictions for the upper edge condition of experiment one based on the weighted mean of the two most prominent harmonics. (a) 5 harmonics (b) 4 harmonics (c) 3 harmonics condition Dashed lines indicate harmonic frequencies



(b) 6 dB/octave preemphasis

14 Model predictions for the upper edge condition of experiment one based on the weighted mean of the two most prominent harmonics. (a) 5 harmonics (b) 4 harmonics (c) 3 harmonics condition Dashed lines indicate harmonic frequencies

pectral estimation procedure as employed in linear are not substantially different in the region above 600 Hz. Preemphasis was therefore applied to all components in the F1 region.

Predictions for the upper edge condition based on the weighted mean of the two most prominent harmonics of the preemphasized magnitude spectrum are shown in Figure 3.14. There is a clear improvement in predictions, with an RMS error of 46.1 Hz, compared with predictions based on unpreemphasized spectra which led to an RMS value of 66.7 Hz. The predictions are closest to the observed means in the two lowest curves; there appears to be a systematic underestimation of the observed F1 matches at higher frequencies. In the lower edge condition, 6 dB/octave preemphasis resulted in predictions which were slightly closer to the observed data than those based on the unpreemphasized spectrum (RMS error=37.9 Hz).

T-tests were carried out to compare the observed medians with predictions based on the FMPH2 hypothesis (preemphasized spectra). Significant differences (p<.05) were found for 10 of the 18 upper edge reference stimuli (2 of the 3 harmonic stimuli, 2 of the 4 harmonic stimuli, and all 6 of the 5 harmonic stimuli), indicating a lack of fit, particularly in the 5 harmonics condition. For the lower edge condition, only one of the 18 predicted values differed significantly from the observed data, suggesting that a good fit is provided by the FMPH2 model.

Excitation and loudness pattern model predictions

The relative auditory prominence of components in the F1 region may be affected by masking, even if each harmonic is separated by a frequency interval greater than the critical bandwidth. Masking studies have demonstrated that the auditory filters do not have infinitely steep attenuation characteristics; instead, they have gradually sloping skirts (for a review, see Patterson and Green, 1976, and section 2.3.4). Masking patterns are typically asymmetrical, with steeper low frequency slopes. Thus the masking effect of a tone which is lower in frequency is greater than one which is higher in frequency, an effect known as the "upward spread of masking". The present data, however, indicate a larger contribution of components which are higher in frequency.

The effects of masking are reflected in the excitation and loudness patterns described above. Harmonic components appear as peaks or local maxima in these patterns; the degree of contrast between the peaks and valleys in the pattern provides an indication of the spectral resolution of the harmonic component. A peak-finding algorithm was applied to each spectral pattern to estimate the frequencies and amplitudes of the harmonics. The most prominent harmonic, i_1, was defined as the largest peak in the F1 region and the second most prominent, i_2 as the largest additional peak in the frequency region f_{i1}±200 Hz. The weighted mean of the two most prominent components was computed according to the model of Carlson, Fant and Granstrom (1975). If no peaks occurred in a 200 Hz region around the largest component, the adjacent harmonics were assumed to have a negligible contribution and the frequency of the largest peak was used.

Predictions of the matching data based on several excitation and loudness pattern models are summarized in Table 3.3, and compared with predictions based on the amplitude spectrum, as described in the previous section.

Model one applied the procedure described by Sekey and Hanson (1982) to obtain excitation patterns specified in linear power versus Bark units. Model two was based on the procedure of Moore and Glasberg (1983), resulting in excitation patterns specified in units of linear power versus frequency in Hz. An equal loudness preemphasis function (Hermansky, Hanson and Wakita, 1985) and .3 power transformation (Stevens, 1966) was applied to the excitation patterns (models one and two) to obtain loudness patterns (models three and four, respectively). For comparison purposes, models five and six show predictions based on unpreemphasized and preemphasized amplitude spectra, respectively (see previous section).

	upper edge condition		lower edge condition	
mo	FMPH	FMPH2	FMPH	FMPH2
1	75.6	66.6	61.5	39.9
2	70.1	66.1	51.8	40.0
3	51.9	45.8	51.8	37.7
4	54.1	57.8	51.8	39.5
5	50.9	66.7	61.5	39.5
6	50.9	46.1	61.5	37.9

Table 3.3. RMS error values for several models.
Model 1: Excitation pattern (I)
Model 2: Excitation pattern (II)
Model 3: Loudness pattern (I)
Model 4: Loudness pattern (II)
Model 5: Amplitude spectrum (no preemphasis)
Model 6: Amplitude spectrum (6 dB/oct preemph.)

FMPH and F2MPH: predictions based on the frequency of the most prominent and weighted mean of the two most prominent harmonics, respectively, as estimated from the excitation or loudness pattern (see text).

Predictions based on the auditory models did not result in substantially better predictions of the matching data. Loudness pattern predictions (model three) using the weighted mean of the two most prominent components gave slightly better results than those based on excitation patterns or amplitude spectra. However, none of the auditory models provided substantially better results than corresponding analyses based on 6 dB/octave preemphasized spectra. Predictions based on a weighting of the two most prominent harmonics generally gave better results than predictions based on the frequency of the single most prominent harmonic.

FMPH3 hypothesis

Carlson, Fant and Granstrom (1975) considered the possibility that boundary shifts along the /i-e/ continuum with increasing fundamental frequency might be accounted for by an equal weighting of the three most prominent harmonics in the F1 region. They found that the weighted mean of the three most prominent harmonics did not coincide closely with the nominal F1 values, and they noted irregularities which were not reflected in the identification data.

The weighted mean of the three most prominent harmonics was computed for each reference and matching stimulus, using the criteria described above to obtain the 2 most prominent harmonics, and defining the third most prominent as the larger of the two harmonics adjacent to the pair. The higher frequency component was selected in case of a tie.

Predictions based on three harmonics led to higher RMS error rates than predictions based on two (RMS=52.8 Hz for the lower edge condition, 79.5 Hz for the upper edge condition). Six dB/octave preemphasis of the spectrum improved the predictions somewhat (RMS error = 48.9 Hz for the lower edge condition, 54.1 Hz for the upper edge condition).

The reason for these lower RMS values is illustrated in Figure 3.15, a plot of the predicted and observed F1 values for the upper edge data using the weighted mean of the three most prominent harmonics in the F1 region of the amplitude spectrum. Predicted F1s are generally lower than observed means. The discrepancies are largest in the 4 and 5 harmonic conditions.

Without preemphasis (panel a) the 4 and 5 harmonic conditions are predicted to have a single discrete change as the upper edge harmonic is attenuated. The upper edge harmonic is not included as one of the three most prominent, except in the 0 dB condition. Predicted F1s are substantially lower in frequency than observed means. With preemphasis (panel b) the predicted F1s follow the general pattern of gradual changes in the matching data, but all of the predictions are too low in frequency. It appears that this model assigns too much weight to the lower frequency harmonics of the reference stimuli. in the 2 harmonics condition.

T-tests were carried out, comparing observed medians and predictions based on the FMPH3 hypothesis. Significant differences (p<.05) were found for 9 of the 18 upper edge reference stimuli and 6 of the 18 lower edge reference

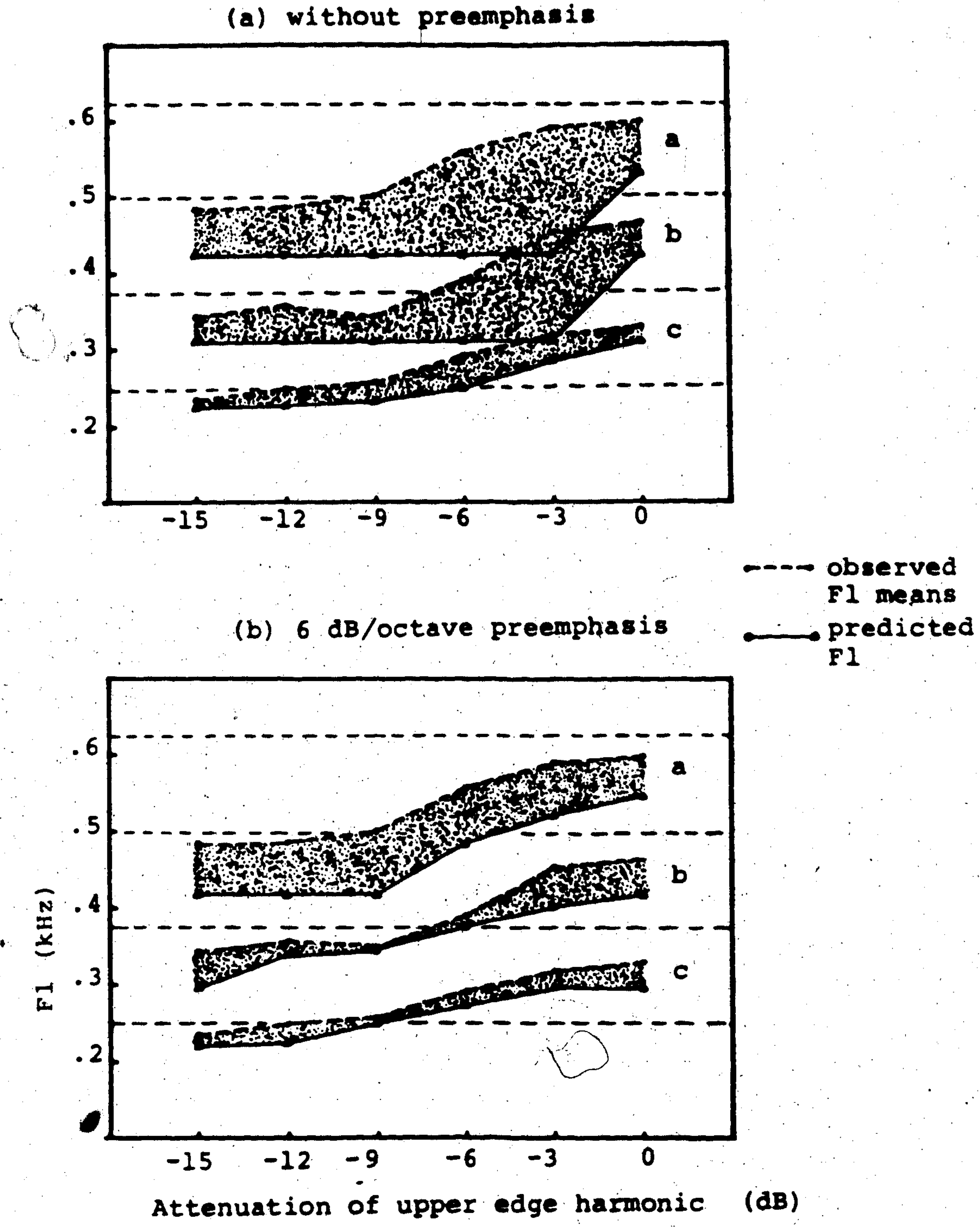


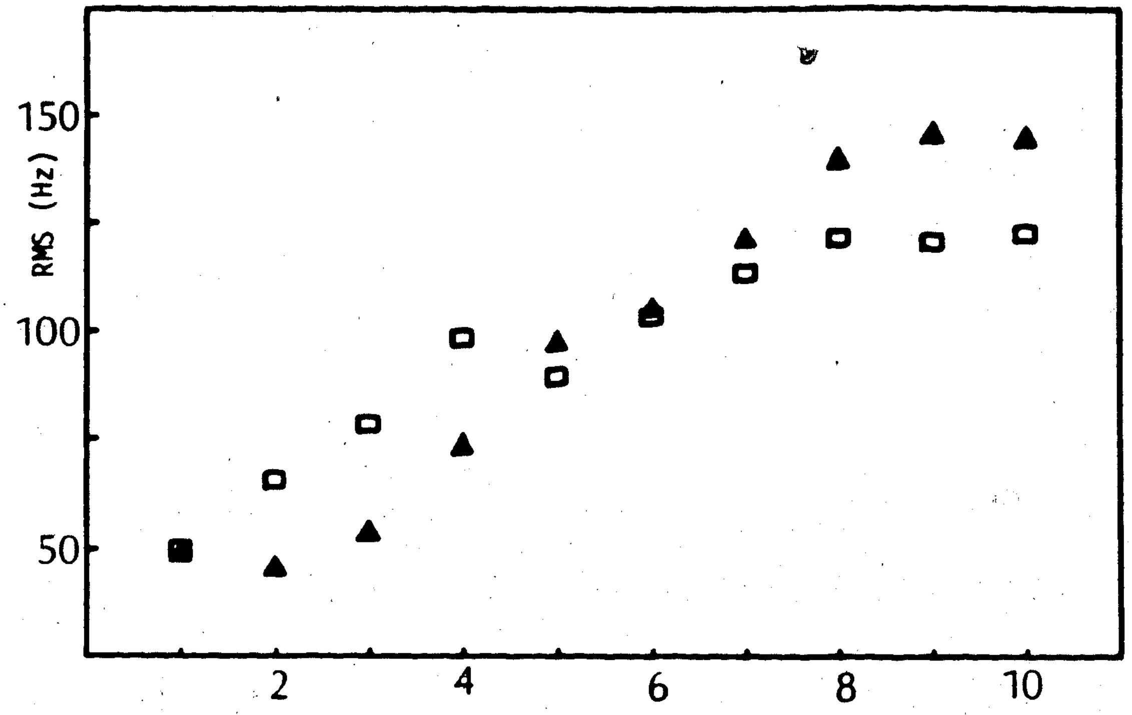
Figure 3.15 Model predictions for the upper edge condition of experiment one based on the weighted mean of the three most prominent harmonics. (a) 5 harmonics (b) 4 harmonics (c) 3 harmonics condition Dashed lines indicate harmonic frequencies

stimuli, reflecting the discrepancies noted above.

Weighted mean in F1 region

If it is assumed that all components in the region of the first formant contribute equally to the perceived vowel height, then the centroid computed over the full F1 range may provide accurate predictions of the matching data. The advantage of this procedure is that there is no need to determine the exact location of harmonics, and that it can be applied to vowels with a noise source or mixed excitation. The main problem with such a procedure is defining the ranges of F1 and F2; in natural speech measurements there is considerable overlap (Peterson and Barney, 1952).

The F1 range for the present study was defined to be 0-1.5 kHz, which includes virtually the entire range of F1 values observed in human speech. This limit excludes F2 in front vowels but not in back and central vowels. Predictions based on the centroid of all components in the F1 range led to higher RMS values than analyses based on a small number of harmonics near the formant peak, for both upper and lower edge conditions. RMS error rates for predictions of the upper edge condition are shown as a function of the number of harmonics used in the weighting function in Figure 3.16. Squares represent predictions based on unpreemphasized spectra, while triangles represent



Number of harmonics used in weighting function

Figure 3.16 RMS distance between observed and predicted F1 match values for the upper edge condition of experiment one as a function of the number of harmonics used to calculate the weighted mean.
Squares represent calculations based on unpreemphasized spectra; triangles, calculations using 6 dB/octave preemphasis

predictions using 6 dB/octave preemphasis of the amplitude spectrum. As more harmonics are added (in order of prominence, using the procedure described for two and three harmonic calculations) the predictions become consistently worse. The lower harmonics of the reference stimuli appear to carry too much weight, and the predicted F1s are consistently lower than the observed. The best predictions (lowest RMS values) were obtained for the weighted mean of the two most prominent harmonics of the preemphasized spectrum.

It is possible that a weighting function which incorporates more than two harmonics, but assigns smaller weights to those further from the peak, may result in improved predictions of the matching data. Further empirical data are needed to determine to what extent these additional components are incorporated in the perception of vowel height.

Choice of amplitude scale

An additional variable which affects calculations of the centre of gravity is the choice of scale in representing spectrum levels. There are a number of possibilities: e.g. amplitude, intensity, or approximate loudness scales. Several alternatives were explored in predictions of the upper edge matching data. Predictions based on the weighted mean of the two most prominent harmonics were little affected by changes in scale (see Table 3.4).

Harmonic selection criteria

A more important variable than the choice of amplitude scale appeared to be the nature of the criteria used in selecting the two most prominent harmonics. The most important factor appeared to be the inclusion of the upper edge harmonic and adjacent lower harmonic of the reference stimulus. With 6 dB/octave preemphasis of the spectrum, these two harmonics were selected as the most prominent for about two-thirds of the upper edge reference stimuli. For the remainder of the upper edge reference stimuli, two contiguous lower frequency harmonics were selected, with the result that the predicted F1s were often much lower than the observed F1 match values. This discrepancy was greatest in the 5 harmonics condition, a consequence of the fact that higher frequency components are affected to a lesser extent by 6 dB/octave preemphasis of the spectrum.

If two harmonics are employed in the weighting function, it appears to be necessary to include the upper edge harmonic over a wide range of attenuation levels to account for the data in both upper edge and lower edge conditions. Although the weighted mean of the two most prominent harmonics in the preemphasized (on a dB/octave basis) spectrum provides a reasonable first approximation to the matching data, it incorporates a frequency dependence in

			RMS error (Hz)		
k	p	q	lower edge	upper edge	nominal F1
1	0	1	61.5	50.9	40.3
2	0	1	39.5	66.7	40.0
3	0	1	52.8	79.5	50.7
1	1	1	51.8	49.1	35.3
2	1	1	37.9	46.1	14.1
3	1	1	48.9	54.1	20.0
1	0	2	61.5	50.9	40.3
2	0	2	40.0	65.6	31.6
3	0	2	50.7	84.2	37.2
1	1	2	51.8	49.1	35.4
2	1	2	38.5	45.2	7.5
3	1	2	43.3	49.9	15.1

k= number of harmonics in weighted mean
p= preemphasis (0= none, 1= 6 dB/octave)
q= amplitude exponent (1= magnitude, 2=power)

Table 3.4. RMS error for predictions of the matching data in the upper edge and lower edge conditions of Experiment 1, using the weighted mean of the k most prominent harmonics, and estimates of the nominal F1 values of the matching stimuli

the resulting predictions which is not present in the observed data.

LPC analysis and F1 estimation

Most of the measurement procedures considered up to this point have involved essentially local operations on the spectrum, taking into account only energy in the immediate vicinity of the first formant peak. An alternative approach is to take into account the entire spectral shape using a global spectral estimation procedure as employed in linear predictive coding (LPC) analysis (Makhoul, 1975) which is often used in formant measurement (e.g., Markel, 1972; McCandless, 1974; Christensen, Strong and Palmer, 1976).

Each of the matching and reference stimuli of Experiment 1 was subjected to an autocorrelation LPC analysis using 12 predictor coefficients, following 6 dB/octave preemphasis of the spectrum (Markel and Gray, 1976).

The LPC log magnitude spectra for the 18 upper edge stimuli are shown in Figure 3.17, and for the 18 lower edge stimuli in Figure 3.18. Several features of these spectra should be noted. First, each one contains a single maximum, in the vicinity of the upper edge harmonic. Second, attenuation of the upper edge harmonic (Figure 3.17) results in a decrease in peak frequency. In Figure 3.18, attenuation of the lower edge harmonic leads to an increase in the peak location, but only in the 2 harmonics condition.

The frequency of the largest peak in the F1 region was determined for the reference and matching stimuli, and the resulting measurements were used to predict the matching data. The predicted F1s and observed means are shown for the upper edge condition in Figure 3.19. The predictions are very close in both conditions. The RMS error was 39.9 Hz in the upper edge condition, and 36.9 Hz in the lower edge condition. A global spectral fitting procedure thus

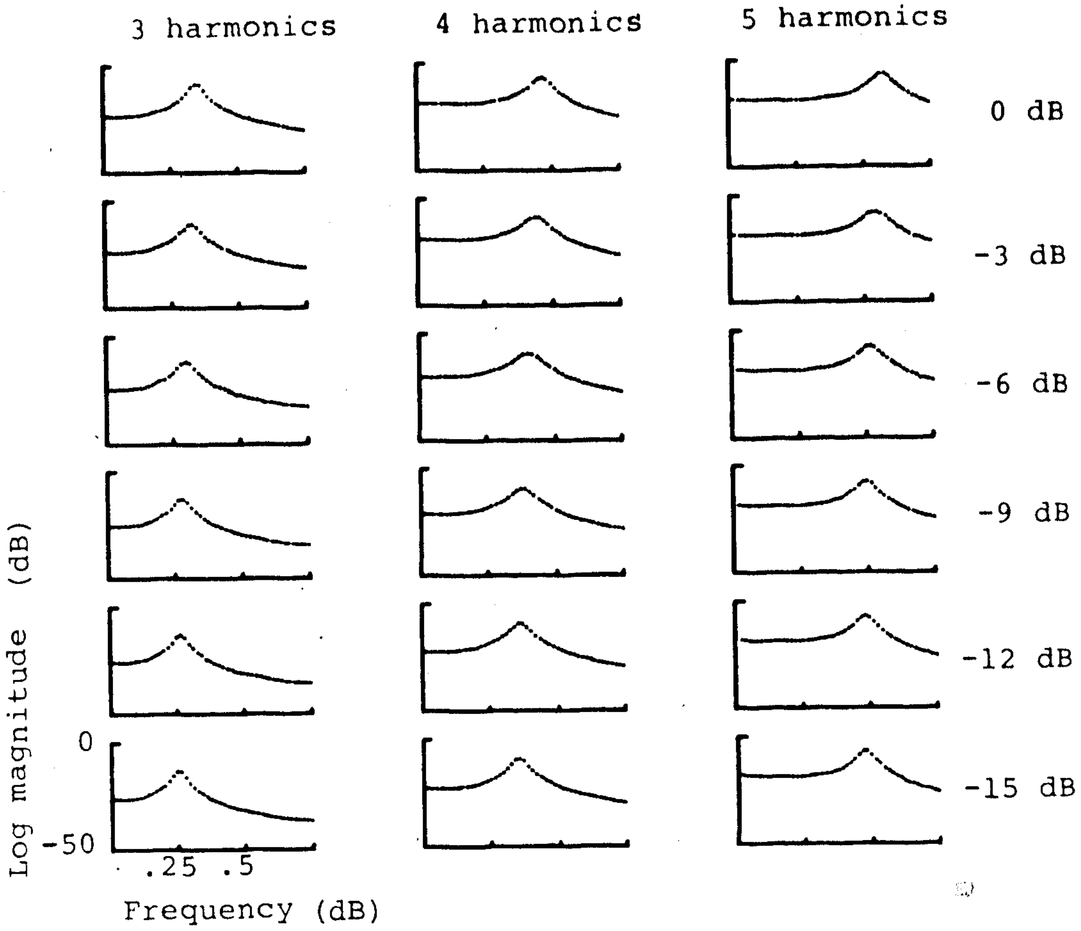


Figure 3.17 LPC log magnitude spectra for the upper edge stimuli of experiment one

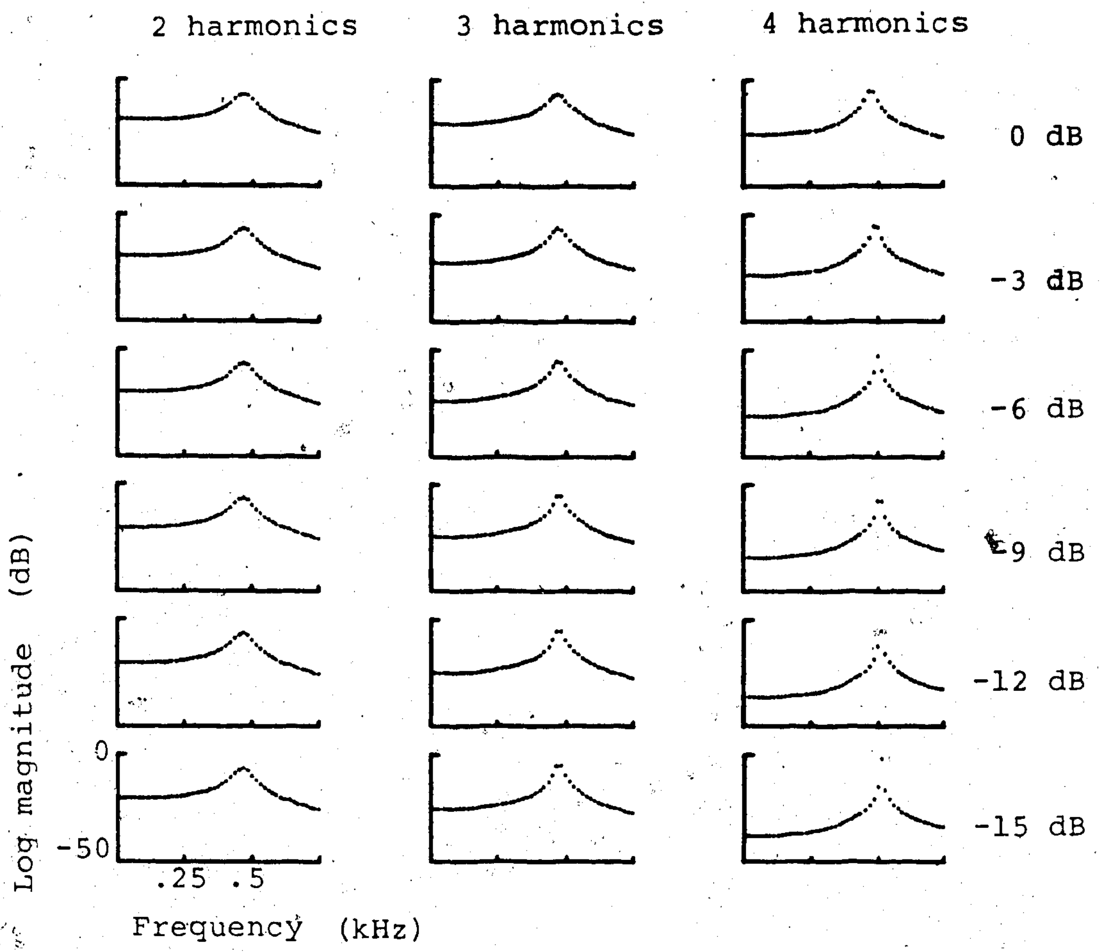


Figure 3.18 LPC log magnitude spectra for the lower edge stimuli of experiment one

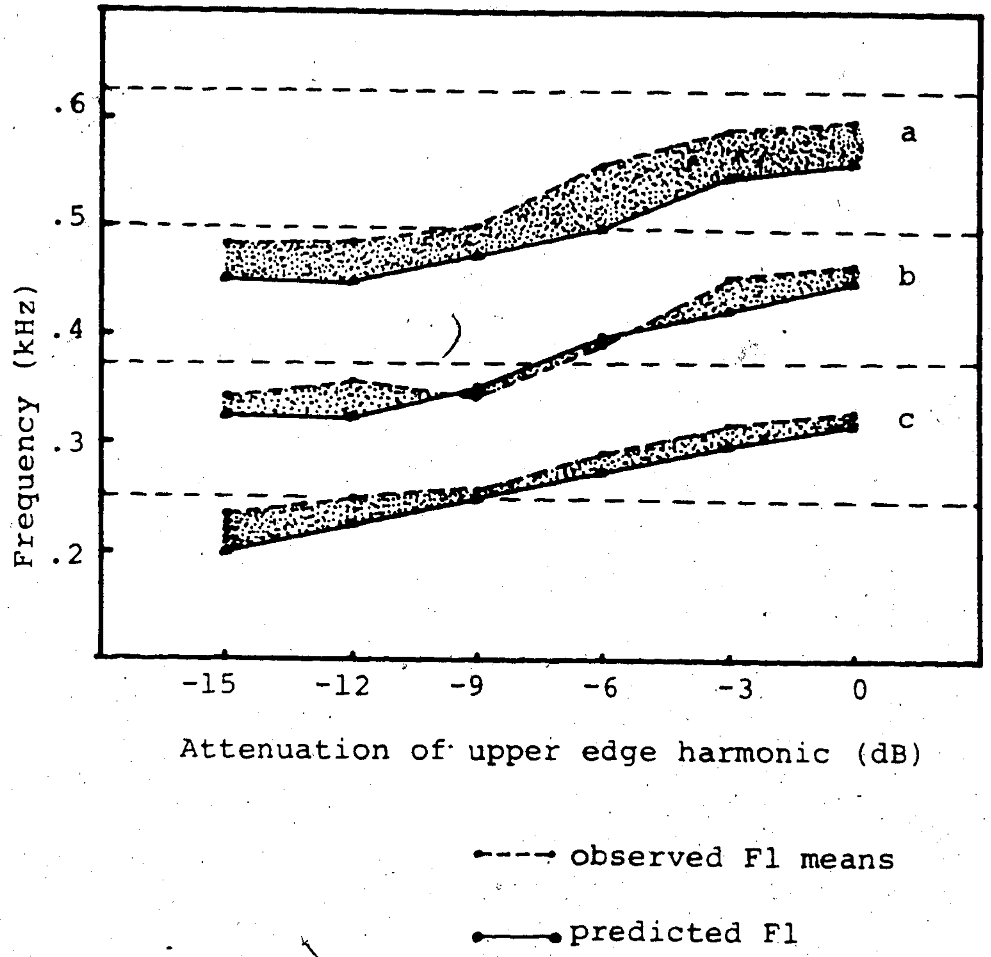


Figure 3.19 Model predictions for the upper edge condition of experiment one based on the frequency of the lowest peak in the LPC spectrum.
(a) 5 harmonics condition
(b) 4 harmonics condition
(c) 3 harmonics condition
Dashed lines indicate harmonic frequencies

provides estimates of the F1 peak location which yield very close predictions of the matching data. in the 2 harmonics condition.

T-tests were carried out to compare the observed medians with predictions based on LPC estimates of F1. Significant differences (p<.05) were found for 11 of the 18 upper edge reference stimuli and 1 of the 18 lower edge reference stimuli. The results are thus fairly similar to predictions based on the FMPH2 hypothesis.

Summary of model predictions

In this section a number of different measures for predicting the matching data are explored. In evaluating how closely each measure could account for perceived vowel height, three general characteristics of the perceptual data were considered: (1) the alignment of F1 in the matching stimuli near the upper edge harmonic in the reference stimuli; (2) the gradual decrease in matched F1 with attenuation of the upper edge harmonic; and (3) the increase in matched F1 with attenuation of the lower edge harmonic. Predictions based on the frequency of the most prominent harmonic could not account for (2) and (3). Local centre of gravity measures (weighted mean of the most prominent harmonics) provided more satisfactory results. Fairly close predictions were provided by measures based on FMPH2. The weighted mean of two or more harmonics led to predicted F1s which were *lower* than observed F1 matches.

Excitation and loudness pattern models did not appear to provide insights which were not already evident from analyses based on amplitude or power spectra. These models are based on *place* representations of the auditory response. It is possible that a temporal coding scheme may give better results (see section 2.3.9)[1].

Accurate predictions of the matching data were also obtained with LPC based estimates of the F1 peak. The location of the low frequency peak in the LPC spectrum was close to the upper edge harmonic in the reference stimuli, and the three characteristics of the matching data noted above were well modelled by LPC analysis.

Comparison of F1 measures

Several alternative hypotheses concerning the perception of vowel height have been considered, and measures based on these hypotheses were used to predict the vowel matching data. If vowel perception depends on the estimation of formant locations, then it might be expected that the best predictors of the matching data would also provide the best measures of formant locations.

[1] Preliminary attempts were made to model the data using the DOMIN model of Carlson, Fant and Granstrom (1975). The resulting patterns did not produce the type of spectral interpolation which appears to be indicated in the matching results; peaks in the DOMIN histograms nearly always corresponded to the frequencies of individual harmonics.

Comparisons were made between the estimates of the first formant peak provided by several of the measures considered above and the nominal (or synthesized) first formant frequency. The RMS difference between estimated and nominal F1 values is indicated for each of the measures in Table 3.4. The first series of measures compute the weighted mean of the k most prominent harmonics in the F1 region. In addition to the number of harmonics, the effects of preemphasis (none versus 6 dB/octave) and scale (amplitude or intensity) were investigated.

In Figure 3.20 the estimated peak frequencies are shown as a function of nominal F1, for the matching stimuli used in Experiment 1. Measures involving a single harmonic show discrete steps at integral multiples of the fundamental frequency. Analyses based on 2 harmonics gave consistently better estimates of the nominal F1 than those based on either 1 or 3 harmonics, consistent with the results of Carlson, Fant and Granstrom (1975) and with the predictions of the matching data. When 2 or 3 harmonics are used, there is still a step-like effect resulting from changes in the frequency of the most prominent harmonic. This influence is reduced when 6 dB/octave preemphasis is applied prior to calculating the weighted mean. The use of preemphasis (which resulted in better predictions of the matching data) also gave better estimates of the nominal F1. Most of the discrepancies between nominal and measured F1 values occurred in the region below the nominal F1 line in Figure

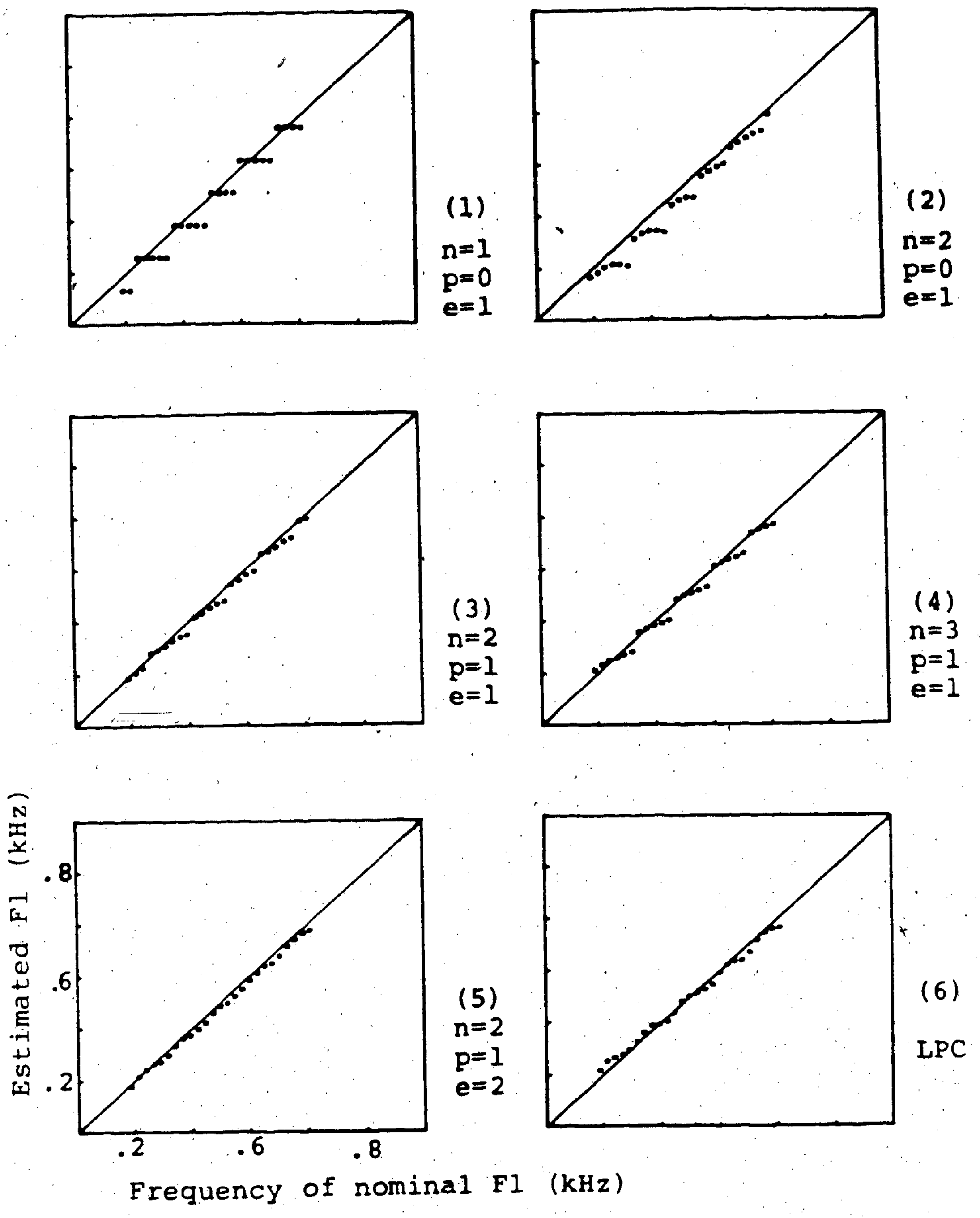


Figure 3.20 Model-based estimates of the nominal F1. Data in(1) to (5) are based on the weighted mean of the n most prominent harmonics;
e=1 for magnitude, 2 for power;
p=0 for no preemphasis, 1 for 6 dB/octave

3.20, indicating that the nominal F1 was generally being underestimated.

The choice of ordinate scale did not have a large effect on the estimation of nominal F1. However, it is interesting to note that the intensity or power scale, which provided slightly better estimates of nominal F1, also gave the best results in predictions of the matching data.

The results of these analyses suggest a close relationship between predictions of the perceptual data and estimates of the nominal F1 of the matching stimuli. Measures which provided the closest estimates of the first formant also predicted fairly accurately which pairs of reference and matching stimuli would be judged by listeners to be closest in vowel height.

3.3.5 General discussion

Experiment 1 has provided evidence against the hypothesis of Mushnikov and Chistovich (1971) that vowel height is determined by the frequency of the most prominent harmonic in the F1 range. This hypothesis predicted at most one change in listeners' matches as the amplitude of the most prominent harmonic was reduced. Instead, a gradual shift was observed as the upper edge harmonic was attenuated in small steps.

It was found that the F1 region of a synthetic front vowel could be replaced by a small number of harmonics of equal amplitude. The resulting stimulus differed slightly in timbre from the original, but could be matched in phonetic quality with a full spectrum vowel drawn from an F1 continuum. Listeners aligned the F1 peak of the matching stimulus near the highest component of the F1 region of the reference. Attenuation of the highest frequency (upper edge) component resulted in a gradual shift toward lower matched F1s. Attenuation of the lowest frequency (lower edge) harmonic did not produce consistent shifts except in the 2 harmonics condition, where listeners selected higher F1 matches. It appears that at least two components are important in vowel height matching. In this experiment it was found that the upper edge and the adjacent lower harmonic of the reference stimulus were generally aligned with the two most prominent harmonics in the F1 region of the matching stimulus.

There are several possible explanations for the apparent similarity between an "upper edge" and a formant peak. One possibility is that the higher frequency harmonics of a band of equal amplitude harmonics are perceived as <u>louder</u>. However, the frequency dependent changes in equal loudness contours were not closely paralleled in the overall pattern of matching results.

A second possibility is that the lower frequency components of the reference stimulus were masked by those higher in frequency. Since masking generally affects higher frequency components to a greater extent than lower (Egan and Hake, 1950), it is unlikely that the the apparent perceptual salience of the upper edge component is due to a masking effect of this sort.

Lateral suppression may have reduced the relative prominence of the lower harmonics. Lateral suppression effects have been reported in studies employing vowel stimuli as maskers (e.g., Tyler & Lindblom, 1982; Moore & Glasberg, 1983). While lateral suppression could operate to enhance the "edges" of a spectral pattern (c.f. the phenomenon of "edge pitch" investigated by Small & Daniloff, 1967, and Rakowski, 1968), it is not clear why F1 matches are very rarely found to coincide with the lower edge of the band of harmonics of the reference stimuli. However, there is some evidence (Houtgast, 1974a,b; 1977) that lateral suppression generally affects spectral components below the masker to a greater extent than components above. It is possible that the upper edge harmonic may be perceptually more prominent than harmonics lower in frequency as a result of suppression effects.

Another possible explanation is based on listeners' implicit "knowledge" of constraints on the shape of the spectrum, an explanation considered by Darwin (1984) to account for the results of a vowel identification experiment which closely parallel some of the findings of the present experiment. Listeners identified vowel stimuli on an F1 continuum in several different conditions in which the intensities of harmonics near the formant peak were boosted by small amounts. Boosting the harmonic immediately above the first formant led to a shift in the /I-ɛ/ boundary toward lower frequencies; boosting the harmonic just below F1 caused a shift toward higher frequencies.

Darwin noted that level changes in the harmonic below F1 had a smaller effect than changes in the harmonic above F1. He proposed that energy associated with the lower frequency component may not have been fully incorporated in the vowel percept. One possible factor considered by Darwin was the shape of the vocal tract transfer function in the low frequency region of the vowel spectrum. The low frequency position of the first formant restricts the range of intensities which may be assumed by harmonics below F1 to a much greater extent than those above F1. The boosted component below the first formant peak may have been partially discounted by the perceptual system, or perhaps associated with characteristics of the glottal source. As pointed out by Darwin, this effect may not be restricted to vowels, since other sound production mechanisms may be constrained in a similar manner.

Darwin's results indicate larger effects on vowel identification by an increase in the intensities of harmonics above the F1 peak than below. The results of the present experiment show that similar asymmetries may be present with harmonic attenuation as well. However, the extent of the asymmetry may have been somewhat larger in Darwin's study. He found that predictions based on LPC estimates of first formant frequency or on a weighting of prominent harmonics could not account for the extent of the asymmetries. Both of these schemes provided fairly successful results in predicting the present matching data. This discrepancy may be due to differences in stimulus design (e.g., attenuation rather than boosting of harmonic levels) or to differences in the response paradigm (vowel matching versus identification). Further empirical study is needed to determine the perceptual basis for these differences.

One possibility is that the asymmetrical effects of changes in harmonic levels stem from a partitioning of the signal into components which form part of the vowel and others which do not (Darwin, 1984). The lower frequency components may have been partially excluded in the perceptual estimation of vowel height. The perceptual weighting of F1 harmonics may be intrinsically asymmetrical. We have seen evidence that an asymmetrical weighting of harmonics can yield accurate estimates of the first formant frequency; perhaps a similar strategy is employed by the auditory system.

3.4 Experiment Two: Effects of amplitude, frequency and number of harmonics on vowel matching

3.4.1 Introduction

The results of Experiment 1 are consistent with the hypothesis that vowel height is determined by the two most prominent harmonics in the F1 region. The reference stimuli of Experiment 1, which were comprised of a band of equal amplitude harmonics in the F1 region, were generally matched with stimuli whose F1 value was intermediate between the frequency of the highest and second highest harmonics in the band. Attenuation of the higher member of this pair led to lower F1 matches, while attenuation of the lower resulted in higher matched F1. Additional harmonics in the band had no significant effects on matching.

The present experiment was designed to establish whether two harmonics are the sole determinants of vowel height or whether other components contribute to the perceived quality. The effects of attenuating either the highest or lowest frequency members of a pair or triplet of harmonics were assessed using the matching task of Experiment 1. Comparisons were made at four different frequencies to determine whether the asymmetries observed in Experiment 1 are dependent on the frequency location (or harmonic rank) of the individual components.

Such a dependency might be postulated, for example, if the observed asymmetries were the result of loudness differences in the individual harmonics. For mid-range intensities, equal loudness curves (Robinson and Dadson, 1956) indicate a gain of about 6 to 10 dB/octave in the region below 600 Hz; between 600 and 1200 Hz the curves are nearly level. If differences in loudness of the individual harmonics underlie the asymmetries in matching performance, it is expected that the degree of asymmetry will decrease with frequency, and will disappear when all of the F1 harmonics of the reference stimuli are above 600 Hz.

3.4.2 Method

The experimental method was similar to that described for Experiment 1.

Stimulus materials

The reference stimuli were constructed from the baseline stimulus as described for Experiment 1. Harmonics between 0 and 1.5 kHz were deleted, and replaced with a band of either 2 or 3 contiguous harmonics. The spectrum in the region above 1.5 kHz was identical to the baseline stimulus. A fundamental frequency of 125 Hz was used.

In the two harmonics condition a single pair of contiguous harmonics occurred in the F1 region. Both harmonics were set to the amplitude of the largest harmonic in the F1 region of the baseline stimulus. Four harmonic pairs were used: f2+f3, f3+f4, f4+f5, and f5+f6.

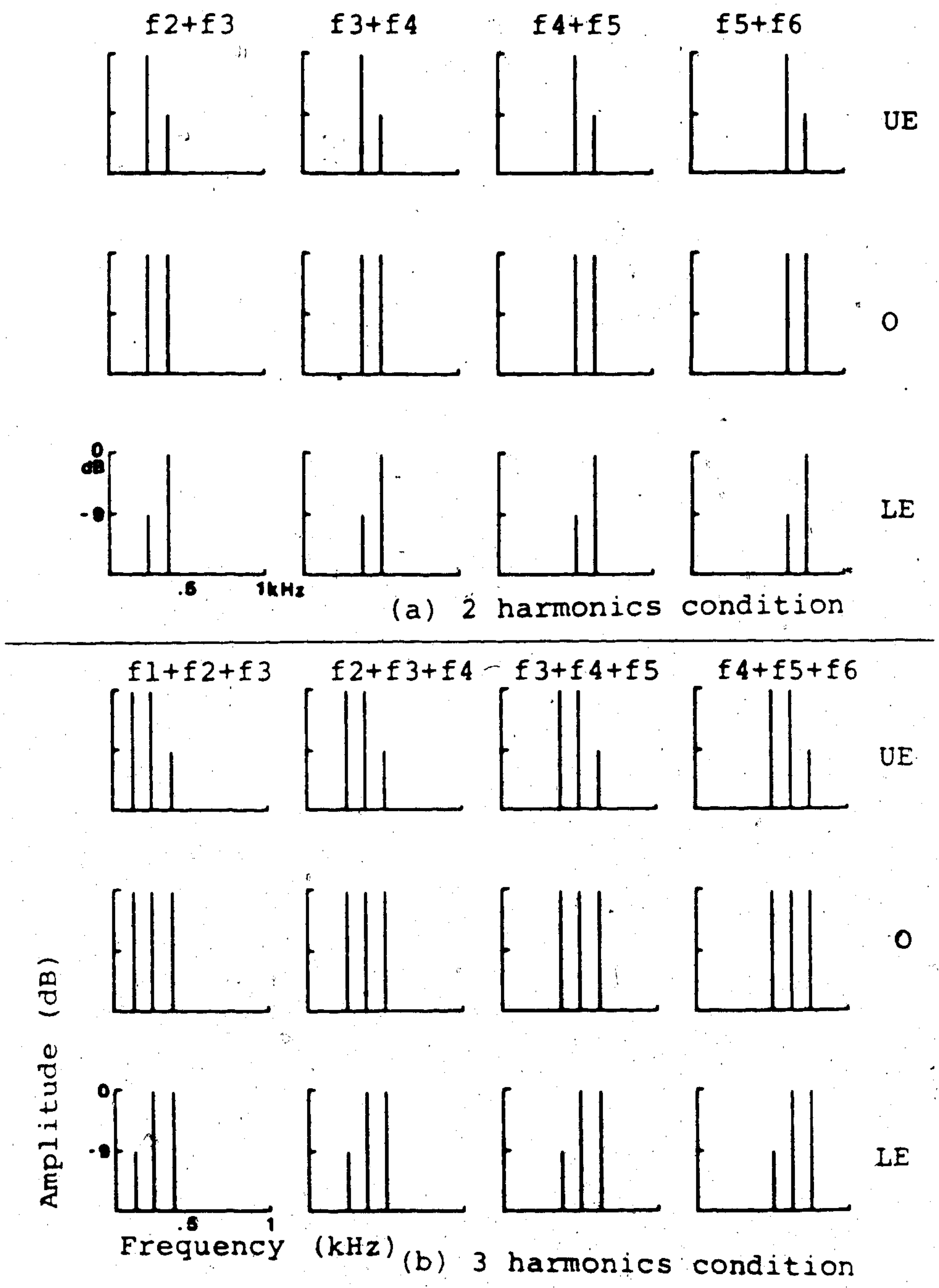
In the three harmonics condition, the F1 region was replaced by a triplet of equal amplitude harmonics. Four combinations were again used: f1+f2+f3, f2+f3+f4, f3+f4+f5, and f4+f5+f6. It should be noted that the highest frequency member of each band coincided with the highest member of the corresponding two harmonic stimulus.

Three amplitude conditions were included: a 0 dB or control condition, with all harmonics equal in amplitude to the peak harmonic in the F1 region of the baseline stimulus; a "lower edge" condition, with the lowest frequency harmonic attenuated by 9 dB, and an "upper edge" condition in which the highest frequency harmonic was attenuated by 9 dB.

Line spectra indicating the F1 region of the reference stimuli for the 2 and 3 harmonics conditions are shown in Figure 3.21. A display of the entire spectrum for two typical stimuli in the 3 harmonics condition is given in Figure 3.22.

Procedure

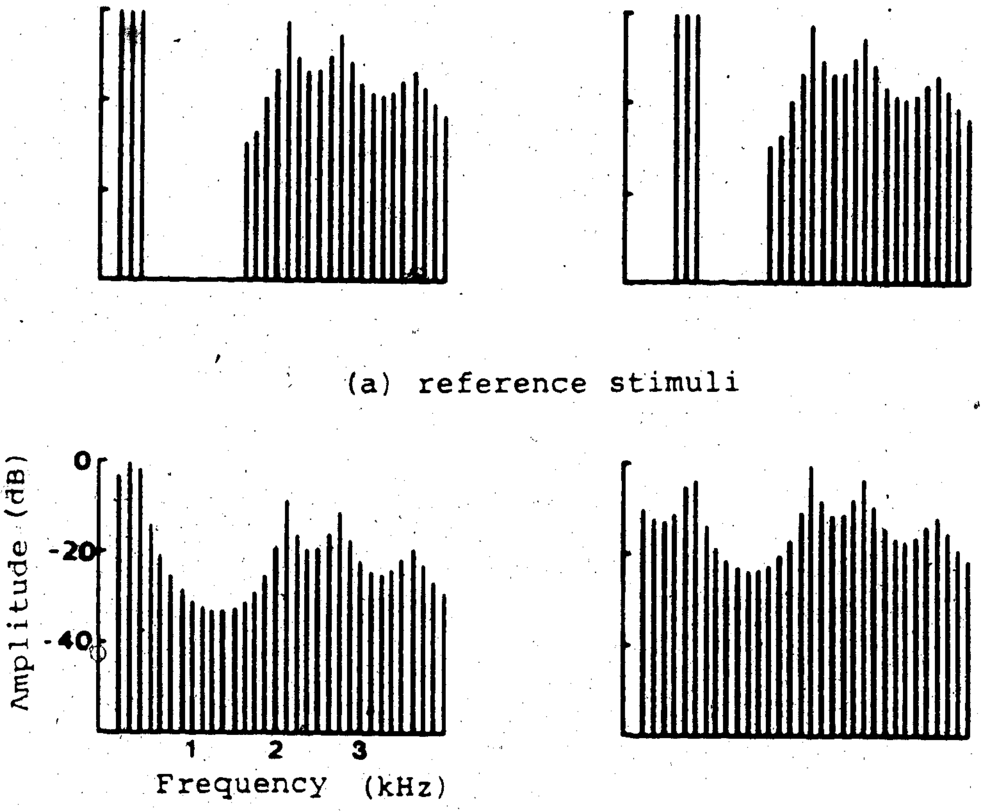
Synthesis of the vowel stimuli, stimulus presentation and recording of responses followed the procedure described in Experiment 1. The same six listeners determined the best matching vowel pairs as described in Experiment 1, with two



(a) 2 harmonics condition

(b) 3 harmonics condition

Figure 3.21 Line spectra of the reference stimuli for experiment two. UE: upper harmonic attenuated by 9 dB; LE: lower harmonic attenuated by 9 dB; O: neither harmonic attenuated



(a) reference stimuli

(b) matching stimuli

Figure 3.22 Line spectra of two typical stimuli in the 3 harmonics condition of experiment two, (a) and 'best' matching stimuli, (b).

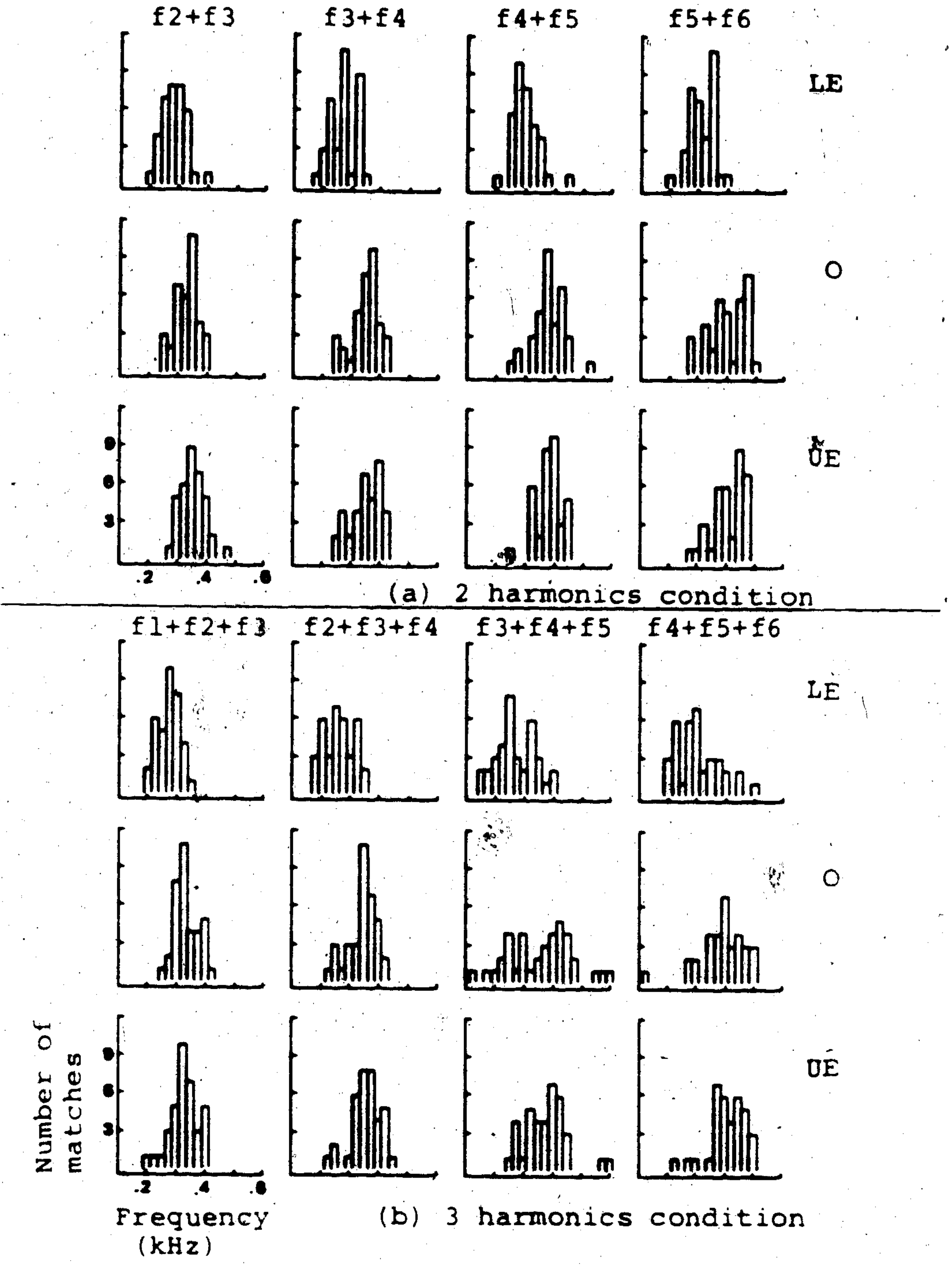
sessions, each comprised of 3 replications of each of the 24 reference stimuli in random order. Each session lasted approximately 15 to 20 minutes.

3.4.3 Results and discussion

Histograms of the matching data are shown in Figure 3.23. In all but one or two cases (discussed below) the matches were clustered around a single peak. The overall pattern of results is similar to that observed in Experiment 1. There is a shift toward lower F1 values in the upper edge condition, (relative to the control) but little change in the lower edge condition. This is consistent with the results of Experiment 1 indicating a considerable asymmetry in the effects of attenuating the upper and lower edge harmonics.

Means and standard deviations of the median F1 matches from each listener (based on 6 trials for each reference stimulus) are shown in Figure 3.24. Asymmetrical effects of amplitude can be seen in both 2 and 3 harmonics conditions. Attenuation of the upper edge had a large effect in both cases, with shifts of approximately the same magnitude. Attenuation of the lower edge harmonic had little effect on matching performance, with one exception to be discussed below.

The 9 dB reduction in amplitude of the upper edge harmonic led to a mean shift of about -80 Hz in matched F1



(a) 2 harmonics condition

(b) 3 harmonics condition

Figure 3.23 Matching histograms for 2- and 3- harmonic stimuli of experiment two. UE: upper harmonic attenuated by 9 dB; LE: lower harmonic attenuated by 9 dB; O: neither harmonic attenuated

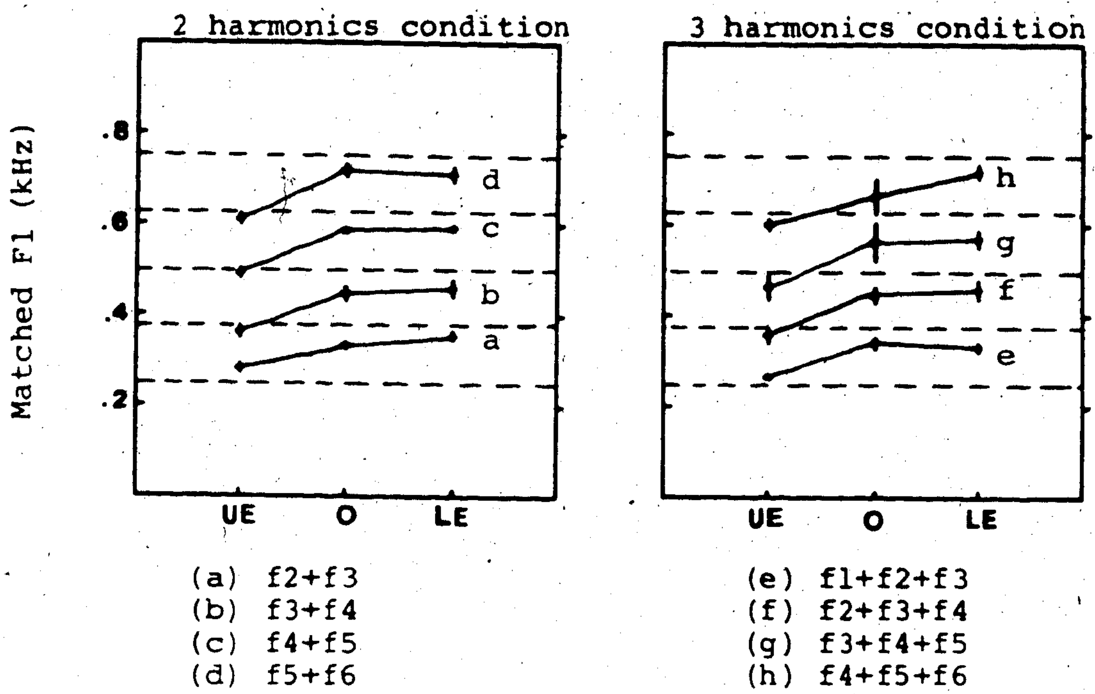


Figure 3.24 Means and standard deviations of matched F1 values for the two harmonics condition (left) and three harmonics condition (right) of experiment two.

Dashed lines indicate frequencies of harmonics.

UE: upper harmonic attenuated by 9 dB
O: neither harmonic attenuated
LE: lower harmonic attenuated by 9 dB

for both conditions, approximately 16 percent of the mean F1 value in the 0 dB condition. Attenuation of the lower edge harmonic by 9 dB led to small positive F1 shifts for some of the stimuli.

The frequency location of the pair or triplet of harmonics did not appear to affect the size of the shifts observed in the upper edge condition: the four curves of Figure 3.24 are nearly parallel. There was a small tendency for the slopes of the lines to increase with frequency.

An analysis of variance was carried out on the median matching data, with factors FREQUENCY (of the highest harmonic), NUMBER OF HARMONICS, and AMPLITUDE (upper edge, lower edge and control conditions). Main effects were found for FREQUENCY (F(3,15)=277.74, p<.001) and AMPLITUDE (F(2,10)=424.01, p<.001) but not for NUMBER OF HARMONICS. No significant interaction of any of the variables was present.

Dunnett's test for comparing treatment means with a control (Winer, 1971) was conducted to evaluate the effects of AMPLITUDE (see Table 3.6). Attenuation of the upper edge harmonic by 9 dB produced significant shifts toward lower matched F1s, relative to the 0 dB condition, in all cases. This result is consistent with the findings of Experiment 1 which indicate a special role for the upper edge harmonic in determining vowel height.

A. 2 harmonics:	upper edge	lower edge
1. f2+f3	-50.0**	20.8
2. f3+f4	-83.3**	12.5
3. f4+f5	-95.8**	6.3
4. f5+f6	-106.3**	-8.3
B. 3 harmonics:	**upper edge**	**lower edge**
1. f1+f2+f3	-72.9**	-8.3
2. f2+f3+f4	-87.5**	-12.5
3. f3+f4+f5	-100.0**	-6.3
4. f4+f5+f6	-62.5**	54.2**

*p<.01

Table 3.6. Mean deviations from the 0 dB condition and significance levels of Dunnett's tests for the matching results of Experiment 2

Attenuation of the lower edge harmonic by 9 dB produced small upward shifts in matched F1 for some of the stimuli, but only one of these shifts was statistically significant. The highest frequency 3 harmonics condition (f4+f5+f6) showed a large upward shift in the lower edge condition. Closer inspection of the data, however, revealed that this difference may have been due to changes in matched F1 in the 0 dB condition. This stimulus was matched with a mean F1 value lower than expected on the basis of the other 3 harmonic stimuli, and also lower than the corresponding stimulus with 2 harmonics. The matching histograms showed a larger matching range (see Figure 3.24). The 0 dB stimulus with harmonics f3+f4+f5 also showed an irregular response profile, but was not matched with a lower mean F1'.

Statistically significant shifts were found in Experiment 1 when the lowest frequency member of a pair of harmonics was attenuated, but not when the lowest of 3 or 4 harmonics was attenuated. Attenuation of the lower edge harmonic in the 2 harmonic condition of the present experiment did not produce statistically reliable shifts, although there was a trend toward higher matched F1 values for several stimuli. Both experiments showed that the lower edge harmonic had a much smaller effect on vowel matching than the upper edge harmonic. The effects of attenuation of lower or upper edge harmonics were largely independent of frequency, making it unlikely that the asymmetric effects of these changes are due to differences in the relative loudness of individual harmonic components.

A comparison of the data from the 2 and 3 harmonic conditions suggests that there may be a small effect of the added component: 20 of the 24 reference stimuli in the 2 harmonics condition of the present experiment were matched with higher mean F1s than the corresponding 3 harmonic stimuli, with a mean difference of 14 Hz, but these differences were not statistically significant. It is not possible to conclude with certainty on the basis of these results that vowel height is determined by more than the two

[1] Informal listening suggested that these stimuli may have been somewhat ambiguous in vowel quality, and were in fact occasionally perceived as back vowels.

most prominent harmonics. A recent study by Darwin and Gardner (forthcoming) presented evidence that additional harmonics may affect vowel identification. In their experiment, listeners identified stimuli along an F1 continuum between /I/ and /e/. Significant shifts in the identification boundaries were found when harmonics near the F1 peak were boosted by either 6 or 12 dB. Increments in harmonics remote from the F1 peak also produced significant shifts, even though the increments were not sufficient to make them more intense than the two harmonics closest to the formant peak. However, it is difficult to ascertain whether the boosted harmonics were not perceptually more prominent than the two components near the first formant peak. It seems likely, however, that there may be a contribution of additional harmonics, although this contribution is small relative to that of the two most prominent harmonics.

3.5 Experiment Three: Effects of fundamental frequency on vowel matching

3.5.1 Introduction

A number of previous studies have shown that the timbre of steady state, periodic sounds depends primarily on the shape of the spectrum envelope, and is largely independent of fundamental frequency (for a review, see Plomp, 1976). However, there is evidence that vowel quality may be affected by fundamental frequency changes. Several studies have reported boundary shifts in vowel identification with changes in fundamental (e.g. Miller, 1953; Fujisaki & Kawashima, 1968; Slepokurova, 1973; Carlson, Fant, & Granstrom, 1975; Traunmuller, 1981). For a constant spectrum envelope, an increase in fundamental frequency results in an upward shift in vowel boundaries in both F1 and F2. A decrease in fundamental frequency has the opposite effect. These boundary shifts may reflect a perceptual compensation for apparent vocal tract size differences: higher-pitched voices are typically associated with higher formant ranges. However, this correlation is not perfect in natural speech (Peterson, 1952; Peterson & Barney, 1952); and the magnitude of fundamental frequency effects on vowel perception appears to be somewhat variable

(Slepokurova, 1973; Ainsworth, 1975).

There are several possible ways to describe how fundamental frequency or spectral fine structure information may be incorporated in a perceptual formant estimation process. One possibility is that formant locations are estimated by a local centre of gravity procedure which applies to a fixed frequency interval around the formant peak; such a procedure does not depend on the perceptual resolution of individual harmonics.

A second possibility is a formant estimation process which depends on the prior estimation of harmonic locations. In this case, the width of the weighting function used to estimate the locations of formant peaks will depend on the fundamental frequency. Components which are not multiples of the same fundamental frequency will not be incorporated into the vowel percept.

It is not clear how the first procedure could account for the identification of vowels in the presence of noise (Pollack & Pickett, 1958) or simultaneous speech, an ability which plays an important role in speech communication (Cherry, 1965). A common fundamental frequency appears to be important in determining which spectral components form part of a vowel sound and which are associated with extraneous sounds (Scheffers, 1979; Brokx & Nooteboom, 1982). Better identification scores are obtained when each member of a pair of simultaneously presented vowels differs in fundamental frequency (Scheffers, 1983). A single formant, added to a vowel, is more likely to affect vowel quality if it shares a common fundamental (Darwin, 1981). This suggests a pitch-matched (e.g., comb filter) analysis which allows components sharing a common fundamental frequency to be grouped together.

However, there is evidence that the presence of a common fundamental frequency is not essential for determining whether spectral components stem from a common source. Scheffers (1983) has shown that simultaneously presented pairs of noise-excited vowels are identifiable at better than chance levels.

One source of evidence bearing on the question of the interaction of fundamental frequency and spectrum envelope characteristics in the perception of vowel height is the effect of harmonic amplitudes on vowel matching at different fundamental frequencies. The effects of harmonic amplitudes on vowel height may be restricted to a fixed frequency region around the formant peak; or to a fixed number of harmonic components (suggesting a pitch matched process).

Experiment 3 involved two stimulus types, one with low pitch (f_0=125 Hz, typical of an adult male voice) and the other with high pitch (f_0=250 Hz, typical of an adult female voice). The effects of a 9 dB attenuation of either the lower or higher frequency members of a pair of harmonics were investigated and compared with a control condition comprised of a pair of harmonics of equal amplitude. However, the control condition differed from that of Experiment 2 in that both harmonics were attenuated by 9 dB, relative to the peak in the F1 region of the baseline stimulus. This change was adopted when it was discovered in post-experiment interviews that some of the stimuli in the control condition of Experiment 2 had a somewhat muffled quality, and were perceived as back vowels by some of the listeners. This was most noticeable for stimuli with triplets of higher frequency harmonics.

3.5.2 Method

The experimental method was similar to that used in Experiment 1.

Stimulus materials

Within each f_0 condition, the matching and reference stimuli had the same fundamental frequency. A low pitch (f_0=125 Hz) and a high pitch (f_0=250 Hz) condition was established. In the high pitch condition, the frequencies of the formants above F1 were scaled upward by 20 percent; formant bandwidths were identical to those used in the low pitch condition.

The matching stimuli described in Experiment 1 were used in the low pitch condition. In the high pitch condition, a matching continuum was produced with F1 ranging from 200 Hz to 1350 Hz in steps of 50 Hz. This larger step size was used to accomodate the wider frequency range needed in matching the high pitch stimuli.

The same procedure was used in producing the reference stimuli in the low and high pitch conditions. Starting with a baseline stimulus from the continuum with F1=500 Hz, all harmonics in the frequency region up to and including 1500 Hz were deleted and replaced with a single pair of adjacent harmonics.

In the reference stimuli of the low pitch condition, pairs of harmonics were included from four frequency settings: f2+f3 (250, 375 Hz); f3+f4 (375, 500 Hz); f4+f5 (500, 625 Hz); and f5+f6 (625, 750 Hz). In the high pitch condition, the four pairs were: f1+f2 (250, 500 Hz); f2+f3 (500, 750 Hz); f3+f4 (750, 1000 Hz) and f4+f5 (1000, 1250 Hz).

Three amplitude conditions were used: a control condition in which the both members of the pair were attenuated by 9 dB relative to the F1 peak in the baseline stimulus; an upper edge condition, in which the lower harmonic of the pair was equal to the peak, and the higher member was attenuated by 9 dB; and a lower edge condition in which the higher member of the pair was equal to the peak while the lower member was attenuated by 9 dB.

Procedure

The syhthesis procedure and stimulus presentation was identical to that described for Experiment 1. Each of the listeners (3 men, 2 women) completed 4 sessions comprised of 3 replications of each stimulus in the low and high pitch conditions. Each session lasted approximately 1/2 hour.

3.5.3 Results and discussion

Means and standard deviations of median matched F1 for the low pitch condition showed a pattern of results very similar to those of Experiment 2 (see Figure 3.25). In the present experiment, the effects of attenuation of either the upper or lower member of the pair of harmonics were more nearly symmetrical. Slightly higher matched F1s were found in the lower edge condition (a mean shift of 15.6 Hz or 3.1 percent of the mean matched F1 in the control condition), while a shift toward lower matched F1s was found in the upper edge condition (a mean shift of 50.0 Hz or 9.9 percent relative to the control). The control stimuli were generally matched with lower F1 values in the present experiment, compared with Experiment 2; this difference may be responsible for the reduced asymmetry.

The matching pattern in the high pitch condition closely paralleled that of the low pitch condition. The

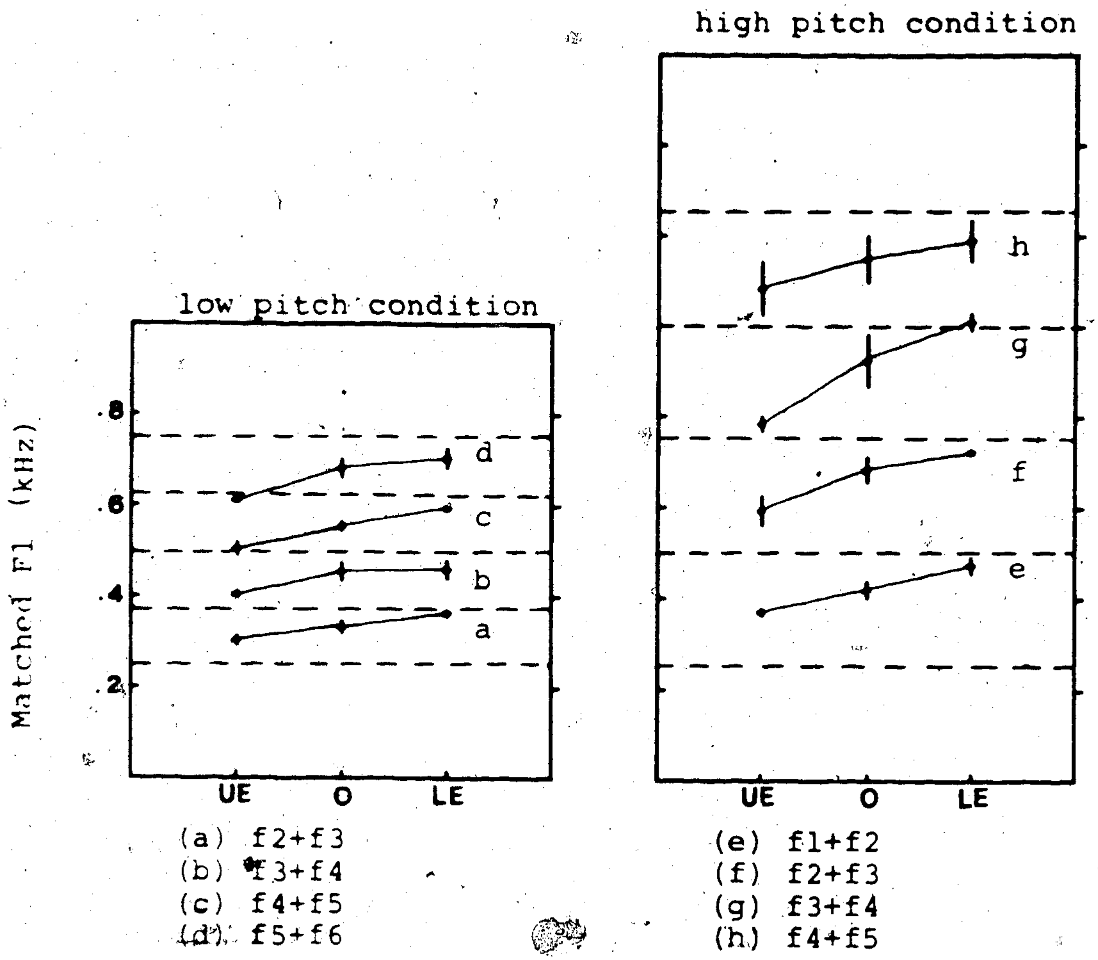


Figure 3.25 Means and standard deviations of matched F1 values for the low pitch condition (left) and high pitch condition (right) of experiment three.
UE: upper harmonic of pair attenuated by 9 dB
O: both harmonics attenuated by 9 dB
LE: lower harmonic attenuated by 9 dB
Dashed lines indicate frequencies of harmonics

asymmetrical effects of harmonic attenuatiion were less pronounced in the high pitch condition, but were still present. In the lower edge condition, 9 dB attenuation led to a mean upward shift of 55.0 Hz (or 6.9 percent, relative to the control) was observed, while in the upper edge condition there was a mean downward shift of 86.3 Hz (or 10.9 percent, relative to the control).

To determine whether harmonic amplitude effects were proportional at different fundamental frequencies, each of the median matched F1 values was divided by the fundamental frequency. A comparison was made between stimuli with the same harmonic in the high and low pitch conditions. This included harmonic pairs f2+f3, f3+f4, and f4+f5 -- the 3 lowest frequency pairs in the low pitch condition, and the 3 highest frequency pairs in the high pitch condition.

The transformed medians were subjected to a repeated measures analysis of variance for the factors FUNDAMENTAL FREQUENCY, HARMONIC RANK, and AMPLITUDE. No significant main effect was found for FUNDAMENTAL FREQUENCY, but there were significant effects of HARMONIC RANK ($F(2,8)=3.16.5$; $p<.001$) and AMPLITUDE ($F(2,8)=104.0$; $p<.001$). None of the two-way interactions was significant, but there was a small three-way interaction ($F(4,16)=3.62$; $p<.03$). This interaction appeared to be due mainly to differences between the high and low pitch conditions for stimuli with harmonics f2+f3 in the lower edge condition. The reasons for this discrepancy are not clear, but one possibility is that this stimulus happened to occur near a category boundary where sensitivity to small acoustic contrasts may be enhanced (Repp, Healy & Crowder, 1979). In addition, as noted above, harmonic attenuation led to smaller asymmetries in the high pitch condition. In spite of these discrepancies, it would appear that the matching data are nearly parallel in the two pitch conditions.

The main results of this experiment can be summarized in terms of three findings:

(1) Matched F1s typically range between fL, the frequency of the lower harmonic of the reference, and fH, the frequency of the higher. Attenuating the lower harmonic relative to the higher causes a shift in matched F1s toward fH, while attenuating the higher component relative to the lower induces a shift toward fL. This pattern may be described as a <u>local centre of gravity effect</u>.

(2) The effects of attenuation are asymmetrical: Attenuation of the higher frequency harmonic produced larger shifts in matched F1 than attenuation of the lower[1].

(3) The ratic of matched F1 tc f_c was very similar in the

[1] Compared with the 2 harmonic stimuli of Experiment 2, listeners matched the control stimuli of the low pitch condition in the present experiment with somewhat lower F1 values. The reasons for this discrepancy are not clear, but the fact that the two harmonics in the control condition of the present experiment were attenuated by 9 dB (relative to the F1 peak in the baseline stimulus) while those of Experiment 2 were not attenuated, suggests that factors such as overall spectral balance in the low and frequency regions may have an effect on perceived height.

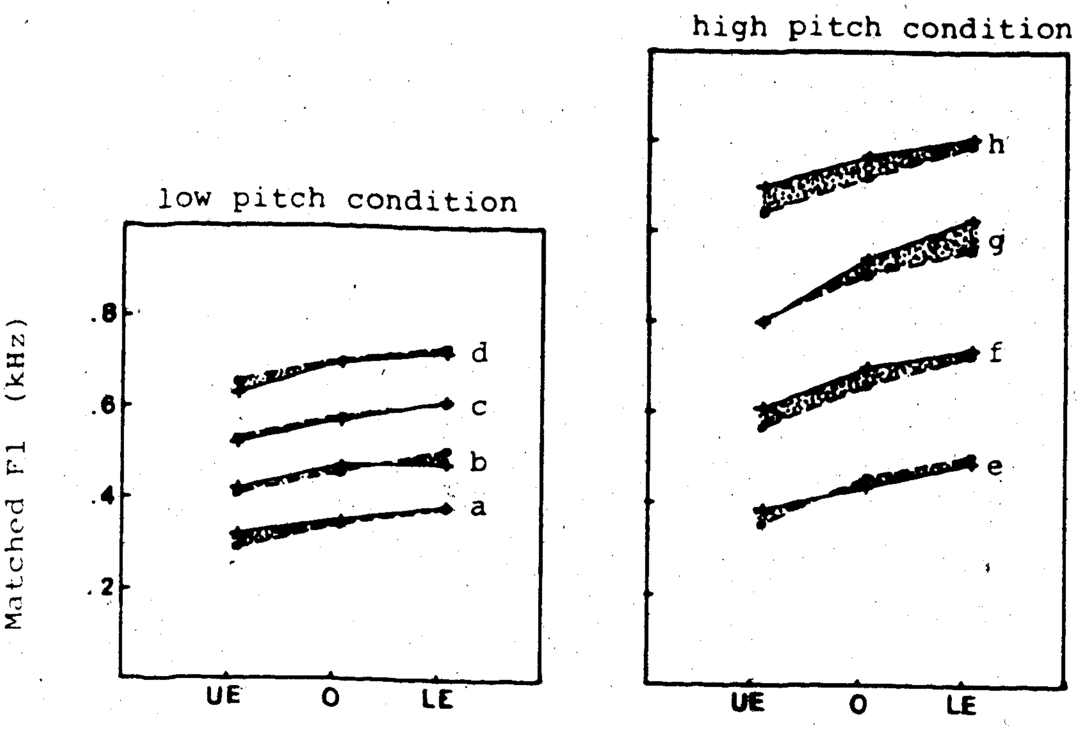
two pitch conditions, for stimuli with fL and fH of the same harmonic rank. It may be concluded that the effects of harmonic amplitude changes on vowel matching are nearly constant at high and low fundamental frequencies.

3.5.4 Predictions of the matching data

What kind of formant estimation procedures would predict the above results? A local centre of gravity procedure which calculates the weighted mean of the two most prominent harmonics might produce fairly accurate predictions, as suggested by the results of Experiment 1. Such a procedure requires that the frequencies and amplitudes of the two most prominent harmonics be known in advance.

Very close predictions of the matching data in both pitch conditions were obtained using the weighted mean of the two most prominent harmonics of the 6 dB/octave preemphasized power spectrum (Figure 3.26), as described in section 3.3.7. The RMS deviation between predicted and observed F1s was 37.68 Hz in the low pitch condition and 75.91 Hz in the high pitch condition.

A global spectral fitting procedure such as LPC analysis effectively excludes the influence of spectral fine structure, producing an interpolated spectrum from which formant locations can be estimated. LPC analyses of the stimuli (using the procedure described in Experiment 1) were



(a) f2+f3
(b) f3+f4
(c) f4+f5
(d) f5+f6

(e) f1+f2
(f) f2+f3
(g) f3+f4
(h) f4+f5

Figure 3.26 Model predictions for the low pitch condition (left) and high pitch condition (right) of experiment three based on the weighted mean of the two most prominent harmonics of the preemphasized power spectrum. Crosses/solid lines indicate observed F1 means, dots/dashed lines indicate predicted F1 values
UE: upper harmonic attenuated by 9 [illegible]
O: both harmonics attenuated by 9 [illegible]
LE: lower harmonic attenuated by 9 [illegible]

carried out and the resulting measurements were used to predict the matching data. Predicted and observed F1 values based on LPC estimates are shown in Figure 3.27.

The predictions in the low pitch condition are very close, with an RMS error of only 35.7 Hz. In the high pitch condition the predicted F1s generally underestimated the observed means, with considerable deviations in the 2 highest curves; the overall RMS error was 90.0 Hz. The discrepancies between observed and predicted F1s in the high pitch condition may be due to the fact that LPC estimates of formant frequencies are generally less accurate at high fundamental frequencies (Markel and Gray, 1976). However, predictions based on LPC estimates of the F1 peak are not substantially worse than those based on a local weighting of harmonic levels.

The successful performance of these two procedures may be due to the precision with which they estimate the location of the F1 peak. Further data, possibly employing a larger range of f_0 values, may help to determine whether the perception of vowel height is better characterized in terms of a local weighting of harmonics or a global spectrum interpolation process followed by peak estimation.

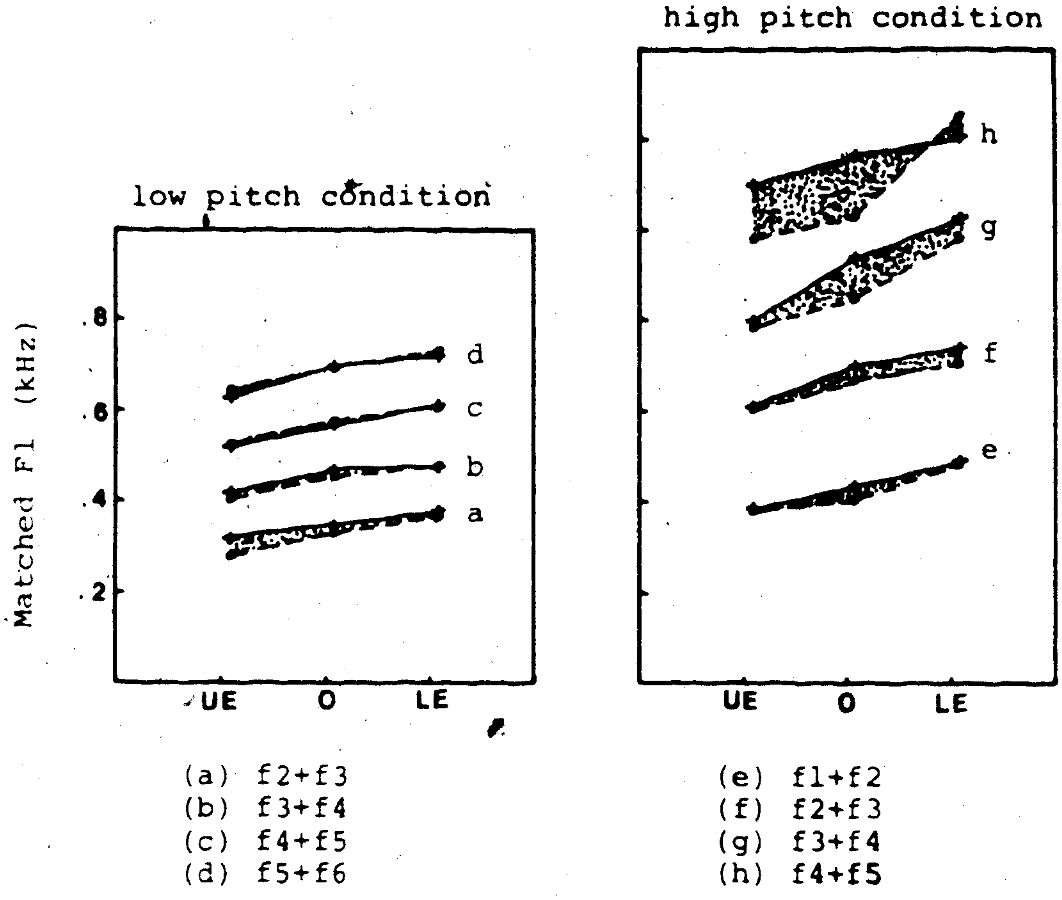


Figure 3.27 Model predictions for the low pitch condition (left) and high pitch condition (right) of experiment three based on LPC estimates of the first spectral peak. Crosses/solid lines indicate observed F1 means, dots/dashed lines indicate predicted F1 values

UE: upper harmonic attenuated by 9 dB
O: both harmonics attenuated by 9 dB
LE: lower harmonic attenuated by 9 dB

3.6 Experiment 4: The effect of filtering on vowel identification

3.6.1 Introduction

The previous matching experiments have shown that vowels in which all but a small number of prominent, low frequency harmonics in the F1 region are deleted can be matched in vowel height with full spectrum vowels from an F1 continuum. It was not possible to establish with certainty that other components (such as the lowest member of a triplet of harmonics) affect vowel matching. The two most prominent harmonics in the F1 region appear to be very important for the perception of vowel height.

It remains a possibility, however, that each of the remaining harmonics in the F1 region also makes a small contribution to perceived vowel height. A vowel identification task may be a more sensitive measure of this contribution, for 3 reasons:
(1) listeners are asked to judge only a single stimulus rather than a pair of stimuli, reducing the memory load;
(2) responses are in terms of natural categories; and
(3) listeners are less likely to be influenced by extraneous, non-phonetic aspects of vowel timbre such as perceived voice quality, characteristics which may add to the difficulty of a matching task.

The contribution of individual harmonics in the F1 region may be determined by comparing listeners' identification responses to vowel stimuli on an F1 continuum in different filter conditions. Changes in the response profiles can be measured in terms of shifts in the vowel boundaries (defined as the 50 percent crossover point between adjacent response categories along the F1 continuum) or vowel centres (defined as the centroid of the identification function for a given vowel response, over the entire F1 continuum).

In each of the previous matching experiments, the effects of a small number of prominent F1 harmonics were investigated. The present experiment set out to measure the effects of all remaining F1 harmonics, above or below the two most prominent. Deleting harmonics below the F1 peak raises the centre of gravity in the F1 region, while deleting harmonics above the peak lowers it. If additional harmonics in the F1 region can influence vowel height, then it might be predicted that increasing the F1 centre of gravity will result in a lowering of vowel height, while decreasing the F1 centre of gravity will raise vowel height (consistent with the effects of changes in the levels of the two most prominent harmonics).

Four conditions were set up to test these predictions using a vowel identification task. The control condition was comprised of synthetic front vowels on an F1 continuum ranging from 250 to 850 Hz. This range includes typical measured F1 values for the vowels /i I e ɛ æ/ of adult male speakers of western Canadian English (Assmann, 1979).

In the *high pass* condition, all harmonics below the two most prominent were deleted. Each stimulus in this condition thus has a *higher* F1 centre of gravity than the corresponding control stimulus. Vowel centres and boundaries in the high pass condition are expected to shift to *lower* values, relative to the control.

In the *notch* condition, all harmonics above the two most prominent were deleted, up to and including 1500 Hz. The F1 centre of gravity is *lowered*, and vowel centres and boundaries are predicted to shift toward *higher* frequency values.

In the *two harmonics* condition, the manipulations of the high pass and notch conditions were combined, resulting in stimuli with only the two most prominent harmonics remaining in the F1 region. The direction and amount of shift in the centre of gravity is dependent on the number and relative amplitudes of harmonics above and below the two most prominent, and identification functions should reflect this.

If the two most prominent harmonics in the F1 region are the exclusive determinants of vowel height, then all four conditions are expected to produce identical identification functions.

3.6.2 Method

Stimulus materials

Four separate F1 continua were constructed, as discussed above. Line spectra for 3 typical stimuli from the continuum in the 4 conditions are shown in Figure 3.28.

Control condition. Stimuli in the control condition were identical to the matching stimuli used in earlier experiments. F1 values ranged from 250 to 850 Hz in 50 Hz steps, for a total of 13 stimuli. The fundamental frequency was constant at 125 Hz. The values of the higher formant frequencies and bandwidths were the same as those in Table 3.1.

Two harmonics condition. The two most prominent harmonics i_1 and ${}_2$ were determined from the 6 dB/octave preemphasized power spectrum using the procedure outlined in section 3.3.7. Frequency values of the two most prominent harmonics, f_1 and f_2 are shown in Table 3.7 for each stimulus.

The amplitude spectrum A(f) for each control stimulus was modified to produce the spectrum A'(f) for the two harmonics condition, in accordance with the following procedure:

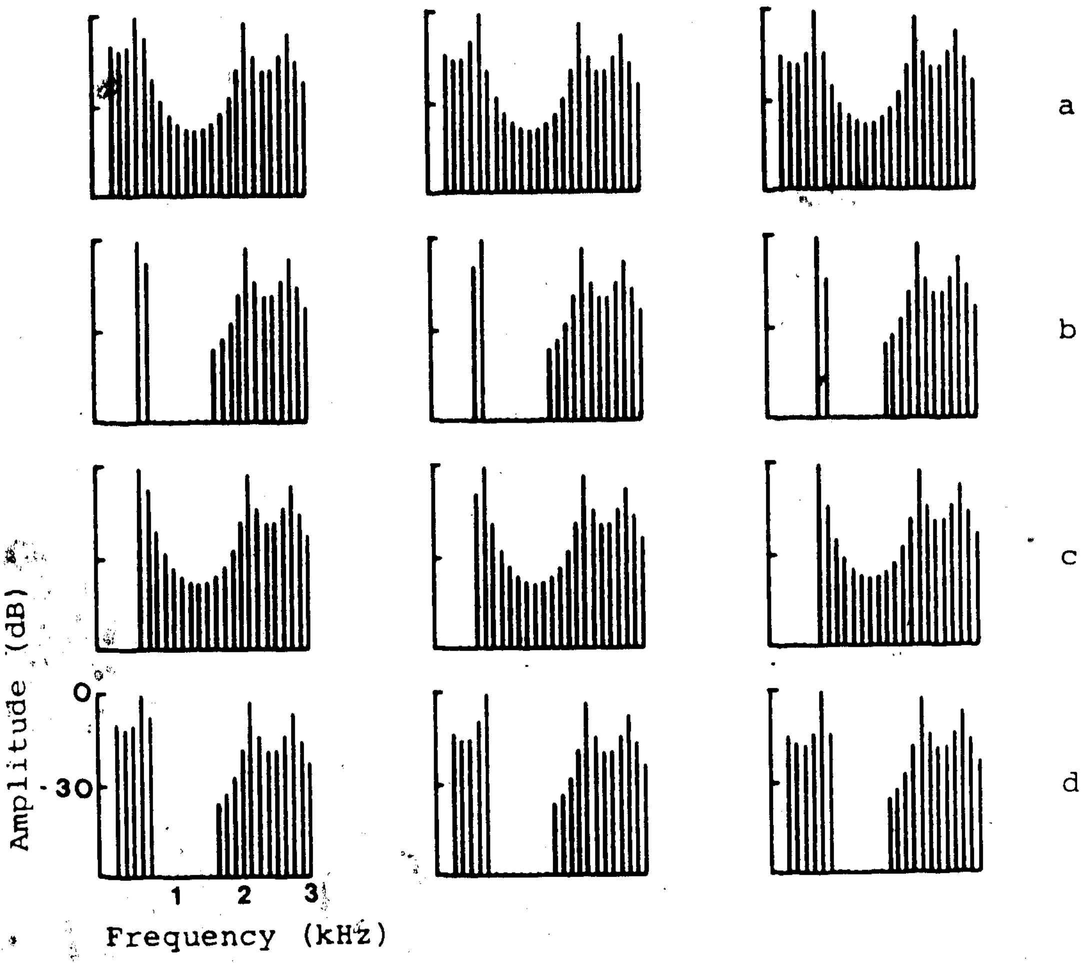


Figure 3.28 Line spectra for three representative stimuli in the 4 conditions of experiment four.
(a) CONTROL condition
(b) TWO HARMONICS condition
(c) HIGH PASS condition
(d) NOTCH condition

$$A'(f) = \begin{cases} 0, & f<\min(f_{i1},f_{i2}) \\ A(f), & f=f_{i1} \text{ or } f=f_{i2} \text{ or } f>1.5 \text{ kHz} \\ 0, & 1.5 \text{ kHz}\geq f>\max(f_{i1},f_{i2}) \end{cases}$$

High pass condition. In this condition all harmonics below the two most prominent were deleted. The two most prominent harmonics were identified as described above, and the following rule was applied to derive the amplitude spectrum A''(f) of each stimulus in the high pass condition:

$$A''(f) = \begin{cases} 0, & f<\min(f_{i1},f_{i2}) \\ A(f), & \text{elsewhere} \end{cases}$$

Notch condition. Harmonics above the 2 most prominent (up to and including 1.5 kHz) were deleted, resulting in a "notched" spectrum:

$$A'''(f) = \begin{cases} 0, & 1.5 \text{ kHz}\geq f>\max(f_{i1},f_{i2}) \\ A(f), & \text{elsewhere} \end{cases}$$

Digital additive harmonic synthesis was used to produce the stimuli in each of the four conditions as described in experiment one. The waveforms were scaled so that each had the same absolute peak level. All other aspects of the synthesis procedure were identical to those descibed for Experiment 1.

Subjects were individually tested in a sound treated room. They were instructed to label each stimulus as one of the 5 front vowels of Canadian English, /i I e ɛ æ/. A

				LPC estimated F1:			
Nominal F1	i1	i2	FMPH2	C	2H	HP	N
250	2	3	257	270	267	271	258
300	2	3	290	320	305	321	305
350	3	2	356	381	362	380	363
400	3	4	391	398	394	410	394
450	4	3	454	476	464	478	462
500	4	3	491	506	497	506	495
550	4	5	544	540	549	567	540
600	5	4	608	617	609	627	611
650	5	6	644	634	640	649	636
700	6	5	708	711	707	723	706
750	6	7	755	760	765	765	751
800	6	7	798	788	797	801	786
850	7	6	860	866	858	865	852

Table 3.7. Harmonic rank and weighted mean frequency of the two most prominent harmonics (in Hz) and and LPC based F1 estimates (in Hz) for the stimuli of Experiment 4. FMPH2 values are the same for corresponding stimuli in the 4 conditions.

multiple choice switchbox was used to record the responses. The labels contained a phonetic symbol and a keyword below it:

/i/	/I/	/e/	/ɛ/	/æ/
heed	hid	hayed	head	had

Each stimulus was repeated, with a 5 second inter-stimulus interval, until a response button was pressed. If a wrong button was pressed by mistake, listeners could cancel the response by pressing a different button immediately afterwards. Otherwise, the response was recorded and the next stimulus item was presented.

Each of the 13 stimuli in the 4 conditions appeared in 4 successive random orderings (for a total of 208 items) in each session. Listeners completed 5 sessions, each lasting approximately 25 minutes.

Listeners

Nine phonetically trained listeners (graduate students and staff in the Department of Linguistics at the University of Alberta) participated in the experiment. All had normal hearing.

3.6.3 Results and discussion

The identification functions (pooled across listeners) for the control, high pass and notch conditions are shown in Figure 3.29. Each point on the curve is based on 20 trials from each of the 9 subjects. It is apparent that the identification functions are very similar in each of these conditions. This suggests that information specifying vowel height differences is largely provided by the 2 most prominent harmonics.

According to the F1 centre of gravity hypothesis discussed above, the response curves may be expected to shift toward lower F1 values in the high pass condition

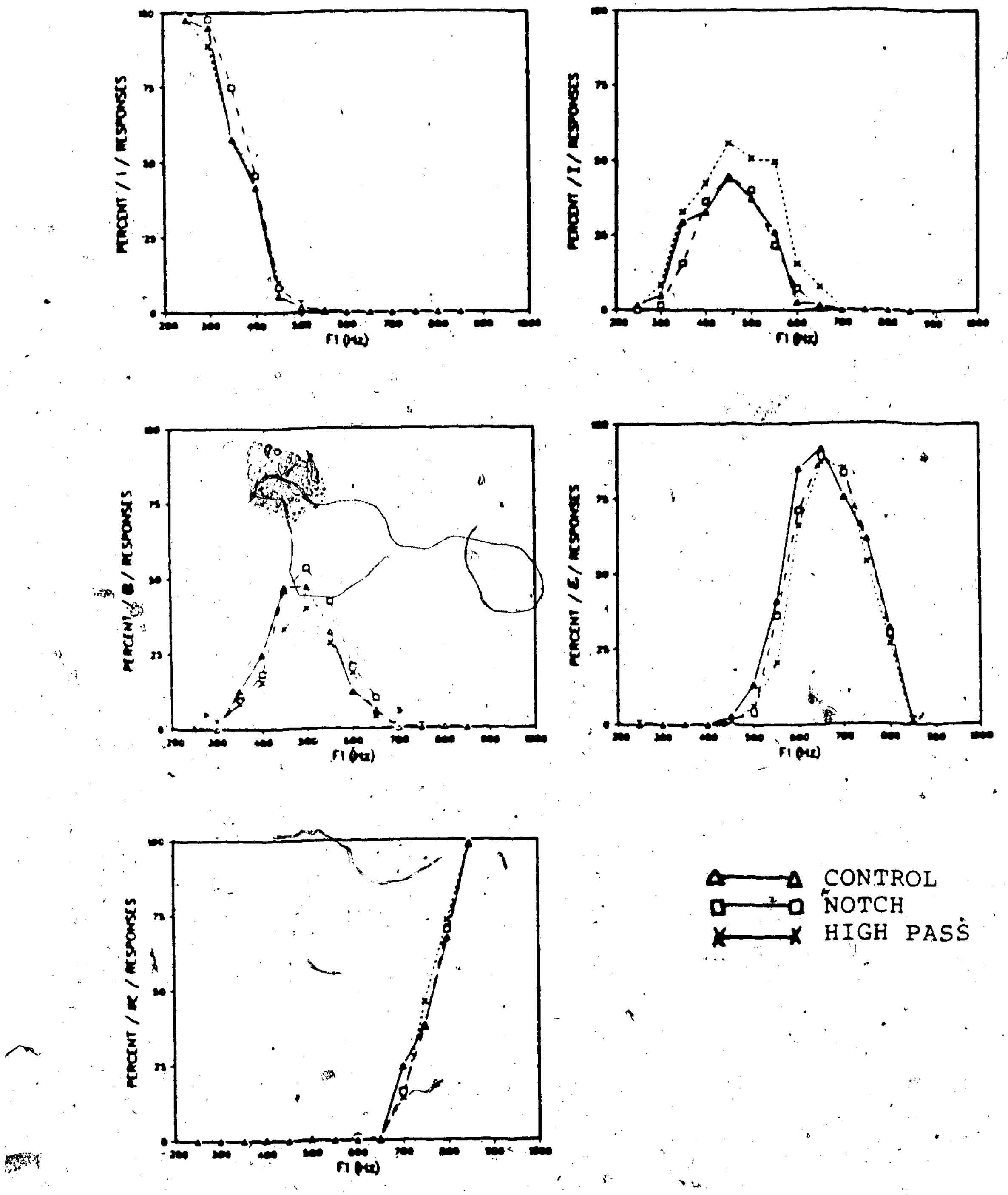


Figure 3.29 Pooled identification functions for the CONTROL, NOTCH, and HIGH PASS conditions of experiment four

relative to the control, and toward higher F1s in the notch condition. Vowel response curves did not shift to lower F1 values in the high pass condition; in fact, there was a small shift in both curves toward higher F1s for some of the vowels. Rather than a shift in F1, the identification functions for some of the vowels appeared to shift along the vertical axis, particularly for the vowels /I/ and /e/. In the high pass condition there were more /I/ responses, and fewer /e/ responses, than in the control condition. A similar change appears for /I/ and /e/ in the two harmonics condition, relative to the control (Figure 3.10). One possibility is that filtering had an effect on vowel identification which was dissociated to some extent from changes in F1.

The vowels /I/ and /e/ did not appear to be fully specified on the F1 continuum, since their identification curves overlapped considerably and their peak values did not exceed more than 55 percent. Natural tokens of /I/ and /e/ in western Canadian English also differ in F2, duration and diphthongization which may help to distinguish these vowels perceptually (Assmann, Nearey and Hogan, 1982; Nearey & Assmann, 1984).

In a perceptual study using natural tokens of isolated vowels (Nearey & Assmann, forthcoming) these differences were neutralized by gating 30 ms segments from near the onset and offset of each vowel. It was found that these two

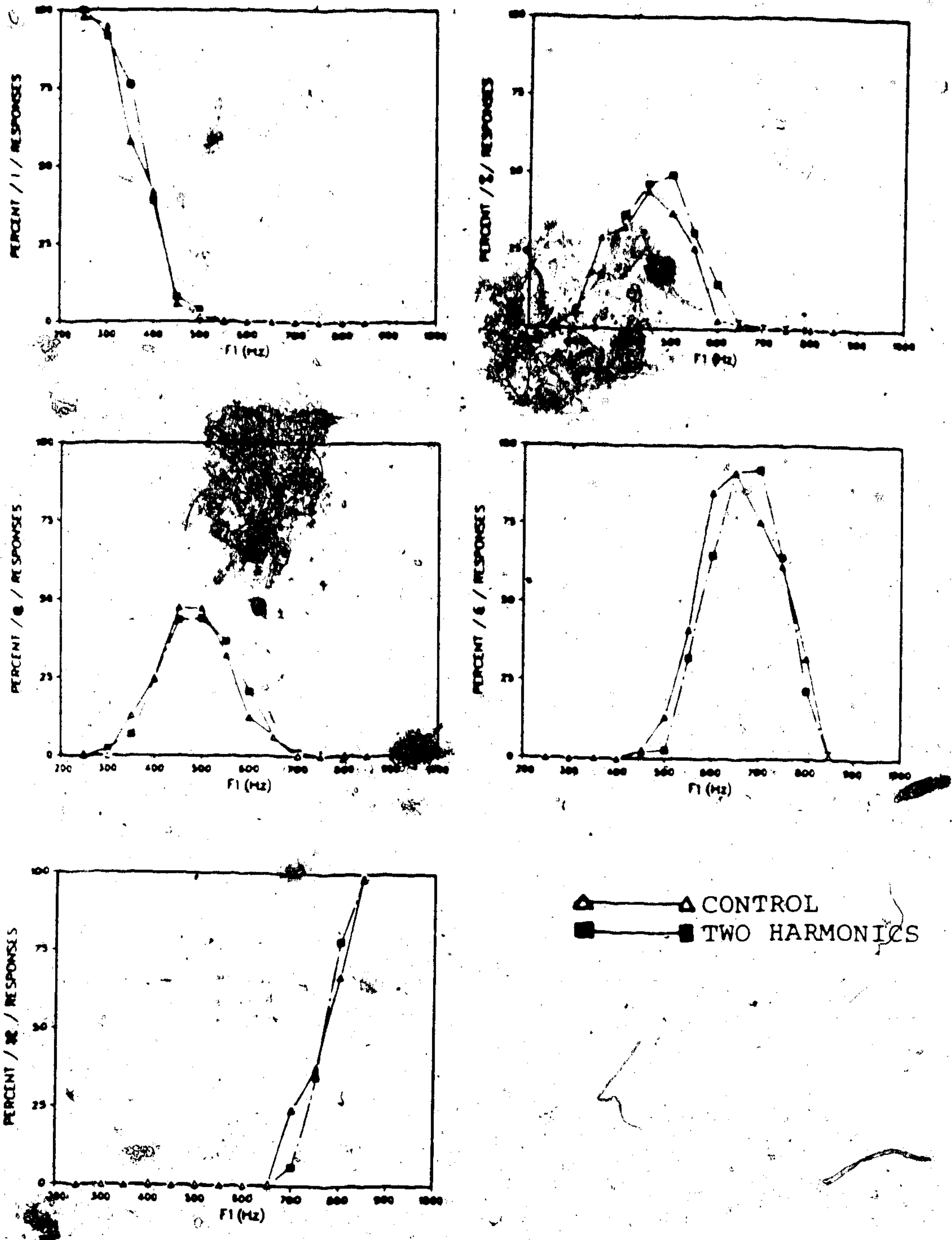


Figure 3.30 Pooled identification functions for the CONTROL and TWO HARMONICS conditions of experiment four

segments (separated by 10 ms of silence) were sufficient for accurate vowel identification. Reversing the temporal order of the segments, or presenting the initial portion twice, resulted in an increase in confusion errors. The vowels /I/ and /e/ showed the largest increase in errors, indicating the potential importance of diphthongization. In the present experiment, the effects of diphthongization and other variables were excluded by using a constant fundamental frequency and formant pattern. This may have been the reason for lower identification scores for these two vowel categories.

To determine the direction and magnitude of the shifts in the identification functions, an analysis of vowel centres and vowel boundaries was carried out.

Vowel centres. The centroid or weighted mean of each vowel response curve was calculated from the identification data from each listener. Means and standard deviations (across listeners) are shown in the top portion of Figure 3.31. Vowel centres do not differ substantially in the four conditions. The mean inter-vowel distance is approximately 125 Hz. The largest shift is only 27 Hz, a little more than one fifth of this range; most vowels differed by less than 15 Hz from the control condition. All of the shifts in the three filter conditions were in the direction of higher F1 values.

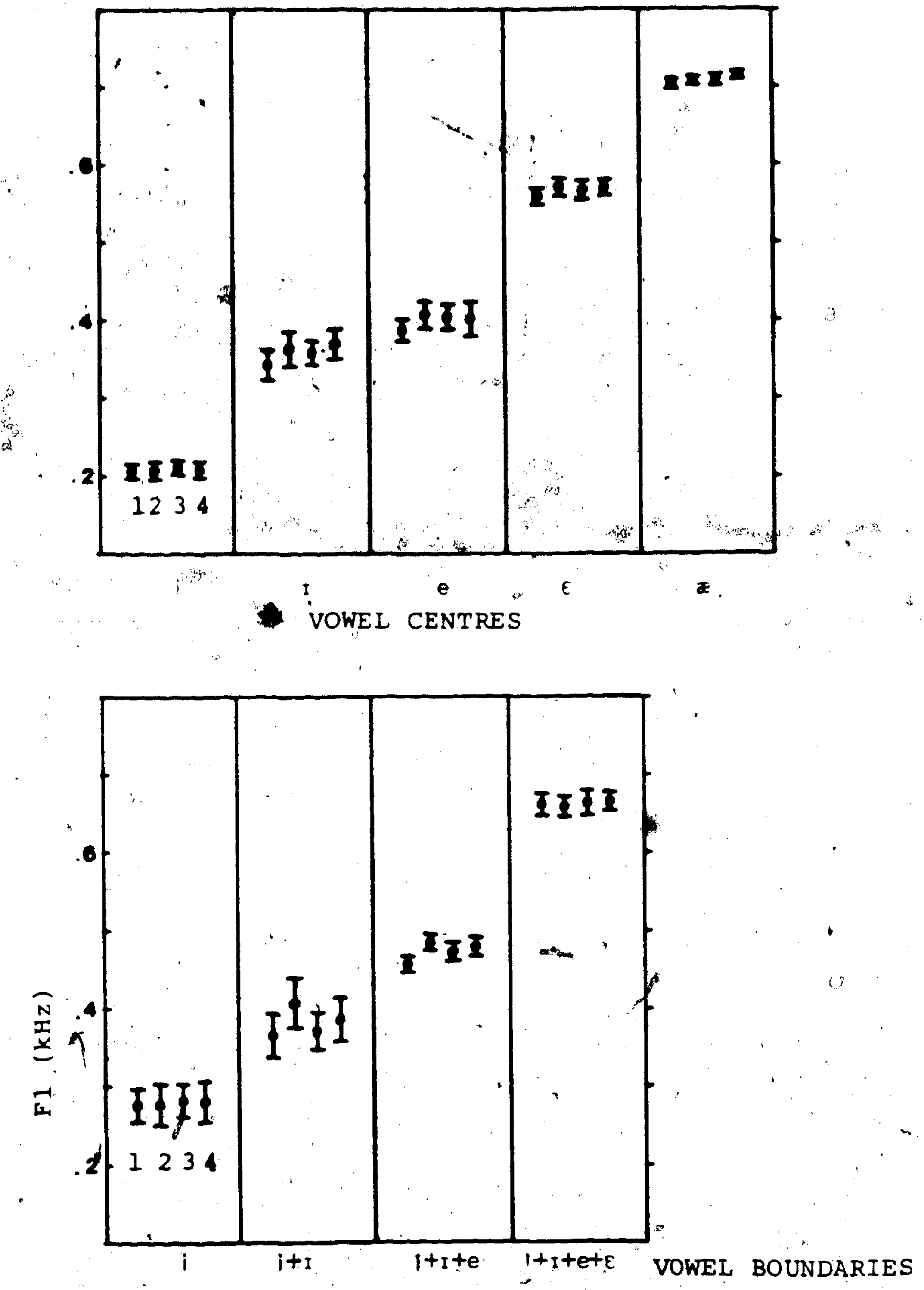


Figure 3.31 Means and standard deviations for vowel centres and vowel boundaries in the 4 conditions of experiment four. 1=CONTROL 2=HIGH PASS 3=NOTCH 4=TWO HARMONICS

A repeated measures analysis of variance was conducted, with the factors VOWEL, CONDITION, and LISTENER. Significant main effects were observed for the factors VOWEL (F(4,32)=760.6; p<.001), CONDITION (F(3,24)=4.61; p<.02) and for the interaction of VOWEL by CONDITION (F(12,96)=2.59; p<.01). Dunnett's test for comparing treatment means with a control was applied, and the results are shown in Table 3.8. The 3 filter conditions differed significantly from the control for the vowels /I/ and /e/. Statistically significant, but smaller effects were found in the high pass and two harmonics conditions for the vowel /ɛ/, and in the two harmonics condition for /æ/. These data do not provide strong support for a formant estimation process based the weighted mean of 3 or more harmonics. The shifts were generally quite small, and were not present for each vowel category, as might be expected. The shifts in the high pass condition were in the opposite direction to those predicted on the basis of a change in the centre of gravity: instead of lower F1 values, the response curves shifted to higher values.

Vowel boundaries. Vowel boundaries were estimated by probit analysis (Finney, 1971). A sequential approach was used to estimate boundaries on the high frequency slopes of the curves, as follows. The vowels were first ordered according to their distribution along the F1 scale in natural speech: 1=/i/, 2=/I/, 3=/e/, 4=/ɛ/, 5=/æ/. The first boundary was estimated as the 50 percent point on the fitted curve for

	/i/	/I/	/e/	/ε/	/æ/
C	305.7	442.9	488.1	659.2	803.8
HP	307.2	464.1**	507.5**	670.5*	807.6
N	310.6	458.6**	505.0**	666.9	808.8
2H	308.0	470.1**	501.9**	669.7*	815.0*

C= control, HP= high pass, N= notch, 2H= two harmonics
*p<.05 **p<.01

Table 3.8. Estimated vowel centres for the identification data of Experiment 4. Means and significance levels for Dunnett's test comparing each of the three filter conditions with the control

/i/ responses. The second boundary was estimated for the pooled /i/ + /I/ response curves; the third was based on the combined /i/ + /I/ + /e/ response curves; and the fourth, on /i/ + /I/ + /e/ + /ε/ response curves.

The four boundary values were calculated for each listener in each of the four conditions. Means and standard deviations are shown at the bottom of Figure 3.31. The results are fairly similar to the vowel centre data. There is again a tendency for higher F1 boundaries in all filter conditions, relative to the control. The largest shifts took place in the high pass and two harmonics conditions.

A repeated measures analysis of variance was carried out on the vowel boundary data. A significant main effect was found for the factor VOWEL (F(3,24)=193.75; p<.001) but not for filter CONDITION. The interaction of VOWEL x CONDITION was significant ($F(9,72)=3.29$; $p<.01$).

Dunnett's test (Table 3.9) indicated significant shifts toward higher F1 values in the high pass and two harmonics conditions for the two "internal" vowel boundaries, delimiting the /ɛ/ response. Shifts toward higher F1 values were also found in the notch condition, but these were not statistically significant. Table 3.9 indicates the means in each condition. The mean shift in the high pass condition was 17.1 Hz; in the notch condition, 7.6 Hz; and in the two harmonics condition, 13.1 Hz. As in the previous analysis, the results in the high pass condition were opposite to those predicted by a broad F1 centre of gravity hypothesis.

The analyses of vowel centres and boundaries presented above have measured shifts in the identification curves in terms of the nominal F1 values of the stimuli, the formant values specified prior to filtering. If listeners estimate spectral peak locations in determining the identity of a vowel, and if they are able to recover nominal F1 values from the filtered vowels, then the identification curves should be identical in each of the four conditions.

If an alternative peak estimation procedure could be devised to accurately determine the perceived or <u>effective F1</u> for each stimulus, then a projection of the identification curves along this scale should produce a close alignment of the four curves. (It would not, however,

	i	i+I	i+I+e	i+I+e+ε
C	376.6	466.5	558.4	761.7
HP	377.7	508.8**	586.2**	758.8
N	382.7	473.3	573.7	764.0
2H	381.9	487.4*	579.9*	766.4

C= control, HP= high pass, N= notch, 2H= two harmonics
*p<.05 **p<.01

Table 3.9. Estimated boundaries for the identification data of Experiment 4. Means and significance levels for Dunnett's test comparing each of the three filter conditions with the control

remove differences in the overall shape or height of the curves).

LPC based predictions

Since the identification functions differed slightly in the 4 conditions, it appears that the FMPH2 procedure (as formulated above) cannot account fully for the identification results. It was shown earlier that LPC based estimates of the F1 peak gave fairly good predictions of the matching data of experiments one and three. It is of interest to determine the extent to which LPC based estimates of the F1 peak can be used to reduce the discrepancies between the response curves in the four conditions.

LPC analyses were carried out for each stimulus [illegible] four conditions, and the first formant frequency was estimated using the procedure described for Experiment 1. The LPC estimated F1 values in the high pass condition were mostly *higher* than those in the control condition, with a mean increase of 7.1 Hz. In the notch condition the measured values were mostly *lower* than those in the control, with a mean difference of -8.0 Hz. Stimuli in the harmonics condition with low nominal F1s generally had LPC estimated F1 values which were lower than those in the control condition, while those with higher nominal F1 had higher LPC estimated F1s. The mean difference over the entire set was -4.1 Hz.

Vowel centres and boundaries were re-calculated, along a frequency scale of LPC estimated F1s rather than nominal F1. If the perceived difference between stimuli in the four conditions is based on peak locations, estimated in a manner analogous to the LPC procedure, then it may be expected that the identification functions will be even more closely aligned when plotted along a scale of LPC estimated F1 rather than nominal F1. The degree of alignment can be assessed in terms of the size of shifts in vowel centres or boundaries in each filter condition relative to the control.

Replotting the identification functions along a scale of LPC estimated F1 led to a closer alignment of the identification curves in each of the notch and two harmonics conditions to the control condition, but a poorer fit of the high pass to the control conditions. In the analysis based on nominal F1, a mean shift of +11.5 Hz in vowel centres and +17.1 Hz in vowel boundaries was observed in the high pass condition. These shifts were even larger when LPC estimated F1s were used (+17.3 Hz for vowel centres, +25.1 Hz in vowel boundaries). Thus an LPC based measure of the first formant peak does not account for observed shifts in the high pass condition.

In the two harmonics and notch conditions, vowel centres and boundaries calculated along a scale of LPC estimated F1 values were closer to those in the control condition than in the original analysis along a scale of nominal F1. On average, the identification functions were still shifted toward higher frequencies, relative to the control condition; but these shifts were smaller in magnitude.

The degree of mismatch between the identification functions was calculated in terms of a summary statistic, the RMS difference between vowel centres or boundaries in each filter condition and the control. Table 3.10 presents the RMS values for vowel centres and boundaries for each filter condition, using nominal F1 and LPC estimated F1 scales. It can be seen that the LPC estimated scale has smaller RMS values for the notch and two harmonics conditions, but higher values in the high pass condition, indicating a poorer fit.

	Vowel centres		Vowel boundaries	
	Nom F1	LPC F1	Nom F1	LPC F1
HIGH PASS	31.6	36.7	19.9	23.7
NOTCH	24.0	15.5	14.9	10.5
TWO HARMONICS	30.1	27.1	22.8	20.0

Table 3.10. RMS distance (in Hz) between the control and filter conditions in estimated vowel centres and boundaries, along a scale of nominal (Nom) F1 and LPC estimated F1

What kind of measure would lead to a closer alignment of the responses profiles in the high pass and control conditions? A stimulus in the high pass condition has the same nominal F1 value as a corresponding stimulus in the control condition, even though its phonetic quality or effective F1 may be different, due to the filtering process. To align stimuli with the same effective F1, identification curves in the high pass condition must be shifted toward lower frequencies. This will take place only if the measure used to estimate the effective F1 assigns lower frequency values to stimuli in the high pass condition relative to those in the control with the same nominal F1. Even though the centre of gravity of F1 components is actually higher in the high pass condition, listeners respond to these stimuli as if they had lower effective F1s than corresponding control stimuli. The centre of gravity measures discussed earlier in this chapter will therefore not work either.

Upward spread of masking by low frequency harmonics might have the effect of reducing the audibility of components on the low frequency side of the F1 peak. When the low frequency harmonics are removed by high pass filtering, a release from masking may enhance the relative prominence of spectral components at or near the cutoff frequency. The perceived height or effective F1 is thereby lowered, resulting in shifts of the identification functions toward higher frequency values. A similar argument could apply to changes in the two harmonics condition. However, deletion of the first and second harmonics did not have a similar effect of lowering the F1 value of best matches to reference stimuli in the lower edge condition of Experiment 1. It is not clear why this effect should be present in a vowel identification task but not in matching.

Further empirical data are needed to determine how high pass filtering affects the perceptual prominence of harmonics in the F1 region, and whether these effects are responsible for the observed shifts. However, some additional points must be borne in mind. One is that the effects of filtering were apparently not constant across all stimuli in the continuum. The largest effects were observed for the response categories /I/ and /e/. At most 5 harmonics were deleted from stimuli with F1 values near the means for the vowels /I/ and /e/ in the high pass condition; while up to 7 harmonics were deleted at the high frequency endpoint of the continuum. Evidently, the number of deleted harmonics does not determine the size of the shifts.

There is some evidence that the perceptual role of low frequency harmonics may be sensitive to context. Sundberg and Gauffin (1982) found that the effects of changes in the amplitude of the fundamental component on vowel identification were larger when stimuli with different source characteristics were randomized within the same listening session; stimulus presentation in a 'blocked' mode (i.e. with each listening session comprised of a single source type) reduced the effects of f_0 amplitude. The context sensitivity found by Sundberg and Gauffin may result from changes in voice quality. Sundberg and Gauffin (1979) showed that the amplitude of the first harmonic may vary by more than 15 dB in spoken vowels from one utterance to another, depending on the type of phonation. It is possible that the perceptual weighting of harmonics may be adjusted to some degree, depending on perceived voice quality or glottal source characteristics suggesting an adaptive filtering process (see also discussion of the findings of Lindquist and Pauli (1968) below). Similar adjustments are suggested by the finding (Darwin, 1981, 1984) that a harmonic which starts or stops at a different time from the remaining harmonics of a vowel is not fully incorporated into the perceptual estimation of vowel quality.

To this point, discussion has focussed on differences between the control condition and each of the filter conditions. It was noted above that vowel centres and boundaries in the three filter conditions all shifted to higher F1 values relative to the control, with means ranging from 10.1 to 13.0 Hz for vowel centres, and from 24.0 to 31.6 Hz for vowel boundaries. Comparisons among the three filter conditions show that these shifts are very similar in extent.

Considering the radical nature of the filtering operations, it is perhaps surprising that the identification curves are so similar in the four conditions. The mean shift in vowel centres ranged from 1.4 to 2.9 Hz, and the mean shift in vowel boundaries ranged from 4.0 to 8.9 Hz. Each filter condition thus differs more from the control than from any other filter condition; and the shifts are all in the same direction. This result is difficult to reconcile with a local centre of gravity account. The identification profiles in the high pass and notch conditions are very similar; vowel centres differ by less than 1.5 Hz and vowel boundaries by less than 8.9 Hz, on the average. Yet stimuli in these two conditions are the furthest apart in terms of a local centre of gravity in the F1 region.

Rather than local changes in the spectrum near the formant peak, listeners may be attending to changes in the overall spectral balance, resulting from the removal of a number of harmonics from the F1 region. These changes may be perceived as a lowering of the effective F1, resulting in shifts of the identification functions toward higher F1s, relative to the control.

To investigate this idea further, the centre of gravity (centroid) of the amplitude spectrum was computed over the region 0 to 4 kHz. The global centre of gravity was higher in each of the three filter conditions, relative to the control, at every position along the continuum. The global centre of gravity was, moreover, a monotonic function of the nominal F1. If listeners use an acoustic parameter of this sort, then a raising of the global centre of gravity must be interpreted perceptually as equivalent to a lowering of the effective F1.

Some evidence that overall spectral balance may play a role in vowel perception is provided in a study by Lindquist and Pauli (1968). They presented data from an identification experiment with stimuli on an F2 continuum. Swedish listeners identified the stimuli as /i/, /y/, or /u/. A 12.5 dB attenuation of F1 relative to higher formants led to shifts in the identification functions toward higher F2 values, compared with 12.5 dB attenuation of higher formants relative to F1. Lindquist and Pauli (1968: p. 14) suggested that spectral balance may be a contributing factor in vowel perception:

> "Evidently we use the overall change in the distribution of spectral energy as an

> important parameter in the decision mechanism. In connected speech with constant source spectrum, the spectral level change is correlated with the changes in formant frequencies so this can be used as a complement to the tracking of the peaks in the spectrum."

Lindquist and Pauli found that filtering had a larger effect when filtered stimuli were randomized rather than presented together in separate blocks. Since the "blocked" condition presents a more natural context (i.e., spectral balance does not vary randomly in natural speech) it might be of interest to investigate the effects of filtering obtained in the present experiment in a "blocked" rather than a "randomized" presentation mode as in the present experiment.

3.6.4 General discussion

The present experiment investigated the hypothesis that vowel height is determined by the two most prominent harmonics in the F1 region. The identification functions were generally quite similar in each of the four conditions, suggesting that a great deal of information may be provided by the 2 most prominent harmonics. However, some statistically significant differences in vowel centres and boundaries were present, indicating that additional harmonics in the F1 region can have an influence on vowel identification.

Some of the observed shifts in the identification functions are consistent with local centre of gravity predictions, suggesting an influence which extends beyond the two most prominent harmonics. In many cases, however, these predictions were not borne out. Shifts did not occur with every response category; the magnitude of the shifts was variable, and on the average quite small. Most shifts were associated with the vowels /I/ and /e/, catgories which may not have been adequately represented along the F1 continuum. Furthermore, the identification functions in the high pass condition did not shift in the direction predicted by the local centre of gravity hypothesis: vowel centres and boundaries were higher, not lower, than those in the control condition.

An attempt was made to reduce the discrepancies between the identification functions in the 3 filter conditions and the control by replotting them along a scale of LPC estimated F1 values. This strategy did lead to a slightly better alignment of the curves in the notch and two harmonics conditions, but a worse fit in the high pass condition.

It was noted that the identification functions associated with the three filter conditions differed less from one another than from the control. Shift in the filter conditions was in the direction of higher F1s, suggesting that the filtering operations may have affected vowel quality in a similar manner in each condition. One possibility is that listeners were attending to the change in global spectral balance resulting from the deletion of low frequency harmonics, rather than to the change in the local centre of gravity. Previous studies have shown that the overall spectral balance of a vowel can affect its perceived quality, but its influence is highly context sensitive.

Additional experiments are needed to investigate the perceptual basis for these effects and to determine their context sensitivity; for example, by comparing randomized and blocked presentation modes. It is worth pointing out that the effects of filtering on vowel perception are often larger in a randomized than in a blocked presentation mode (Lindquist and Pauli, 1968; Gauffin and Sundberg, 1982). It seems likely, therefore, that the perceptual consequences of filtering explored here are relatively small, second-order effects; the global similarity between the identification functions suggests that listeners attend primarily to the two most prominent harmonics in the F1 region when determining the phonetic identity of steady state front vowels.

3.7 Summary and conclusions

The experiments described in this chapter have provided evidence that vowel height perception is determined largely by the frequencies and amplitudes of the two most prominent harmonics in the F1 region. The major findings can be summarized as follows:

(1) When a band of 2 to 5 adjacent harmonics of equal amplitude was subtituted for components in the F1 region, listeners aligned the F1 peak of a matching stimulus with the highest frequency harmonics in the band. The presence of additional low frequency harmonics had little effect on matching.

(2) Attenuation of the highest frequency member of the band raised vowel height (over a range of at least 9 dB), and listeners selected lower F1 matches. Small shifts in matched F1 were observed as the harmonic was attenuated, rather than a single large discrete change, indicating that judgements were not based exclusively on the most prominent harmonic.

(3) Attenuation of the lowest frequency component of a band of 3 or 4 contiguous harmonics did not lead to significant shifts in matched F1.

(4) Attenuation of the lowest frequency member of a pair of harmonics lowered vowel height, leading to higher F1 matches. The size of these shifts was typically smaller than those obtained with attenuation of the highest frequency component. Several possible explanations for these asymmetrical effects were considered, including masking by adjacent components, lateral suppression, frequency-dependent changes in loudness, and listeners'

expectations of the shape of the spectrum in the F1 region of front vowels. Although more than one factor may be involved, the latter hypothesis provides the most plausible account for the overall pattern of results.

(5) The effect of harmonic level changes on vowel matching was constant with changes in rank numbers of a pair of harmonics; no significant interaction of harmonic rank and harmonic attenuation was found in the matching data.

(6) The effects of harmonic attenuation were similar when the fundamental frequency was raised by an octave (from 125 to 250 Hz). The ratio of matched F1 to f_0 was nearly the same in the two pitch conditions, for corresponding stimuli of the same harmonic rank. (7) Spectral distance measures were investigated for predicting vowel matching data. The best predictors were those which also provided good estimates of the nominal F1 of the matching stimuli. A local peak estimation procedure (weighted mean of the two most prominent harmonics in the F1 region of the preemphasized spectrum) gave the best overall predictions. Similar results were obtained using LPC-based estimates of the F1 peak, except in the high fundamental frequency condition of Experiment 3.

(8) An identification task was used to assess the hypothesis that vowel height depends on a local peak estimation process involving only the two most prominent harmonics. Results indicated that front vowel identification is largely unaffected by the gross changes in spectral shape resulting

from the removal of all harmonics above and/or below the two most prominent. However, there were some differences in the identification functions which were not fully accounted for by either an LPC-based F1 measure, or a local centre of gravity measure incorporating energy in a wider frequency interval around the first formant peak. Each of the filter conditions appeared to have a similar effect on vowel height, possibly the result of the change in spectral balance in the low and high frequency regions of the spectrum.

These results, taken in conjunction with the asymmetrical effects of harmonic levels in matching, suggest that a local peak estimation hypothesis cannot account entirely for vowel height perception; additional aspects of spectral shape may influence vowel height under certain conditions. As a measure of perceived vowel height, however, a local peak estimation procedure provides a close approximation to the perceptual data.

CHAPTER FOUR

A PERCEPTUAL STUDY OF BACK VOWELS

4.1 Introduction

The evidence from matching and identification experiments presented in chapter three indicates that height differences in front vowels are determined largely by a small number of prominent harmonics in the first formant region. While front vowel spectra are characterized by a broad valley separating the first and second formant peaks, in back vowels these peaks can be very close together. Consequently it may be difficult to identify two separate peaks in the low frequency region. This raises a problem for the formant hypothesis of vowel perception: How can the formant locations be determined when the line spectrum contains only a single maximum?

One possibility is that the two formants are not resolved; two formants in close proximity may be effectively "merged" into a single spectral prominence whose centre of gravity determines the vowel quality. This 'formant centre of gravity' (FCOG) hypothesis is evaluated in the present chapter by means of vowel matching and vowel identification experiments, examining the effects of F1-F2 proximity and formant amplitude.

Two alternatives to the FCOG hypothesis can be envisioned. The frequency locations of the first and second formants may be estimated by a process which is based only in part on information given by the locations of formant peaks. Other cues, such as formant amplitudes, overall spectral balance and spectral slope in the region above the formant cluster may provide additional information. Another possibility is that the locations of the two formants are estimated as peaks in a "sharpened" version of the spectrum.

An illustration of the problem. Figure 4.1 indicates the changes in the spectrum which occur when the separation of F1 and F2 is reduced. Line spectra are shown for three vowels calculated using the cascade synthesis model described in section 3.2. In part (a), the formant pattern is appropriate for a front vowel (F1=500 Hz; F2=2500 Hz). In (b) and (c) a back vowel is shown (F1=464 Hz; F2=863 Hz). In (a) and (b) the vowel is produced with a high fundamental frequency (f_0=250 Hz) while in (c) a low fundamental is used (f_0=125 Hz)

The back vowel with a high fundamental frequency shown in (b) does not have separate peaks corresponding to F1 and F2; a simple peak-finding algorithm, applied to the amplitude spectrum, would be unable to determine that two formants were present rather than one. By contrast with the

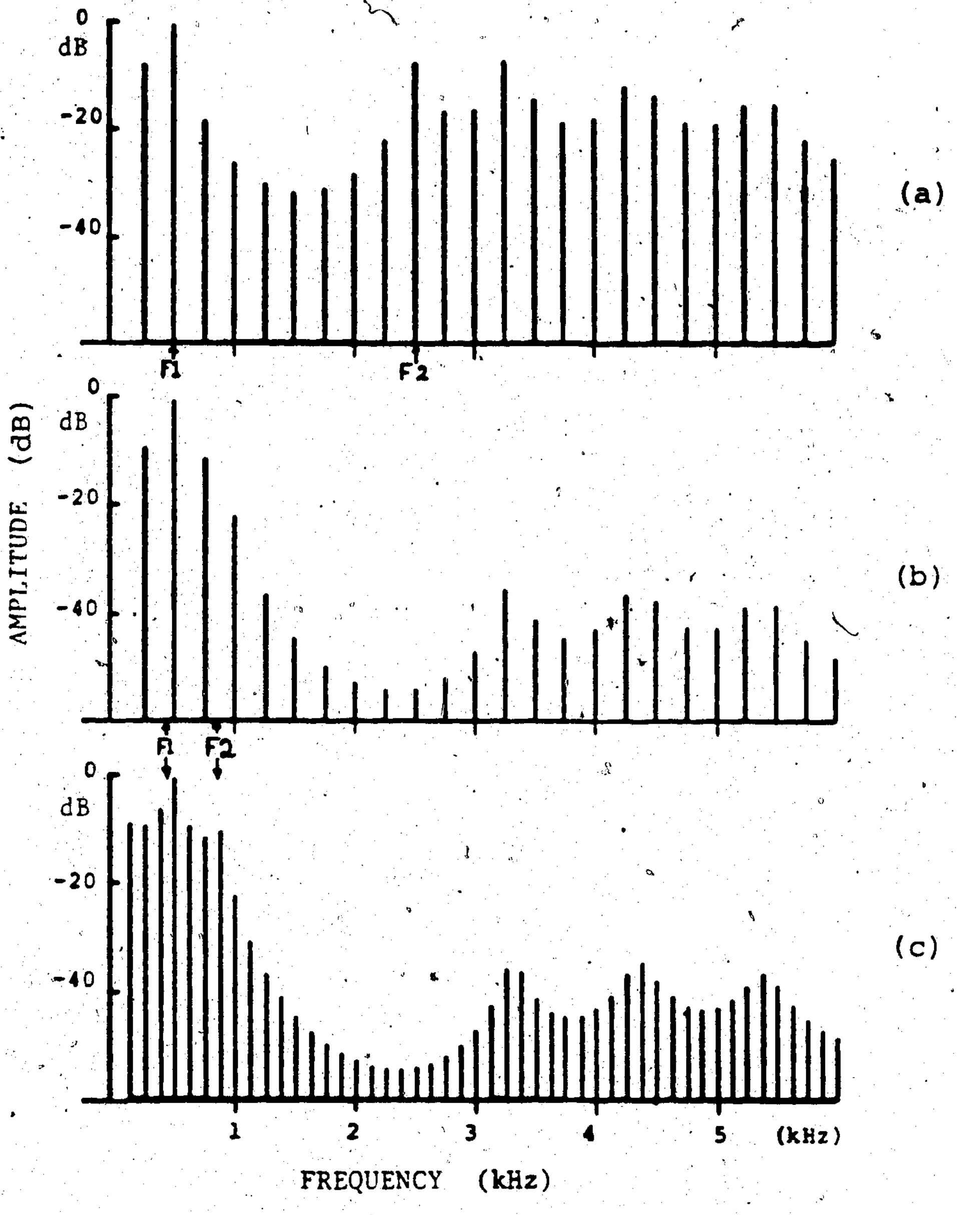


Figure 4.1 Effects of F1-F2 separation and fundamental frequency on vowel spectrum (see text)

front vowel in (a), however, there are several indicators that two formants are present in the low frequency region. The high frequency side of the peak representing the F1-F2 cluster is steeper, and the levels of all components in the high frequency region are reduced. The difference in spectral pattern between (a) and (b) might be described in terms of a shift in spectral balance. Listeners have no difficulty in separating these vowels: (a) is perceived as a front vowel similar to /ɛ/, while (b) and (c) are heard as a back vowel similar to /o/.

4.2 Measuring two formants in close proximity

Several techniques have been proposed for isolating two formant peaks in close proximity. If the locations of formant peaks are important in vowel perception, as suggested by the data on front vowels, it may be instructive to compare the methods by which these techniques estimate the locations of two formants in close proximity.

One approach to the measurement of F1 and F2 in close proximity involves spectral envelope estimation in conjunction with analysis procedures which enhance the degree of contrast in the spectrum. A peak-picking algorithm is applied to the 'sharpened' spectrum to determine the frequency locations of F1 and F2. This includes techniques based on the chirp Z-transform (Schafer and Rabiner, 1970) and the related method for identifying peaks in linear prediction spectra computed at radii inside the unit circle (McCandless, 1974). Another proposal involves a lateral inhibitory network based on a neural model for enhancing peaks (and edges) in the spectrum (Weston, 1974; Ujihara and Sakai, 1974; Karnickaya, Mushnikov, Slepokurova and Zhukov, 1975). Christensen, Strong and Palmer (1976) suggested that an examination of the negative second derivative of the log spectrum (which provides a measure of spectral curvature) might help to resolve 'merged' formant peaks.

A second approach compares the vowel spectrum with a series of internally generated spectral patterns based on the model of speech production. This approach, known as analysis-by-synthesis, incorporates fairly detailed information about the speech spectrum; however, it provides very robust estimates of the formant frequencies, and does not depend on the presence of distinct peaks in the spectrum. Bell, Fujisaki, Heinz, Stevens and House (1961) proposed an analysis-by-synthesis scheme which uses an iterative technique to adjust the parameters (formant frequencies and bandwidths) of a model of speech production to determine the best match (using a minimum mean squared error criterion) to the input speech spectrum. Since the criterion of fit is based on the entire spectrum, the proximity of F1 and F2 does not present a difficulty for this model. A related approach was suggested by Coker (1965). This procedure locates the absolute maximum in the spectrum and then by inverse filtering effectively removes the formant resonance (using an idealized formant shape based on the speech production model, positioned at the frequency of the maximum) from the input speech spectrum. Repeated application allows merged formants to be identified and measured fairly accurately.

Yet another approach, which does not involve the use of a model of speech production, estimates the frequency of a formant in terms of the spectral centre of gravity within a selected frequency region. Suzuki, Kadokawa and Nakata (1963) used the first moment of the entire vowel spectrum to separate F1 from the ranges of the higher formants. The first formant was then estimated as the first spectral moment, calculated over the low frequency region bounded by the global moment. This procedure did not give very accurate results for back vowels, where the proximity of F2 resulted in estimated F1 values which were biased toward higher frequencies.

These techniques suggest several possible mechanisms by which formant information could be extracted in back vowels. First, it is possible that the auditory projection of a vowel stimulus involves a sharpened version of the input spectrum, allowing a simple peak extraction procedure to determine the formant frequencies. A second possibility is that the formant locations are determined on the basis of an auditory spectral matching process which compares the input speech spectrum with internally generated spectral patterns. It is difficult to see how such a model could account for the effects of fundamental frequency on vowel perception, however, particularly if spectral fine structure is removed by an interpolation process.

It is possible, however, that other cues are present in the spectrum which help to specify the frequency locations of closely spaced formants. If so, we might expect to find that changes in F1 and F2 have independent effects on the perception of back vowels. This hypothesis depends on the assumption that F1 and F2 are independently resolvable by the auditory system.

4.3 Auditory resolution of F1 and F2 in back vowels

In this section the ability of the ear to discriminate or resolve formant changes with F1 and F2 in close proximity will be considered. Although very few psychophysical studies have investigated vowels with close spacing of F1 and F2, the available evidence suggests a high degree of sensitivity to formant changes in back vowels.

Discrimination studies. The data on formant detection suggest an increased sensitivity to small changes in formant frequency for back vowels with close spacing of F1 and F2. Flanagan (1955) measured difference limens (DLs) for the first and second formant frequencies in synthetic vowels. The test conditions for F1 involved only stimuli with wide separations of F1 and F2. However, one reference stimulus used to measure DLs for F2 was a back vowel, with F1=500 and F2=1000 Hz. An asymmetrical pattern of results was obtained: small decreases in F2 were more readily detected than small increases of the same magnitude. This asymmetry decreased with larger formant separations.

Formant frequency and formant amplitude are not independently manipulated by the vocal tract during vowel production (Stevens and House, 1961). The asymmetry in formant frequency DLs may thus be due in part to detection of formant amplitude changes. This is reflected in cascade formant synthesis in that changes in the frequencies of closely spaced formants result in simultaneous changes in formant amplitude.

Carlson and Granstrom (1976) measured DLs for amplitude and spectral slope in synthetic vowels with formant patterns typical of /i/, /ɑ/, /u/ and the neutral vowel, schwa. Two of these vowels (/u/ and /ɑ/) had closely spaced formants. These stimuli showed larger DLs for overall amplitude and spectral slope (both with and without a correction for differences in overall amplitude). Carlson and Granstrom suggested that detection thresholds for spectral slope changes might depend on changes in spectral levels near the formant peaks, rather than a sensitivity to spectral slope *per se*. These results indicate that although listeners may be more sensitive to formant *frequency* changes in vowels with closely spaced formants, their sensitivity to formant amplitude changes may be decreased because of the proximity of the formants.

Klatt (1982a) compared 'phonetic' and 'psychoacoustic' distance judgements for two reference vowels, one similar to /æ/ (with relatively wide separation of F1 and F2) and the other similar to /a/ (with closely spaced formants). There was no evidence for a difference between the two vowels in the apparent sensitivity to formant frequency and bandwidth changes; however, the introduction of a spectral notch between F1 and F2 had a larger effect on judgements involving the vowel /a/.

Kakusho, Hirato, Kato and Kobayashi (1971) investigated DLs for individual harmonic levels in steady state approximations to Japanese vowels /i e a o u/. For all of these vowels, DLs were smallest near formant peaks (including both F1 and F2 in the back vowels /a/ and /u/). DLs for an increment in harmonic amplitudes tended to follow a curve similar to the inverted spectrum envelope. DLs for harmonic attenuation were often so large as to be unmeasureable, particularly in the broad spectral valley between F1 and F2 in the front vowels; this was not the case for the back vowels. Evidently, changes in the levels of spectral components in the F1-F2 region of back vowels can be readily detected by listeners.

Masking studies. Masking studies of auditory frequency selectivity provide another source of evidence concerning the perception of F1 and F2 in close proximity. Houtgast (1974) obtained pulsation threshold measurements for synthetic vowel maskers /i a u/ with 32 harmonics of a 125 Hz fundamental frequency. Peaks were present in the resulting masking patterns for both F1 and F2 of /a/, and were in fact more prominent than those present in one-third octave band spectra. Tyler and Lindblom (1983) compared simultaneous and pulsation threshold masking patterns for vowel maskers. Although both procedures generally showed distinct peaks corresponding to F1 and F2, the pulsation threshold method resulted in sharper peaks in the formant regions, possibly reflecting lateral suppression.

Sachs and Zurek (1979) used contralateral probe measurements of vowel spectra in which subjects were asked to centre the tonal image by adjusting the amplitude and phase of a probe tone corresponding to each of the harmonic components of a test vowel presented to the opposite ear. The data for the vowels /i ɛ a u/ indicate clear resolution of the formant peaks, in fact resulting in enhanced peak-to-valley differences, relative to the physical spectra, in the low frequency region. These results provide further indications that F1 and F2 are independently audible, even in close proximity.

Dichotic studies. Another source of evidence concerning the auditory resolution of F1 and F2 comes from dichotic listening studies which deliver separate formants to each ear. If there is an interaction or fusion of spectral peaks in close proximity which occurs at a peripheral level of the auditory system, it may be expected that dichotic presentation will prevent this interaction and cause a loss or change in vowel quality. Several studies comparing dichotic and binaural performance have failed to show disruptive effects of dichotic presentation (Ladefoged & Broadbent, 1957; Carlson, Fant & Granstrom, 1975).

One possible exception was a study by Ainsworth (1981) who compared normal and short duration vowels using dichotic and binaural presentation modes. In the dichotic condition F1+F3 were presented to one ear, F2+F4 to the other. No changes in the identification matrices were observed for normal duration vowels. However, there were some differences between the front and back vowels at durations of 50 and 100 ms. Under dichotic presentation, there were frequent confusions of the vowels /I/ and /U/ which did not appear at normal durations. Ainsworth suggested that the F1+F3 pattern by itself might be perceptually similar to the F1+F2 pattern of another vowel. This situation is analogous to a "formant slotting error" in automatic formant measurement in which one formant peak is missed and its 'slot' is filled by a higher frequency formant (McCandless, 1974). Such errors are relatively uncommon in the identification of natural vowels (Assmann, Nearey & Hogan, 1982). Ainsworth attributed these errors to a failure to integrate the inputs from each ear at short durations, resulting in the disappearance of the auditory image from short term memory. The absence of a disruptive effect of dichotic presentation on normal duration vowels is inconsistent with the hypothesis that the perception of back vowels depends critically on fusion or integration of closely spaced formant peaks at a peripheral level of the auditory system.

Auditory nerve responses. There is little evidence in the series of studies by Sachs and Young (1979), Young and Sachs (1979) or Delgutte and Kiang (1984a) for qualitatively different behaviour in front and back vowels. Sachs and Young employed synthetic tokens of the vowels /i ɛ a/, with formant frequencies adjusted to coincide with harmonics[1] ,while Delgutte and Kiang used a larger sample of vowels whose formant frequencies were not constrained to coincide with harmonics. In the rate-place representation (discussed in section 2.3.8) separate peaks appeared corresponding to F1 and F2 for all vowels at low presentation levels. At higher intensities, because of rate suppression and two-tone suppression, the formant peaks were no longer discernible in the rate profiles. This occurred at lower levels for the back vowels than for the front vowels, possibly because of the proximity of F1 and F2.

[1] This tends to result in "peakier" spectra, with the most prominent harmonic in the region of the formant peak at least 5 dB higher than either of its neighbours. (See Figure 3.1)

At high levels, however, all vowels showed a loss of contrast and a disappearance of the formant peaks. Temporal responses, measured as the average magnitude of each component of the Fourier transform of the period histograms, retained peaks corresponding to F1 and F2 and low order harmonics even at the highest presentation levels. These findings suggest that components in the F1 and F2 regions of back vowels may be resolved at the level of the auditory nerve.

4.4 Determinants of phonetic quality in back vowels

The evidence reviewed to this point suggests that F1 and F2 are independently resolved by the auditory system and that small changes in formant frequency can be readily detected. This conclusion, however, does not imply that F1 and F2 have independent phonetic consequences. A number of recent studies have challenged the hypothesis that the phonetic quality of back vowels is based on an independent contribution of F1 and F2.

Formant centre of gravity (FCOG) hypothesis

Delattre, Liberman, Cooper and Gerstman (1952:p. 201) advanced the hypothesis that

> "the ear effectively averages two vowel
> formants which are close together, receiving
> from these two formants an impression which
> is highly similar to that which would be

> heard from one formant placed at a position somewhere intermediate between them."

This statement was based on the observation that a single formant could substitute for both F1 and F2 in back vowels with little apparent change in vowel quality; and on the qualitatively different effects of a reduction in formant amplitude in front and back vowels.

4.4.2 Single formant approximations to back vowels

Delattre et al. (1952) synthesized a series of 235 two formant vowels covering the F1-F2 plane, to obtain best approximations to each of the 16 cardinal vowels. Fundamental frequency was fixed at 120 Hz, and formant locations were adjusted by changing the levels of three prominent harmonics in the desired formant region. Listening tests showed that selected combinations of F1 and F2 values could be obtained to produce all of the desired vowel qualities. Delattre et al. then attempted to replace each of the peripheral vowels in the F1-F2 plane with a single formant, whose frequency was intermediate between F1 and F2. It was found that the cardinal back vowel series could be matched with a single, relatively low frequency formant placed between F1 and F2; closer to F1 of the original for /u/ and /o/, and progressively closer to the midpoint of F1 and F2 for more open vowels. The remaining (non-back) vowels could not be satisfactorily approximated by a single formant, with the exception of /i/ which was matched with a high frequency formant at or above F2'.

4.4.3 Effects of formant amplitude

Delattre, Liberman, Cooper and Gerstman (1952) also investigated the effects of reducing the amplitude of the first and second formants in small steps. Their informal observations on the effects of these changes suggested a qualitative difference between back vowels with F1 close to F2, and front or central vowels with wide separations of F1 and F2. Non-back vowels did not show a gradual shift in quality as formant amplitudes were lowered. A decrease in the level of the first formant resulted in a "dulling" effect on the quality of the vowel, followed by a complete loss of vowel colour. Lowering the second formant level also caused a "dulling" of vowel quality; with considerable attenuation there was a shift in phonetic quality to a back vowel. These results suggested to Delattre et al. that the second formant was no longer audible, and that the stimuli with attenuation of F2 were effectively single formant vowels for listeners.

A different pattern of results was found in the back vowels. A reduction in first formant amplitude for the

'The vowel /i/ can be separated from all other English vowels on the basis of its extreme F2 value; it is rarely confused with other vowels in identification experiments (e.g., Assmann, Nearey and Hogan, 1982). /i/ may have a somewhat special status in vowel perception because of its extreme formant values (Nearey, 1977).

vowels /u/, /o/ and /ɑ/, for example, resulted in a gradual shift in the direction of a lower vowel quality. Reduction in the amplitude of the second formant in back vowels generally led to a higher vowel quality.

Delattre et al. suggested that these effects could be explained by a centre of gravity or formant averaging mechanism. They proposed that phonetic quality in back vowels depends on the location of the formant cluster rather than on individual formant locations. However, they did not present detailed results in support of this hypothesis; their report was largely anecdotal. Furthermore, there were some aspects of their reported findings which are not consistent with this hypothesis. For example, they also found shifts as a function of formant amplitude for /æ/, which has a relatively wide separation of F1 and F2, and is not a back vowel.

Miller (1953) also reported shifts in vowel quality as a function of second formant amplitude in back vowels. He noted that the amplitude of the second formant was considerably lower than that of the first in naturally spoken high back vowels /u/ and /o/. More open vowels such as /ɑ/ had nearly equal formant amplitude levels. The best single formant approximations for /u/ and /o/ had frequencies close to their typical F1 values, while for the low back vowels, the best approximation was close to the midpoint or centre of gravity of F1 and F2.

Bernstein (1981) obtained auditory dissimilarity judgements of vowel stimuli and attempted to model the data in terms of an additive combination of scaled formant frequencies. It was not possible to obtain satisfactory results using a metric based solely on formant frequency values. However, the inclusion of formant amplitudes led to a marked improvement in the predictions. Although listeners may not have responded in a 'phonetic mode', the results suggest that formant amplitude may play a role in the auditory coding of vowel sounds.

The studies reviewed above suggest that formant amplitude may have an important influence on back vowel quality when F1 and F2 are in close proximity. However, there is one study which suggests that formant amplitude effects are minimal, even in back vowels. Ainsworth and Millar (1972) reported a vowel identification experiment in which F1 and F2 were varied in small steps to create a continuum spanning typical values for all of the English vowels. A set of standard vowels was also produced as exemplars of each vowel category; their F1 and F2 values were based on measured formant means from natural speech data. The level of the second formant was reduced in 5 steps: 0, -14, -28, -42 and -∞ dB, relative to that of F1. In each listening session only one of the five amplitude conditions was presented, i.e., sessions were 'blocked' with respect to formant amplitude changes[1].

----

[1] Range effects on the identification of a continuum are

Ainsworth and Millar (1972) measured vowel centres (centroids of the identification functions) for F1 and F2 in each amplitude condition. They found that F1 centres were nearly constant across all conditions. Listeners did not appear to compensate for attenuation of the second formant by a shift in response curves toward higher F1 values, as predicted by the formant centr of gravity hypothesis of Delattre, et al. (1952). Vowel centres for F2, however, showed fairly systematic changes as the amplitude of the second formant was reduced. For low F2 values, F2 centres increased slightly as the level of the second formant was reduced. Some of the front vowels (with high F2 values) showed a small decrease in F2 centres.

One possible explanation for these results is a general tendency for F2 centres to shift with decreasing amplitude toward the middle of the F2 range, around 1.8 kHz. Ainsworth and Millar stated that most listeners heard only back vowels with more than 28 dB attenuation of the second formant. They suggested that this value is close to the F2 detection threshold. F2 centres did not appear to depend

----

[1](cont'd) occasionally observed when speech stimuli are presented in a 'blocked' mode (Brady & Darwin, 1978). Differences between blocked and randomized conditions have been reported in vowel identification experiments for the effects of changes in speaker identity (Assmann, Nearey & Hogan, 1982); high and low pass filtering (Lindquist & Pauli, 1968): and changes in source characteristics (Gauffin & Sundberg, 1982). Perceptual compensation may take place in a blocked presentation mode, reducing the effects of the experimental variable under investigation. The blocked mode of presentation may be more representative of natural speech, since formant amplitudes do not vary randomly over the course of an utterance.

substantially on second formant amplitude provided that the amplitude difference between the first and second formants was smaller than 30 dB, a value which appears to represent the detection threshold for the second formant (see also Nearey & Levitt, 1974).

These results leave a number of unresolved questions. Ainsworth and Millar (1972) argue in favour of a peak-picking mechanism in vowel perception, on the strength of the evidence that vowel identification is unaffected by formant level over a broad range. The gradual shifts in vowel quality reported by Delattre et al. (1952) and Miller (1953) are not evident in the identification data presented by Ainsworth and Millar. It is possible that their measure of vowel identification (vowel centres) may not be sensitive enough to detect the effects reported by Delattre et al. Although both studies used identification responses, the Ainsworth and Millar study presented each amplitude condition in a blocked presentation mode which may have reduced the influence of formant amplitude.

4.4.4 Single formant matching studies

The formant averaging or centre of gravity hypothesis advanced by Delattre et al. (1952) suggests that a single formant can substitute for both F1 and F2 in back vowels. It should therefore be possible to adjust the frequency of a single formant to match a two formant or multiformant back vowel in phonetic quality.

Cohen (1974) presented data from 3 listeners on single formant matching of two formant reference stimuli. Parallel formant synthesis was used, with noise excitation of the filters. The two formants were equal in amplitude. F1 ranged from 200 to 1000 Hz; F2 was positioned near F1, with a 100 Hz interval in most stimuli. The best match always occurred in the region spanning F1 and F2, with about half of the matches coinciding with either F1 or F2; the remainder occurred in the region between F1 and F2, with no obvious pattern to account for this variability. It was reported that listeners could not obtain satisfactory matches for stimuli with F2/F1>1.5.

The high number of matches to F1 or F2 might have been the result of a perceived pitch corresponding to the peaks in the filtered noise spectrum, according to Cohen (1974). Steeply filtered noise bursts may give rise to the perception of an "edge pitch" (von Bekesy, 1960; Small & Daniloff, 1967; Rakowski, 1968). The variability in responses may have been the result of different matching criteria, involving both pitch and vowel quality differences. Cohen's results therefore do not clearly support a formant centre of gravity hypothesis.

A detailed investigation of the formant centre of gravity hypothesis was conducted by Bedrov, Chistovich and Sheikin (1978), Chistovich and Lublinskaya (1979), and

Chistovich, Sheikin, and Lublinskaya (1979). Bedrov et al. reported an experiment in which the frequency of a single formant was adjusted to find the best match to a two-formant reference. The amplitude of the first formant of the reference was either 5 dB below or 5 dB above that of the second formant; F2 was fixed at F1+350 Hz. The frequency of the best matching single formant, F*, generally occurred between F1 and F2, closer to the more intense formant. No indications of bimodal matching were found in the data from the two listeners. When the listeners matched for pitch, on the other hand, F* was placed near F1 in all cases. Bedrov et al. concluded that matching depends on the alignment of the spectral centres of gravity in reference and matching stimuli.

Chistovich and Lublinskaya (1979) reported a matching experiment in which listeners adjusted the amplitude ratio A2/A1 in a two formant vowel to match a single formant whose frequency was set equal to the midpoint of F1 and F2. A continuous relationship between F*, the frequency of the best matching single formant, and the formant amplitude ratio was reported for small separations of F1 and F2. As the separation between the formants was increased there was a point beyond which satisfactory matches in vowel quality could not be obtained by amplitude ratio adjustments. This critical separation of F1 and F2 was estimated to be between 3 and 3.5 Bark.

In a subsequent experiment the effects of formant amplitude were investigated using single formant matches to two formant vowels. The formant amplitude ratio A2/A1 ranged from -20 dB to +30 dB in 10 dB steps. A nearly continuous shift was observed for vowels with a small separation of F1 and F2. Unsystematic results were obtained for reference vowels with large separations of F1 and F2. One of the two listeners showed a bimodal pattern of matching, with F* positioned near F1 for -20 and -10 dB, and an abrupt shift to F2 with higher ratios. The other listener showed a continuous shift but with increased variability near the middle of the range, suggesting a degree of uncertainty in the decision process. Chistovich and Lublinskaya argued that the increase in variability and the discontinuity in matching as a function of formant level was evidence that the phonetic quality of vowels with a large separation of F1 and F2 could not be adequately reproduced by adjusting the frequency of a single formant, and hence a centre of gravity characterization was not appropriate for such stimuli. They claimed that vowel quality in stimuli with greater than critical separation of F1 and F2 is largely independent of formant amplitude, provided that neither formant is attenuated below threshold levels'.

'In spite of the increase in variability near the centre of the range, however, the matching data from the second subject shows a systematic dependence on formant amplitude which is not accounted for by their hypothesis. In later sections an alternative possibility will be considered. Formant amplitude may provide a secondary source of vowel

Some of the effects of formant amplitude changes observed by Delattre et al. (1952), Miller (1953) and Ainsworth and Millar (1972) are suggestive of a formant threshold effect. For example, considerable attenuation of F2 in their front vowel stimuli caused an abrupt shift to the perception of a back vowel of corresponding vowel height. On the basis of matching experiments using reference stimuli with large separations of F1 and F2, Chistovich and Lublinskaya (1979) estimated that F1 and F2 could be detected over nearly a 50 dB range of formant amplitude ratios.

Chistovich et al. summarized their main findings as follows:

(1) in stimuli with <u>more</u> than 3 to 3.5 Bark separation of F1 and F2, vowel quality is independent of changes in the amplitude ratio of the two formants, over a wide range bounded by the detection thresholds for each formant.

(2) in stimuli with <u>less</u> than the critical separation of F1 and F2, vowel quality is shifted by changes in the formant amplitude ratio. A gradual shift in vowel quality occurs as the formant ratio A2/A1 is altered, until one formant or the other is longer detectable; the stimulus is then effectively a single formant vowel.

Beddor (1984) and Beddor and Hawkins (1984) presented the results of a single formant matching study which

[1] (cont'd) information whose role is listener- and and task-dependent.

replicated the finding of Chistovich et al., indicating that the best single formant match to a back vowel is positioned between the F1 and F2 values. However, the frequency of the best matching single formant was closer to F1 than predicted on the basis of centroid calculations spanning the F1-F2 region. In section 4.5.4 several alternative models for predicting centre of gravity effects will be explored.

Critical distance between F1 and F2

Fant (1983) proposed that a 3-3.5 Bark separation of F1 and F2 might account for the universal class of back vowels. He suggested that the perceptual differences between vowels with less than and greater than this critical formant separation might correlate directly with the phonological distinction between back and non-back vowels.

Figure 4.2 is a graph of the ten vowels of western Canadian English in the F1-F2 plane (in Bark units; c.f. Schroeder, Atal & Hall, 1979). The plotted values are average values from 5 female and 5 males speakers obtained in a previous study (Assmann 1979). The solid and dashed lines represent F1 and F2 separations of 3 and 3.5 Bark, respectively.

It can be seen that the 3 Bark line does not clearly separate the vowel set into two phonological classes, back versus non-back vowels. A number of tokens of the vowels /o/ and /ʌ/ occur on both sides of the F2-F1=3 Bark line.

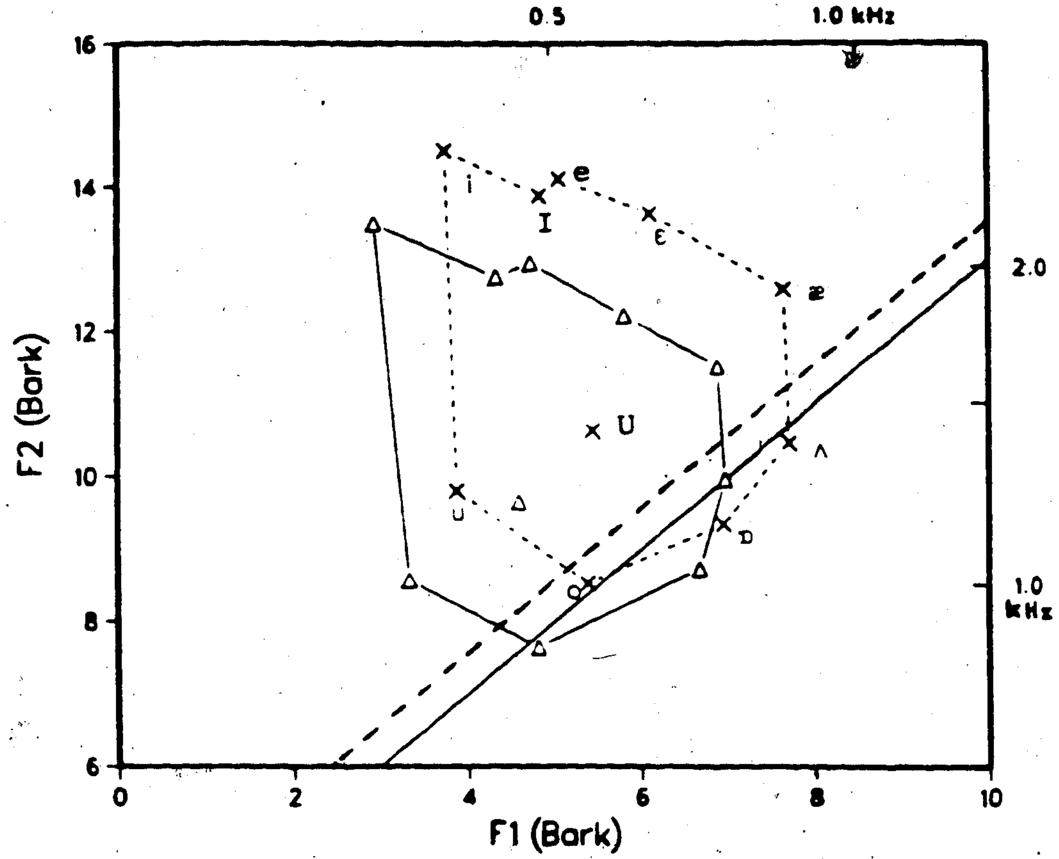


Figure 4.2 Formant frequency means for Western Canadian English. Male averages are indicated by triangles, female averages by crosses. Solid line indicates F2-F1=3 Bark; dashed line indicates F2-F1=3.5 Bark

The vowels /u/ and /U/ have relatively high F2 values in this dialect and often show greater than 3 Bark separation of F1 and F2. Similar violations of Fant's hypothesis were reported for American English by Syrdal (1982).

It seems unlikely that different perceptual processes are involved when listeners attend to different tokens of the same vowel, uttered by speakers with slightly different vocal tracts sizes. There does not appear to be a difference in phonetic quality or timbre between tokens of the same vowel which happen to occur on different sides of the F2-F1=3.5 Bark line. The phonetic relevance of the notion of a critical 3-3.5 Bark distance should therefore be examined very closely.

4.5 Experiment 5: Effects of formant amplitude on matching in back vowels

4.5.1 Introduction

The evidence for centre of gravity effects in back vowels is based almost exclusively on experiments in which single formant stimuli are matched to multiformant stimuli with altered amplitudes of the first and second formants. Single formant stimuli are different in several respects from multiformant stimuli, as discussed in section 4.1. They share with low back vowels the characteristic that most of the energy is located in the low frequency region of the spectrum, exhibiting only a single peak, but they differ in other characteristics, such as the amount of energy in the high frequency region, the slope of the spectrum on the high frequency side of the peak, overall amplitude, etc. These differences make it important to verify the results with matching experiments in which both matching and reference stimuli are multiformant stimuli. A second reason for investigating the effects of formant amplitude on the matching of multiformant vowels is that at least two previous studies have indicated only minor effects of formant amplitude on phonetic dissimilarity judgements (Klatt, 1982a) and back vowel identification (Ainsworth and Millar, 1972).

If the spectral centre of gravity is the exclusive determinant of vowel quality when F1 and F2 are close together, it should be possible to manipulate formant amplitudes to produce multiformant stimulus pairs which differ in their respective F1 and F2 values but have the same centre of gravity, or vice versa. It is important to determine whether such pairs are judged as more similar (in a matching task) than pairs with identical formant frequencies but different formant amplitudes.

A matching task was used to evaluate several predictions of the formant centre of gravity hypothesis concerning the perception of multiformant vowel stimuli. One aspect of the FCOG hypothesis discussed by Chistovich (1985) is that formant frequency and formant amplitude trade off against each other when the formants are in close proximity. An increase or decrease in the frequency or amplitude of either formant will result in a lowering or raising of the centre of gravity which in turn determines the identity of the vowel. This hypothesis was tested using multiformant matching and reference stimuli.

Fixing the position of F2 in back vowel matching. One problem which arises in the design of back vowel matching experiments concerns the constraints on formant movement when F1 and F2 are in close proximity (assuming that formants are not allowed to cross in frequency). If F2 is set to a constant value, F1 is constrained to vary over only a small frequency range. Similarly, if F1 is fixed, F2 will be adjustable only over a narrow frequency region. One possible solution is to allow listeners to adjust each formant separately. This task may be somewhat more difficult for the listener and requires a longer time period to arrive at a satisfactory match.

A second alternative (adopted in the present experiment) is to ask listeners to covary F1 and F2 in the matching stimuli. A fixed separation of F1 and F2 may be used in both matching and reference stimuli. Listeners adjust F1 and F2 simultaneously to compensate for changes in the amplitude ratio of F1 and F2 of the reference. In the present experiment it was decided to use a 250 Hz separation of F1 and F2, a frequency distance less than 3 Bark at all positions along the matching continuum[1].

A schematic diagram illustrating this paradigm is shown in Figure 4.3. Part (a) represents the single formant matching paradigm used by Chistovich, Sheikin and Lublinskaya (1979). In part (b) both matching and reference stimuli are two formant vowels, and listeners adjust F1 and F2 together to match changes in the amplitude ratio of the formants in the reference.

[1] 250 Hz represents approximately 2.4 Bark separation of F1 and F2 at the low frequency endpoint of the stimulus continuum, (F1=175 Hz) and approximately 1.4 Bark at at the high frequency endpoint (F1=950 Hz) (Schroeder, Atal & Hall, 1979).

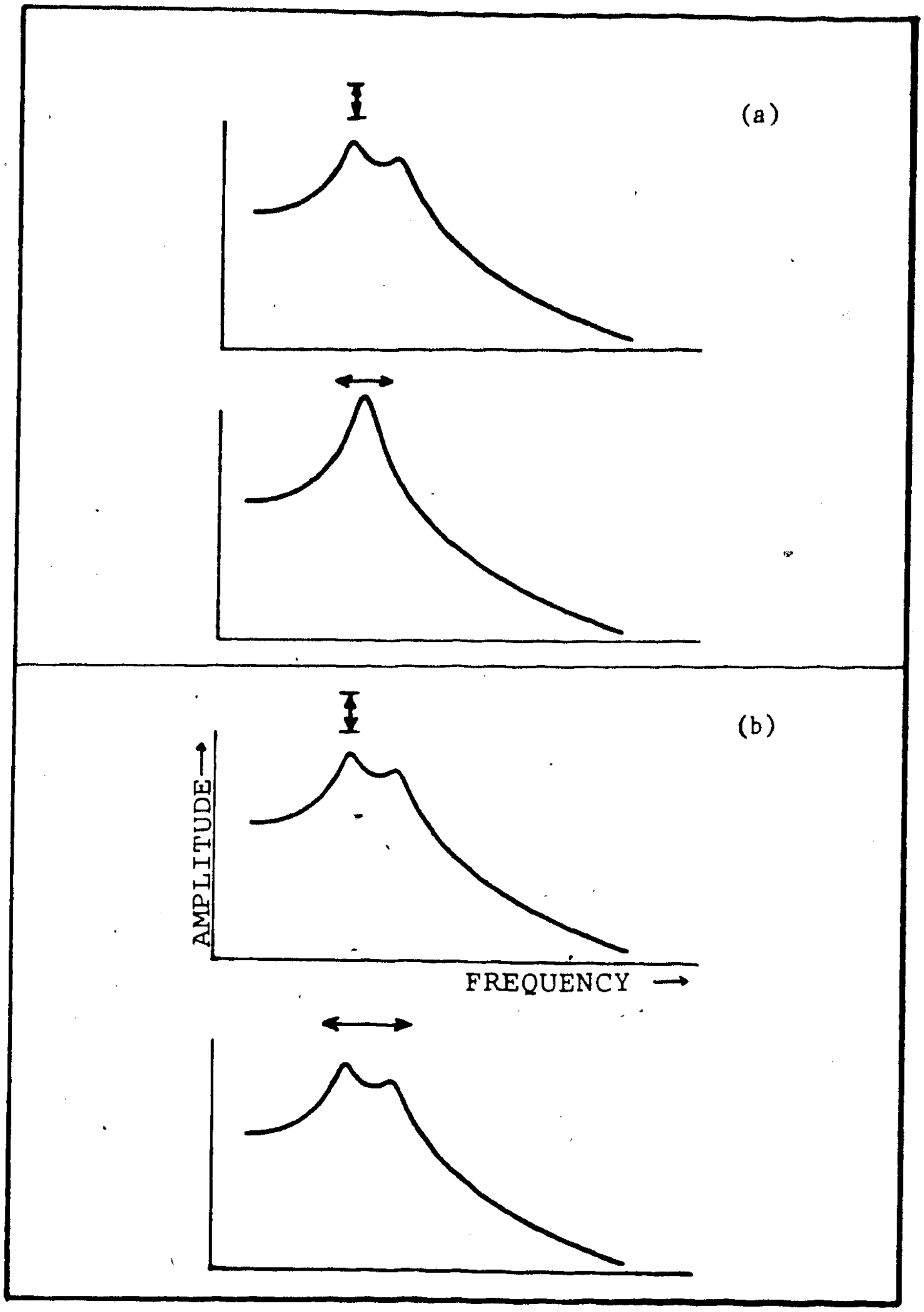


Figure 4.3 Schematic amplitude spectra for vowel matching experiments. (a) 2 formant reference, 1 formant matching stimulus (b) 2 formant reference, 2 formant matching stimulus

The decision to covary F1 and F2 can be justified on two grounds. First, there is a correlation of F1 and F2 in measurements of natural back vowels: higher F1 values are associated with higher F2 values (see Figure 4.2). Second, by varying F1 and F2 together listeners are effectively controlling the arithmetic mean of the formant frequencies. If matching depends on a centre of gravity or formant averaging process, a change in the formant amplitude ratio should be phonetically equivalent to a shift along the (F1+F2)/2 dimension. A shift in both F1 and F2 will affect the centre of gravity in the low frequency region in a similar manner to a change in the formant amplitude ratio. Formant amplitude changes are expected to produce changes in matching even if these changes result in misalignment of the formant peak locations in the matching and reference stimuli.

Formant amplitude manipulation. A second problem concerns the nature of amplitude manipulation in the reference stimuli. When formant amplitudes are manipulated using parallel formant synthesis, there are correlated changes throughout the spectrum (Klatt, 1980; Holmes, 1983). A change in second formant amplitude, for example, may alter the levels of prominent harmonics in the F1 region, particularly if the formants are very close together in frequency. Experiments one to three have indicated that vowel height may be altered by changes in the levels of prominent F1 harmonics. It is important to keep the levels of these components fixed while manipulating the amplitudes of the peaks.

In the present experiment, amplitude was manipulated by directly altering the levels of individual harmonic components using additive harmonic synthesis. Components in the region of F2 and higher formants were either raised or lowered relative to spectrum levels in the unmodified or control condition.

According to the FCOG hypothesis the best match will be the one closest to the reference in terms of the centre of gravity of F1 and F2. The centre of gravity in the reference stimuli will increase monotonically as the amplitude level of the second formant is raised, and decrease when it is lowered. A limit may be reached when one formant is substantially higher in level than the other; the vowel is then effectively comprised of a single peak in the low frequency region. Formant amplitude changes are expected to result in phonetic changes which are equivalent to altering the frequencies of both formants. Raising or lowering the second formant amplitude should lead to parallel changes in the F1-F2 values of the preferred match.

Alternatively, the two spectral peaks may have some independent effects on vowel quality. One possibility is that listeners disregard peak amplitude changes entirely, selecting the same matching stimulus in each formant amplitude condition. We will refer to this as the 'no change' hypothesis. A second possibility is that formant amplitudes may affect vowel quality, but as a supplemental or secondary cue. Some conflict between the two cues may be present, and the matching data may show an increase in variability or multimodal matching patterns. A third possibility is that formant amplitudes may have minimal effects until one of the two formant peaks is no longer audible. Masking of one formant by the other will then lead to an abrupt change in matching performance.

4.5.2 Method

Stimulus materials

The stimuli were generated using digital additive harmonic sythesis. Amplitude spectra for the matching stimuli and reference stimuli were calculated using a cascade formant synthesis procedure as in Experiment 1. The bandwidths of the formants and the higher formant frequencies were identical with those indicated in Table 3.1. A constant fundamental frequency of 125 Hz was used for all stimuli.. Additional parameters (higher pole correction, glottal source and radiation characteristic) were as described in chapter three.

Reference stimuli. Three baseline stimuli were produced, with F1 = 350, 450 and 550 Hz. F2 was fixed in all stimuli at F1+250 Hz. The amplitudes of harmonics were altered to produce five conditions for each of the 3 baseline stimuli.

The weighting functions used to produce each of the amplitude conditions are shown in Figure 4.4. For each stimulus, three constraints were defined:

(1) components lower than or equal to the frequency value F1 + 125 Hz were unmodified. This condition ensures that the two components of largest amplitude in the vicinity of F1 will not be altered.

(2) components above F2 were altered by -20, -10, 0, +10, or 20 dB.

(3) components in the intermediate region were linearly interpolated in amplitude to avoid edges or discontinuities in the spectral envelope. The amplitude of each harmonic A_i was multiplied by a scale factor S_i, computed as follows:

$$S_i = \begin{cases} 1, & f_i < fc \\ ((f_i - fc)(M-1)/(F2-fc))+1, & fc < f_i < F2 \\ M, & f_i > F2 \end{cases}$$

where f_i is the frequency in Hz of the ith harmonic component;

$fc = F1 + 125$ Hz

$M = 10^{(d/20)}$

d represents the change in amplitude level (-20, -10, 0, +10, or +20 dB).

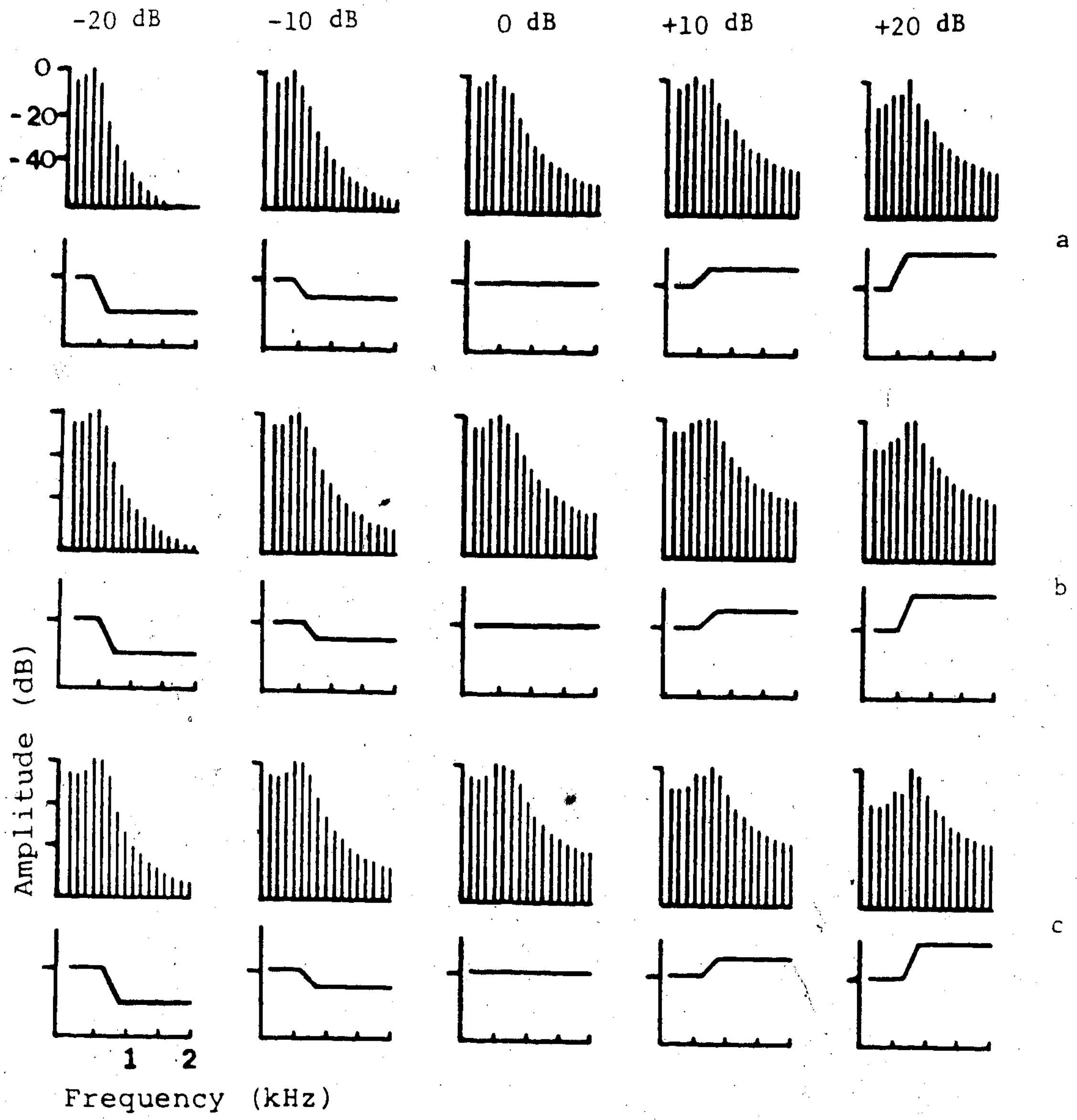


Figure 4.4 Experiment 5. Weighting functions and line spectra for reference stimuli.
(a) F1=350 Hz, F2=600 Hz
(b) F1=450 Hz, F2=700 Hz
(c) F1=550 Hz, F2=800 Hz

Matching stimuli. A continuum of vowels was constructed with F1 ranging from 175 to 950 Hz in steps of 25 Hz. F2 was fixed at 250 Hz above F1 as in the reference stimuli. Figure 4.5 indicates the F1 and F2 values, in Bark units, for each of the matching and reference stimuli. All stimuli had less than 3 Bark separation of F1 and F2.

Digital additive harmonic synthesis was used to produce the stimuli as described in section 3.3.2.

Subjects

Five phonetically trained listeners completed the task. All had participated in one or more of the experiments descibed in Chapter 3. All of the listeners had normal hearing, and all were native speakers of Canadian English.

Procedure

Stimulus presentation, instructions to the subjects, and the recording of responses followed the procedure described in Experiment 1, section 3.3.2.

4.5.3 Results and discussion

The data from each listener was summarized in terms of the median F1 value of five matches. Means and standard deviations of these medians are shown in Figure 4.6. The lowest curve represents reference stimuli with a nominal F1

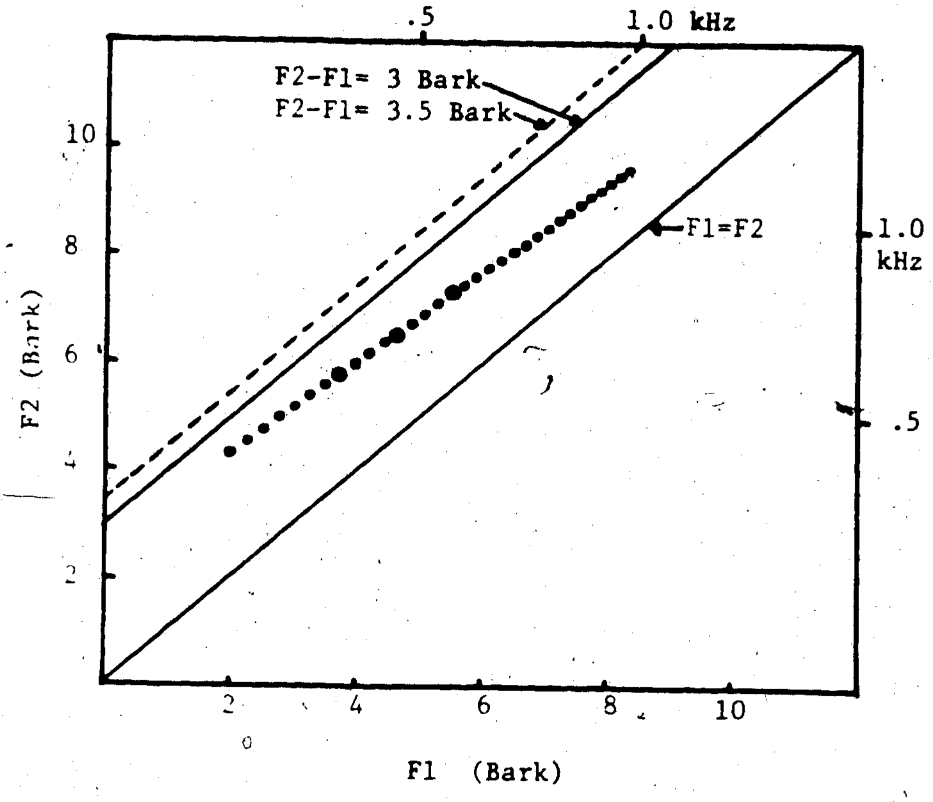


Figure 4.5 Experiment 5. Formant frequencies of matching stimuli (●) and reference stimuli (o)

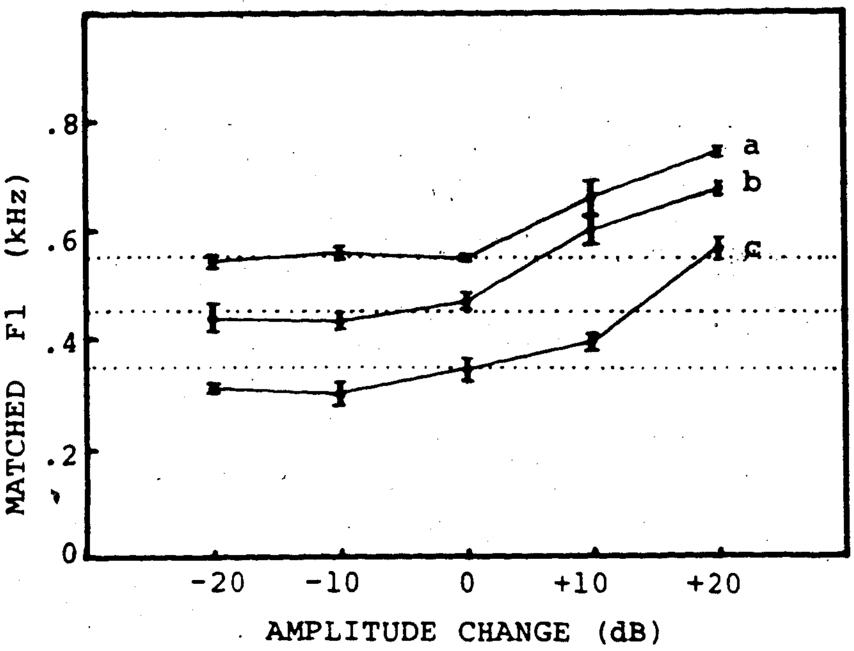


Figure 4.6 Means and standard deviations of median matched F1 values for experiment five. Dotted lines indicate nominal F1 values.
(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

and F2 of 350 and 600 Hz; the middle curve, 450 and 700 Hz; and the top curve, 550 and 800 Hz. The dotted lines at 350, 450 and 550 Hz indicate the frequency locations of the nominal F1 values for stimuli associated with each of the 3 curves. These lines indicate the expected F1 matches under the 'no change' hypothesis of formant amplitude effects. There are substantial deviations from these lines in the two highest amplitude conditions, +10 and +20 dB, indicating that a change in higher formant amplitudes does affect matching.

The mean match in the 0 dB condition is near the nominal F1 values in all three cases, as expected. In the +20 dB condition, the mean matched F1 has shifted to a higher frequency, just below F2 of the reference. Matches in the +10 dB condition are intermediate between these two, suggesting a continuous change as a function of formant amplitude. In the opposite direction, comparing the 0, -10 and -20 dB conditions, amplitude does not appear to have an effect on matching performance.

An analysis of variance was performed on the median matched F1 data. A significant main effect was found for the factors AMPLITUDE (F(4,16)=87.05; p<.0001) and FORMANT FREQUENCY (F(2,8)=152.22; p<.0001). The interaction of FORMANT FREQUENCY and AMPLITUDE was also significant (F(8,32)=3.04; p<.02). Each of the conditions with modified amplitude were compared with the control or 0 dB condition using Dunnett's test. Means and significance levels for each of the modified amplitude conditions are shown in Table 4.1.

None of the stimuli with attenuation of higher formants was significantly different from the control, although there appears to be a small trend toward lower F1 matches with attenuation of F2 and higher formants in the 2 lowest curves of Figure 4.6. The absence of a significant shift to lower F1 matches with a decrease in the amplitudes of higher harmonics is inconsistent with the predictions of the formant centre of gravity hypothesis. Raising the levels of higher formants did lead to significant upward shifts in matched F1, except for the +10 dB stimulus with the lowest formant values (F1=350, F2=600Hz).

Histograms of the individual matches from each subject are shown in Figure 4.7. A clearly multimodal matching pattern can be seen for the reference stimuli with F1=450 and 550 Hz in the +10 dB condition. This result is difficult to reconcile with the formant centre of gravity hypothesis which predicts a continuous shift in vowel quality as the amplitude of the second formant is raised relative to the amplitude of the first.

This finding indicates an important difference between single formant and multiformant matching. Chistovich, Sheikin and Lublinskaya (1979) reported that single formant matching was a continuous function of amplitude over a range

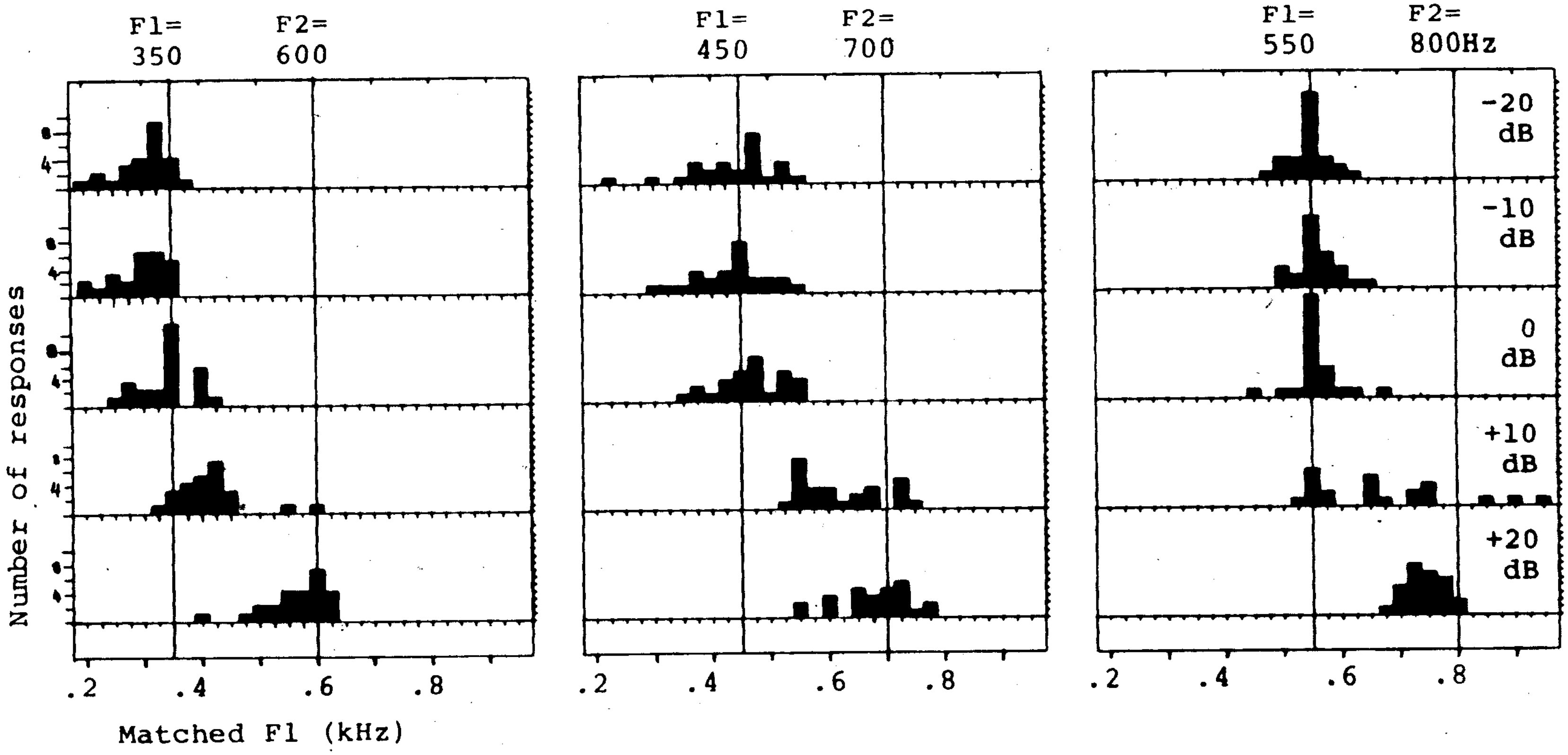


Figure 4.7 Histograms of matching data for experiment five. Solid lines indicate F1 and F2 values of the reference stimuli

	-20	-10	0	-10	+20 dB
F1=550	545	560	550	660**	745**
F1=450	440	435	470	600**	675**
F1=350	310	300	345	395	565**

**p<.01

Table 4.1. Mean matched F1 values and significance levels for Dunnett's test comparing modified amplitude conditions with the control or 0 dB condition

of approximately 40 dB, provided that the F2-F1 distance in the reference was less than 3 Bark. Their reference stimuli (Figure 4, p. 148) were altered in 10 dB steps from a 0 dB baseline condition with equal amplitudes of F1 and F2. In the present experiment, 0 dB was used to describe the relative formant amplitudes given by the cascade formant synthesis model without any filtering. This procedural difference is unlikely to be responsible for the discrepancy in results, since the amplitudes of the first and second formants in the 0 dB condition of the present experiment differed by less than 10 dB. The formant amplitude ranges in the present study and that of Chistovich et al. overlapped to a large degree. It seems more likely, therefore, that the difference in results is due to the use of multiformant rather than single formant matching stimuli in the present experiment.

The absence of significant changes in matching performance with attenuation of F2 and higher formants suggests that it may be important to align both F1 and F2 when both formants are present in the matching stimuli; amplitude differences are largely ignored. Yet substantial shifts do occur in the +10 dB and +20 dB conditions. Why should vowel quality be affected more by an increment than by attenuation of the higher formant amplitudes?

One possibility is that the first formant in the +20 dB condition is masked by energy at higher frequencies. Masking of F1 by higher formants may also have affected matching performance in the +10 dB condition, where extremely variable matching behaviour was observed. The reference stimuli with +10 or +20 dB boost of higher frequency components might then be perceived as having only one low frequency formant. The result is that F1 of the matching stimulus is often aligned by listeners with F2 (or with a frequency value just below F2) in the reference stimulus.

4.5.4 Predictions of the matching data

Several procedures for predicting the matching data were investigated. The first stage of each analysis involved the computation of spectral measures for each of the reference and matching stimuli. For each reference stimulus, the closest predicted match along the formant continuum was determined, using spectral measurements in conjunction with the distance model described in section 3.3.4. The degree of fit of the predictions was summarized in terms of the RMS distance between observed median matched F1 values, using the procedure described in Experiment 1'. The first set of analyses was based on measures of the centre of gravity of the F1-F2 cluster. The frequency location of the spectral centre of gravity was estimated using procedures based on the results of Chistovich and Lublinskaya (1979). Estimated centre of gravity locations ranged between F1 and F2, and shifted continuously as a function of changes in the amplitude of either formant.

Centre frequency and centroid measures. Several explicit formulations of a model outlined by Chistovich, Sheikin and Lublinskaya (1979) were first considered. This model involved the summation of energy over a series of highly overlapping 3 Bark regions of the spectrum. The centre frequency of the 3 Bark channel with the maximum summed output in the F1-F2 region was used as a measure of the centre of gravity.

'It should be pointed out that the median matched F1 coincided in most cases with the largest mode in the matching histograms. A multimodal pattern of matching was present for two stimuli in the +10 dB condition (F1=450 and 550 Hz). In such cases, the median may not adequately characterize the matching response profiles. No attempt was made to predict the entire vector of responses (the proportion of matches drawn from each position on the continuum, rather than just the median response). Further refinements of the model might attempt to account for the major modes of the matching response profiles in cases of multimodal matching.

One possible shortcoming of this procedure lies in its lack of sensitivity to changes in the distribution of energy within the 3 Bark range. A second procedure, which is sensitive to such changes, used the first moment or centroid (calculated over the frequency region spanned by the 3 Bark channel with maximum output) as a measure of the centre of gravity[1].

An additional refinement of these centre of gravity measures may result from a consideration of the transformation of the spectrum in the auditory system. The question of a spectral representation compatible with psychophysical data on frequency resolution and loudness has been addressed in earlier chapters. In the present analyses centre of gravity measures were obtained using either the power spectrum or transformations of the spectrum using the excitation and loudness pattern models described in section 3.3.4.

Three pairs of analyses were conducted. The first member of the pair employed the centre frequency measure, the second employed the centroid to estimate the centre of gravity of the F1-F2 cluster. The first pair (analyses 1 and 2 in Table 4.2) used measurements derived from the power spectrum. The second pair (analyses 3 and 4) used

[1] Chistovich (1985) employed such a measure for estimating the spectral centre of gravity from the 'auditory' spectrum, but used an integration range spanning the entire spectrum. Ranges wider than 3 Bark were also investigated in attempting to account for the present matching data, but these led to poorer results.

excitation patterns computed using the critical band filterbank model proposed by Sekey and Hanson (1984). The final pair (analyses 5 and 6) employed the excitation pattern model of Moore and Glasberg (1983). Further details concerning the auditory models are given in Chapter 3.

Excitation patterns calculated according to the model of Sekey and Hanson (1984) are shown in Figure 4.8 for each of the 15 reference stimuli. Excitation level (in dB) is plotted as a function of critical band rate (in Bark) spanning the 0-10 Bark (about 1280 Hz) range. One important feature of these patterns is the presence of 3 to 6 peaks corresponding to low-order harmonics of the vowels. The excitation pattern for the 0 dB stimulus in the bottom row (F1=350 Hz) has its maximum near 4 Bark, which reflects the presence of the third harmonic (375 Hz), close to the first formant frequency. Harmonics in the vicinity of the second formant are not clearly resolved; there is little evidence for two distinct formant peaks. The same appears to be true for the 0 dB stimulus in the two higher rows. It would be difficult to estimate the location of F2 from these patterns. However, there are other aspects of these patterns (such as the overall spectral balance and the slope above and below the peak) which may convey information concerning the location of the second formant.

Attenuation of higher formants leads to a decline in the slope of the excitation patterns above F1. An increase

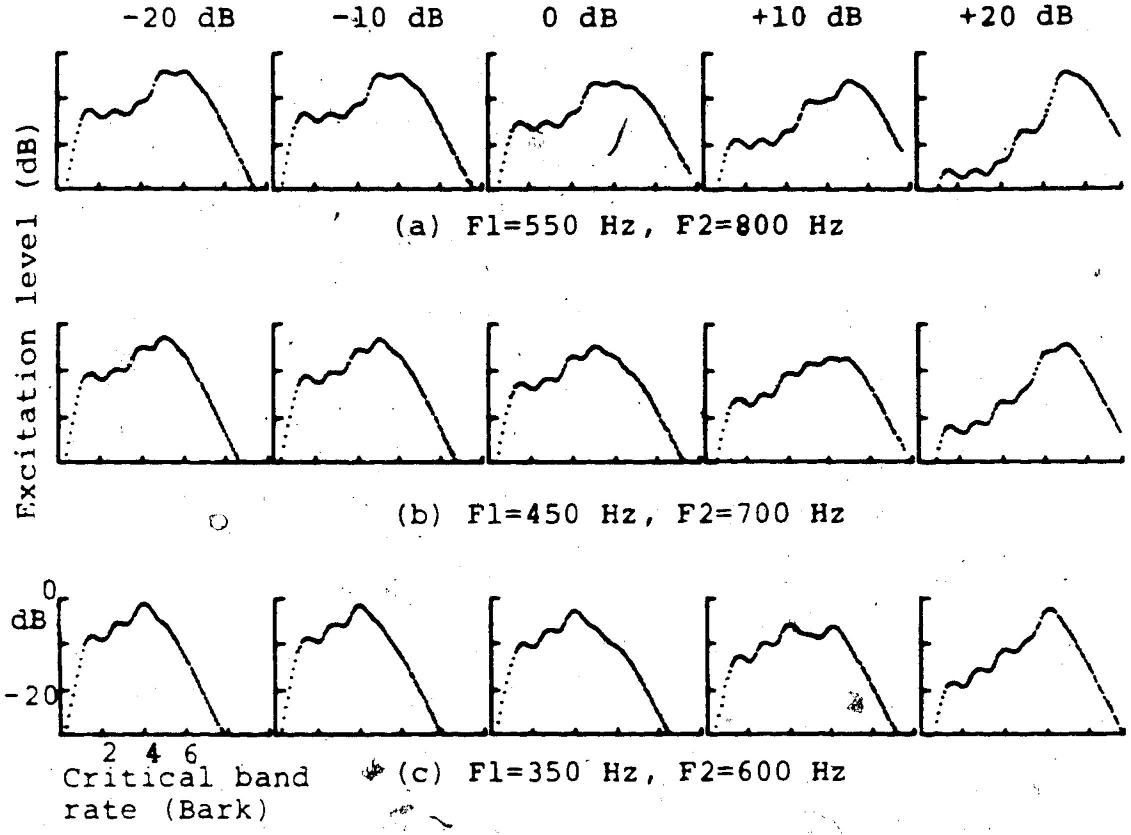


Figure 4.8 Excitation patterns for the fifteen reference stimuli of experiment five

in amplitude has a different effect: additional harmonics above F1 appear as distinct peaks. In the +10 dB condition the patterns are less 'peaky' in appearance, although there are distinct maxima corresponding to harmonics near both F1 and F2. In the +20 dB condition there are strong peaks corresponding to prominent harmonics in the F2 region; F1 harmonics are poorly resolved.

Table 4.2 presents the RMS deviations between observed and predicted F1 values according to each of the centre of gravity models. A useful comparison is provided by the "no change" hypothesis which states that all matched F1s are positioned at the nominal F1 in the reference stimulus. Predictions based on the nominal F1 resulted in an RMS value of 113.91 Hz, almost twice the values for the other analyses. The centre of gravity measures thus appear to account for some aspects of matching performance.

The degree of success in predicting the matching data is fairly similar in each of the models. The centroid measure led to smaller RMS values than the centre frequency measure. Predictions based on auditory models (analyses 3 to 6) yielded marginally better results than those based on comparable measures applied to the power spectrum (analyses 1 and 2).

The lowest RMS values were obtained using analysis 4, the centroid of the maximum 3 Bark region of the excitation pattern. Comparable results were obtained by using a

	ANALYSIS	RMS (Hz)
	Nominal F1 (no change hypothesis)	113.91
1.	Centre Frequency: Power spectrum	69.28
2.	Centroid: Power spectrum	48.64
3.	Centre Frequency: Excitation pattern (I)	48.57
4.	Centroid: Excitation pattern (I)	44.45
5.	Centre Frequency: Excitation pattern (II)	54.34
6.	Centroid: Excitation pattern (II)	47.07
7.	Weighted mean of LPC estimated F1 and F2	47.34

Table 4.2. RMS deviations (in Hz) between observed and predicted matching data for Experiment 5 based on centre of gravity measures

loudness preemphasis function (Hermansky, Hanson, and Wakita, 1985) and/or a subsequent .3 power transformation (Stevens, 1966).

Centre of gravity locations based on analysis 4 are shown in Figure 4.9. The nominal F1 values for the vowel continuum are indicated by the solid line along the diagonal; F2 values coincide with the dashed line. Fmean values are close to F1 in the low frequency range, but move progressively closer to F2 at higher frequencies.

Observed and predicted F1s based on analysis 4 are shown in Figure 4.10. The 3 curves from lowest to highest represent stimuli with F1=350, 450, and 550 Hz. The predicted values are in most cases fairly close to observed means. The largest discrepancy occurs in the 10 dB

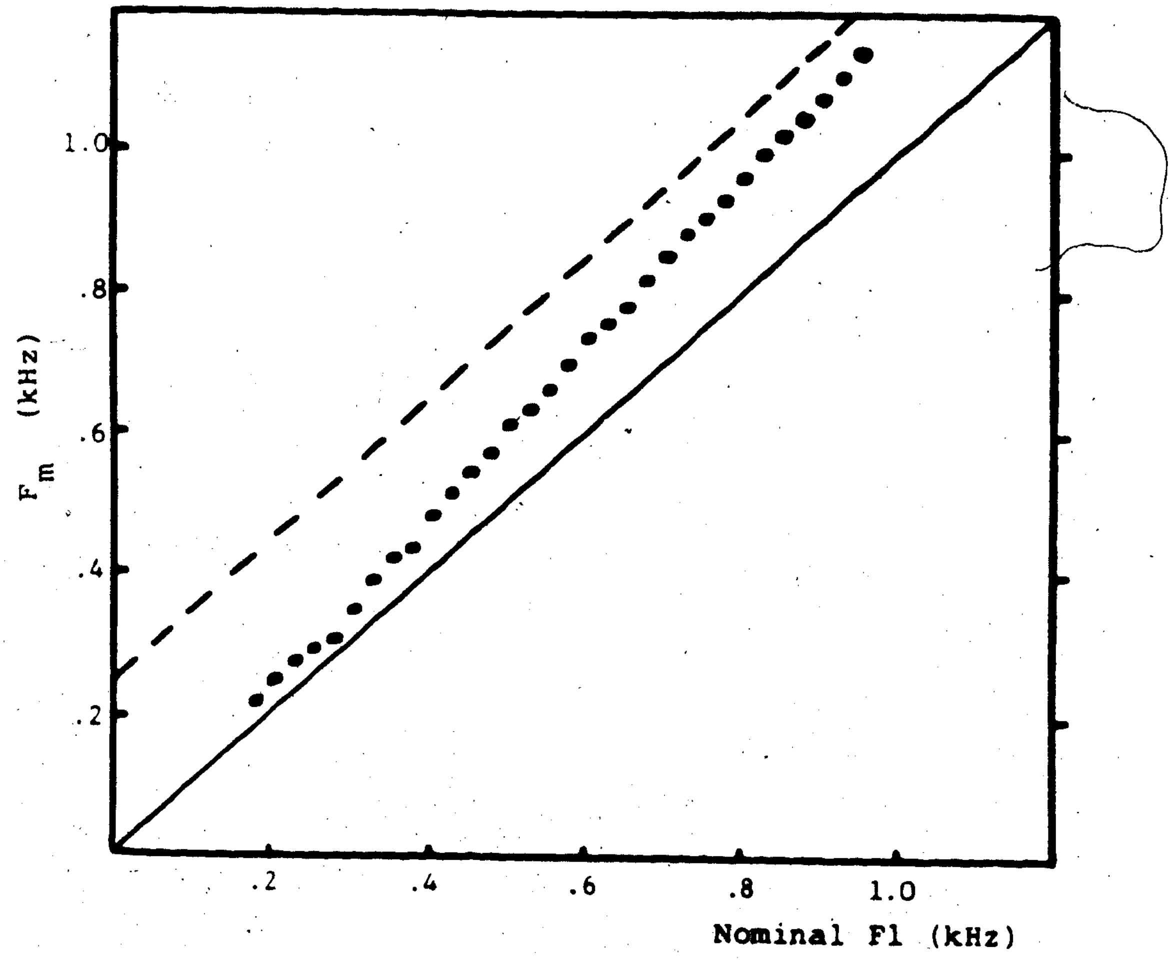


Figure 4.9 Formant frequencies and measured centre of gravity, F_m , based on analysis four (centroid of the maximum 3 Bark region of the excitation pattern, using the model proposed by Sekey and Hanson, 1984), for the matching stimuli of experiment five. Solid line indicates F1 locations, dashed line indicates F2.

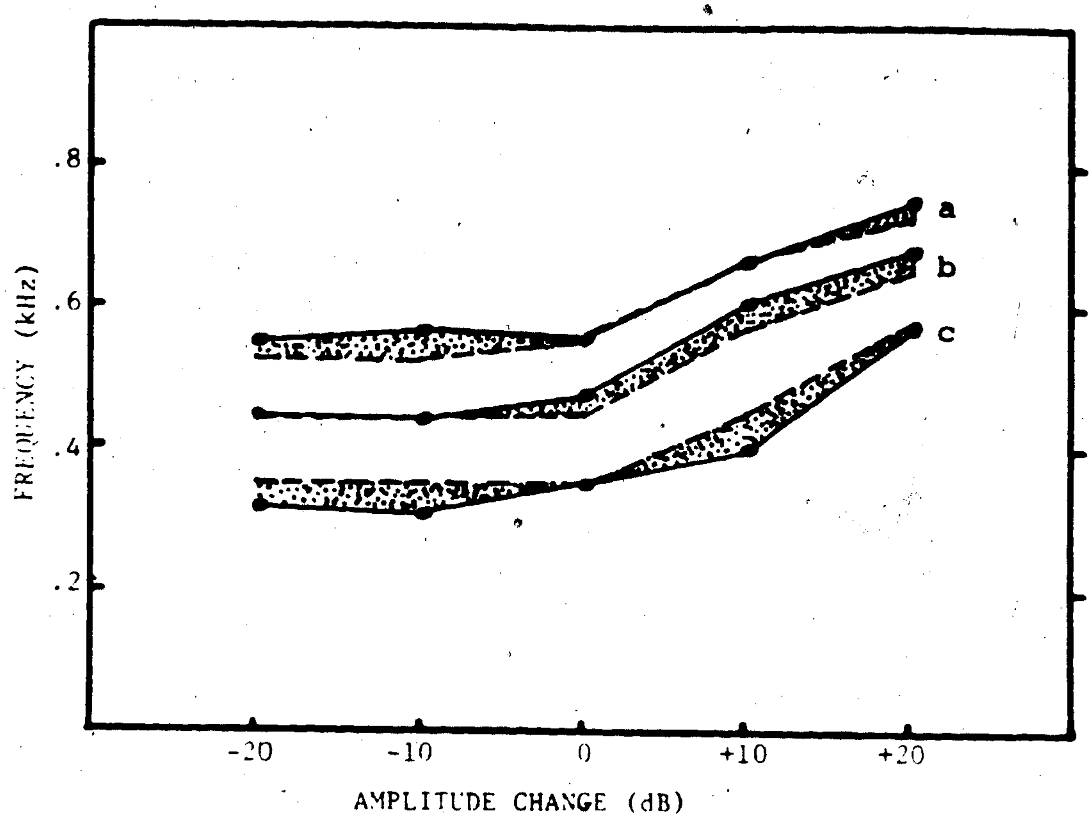


Figure 4.10 Mean matched F1 values (solid lines) and predicted F1 values (dashed lines) based on analysis four

(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

condition, for a stimulus with F1 = 350 Hz; the predicted F1 value is 55 Hz higher than the observed mean. Reference stimuli in the +10 and +20 dB conditions were predicted to have higher values than the control; in the +20 dB condition this value is near F2 of the reference. This aspect of the data is well modeled by the centre of gravity predictions. Attenuation of the second formant does not lead to substantially lower predicted F1s; this is also consistent with observed means. The reason for this appears to be the small influence of F2 on the excitation patterns, as noted above. In fact, F2 lies just above the upper cutoff of the maximum 3 Bark region for all attenuation levels, even in the 0 dB condition. A further problem is that F1 occurs below the lower cutoff of the maximum 3 Bark region in the +20 dB condition. Similar problems emerged with other centre of gravity measures as well. A wider integration range would ensure that both F1 and F2 were included, but it was found that matching predictions based on a wider integration

<u>Weighted mean of LPC estimated F1 and F2</u>. Another type of procedure for estimating the formant centre of gravity is an averaging process which is applied only at the frequencies of peaks corresponding to formant locations (Chistovich, 1985). A simple version of this model was implemented using LPC estimates of the frequencies and amplitudes of the spectral peaks. LPC spectra were computed as described in section 3.3.7. Formant frequencies (F1 and F2) and

amplitudes (A1 and A2) were measured from the LPC spectra using a peak estimation procedure based on the negative second difference method of Christensen, Strong and Palmer (1976). The weighted mean of F1 and F2 was calculated as follows:

$$F = [(F1 \cdot A1)+(F2 \cdot A2)]/(A1+A2)$$

Predictions based on the weighted mean of LPC estimated formant frequencies (analysis 7 of Table 4.2) are shown in Figure 4.11. While this procedure ensures the centre of gravity is determined by both F1 and F2, it was found that this formant estimation procedure did not give accurate estimates of the formant locations in some of the amplitude conditions. In the +20 dB condition, F1 was invariably 'missed' and its 'slot' filled by F2. This also happened at the two highest frequencies in the +10 dB condition.

It may be that more sophisticated procedures will provide more accurate measurements of the formants. Even so, it is not clear how such a procedure is to operate when there is only a single peak present, as in the matching stimuli used by Chistovich et al. (1979). In the light of these difficulties, an explicit formant-averaging procedure does not seem well-motivated. Predictions based on centre of gravity measures in the region of highest energy concentration provided fairly accurate predictions, but it was noted that slightly better predictions could be obtained

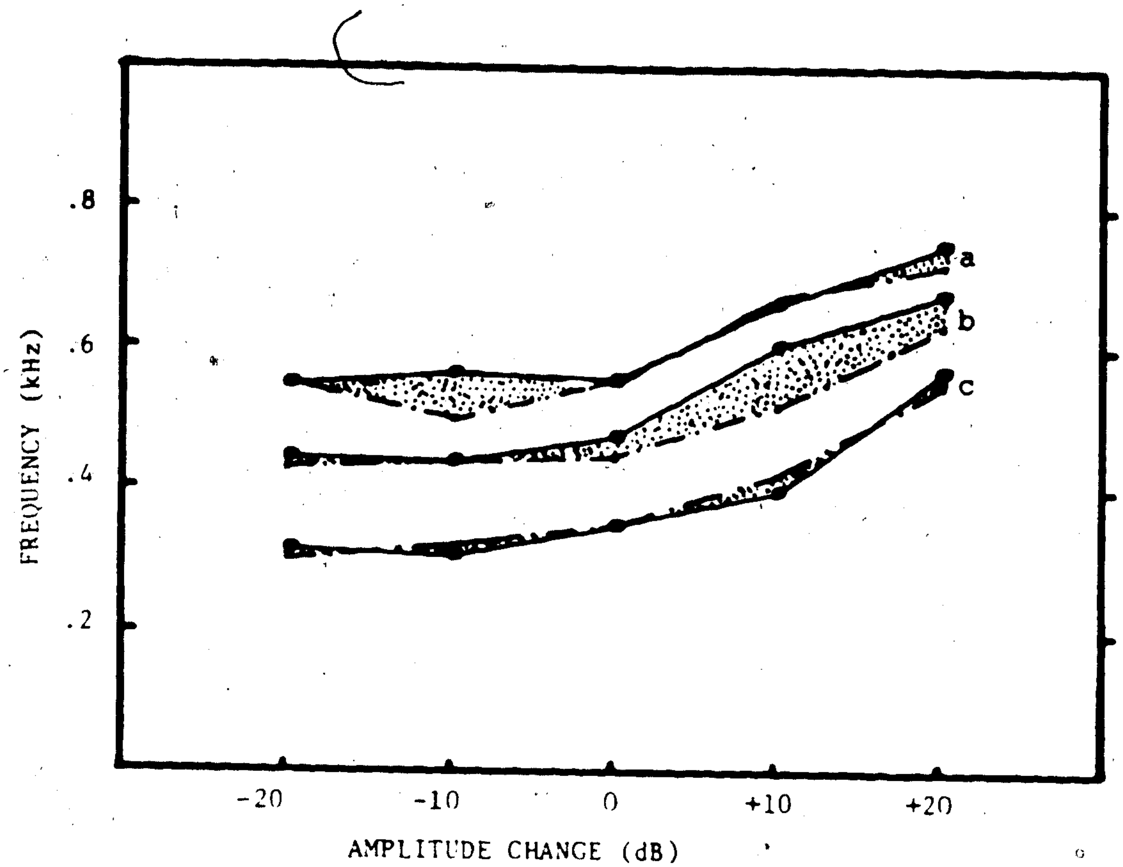


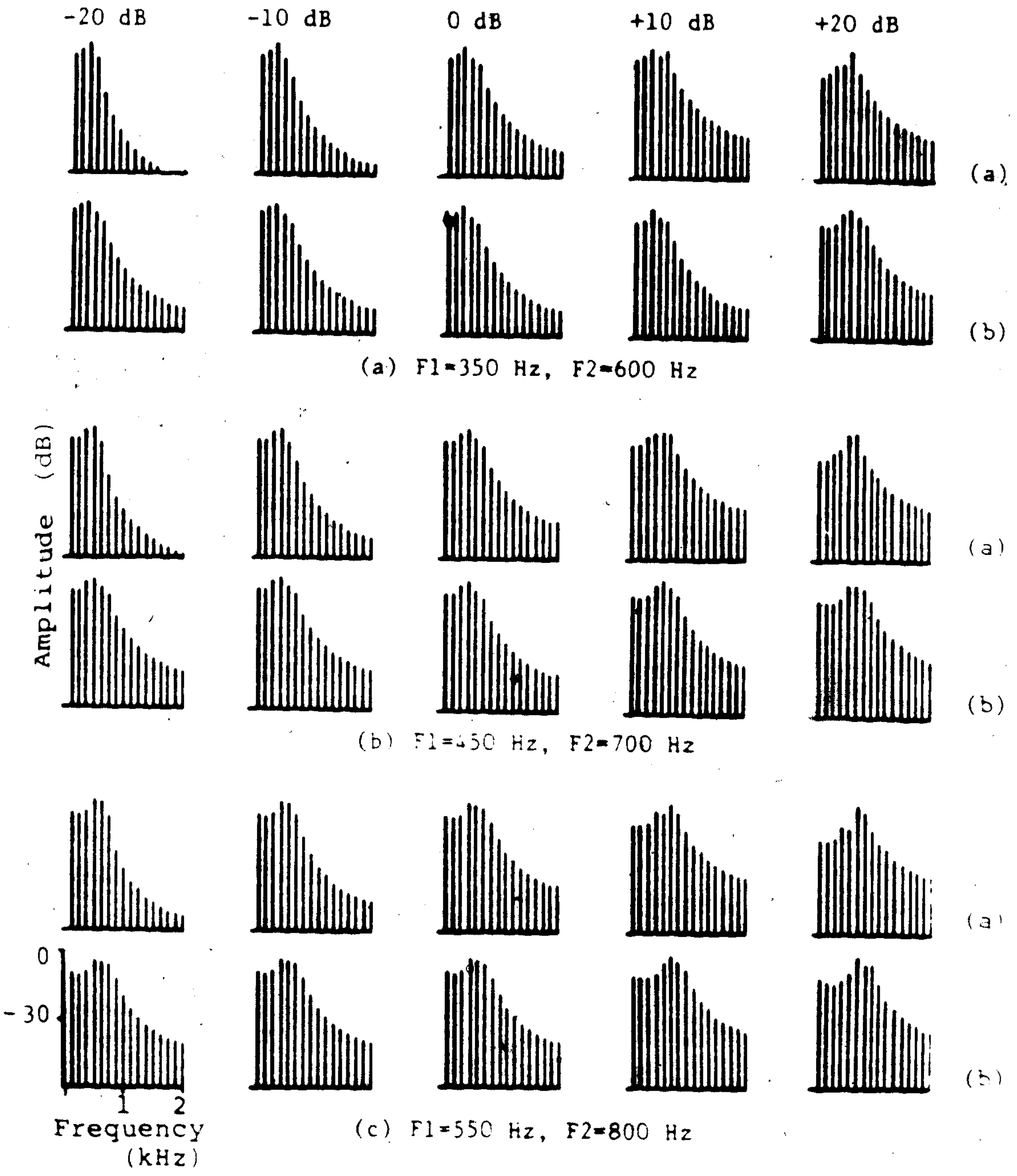
Figure 4.1 1 Mean matched F1 values (solid lines) and predicted F1 values (dashed lines) based on analysis seven (weighted mean using LPC estimated formant values).

(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

by calculating the centre of gravity over a range smaller than 3 Bark.

Comparison of line spectra of best matches. A comparison of the line spectra of 'best matches' suggests that matching may be based on the alignment of a small number of prominent harmonics in the reference and matching stimuli. Figure 4.12 shows line spectra for each reference stimulus (upper plot) and its 'best match' (i.e. the match most frequently selected by listeners). It should be noted that the single most prominent harmonic in the spectrum of the reference stimuli occurs near the F1 peak for stimuli in the 0, -10 and -20 dB conditions; and near F2 in the +20 dB condition. In the +10 dB condition the most prominent harmonic occurs near F1 for F1=350 Hz, but near F2 for F1=450 and 550 Hz.

It can be seen that the major peaks in the spectra are closely aligned in the best match pairs. All of the reference stimuli have in common with their best match the frequency location of the most prominent harmonic. However, this did not appear to be the exclusive basis for matching judgements. It was noted, for example, that although the most prominent harmonic may occur at the same frequency location in a number of potential matching stimuli, some are selected by listeners more often than others. For example, using a single harmonic matching criterion, all of the stimuli in the region from F1=325 Hz to F1=425 Hz should be equally good matches for the reference stimulus with F1=350



(a) F1=350 Hz, F2=600 Hz

(b) F1=450 Hz, F2=700 Hz

(c) F1=550 Hz, F2=800 Hz

Figure 4.12 Line spectra for each of the reference stimuli (a) and the 'best' matching stimuli (b) of experiment five.

Hz and a +10 dB change in amplitude, since they have the most prominent harmonic (at 375 Hz) in common. Yet there is a definite preference for matching stimuli with F1=400 or 425 Hz. The levels of harmonics adjacent to the most prominent also appear to be important in matching.

In Chapter 3 several local centre of gravity measures were investigated in the prediction of front vowel matching data. It was found that the weighted mean of the levels of the two most prominent harmonics could account fairly well for the selection of matching stimuli by listeners. A further series of analyses was carried out to determine how well the local centre of gravity measures based on the weighted mean of prominent harmonics can predict the back vowel matching data, and to compare the results with the predictions of the centre of gravity models compared in Table 4.2.

The weighted mean of the k most prominent harmonics, F_m was calculated according to the procedure described for front vowels in section 3.3.7. The exponent q determines the choice of scale (1=magnitude, 2=power); p determines the degree of preemphasis (0=none, 1=6 dB/octave).

The effects of the parameters k, p, and q can be seen in Table 4.3, which presents the RMS error rates for several different analyses.

	p=0		p=1	
k	q=1	q=2	q=1	q=2
1	63.4	63.4	80.9	80.9
2	41.4	42.0	65.8	66.4
3	63.6	51.0	50.4	48.6

Table 4.3. RMS distance (in Hz) between observed F1 matches and predictions based on the weighted mean of the k most prominent harmonics; q represents the amplitude exponent, p represents the degree of preemphasis

Predictions of the back vowel matching data based on 2 harmonics (k=2) yielded the smallest RMS values, consistent with the model of Carlson, Fant and Granstrom (1975) and the results of the front vowel matching experiments. The lowest RMS value (41.4 Hz) was obtained with k=2, p=0 and q=1.

The relationship between nominal F1 values and F_m, the weighted mean of the two most prominent harmonics (k=2, p=0, q=1) is indicated in Figure 4.13. F_m follows F1 closely in the low frequency region, with discrete steps at intervals separated by the fundamental frequency. This tendency was also observed in Figure 3.20 for the front vowels. For the back vowel continuum, however, F_m values were higher than F1 (close to the midpoint of F1 and F2) at higher frequency positions on the continuum.

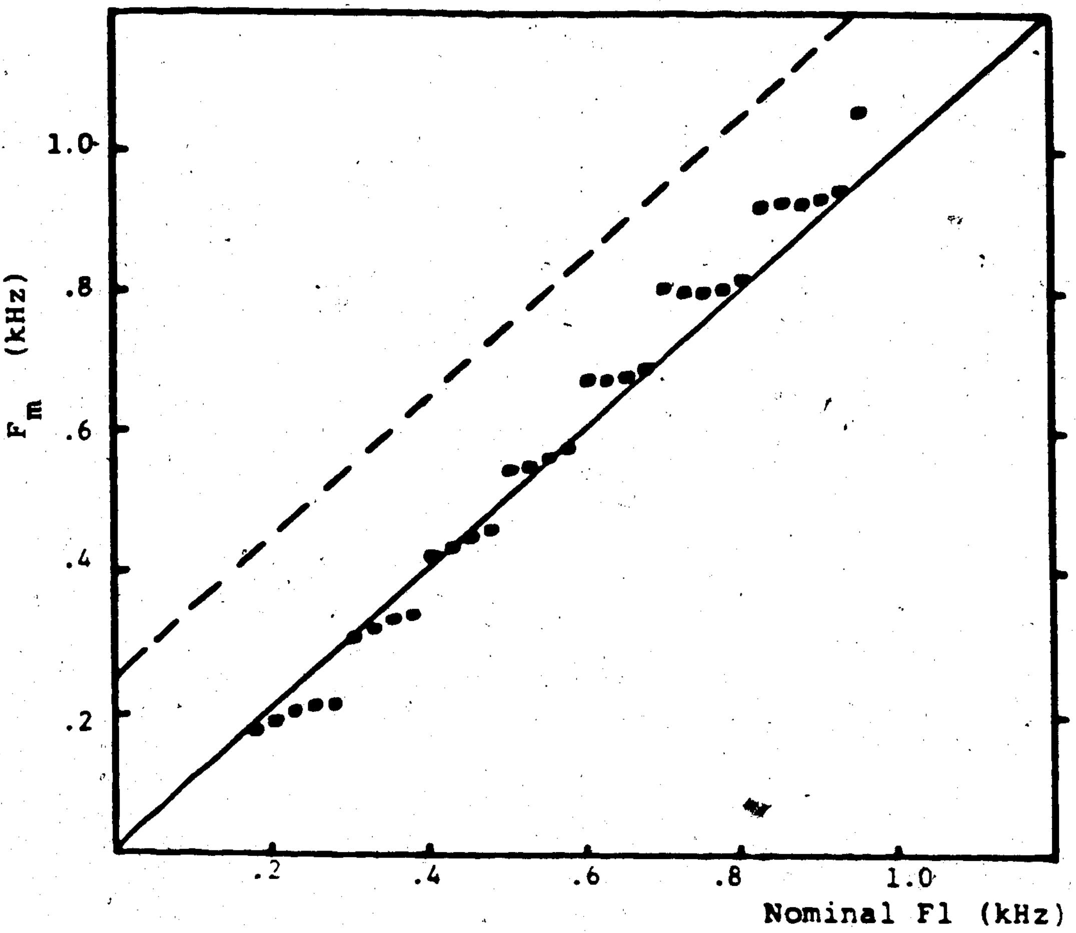


Figure 4.13 Formant frequencies and measured centre of gravity using the weighted mean of the two most prominent harmonics (p=0,q=1), for the matching stimuli of experiment 5. Solid line indicates F1 locations, dashed line indicates F2.

Matching predictions based on this model are shown in Figure 4.14. Predicted F1s for the lowest curve (F1=350 Hz) are somewhat higher than the observed means. In the middle curve (F1=450 Hz) predicted F1 values in the +10 dB condition are somewhat lower than observed means. The minimal effects of a decrease in amplitude and the large effects of an increase in amplitude are evident in the model predictions and follow the observed data fairly closely.

Table 4.3 indicates that predictions based on preemphasized spectra (p=1) were generally worse than those based on unpreemphasized spectra (p=0), unlike the results for front vowels for which weighted mean calculations based on preemphasized spectra gave better results. The effects of preemphasis were discussed in section 3.3.4, where the possibility was raised that preemphasis may provide a better characterization of the relative auditory prominence of harmonic components than the raw spectral magnitudes. Why then does preemphasis not result in improved predictions of the back vowel matching data?

It was found that weighted means based on preemphasized spectra (k=2, p=1, and q=1) were generally higher than the nominal F1. The two most prominent harmonics were often closer to the nominal F2, particularly with higher formant frequencies. F1 harmonics were thus excluded from the calculations; these components may play an important role in the perception of back vowels.

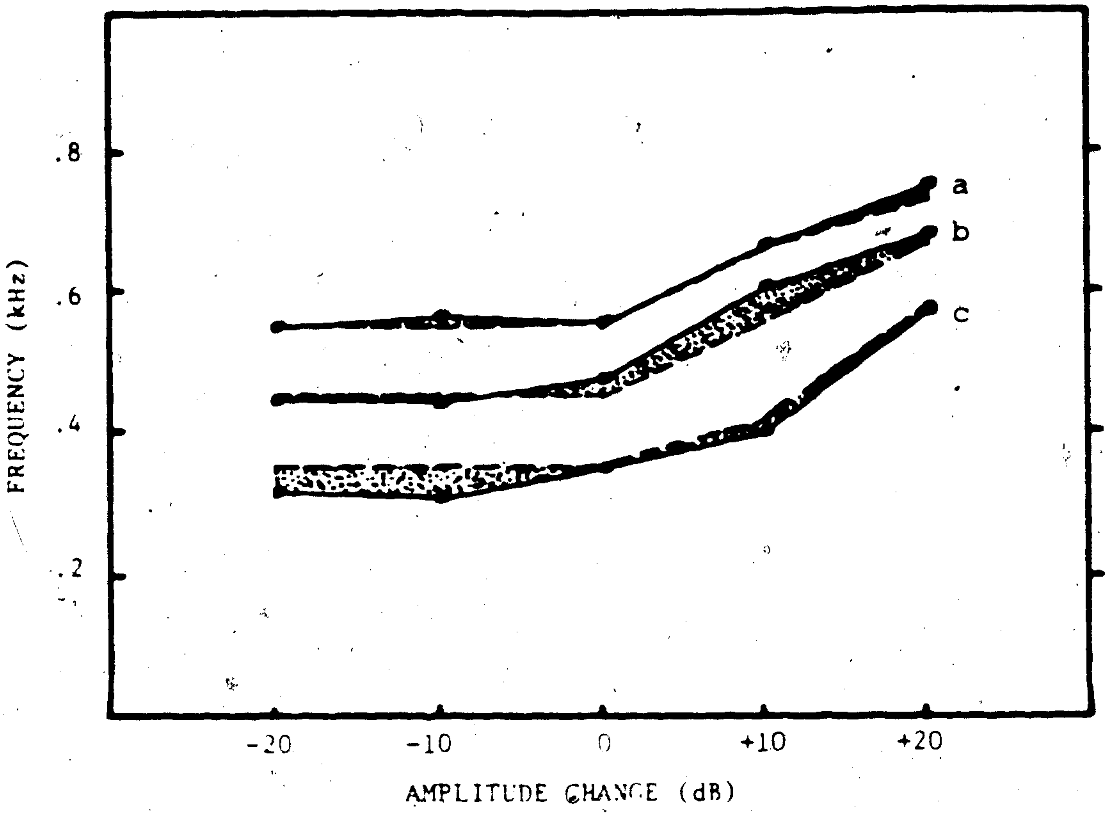


Figure 4.14 Mean matched F1 values (solid lines) and predicted F1 values (dashed lines) based on the weighted mean of the two most prominent harmonics (p=0,q=1).
(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

One factor to be considered is the shape of the vocal tract transfer function in the low frequency region (Darwin, 1984). In chapter three it was noted that harmonic levels below F1 in front vowels are much more constrained in the range of intensities they can assume than those occurring above the peak. In back vowels, harmonics around the F1 peak show smaller variations in intensity, and the degree of contrast in this region of the spectrum is reduced, relative to front vowels. If the perceptual system assigns a weight to each harmonic reflecting an implicit knowledge of constraints of this kind, the effective contribution of harmonics below and above F1 may be more nearly symmetrical.

A second possible factor is a masking effect of harmonics in the vicinity of F2 on lower harmonics near F1. The masking effects of F2 harmonics on those near the F1 peak region can be determined to some extent from an examination of the excitation patterns, as discussed earlier. The degree of resolution of individual harmonic components in the F1 region may vary considerably depending on relative formant amplitudes and the distance between F1 and F2.

To investigate these effects, the most prominent harmonics were estimated from excitation patterns calculated according to the procedure of Sekey and Hanson (1984). Harmonics were identified by a peak extraction algorithm

[1] Peaks were identified using the criterion $S_{i-1} < S_i > S_{i+1}$, where S_i is the intensity of the ith frequency bin; each bin

Up to 8 peaks were located in the frequency region between 0 and 1500 Hz. The frequencies of the peaks were in all cases close to the expected harmonic frequencies. Next, a search was made for the most prominent peak and the larger of the two adjacent peaks, corresponding to the most prominent and second most prominent harmonics.

The weighted mean of the two most prominent peaks in the excitation pattern was calculated for each of the matching and reference stimuli. In some cases there was only a single harmonic in the peak region. For these stimuli, the frequency of the peak was substituted for the weighted mean[1]. The RMS value for predictions based on this analysis was 50.42 Hz. This value is somewhat higher than that obtained using the weighted mean of the two most prominent harmonics of the amplitude spectrum. Figure 4.15 presents the predicted F1 values and observed means. It can be seen that the predictions are fairly close, except in the +10 dB condition where predicted F1s are consistently higher than observed means. Although predictions based on excitation patterns were not better than those based on magnitude spectra, the results were fairly close, suggesting that the measurement procedure is insensitive to the effects

[1](cont'd) was separated by 0.1 Bark.

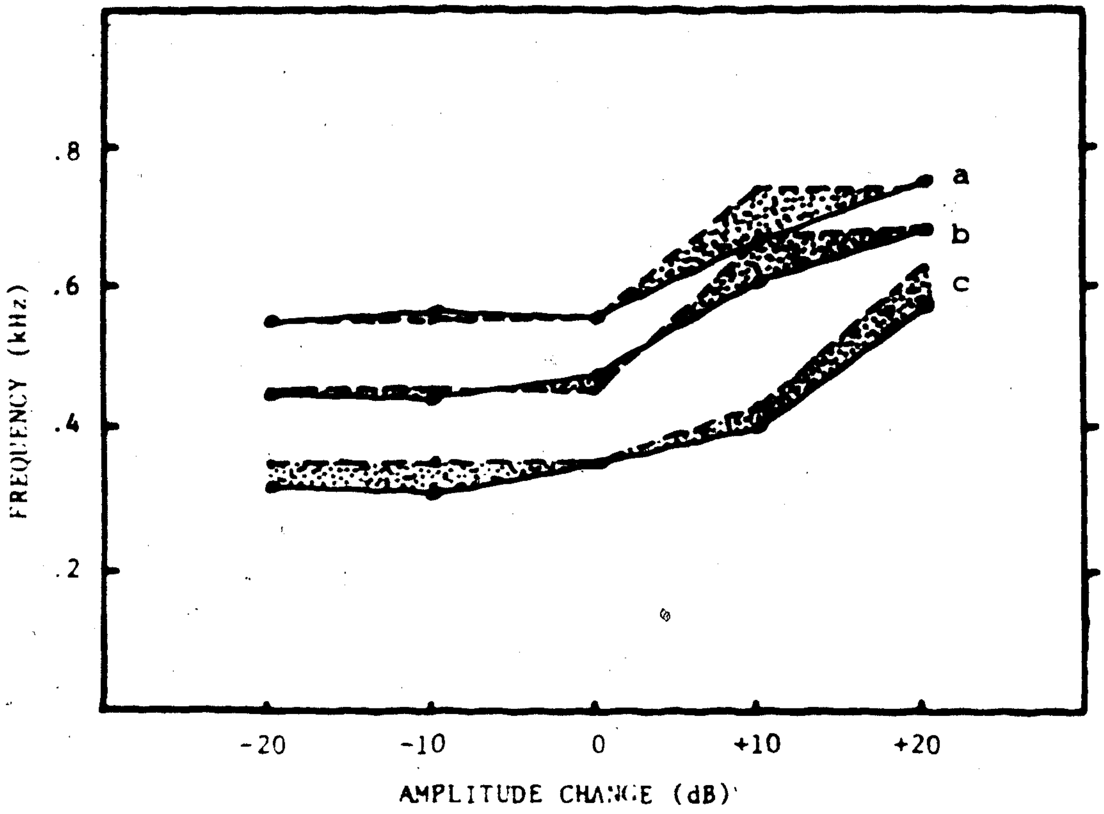


Figure 4.15 Mean matched F1 values (solid lines) and predicted F1 values (dashed lines) based on the weighted mean of the two most prominent harmonics estimated from excitation patterns based on the model proposed by Sekey and Hanson (1984).
(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

of critical band filtering.

4.5.5 General discussion

The effects of formant amplitude on back vowel matching with F1 and F2 in close proximity were qualitatively different from those reported by Chistovich and Lublinskaya (1979). In their study, listeners matched the frequency of a single formant to a two formant stimulus whose F1 and F2 values were appropriate for back vowels. The single formant was positioned between F1 and F2, and its preferred location depended on the relative amplitudes of the two formants. Chistovich and Lublinskaya proposed that the phonetic quality of back vowels is determined by the centre of gravity of the F1-F2 cluster. Chistovich (1985: p. 793) suggested that "a change of relative formant amplitudes in two-formant back vowels should be equivalent to a displacement of the formants along the frequency scale."

The present experiment tested this hypothesis using multiformant matching and reference stimuli. Listeners were required to adjust F1 and F2 simultaneously (with a fixed frequency separation) to compensate for an increase or decrease in the amplitudes of F2 and higher formants. If the centre of gravity is the sole determinant of back vowel quality, it should not matter if the formant peaks are misaligned, provided that the centre of gravity location [illegible] the reference and matching stimuli is the same.

It was found that listeners did not select significantly lower frequency matches as F2 was attenuated. An increase in amplitude did lead to shifts toward higher F1-F2 values. For the largest increment (+20 dB) F1 in the matching stimulus was positioned near F2 of the reference. With a +10 dB increment, increased variability and a multimodal matching pattern was observed. One possible explanation for this finding is a masking effect of F1 by higher frequency components. The F1-F2 centre of gravity cannot be the exclusive determinant of back vowel matching: alignment of the formants, or of prominent harmonics near the major formant peaks, also appears to be important.

Centre of gravity measures based on the centroid of the maximum 3 Bark region of the excitation pattern resulted in fairly accurate predictions; however, similar results were obtained using the weighted mean of the two most prominent harmonics. The overall pattern of results does not support the proposal that back vowel quality is determined by the frequency location of the centre of gravity of the spectrum. Instead, it suggests that listeners align major peaks in the spectra when matching stimulus pairs which are both multiformant vowels.

4.6 Experiment Six

4.6.1 Introduction

Bedrov, Chistovich and Sheikin (1978) reported the results of an identification experiment as further evidence for the operation of a centre of gravity mechanism in back vowels. Two formant vowels with close spacing of F1 and F2 were used, with three conditions of relative formant amplitude: the second formant was either (1) boosted by 5 dB, relative to the first formant; (2) attenuated by 5 dB, relative to F1; or (3) attenuated to silence. Identification functions were shifted in each of the amplitude conditions. However, these shifts were not simply changes in the category boundaries; there were also radical changes in the shapes and heights of the curves. It is not clear that a linear shift in F1 of the identification functions presented by Bedrov et al. will result in a satisfactory alignment of the identification curves.

These results were interpreted in terms of the FCOG hypothesis, which predicts an effect of amplitude on the phonetic quality of back vowels when the formants are separated by less than an estimated critical distance of 3-3.5 Bark. Chistovich and Lublinskaya base these estimates on matching data; no evidence for critical distance was presented using identification data.

It is difficult to reconcile the findings of Ainsworth and Millar (1972) with this hypothesis. They investigated the effects of formant amplitude on vowel identification, using a larger sample of vowels covering the entire F1-F2 plane. Their data show relatively minor effects of formant amplitude changes; however, only the effects of F2 attenuation were investigated. There was little evidence that only vowels with very close spacing of F1 and F2 were affected by formant amplitude changes.

The present experiment was designed to investigate the effects of formant amplitude changes on vowel identification as a function of the frequency separation of F1 and F2. According to the FCOG hypothesis, the phonetic quality of vowels should be unaffected by changes in the formant amplitudes unless they are separated by less than 3-3.5 Bark. The resulting shifts in phonetic quality are predicted to follow the change in centre of gravity. It should be possible to project the identification functions specified in a two-dimensional plane whose coordinates are F1 and F2 onto a single dimension specifying the centre of gravity location, with no loss of phonetic information.

4.6.2 Method

Stimulus materials

The vowel stimuli formed a two dimensional continuum as shown in Figure 4.16. There were 8 steps in F1 (440, 480, 520, 560, 600, 640, 680, and 720 Hz) and 3 steps in F2 (850, 1000 and 1150 Hz). Close to half of the stimuli had more than 3 Bark separation of F1 and F2; about one-third had greater than 3.5 Bark separation of the formants.

The amplitude spectrum for each vowel was calculated according to Fant's (1960) cascade formant synthesis model as described in section 3.2. Two sets of vowels were produced. The first set was synthesized with a fundamental frequency of 125 Hz, typical of an adult male voice; the second set had a fundamental of 250 Hz, characteristic of an adult female voice. In the low pitch condition, formant bandwidths and and formant frequencies (F3 to F6) were identical to those indicated in Table 3.1. In the high pitch condition, formant frequencies above the second were scaled upward by a factor of 1.2, to take into account the observed association between f_0 and formant frequencies in natural speech (Peterson and Barney, 1952). Formant bandwidths were identical with those in the low pitch condition, as it was found that wider bandwidths did not result in more natural stimuli.

The amplitudes of components in the region above the F1 peak were manipulated to produce 3 different conditions (-6 dB, 0 dB, +6 dB) for each vowel, using the procedure described in section 4.5.2. Figure 4.17 illustrates the weighting functions used to produce the 3 amplitude conditions for a typical vowel from the continuum

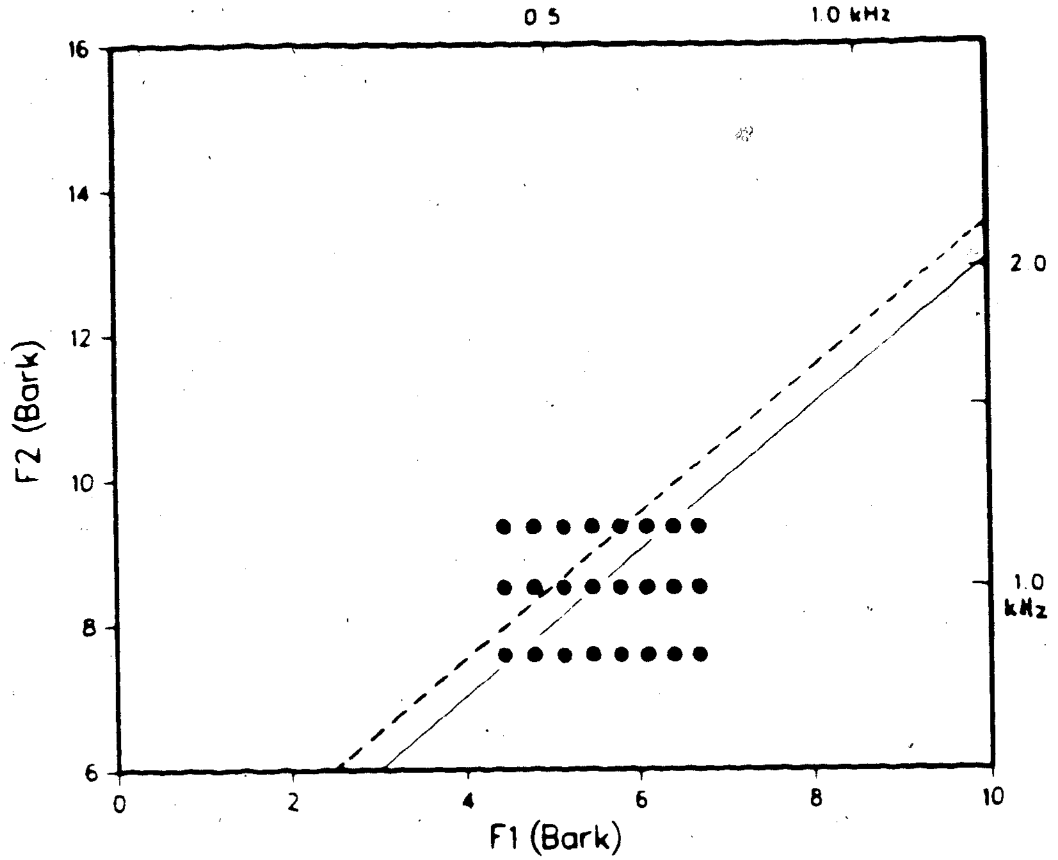


Figure 4.16 Experiment six. Formant frequency values for vowel continuum. Solid line indicates where F2-F1=3 Bark, dashed line, F2-F1=3.5 Bark

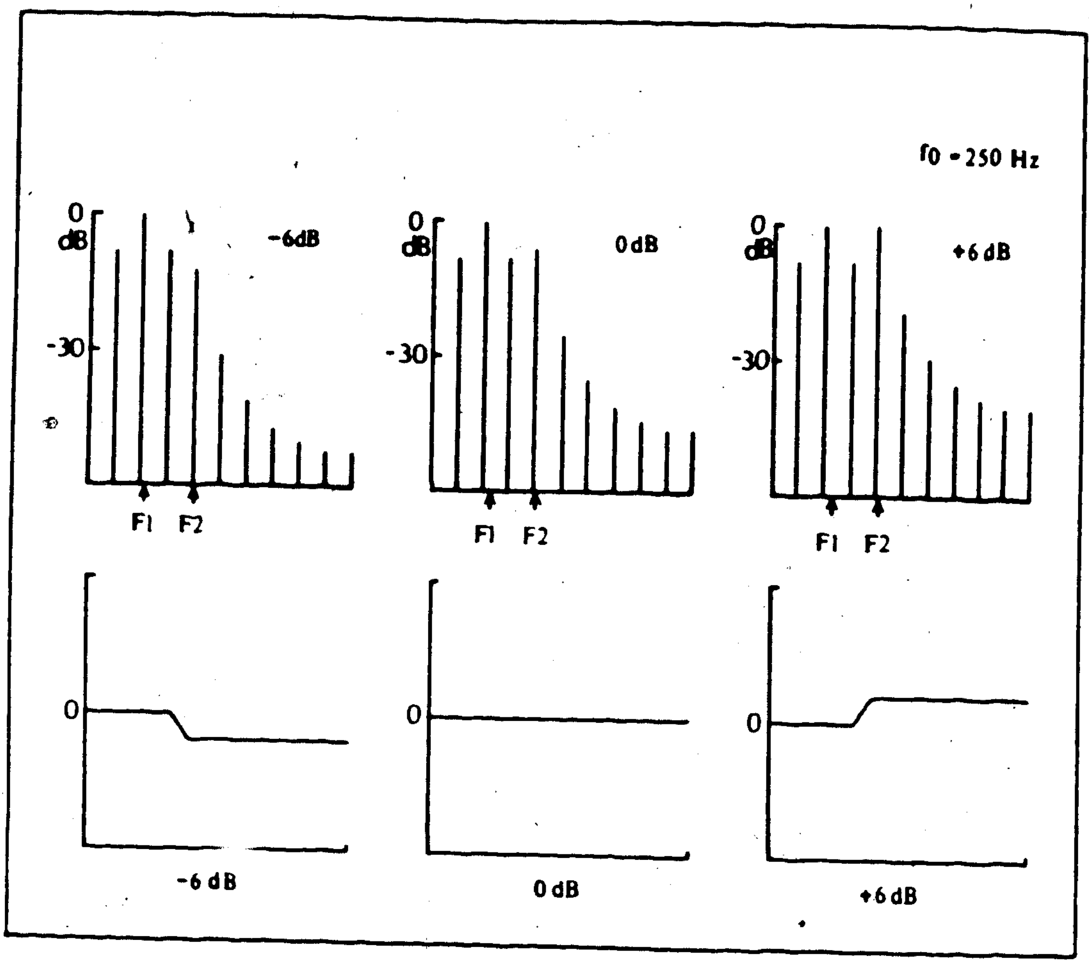


Figure 4.17 Experiment 6. Weighting functions and line spectra for a typical stimulus from the continuum (f0=250 Hz, F1=600 Hz, F2=1000 Hz)

(f_0=250 Hz; F1=600 Hz; F2=1000 Hz). The middle figure represents the 0 dB or "no change" condition; on the left, all components above F2 are attenuated by 6 dB; on the right, they are raised by 6 dB.

Digital additive harmonic synthesis was used to produce steady state stimuli 250 ms in duration, with a 15 ms rise/fall as described in section 3.3.2. Phase angles were specified as follows:

$$\varnothing_j = \pi \cdot j^2/n$$

where $\varnothing_j$ is the phase angle, in radians, corresponding to the jth harmonic component, and n is the total number of harmonics (Schroeder and Mehrgardt, 1982)[1].

Listeners

Eight listeners (4 men, 4 women) took part in the experiment. All had received prior training in phonetics and had participated in one or more experiments involving natural and synthetic speech sounds. None of the listeners exhited any hearing loss, as determined by audiometric testing. All of the listeners were native speakers of Canadian English.

Procedure

[1] This formulation avoids the strong peak factor arising from all-sine or all-cosine phases, and produces a mellower quality when used as an excitation signal for a source-filter vocoder.

Listeners were tested individually in a sound treated room. Stimulus presentation and response collection were both conducted on-line using the procedure described in experiment 4. A multiple choice switchbox was used to record listeners' responses. A five-alternative forced choice task was used. Beside each button was a label with a phonetic symbol and a keyword in standard English orthography:

/ɑ/	/U/	/ʌ/	/o/	/u/
'pot'	'hook'	'cup'	'boat'	'tool'

Each session was comprised of 144 items (8 steps in F1, 3 steps in F2, 3 amplitude conditions, and 2 replications). The 144 items were randomized, with a different randomization list for each subject. Each listener completed 4 sessions, 2 for the low pitch condition and 2 for the high pitch condition. Four of the listeners completed two sessions of the high pitch condition followed by two sessions of the low pitch; for the other four listeners this order was reversed. Individual sessions lasted about 20 minutes.

None of the listeners had any difficulty in labelling the stimuli using the 5 available response categories[1].

[1] One of the subjects reported occassionally hearing a stimulus which sounded like a front vowel but had no difficulty assigning it to one of the five back vowel

4.6.3 Results and discussion

If the primary cue ~~for vowel~~ identification in stimuli with close spacing of F1 and F2 is the formant center of gravity, or some weighted average of the energy associated with each formant peak, then it should be possible to make some general predictions about the effects of formant amplitude changes on vowel identification. An increase in the level of F2, for example, will lead to an upward shift in the centre of gravity, and a decrease in F2 level will pull the centre of gravity to a lower frequency value; changes in F1 level will have the opposite effect. As the formant centre of gravity is decreased, we may expect an increase in the proportion of responses to vowel categories associated with a lower centre of gravity, and diminished responses to vowels characterized by a higher centre of gravity.

As a first approximation to the centre of gravity associated with each vowel category, a set of measurements was carried out:

(1) F1 and F2 means from 10 speakers (see Figure 4.2) were used to produce synthetic vowels similar to each of /u U ɔ ɑ ʌ/. Separate spectra were calculated for adult male and adult female averages.

(2) F1 and F2 values were transformed to Bark units

[1] (cont'd) categories.

(Schroeder, Atal and Hall, 1979) and the vowels were separated into two classes; those with less than critical spacing of the formants (F2-F1<3 Bark) and those with greater than critical spacing of the formants (F2-F1>3 Bark).

(3) The centre of gravity was calculated as the weighted average of F1 and F2, using LPC analysis to estimate F1, F2, A1 and A2 values (model 7 in Table 4-2).

Male and female averages for /u/ and /U/ both showed more than 5 Bark separation of F1 and F2, while /o/, /ɑ/ and /ʌ/ had less than 3.5 Bark separation. For both male and female averages, the following relationship between the vowel centres of gravity was obtained:

/o/ < /ɑ/ < /ʌ/

(where 'x<y' means 'y has a higher centre of gravity than x'; in all cases, these differences exceeded 100 Hz.)

A stimulus in the -6 dB condition has a lower centre of gravity than a corresponding stimulus in the 0 dB condition; the same stimulus with a +6 dB amplitude change has a higher centre of gravity. Thus we might expect that a number of /ʌ/ responses to a given stimulus in the 0 dB condition will change to /ɑ/ or /o/ responses in the -6 dB condition; and that some /ɑ/ responses will be replaced by /o/. The reverse is predicted to occur in the +6 dB condition. No changes are expected for any stimulus with more than 3.5

Bark separation of the formants.

Modal vowel response profiles. Figures 4.18 and 4.19 present the modal, or most frequently reported, vowel response for each of the 72 stimuli in the low (125 Hz) and high (250 Hz) pitch conditions. Each cell is identified by a number from 1 to 24 above the vowel symbol, to facilitate comparisons among the three amplitude conditions. Below each vowel symbol is the percentage of responses assigned to the modal category; these percentages are based on 4 replications from 8 listeners. In most cases the modal response was also a majority response; the stimulus was assigned to the modal response class on greater than 50 percent of all trials.

In the low pitch condition, amplitude changes had little effect on the modal responses. Only one of the predicted changes was observed, a shift of /ɑ/ to /o/ in the -6 dB condition (item 20). Stimulus item 13 shifted from /ʌ/ in the 0 db condition to /ɑ/ in the +6 dB condition; it was predicted to change in the opposite direction. Several other changes involved stimuli with more than 3 to 3.5 Bark separation of F1 and F2. For example, stimulus items 4 and 12 shifted from /ʌ/ in the 0 dB condition to /U/ in both -6 and +6 dB conditions, contrary to the centre of gravity predictions.

The high pitch condition produced some of the predicted changes (for example, /ʌ/ changed to /ɑ/ in the -6 dB

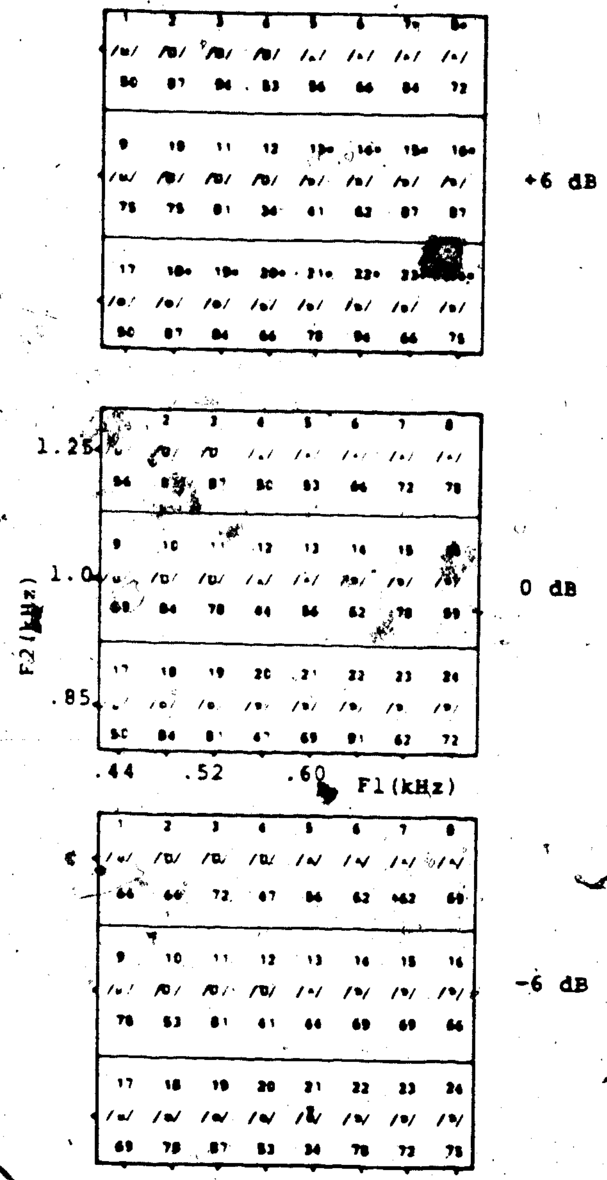


Figure 4.18 Modal vowel responses for the low pitch condition of experiment six. Percentage of responses indicated below each vowel symbol. Double symbols indicate tied values. * F2-F1 < 3 Bark

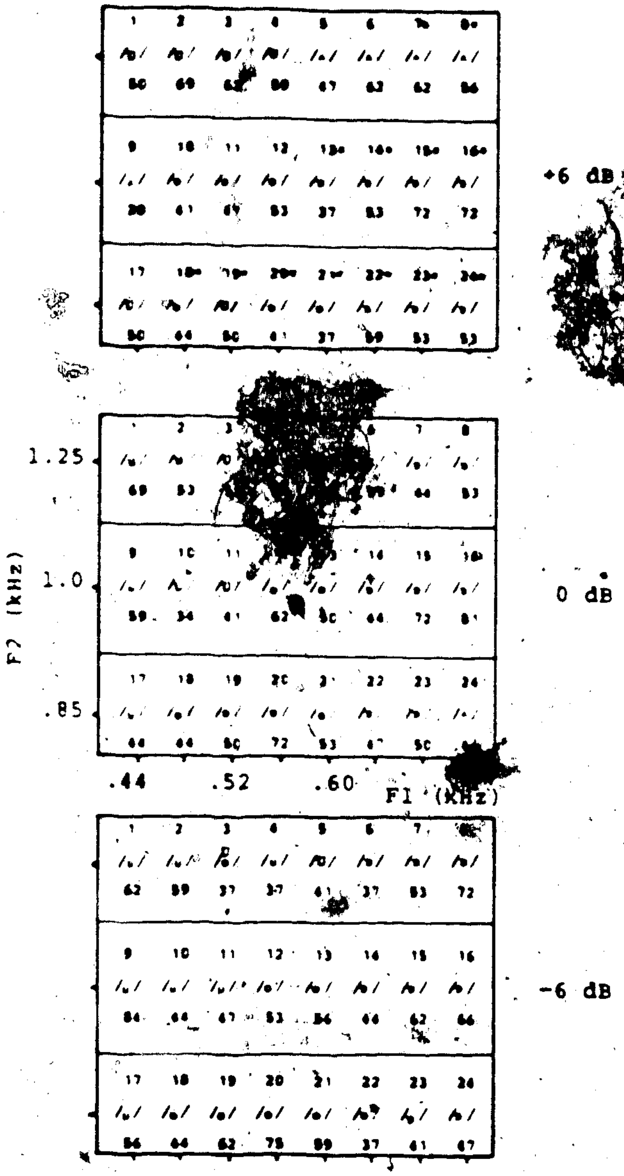


Figure 4.19 Modal vowel responses for the high pitch condition of experiment six. Percentage of responses indicated below each vowel symbol. Double symbols indicate tied values. * F2-F1 < 3 Bark

condition (items 6 and 24)[1]; /ɑ/ went to /ʌ/ in the +6 dB condition (items 7 and 8); and /o/ became /ɑ/ in the +6 dB condition (item 13).

It may be noted that the percentages are generally lower in the high pitch condition than in the low pitch condition. Some of the listeners also reported greater uncertainty in labelling the high pitched vowels. Ryalls and Lieberman (1982) have presented evidence that synthetic vowels with high fundamental frequency are identified less accurately than vowels with low fundamental frequency. They suggested that high pitched vowels are more frequently misidentified because the spectrum envelope is not as well defined as that of a low pitched vowel since there are fewer harmonics in the regions of the formant peaks. In natural speech, however, vowels with high fundamental frequency do not appear to be less intelligible than those with low pitch. Changes in fundamental and formant frequencies over the course of a vowel, and the presence of aperiodic vocal source components (Makhoul, Viswanathan, Schwartz & Huggins, 1978) may improve the specification of the spectrum envelope in natural high-pitched vowels.

Individual vowel response profiles. Although the modal

[1] It is surprising that item 24 was classified as /ʌ/ in the 0 dB condition, since its formant frequencies are more appropriate for an /ɑ/. This stimulus has very close spacing of the formants (F2-F1=130 Hz), which resulted in a somewhat shrill quality. The second highest response was /d/ (41 percent), as might be expected on the basis of its formant pattern.

response plots provide a convenient summary of the identification data, it is necessary to examine the complete response profiles for a more detailed picture of the pattern of amplitude effects on vowel identification. Individual response curves for the vowels /u U o ɑ ʌ/ are shown in Figures 4-20 to 4-24. The top part of each figure depicts the data for the low pitch condition, while the lower part represents the high pitch data. Each curve is plotted as the percentage of responses versus the frequency of the first formant. Each of the 3 rows presents a different F2 value, while the columns present the 3 amplitude conditions.

Effects of fundamental frequency. A comparison of the upper and lower panels indicates that fundamental frequency has an effect on vowel identification, consistent with the findings of Miller (1953) Fujisaki and Kawashima (1968), Slepokurova (1973), Carlson, Fant and Granstrom (1974) and Ainsworth (1974, 1975). The identification curves appear to shift to higher values of F1 and F2 from the low to the high pitch condition. These shifts may be attributed to a change in perceived speaker identity: in natural speech, an octave increase in voice pitch is associated with an upward shift of about 15-20 percent in F1 and F2 (Peterson and Barney, 1952).

Effects of formant amplitude. If the centre of gravity in the F1-F2 region is the perceptual variable determining maximum response regions and boundaries when the two

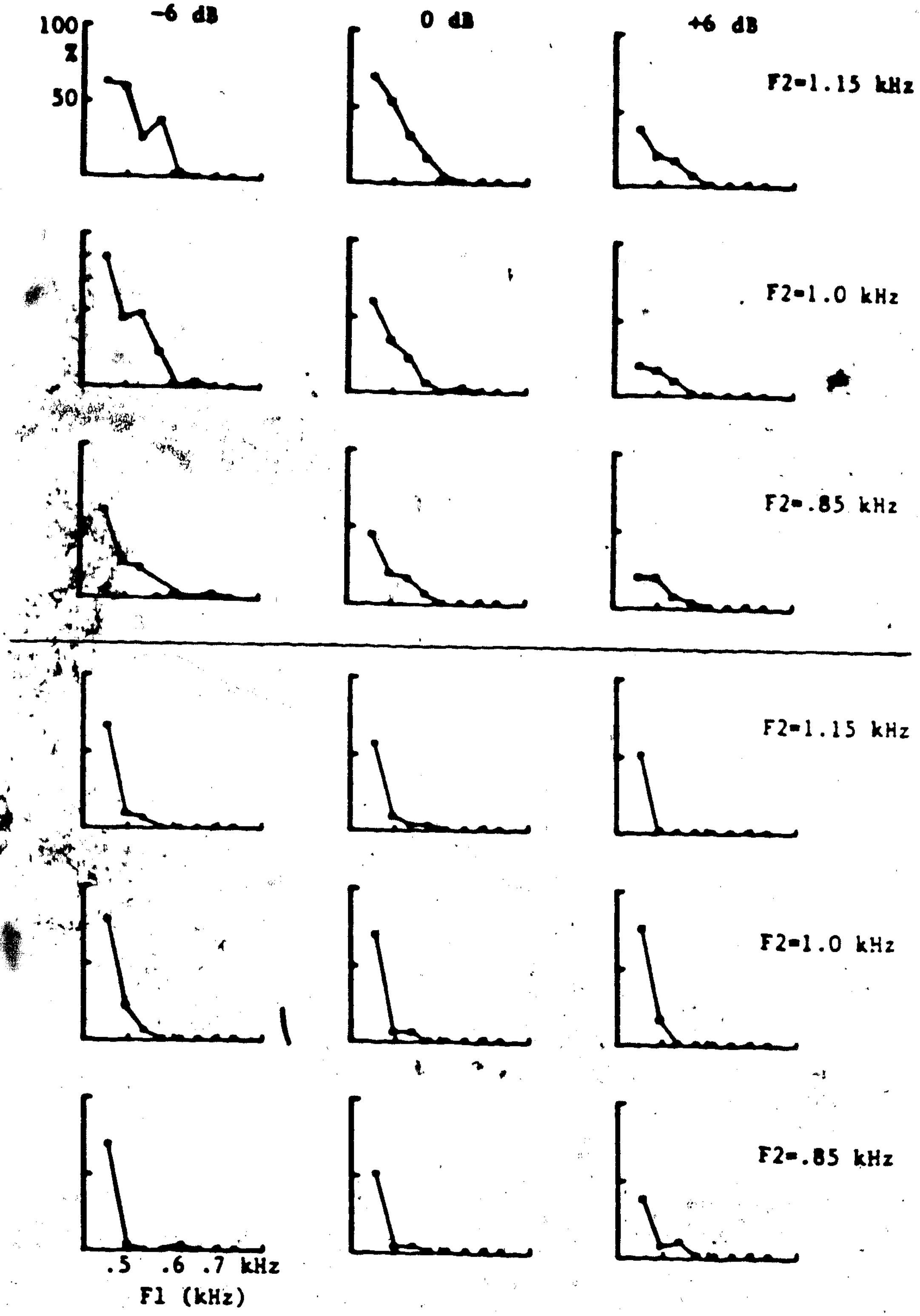


Figure 4.20 Experiment 6. Percentage of / u / responses
Upper panel: f0=250 Hz
Lower panel: f0=125 Hz

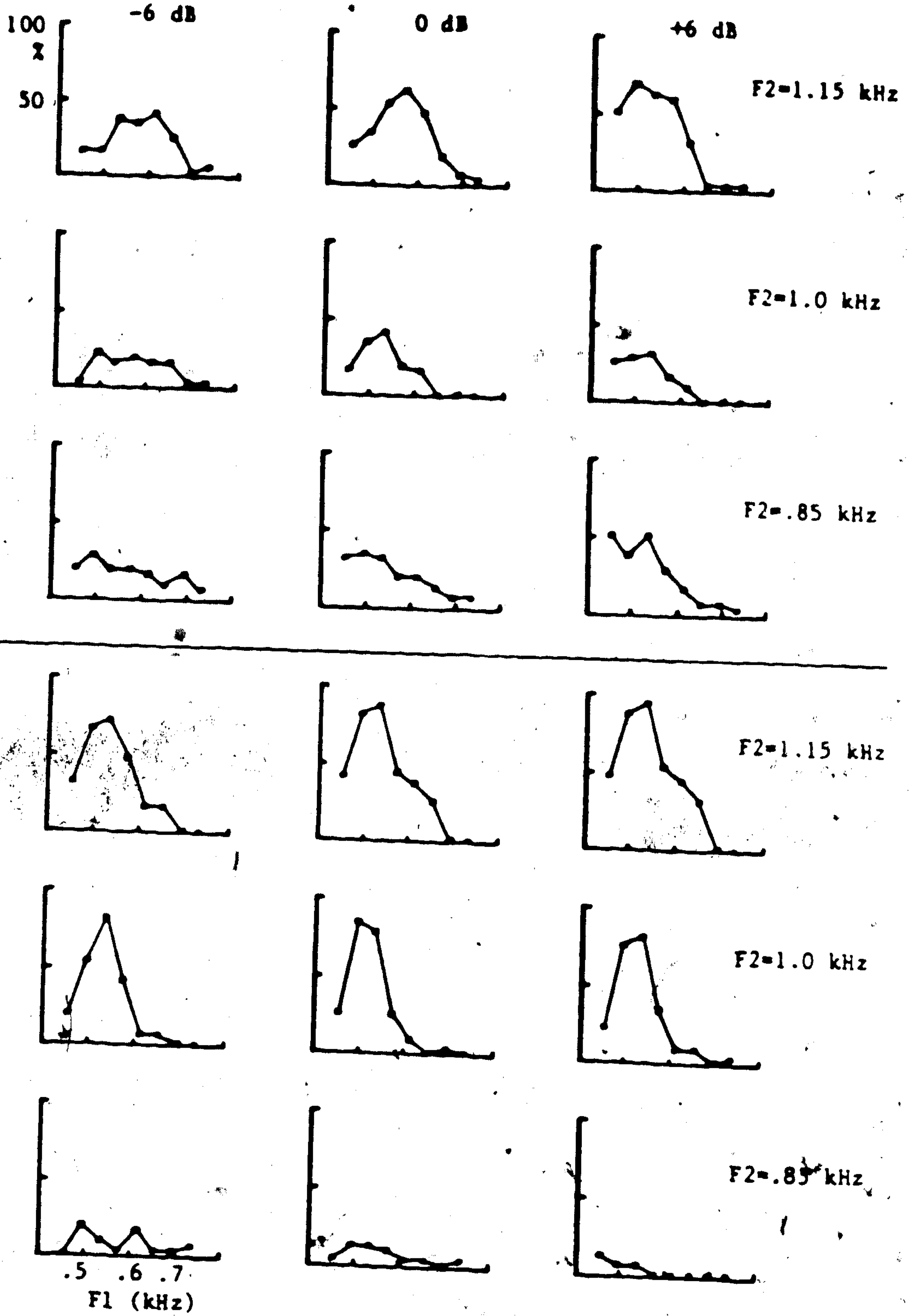


Figure 4.21 Experiment 6. Percentage of / U / responses
Upper panel: f0=250 Hz
Lower panel: f0=125 Hz

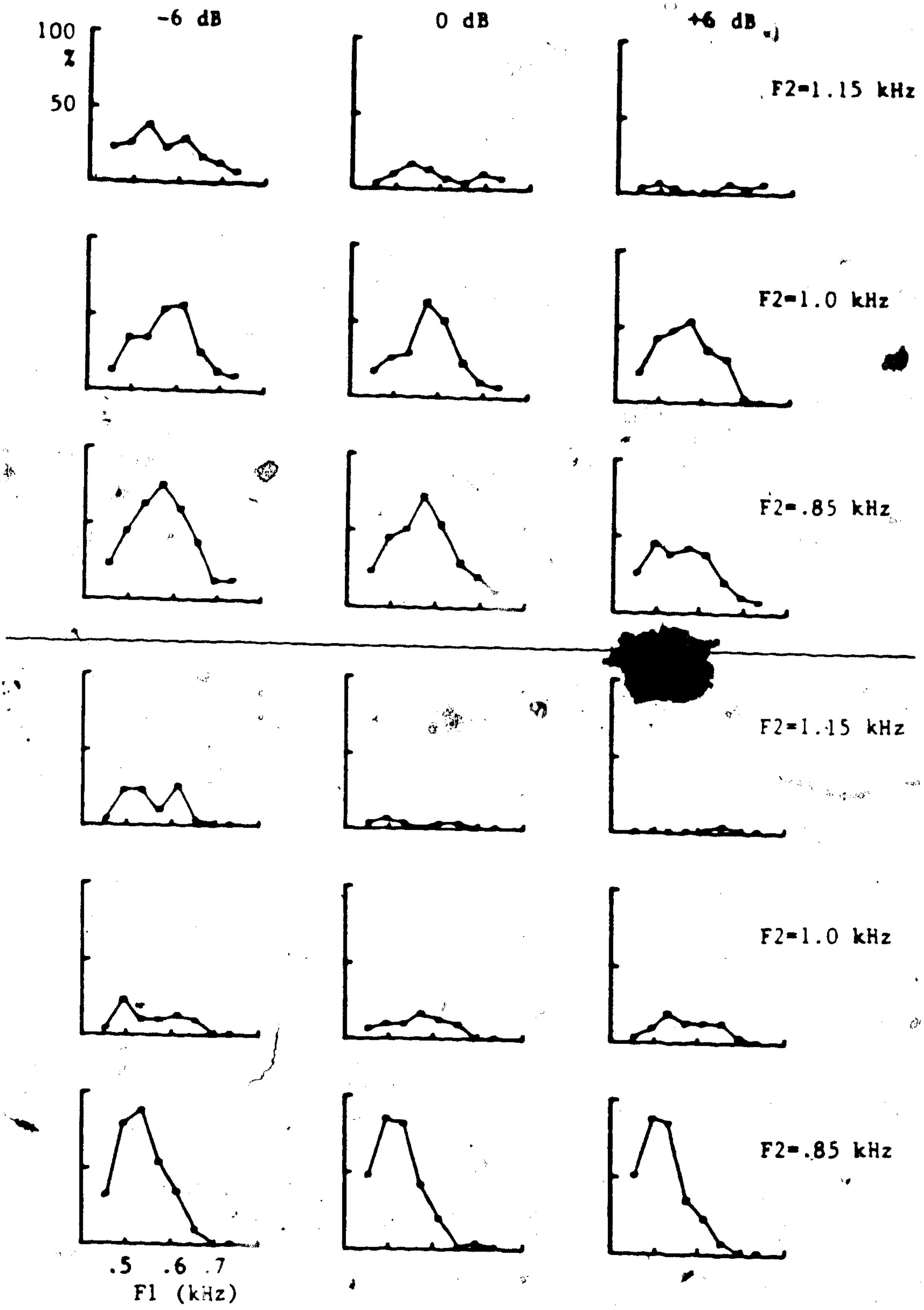


Figure 4.22 Experiment 6. Percentage of / o / responses
Upper panel: f0=250 Hz
Lower panel: f0=125 Hz

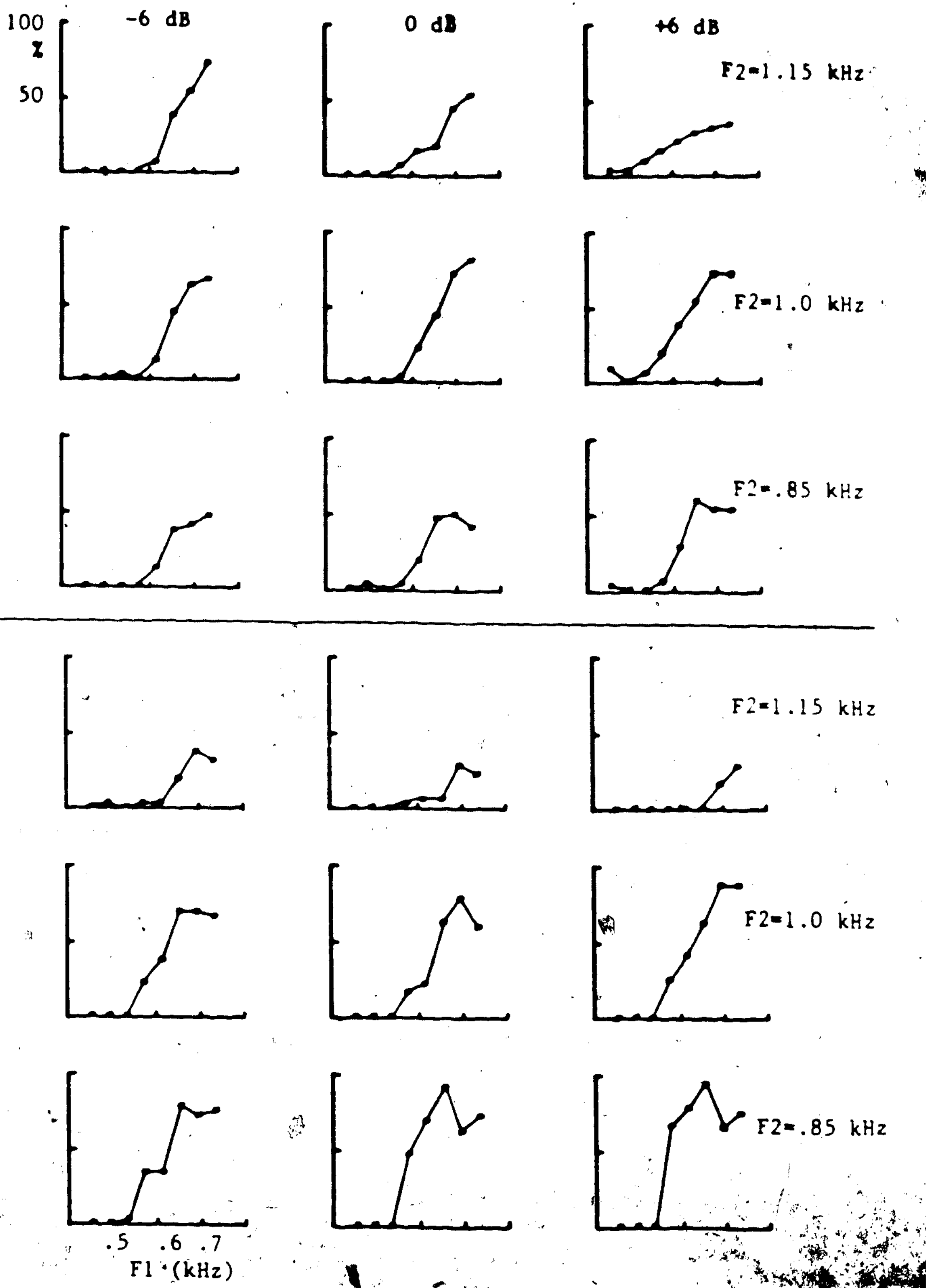


Figure 4.23 Experiment 6. Percentage of / ɒ / responses
Upper panel: f0=250 Hz
Lower panel: f0=125 Hz

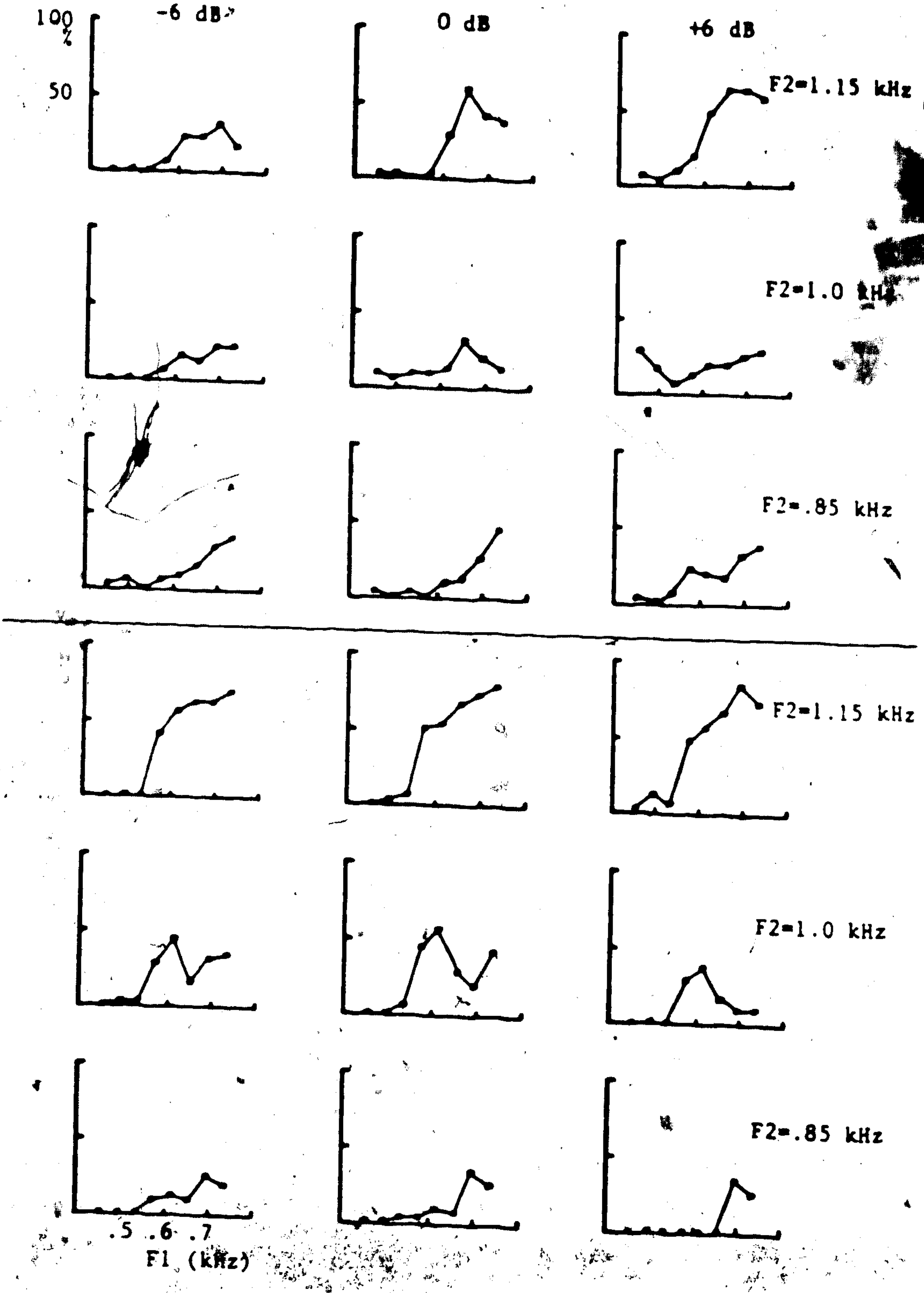


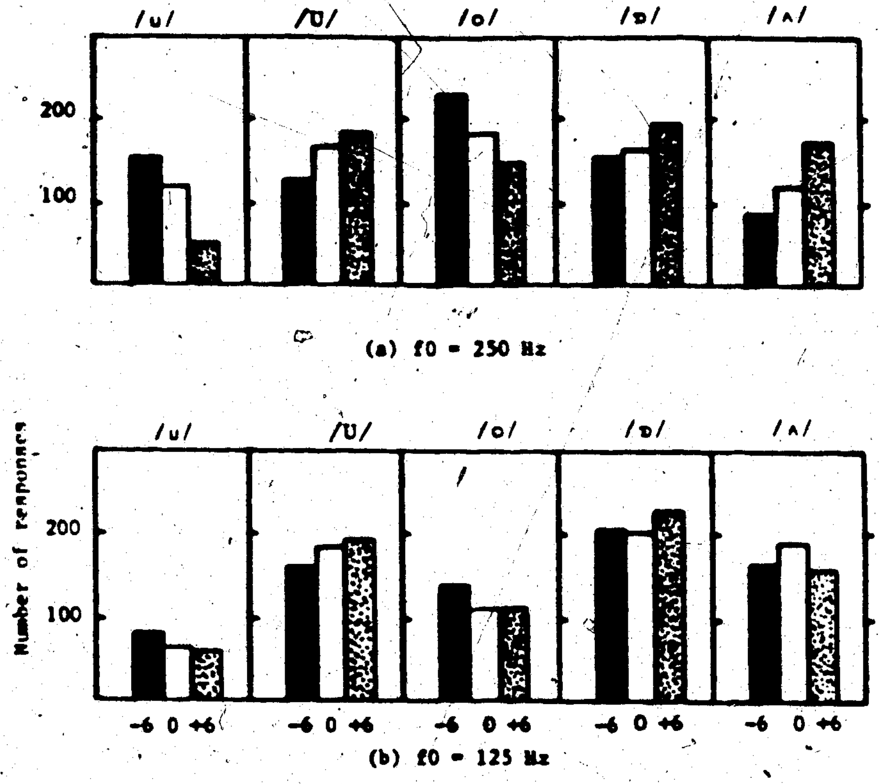
Figure 4.24 Experiment 6. Percentage of / ʌ / responses
Upper panel: f0=250 Hz
Lower panel: f0=125 Hz

formants are close together, then a horizontal shift along either or both of the F1 and F2 axes is expected as the amplitude relations between the formants are altered. Natural back vowels typically differ in both F1 and F2 (see Figure 4.2). A lowered centre of gravity may come about as a result of a lower F1, a lower F2, or both. It might be expected that changes in the centre of gravity resulting from altered formant amplitudes affects the locations of the identification functions along both F1 and F2 dimensions. These changes should only appear in regions where the stimuli have less than 3-3.5 Bark separation of F1 and F2.

The effects of formant amplitude changes are difficult to assess by visual inspection of the response curves. While amplitude is clearly having an effect in some cases, its most salient characteristic is a change in the overall height of the curves rather than a shift along the F1 and F2 scales. Examining the regions where boundary shifts are predicted (e.g, the high F1 slope of the identification functions for the vowel /o/ and the low F1 slopes for /ɑ/ and /ʌ/), it does not appear that the response curves are shifted toward higher frequencies in the -6 dB condition or toward lower frequencies in the +6 dB condition in any systematic fashion. However, it is difficult to estimate category boundaries under these conditions, since there are many cases in which the maxima on the response curves do not exceed 50 percent (i.e., there are no majority responses).

These results raise an interesting possibility. If the three amplitude conditions are primarily affecting the shape and heights of the curves (rather than their location along the F1 and F2 scales), it may be that amplitude effects on vowel identification are to some degree independent of changes in F1 and F2. This is in direct opposition to the hypothesis of Chistovich and Lublinskaya (1979) which states that F1 and F2 and their amplitudes A1 and A2 do not act independently but have a combined influence on the centre of gravity, which determines the perceived vowel quality.

The number of responses to each vowel category was used as a measure by which changes in the heights of the response curves in each amplitude condition was estimated. Means for each vowel are shown in Figure 4.25. Analyses of variance were performed on the data from the low and high pitch conditions. For the low pitch condition, a significant main effect was found for the factor VOWEL (F(4,28)=8.06; p<.001) but not for AMPLITUDE, or the interaction of VOWEL x AMPLITUDE. In the high pitch condition, there was a significant AMPLTIUDE x VOWEL interaction (F(8,56)=5.97; p<.001). Inspection of the data in Figure 4.25 reveals that more /u/ and /o/ responses were given as the amplitude changed from +6 to 0 to -6 dB; while fewer /U/, /ɑ/ and /ʌ/ responses were given. The direction of these changes was consistent across all vowel categories, indicating that amplitude has a systematic effect on the heights of the response curves in the high pitch condition.



(a) f0 = 250 Hz

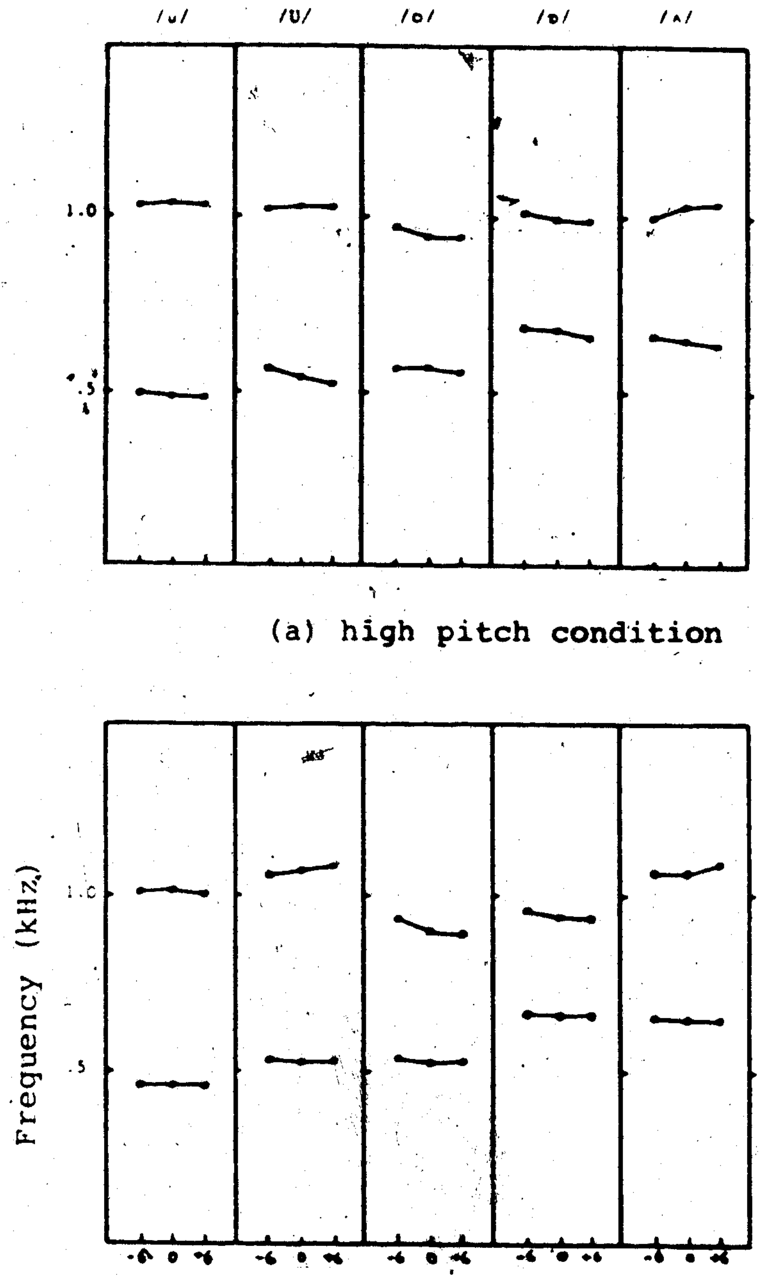
(b) f0 = 125 Hz

Figure 4.25 Number of responses to each vowel category in each amplitude condition for the identification data of experiment six.

Although for reasons discussed above, it was not possible to estimate vowel boundaries accurately, an attempt was made to estimate shifts in location of the identification functions along the F1 and F2 scales by calculating vowel centres, using the procedure described for Experiment 4. This measure was also used by Ainsworth and Millar (1972). Vowel centres were computed for both F1 and F2; means are presented in Figure 4.26.

Consistent with the findings of Ainsworth and Millar (1972), vowel centres do not change substantially as a function of amplitude. In some cases vowel centres shift in opposite directions for F1 and F2. This appears to conflict with centre of gravity predictions that lowering the centre of gravity will result in upward shifts in the identification curves for both F1 and F2, while raising the centre of gravity will shift the curves toward lower F1 and F2 values. Some of the observed changes are in the direction predicted by the centre of gravity hypothesis: for example, there are shifts toward lower F1 and F2 values for the vowels /ọ/ and /ɑ/ as the amplitude of higher components is increased. However, the shifts for the vowel /ʌ/ are not in the expected direction. Substantial changes also occur with the vowel /U/; these changes are unexpected since this vowel has a wide separation of F1 and F2.

Four separate analyses of variance were conducted on the vowel centres for F1 and F2 in the low and high pitch



(a) high pitch condition

(b) low pitch condition

Figure 4.26 Vowel centres for the identification data of experiment six

conditions. In the low pitch condition, only the factor VOWEL led to significant F-ratios (F1 vowel centres: F(4,28)=182.05; p<.001; F2 centres: F(4,28)=46.29; p<.001). No significant effect was found for either AMPLITUDE or for the VOWEL x AMPLITUDE interaction. In the high pitch condition, F1 centres resulted in significant main effects for AMPLITUDE (F(2,14)=23.8; p<.001) and VOWEL (F(4,28)=83.18; p<.001); and a significant interaction of VOWEL x AMPLITUDE (F(8,56)=2.20; p<.05). F2 centres in the high pitch condition resulted in a significant effect only for the factor VOWEL (F(4,28)=8.82; p<.001); neither AMPLITUDE nor VOWEL x AMPLITUDE effects were significant.

Table 4.4 shows the means and significance levels resulting from the application of Dunnett's test (Winer, 1971) comparing the -6 dB and +6 dB condition with the control, for F1 vowel centres in the high pitch condition. Significant differences were found only for /U/ in the -6 dB condition, and for /o/ and /ʌ/ in the +6 dB condition.

Shifts in the +6 dB condition were generally larger (with a mean upward shift of 17.7 Hz) than shifts in the -6 dB condition (with a mean downward shift of 9.8 Hz). This asymmetry in the effects of raising or lowering the amplitude of higher formants is similar to that observed in Experiment 5.

A consistent trend toward higher F1 centres in the -6 dB condition and lower F1 centres in the +6 dB condition

	0 dB	-6 dB	+6 dB
/u/	475.9	481.1	470.8
/U/	531.1	568.3**	514.0
/ọ/	571.0	569.0	542.6**
/ɑ/	673.6	677.8	656.2
/ʌ/	653.5	657.9	633.2*

*p<.05 **p<.01

Table 4.4. Vowel centres: means and signifcance levels for Dunnett's test comparing amplitude conditions (-6 and +6 dB) with the control (0 dB) for F1 in the high pitch condition of Experiment 6

can be observed in Figure 4.26. Such a trend is absent in the F2 centres, although it is predicted by the centre of gravity hypothesis.

Ainsworth (1975) argued that vowel centres (estimated as the weighted means of the response profiles along the F1 and F2 scales) may not give accurate estimates of the effects of a variable on vowel identification if the complete response surface for a given vowel category is not represented in the stimulus continuum. This difficulty, which Ainsworth referred to as the "window problem" may be severe if the response centre occurs near the edges of the sampled stimulus continuum. The effects of fundamental frequency (also investigated by Ainsworth) provide an illustration of this problem. Figure 4.26 indicates that vowel centres for F1 and F2 shift to higher frequency values in most cases, from the low to high pitch conditions. The shifts are largest for the vowel /o/ whose estimated F1 and F2 centres lie near the midpoint of the continuum. F2 centres for /U/ and /ʌ/ are actually lower in the high pitch condition than in the low pitch condition. This may be because the 'actual' response centre occurs at a higher F2 value than the highest value sampled in the continuum (1150 Hz). Listeners may adjust the criteria by which they select a particular response category, resulting in range-frequency effects (Brady & Darwin, 1978).

Even though vowel centres may underestimate the effects of amplitude for the "peripheral" vowel categories, it is apparent that no vowel shows a very large or systematic effect of formant amplitude. There are some clear violations of centre of gravity predictions, and it appears that formant amplitude effects are not restricted to stimuli with closely spaced formants.

F1-F2 separation and the effects of formant amplitude. The effects of amplitude can be described in terms of a distance measure based on the deviations between the identification functions. Distance (between the -6 or +6 dB and the 0 dB condition) for the jth stimulus on the continuum, D_j, is defined as follows:

$$D_j = \Sigma_i \mid P_{ij} - Q_{ij} \mid$$

where P_{ij} is the number of responses to stimulus j for vowel response category i in the -6 or +6 dB conditions; and Q_{ij} is the number of responses to stimulus j for response category i in the 0 dB condition.

Figure 4.27 demonstrates the effects of the amplitude changes as a function of formant separation in Bark units. Distance D_j between the response profiles is taken as an index of the size of the formant amplitude effect on vowel identification profiles. The distances between -6 dB and 0 dB conditions are indicated with squares, while the distances between +6 dB and 0 dB conditions are indicated with triangles. The larger the effects of amplitude on the response profiles, the greater the expected distances will be.

Formant amplitude effects should only be present for stimuli with less than the critical separation of F1 and F2. A step-like decline is expected when the critical separation is exceeded. The dashed line in Figure 4.27 indicates a 3 Bark separation of the formants. It is apparent that the distances do not show a sudden decline, even in stimuli with more than 3.5 Bark separation. Amplitude appears to have a larger effect on stimuli with more than 3 Bark separation of the formants in the high pitch condition. There appears to be a systematic trend in the high pitch condition for distances to get progressively larger as the formant separation increases; amplitude appears to have its largest effects at the widest separations of F1 and F2, close to 5

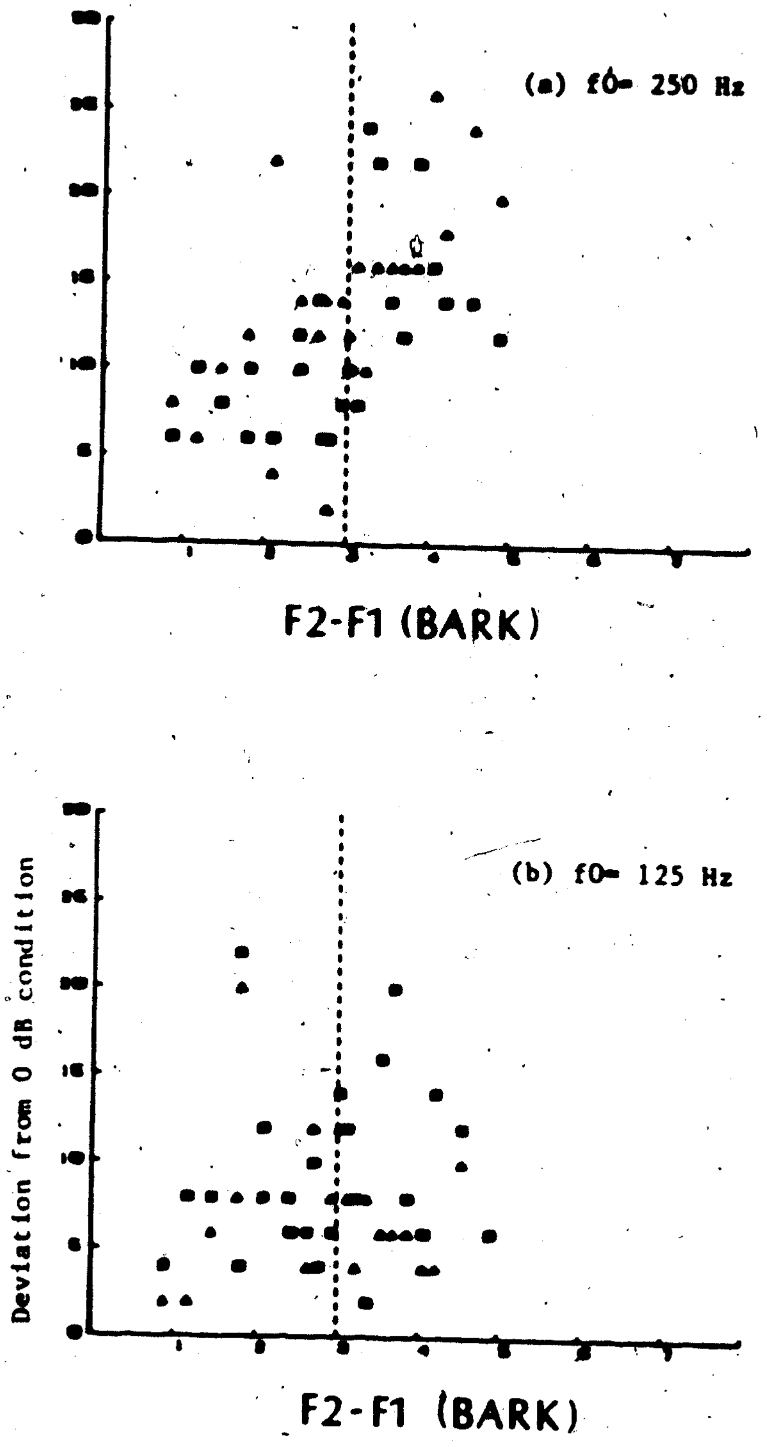


Figure 4.27 Experiment 6. Distances between response profiles. Sum of absolute deviations of -6 dB condition from 0 dB condition (□); sum of absolute deviations of +6 dB condition from 0 dB condition (▲).

Bark or about 700 Hz.

The size of the formant amplitude effect (expressed as the summed absolute deviations between the response profiles) is shown in Table 4.5 for each stimulus in the -6 and +6 dB conditions. The effects of amplitude on the response profiles are somewhat unsystematic, but several trends are apparent. Distances are larger in the high pitch condition, on the average, than in the low pitch condition. No such dependency of formant amplitude on fundamental frequency is postulated by the centre of gravity hypothesis of Chistovich and Lublinskaya (1979). The effects of amplitude are largest for stimuli with wide separation of the formants, and become increasingly smaller as the formants approach one another. There is thus little evidence that formant amplitude effects are restricted to stimuli with less than a critical separation of F1 and F2.

There is one possible objection which might be raised to the use of this measure of distance as an index of the size of amplitude effects on <u>individual</u> stimulus items. While a 6 dB change in formant amplitude may have approximately the same psychophysical effect on stimuli from two different positions along the vowel continuum, the changes in the identification profiles may differ depending on the proximity of each stimulus to a category boundary on the continuum. Larger shifts may be expected for stimuli occurring near vowel boundaries; and smaller shifts for

f_0=125 Hz

	F1=440	480	520	560	600	640	680	720 Hz	
F2=850	8	4	6	12	8	6	2	2	
1000	4	6	8	12	12	8	8	20	+6 dB
1150	6	10	4	6	6	4	8	4	
F2= 850	12	4	6	12	22	8	8	4	
1000	6	20	2	14	10	8	8	4	-6 dB
1150	6	12	14	8	16	8	6	6	

f_0=250 Hz

F1=	440	480	520	560	600	640	680	720 Hz	
F2=850	16	2	14	22	12	10	6	8	
1000	26	16	16	12	14	10	4	10	+6 dB
1150	20	24	18	16	16	10	14	12	
F2=850	8	6	10	6	6	8	10	6	
1000	16	12	22	10	6	12	6	10	-6 dB
1150 Hz	12	14	14	22	14	24	8	14	

Table 4.5. Amplitude effect (summed absolute deviations between the identification functions) in the +6 and 0 dB conditions, and between the -6 and 0 dB conditions, for Experiment 6

those located in 'plateau' (or stable) regions, where one category dominates over the others. If the proportion of stimuli from 'boundary regions' and 'plateau regions' is evenly distributed across the F1-F2 continuum and on either side of the F2 - F1 = 3 Bark line, then this effect may be negligible. However, if there are fewer stimuli from 'plateau regions' on one side or the other, the size of the estimated amplitude effect may appear to be larger even though in perceptual terms, the effects are similar.

One indicator of the relative proportion of stimuli from 'boundary regions' and 'plateau regions' is the percentage of responses to the dominant or modal response category. A stimulus near the boundary between two categories will generally have fewer responses to the dominant or modal category than a stimulus situated in a stable or plateau region.

Figure 4.28 indicates the percentage of responses to the modal category for individual stimuli, plotted as a function of the separation of F1 and F2 in Bark units. It is clear that the stimuli with less than 3 Bark separation of F1 and F2 are not characterized by smaller percentages of responses.

4.6.4 FCOG predictions

If F1 and F2 both exert an influence on the perception of back vowel differences by means of altering the centre of gravity of the formant cluster, then a projection of the response profiles along a scale of F_m, the estimated F1-F2 centre of gravity, should result in a closer alignment of the identification functions than a projection along either F1 or F2 or both.

The centre of gravity was estimated as the centroid of the maximum 3 Bark region of the excitation pattern, calculated according to the model of Moore and Glasberg (1983). Excitation patterns for a typical stimulus

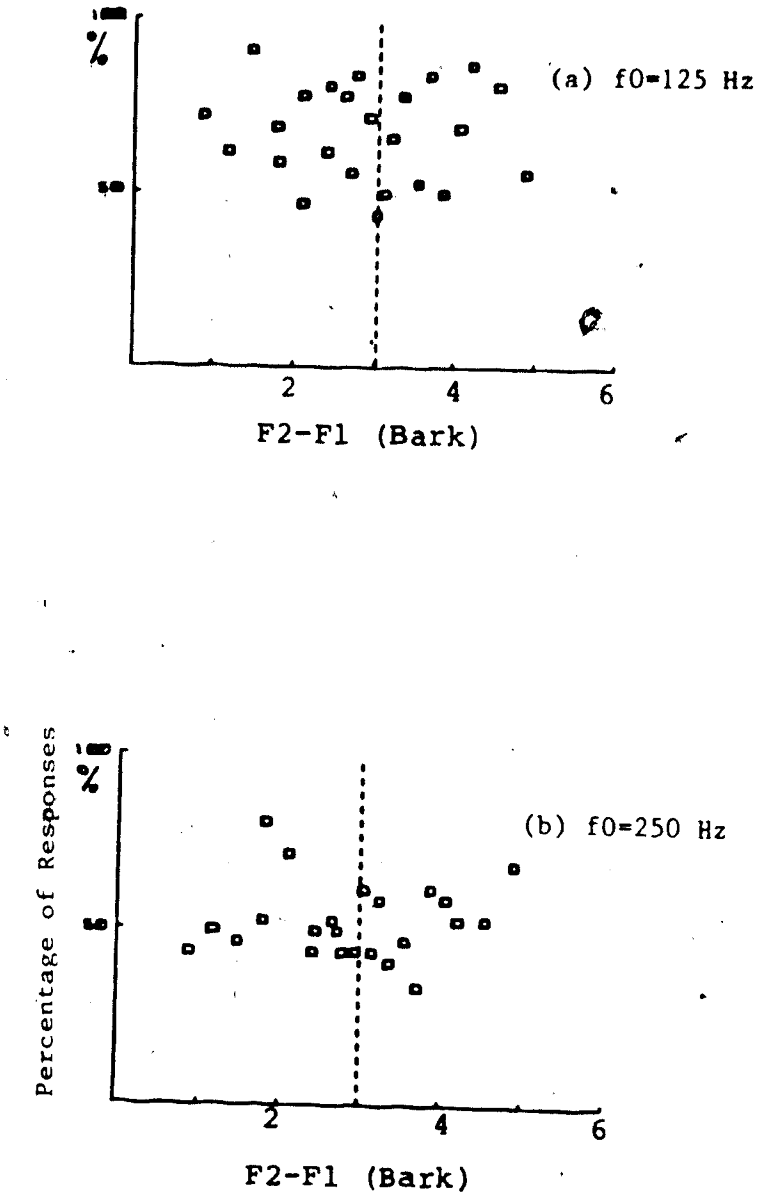


Figure 4.28 Experiment 6. Total number of responses to modal vowel category as a function of formant separation in Bark. Dashed line indicates F2-F1= 3 Bark

(f_0=250 Hz, F1=600 Hz, F2=1000 Hz) are shown for the three amplitude conditions in Figure 4.29. The dashed lines on each curve indicate the upper and lower cutoffs of the 3 Bark region with maximum summed intensity. F1 and F2 locations are indicated with arrows on the 0 dB plot. Note the abrupt change from the -6 to 0 dB conditions in the maximum 3 Bark region. In the -6 dB condition this region includes the second and third harmonics (note that F2 lies outside the 3 Bark interval), while in the 0 dB and +6 dB conditions it includes the third and fourth harmonics. As noted earlier, it is difficult to see how F1 and F2 locations could be estimated from these patterns.

Figure 4.30 indicates the centre of gravity locations (crosses) along with the first and second formant frequencies. The centres of gravity range between F1 and F2. For higher F1 and F2 values, F_m is displaced away from F1. In the high pitch condition, stimuli with F2=1000 Hz show a <u>decrease</u> in F_m as a function of increasing F1. This is because higher F1s have an increasingly larger effect on the centre of gravity, pulling F_m away from F2. These effects are fairly complex and appear to depend on the relative amplitudes of individual harmonics in the vicinity of F1 and F2.

Response plots for stimuli in the high pitch condition with less than 3 Bark separation of F1 and F2 are shown as a function of the nominal F1 and F_m, the frequency of the

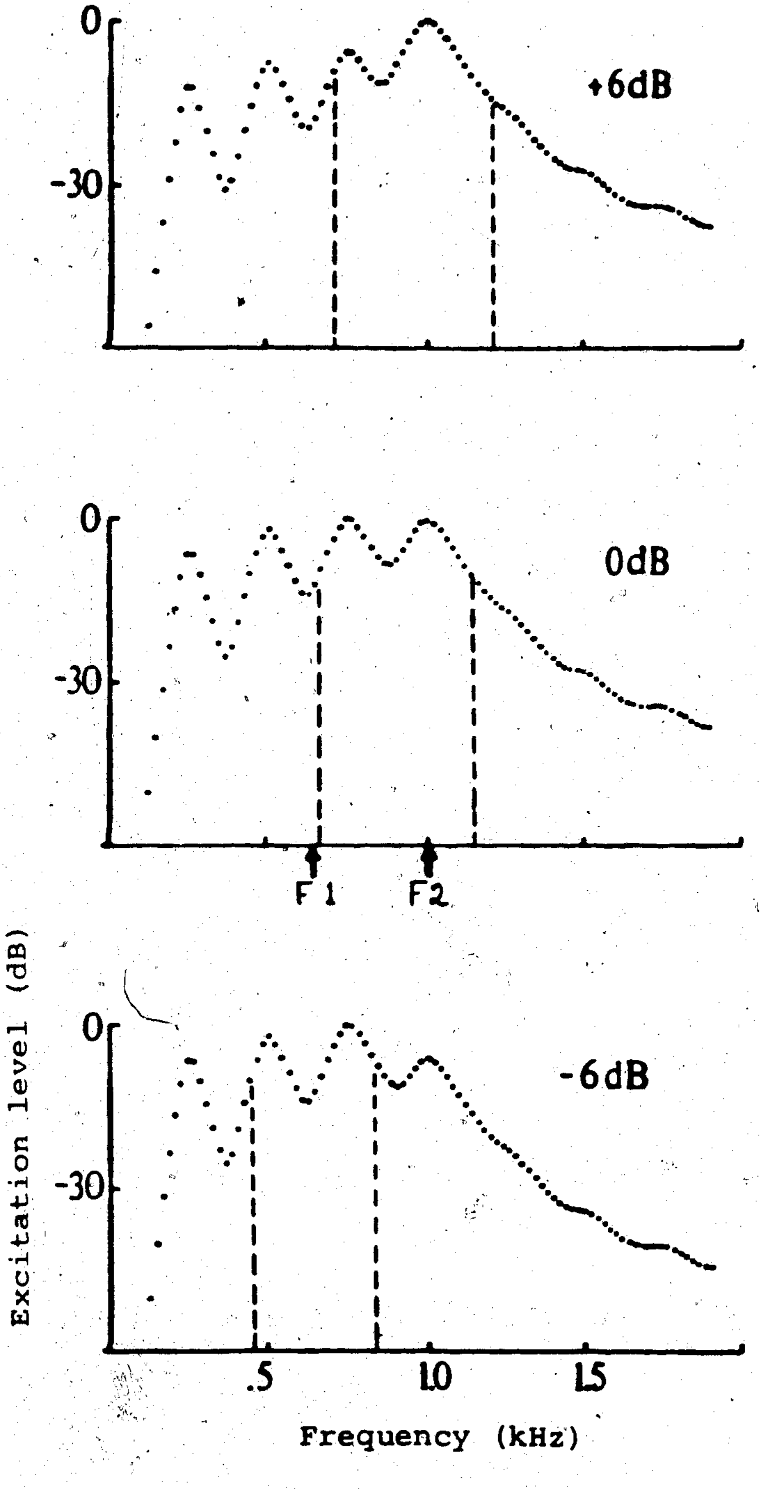
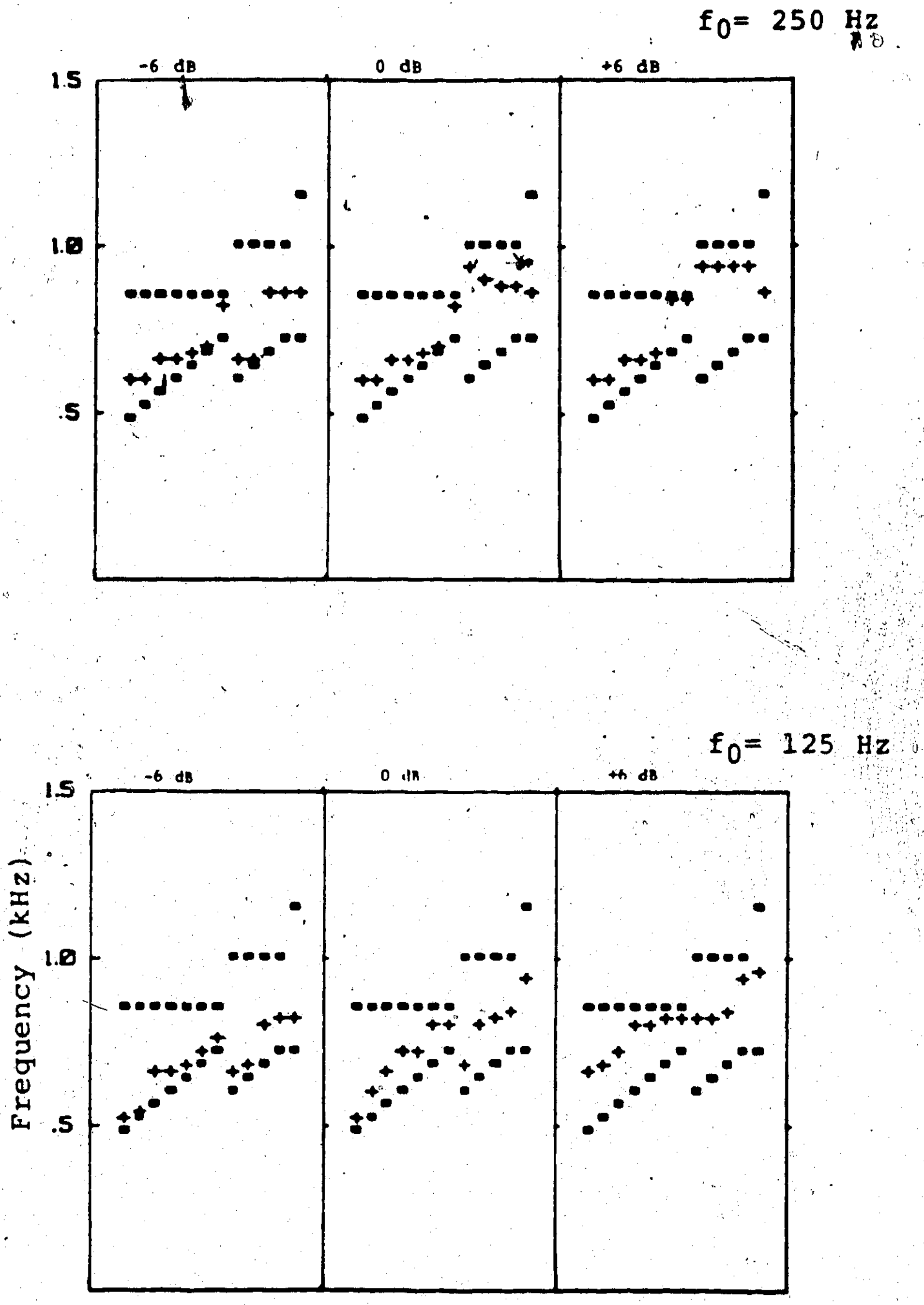


Figure 4.29 Excitation patterns based on the model of Moore and Glasberg (1983) for stimuli in each amplitude condition. f_0=250 Hz; F1=600 Hz; F2=1000 Hz. Dashed lines indicate maximum 3 Bark response region



(b)

Figure 4.30 F1 and F2 (bars) and measured centre of gravity, F_m (crosses), computed as the centroid of the maximum 3 Bark region of the excitation pattern based on the model of Moore and Glasberg (1983) for stimuli with less than 3 Bark separation of F1 and F2

estimated centre of gravity location (Figure 4.31). Included in this set are all 3 levels of amplitude and the 3 F2 conditions. The upper plots indicate the proportion of /o/ and /ɑ/ responses as a function of F1. Squares indicate stimuli in the -6 dB condition, while triangles represent the +6 dB condition. The data are quite variable, particularly for the high F1 values. The bottom plots represent /o/ and /ɑ/ responses as a function of F_m. Even greater scatter in the identification data can be observed, and some non-monotonicities in the response curves are present, especially at high F_m values for the vowel /o/. Clearly, this measure cannot substitute for F1 and F2, nor does it appear to characterize the effects of amplitude changes accurately.

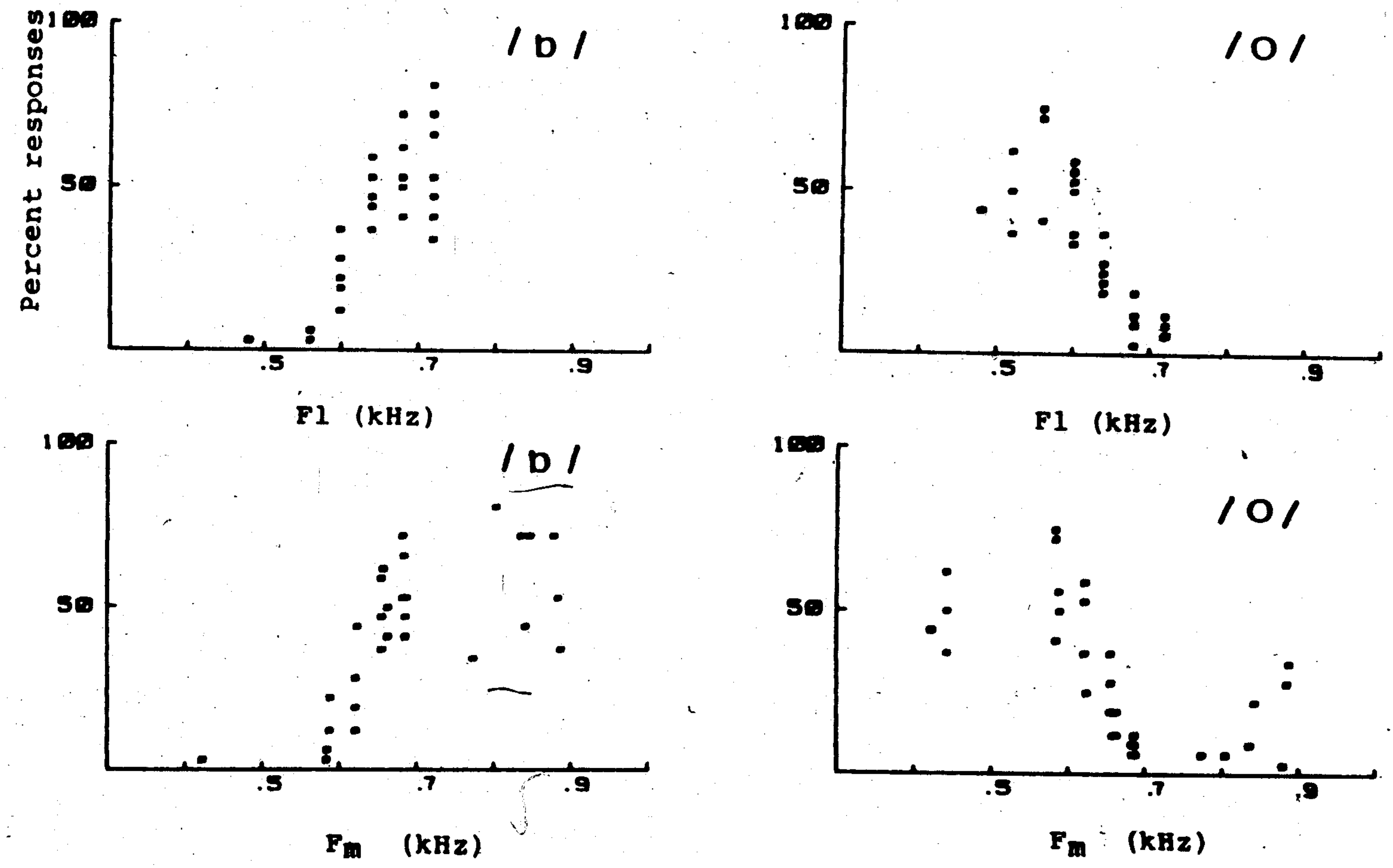


Figure 4.31 Percentage of /o/ and /ɒ/ responses, plotted as a function of nominal F1 (top) and F_m , measured centre of gravity computed as the centroid of the maximum 3 Bark region of the excitation pattern based on the model of Moore and Glasberg (1983)

4.7 Summary and conclusions

This chapter investigated the hypothesis advanced by Chistovich and Lublinskaya (1979) that back vowels (with a small separation of F1 and F2) are differentiated perceptually by the frequency location of the centre of gravity (FCOG) of the F1-F2 cluster. Two aspects of their FCOG hypothesis were investigated in this chapter: the effects of formant amplitude and the effects of frequency separation of F1 and F2.

Experiment 5 employed a matching task in which reference and matching stimuli were both multiformant back vowels, with a fixed, 250 Hz separation of F1 and F2 in all stimuli. Listeners adjusted both F1 and F2 simultaneously as a function of the relative amplitudes of F1 and higher formants. Significantly lower frequency matches were not selected when the higher formants of the reference were attenuated by either 10 or 20 dB, even though this operation lowered the centre of gravity. Boosting the amplitudes of the higher formants did result in upward shifts in F1-F2 matches; however, the shifts were not continuous. In the +10 dB condition, matching histograms showed a large amount of variability, while in the +20 dB condition matched F1 values were generally positioned very close to F2 of the reference stimulus. The centre of gravity therefore cannot be the sole determinant of back vowel matching judgements in multiformant stimuli with close spacing of F1 and F2.

Two possible alternative explanations for the shifts in the +10 and +20 dB conditions were considered. Matching may be based on the alignment of detectable formant peaks; when F2 and higher formants are increased in amplitude, F1 may be masked. A second possibility is that both formant peak locations and their amplitudes are perceptually important, but that formant amplitude has the role of a secondary variable which is only of importance when F1 is much higher in amplitude than F2 or vice versa. Large asymmetries in formant amplitude may result in a matching strategy of aligning the largest peak in the matching stimulus with the largest peak in the reference.

A second aspect of the centre of gravity hypothesis proposed by Chistovich and Lublinskaya (1979), the concept of critical distance, was assessed in Experiment 6 by means of a vowel identification task. Chistovich and Lublinskaya postulated that changes in the relative amplitudes of F1 and F2 are phonetically equivalent to changes in the frequency of a single formant, provided that F1 and F2 were less than 3-3.5 Bark apart. They found that stimuli with more than 3-3.5 Bark separation of F1 and F2 could not be matched in phonetic quality with a single formant, and changes in formant amplitude did not have a systematic effect on matching.

Experiment 6 set out to determine whether formant amplitude has a systematic effect on vowel identification, and whether this effect is restricted to stimuli with closely spaced formants. The stimuli formed a continuum in F1-F2 in the back/central region with a range of formant separations spanning both sides of the F2-F1=3 Bark line. The amplitudes of F2 and higher formants were either attenuated or boosted by 6 dB, relative to a control condition. Two sets of stimuli were used: a high pitch condition (f_0=250 Hz) and a low pitch condition (f_0=125 Hz).

Formant amplitude changes resulted in small shifts in some of the identification functions. However, no compelling evidence was found for a critical distance constraint. Systematic effects of formant amplitude were found only in the high pitch condition. Rather than exclusively a displacement of the identification functions along the F1 or F2 axes, however, formant amplitude changes frequently altered the heights of the identification curves, which suggests that its effects are not equivalent to formant frequency changes. A larger effect of formant amplitude was found in the +6 dB condition than in the -6 dB condition, consistent with the findings of Experiment 5. Significant changes were observed in the proportion of responses to each vowel category as a function of formant amplitude. With increasing higher formant amplitude, there were more /U/, /ɑ/ and /ʌ/ responses but fewer /u/ and /o/ responses.

The size of the formant amplitude effect on vowel identification was assessed in terms of the sum of the absolute deviations between response profiles. The effects of formant amplitude were not reduced in stimuli with more than 3-3.5 Bark separation of F1 and F2. In fact, there was a tendency in the high pitch condition for stimuli with more than 3.5 Bark separation to show higher deviation scores. The largest effects were found for stimuli with close to 5 Bark separation of the formants. Predictions of the identification data based on estimates of the centre of gravity did not lead to encouraging results. Plotting responses as a function of the estimated centre of gravity location did not reduce the scatter in the identification data.

These findings, taken together, provide further indications that the centre of gravity hypothesis does not successfully account for phonetic differences between back vowel stimuli. The effects of formant amplitude on vowel identification were not limited to stimuli with less than 3-3.5 Bark separation of F1 and F2. The nature of the shifts observed in this experiment suggests that this formant amplitude has a fairly small, context-dependent influence on vowel quality. Since formant frequency and formant amplitude are correlated in natural speech (Fant, 1956), the latter may function as a secondary cue to vowel identity. The perceptual system may attend to relative formant amplitudes to determine the locations of formant peaks; e.g. the relatively low spectral energy above the F1-F2 cluster may indicate a back vowel, and the degree of asymmetry in spectrum levels in the region of the F1-F2 cluster may provide a clue to the locations of F1 and F2. The present experiments suggest, however, that formant amplitude effects are fairly small, and are generally overridden by changes in formant frequency.

It is of interest that only the high pitch condition of Experiment 6 indicated substantial effects of formant amplitude. One possibility is that formant amplitude serves as a supplementary cue when the vowel spectrum is not well specified, as in stimuli with high fundamental frequencies.

Natural back vowels differ in several respects from single formant stimuli, some of which may be perceptually important. Although two distinct peaks do not necessarily appear in the line spectrum when F2 is close to F1 there are other potential cues, such as the overall spectral shape and the slope of the spectrum above F2, which may convey phonetic information. Further psychophysical experiments are needed to evaluate the contribution of these additional cues.

CHAPTER FIVE

SUMMARY AND CONCLUSIONS

5.1 Introduction

The present study has investigated several aspects of the auditory pattern recognition process which enable a listener to (1) assign a sound stimulus to a vowel class (using vowel identification responses); and (2) judge the relative similarity of pairs of stimuli in terms of perceived vowel quality (using a vowel matching task). The close correspondence between the perceptual dimensions of vowel quality and the frequencies of the lowest 2 or 3 formants suggests the possibility that the auditory system may be able to infer the frequencies of vocal tract resonances from the acoustic signal. Several problems stemming from the formant hypothesis of vowel perception were investigated in the present study.

The overall pattern of results was generally compatible with the formant hypothesis, although it seems likely that some characteristics of overall spectral shape also contribute to vowel perception under certain conditions. It was found that predictions of the experimental data based on excitation and loudness pattern models did not offer a significant advantage over predictions based on preemphasized power spectra. The findings are consistent with the results of Carlson and Granstrom (1979) indicating that more detailed models (possibly incorporating a temporal coding scheme, as suggested by recent neurophysiological studies) may be needed.

5.2 Perception of front vowel contrasts

The perception of height in front vowels (associated with changes in F1) appears to depend on a very narrow spectral region around the F1 peak. The matching experiments reported in chapter three indicated that vowel height is largely determined by the two most prominent harmonics in the F1 region. It was found that the contribution of these two harmonics to perceived height was not equal: attenuation of the higher frequency member of the pair had a larger effect. Additional harmonics in the F1 region appeared to have only a marginal effect.

The asymmetry in the contribution of harmonics is consistent with the findings of Darwin (1984) who reported that an increment in the level of the higher frequency member of a pair of harmonics near the F1 peak shifted vowel identification boundaries to a greater extent than an increment in the lower frequency component. Although this asymmetry may be due to auditory factors such as masking or suppression, it may also reflect listeners' implicit "knowledge" of constraints on the shape of the spectrum in the first formant region.

The matching data were closely predicted by a model which computes the weighted mean or centre of gravity of the two most prominent harmonics in the first formant region of the preemphasized spectrum. The contribution of harmonics was largely independent of harmonic rank (harmonics 1 to 4); parallel results were found for both low pitched (f_0=125 Hz) and high pitched (f_0=250 Hz) vowels. It was found that the best predictors of the matching data also produced the best estimates of the nominal F1 values of the matching stimuli.

The effects of deleting components from the F1 region of the spectrum was studied in an identification experiment, which demonstrated that front vowel identification is largely unaffected by the removal of all but the two most prominent harmonics. There were, however, some shifts in the identification functions indicating that additional F1 components may be perceptually relevant. These shifts were not consistent with local centre of gravity predictions: removal of all components either above or below the two most prominent had a similar effect on the identification boundaries, suggesting that factors such as overall spectral balance may provide a secondary cue in the perception of vowels.

5.3 Perception of back vowel contrasts

In back vowels, F1 and F2 may be very close together with only a single peak in the line spectrum. However, the presence of two closely spaced formants in back vowels may be signaled in other ways, for example by the relative density of energy in the low and high frequency regions, by the slope of the spectrum above the F1-F2 region of prominence, and by the presence of a small shoulder on the high frequency slope of the spectrum envelope above F1.

The experimental studies reported in chapter four were designed to evaluate the formant centre of gravity (FCOG) hypothesis advanced by Delattre, Liberman, Cooper and Gerstman (1952) and recently supported by Chistovich and Lublinskaya (1979). According to this hypothesis, F1 and F2 in close proximity combine to form a single "effective" peak; vowel quality is then determined by the centre of gravity of the formant cluster. On the basis of their matching data, Chistovich and Lublinskaya estimated that these effects are only present when the formant separation is less than 3-3.5 Bark.

The results of a matching experiment using multiformant stimuli did not support the FCOG hypothesis: the formant frequency - formant amplitude tradeoff predicted by the centre of gravity hypothesis was not found. When F2 and higher formants were attenuated by as much as 20 dB, no shifts in matching were observed. An increase in formant amplitudes did produce shifts, but the large variability associated with the matching profiles for some of the stimuli suggested a conflict between formant frequency and formant amplitude cues, rather than an integration effect.

The effects of formant amplitude were further investigated in a vowel identification experiment using an F1-F2 continuum with a wide range of separations of the two formants. A 6 dB increase or decrease in the amplitudes of F2 and higher formants resulted in small shifts in some of the identification functions. These effects were present only in a high pitch condition, and did not follow the predictions of a formant centre of gravity model. Formant amplitude effects on vowel identification were not restricted to closely spaced formants. In fact, the largest effects appeared for stimuli with the widest separations of F1 and F2 on the continuum.

The results of these two experiments are inconsistent with the formant centre of gravity hypothesis. Both experiments indicated, however, that formant amplitude may have an effect on vowel quality under certain conditions. Since formant amplitudes can be predicted from a knowledge of formant frequencies (Fant, 1956) the auditory system may be able to exploit these relationships in attending to vowel sounds. The presence of F1 and F2 may be signalled by other means when the formants are close together; further research is needed to determine which cues are exploited by the auditory system.

5.4 Directions for future research

These studies suggest a number of possible avenues for further research. It would be useful to investigate the contributions of individual harmonics in the first formant region of front vowels using a wider range of fundamental frequency values. It may be expected that vowel quality may be poorly defined with high fundamental frequency values; to what extent can these limits be characterized by the models investigated in this study? Is there a change in performance for stimuli with fundamental frequencies smaller than the critical bandwidth?

The vowel matching data of Experiment 3 suggest that the effects of attenuation are nearly constant when the fundamental frequency is raised by one octave. Several possible explanations for this finding may be considered, including a pitch-matched weighting of spectrum levels. Such a procedure is also suggested by the data on the perceptual separation of simultaneous voices (c.f. Darwin, 1983). It is therefore of considerable interest to determine the contribution of inharmonic components. Are inharmonic components 'fused' with other vowel harmonics, or are they excluded from phonetic quality judgements?

Further data are needed to determine the perceptual basis for the asymmetries observed in front vowel matching. It would be interesting to determine whether these asymmetries become less pronounced as the frequency of the second formant is lowered, as suggested by the results of Experiment 5. Do these differences depend on relationships which are inferred by listeners on the basis of spectral properties of natural vowels, and if so, how are these spectral properties encoded by the auditory system?

At present, a great deal remains to be learned about the perceptual processing of spectral information in back vowels, where the close spacing of F1 and F2 presents a problem for the formant hypothesis of vowel perception. The evidence presented in chapter four did not support the alternative hypothesis proposed by Chistovich and Lublinskaya (1979) that energy is integrated over a broad region of the spectrum. It was suggested that listeners may be able to recover formant frequency information from other cues such as relative formant amplitudes or spectral balance, or the slope of the spectrum above the F1-F2 cluster in back vowels.

Further experimentation is needed to determine whether these cues are employed in the perception of back vowels. Methods can be devised to test whether these cues have an independent contribution, or whether they help to specify the locations of the formants.

REFERENCES

Ainsworth, W. A. 1972. Duration as a cue in the recognition of synthetic vowels. *J. Acoust. Soc. Am. 51*: 648-651.

Ainsworth, W. A. 1974. Influence of f0 on perceived English vowel boundaries. *Speech Communication Seminars 3*: 123-129.

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

Ainsworth, W. A. 1981. Perception of brief, split-formant vowel sounds. *Acustica 48*: 254-259.

Ainsworth, W. A. and Millar, J.B. 1972. The effect of relative formant amplitude on the perceived identity of synthetic vowels. *Language and Speech 15*: 328-341.

Assmann, P. F. 1979. The role of context in vowel perception. Unpublished M.Sc. thesis, Department of Linguistics, University of Alberta.

Assmann, P. F., Nearey, T. M. and Hogan, J. T. 1982. Vowel identification: Orthographic, perceptual and acoustic aspects. *J. Acoust. Soc. Am. 71*: 975-989.

Atal, B. S. and Remde, J.R. 1982. A new model of LPC excitation for producing natural-sounding speech at low bit rates. *Proc. 1982 Int. Conf. Acoust., Speech, Signal Processing*, pp. 614-617.

Beddor, P. 1984. Formant integration and the perception of nasal vowel height. *Haskins Laboratories Status Report SR-77/78*:

Beddor, P. and Hawkins, S. 1984. The 'centre of gravity' and perceived vowel height. *J. Acoust. Soc. Am. 75*: S86a.

Bedrov, Y.A., Chistovich, L.A., and Sheikin, R.L. 1978. Frequency position of the "centre of gravity" of formants as a useful feature in vowel perception. *Soviet Physics-Acoustics 24*: 275-281.

Bekesy, G. von 1960. *Experiments in hearing*. McGraw

Hill: New York.

Bell, C.G., Fujisaki, H., Heinz, J.M., Stevens, K.N. and House, A.S. 1961. Reduction of speech by analysis-by-synthesis techniques. *J. Acoust. Soc. Am. 33*: 1725-1736.

Bennett, D.C. 1968. Spectral form and duration cues in the recognition of English and German vowels. *Language and Speech 11*: 65-85.

Bernstein, J.C. 1981. Formant-based representation of auditory similarity among vowel-like sounds. *J. Acoust. Soc. Am. 69*: 1132-1144.

Bismarck, G. von 1973. Vorschlag fur ein einfaches Verfahren zur Klassifikation stationarer sprachchalle. *Acustica 28*: 186-188.

Bladon, R.A.W. 1983. Two-formant models of vowel perception: shortcomings and enhancements. *Speech Communication 2*: 305-314.

Bladon, R.A.W. and Fant, G. 1978. A two-formant model and the cardinal vowels. *STL-QPSR 1/1978*: 1-8.

Bladon, R.A.W. and Lindblom, B. 1981. Modeling the judgement of vowel quality differences. *J. Acoust. Soc. Am. 69*: 1414-1422.

Blomberg, M., Carlson, R., Elenius, K., and Granstrom, B. 1982. Experiments with auditory models in speech perception. In *The representation of speech in the peripheral auditory system*. R. Carlson and B. Granstrom, eds. ELsevier, Amsterdam, pp. 197-201.

Blomberg, M., Carlson, R., Elenius, K., and Granstrom, B. 1983. Auditory models and isolated word recognition. *STL-QPSR 4/1983*: 1-15.

Brady, S.A. and Darwin, C.J. 1978. Range effect in the perception of voicing. *J. Acoust. Soc. Am. 51*: 483-502.

Brokx, J.P.L. and Nooteboom, S.G. 1982. Intonation and the perceptual separation of simultaneous voices. *J. Phonetics 10*: 23-26.

Brugge, J., Anderson, D., Hind, J. and Rose, J. 1969. Time structure discharges in single auditory nerve fibres of the squirrel monkey in response to complex sounds. *J. Neurophysiology 32*: 386-401.

Carlson, R. and Granstrom, B. 1976. Detectability of changes of level and spectral slope in vowels. *STL-QPSR*

2-3/1976: 1-4.

Carlson, R. and Granstrom, B. 1979. Model predictions of vowel dissimilarity. STL-QPSR 3-4/1979: 84-104.

Carlson, R., Fant, G. and Granstrom, B. 1975. Two-formant models, pitch and vowel perception. In Auditory analysis and perception of speech. G. Fant and M.A.A. Tatham eds. Academic Press: London, pp. 55-82.

Carlson, R., Granstrom, B. and Fant, G. 1970. Some studies concerning the perception of isolated vowels. STL-QPSR 2/3/1970: 19-35.

Carlson, R., Granstrom, B. and Klatt, D. H. 1979. Vowel perception: The relative perceptual salience of selected acoustic manipulations. STL-QPSR 3-4/1979:73-83.

Cherry, C. 1965. On human communication. MIT Press: Cambridge, Mass, Second edition,

Chistovich, L.A. 1971. Auditory processing of speech stimuli - evidences from psychoacoustics and neurophysiology. Proc. 7th Int. Congress Phonetic Sciences I: 83-92.

Chistovich, L.A. 1985. Central auditory processing of peripheral vowel spectra. J. Acoust. Soc. Am. 77: 789-805.

Chistovich, L.A. and Lublinskaya, V.V. 1979. The 'centre of gravity' effect in vowel spectra and critical distance between the formants: Psychoacoustic study of the perception of vowel-like stimuli. Hearing Research 1: 185-195.

Chistovich, L.A., Sheikin, R.L. and Lublinskaya, V.V. 1979. "Centres of gravity" and spectral peaks as the determinants of vowel quality. In Frontiers of speech communication research. B. Lindblom and S. Ohman eds. Academic Press: London, pp. 143-157.

Chistovich, L.A., Kozhevnikov, V.A., Lesogar, L.W., Shupljakov, V.S., Taljasin, P.A. and Tjulkov, W.A. 1974. A functional model of signal processing in the peripheral auditory system. Acustica 31: 349-353.

Christensen, R., Strong, W. and Palmer, E. 1976. A comparison of three methods of extracting resonance information from predictor-coefficient coded speech. IEEE Trans. Acoust., Speech, Signal Processing, vol ASSP-24: 8-14.

Cohen, A. 1974. Formant discrimination in the auditory system. *Speech Communication Sem.* , Stockholm, pp. 111-116.

Coker, C.H. 1965. Real-time formant vocoder using a filter bank, a general purpose computer and an analog synthesizer. *J. Acoust. Soc. Am. 38*: 940a.

Darwin, C.J. 1981. Perceptual grouping of speech components differing in fundamental frequency and onset time. *Quart. J. Exp. Psychology 36*: 193-208.

Darwin, C.J. 1983. Auditory processing and speech perception. In *Attention and Performance X*. H. Bouma and D.G. Bouwhuis, eds. Erlbaum: Hillsdale, N.J.

Darwin, C.J. 1984. Perceiving vowels in the presence of another sound: Constraints on formant perception. *J. Acoust. Soc. Am. 76*: 1636-1647.

Darwin, C.J. and Gardner, R.B. In press. Which harmonics contribute to the estimation of first formant frequency? *Speech Communication*

Delattre, P., Liberman, A.M., Cooper, F.S., and Gerstman, L. 1952. An experimental study of the acoustic determinants of vowel colour. *Word 8*: 195-210.

Delgutte, B. 1984. Speech coding in the auditory nerve: I. Vowel-like sounds. *J. Acoust. Soc. Am. 75*: 879-886.

Delgutte, B. and Kiang, N.Y.S. 1984a. Speech coding in the auditory nerve: I. Vowel-like sounds. *J. Acoust. Soc. Am. 75*: 866-878.

Duifhuis, H., Willems, L.F. and Sluyter, R.J. 1982. Measurement of pitch in speech: An implementation of Goldstein's theory of pitch perception. *J. Acoust. Soc. Am. 71*: 1568-1580.

Egan, J.P. and Hake, H.W. 1950. On the masking pattern of a simple auditory stimulus. *J. Acoust. Soc. Am. 22*: 622-630.

Evans, E.F. 1982. Representation of complex sounds at cochlear nerve and cochlear nucleus. In *The representation of speech in the peripheral auditory system*. R. Carlson and B. Granstrom, eds. Elsevier: Amsterdam, pp. 27-42.

Evans, E.F. and Wilson, J.P. 1973. The frequency selectivity of the cochlea. In *Basic mechanisms in hearing*. A.R. Moller, ed. Academic Press: New York pp. 519-551.

Fairbanks, G. and Grubb, P. 1961. A psychophysical investigation of vowel formants. J. Speech Hearing Res. 4: 203-219.

Fant, G. 1956. On the predictability of formant levels and spectrum envelopes from formant frequencies. In For Roman Jakobson. M. Halle, ed. Mouton: The Hague, pp. 109-120.

Fant, G. 1960. Acoustic theory of speech production. Mouton: The Hague.

Fant, G. 1973. Speech sounds and features. MIT Press: Cambridge, Mass.

Fant, G. 1983. Feature analysis of Swedish vowels - a revisit. STL-QPSR 2-3/1983: 1-19.

Fant, G. and Risberg, A. 1963. Auditory matching of vowels with two-formant synthetic sounds. STL QPSR 4/1963: 7-11.

Fidell, S., Horonjeff, R., Teffeteller, S. and Green, D.M. 1983. Effective masking bandwidths at low frequencies. J. Acoust. Soc. Am. 73: 628-638.

Finney, D.J. 1971. Probit analysis. Third edition, Cambridge Univ. Press: Cambridge.

Flanagan, J.L. 1955. A difference limen for vowel formant frequency. J. Acoust. Soc. Am. 27: 613-617.

Flanagan, J.L. 1972. Speech analysis, synthesis and perception. Second edition, Springer Verlag: Berlin.

Fletcher, H. 1940. Auditory patterns. Review of Modern Physics 12: 47-65.

Fox, R.A. 1983. Perceptual structure of monophthongs and diphthongs in English. Language and Speech 26: 21-60.

Fujimura, O. 1967. On the second spectral peak of front vowels: A perceptual study of the role of the second and third formants. Language and Speech 10: 181-193.

Fujisaki, H. and Kawashima, T. 1968. The roles of pitch and higher formants in the perception of vowels. IEEE Trans AU-16: 73-77.

Gauffin, J. and Sundberg, J. 1982. Amplitude of the voice source fundamental and the intelligibility of super pitch vowels. In The representation of speech in the peripheral auditory system. R. Carlson and B. Granstrom eds. Elsevier: Amsterdam, pp. 223-228.

Gerstman, L.J. 1968. Classification of self-normalized vowels. IEEE Trans AU-16: 78-80.

Glasberg, B.R., Moore, B.C.J. and Nimmo-Smith, I. 1984. Comparison of auditory filter shapes derived with three different maskers. J. Acoust. Soc. Am. 75: 536-544.

Haggard, M.P. 1978. Mechanisms of formant frequency discrimination. In Psychophysics and physiology of hearing. E.F. Evans and J.P. Wilson, eds. Academic Press: London pp. 499-507.

Hanson, G. 1967. Dimensions in speech perception: An experimental study of vowel pereption. Ericsson Technics 23: 3-175.

Hawkins, J.E. and Stevens, S.S. 1950. The masking of pure tones and of speech by white noise. J. Acoust. Soc. Am. 22: 6-13.

Helmholtz, H.L.F. von 1954. On the sensations of tone. Dover: New York, English translation of 1877 edition.

Hermansky, H., Hanson, B. and Wakita, H. (in press). Critical evaluation of the PLP method. Speech Communication

Hess, W.J. 1983. Pitch determination of speech signals: Algorithms and devices. Springer-Verlag: Berlin.

Holmes, J.N. 1983. Formant synthesizers: Cascade or parallel? Speech Communication 2: 251-273.

Hose, B., Langner, G. and Scheich, H. 1983. Linear phoneme boundaries for German synthetic two-formant vowels. Hearing Research 9: 13-25.

Houtgast, T. 1972. Psychophysical evidence for lateral inhibition in hearing. J. Acoust. Soc. Am. 51: 1885-1894.

Houtgast, T. 1974a. Lateral suppression in hearing. A psychophysical study on the ear's capability to preserve and enhance spectral contrasts. Ph.D. Dissertation, Vrije Universiteit, Amsterdam. Academische Pers B.V.: Amsterdam.

Houtgast, T. 1974b. Auditory analysis of vowel-like sounds. Acustica 31: 320-324.

Houtgast, T. 1977. Auditory filter characteristics derived from rippled noise. J. Acoust. Soc. Am. 61:

Huggins, W.H. and Licklider, J.C.R. 1951. Place mechanisms of auditory frequency analysis. *J. Acoust. Soc. Am. 23*: 290.

Jones, D. 1956. *An outline of English phonetics.* Heffer: Cambridge, Eighth edition,

Joos, M. 1948. *Acoustic phonetics. Language Suppl. 24*: 1-136.

Kakusho, O., Hirato, H., Kato, K., and Kobayashi, T. 1971. Some experiments of vowel perception by harmonic synthesizer. *Acustica 24*: 179-190.

Kakusho, O., Nakashima, H., Yanagida, M, and Mizoguchi, R. 1983. Perceptual prominence of harmonic components in vowel-like sounds. *Acustica 53*: 132-142.

Karnickaya, E.G., Mushnikov, V.N., Slepokurova, N.A. and Zhukov, S.J. 1975. Auditory processing of steady state vowels. In *Auditory analysis and perception of speech.* G. Fant and M.A.A. Tatham eds. Academic Press: London pp. 37-53.

Kiang, N.Y.-S. and Moxon, E.C. 1974. "Tails of tuning curves" of auditory-nerve fibres. *J. Acoust. Soc. Am. 55*: 620-630.

Klatt, D.H. 1980. Software for a cascade/parallel formant synthesizer. *J. Acoust. Soc. Am. 67*: 971-995.

Klatt, D.H. 1982a. Prediction of perceived phonetic distance from critical band spectra: A first step. *IEEE Acoust., Speech, Signal Processing, vol ASSP-30*: 1278-1281.

Klatt, D.H. 1982b. Speech processing strategies based on auditory models. In *The representation of speech in the peripheral auditory system.* R. Carlson and B. Granstrom, eds. Elsevier: Amsterdam, pp. 181-196.

Klein, W., Plomp, R., and Pols, L.C.W. 1970. Vowel spectra, vowel spaces, and vowel identification. *J. Acoust. Soc. Am. 48*: 999-1009.

Koenig, W., Dunn, H.K. and Lacy, L.Y. 1946. The sound spectrograph. *J. Acoust. Soc. Am. 17*: 19-49.

Koopmans-van Beinum, F.J. 1980. Vowel contrast reduction: An acoustic and perceptual study of Dutch vowels in various speech conditions. Ph.D. Dissertation, University of Amsterdam Academische Pers B.V.: Amsterdam

Ladefoged, P. and Broadbent, D. 1957. Information conveyed by vowels. J. Acoust. Soc. Am. 29: 98-104.

Ladefoged, P. 1975. A course in phonetics. Harcourt, Brace and Jovanovich: New York.

Ladefoged, P., DeClerk, J., Lindau, M., and Papçun, G. 1972. An auditory-motor theory of speech production. UCLA Working Papers in Phonetics 22: 48-75.

Lehiste, I. and Meltzer, D. 1973. Vowel and speaker recognition in natural and synthetic speech. Language and Speech 16: 356-364.

Liberman, A.M. and Studdert-Kennedy, M. 1978. Phonetic perception. In Handbook of Sensory Physiology Vol. VIII: Perception. Springer-Verlag: Berlin, pp.143-178.

Licklider, J.C.R. 1952. On the processes of speech perception. J. Acoust. Soc. Am. 24: 590-594.

Licklider, J.C.R. 1957. Effects of changes in the phase pattern upon the sound of 16-harmonic complexes. J. Acoust. Soc. Am. 29: 780a.

Lieberman, P. 1976. Phonetic features and physiology: A reappraisal. J. Phonetics 4: 91-112.

Lindquist, J. and Pauli, S. 1968. The role of relative spectrum levels in vowel perception. Proc. 6th Int. Cong. Acoustics, pp. B91-B94.

Lloyd, R.J. 1896. The genesis of vowels. J. Anatomy and Physiology 31: 234-251.

McCandless, S.S. 1974. An algorithm for automatic formant extraction using linear prediction spectra. IEEE Trans Acoust., Speech, Signal Processing, vol ASSP 22: 135-141.

Makhoul, J. 1975a. Linear prediction: A tutorial review. Proc. IEEE Acoust., Speech, Signal Processing, vol ASSP 63: 561-580.

Makhoul, J. 1975b. Spectral linear prediction: Properties and applications. IEEE Trans Acoust., Speech, Signal Processing, vol ASSP 23: 283-295.

Makhoul, J., Viswanathan, R., Schwartz, R. and Huggins, A.W.F. 1978. A mixed-source model for speech compression and synthesis. J. Acoust. Soc. Am. 64: 1577-1581.

Markel, J.D. 1972. Digital inverse filtering - a new tool for formant trajectory estimation. IEEE Trans Audio Electroacoustics AU-20: 129-137.

Markel, J.D. and Gray, A.H. 1976. Linear prediction of speech. Springer-Verlag: Berlin.

Miller, R.L. 1953. Auditory tests with synthetic vowels. J. Acoust. Soc. Am. 25: 114-121.

Moore, B.C.J. 1982. An introduction to the psychology of hearing. Second edition, Academic Press: London.

Moore, B.C.J. and Glasberg, B.R. 1983a. Masking patterns for synthetic vowels in simultaneous and forward masking. J. Acoust. Soc. Am. 73: 906-917.

Moore, B.C.J. and Glasberg, B.R. 1983b. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74: 750-753.

Moore, B.C.J. and Glasberg, B.R. 1981. Auditory filter shapes derived in simultaneous and forward masking. J. Acoust. Soc. Am. 70: 1003-1014.

Mushnikov, V.N. and Chistovich, L.A. 1971. Method for the experimental investigation of component loudnesses in the recognition of a vowel. Soviet Physics - Acoustics 17: 339-344.

Mushnikov, V.N. and Chistovich, L.A. 1973. Experimental testing of the band hypothesis of vowel perception. Soviet Physics - Acoustics 19: 250-254.

Mushnikov, V.N., Slepokurova, N.A. and Zhukov, S.J. 1974. On the auditory correlates of the second formant of vowels. Eighth Int. Cong. Acoustics, p. 323.

Nearey, T.M. 1977. Phonetic feature systems for vowels. Ph.D. Dissertation, University of Connecticut. Reprinted by the Indiana University Linguistics Club, 1978.

Nearey, T.M. 1980. On the physical interpretation of vowel quality: Cineflourographic and acoustic evidence. J. Phonetics 8: 213-241.

Nearey, T.M. and Assmann, P.F. 1984. Listeners' identification of brief segments of natural isolated vowels. J. Acoust. Soc. Am. 76: KK1a.

Nearey, T.M. and Levitt, A.G. 1974. Evidence for spectral fusion in dichotic release from upward spread of masking. Haskins Lab. Stat. Rep. Speech Res. SR-39/40: 81-89.

Parsons, T.W. 1976. Separation of speech from interfering speech by means of harmonic selection. J. Acoust. Soc. Am. 60: 911-918.

Patterson, R.D. 1976. Auditory filter shape derived with noise stimuli. J. Acoust. Soc. Am. 59: 640-654.

Patterson, R.D. and Green, D.M. 1978. Auditory masking. In Handbook of Perception Vol. IV: Hearing. Academic Press: New York, Chapter 9.

Patterson, R.D. and Nimmo-Smith, I. 1980. Off-frequency listening and auditory filter asymmetry. J. Acoust. Soc. Am. 67: 229-245.

Patterson, R.D., Nimmo-Smith, I., Weber, D.L. and Milroy, R. 1982. The deterioration of hearing with age: Frequency selectivity, the critical ratio, the audiogram and speech threshold. J. Acoust. Soc. Am. 72: 1788-1803.

Peterson, G.E. 1951. Frequency detection and speech formants. J. Acoust. Soc. Am. 23: 668-674.

Peterson, G.E. 1952. The information-bearing elements of speech. J. Acoust. Soc. Am. 24: 629-633.

Peterson, G.E. 1959. Vowel formant measurements. J. Speech Hear. Res. 2: 173-183.

Peterson, G.E. and Barney, H. 1952. Control methods used in a study of vowels. J. Acoust. Soc. Am. 24: 175-184.

Plomp, R. 1964. The ear as a frequency analyzer. J. Acoust. Soc. Am. 36: 1628-1636.

Plomp, R. 1970. Timbre as a multidimensional attribute of complex tones. In Frequency analysis and periodicity detection in hearing. R. Plomp and G.F. Smoorenburg eds. A.W. Sijthoff: Leiden.

Plomp, R. 1976. Aspects of tone sensation: A psychophysical study. Academic Press: London.

Plomp, R. and Mimpen, A.M. 1968. The ear as a frequency analyzer II. J. Acoust. Soc. Am. 43: 764-768.

Plomp, R. and Steeneken, H.J.M. 1969. Effects of phase on the timbre of complex sounds. J. Acoust. Soc. Am. 46: 409-421.

Plomp, R., Pols, L.C.W., and van der Geer, J.P. 1967. Dimensional analysis of vowel spectra. J. Acoust. Soc. Am. 41: 707-712.

Pollack, I. and Pickett, J.M. 1958. Masking of speech by noise at high levels. J. Acoust. Soc. Am. 30: 127-130.

Pols, L.C.W. 1977. Spectral analysis and identification of Dutch vowels in monosyllabic words. Ph.D. Dissertation, University of Amsterdam. Academische Pers B.V.: Amsterdam.

Pols, L.C.W., van der Kamp, L.J.T., and Plomp, R. 1969. Perceptual and physical space of vowel sounds. J. Acoust. Soc. Am. 46: 458-467.

Potter, R.K. and Steinberg, J. 1950. Towards the specification of speech. J. Acoust. Soc. Am. 22: 807-820.

Rabiner, L.R. and Schafer, R.W. 1978. Digital processing of speech signals. Prentice-Hall: Englewood Cliffs.

Rakowski, A. 1968. Pitch of filtered noise. Sixth Int. Cong. Acoustics A-5-7: 105-108.

Repp, B.H. 1983. Trading relations among acoustic cues in speech perception are largely a result of phonetic categorization. Speech Communication 2: 341-361.

Repp, B.H., Healy, A.F. and Crowder, R.G. 1979. Categories and context in the perception of isolated steady-state vowels. J. Exp. Psych.: Human Perception and Performance 5: 129-145.

Robinson, D.W. and Dadson, R.S. 1956. A redetermination of the equal-loudness relations for pure tones. Brit. J. Applied Physics 7: 166-181.

Rosenberg, A.E. 1971. Effect of glottal pulse shape on the quality of natural vowels. J. Acoust. Soc. Am. 49: 583-590.

Rose, J.E., Brugge, J.F., Anderson, D.J. and Hind, J.E. 1967. Phase-locked response to low frequency tones in single auditory fibres of the squirrel monkey. J. Neurophysiology 30: 769-793.

Rose, J.E., Kitzes, L.M., Gibson, M.M. and Hind, J.E. 1974. Observations on phase-sensitive neurons of anteroventral cochlear nucleus of the cat: Nonlinearities of cochlear output. J. Neurophysiology 37:218-253.

Russell, G.O. 1928. The vowel. Ohio State University Press: Columbus.

Ryalls, J.H. and Lieberman, P. 1982. Fundamental frequency and vowel perception.

J. Acoust. Soc. Am. 72: 1631-1634.

Sachs, M.B. and Young, E.D. 1979. Encoding of steady-state vowels in the auditory nerve: Representation in terms of discharge rate. J. Acoust. Soc. Am. 66: 470-479.

Sachs, M.B., Young, E.D. and Miller, M.I. 1982. Encoding of speech features in the auditory nerve. In The representation of speech in the peripheral auditory system. R. Carlson and B. Granstrom, eds. Elsevier, Amsterdam.

Sachs, R.M. and Zurek, P.M. 1979. Contralateral probe measurements of auditory vowel spectra. J. Acoust. Soc. Am. 65: S55a.

Schafer, R.W. and Rabiner, L.R. 1970. System for automatic formant analysis. J. Acoust. Soc. Am. 47: 634-648.

Scarf, B. 1970. Critical bands. In Foundations of modern auditory theory: Vol. I. Academic Press: New York Chapter 5.

Scheffers, M.T.M. 1982. The role of pitch in the perceptual separation of simultaneous vowels: II. IPO Annual Prog. Rep. 17: 41-45.

Scheffers, M.T.M. 1983. Sifting vowels: Auditory pitch analysis and sound segregation. Ph.D. Dissertation, Groningen University.

Schouten, M.E.H. 1980. The case against a speech mode of perception. Acta Psychologica 44: 71-98.

Schroeder, M.R. 1956. On the separation and measurement of formant frequencies. J. Acoust. Soc. Am. 28: S159a.

Schroeder, M.R. 1959. New results concerning monaural phase sensitivity. J. Acoust. Soc. Am. 31: S1579a.

Schroeder, M.R. 1975. Models of hearing. Proc. IEEE 63:1332-1350.

Schroeder, M.R. 1977. Speech processing by man and machine. In Dahlem Workshop on the recognition of complex acoustic signals. T.H. Bullock, ed. Abakon: Berlin pp. 307-352.

Schroeder, M.R. and Mehrgardt, S. 1982. Auditory masking phenomena in the perception of speech. J. Acoust. Soc. Am. 72: S a.

Schroeder, M.R. , Atal, B.S. and Hall, J.L. 1979. Objective measures of certain speech signal degradations based on masking. In Frontiers of speech communication research. B Lindblom and S. Ohman, eds. Academic Press: London pp. 217-229.

Searle, C.L., Jakobson, J.Z. and Rayment, S.G. 1979. Stop consonant discrimination based on human audition. J. Acoust. Soc. Am. 65: 799-809.

Sekey, A. and Hanson, B.A. 1984. Improved 1-Bark auditory filter. J. Acoust. Soc. Am. 75: 1902-1904.

Singh, S. 1974. A step towards a theory of speech perception. Speech Communication Sem. Stockholm, Aug. 1974, pp. 55-66.

Slawson, A.W. 1968. Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. J. Acoust. Soc. Am. 43: 87-101.

Slepokurova, N.A. 1973. Effects of fundamental tone pitch on the position of the phoneme boundary between vowels. Soviet Physics-Acoustics 18: 356-361.

Small, A.M. and Daniloff, R.G. 1967. Pitch of noise bands. J. Acoust. Soc. Am. 41: 506-512.

Srulovitz, P. and Goldstein, J.L. 1983. A central spectrum model: A synthesis of auditory-nerve timing and place cues in monaural communication of frequency spectrum. J. Acoust. Soc. Am. 73: 1266-1276.

Stalhammar, J.U.J. 1978. Form factors for power spectra of vowel nuclei. II. STL-QPSR 2-3/1978: 23-34.

Stelmachovicz, P.G., Small, A.M and Abbas, P.J. 1982. Suppression effects for complex stimuli. J. Acoust. Soc. Am. 71: 4[illegible]420.

Stevens, K.N. and House, A.S. 1961. An acoustic theory of vowel production and some of its implications. J. Speech Hear. Res. 6: 111-128.

Stevens, S.S. 1966. On the psychophysical law. Psych. Rev. 64: 153-181.

Stevenson, D. and Stephens, R. 1979 The Alligator Reference Manual. Unpublished manuscript.

Studdert-Kennedy, M. 1976. Speech perception. In Contemporary issues in experimental phonetics. N.J. Lass, ed. Academic Press: New York pp. 243-293.

Summerfield, Q., Haggard, M., Foster, J. and Gray, S. 1984. Perceiving vowels from uniform spectra: Phonetic exploration of an auditory aftereffect. Perc. and Psychophys. 35: 203-213.

Sundberg, J. and Gauffin, J. 1979. Waveform and spectrum of the glottal voice source. In Frontiers of speech communication research. B. Lindblom and S. Ohman, eds.)

Sundberg, J. and Gauffin, J. 1982. Amplitude of the voice source fundamental and the intelligibility of super pitch vowels. In The representation of speech in the peripheral auditory system. R. Carlson and B. Granstrom, eds. Elsevier: Amsterdam.

Suomi, K. 1984. On speaker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem. Speech Communication 3: 199-210.

Suzuki, J., Kadokawa, Y. and Nakata, K. 1963. Formant frequency extraction by the method of moment calculations. J. Acoust. Soc. Am. 35: 1345-1353.

Syrdal, A.K. 1982. Frequency analysis of American English vowels. J. Acoust. Soc. Am. 71: S105a.

Terbeek, D. 1977. A cross-language multidimensional scaling study of vowel perception. UCLA Working Papers in Phonetics 37: 1-271.

Terhardt, E. 1979. On the perception of spectral information in speech. In Hearing mechanisms and speech. O. Creutzfeld, H. Scheich and C. Schreiner, eds. Springer: Berlin, pp. 281-291.

Tiffany, W.R. 1959. Nonrandom sources of variation in vowel quality. J. Speech Hearing Res. 2: 305-317.

Traunmuller, H. 1981. Perceptual dimensions of openness in vowels. J. Acoust. Soc. Am. 69: 1465-1475.

Tyler, R.S. and Lindblom, B. 1982. Preliminary study of simultaneous masking and pulsation threshold patterns of vowels. J. Acoust. Soc. Am. 71: 220-224.

Ujihara, J. and Sakai, H. 1974. Lateral inhibition of the auditory nervous system for monosyllabic vowel speech sounds. J. Acoust. Soc. Japan 30: 133-143.

Weston, R.H. 1974. Digital modelling of lateral inhibition with reference to the auditory system. J. Sound and Vibration 35: 309-341.

White, G.M. and Neely, R.B. 1976. Speech recognition experiments with linear prediction, bandpass filtering, and dynamic programming.

IEEE Trans. Acoust., Speech, Signal Processing vol ASSP 24: 183-188.

Winer, B.J. 1971. Statistical principles in experimental design. Second edition, McGraw Hill: New York.

Wood, S. 1979. A radiographic analysis of constriction locations for vowels. J. Phonetics 7: 25-43.

Young, E.D. and Sachs, M.B. 1979. Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory nerve fibres. J. Acoust. Soc. Am. 66: 1381-1403.

Youngberg, J.E. and Boll, S.F. 1978. Constant-Q signal analysis and synthesis. IEEE Trans. Acoust., Speech, Signal Processing, vol ASSP-26: 375-378.

Zwicker, E. 1952. Die Grenzen der Hörbarkeit der Amplitudenmodulation und der Frequenzmodulation eines Tones. Acustica 2: 125-133.

Zwicker, E. 1963. Uber die Lautheit von ungedrosselten und gedrosselten Schallen. Acustica 13: 194-211.

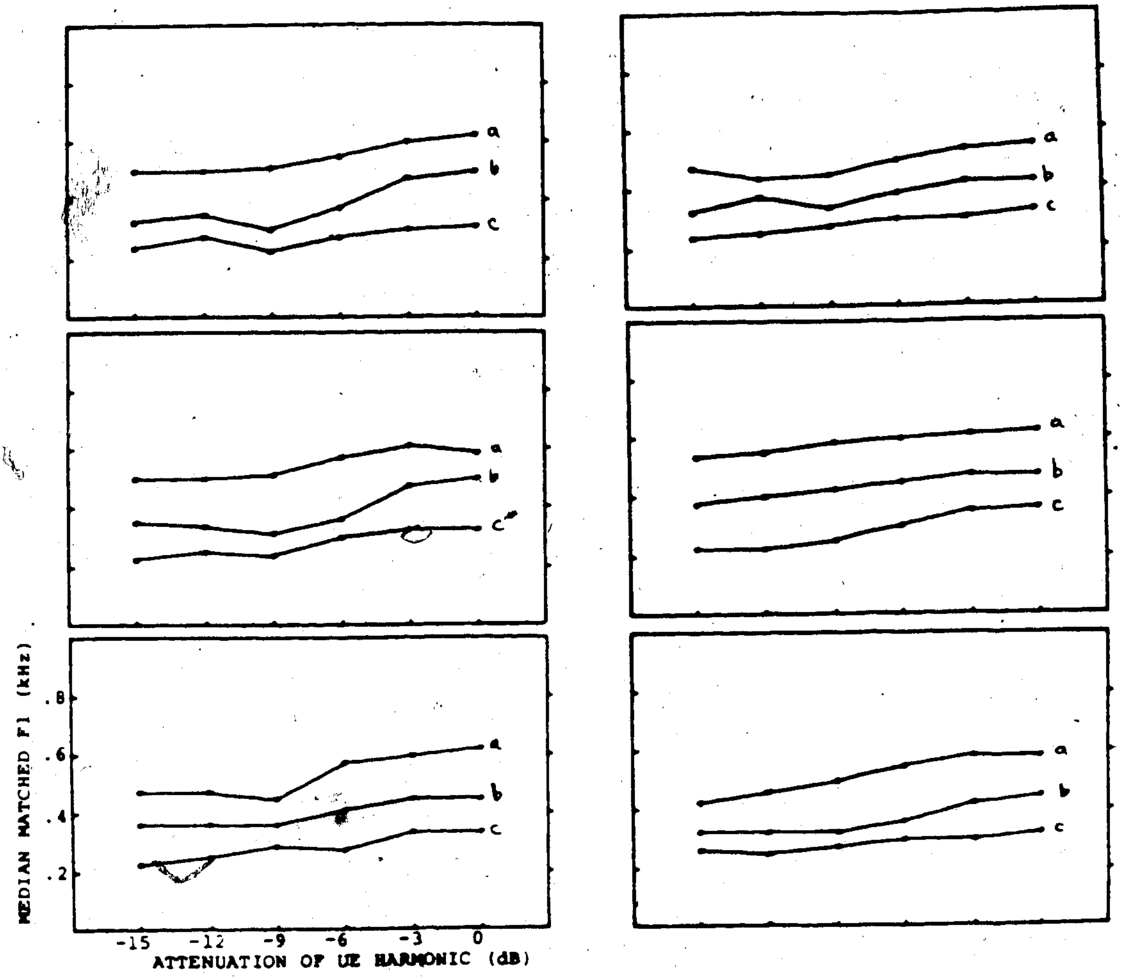
Zwicker, E. and Feldtkeller, R. 1967. Das Ohr als Nachtrichtenempfanger. Second edition, Hirzel-Verlag: Stuttgart.

Zwicker, E. and Terhardt, E. 1980. Analytic expressions for critical band rate and critical bandwidth as a function of frequency. J. Acoust. Soc. Am. 68: 1523-1525.

Zwicker, E., Flottorp, G., and Stevens, S.S. 1957. Critical bandwidth in loudness summation. J. Acoust. Soc. Am. 29: 548-557.

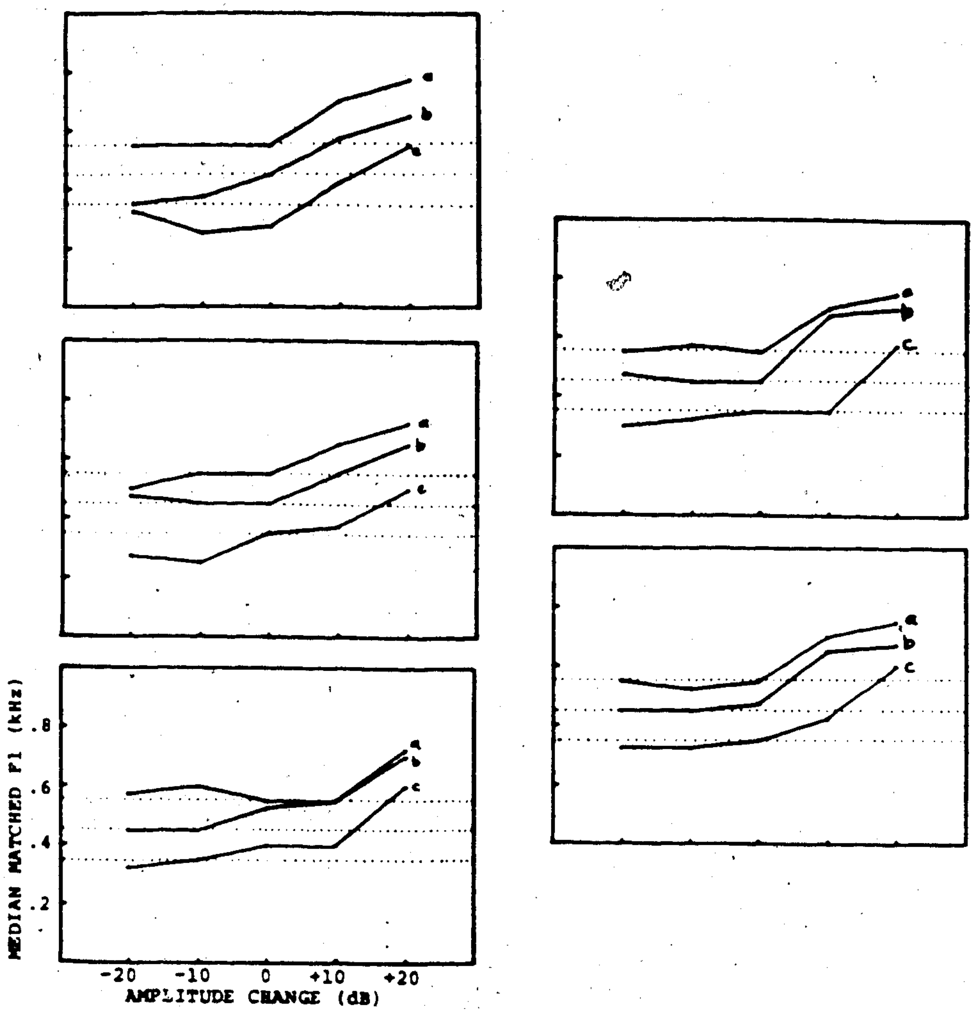
Zwicker, E., Terhardt, E. and Paulus, E. 1979. Automatic speech recognition using psychoacoustic models. J. Acoust. Soc. Am. 65: 487-498

APPENDIX A



Individual listeners' matching data (median matched F1 values across 10 trials) for the upper edge condition of Experiment one.
(a) 5 harmonics condition
(b) 4 harmonics condition
(c) 3 harmonics condition

APPENDIX B



Individual listeners' matching data (median matched F1 values across 5 trials) for Experiment five.

(a) F1=550 Hz; F2=800 Hz
(b) F1=450 Hz; F2=700 Hz
(c) F1=350 Hz; F2=600 Hz

Dotted lines indicate nominal F1 values