10-28-1994

Spectrographic Analysis of Second Language Speech: Investigating the Effects of L1

Troy D. Bailey
Portland State University

THESIS APPROVAL

The abstract and thesis of Troy D. Bailey for the Master of Arts in Teaching English to Speakers of Other Languages were presented October 28, 1994, and accepted by the thesis committee and the department.

COMMITTEE APPROVALS:

[redacted]
Beatrice Oshika, Chair

[redacted]
Tom Dieterich

[redacted]
John Tetnowski
Representative of the Office of Graduate Studies

DEPARTMENTAL APPROVAL

[redacted]
Beatrice Oshika, Chair
Department of Applied Linguistics

****

ACCEPTED FOR PORTLAND STATE UNIVERSITY BY THE LIBRARY

by [redacted] on [redacted]

ABSTRACT

An abstract of the thesis of Troy D. Bailey for the Master of Arts in Teaching English to Speakers of Other Languages, Department of Applied Linguistics, presented October 28, 1994.

Title: Spectrographic Analysis of Second Language Speech: Investigating the Effects of L1

Technological advances in Digital Signal Processing over the last decade have provided applied linguists with a number of computerized applications for speech analysis which can be of benefit to both the researcher and the instructor. This research project explores the techniques of speech spectrography and implements methods of acoustic phonetics to current issues in Second Language Acquisition theory.

Specifically, the effects of vowel production in one's native language on the targets in a second language are investigated. Acoustic measurements of English vowels spoken by Japanese students were compared with measurements of native Japanese vowels and American English vowels. In addition, these data were compared with measurements of learner speech from a variety of native language backgrounds. Vowels from both groups of non-native English speakers showed tendencies toward the center of the vowel space. The less-experienced group showed greater token-to-token variability across height parameters than across frontedness parameters while the more experienced group showed no difference for parameters. Both groups exhibited greater frontedness than heigth variability between speakers which can be explained in part by differences in vocal tract size. In addition, Flege's Speech Learning Model was tested.

Data did not support the hypothesis that similar vowels are more difficult to produce than different vowels. ANOVA tests showed that large L1 vowel inventories do not advantage learners of languages with many vowels. The results suggest that the unique qualities of L2 speech may have more to do with developmental processes than L1 interference.

SPECTROGRAPHIC ANALYSIS OF SECOND LANGUAGE SPEECH: INVESTIGATING THE EFFECTS OF L1

by

TROY D. BAILEY

Adviser: Beatrice Oshika, Ph.D.
Department of Applied Linguistics

A thesis submitted in partial fulfillment of the
requirements for the degree of

MASTER OF ARTS
in
TEACHING ENGLISH TO SPEAKERS OF OTHER LANGUAGES

Portland State University
1994

Acknowledgments

Throughout this research project I have benefitted from the insights and encouragements of a number of individuals, some of which I would like to mention here. First of all, I would like to thank my committee for taking the time to give me valuable input. The committee consisted of Tom Dieterich, an expert in perception and psycholinguistics, John Tetnowski, a specialist in the speech and hearing sciences and expert in acoustic instrumentation, and my advisor, Beatrice Oshika, a computational linguist with expertise in acoustic phonetics and speech recognition. I am honored to have such an outstanding group of researchers informing my work.

Also, a number of departmental institutions contributed to my research by allowing me access to needed equipment. I would like to thank the Portland State University department of Speech Communication, Division of Speech and Hearing Sciences for providing me access to the acoustic equipment and the Audio-Visual Services center for access to the recording studio and equipment. Additionally I would like to thank those at the Center for Spoken Language Understanding at the Oregon Graduate Institute, specifically Mike Noel and Terri Lander, for their flexibility regarding my work schedule in the lab.

Finally, I extend my gratitude to the Portland State University ESL students who participated in the study. Without their willing and eager participation I could not have accomplished this. I am grateful for the kindergarten children at St. Anthony's School whose curious spontaneity served as a constant reminder throughout the year of the wonder of developing speech. I would like to thank Savitri Simanjuntak for her help in the preparation of elicitation lists, my father for help with the diagrams, and, of course, my

mother who patiently endured crash courses in word processing in order to help her distressed son with last minute revisions.

TABLE OF CONTENTS

TABLE OF CONTENTS: DETAILED

TABLES AND FIGURES

Tables

Figures

Abbreviations

JENG	Japanese English (as a second language)
SAE	Standard American English
ENG	American English as measured by Peterson and Barney (1952).
IL	Interlanguage
L1	First, native Language
L2	Second or additional language.
CAH	Contrastive Analysis Hypothesis
VSD	Vowel Space Diagram
F1	Formant # (from lowest frequency to highest)
cps	Cycles per second
Hz	Hertz
CLA	Child Language Acquisition
SLA	Second Language Acquisition

Description of Phonetic Symbols

For this paper I have chosen to use the "Worldbet" developed by Jim Hieronymus (1993) because each of the phones can be represented using ASCII characters available on a standard keyboard without using control characters. It should be noted, however, that there are a some inconsistencies introduced by this system. One such example, the English high front vowel, appears from the conventions to be a long vowel (note [i:]) when actually the key feature is not length but rather the diphthong-like transition. In the Japanese data, however, the Worldbet symbol [i:] indicates phonemic length.

ENGLISH SYMBOLS

VOWELS			DIPTHONGS	
	Example	Description		Example
i:	eat	high front tense		
I	hit	high front lax		
ei	bait	mid front tense		
E	head	mid front lax		
@	had	low front		
A	hot	low back	aI	tie
^	hut	mid central	aU	cow
>	caught	low back	>i	toy
u	hoot	high back tense	oU	toe
U	hood	high back lax		
3r	hurt	retroflexed		

JAPANESE SYMBOLS

CONSONANTS			DIPHTHONGS	
	Example	Description		
i	ichi "one"	high front unrounded		
i:	iie "no"	high front unrounded long		
e	koe "voice"	mid front	eI	kirei "pretty"
e:	sensei "teacher"	mid front long		
4	uta "song"	high back unrounded		
4:	futsuu "ordinary"	high back unrounded long		
o	igo "Igo game"	mid back rounded		
o:	tookyoo "Tokyo"	mid back rounded long		
a	san "three"	low central	aI	hai "yes"
a:	apaato "apartment"	low central long		
&		mid central		
&_0		voiceless mid central		

PART I:

THEORETICAL ISSUES

1
Introduction

The topic of "Foreign accent" has often been treated by academics and non-academics alike as a kind of debilitation, an inevitable part of second language acquisition[1] (SLA). Commonly, the difficulties of acquiring native-like[2] pronunciation in a second language are attributed to relatively inflexible nuero-muscular patterns of articulation. Foreign accent is therefore thought of as random articulatory fumbling without pattern (except that it often resembles the characteristics of the speaker's native language). The purpose of this study is to examine some of theoretical assumptions based on these popular notions of "randomness" and "interference" in second language (L2) speech production.

The study of foreign accent and other aspects of L2 speech has often focused on consonants. One reason for this is that in many ways consonants are more amenable to traditional phonetic analysis. Much of the significant information conveyed in consonants is based on the presence or absence of largely categorical features such as voicing and manner of articulation. Even the notion "place of articulation" has been traditionally described categorically as a series of eleven distinct positions across the mid-saggital region. Vowels, on the other hand, differ from each other along continua of vowel "quality" parameters. These are much more difficult to distinguish, even among trained phoneticians. As a result, consistent and reliable methods of comparison have been scarce.

Vowels, however, comprise much of the nucleus of the syllable and consequently contain volumes of information about the speaker's background, dialect, psychological state as well as discourse level information. This study is concerned with the transitive nature of learned vowels in a second or foreign language. Fortunately, the techniques of spectrographic analysis make it possible to investigate many of these topics empirically. It may be that an increased awareness of acoustic spectrograms among applied linguists could yield many powerful insights into language learning processes.

Part I of this presentation develops theoretical viewpoints from both speech research and SLA. Chapter 2 provides the groundwork for the acoustic phonetic study by highlighting fundamental theoretical perspectives within the speech sciences with particular focus on the perception/production debate.

In Chapter 3 is a description of various theoretical perspectives of in the fields of SLA and adult speech learning (ASL). The chapter particularly highlights historical trends away from "product-oriented" models characterized by the Contrastive Analysis theory and methods of error analysis toward "process-oriented" models emphasizing psychological and linguistic universals that make L2 acquisition similar regardless of the learner's L1. Typically, such approaches describe the process of SLA as the development of an "interlanguage" (IL) or "approximative system" which is said to be rule-governed and common to all learners. Among the IL approaches, the notion of "transfer" was not ignored but rather integrated into a broader conceptualization of the processes of speech learning. This shift has entailed a move away from Contrastive Analysis in all areas *except* for speech learning research where it has until recently remained the primary theoretical framework.

Part II describes some of the essentials of Acoustic Phonetics methodology which will be of importance to this study. As with all empirical sciences, there are critical issues

regarding measurement procedures and the quantification of variables which need to be settled before discussing the study at hand. Chapter 4 focuses on the tools of spectrographic analysis, while Chapters 5 describes issues pertaining to the study of vowels especially as it relates to SLA research. Here, references will be made to former studies in the literature with regard to L2 vowel production.

Part III describes the study of L2 vowels. Chapter 6 provides the methodological framework for the experiments to follow in Chapters 7 and 8. Chapter 7 provides phonetic descriptions of English and Japanese vowels comparing traditional linguistic descriptions from the literature with actual measured data of native-proficiency Japanese vowels. Chapter 8 comprises the core of the study. Two groups of ESL students are studied: Japanese speakers and a quasi-control sample of speakers from 5 different languages. The experiments in the Chapter are divided into four areas of inquiry: 1) Intra-speaker variability, 2) IL Vowel positions and spread for both groups, 3) Inter-speaker Variability, and 4) Flege's Speech Learning Model. Finally, Chapter 9 provides a discussion of the experiment results in the light of the theoretical issues raised in Part I.

One of the underlying goals of this study has been to address a fairly traditional phonetics issue (influence of native language on second language speech) using an experimental approach with quantifiable data. However, there have been controversies among phoneticians about the validity of data gathered from acoustic instrumentation. The claim is that speech is primarily a human event characterized by human perceptual tendencies. Attempts to reduce the phonetics task to an engineering problem of acoustic analysis, the argument proceeds, introduces an artificial heuristic to the data. Additionally, it is claimed that an element of the *art* of phonetics is lost when problems are addressed mechanically.

Language, in its natural form defies even the best of theoretical models. Most linguists agree that the fundamental criterion for testing descriptive statements about language lies in the intuition of the speaker. At the same time, these claims must be tested if they are to be accepted widely and if they are to be respected by researchers in other disciplines. Linguistic phonetics is both art and science. The process of "acquiring an ear" for subtle distinctions within speech involves skill worthy of the title "art" yet phoneticians have found it beneficial to make use of the fundamentals of scientific method (i.e., observation, hypothesis generation, systematic testing). These two aspects of art and science need not be seen as contradictory but complementary.

Interestingly enough, the most heated arguments of this debate have focused on the very topic of this study: *vowel measurement.* British and American linguistic traditions have been strongly influenced by the notion of the "Cardinal Vowels" first discussed by the British phonetician Daniel Jones. The cardinal vowel system was introduced as an analytical framework for the identification and placement of vowels from

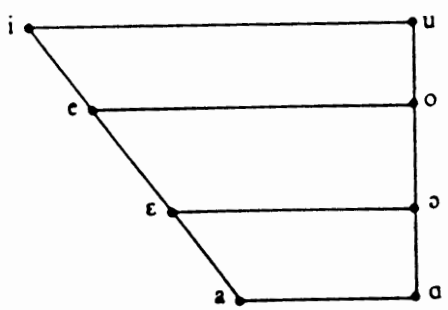


Figure 1.1 *Daniel Jones's system of "Cardinal Vowels"*

any language. The model posited originally 8 and then 14 equally spaced vowels positioned across two perpendicular axes: height and frontedness (Abercrombie, 1985). If these theorized vowels approximated the *real* vowels of any given language, it was only by coincidence. Their primary utility was that they served as a grid of possible points of vowel articulation and could be used for the purpose of classification and measurement across languages and of speech variations within a given language. According to Daniel Jones' approach, the phonetician had to be trained to recognize these vowels upon hearing them. This skill could only be learned from one who "knows" the cardinal vowels.

Recently, however, this framework has come under much criticism by researchers making use of technological methods of measurement. It has become clear from X-ray studies that the Cardinal Vowels diagram does not accurately represent the motions of the tongue within the mouth (Fischer-Jorgensen, 1985). Thus, researchers of this camp have criticized the cardinal vowel system for being "unscientific" and "inaccurate". Also, some have become skeptical about the purported reliability of phoneticians' subjective judgments based on this framework.

While the evidence confounding the relationship between tongue motions and the theorized parameters of "height" and "frontedness" is clear, Jones's model has not been made obsolete. It must be remembered that the purpose for the cardinal vowel system was to provide a mental heuristic for phoneticians making qualitative judgments "in the field". In many ways the model closely approximates the perceptual parameters of height and frontedness. To this end, the model is empirically defensible.

FOOTNOTES:

[1]While Krashen (1983) makes a distinction between the terms language learning and language acquisition so that learning implies a high degree of self-monitoring while acquisition indicates a more "natural" process toward proficiency, I will use the terms interchangeably without regard to the notion of "monitoring" unless addressed specifically.

[2]The term "native-speaker" is currently under scrutiny by many within the disciplines of applied linguistics and language teaching. Some claim that the term discriminates against speakers of a given language who did not learn the language in childhood yet who have "native-like" mastery over the language. I have chosen to use the term "native-like", not because I agree with the proponents of the term, but because I do not wish to make it an issue in this paper. For a fuller explanation of the arguments in this controversy see Paikeday (1985).

2

Theories of Speech Perception and Production

In this chapter I will survey some of the core issues of psycholinguistic research giving special attention to the perception/production controversy. Of course, space does not allow a thorough treatment of the issues at hand, nevertheless, the nature of this present study requires that I delineate exactly what is *not* being investigated. This chapter should serve to clarify the scope and limitations of the present research. Concepts described here will form the foundation of the research in this study and will provide a platform for the discussion of theoretical issues of language acquisition in the following chapter.

A. TWO MODES OF SPEECH

Denes and Pinson (1963) describe speech communication as a communicative model with two essential modes: acoustic encoding and decoding. Early notions of linguistic phonology often linked the two as if they were psychologically linked at the level of the phoneme. Thus, proponents of the Motor Theory of Speech perception claimed that phonemes were actually psychological primitives that served as the input to hearing speech perception and articulation. Thus, it is claimed, people map acoustic input

to a psychological representation of the oral cavity before identifying the component parts. Today, most psycholinguistic researchers agree that this linkage is untenable and that cognitive, perceptual and proprioceptive processes seem to point to two distinct modes.

1. Speech Perception

1.1 The Auditory System as a Pre-Processor for the Signal.

The human auditory system seems remarkably suited for speech decoding. Frequency resolution, the most important acoustic dimension of the phonetic signal is most precise at the 0-5kHz bandwidth of speech. This heightened awareness is vital since the speech signal is very quiet relative to other audible sounds. It is well-attested that the speech signal is quite redundant (O'Shaughnessy, 1987). Such observations rise from our understanding of the auditory system's unique ability to analyze this 5kHz speech spectrum. In this section I will briefly summarize the primary elements of the auditory system pointing to their respective functions and, where possible, their contribution to speech understanding.

The function of the ear's anatomical structure is to convert atmospheric sound waves into electrochemical neural responses. Hearing in the human ear involves wave conduction through three types of media: gas, solid, and liquid corresponding to three subdivisions of the ear: the outer, middle and inner ear. First, the signal reaches the outer ear and funnels down the ear canal toward the tympanic membrane. At this point, the ear canal forms an acoustic resonator amplifying sound waves near 3kHz (the resonant frequency of this external ear structure).

Next, acoustic energy is amplified by about 80 times (38dB) in the air-filled middle ear through a complex mechanical interaction of the malleus, incus, and the stapes which enable about half of the energy to be transferred to the inner ear. This intermediary

mechanical process helps match the impedance levels between the air-filled middle ear and the liquid-filled inner ear. Without this impedance match, the total energy entering the liquid medium of the inner ear would be deflected, halting the hearing process.

Finally, this filtered and amplified signal enters the inner ear where it is processed and analyzed by the cochlea. The structural shape of the spiraled organ causes vibrations of various frequencies to be spread out along the tube proportionate to the diameter of the tube. Thus, low frequencies excite the widest part of the basilar membrane while high frequencies excite the terminal narrow portion. This is known as the "place theory" of hearing. Throughout the length of the basilar membrane tiny bundles of hair cells extend into the liquid medium of the canals. These constitute the organ of Corti. When excited, the cells convert the wave energy into electro-chemical impulses transmittable by the central nervous system.

Thus we can see that the anatomy of the ear functions as a "front-end" signal filtering and amplification system. Following this stage is a higher-order "back-end" level of perception which serves as a kind of data-reduction process allowing the brain to attend primarily to the range of relevant information. The next section deals with this perceptual back-end and the linguistically important parts of the signal.

1.2 Perceptually Salient Acoustic Features

Perception is difficult to analyze directly. The best that researchers can do is to test people's performance on given tasks which control various kinds of "input". Such modifications include bandpass filtering out various frequency ranges in the spectrum, sampling in additive noise, and temporal masking techniques which add or subtract elements of the signal. Such experiments have revealed that the first 1 kHz helps to identify voicing and manner of articulation in weak obstruents such as /p/ vs. /b/ vs. /v/

while frequencies above 1.2kHz contribute to the identification of place of articulation (O'Shaughnessy, 1987).

Fortunately, vowel perception is relatively simple in that position of the first three formants directly relate to what people hear. This is well established in studies of listener's classifications of synthetically generated vowels. It is also clear that vowel identification requires some level of vocal tract normalization from data gained from coarticulation of surrounding consonants. Inter-speaker variability is very large for vowels.

1.3 Auditory Warping

Much is known about human auditory thresholds for the acoustic dimensions of frequency, time, and intensity. Research in audition confirm findings in other areas of perceptual research that suggest that human sensing involves a "warping" of the physical signal to levels usable by the brain. Intensity acuity is related to frequency. For example, a 80Hz tone must have an amplitude of at least 45 dB-SPL to be heard, whereas a tone at 1000Hz can be heard at intensity levels of -5dB-SPL. Detection of frequency differences begins to weaken at 1000Hz and becomes progressively worse as frequencies increase. Highest frequency resolution lies between 100Hz and 5000Hz. Temporal resolution is related to "critical bandwidths" across the basilar membrane of the cochlea.

Because of this perceptual warping it is necessary to be able to measure differences between actual physical phenomena (i.e., frequency and intensity) and human perception of these (, pitch and loudness) phenomena. Thus, loudness is measured in decibels rather than dynes per square centimeter, and pitch is measured in mels or barks rather than Hertz (cycles per second).

Each of these logarithmic scales is based on either perceptual research or theoretical models. The mel scale was designed to model listener judgments of pitch

differences. A tone of 1000Hz was selected as an arbitrary reference point for listeners and called 1000Mels (Stevens, Volkmann, and Newman, 1937). From there listeners are asked to judge when a tone has reached half its frequency, double its frequency and so on in order to plot the pitch curve. Results of these perceptual experiments have been used to interpolate the mel scale. A disadvantage of this scale is that its authenticity depends upon the extent to which perceptual difference limens tests represent true psychoacoustic pitch perception as it operates naturally.

An alternative logarithmic scale is the Critical Band Rate (Bark) scale based on cochlear modeling theory. A "critical band" is described as a bandpass filter whose frequency response corresponds roughly to the tuning curves of auditory neurons (O'Shaughnessy, 1987). Experimentally derived measures of such critical bandwidths were derived using noise threshold and frequency modulation. From these data it has become possible to model the frequency responses of the basilar membrane.

2. Speech Production

Speech begins as a release of pulmonic air which flows through the glottis. In voiced speech the folds then begin to vibrate at a frequency relative to their tension. In unvoiced speech the semi-random noise occurs resulting from friction against the walls of the larynx. Here I will concentrate on voiced speech of vowels. The effect of the glottis is to produce a periodic excitation of the air stream as it passes through the larynx into the oral cavity. Here the energy interacts with the structure and position of the articulators to form a set of oral resonants. This constitutes the secondary source of excitation in the signal. As the air stream exits the mouth the energy is further radiated by the lips.

Vowels have received much attention in the area of mathematical speech production modeling. The dominant model describes the set of vowel resonators as a

series of concatenated lossless tubes (the *lossless* assumption is imperfect but it greatly simplifies computation). Tongue motions across the horizontal plane (place of articulation) are modeled as variations in the tube length. Vertical tongue motions are modeled as variations in tube diameter. Figure 2.1 shows vocal tract configurations for three vowels: [i:], [A], and [u] here the frequency output is a function of the area, A, and length, l, of the component tubes. Tubes open at both ends are *quarter-wave* resonators. Tubes closed off at the back and open in the front are *half-wave* resonators. Other vowels with intermediary tongue positions are difficult to model because they lack abrupt boundaries between these theoretical tubes. For a more complete explanation see Zue, 1985.

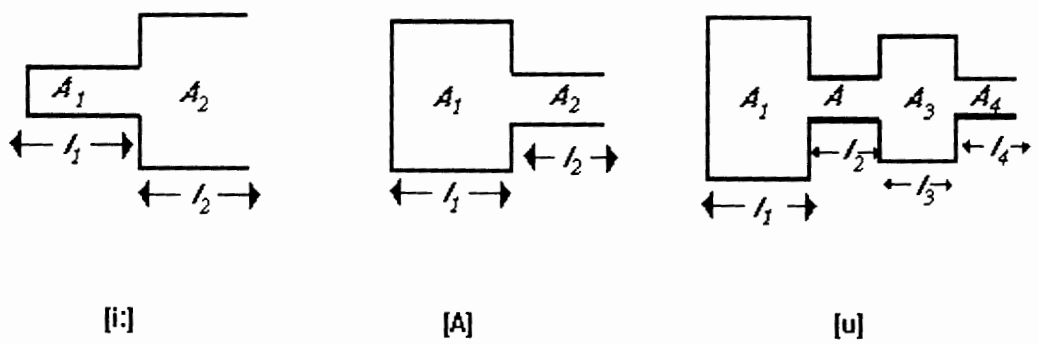


Figure 2.1 Three Vocal Tract Models

3. Perception/Production Relationships and Discontinuities

Peterson and Barney (1952) measured the English vowel productions of individuals and then subjected the speakers to a listening test that required them to classify vowels spoken by other speakers. They found that in some contiguous F1xF2 space vowel pairs where classification was difficult, listeners tended to classify the vowels

according to the way they themselves produced the vowel. For example, if a speaker produces [I] for [E] before nasals such as in /pIn/ for /pEn/ as currently heard in some dialects; then such an individual when serving as a listener will be inclined to write "pen" when he hears /pIn/.

The observation that a linguistic segment must be perceived before it can be produced is apparent from studies investigating the order of acquisition of these two modes both in Child Language Acquisition (CLA) (Berndt 1992) and Second Language Acquisition (SLA) (Mochizuki-Sudo 1991; Sakow and McNutt 1991). However, evidence from developmental order alone does not constitute evidence that the one mode explains the ontogeny of the second. Liberman and other proponents of the Motor Theory of Speech Perception (Mattingly 1991) also linked perception and production in a unified theory of speech, positing that the auditory system mapped acoustic information onto deep phonological gestural primitives that would then drive production. However such views that claim perception relies on knowledge of production have received much criticism (Leather and James 1991). Straight (1980) argued that there are no acoustic features which map directly to articulatory features nor articulatory features which directly map onto acoustic features. Many researchers now view the two systems as distinct psycholinguistic modes with distinct underlying representations.

What is important, however, is that information learned in either of the two systems does to some degree transfer to the other. Leather (1986) conducted an experiment where adult Dutch and English speakers were given computer-managed individualized training in the perception of Chinese tone. Subjects who achieved a minimal perceptual proficiency were then administered a tone production test. Other groups of subjects were trained using a computer-managed, interactive visual feedback system to produce the tones without any auditory exposure to tone exemplars, and those

subjects who attained a proficiency criterion were then tested in their tone perception abilities. The results indicated that learners do not need to be trained in production to be able to produce or in perception to be able to perceive the sound patterns of the target system.

B. VOWEL SYSTEMS AND LINGUISTIC FORCES

In the following section I will discuss vowel systems and linguistic forces which could serve as explanatory models integrating knowledge gained in the above section (Perception/Production theories) into a set of hypotheses.

1. Linguistic Forces

What determines the structure of a given language's vowel system? Ladefoged describes two linguistic forces which he believes shape the vowel inventories and configurations of every language (1993). The first of these "linguistic forces" is the principle of "ease of articulation". The desire to simplify utterances is realized in coarticulation. Targets become fuzzy and meld with those of the neighboring sounds. This coarticulation process, says Ladefoged, is one of the causes of language change. Examples of this include the various types of context driven assimilation such as the voicing of an alveolar fricative between voiced vowels such as in "resist" and "result". Here the [s] has become a [z] because of the surrounding sonorant vowels.

If "ease of articulation" were unchecked, soon language would become unintelligible. Thus, a second linguistic force helps maintain intelligibility: the principle of "sufficient perceptual separation". This involves the notion that phonetic targets are

organized in such a way as to be sufficiently distinct from each other. The purpose of phones is to make words sound differently. This principle operates especially in word contexts where various segments could be phonetically inserted. Languages tend to set their vowels toward the outside of the vowel space in order to make them maximally distinct. Also, linguists often observe that vowel systems tend to be balanced and symmetrical. Speakers of all languages recognize the limits of the vowel space and make use of its range by spreading the vowels across the acoustic area.

There are, however, many contexts where proper word identification does not require hyper-accurate articulations because phonotactic constraints rule out certain vowels occurring in certain contexts. Such constrained contexts seem to allow vowels to succomb more to coarticulation. Ladefoged has explained "whenever a language does not distinguish between two similar sounds, the actual sound produced will tend to be in between the two possibilities." (1993, p 269). For example, English does not allow tense vowels before the velar nasal [N], consequently most people utter the wordfinal -ing suffix such that the vowel falls between [i:] and [I]. Similarly, phonotactic constraints rule out the occurrence of tense vowels before the rhotacized approximate [9r]. Consequently, most English speakers utter the vowel in "here" somewhere between the [i:] and [I] targets.

If these processes which aid communication and provide rules for economy of effort are at work among native speakers of a language, it seems reasonable to predict that they would also be at work among learners of a second language. Learners wish to make themselves understood and yet they want to sound natural. Both of these principles are needed to achieve these goals. As mentioned before, these forces operate with varying degrees of strength depending on the word context. Since L2 learners may not be familiar with the various phonotactic rules of the target language they are unable to know when to

apply perceptual separation to their productions. This is related to the common hypothesis that IL errors take place in contexts which, in the L1, needed no distinction. In such situations, the learner would coarticulate right through the needed target distinction. If this is true, universal linguistic forces could be driving L1 transfer (the process of substituting native language forms for target language structures).

2. Phonetic Inventory Universals

It is popularly observed that learners' whose L1 contains a large inventory of vowels (i.e., German or French) will find it easier to approximate the vowels of English than learners whose L1 inventory is small (i.e., Spanish, Japanese) since English has a relatively large set (11). The assumption is that a larger inventory increases the likelihood that the learner will have access to transferable vowels. According to this reasoning, I have heard it said that Korean ESL students find it easier to pronounce English vowels than do Japanese because the former have a much larger vowel inventory.

Flege (1989) demonstrated that a language's vowel inventory size may influence the *positioning* of vowels in the acoustic vowel space but does not affect the *accuracy* of tongue motions in production. Using glossometry techniques, Flege measured tongue heights of eight native Spanish speakers and eight native English speakers producing the vowels [i], [u], [a], [e] ([eI] for English) and [o] ([oU] for English). Token-to-token variability for these vowels was no greater for the Spanish than for the English despite the smaller vowel inventory of Spanish.

Flege's research focused on the intra-speaker (utterance-to-utterance) variability (*precision*) of learners' approximations. However it is not clear whether inter-speaker variability (*accuracy*) is affected by the size of a learner's L1 inventory.

3. Validity of the Features "Height" and "Front-back"

It has long been the tradition within articulatory phonetics to classify vowels using the features of "height" (or degrees of openness) and "frontedness". The assumption has been that auditory qualities of vowels correspond directly to articulatory positions along two perpendicular axes--the one being vertical and the other being horizontal. Evidence from a large number of X-ray studies has contradicted this assumption and has caused researchers to question the validity and, or accuracy of this two-parameter description of vowels (Meyer 1910; Russel 1928, 1936; Ladefoged 1962, 1971, 1975, 1976; Ladefoged, DeClerk, Lindau, and Papcun 1972; Joos 1948; Nearey 1978; Wood 1975, 1982). Researchers have tried to improve the accuracy of such descriptions by defining vowel placement by the "point of the highest part of the tongue" but even this is quite variable. Fischer-Jorgensen says that this description is "a much too precise concept to be used in a general vowel system" (1985). Ladefoged has said that he often has considered introducing new terms for vowel features but did not do so because the old system has become so familiar to linguists throughout (1993).

Table 2.1 Wood's Binary system (1982:168)

CONSTRICTION	PALATAL				PALATOVELAR		PHARYNGOVELAR			LOW
PHARYNGEAL										
VOWEL:	i:	I	eI	E	u	U	oU	>	A	a
palatal	+	+	+	+	+	+	-	-	-	-
velar	-	-	-	-	+	+	+	+	-	-
pharyngeal	-	-	-	-	-	-	+	+	+	+
open	-	-	+	+	-	-	+	+	+	+
round	-	-	-	-	+	+	+	+	(-)	-
tense	+	-	+	-	+	-	+	-	+	-

Wood (1982) proposes a binary feature system with four primary places of maximal constriction (Table 2.1). This is, however becomes innaccurate for modeling rapid continuous speech. Ladefoged has worked on a more thorough description of vowels using the additional parameters of lip rounding and ATR (advanced tongue root) which seem to have wide application across languages. He suggests not an abandonment of the original two features but rather a greater degree of specificity (18 cross sections of the mouth for 10 vowels). Some have even suggested a return to the simplified one-dimensional system of brightness-darkness. This is a purely auditory dimension that shows up in many vowel classification studies of linguistically-naive listeners (Fischer-Jorgensen, 1985).

Then, do the parameters "height" and "front-back" accurately identify vowel quality differences? While the terms may be somewhat misleading, there is much evidence that the way in which this two-feature system (excluding lip rounding and ATR) has classified vowels is both linguistically and perceptually accurate. First of all, the acoustic relationships between F1 and vowel height and between F2 and frontedness is strong evidence. Ladefoged points out that it is possible to *hear* these oral resonances. For example, when whispering the vowels in front-to-back order, it is possible to hear the descending pitch of F2. Conversely, when saying the four front vowels in descending order with extreme glottalization it is possible to hear the ascending pitch of F1. It may be that the perceptual separability of these two parameters suggests that they may serve as perceptually valid indices for vowel descriptions. In addition, well-known phonetic universals of vowel support the maintenance of these features. "Low vowels are longer than high vowels; they are pronounced on a lower pitch and have more intensity." (Fischer-Jorgensen, 1985).

3

Theoretical Issues in Adult Speech Acquisition

A. TRANSFER

The unique characteristics of one's "foreign accent" stem in part from the influence of his or her native language. This unmistakable influence of the L1 (first Language) is partly responsible for the beauty as well as the perplexity of one's L2 speech. The reality of such L1 transfer (sometimes referred to as "interference") is both intuitively apparent and supported by decades of research. Nevertheless, many researchers now feel that the role of one's native language in the acquisition of L2 speech has been overstated. In fact, second language speech errors, that is, any realized deviation from a native speaker target may be caused by a variety of factors: linguistic, sociological, and affective variables that each uniquely influence L2 productions (Gass et al, 1989).

Recently, phoneticians and applied linguists have begun to investigate another type of L2 variability: developmental variability. Second language learners produce intermediary forms as well as unique forms which have no realization in their L1 (Major 1987; Flege 1980, 1981). L2 Learners engage in many of the same processes of simplification that children use in acquiring their first language (Major 1986; Wode 1981; Flege 1980, 1981; Mulford and Hecht 1980). It is widely attested that some L2

substitutions parallel substitutions made by children acquiring their L1 (Major 1986; Wode 1981; Flege 1980, 1981; Mulford and Hecht 1980; Flege and Davidian 1985). Such findings suggest that "foreign accent" is more complex than was originally thought.

1 The Contrastive Analysis Hypothesis

During the intellectual era from 1940-1960, the way in which people viewed learning a second language was shaped by behaviorist psychology and linguistic structuralism. Learning was viewed as patterned practice. Language instruction typically began with a systematic comparison of the form of the base (first) language and the target language (TL) in order to map out a path for learning which would indicate those forms of the native language could be transferred to a knowledge of the second language and those forms which could not be transferred (Larsen-Freeman and Long, 1991, p52). These views led to the Contrastive Analysis Hypothesis (CAH) first articulated by Lado: "those elements which are similar to his native language will be simple for him and those elements that are different will be difficult" (1957, p2).

Since then, years of research in syntactic acquisition have disproved the hypothesis that learner errors could be predicted from a structural analysis of the first language (Larsen-Freeman and Long, 1991). As Wardaugh writes (1971, p123), "The CAH experienced a period of quiescence" over the following two decades--at least in the fields of grammar and syntax acquisition. Theoreticians in Second Language Acquisition (SLA) research began to direct their attention to an understanding of the influence of cognitive processes in second language acquisition and to a more thorough understanding of universal tendencies in language learning.

Despite advances in the fields of grammar and syntax, phonological theories of SLA remained heavily influenced by notions of L1 transfer. A "foreign accent" was often

described from the behaviorist standpoint which focused on neuromuscular development. Findings from studies in the muscular development of various parts of the body showing that adults are less able to acquire some motor control skills was taken as evidence that the neuromuscular systems of speech also experienced the same loss in ability to learn motor fine motor skills. This explained why adult learners find it necessary to resort to the familiar articulatory patterns of their L1 to a greater extent than do young children. While some contest the validity of such conclusions (Flege 1981), most theoreticians subscribe to at least a moderate view of the influence of diminished motor skills as a cause of a foreign accent.

The CAH, while providing some explanation for learner errors, could not by itself accurately predict learners productions. Leather and James (1991) in an evaluation of current theories on SLA write "the monolithic contrastive analysis hypothesis provided by classic structural phonology is no longer believed to be adequate to account for acquisition data and has been largely supplanted by models that are sensitive to longitudinal change as well as to those 'sub-phonemic' variations that may be of developmental importance"(p332).

Major (1987) cites five main weaknesses of an interference description of second language acquisition: 1) The CAH could not reliably predict errors. It could only explain them *a posteriori*. 2) Interference could not explain why learners produced sounds which occurred neither in the native language (NL) nor in the target language (TL). For example, Hungarian learners substitute [sth] which does not occur in their L1 for [T]. 3) It could not explain the wide variation between speakers or within speakers, 4) It could not explain why learners gradually progress rather than make a categorical jump between L1 forms and target forms, and 5) Interference theory could not relate L2 acquisition to other aspects of linguistic theory such as universals or developmental sequence.

2 Learning Theory and Transfer

Psychological theories of perception have come to influence SLA phonological theories. Oller and Ziahosseiny (1970) proposed a version of the CAH which would incorporate perceptual factors in the prediction of difficulty. They argued that "the categorization of abstract and concrete patterns according to their perceived similarities and differences is the basis for learning; therefore, wherever patterns are minimally distinct form or meaning in one or more systems, confusion may result" (p186). Thus, they argued that intralingual errors (errors resulting from difficulty within the L2) caused greater difficulties than interlanguage errors (errors resulting from L1 interference) for learners because forms within the second language are often perceived as less distinct from each other than when L1 forms are compared with L2 forms.

Flege (1981), in his Speech Learning Model (SLM), elaborates on the transfer process by incorporating an understanding of these learning theory concepts. In the SLM, Flege hypothesizes that learners undergo an "equivalence classification" of the sounds of their target language comparing them to the native sounds of their L1. When finding a sound which learners perceive as similar to a sound in their L1 they transfer their native sound to the developing L2 phonological set. If the target sound is classified as equivalent to sounds in the L1 then they superimpose their native sound onto their developing L2 phonological set. If the target sound is not classified as equivalent to sounds present in the learners' L1 then they develop a new category for the sound. Accordingly, argues Flege (1980, 1981, 1983, 1989), learners find it more difficult to produce sounds which correspond to sounds in their native language than sounds which are not present in their L1. This is explained as a sort of "data reduction" device (my term) which enables the learner to shift focus from the familiar to the new, concentrating first on the gross

differences before (if at all) moving to the finer discrimination tasks. Flege's model dealt with learner judgment of L2 forms (as either "equivalent" to or "different" from L1 forms) Best, in his Perceptual Assimilation Model (1992) argues that learners' judgment are not categorical but rather graded. In this way, learners classify L2 forms in degrees of "goodness of fit" with their L1 sounds. Thus, each sound of the target language must be evaluated in terms of proximity to native language sounds. Whatever the nature of these perceptual comparisons may be (graded or categorical), what is common to both of these lines of research is the belief that ignoring the finer differences between L2 sounds and corresponding L1 sounds may eventually pose difficulties for learners because they may make their L1 phonetic categories the basis for their L2 speech perception.

In order to link phonological acquisition theory with current perspectives in syntax acquisition studies, researchers in L2 speech acquisition began investigation of error sources from a more process-oriented perspective. As evidence was quickly being amassed from studies in morpheme acquisition order (Larsen-Freeman and Long, 1991), researchers began to ask whether phonological acquisition also proceeded in a systematic fashion. The dominant terms used in syntax studies were carried into the phonological arena. As such, L2 speech errors have been classified as either interlingual or intralingual errors (Dickerson 1975; Tarone 1980). Interlingual variability entails all of the unique characteristics of learner speech which are influenced by the L1 (i.e., transfer). Intralingual variability, on the other hand, entails all of the characteristics of learners' speech which stem from general processes of acquisition (Leather and James, 1986). Thus, an interlanguage (IL) is considered a systematic language in its own right with forms which may be separate from both the L1 and L2.

3 Typological Universals and Transfer

This trend toward process-oriented research did not, however, lay to rest the entrenched views of L1 interference. Instead many of the interlanguage studies have integrated Universalist perspectives with contrastive analysis theories (Cichocki et al, 1993). As such, they presented second language acquisition as essentially a process of transfer directed by universal principles. Eckman (1977) in his Markedness Differential Hypothesis (MDH) argues that universal principles of typological markedness direct the way in which learners transfer forms from the l1. Broselow (1992) gives examples from L1 Arabic speakers learning English illustrating how epenthesis errors (insertion of an unnecessary vowel) cannot always be predicted by analysis of the L1 structure. In such cases, Broselow argues, errors are explained by the universal principle of sonority hierarchy which classifies onset obstruent-sonorant clusters as "unmarked" while-stop clusters are considered "marked". This explains why epenthesis errors made by L2 learners occur most in the less universal s-stop clusters.

Tarone (1980) investigated universal tendencies toward simplifying the L2 syllable structure to an open syllable (CV). She studied Cantonese speakers and Portuguese speakers learning English and found that learners show a preference of either deletion or epenthesis strategies when negotiating difficult consonant clusters and that this preference was influenced by the student's particular L1. Other subjects spoke Korean as their L1--a language with many of the same syllable structures as English. The errors committed by these students were attributed to non-transfer sources since these learners already were accustomed to complex syllable structures from their native language.

B. INTERLANGUAGE VARIABILITY

Central to the concept of interlanguage (IL) is the notion that part of the variability is rule-governed and systematic while part is idiosyncratic, resulting from the individual experiences of the learner. As such, this intermediary form of speech has been called an "idiosyncratic dialect" and "an approximative system" (Larsen-Freeman and Long, 1991). Dickerson (1975) discusses the relative instability of interlanguage forms describing L2 acquisition process as the acquisition of a set of variable rules. A learner may substitute any one of 5 possible sounds for a target sound for a particular environment depending on the stylistic context or perhaps as the result of simply idiosyncratic acquisition processes.

Researchers studying interlanguage normally classify L2 errors as either transfer errors or developmental errors. When the error substitutes an L1 form for the L2 where such a form does not exist, this (over-generalization) is obviously a transfer error. When an error involves the substitution of a form neither present in the L1 or the L2, then this is clearly a developmental error. I have noticed that in the less clear situations, researchers tend to be conservative and classify errors as transfer rather than developmental.

1 Developmental Processes and Strategies

A number of IL studies have focused on the universal strategies that L2 learners use when acquiring a second language (Major 1986, 1987; Tarone 1980; Dickerson 1975; Altenberg and Vago 1983; Macken and Ferguson 1981). Researchers investigate how learners use simplification, substitution, deletion, and epenthesis to negotiate difficult target utterances. Major (1986) investigates the interrelationship between transfer influences and developmental factors over time. In his Ontogeny Model, Major hypothesizes that beginners at first make more transfer errors than developmental errors, and in time the number of transfer errors decreases. Developmental errors, on the other hand, first increase then eventually decrease. To test this hypothesis, he conducted a

longitudinal investigation of English-speaking learners of Spanish. The results validated the hypothesis and revealed a "temporal hierarchical organization" of L2 acquisitional processes similar to the L1 hierarchy as proposed in Natural Phonology (Stampe 1969). According to Major, L2 acquisition begins with transfer, then moves to developmental processes which in turn, are eliminated as acquisition approaches native-like production.

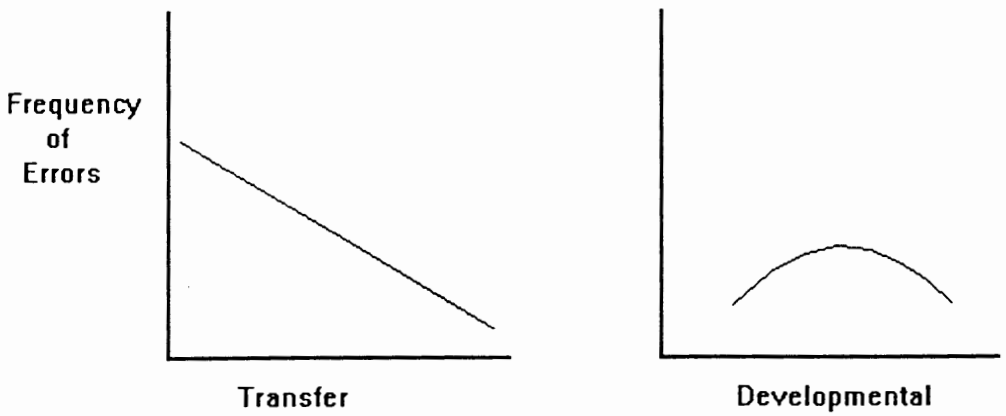


Figure 3.1 Major's Ontogeny Model

Tarone (1980) suggests that both L1 and L2 learners gravitate towards producing open CV syllable structure. Since children tend to rely more on consonant deletion than on epenthesis when simplifying syllable structure, she wanted to see if L2 learners would tend toward deletion as well. Analyzing the speech of English learners from different

languages (Portuguese, Cantonese, and Korean) Tarone investigated the relative proportion of consonant deletion and vowel epenthesis made by the learners. Since the Korean speaking learners already were familiar with complex syllable structures found in English, any errors committed would be developmental rather than transfer in nature. She found that the syllable structures of each L1-L2 pair could not predict strategy preference. For example, if transfer does influence the simplification process, it is not revealed in structural comparisons. Unfortunately, the small sample sizes of the study prevent us from making any strong assertions from Tarone's data.

2 Developmental Substitutions

Not only do second language learners make use of strategies common among children acquiring their first language, they also make some of the same substitutions that children make when learning their L1. (Johannson 1973; Wode 1981; Tarone 1980; Macken and Ferguson 1981; Hecht and Mulford 1982; Altenberg and Vago 1983; Flege and Davidian 1984). Some have asked whether L2 learners re-activate child language acquisition (CLA) processes. It is obvious that L2 acquisition does not replicate CLA in a wholesale fashion. Nonetheless Wode (1983) argues that there are interesting similarities between the developmental products and processes of L1 and L2 acquisition. To support this view, he presents data from both L1 and L2 acquisition showing the same substitutions (between L1 and L2 learners) for 5 varieties of target [r]. He also presents data comparing the substitutions of children learning the English interdental fricative [T] (in their L1) with substitutions made by adult speakers of other languages. He noted that among children, [s] was the major substitution, however for the L2 speakers, the substitution varied systematically as a function of L1, yet each of the L2 speakers substituted phones with acoustic similarities. Wode argues that the crucial property for

each of these was not place or manner of articulation, but rather the "hiss". The Hindi speakers substitute a stop [T] for the target continuant [T] (even though they have [s] in their L1 inventory) and in so doing they maintain the "hiss component".

Learners vary in the substitutions they make according to their language background yet the preferences for a particular substitution cannot be accounted for by an examination of the L1 phonemic inventory alone (Weinberger 1990; Wode 1983). Weinberger (1991) notes that learners of different L1's favor different substitutions for the same target, regardless of the phonemic inventory of the language. For example, French speakers substitute [s] and [z] for the English voiceless interdental fricative [T] and voiced [t] while Russian speakers favor [t] and [d] for the same phonemes even though both languages have [t], [d], [s], and [z]. Weinberger appeals to Underspecification Theory (a phonological theory which reduces phonemes down to a minimal number of features) to account for the speaker's preference.

Certainly such a phonological approach to substitutions seems reasonable, however it does so without reference to psycholinguistic principles such as perceptual salience. As mentioned above, Wode argues that the "hiss" feature (an acoustic feature) seemed to be the guide in selecting substitutions rather than articulatory approximations. This observations is in line with my hypothesis that learners fashion their substitutions not on the basis of articulatory difficulty alone, but also on the notion of "acoustic plausibility".

3 Metalinguistic Knowledge and the Monitor Hypothesis

How learner acquire L2 speech may be influenced by what they know about the second language speech patterns in general. Recently, researchers have been exploring the role of "metalinguistic knowledge" such as an awareness of how to articulate sounds based

on knowledge of articulatory principles (Gombert, 1992). This "meta-knowledge" varies from learner to learner based on their experience. If learners are at least somewhat aware of their productions they will be able to monitor their speaking through kinesthetic and aural sensations.

Unless trained in articulatory phonetics, learners do not focus on the physiological gestures of their speech but on the acoustic product, comparing it with some idealized prototype of native speech. Without this metalinguistic knowledge it becomes difficult to acquire new sounds in a second language. Since there may be a variety of gestural features which may correspond to a particular acoustic feature (Straight 19809) learners may be able to perceive the L2 sound accurately (Mochizuki-Sudo 1991) and yet have trouble articulating the features they hear.

The process of learning to produce a new sound not found in the L1 must be tied to our capacity to "mimic" sounds resident in the auditory memory. Whatever the representation of this auditory model may be, learners must control their own articulatory gestures in a self-aware fashion so as to approximate the auditory target. Leather & James (1991) discuss the role of various sources of feedback in learning a new set of articulatory gestures. Feedback from kinesthesia of articulators, changes in bone-conducted pressure within the cranium, as well as external acoustic projections all assist the speaker in modifying the speech production. It is suggested that motor mechanisms are of two types: (1) pre-planned, centrally driven "open-loop", and (2) moment-to-moment feedback regulated "closed-loop". Open-loop control involves instructions for the motor gestures while closed-loop control involves the feedback from the various "head-internal" sources mentioned above.

These constructs have given rise to the hypothesis that as speakers gain experience in the new sound productions, their articulation comes to be guided less by closed-loop

control and commensurately more by open-loop control (Straight 1980, p.316). According to this model, the acquisition of native-like pronunciation would entail some cognitive linking between the perceptual system and the productive system. This linking may come through trial-and-error comparisons of the learner's output with his or her internalized models.

4 Compensatory Strategies and Substitution

Working with kindergarten children I have observed some interesting relationships between CLA and SLA. Normally, children require the perceptual discrimination of a speech sound before they are able to produce the sound (Straight 1980). There comes a point in the child's development (about age 5-6yrs) when their perceptual categories are almost completely developed, yet they lack the ability to produce a few of the consonants such as the approximates [r], [l], and fricative distinctions [T] and [s] which may not be acquired until roughly age 7-8.

I have noticed some interesting compensatory strategies which the children employ during this period to make their speech intelligible until they are physiologically able to articulate the target sounds. The substitutions that children make are not merely phonetic sounds that are part of their attained linguistic repertoire, but rather inventions which acoustically approximate the target sounds. For example, kindergarten-aged children (5-6 years old) often substitute invented vowels for the liquid phonemes [r] and [l] word-finally. Thus, we may see children make the types of substitutions such as are shown in Table 3.1. An inventory approach does not explain this phenomenon (Wode 1983) since the children use [r] in other contexts.

Table 3.1 Substitutions Made by Children

[3r] --> [& or oU] word finally				
	"there"	/D eI &/	or	/D eI oU/
	"here"	/h i: &/	or	/h i: oU/
[l=] --> [oU] word finally				
	"bottle"	/b A T oU/		
	"table"	/T eI b oU/		
	"people"	/ph i: ph oU/		

When examining the spectral patterns of these substitutions, one notices remarkable similarities between acoustic shape of the approximation and that of the target. However, an articulatory analysis of these fails to account for their regularity. Beyond sonority, there is little similarity between the articulatory features of these vowels and their target liquids. I would hypothesize that second language learners also undergo a similar process of approximating the auditory target when attempting to produce a new sound. It seems that learners are able to engage a wide range of articulatory gestures in order to achieve some a acoustic plausibility in their substitutions. As mentioned before, the acoustic features seem to influence learners' choices in substitutions more than articulatory features. (Native language Hindi speakers substitute the strongly aspirated [T] for an English [s] even though they have a quiet fricative [s] in their L1.)

C. SUMMARY

Much of the work done in IL theory has been done from a phonological perspective examining phonemic processes and phonotactic contexts. Often the conclusions presented are not new findings but rather a re-analysis of the same principles

formerly examined by transfer perspectives. For example, Major (1986) classifies developmental errors as the substitution of an L1 sound for an L2 sound when the substitution places the native sound in a context not possible in the L1. Thus, at the phonetic level, no "development" needs to take place in order to be called "developmental acquisition". Certainly such an explanation is reasonable, but it adds little to our overall knowledge of acquisitional process.

A large gap exists within IL research. Some studies recognize non-transfer development only at the phonological level. A few studies recognize that non-transfer development exists at the phonetic level but is minimal (Johansson 1973). Yet, to my knowledge, no one has investigated the process of acquiring new sounds at the phonetic level. Wode (1983) argues that in order to discover why speakers of various languages show preferences for certain substitutions which are independent of structural factors researchers "will have to give prime attention to phonetic substance." It may be that an analysis of acoustic features provide better explanations of transfer and development than do the former articulation-based feature analyses. Perhaps s such acoustic features have more perceptual salience for learners as they focus their pronunciation on target forms.

In my opinion, a more significant area of inquiry would deal with the phonetic acquisition of new sounds and the substitutions that learners make along the way. A number of researchers have begun to use acoustic instrumentation to study the phonetic approximations (Flege 1980, 1981, 1987a, 1987b, Flege and Davidian 1984; Bohn and Flege 1992; Munro 1993). In these studies they often find intermediary forms which belong neither to the L1 nor to the L2 nor can they be said to be simply linear approximations of the target (although few comment on this in their analysis). For example, Munro 1993) shows vowels articulated by native Arabic speakers which appeared to move away from both the L1 and L2 forms.

PART II:

TOOLS FOR ANALYSIS

4

Analyzing the Speech Signal

Denes and Pinson's foundational work The Speech Chain (1993) describes speech as a model of encoding and decoding with the acoustic signal as the message. In Chapter 2, I discussed theories of speech production and perception which comprise the encoding and decoding of the signal respectively. Each of these speech modes is complex and presents many challenges for the building of accurate models. However, before such work can begin, a thorough understanding of the *signal* is needed. In this chapter I will give a brief definition of terms used and then turn to a description of well-established methods of analysis.

A. DEFINITION OF TERMS:

1. The Signal

Sound waves are most often graphically represented as sine waves although it is understood that the true nature of wave propagation in gas does not resemble a rising and falling motion such as through a liquid medium, but rather the compression and rarefaction of atmospheric molecules. Nonetheless, sine wave representations provide simple models of the actual phenomenon.

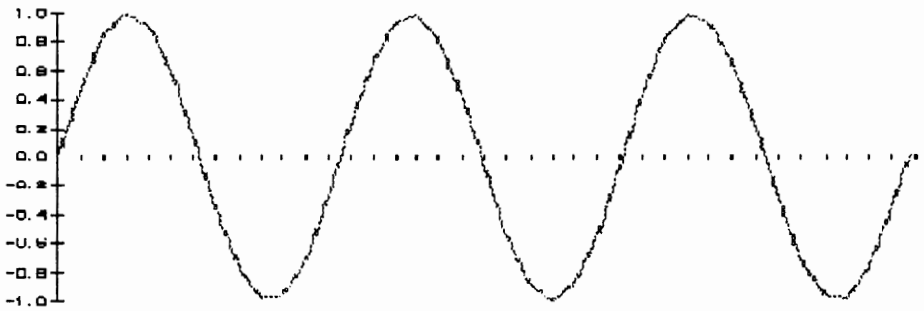


Figure 4.1 Sinusoidal Waveform

The focus of this study, however, is on the spectral structure of *complex waves*. Sound waves rarely occur as simple *sinusoidal* patterns (as depicted in Figure 4.1). Instead, waves overlap each other and interact with each other giving signals a complex structure. At times this structure has observable regularities called *periodicity*. Often, however, sound waves are *aperiodic* being characterized by random noise. Human speech exploits this basic distinction of sound patterns in order to distinguish the broad phonetic categories of sonorants versus fricated segments (see Figure 4.2).

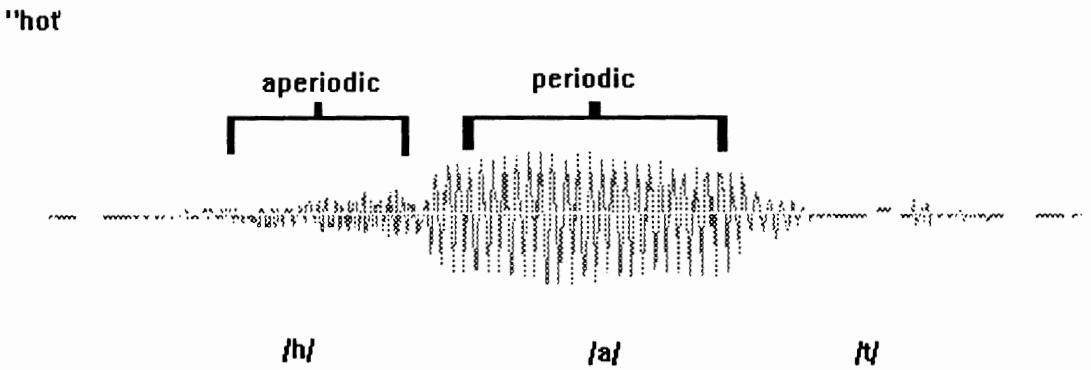


Figure 4.2 Complex Waveform

The *spectrum* of a complex wave specifies the amplitude, frequency, and phase of each of its internal component waves. However, since the particular phase of waves within the spectrum is normally imperceptible to humans, it is common to speak of only the amplitude and frequency when referring to the speech "spectrum" (Denes and Pinson, 1993). At any given point in time along the waveform, the spectrum can be represented as a "spectral slice" showing amplitude (often referred to as "intensity") on the y-axis and frequency on the x-axis. Figure 4.3 shows a cross section of the phoneme [i:] occurring in natural speech. Many of the meaningful acoustic variations within the signal take place not as amplitude variations but frequency variations (O'Shaughnessy, 1987). Two very distinct time waveforms of the same utterance by a speaker can have similar frequency spectra. In the following section I will describe methods of analysis which highlight various aspects of the frequency parameter.

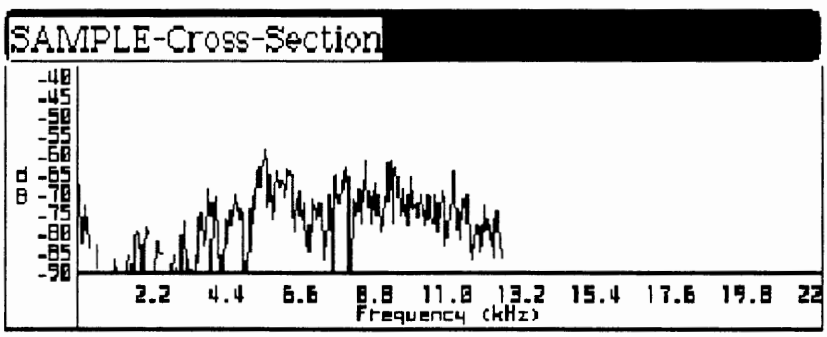


Figure 4.3 Spectral Slice (Sample output from CSRE software, series 4.2.)

2. Digital Sampling

Complex sound patterns can be represented digitally as a series of numeric values. Analogue sound waves can be "sampled" and converted to digital representations using Digital Signal Processing (DSP) techniques. Once converted to digital format, the signal can be easily manipulated and is available for any analysis procedure. This analogue to

digital (A/D) process can occur at various rates. The Nyquist Theorem specifies that a signal can be adequately represented by a sampling rate at least twice the frequency range of the signal. Thus, in order to capture a signal with components as high as 20kHz, it is necessary to sample at 40kHz.

B. ACOUSTIC FEATURES OF SPEECH

1. Time, Amplitude, and Frequency

The speech signal exploits each of the parameters time, amplitude, and frequency to produce a robust range of sounds. As mentioned in chapter 2, each of these features is subject to perceptual warping. The speech signal remarkably exploits the ranges of highest hearing acuity for each of these dimensions. Accordingly, speech amplitude levels fall between 60-70 dB in normal conversations (at 3 feet apart) (Denes and Pinson, 1993). Most of the linguistically meaningful information falls between 100-3200Hz.

2. Broad Phonetic Categories

Observations of the speech spectrum reveal a number of vary distinct categories of speech sounds: clusters of random noise, intermediate pauses within the signal, robust segments of multi-componential time-varying sound, and multi-componential segments with strong low-energy concentrations. Each of these represents a "broad phonetic category" roughly corresponding to the articulatory dimension "manner of articulation". The articulatory dimension "place of articulation" is most directly evidenced in the various frequency locations of the spectral energy for each of the mentioned "broad phonetic categories". I will briefly describe these categories in the section that follows.

Stops are clearly identified in continuous speech as momentary pauses usually between 30 and 100ms. The articulators may still be in motion but most of the acoustic signal is cut off in the oral cavity. Sometimes, however, a low frequency band of energy derived from voicing can be seen. Voicing contrasts in stops are most clearly identified by the length of the "plosive" portion of the segment (i.e., the stop release). Voiceless stops show long periods of high frequency frication following the occluded segment. Voiced stops have a very brief, sometimes unnoticeable release.

Fricated segments are displayed as clusters of random noise. Place of articulation, as with all of these categories, is evidenced in the relative frequency band covered by signal (here, the random noise). Also, alveolar fricatives are most always louder than labials. Affricates appear as fricatives preceded by stops except with a sharp division between the stop portion and the fricated portion (i.e., the release).

Nasals show much more acoustic structure. This category is characterized by a low-frequency band of energy with faint formant patterns at higher frequencies. The majority of the bottom half of the spectrum (assuming analysis focuses on the first 4kHz of the signal) is filled with scattered low energy.

Finally, the category "sonorants" includes the linguistic categories of vowels, liquids, and glides. The major feature of this class of sounds is the presence of "formants" shown as strips of high energy with surrounding spectral decay. The relative positions of these bands evidences the various linguistic phonemes being uttered. Glides are evidenced by formant transitions. The presence of formant transitions, however, is not a unitary feature of glides since transitions can also be seen on coarticulated vowels if they are located between consonants with extreme points of articulation.

3. Formants

That these sub-phonetic features contribute to vowel identification is well established by research. Particularly, the first three formants (F1, F2 and F3) are important to the linguistic identification of vowels and other vowel-like articulations. Of these, F1 and F2 closely relate with the acoustic features of "height" and "front-back" relations among vowels. F1 is inversely related to tongue height while F2 is related to frontedness. As mentioned in chapter 2 these articulatory and auditory features are not entirely perfectly related, nevertheless, there are enough similarities to make features predictable.

C. SPECTRAL ANALYSIS

1. The Spectrogram

The term "spectrogram" comes from the original electro-magnetic mechanism produced in the 1940's for giving visual output of spectral information. Over the decades it has been used to refer to a variety of different tools designed with the same purpose in mind. Today, digital techniques make the process of generating visual output of sound quicker and preservable. (Originally, the spectrogram wrote to a phosphorous belt which would eventually be over-written on the next cycle). All of these techniques plotted Frequency against Time with Amplitude as darkness in the plot. There are various methods for analyzing the complex signal once it is in digital format. The most common techniques are the Fast Fourier Transform (FFT) and the Discrete Fourier Transform (DFT) named after Jean Baptiste Joseph Fourier. The FFT is based on the principle that any non-sinusoidal signal can be represented as the sum of its component sinusoids.

The spectrogram computes its output in a series of overlapping "**windows**" of the signal. Since motion of the articulators is relatively slow compared to the degree of granularity (resolution) contained in the sampled signal, the spectrogram need only to compute the signal parameters about once every 10 ms. This rate varies depending on the speed of the meaningful signal changes. For example, stop bursts occur in a very short window of time.

Consequently, an analysis window must be small enough (i.e., sampled often enough) to identify the timing of the burst. Spectrograms with long time windows often display "ghost bursts" which actually spread portions of the burst out across the previous window. Also, the analysis of low frequencies requires long windows which reduces the granularity of transition representations. Specifying the sampling window size (this time "sampling" refers to the extraction of portions from the digital signal file, rather than the actual audible signal) achieves two goals: 1) maintains an economy of data transformation and thus speeds up the processing time for the computer, and 2) makes it possible to highlight either frequency resolution or time resolution.

Thus the primary dilemma presented by an FFT analysis is that the researcher is confronted with a tradeoff between good frequency resolution and good time resolution. This stems from the inverse relationship of spectrogram bandwidth and window size (the amount of time between samples). Good frequency resolution is provided by a long analysis window (narrow band spectrograms), while good timing resolution is provided by a short analysis window (broad band spectrogram). Vowels are best analyzed with narrow band spectrograms while stops and other highly time-variant segments should be analyzed using broad band spectrograms. Also, long windows are needed in order to capture low-frequency spectral information. This is because the sampling algorithm must wait for at least the duration of the longest wave period to complete its cycle if that base wave is to

be analyzed as a complete wave. In other words, to capture a 100Hz signal, the sampling window must be at least 10ms long.

When using digital equipment, window sizes and bandwidth are measured in bytes. Thus, the possible window sizes increment by factors of 2 somewhat limiting the choices when the signal has already been digitized. Greater flexibility in determining window size can be obtained by changing the sampling frequency of the A/D conversion. The relation is expressed as follows:

$$\text{Window length (ms)} = \frac{\text{Bandwidth (bytes)}}{\text{Sampling bit-rate (kHz)}} \qquad (4.1)$$

Thus, for a signal sampled at 44kHz, a 256 byte/sec. (83Hz) bandwidth yields a window of approximately 6 ms. Vowels should be analyzed with a window of 15-25ms. while bursts and other segments yielding sharp spectral changes require a window of 3-6 ms.

One way to mediate between frequency resolution and time resolution is to vary the amount of overlap between windows. Increasing the overlap causes a smoothing of the frequency resolution without necessarily increasing the *frame rate* (the number of times per second that the analysis is performed). Another way to confront the problem is to change the shape of the window. Most commonly a rectangular window is used for computational simplicity, however windows with tapered edges (i.e., Hamming, Hanning, Blackman, Bartlett, and Kaiser windows) are also used. The effect of the tapered edges is to emphasize the middle portion of the window causing the algorithm to become sensitive to sharp spectral changes (in frequency or energy) while maintaining a long enough window to provide good frequency resolution.

2. Formant Analyses

The precise location of formants is the goal of many lines of investigation in speech science ranging from the building of speech synthesis models to making measurements for acoustic phonetic research. The Fourier transforms are very useful for acquiring a kind of "gestalt" perspective of the entire wave spectrum, however the method makes precise identification of formant bands difficult since there are often many spectral peaks displayed which do not correspond to vocal tract resonances (formants). Also, the thick formant bands reconstructed by Fourier transform spectrograms make it difficult to find the precise frequency of the formant.

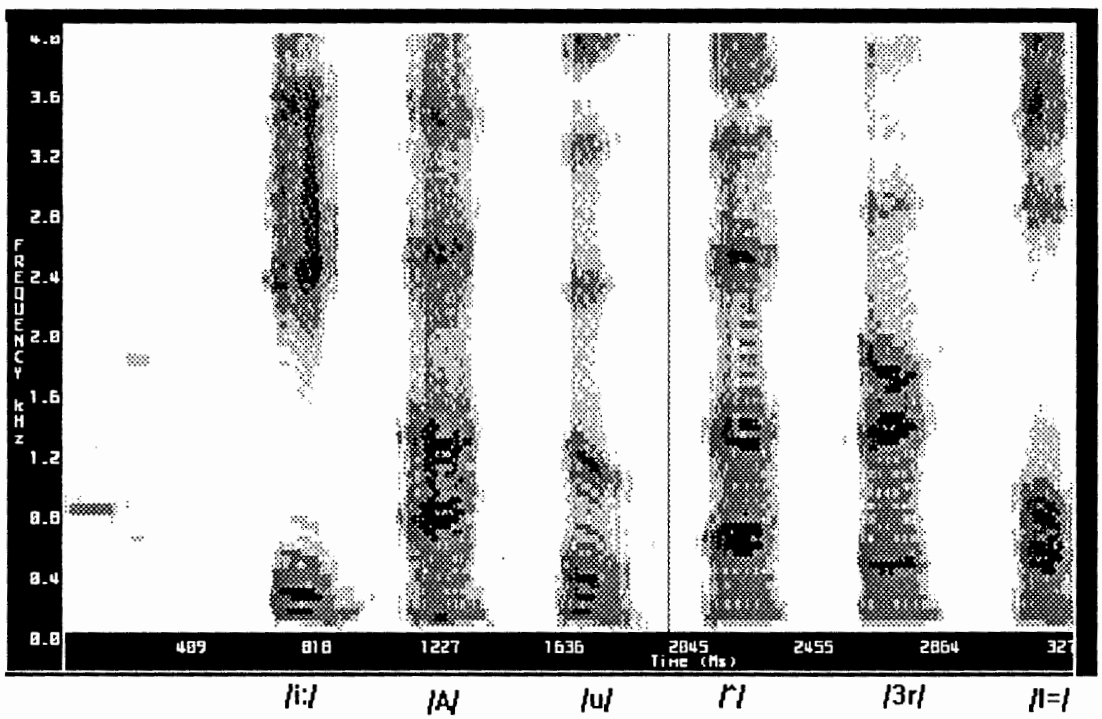


Figure 4.4 Fast Fourier Transform (Sample output from CSRE software, series 4.2.)

A common method of *peak-picking* involves simply identifying the central frequency of the formant band as the formant peak. Such methods allow an accuracy of within +/- 60Hz for F1 and F2 but accuracy degrades to +/- 110Hz for F3. This, however, makes the assumption that the spectral skirts of the formants are symmetrical--an assumption which is rather inexact. "The automatic tracking of formants has been an elusive task despite the typical spacing of formants every 1kHz (for a vocal tract length 17 cm long), the limited range of possible bandwidths (30-500Hz), and the generally slow formant changes." (O'Shaughnessy, 1987, p225). Ultimately, the investigator must concede that the notion of "formant" is a construct and not an isomorphic acoustic feature.

An alternative approach involves smoothing the spectrum so that the peaks are more easily resolved. Linear Predictive Coding (LPC) is a commonly used technique belonging to a class of *analysis-by-synthesis* approaches. Among theses approaches, the goal is to estimate a set of parameters (*analysis*) from a theorized speech model. (*synthesis*). Specific mathematical details can be found throughout the literature. Essentially, the LPC speech model consists of a glottal source excitation as input to a vocal tract transfer function. The economy of the LPC analysis comes from the computed *flattening* of the glottal source excitation so that the effects of oral resonances are highlighted. Note in Figure 4.5 how the spaces between the formants are empty while in the Fourier transform shown in Figure 4.4 exhibits much low energy scatter. The difference in the LPC is that the signal is treated as an all-pole model and assumes no zeros (treats the zero's as noise). This assumption, however runs into conflict when attempting to capture the anti-resonances characteristic of nasalization. LPC's are very effective for modeling vowels and other sonorant segments.

When "predicting" the output parameters of the signal, the LPC requires a pre-stated number of coefficients based on the number of anticipated formants within the

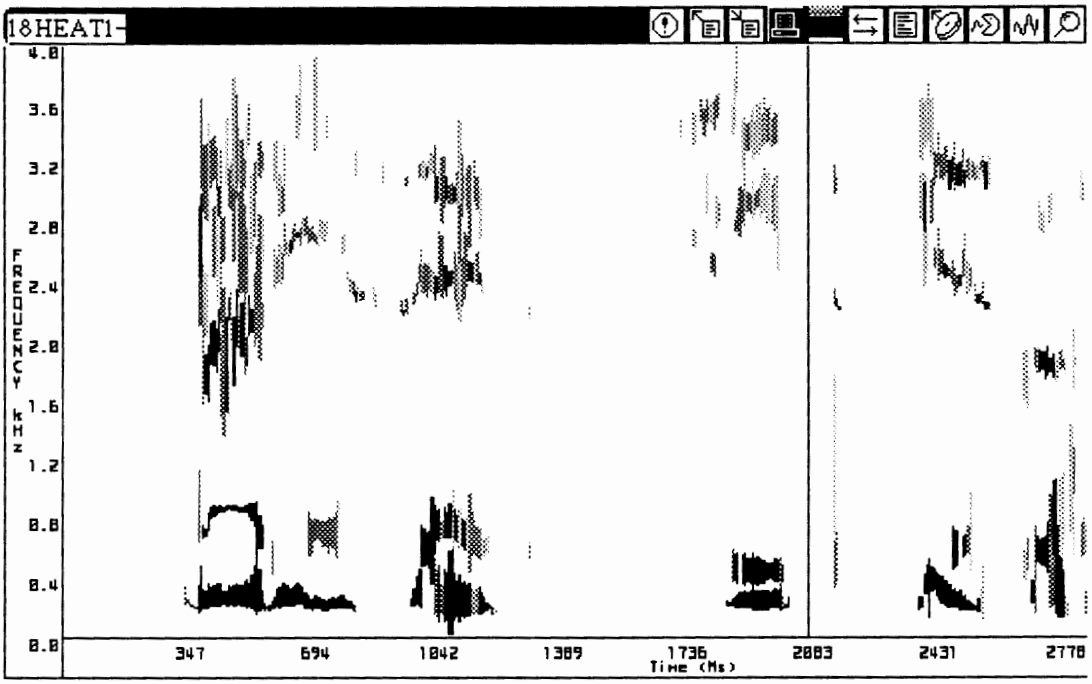


Figure 4.5 Linear Predictive Coding (Sample output from CSRE software, series 4.2.) Analysis of a sentence "Larry said *heat* yesterday".

analyzed bandwidth. This number is best altered depending on the size of the analyzed vocal tract (i.e., differences between males, females, and children) since male formants lay lower on the frequency scales and thus are apt to fit more formants into the analysis bandwidth. Each of the vocal tract resonances requires 2 coefficients plus the model requires additional coefficients to account for the possible zeros in the spectrum as well as effects from glottal and lip radiation. Thus, for an analysis bandwidth of 4kHz, 11-12 coefficients should be typically be used for females and 12-14 for males.

One of the challenges to locating formant peaks with low energy has to do with the irregular forms of spectral decay present on various speech segments. Glottal source radiation and additional radiation at the lips confounds the interpretation of oral resonants. Some LPC algorithms *pre-emphasize* the signal prior to performing the

computation. This effectively reduces the dynamic range by cancelling the low frequency "rolloff" caused by the glottal and labial radiation. The result is that low energy peaks are better represented.

5

Considerations in the Acoustic Study of Vowels

Acoustic phonetics, like all scientific disciplines has its repertoire of analytical tools. Through decades of use, researchers have explored a variety of ways in which to present data efficiently and accurately. In this chapter I will begin by defining the notion of "vowel" as it relates to the measurement of second language speech and then proceed to an explanation of issues surrounding appropriate methods of vowel comparisons describing one of the most commonly used vowel representations--the Vowel Space Diagram (VSD).

A. VOWELS DEFINED

The study of vowels has proved important to a number of different scientific disciplines including psychology, acoustic physics, linguistics, computer science, and engineering. Each discipline has focused on different aspects of vowels and consequently, the term is often ambiguous for those doing cross-disciplinary research.. It is very important that these various perspectives be kept separate because many of the inherent assumptions of each are not compatible across frameworks. The linguist, who is

interested in description, analyzes vowels as unitary target segments each with its own unique quality and articulatory features which can be located in time. An acoustic analysis of vowels will yield vague (if not non-existent) boundaries circumscribing vowels. Because the acoustic parameters defining vowels overlap each other from speaker to speaker. Psychologically, the evidence shows that the notion of vowel involves more than an identifiable unit in time, but rather a base unit whose identity lies partially in the information provided by contiguous segments. The computational engineer building speech recognizers may borrow elements from each of these perspectives and integrate them in a manner most conducive to his or her design approach.

For the purposes of this study, the term "vowel" shall refer to the linguistic phonetics notion which posits that phonemic vowel targets exist for every language and that these targets differ from language to language, and that the set of vowels of a given language are form a system of phonemes sensitive to perceptual and neuromuscular information integrated to serve the goals of *sufficient perceptual separation* and *economy of effort* (Ladefoged, 1993). These idealized vowel phonemes are subject to phonological transformation processes that serve these goals to make words distinct and allow for ease of articulation.

B. VOWEL COMPARISONS

1. Vowel Positions and Spread

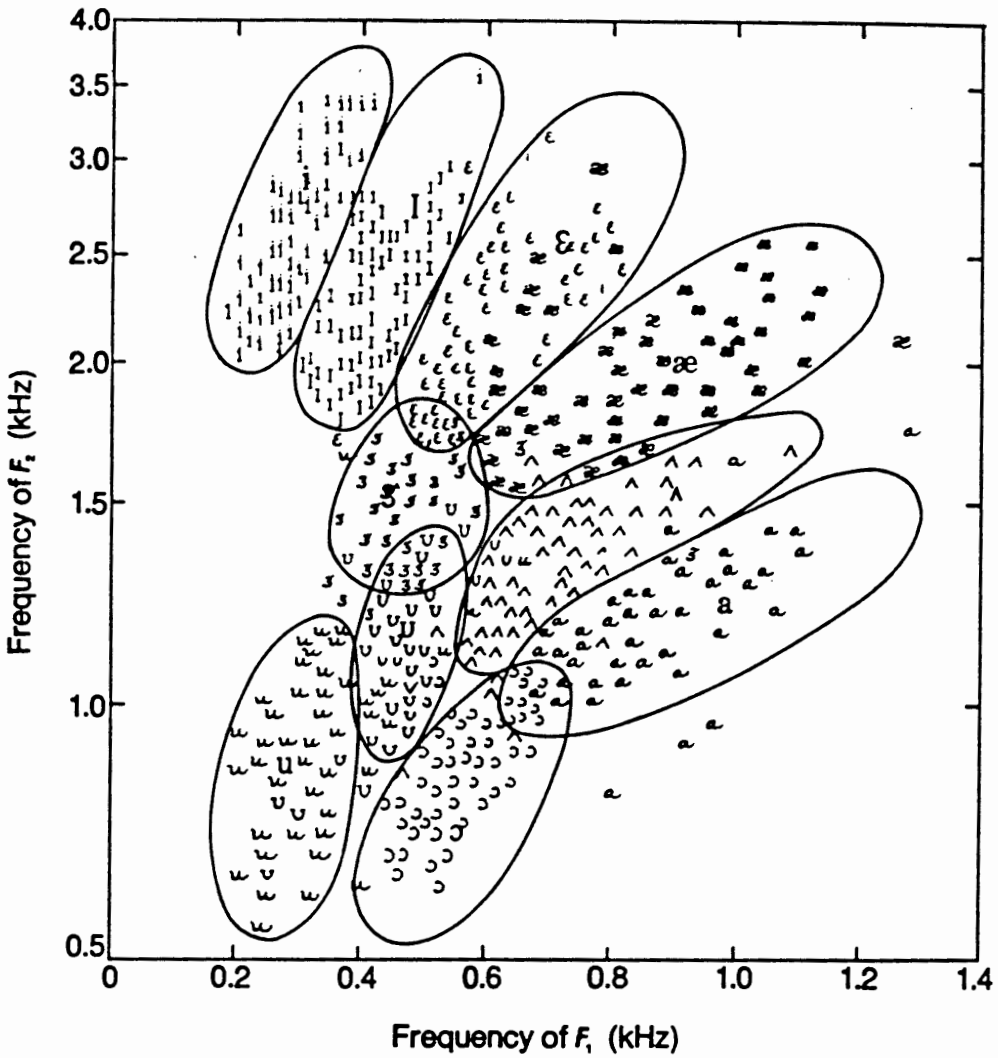
The Vowel Space Diagram (VSD) is essentially a plot of the relative acoustic qualities of vowels and locates the vowels in the acoustic space. Relationships on the vowel charts involve two dimensions: the F1 value on the ordinate and, depending on the

kind of plot, the F2 value or the F2-F1 value on the abscissa. If we conceive of the VSD as analogous to a vowel production chart, we could imagine that it resembles the cross-section of a speaker's mouth with the lips to the left of the diagram and the back of the mouth to the right. It is important to point out, however, that VSD's, like vowel charts, do not accurately represent tongue positions of articulated vowels (Ladefoged, 1975). Instead, these diagrams model vowel quality as measured by formant values and only indirectly do they represent the proportions within a speaker's mouth.

The purpose of the VSD is to model the acoustic space in which vowels occur. The most direct way to do this is to represent the axes with a linear scale such as cycles per second (cps), commonly called Hertz (Hz). This scale is based on observable physical phenomena--the cycles in a sound wave. Using the Hertz scale requires no additional computation. However, it does not accurately represent relationships between pitch components as perceived by the human ear. The most common way of resolving this problem is by plotting the frequency values on a logarithmic scale. In this way it is easy to get a visual image of the auditory proportions. This is very useful for VSD's when the sole purpose is to examine relative vowel positions on a chart.

A study of vowels must begin with an understanding of vowel quality (and indirectly, vowel positions). As mentioned earlier, acoustic measurements enable us to chart spoken vowels on the auditory space with high precision. Vowel charts also allow researchers to observe trends across the auditory space. When multiple levels of information are plotted on the same chart, their power becomes even more apparent. In chapter 8, I have plotted various cross-linguistic vowel patterns. Caution needs to be exercised, however, when making conclusions about articulation solely from the VSD. This is because a number of intervening factors such as lip activity and tongue root position influence the location of vowels on the diagram. Statements made about

Figure 5.1 Vowels of American English from *Peterson and Barney (1952)*



articulation tendencies based solely on data from the VSD could be inaccurate. However, with information provided by acoustic theories of speech production (see chapter 2), much of the ambiguities can be resolved. Since Stevens and House (1955) research has repeatedly confirmed the predictability of formant values from gross positions of the articulators though there remain many gaps in our understanding of the finer elements of production.

Originally, VSD's consisted of a simple plot of F2 against F1. When presented on a logarithmic graph this gave a reasonably accurate view of the auditory relationships between vowel points. Later it was found that plotting the difference between F2 and F1 on the horizontal axis more closely approximated the frontedness differences between the back vowels as revealed in X-ray studies. Accordingly I refer to plots of F2xF1 as "auditory plots" and plots of (F2-F1)xF1 as "articulatory plots".

2. Distance Between Vowels

Understanding distance relationships among vowels is important for describing intact vowel systems as well as more transient phenomena such developing IL vowel systems. The ability to properly and accurately measure such distance relationships enables researchers to ask questions like:

How symmetrical is the overall vowel system of language X?
How similar are the vowels of language X compared to the vowels of language Y?
How closely does learner X approximate the sounds of the target language?
How much has learner X improved in his or her approximation of the target?
Which does the learner's IL vowel more closely approximate: the target language or the native language?
Which group of learners most closely approximates the target vowel(s)?

2.1 Distance and Auditory Warping

While linear frequency data (in the Hertz scale) can be accurately presented on a logarithmic scale, any statistical comparisons between formant positions require that the data be pre-warped. For instance, the study that follows compares error distances of accented vowels from the target form. For the sake of illustration, let us imagine that in approximating the English [i:] and [u] vowels, speaker X tends to deviate about 150Hz from the English targets on both vowels. The standard[1] English F2 values for these vowels are 2290Hz and 870Hz respectively. Even though the error distance was 150Hz for both vowels the approximation of [i:] is perceptually further from the target due to the spectral warping of the higher frequencies. The 150Hz error on [u] had much less effect because it involved the top front of the vowel space where F2 values are low whereas the error distance on [i:] where F2 is high. A more accurate numeric comparison of these frequency values requires a conversion of the data to a logarithmic frequency scale--the two most common of which are 1) the Mel scale, and 2) the Bark scale. Chapter 2 provides a summary explanation of perceptual warping and these psychoacoustic scales.

2.2 Pythagorean Distance

When comparing the positions between two vowels on the vowel chart, it is important to be able to consider both F1 and F2 together. Often in vowel studies these two values are separated, but that approach is somewhat artificial because in reality the two are interdependent. While the movements of F1 and F2 are strongly related to tongue motions across vertical and horizontal axis within the mouth, there are many other factors which confound this relationship such as lip rounding or compression and tongue root motion. Thus, to make simple measurements of formants individually is to pull the data even one more level away from the actual physical phenomena operant in the oral cavity.

The distance between two vowels can be calculated using Pythagorean distance. In this method, distance is solved for as the hypotenuse of a right triangle using the formula $\mathbf{a^2+b^2=c^2}$ where F1 and the difference between F1 and F2 are **a** and **b**. In this way we may treat IL development as a sort of orbit around the target form. Linear distance from a target can be measured without reference to direction. The importance of this will be seen in chapter 8 when we analyze IL vowels and their relative approximative distances which may not move directly toward the target.

3. Variance Among Vowel Utterances

Central to the questions of a number of theories in acoustic phonetics is the notion of variability. Speakers vary their speech from utterance to utterance. The same utterance will undoubtedly vary from speaker to speaker. Using simple statistical measures of variance it is possible to make valid comparisons either across groups of speakers (inter-speaker variability) or across a set of utterances by a particular speaker (intra-speaker variability). In this study, a number of questions focus on the notion of variability.

How large is the overall vowel space of speaker X (or group X)?
How large is the overall L2 vowel space of speaker x (or group X)?
Do the learners have greater variance than native speakers?
Would a group of inexperienced learners have a greater variance than experienced learners?
Would a linguistically heterogeneous group have a greater variance than a homogenous group?

FOOTNOTES:

[1]Here "standard" refers to the average formant values for a population of native-proficiency speakers. While some contest the theory that there exist phonemic primitives within the linguistic competence of speakers, many vowel studies since Peterson and Barney (1952) have proceeded on the assumption that vowel targets can be represented as the average acoustic value for the population of native speakers.

6
Methods and Procedures

Chapter 4 was concerned with an explanation of widely used analytical measures and chapter 5 dealt with an acoustic theory of vowels and important considerations for their measurement and analysis. In this chapter I will briefly outline the procedures executed within this study highlighting any methodological choices made at each step.

A. DATA COLLECTION

1. Groups:

The research involved two groups: one very homogenous group of Japanese students and one very heterogeneous group of bilinguals with a variety of first languages which I will refer to as the "mixed" group. Subjects from both groups were given $3.00 gift certificates for participating restaurants near the university. The Japanese group consisted of 11 females between the ages of 19 and 21 years old who were part of an exchange program. The students lived in campus housing. None of the speakers spoke languages other than English and Japanese. This was the first visit to the U.S. for all of the speakers and none had spent time in other English speaking countries. All of the students were from the same region in Japan and spoke the same (Western) dialect of Japanese. Table 6.1 shows the Japanese scores on an index of English proficiency. It is important

to note that the value for "years speaking English" may have been interpreted as "years studying English" and therefore may be, when taken by itself, an unreliable measure of proficiency for the Japanese who rarely use English outside of academic language class settings when in their home country.

Table 6.1 English Proficiency Indicators for the Japanese Group

Years living in the U.S.	mean=0.2 (SD=0.00)
Age began learning English	mean=10.04 (SD=2.00)
Years of college completed	mean=2.36 (SD=0.67)

The mixed group began with 11 males and females. Three cases were omitted from the data because they had substantially more experience than the rest of the group either in years speaking English or in years living in the US. One additional subject was omitted because of difficulty reading the prompts correctly. After these outliers were removed, the group total was 7. Ages ranged from 19-26 years old. Six different mother tongues were represented: Chinese, Spanish, German, Indonesian, Vietnamese, and French. The subjects came from five different countries. Two of the speakers were trilingual (including English).

This is not a true control group since it is not similar to the test group in the amount of exposure to spoken English. The Japanese group had been in the US for only 2 months while the average time spent in the US by the mixed group was 6.7 years. Nevertheless, it is helpful to compare the Japanese group and the mixed group with native

speaker values in order to get a more complete picture of second language acquisition across time and language backgrounds.

Table 6.2 English Proficiency Indicators for the Mixed group	
Years living in the U.S.	mean=6.7 (SD=3.70)
Age began learning English	mean=12.1 (SD=1.34)
Years speaking English	mean=8.0 (SD=3.29)
Years of college completed	mean=3.3 (SD=1.98)

The relatively small size of the groups is not unique to this study. Linguists have repeatedly resorted to smaller groups for similar studies, due to the volume of data that must be analyzed. The largest study ever made of vowel measurements was done by Peterson and Barney (1952) and their study involved three groups: 33 men, 28 women and 15 children. While the size of these sample groups may not be sufficient to make categorical statements about the populations in general, they certainly represent a sample of the groups being studied and therefore are useful for giving reliable indications about the larger populations.

2. Elicitation Procedures

The subjects were asked to read a list of English carrier phrases containing monosyllabic [CVC] words exhibiting the vowels /i:/, /I/, /eI/, /E/, /@/, /A/, /oU/, /U/, /u/, /3r/, and /u/. A description of the phonetic notation used in this paper can be found after the table of contents. Copies of the test lists can be found in Appendix 1) . Each phrase was spoken twice, yielding two vowel tokens. The purpose of the carrier phrase was to

normalize the rate and amplitude of the test words. This has been a standard procedure for measuring vowels since the Peterson and Barney study (1952). Of course, the ideal situation would be to use open [CV] syllable words, however, English lacks such words containing the lax vowel phonemes such as /I/, /E/, /@/, /U/, and /^/. Thus, it was necessary to measure the vowels as uttered in the [hVt] and [hVd] monosyllables in order to control for the effects of coarticulation influenced by the surrounding consonants. Since the syllable final /t/ and /d/ segments have the same place of articulation we can assume that coarticulation will produce a standard effect on every vowel.

The Japanese subjects were asked to read the Hiragana syllabary--a list of the 46 essential writing units (kana) made up of the 5 primary vowels and the 41 mora (roughly speaking, "syllables") each consisting of a consonant and a vowel (except the mora /n/). Of these, only the vowels and the kana beginning with /h/ were used for measurement. Reading the Hiragana characters proved to be a convenient way to minimize coarticulation effects. Since the primary measurements were concerned with vowel quality and not length, long vowels were not elicited (there is no difference in place of articulation for short-long vowel contrasts).

Additionally, two other sets of data were collected during the elicitation sessions (same 22 speakers under identical conditions). One set included a list reading of 55 monosyllabic [cVc] English words exhibiting 5 tokens of each of the 11 English vowels mentioned above. The other set consisted of three paragraphs of an ESL text book read by each subject totalling approximately 96 minutes of high quality speech data. Unfortunately, time constraints made it impossible to analyze these data sets.

B. INSTRUMENTATION

Word lists were recorded in an anechoic, noise reducing recording studio. Subjects were seated at a table with a suspended microphone approximately 12" away from their lips. Word lists were placed on the table in front of the subjects. Recordings were made using a Neuman U-47 bi-directional microphone. Sound was then channeled through a Tascam M-50 mixer board directly (bypassing the equalizer) to a Technics 4-Track 1506 Reel-to-Reel analogue recorder (isolated loop/direct drive/tension control) at a tape speed of 7.5 ips. The recordings were then played back on a Sony 3-Head Stereo Reel-to-Reel recorder at the same rate. This signal was channeled through a Phillips 900 Series Integrated Stereo Amplifier set at minimal (-30dB) amplification. From there the signal was directed (through a 20dB cord) to a Tucker Davis Technologies amplifier on a DSP rack. The TDT amplifier gain was modified to account for varying speaker volumes so that the signal throughput to the A/D board would always be at a level between -20dB and -10dB.

The analogue-to-digital conversion was done directly on the output from the TDT DSP rack. The board and software for this A/D process were part of the CSRE 4.2 package (CSRE being an acronym for Canadian Speech Research Environment) which ran on a Gateway 2000 66MHz 486 PC. Sampling rates, methods of analysis and specific parameters of the computations were all driven by the CSRE software.

C. SPECTRAL ANALYSIS

1. LPC Analysis Parameters

Except for quick FFT references to aspects of a particular speech segment, the data were all analyzed using an LPC algorithm provided in the software. Naturally, decisions had to be made between degrees of frequency resolution and time resolution. (Chapter 4 provides a discussion of these issues) Best results for optimal frequency resolution were attained using a 1024 byte (approximately 23ms) Hanning window corresponding to a 512 byte/sec. (approximately 86Hz) bandwidth with 70% overlap and a 98% pre-emphasis. Although other studies have found better results using more coefficients for modeling male voices, I found that 13 coefficients provided optimal results for both sexes. These analyses were performed on speech sampled at 44kHz. This high sampling rate was unnecessary. Adequate results could have been obtained at 8-12kHz since only the first 3 formants were measured, all of which typically fall between 100Hz-4000kHz.

2. Measurement Conventions and Reliability

Looking at the spectral display it becomes apparent that vowels embedded in speech are not steady state. In fact, vowels form a continuum of spectral patterns which transition from the locus of the preceding consonant through the nucleus of the vowel target to the locus of the following consonant. In situations where there was an observable steady-state portion of the nucleus, the midpoint of this region was selected for formant measurement. However, in such cases where there was no observable steady-state portion, the midpoint of the entire duration was selected as the vowel target. Both of these conventions serve to minimize coarticulation distortions introduced from

neighboring segments. Thus, it would be possible for another researcher to achieve the same values for the data. Wherever possible, a steady-state portion of the vowel nucleus was selected, however vowels vary greatly with context. The CSRE software simplifies the tedious process of measuring the exact midpoints of formants by producing automatic readout of the formant value for every position of a vertical line cursor.

Some of the hypotheses generated in the literature review (chapter 3) called for an analysis of English diphthongs (especially [eI] and [oU]). Because the L1 forms closest to the English diphthongs were static vowels it was necessary to reduce the diphthong to a single measure rather than a series of measures along the duration. Accordingly, the midpoint of the formant transition was taken as the representative measure. This same technique was used to obtain formant values for the native speaker English data (Holbrook and Fairbanks, 1962). Obviously, caution must be exercised in the interpretation of these formant figures since they do not correspond to the theorized perceptual models of the diphthong which make transition data necessary.

Even with the smoothed output generated by the LPC, at times the formants were split either due to nasalization or radiation from glottal source or the lips. When this was the case, the speech segment was re-analyzed using a different number of coefficients. If this failed to produce smoothed unitary formants, the values of the split portions were measured and an intermediary position was interpolated.

D. DATA VERIFICATION

As the formant values were being recorded it was immediately apparent when a measured value was extreme for the given vowel spoken. Such files were played again to

see if the vowel sounded extreme in the direction that the measured value suggested. At times, it was apparent that the subject had made an oversight or other type of reading error. These were immediately noticed through play back and analysis. Additionally, once the formant values were entered into the MYSTAT statistical software program, extreme values could be easily spotted either by glancing down the columns of the spreadsheet or histogram plots. All such extreme values were checked in order to eliminate investigator-induced error and other non-meaningful deviations such as reading errors.

PART III:

THE STUDY OF SECOND LANGUAGE SPEECH

7

The Vowels of English and Japanese

In chapter 6 I described the specific procedures and methodological choices made for this study of IL vowels. Before turning to an analysis of what is "variability" in second language speech, it is crucial that we establish the basis for first language norms and second language targets. I will attempt to do so using both traditional descriptions of vowels provided in the literature and results from acoustic measurements made in this study. Insights from these comparisons will provide a basis of analysis in chapter 8. Because an analysis of Japanese vowels was made from actual empirical data collected in the study, I chose to place this chapter after chapter 6 so that the reader may reference the methodological chapter for this portion of the study as well.

A. LINGUISTIC DESCRIPTIONS

Before turning to a discussion of acoustic phonetics, it is important to understand what the literature tells us about English and Japanese vowels. Following a discussion of the each language's respective vowel inventories I will discuss the notion of articulatory setting and how it has been used to describe English and Japanese speech.

1. English Vowels

For many varieties of English, the front vowels [i:, I, E, @] are virtually equidistant from each other. However, as Ladefoged points out (1993), in the Midlands and the North of England they make a lower vowel in such words as "had" so that it sounds more like the [A] in father. Also, some Eastern American speakers distinctly make a diphthong in "heed" starting from the [I] position and raising up to the [i:] position. In such exceptions, the front vowels may not form a series of equally spaced steps. However, acoustic measurements (Peterson & Barney, 1952) show that the front vowels are indeed evenly spaced for much of American English..

The back vowels, however, do not show such consistent even spacing or linearity. The low back vowels [A] and [>] are further back than the high back vowels [U] and [u]. Also, the back vowels seemed to be paired off (by proximity) into the high versus low pairs. Thus, the low pairs seem to be closer to each other than to either of the high vowels, while the high pairs seem to be closer to each other than either of the low vowels. In addition, there is a certain similarity in quality that is shared between the [A] and [>] vowels while a different quality is shared by the [U] and [u] vowels.

English vowels can be divided into two distinct classes. The **tense vowels** are distributed across both open (vowel-final) and closed (consonant-final) syllables structures. The **lax vowels** are found only in closed syllables. Distribution is not the only difference between these two classes of vowels. Lindau (1978) has shown from acoustic evidence that the tense vowels are actually positioned peripherally while the lax vowels lie more centrally. It is also interesting to note that the more widely distributed tense vowel class is also that group of vowels found more commonly across languages

Table 7.1 Tense/Lax Distinction in English Vowels

TENSE VOWELS	LAX VOWELS	OPEN SYLLABLES	CLOSED SYLLABLES
i:		bee	beat
	I		bit
eI		bay	bate
	E		bet
	@		bat
A		pa	hot
>		saw	bought
	U		good
u		do	dude
	^		hut
	3r		hurt

2. Japanese Vowels

Japanese has five short vowels and five long vowel phonemes ([a], [i], [4], [e], [o]). The five short vowels correspond roughly to the cardinal vowels. The short and long vowels differ primarily in length. No appreciable difference in quality has been observed among short-long pairs (Vance, 1987) with the exception of [e:] and [e]. Many Japanese words are comprised of Chinese morphemes brought over as early as 6th century C.E.. Of these, a large number of morphemes were written with [eI] sequences in order to preserve lexical distinctions. While many have assimilated to the Japanese long phoneme [e:], some still are pronounced as [eI] such as *keiki*.

The sequences [aI], as in *kaimono* ('food') and [aU] in *aushingo* ('blue' light) are phonemic diphthongs. The vowel sequences [e + i] and [o + u] when bridging morpheme boundaries are pronounced as diphthongs. For example [*ke* + *iro*] (*ke* meaning 'hair' and *iro* meaning 'color') is pronounced [k eI r o]. (Maeda, 1971:172)

Positional allophones are quite rare in Japanese. Of course, the effects of coarticulation change the shape of all vowels to some degree in continuous speech. Often the allophone [&] appears in natural speech. Often [4] becomes fronted before [s] and [z] (Sakuma, 1973:35). Hieronymus transcribes this phone as a voiceless central vowel [&_0] implying that it is shorter than its canonical phoneme. There are a number of morphophonemic alternations described in Martin (1987). One example occurs in a number of root words ending with [e]. When these roots are combined with other roots to form compounds, the [e] on the first root often becomes /a/:

Table 7.2 Morphophonemic Alternation in Japanese

take 'bamboo'	takamura 'bamboo grove'
kane 'metal'	kanagu 'metal fittings'
mune 'ridge'	munagi 'ridge-pole'

The high front vowel [i] is described by Sakuma (1973:32) and Kawakami (1977:21) as equivalent to the cardinal [i]. The mid front vowel [e] is described as between the cardinal [e] and [E] and slightly more central. Kawakami describes this Japanese vowel as similar to the American English vowel in *set*. The Japanese low vowel [a] lays between the low front [a] in American English as in *father* and the low back British vowel [A] spoken in the same word, *father*.

The mid back vowel [o] lays between the cardinal vowels [>] and [o] and slightly anterior to both. One salient feature of the Japanese vowel is the lack of noticeable lip rounding. The Japanese high back vowel [4], however, does shows a form of lip activity which Ladefoged (1975) describes as **lip compression** (protrusion of the corners of the mouth while tensing the lips vertically) differentiating it from common lip rounding which is seen in the American English counterpart [u].

3. Articulatory Setting

Integral to a description of vowels should be a discussion of the "articulatory setting". Laver (1978) describes two kinds of voice quality features: 1) those which are intrinsic to the speaker and identify the person talking and 2) those features which are extrinsic, "long-term muscular adjustments of the intrinsic vocal apparatus which are once acquired by social imitation or individual idiosyncrasy and have become habitual." Among the extrinsic features of vocal quality lies the concept of "articulatory setting" named by Honikman (1964) to refer to those learned muscular adjustments which a group of speakers from a particular language share in common.

Parameters contributing to an articulatory setting include activity by the lips, jaw and tongue and velum. The effect of articulatory setting is seen on individual segments but not all segments in a language are equally effected. Instead, Honikman points out, the setting is derived from the contribution of those segments which most frequently occur in the language. For example, if the most frequent consonants in language X tended to have a secondary articulation of lip rounding then lip rounding would be considered one of the parameters of the articulatory setting for language X. More than a tally of features from the phonetic inventory, the "setting" is a sort of gestalt impression observed by the phonetician derived from a composite of statistical weightings within actual spoken language (Laver 1978).

The setting of English can be described as having moderate lip activity, little jaw movement (except in the low vowels like [a]), and "roof-tethered" [my quotes] tongue activity (Honikman 1964:76-77). Japanese, however, has even less lip activity, considerably more jaw activity and "root-anchored" tongue activity. Again, these

descriptions will not be true of every segment but should be true of the complex of spoken language.

B. ACOUSTIC MEASUREMENTS

Figure 7.1 illustrates the measurements of English and Japanese native speaker vowels. As described in chapter 6, English diphthongs values were derived from the midpoints of a series of measurements throughout the duration of the dipthong. The overall spread of the Japanese vowels was proportionate to the descriptions in the literature. The only major difference between the measured vowels and the described vowels was that the measurements showed [u] lower than described by Vance (1987). Also, in the literature, [o] is described as substantially lower than [e]. Measurements from this study show an [o] that is about the same height as [e].

According to Figure 7.1, Japanese vowel spread seems much more shallow than the English, especially with regard to the central and back vowels. One noteworthy example of this tendency is the low central [a]. From the data, this vowel seems only slightly lower than the Japanese [o] in the acoustic space. The gathered data show a Japanese [a] which is quite anterior, well into the mid-section of the vowel space, whereas in English the analogous [a] appears low and to the back of the vowel space. The Japanese [u] also appears quite anterior to the English analogue in the acoustic space due to the unrounded lip position of the Japanese [u]. While the tongue positions may be similar in both languages, the effect of lip rounding on the English [u] causes the formant frequencies to lower and consequentially the English analogue appears further back.

Earlier I cited work in the area of articulatory setting done by Honikman and Laver. These suggest that Japanese has less lip activity and considerably more jaw activity than English. English speech can be characterized by relatively little jaw activity, slightly more lip activity and "roof-tethered" motions of the tongue. I was curious to see whether these descriptions were supported by the acoustic measurements of the English and Japanese vowels. Looking at the VSD (Figure 7.1) the descriptions seem plausible. By comparison, the English setting has moderately more lip activity than Japanese. This corresponds to the

Table 7.3 Formant Values for English and Japanese (Hz)

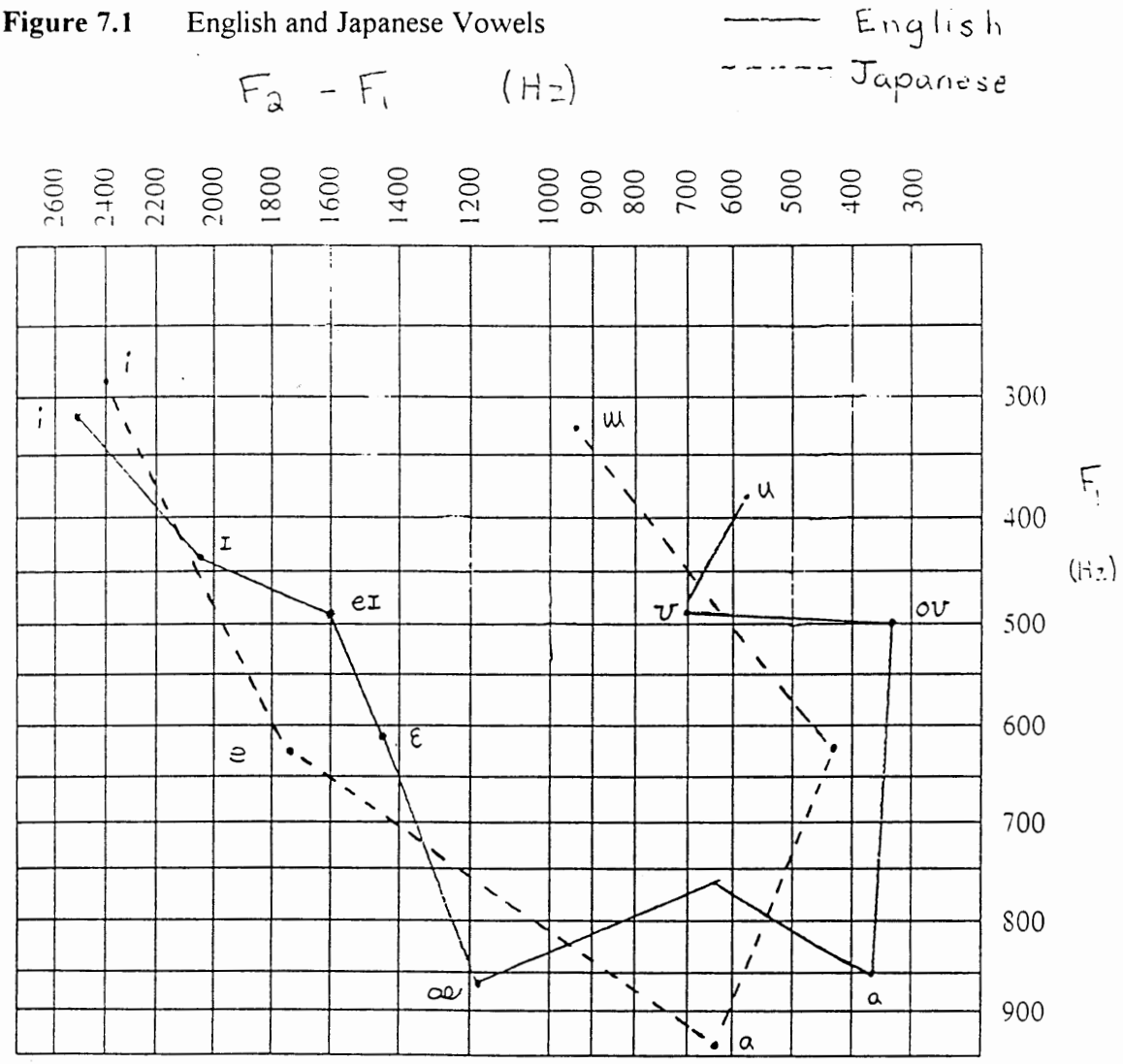
Vowel	English*		Japanese	
	F1	**F2**	**F1**	**F2**
i:	310	2790	292	2695
I	430	2480		
eI*	469	2102	620	2332
E	610	2330		
@	860	2050		
A	850	1220	933	1558
oU*	504	803	610	1046
U	470	1160		
u	370	950	333	1277
^	760	1400		
3r	500	1640		

*English data from Peterson and Barney, 1952

slightly lower F1 and F2 values for the English [u] than the Japanese [u]. In terms of jaw position, the back vowels [a] and [oU] especially demonstrate a difference between the two settings. The Japanese vowels register lower on the vowel space which corresponds

to a more open jaw position. The VSD also corroborates these claims with regard to tongue position. The "roof-tethered" tendency of English is a result of the tongue becoming bunched up in the back of the mouth. Again, the notion of articulatory setting is something which has been theoretically defined as realizable on the aggregate. Consequently, the observations made from the VSD, while consistent with Honikman's predictions, could not be said to provide evidence for the theory.

Figure 7.1 English and Japanese Vowels



8
Tests and Results

This chapter shall be devoted to the quantitative and qualitative description of the data gained from 1) native speaker Japanese, 2) Japanese learners of English, and 3) the mixed group of ESL learners. The purpose of the study is to investigate second language variability noting particularly the influence of the first language through the locations of L1 vowels and the size of the L1 vowel inventory. Additionally, hypotheses from Flege's SLM are tested. Part A contains tests for vowel stability across utterances for each of the speech groups. Part B contains vowel measurements and VSD plots for each group. Part C examines the inter-speaker variability characteristics of each of the studied speech groups. Finally, Part D is devoted to the issues raised by Flege in the Speech Learning Model (see Chapter 3 for further information) and specifically tests Flege's Equivalence Classification Hypothesis.

A. INTRA-SPEAKER VOWEL STABILITY

Individuals vary from utterance to utterance in the way that they produce their vowels. Usually these differences are minimal and fit within a range characteristic of the individual speaker, Nonetheless, this kind of "intra-speaker variability" is a physically real characteristic of all human speech and must therefore be accounted for in any kind of spectral study.

1. English Vowels

Peterson and Barney (1952) in their seminal study of American English vowels used statistical methods to capture speaker-internal variation. At first the method of recording two samples of each of the test words for each subject was done as a method of verifying the accuracy of their data by screening investigator-induced errors. By plotting the formant value of the first token of the test word against the values for the second they were able to quickly detect large discrepancies between utterance values. Figures 8.1 and 8.2 give examples of this. Points that lay directly on the 45-degree line would indicate that the vowels were identical in formant value. If the formant for the first utterance was greater than that of the second utterance, the point would be located above the line. If the formant was less than that of the second utterance the point would be located below the line. By examining the gross differences (+/- 3 Standard Deviations) between utterances Peterson and Barney (1952) could check for erroneous measurements or typographical errors. After these outliers were omitted or corrected the remaining variability represented the true intra-speaker variability.

Once the investigator-induced errors were corrected, Peterson and Barney (1952) found that intra-speaker variability was noticeable but not statistically significant. Unfortunately, Peterson and Barney (1952) did not present the exact correlation scores for this data. However, the discussion of these"Accuracy-Precision Charts" (such as Figure 7 in the paper) indicated that the variability was linear for each of the formants.

Unfortunately, Peterson and Barney (1952) did not measure token-to-token variability for each vowel but rather lumped all of the data into one measure. An investigation of the range of articulations for specific vowels would be valuable to vowel space theories and our understanding of vowel perception. It would be helpful to have

this kind of data for each vowel in order to test some of the explanatory statements made in their investigation of listener judgment. For example, Peterson and Barney (1952) predict that [i] and [u] would be the most stable vowels because of their terminal positions in the mouth and yet they present no substantiating evidence for this. "In the formation of [i] the tongue is humped higher and farther forward than for any other vowel; in [u] the tongue hump takes the highest posterior position in the mouth and the lips are more rounded than for any other vowel. The vowels [u] and [i] are thus much more difficult to displace...". This question of mouth positioning can be tested with the data from the IL groups.

Additionally Peterson and Barney (1952) claim "The vowels [u] and [i] are thus much more difficult to displace, and a greater stability in the organic formation of these sounds would probably be expected which in turn would mean that these sounds are recognized by a listener" (p 120).This statement was based not on articulatory data but on listener judgment frequencies. Today, most researchers would agree that such a statement is untenable, yet we must remember that these statements were written in 1952 before the debates between perception and production began. Since four decades of research have confirmed the disparities between perception and production it is now necessary to investigate whether vowel differences in intra-speaker variability truly can be attributed to motor processes or acoustic processes.

2. Guided Variability Hypothesis (GVH)

It is interesting to me that the majority of work in speech production has focused on the relationship between neuro-muscular patterns and acoustic product while neglecting a detailed study of the acoustic variation that a given speaker exhibits from utterance to utterance. It seems that much information about target approximation could

be gleaned from studies of developing phoneme systems such as in child language acquisition and second language acquisition. Often intra-speaker variability is ignored ("controlled for" or"normalized") in the pursuit of other sources of information. It maybe that this level of variability is meaningful to productive processes. It is my opinion that researchers neglect intra-speaker variability because it has less magnitude than the "linguistic" sources of inter-speaker variability. As a result, the misconception exists that intra-speaker variability consists primarily of motor program "accidents" and clumsy tongue motions.

It is my contention that intra-speaker variability among developing phonological systems reveals systematic approximation across predictable parameters. It seems reasonable to think that the direction of utterance-to-utterance variability could be predicted from the position of the target on the vowel space. Since the goal of vowel production is to achieve the widest possible perceptual separation (Ladefoged, 1993), one would expect learners (and native speakers for that matter) to vary their productions across dimensions which are perceptually meaningful. Some vowels such as [i:], [I], [u], and [U] depend more heavily on height accuracy than on frontedness accuracy to maintain perceptual separation from contiguous vowels. (Chapter 6 in Fromkin, 1985) Because these vowels differ from their closest neighbors mainly in height, it is predictable that token-to-token comparisons would yield greater variability (less stability) across the acoustic parameters for frontedness parameter (approximately F2) than across the height (chiefly F1 variation).

The converse is not necessarily true. As mentioned in Chapter 2, very few languages contain vowels which differ only in frontedness (Ladefoged, 1993). English distinguishes between [@]and [A] primarily in frontedness. However each of these must maintain sufficient distance along the height axis in order to prevent confusion with

neighboring vowels. The alternative hypothesis, thus states that the Type II vowels (see table) may vary along either axes depending on phonological context. The GVH does not predict which vowels will have greatest utterance-to-utterance variation, only which parameters will be most stable for particular vowels. The specific classification of vowels into Type I and Type II, of course would vary from language to language as inventory shapes and relationships differ.

Table 8.1 Predictions of the GVH (Inter-speaker variation)

	Type I: Variation primarily on the height axis (F1)	Type II: Variation across both axes. (F1 and F2)--context dependent
	i:, I, u, U	eI, E, oU, A, @, ^, 3r

3. Japanese-English

If native speakers of a language vary from utterance to utterance in their speech certainly second language learners who have not yet fossilized would be expected to demonstrate a greater degree of variability. The group of Japanese beginner students of English provided such a sample of formative IL speech.

The data showed substantial variation between phones and between formants. F1 values had substantially lower token-to-token agreement than F2 values (Compare the average R for F1=.566 while the average R for F2 is .755). If F1 is taken to be the primary indicator of vowel height, then we can deduce that for this group, learners have

less stability in producing the appropriate vowel heights than in producing the appropriate front-back positions.

The Guided Variance Hypothesis was supported by the scores of [i:], [I], and [U]--all of which had a greater F1 variability than F2 variability. The exception was [u]. Perhaps learners were trying to compensate for the difference between [u] and [U] by modifying frontedness (or lip rounding/unrounding which would achieve similar goals) rather than height. That this is taking place may be supported by the evidence that [u] was the only vowel besides [oU] to have a lower F2 correlation than F1.

Peterson and Barney (1952) predictions about the terminal positions being "more difficult to displace" is not supported by the Japanese-English data. In fact, [i] ranks 8th in F1 stability and 10th in F2 stability while [u] ranks 7th and 11th respectively. The

Table 8.2 Ranked Stability of Japanese-English Vowels (Pearson R)

	Ranked by F1			**Ranked by F2**		
Rank		F1	F2	F1	F2	
1	oU	.953	.774	U	.753	.979
2	E	.938	.949	E	.938	.949
3	^	.897	.915	^	.897	.915
4	U	.753	.979	eI	.476	.915
5	eI	.476	.915	I	.342	.848
6	A	.427	.843	A	.427	.843
7	u	.371	.183	@	.348	.786
8	i:	.367	.413	oU	.953	.774
9	3r	.349	.746	3r	.349	.746
10	@	.348	.786	i:	.367	.413
11	I	.342	.848	u	.371	.183
Mean F1 correlation: .566						
Mean F2 correlation: .755 (excluding outliers [i] and [u])						

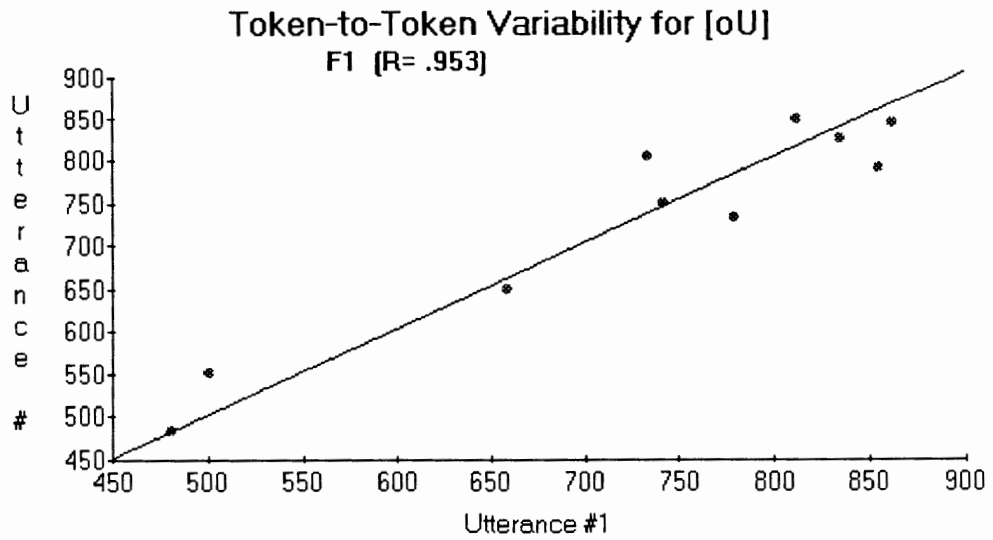


Figure 8.1 Japanese speakers of English.

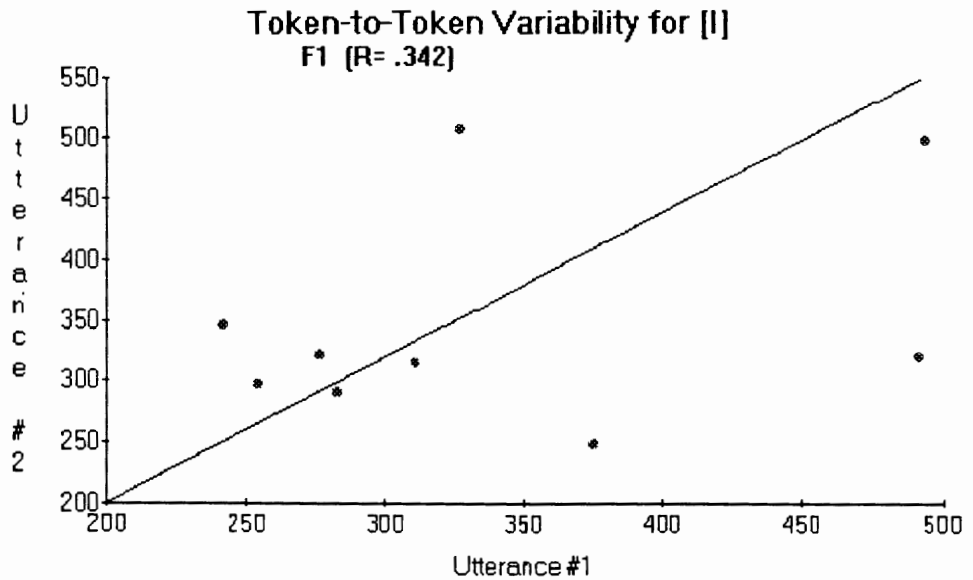


Figure 8.2 Japanese speakers of English.

evidence does not support the Peterson and Barney (1952) claim for the Japanese ESL students. The low correlation score for F1 values of [i:] and [u] does, however support my GVH prediction that height would be the axis of approximation for these vowels. Thus, for the majority of vowels, speakers varied significantly between utterances suggesting that, for this group, intra-speaker variability may exceed inter-speaker variability (see Part D).

4. Mixed group

The same fundamental question was asked of the Mixed data: 1) Does intra-speaker variation reveal parametric tendencies (Guided Variance) toward maximal perceptual separation as provided in the features "height" and "frontedness"?.

Table 8.3 Ranked Stability of Mixed Group Vowels (Pearson R)

	Ranked by F1			Ranked by F2		
Rank		**F1**	**F2**		**F1**	**F2**
1	^	.956	.817	u	.494*	1.00*
2	@	.955	.932	eI	.411	.976
3	A	.941	.967	A	.941	.967
4	oU	.917	.797	i:	.731	.936
5	3r	.859	.840	@	.955	.932
6	i:	.731	.936	I	.389	.916
7	U	.699	.644	3r	.859	.840
8	E	.641	.793	^	.956	.817
9	u	.494*	1.00*	oU	.917	.797
10	eI	.411	.976	E	.641	.793
11	I	.389	.916	U	.699	.644

* The scores for [u] should be taken with caution since only two cases were usable.

Mean F1 correlation: .749 (excludes [u] values)
Mean F2 correlation: .816 (excludes [u] values)

Again, F2 values show more stability across utterances than F1 values, although the difference is much smaller for the mixed group than for the Japanese students. Still, only [oU] and [^] had substantially lower F2 than F1 values. The mixed group also failed to substantiate the Peterson and Barney (1952) claim that [i] and [u] would be formed more consistently although the data for [u] here cannot be taken with much confidence since they consist of only two usable cases. As with the Japanese-English data, there is no observable influence of tongue position on stability. The GVH, which is concerned with perceptual targets, is again supported by the vowels [i:] and [I] that have greater F1 variability than F2 variability. [U] shows no substantial difference between the formants.

Table 8.4 Comparison of F1/F2 Pearson Scores for Each Vowel

Japanese-English			**Mixed ESL English**		
	F1	**F2**		**F1**	**F2**
i:	.367	.413	i:	.731	< .936
I	.342	< .838	I	.389	< .916
eI	.476	< .915	eI	.411	< .976
E	.938	.949	E	.641	< .793
@	.348	< .786	@	.955	.932
A	.427	< .843	A	.941	.967
oU	.953	> .774	oU	.917	> .797
U	.753	< .979	U	.699	.644
u	.371	> .183	u	.494*	1.00*
^	.897	.915	^	.956	> .817
3r	.349	< .746	3r	.859	.840
Mean:	.566	< .755		.749	.816

> F1 is substantially greater than F2 (1)
< F1 is substantially less than F2 (.1)

The data shows that intra-speaker variability is not sporadic but guided by targets and the goal of perceptual separation. Some of the vowels which, when averaged, were very close to the target ([^], [E]) were also the most stable suggesting that the learners were aware when they had accurately formed the vowel. However, [i] and [I] being on the average quite close to the targets were not very stable. The higher accuracy of F2 in vowel pairs such as [i:] / [I] and [eI] / [E] supports the notion that tense/lax distinctions are perceived with information about frontedness more than height.

5. Discussion

The assumption that a less experienced ESL group (i.e., the Japanese group) would have greater utterance-to-utterance variability was confirmed by the data. What is interesting is that some IL vowels seem to be produced more consistently than others. [^] was ranked among the top three most consistent vowels for both groups (with the exception of F2 for Mixed).

Peterson and Barney's (1952) prediction that the endpoint vowels will be articulated more consistently was not supported by the data which showed [i] and [u] as having some of the greatest utterance-to-utterance disparities across both groups. Acoustic parameters of height and frontedness also could not predict intra-speaker variability for either group. The "corner" vowels of [A] and [@] ranged from quite consistent in the Mixed group to quite variable in the Japanese group.

Within the vowel, F1 tends to vary more than F2, but more so among the Japanese. The reason for this difference in formant values is not yet known but I would speculate that the cause has more to do with structural patterns of L1 than developmental issues. The "Guided Variance" hypothesis accurately predicted that those vowels whose targets differ from contiguous vowels primarily in height would vary from utterance-to-

utterance more in F1 than in F2, was supported by the data in Table 8.4. The high back IL vowels of the Mixed group ([u] and [U]), however failed to support the hypothesis.

B. SPEAKER-AVERAGED VOWEL POSITIONS

1. Japanese-English

In order to discuss linguistic issues it is important to control for the intra-speaker variability mentioned in Part A. To accomplish this, formant values for both vowel utterances of a given speaker were averaged (normalized) as in Peterson and Barney (1952). Chapter 7 contains a discussion of the English data and the Japanese vowels collected in this study, as well as comparisons of the empirical data with comparable descriptions in the phonetics literature. Table 8.5 displays the speaker-averaged formant values for English and Japanese-English. Data are given in both linear and logarithmic scales (Hz, Barks) in order to facilitate comparisons with the various charts in this work.

Earlier I showed how it is possible to normalize for intra-speaker variability. In this section I discuss the mean formant values for each group treating them as if they were all homorganic. This computed figure will be referred to as the "prototype". In this way it is possible to normalize the inter-speaker variability. This, of course is an idealization for the purpose of inquiry and overlooks the demonstrable amounts of inter-speaker variation discussed in Part C.

The front vowels are not compact as would be expected of IL vowels. Instead, they are quite forward and seem to maximize the horizontal axis in their spread. The Japanese (IL) approximation of [i:] is very close to the target form. In fact, this IL vowel is the closest approximation of any of the IL English vowels spoken by the Japanese. The

[I] form is quite far from its target. It seems that the Japanese are substituting an [i:] for the [I], as the IL form most closely approximate the [i:] region.

It is very interesting that the Japanese-English form of English [eI] most closely approximated the English [I]. Perhaps there is some perceptual confusion between the vowels. It could be that learners are listening to the final [I] portion of the dipthong [eI] and trying to replicate that without the dipthongized [eI] base. It is clear, however, that the IL form moves in the opposite direction of Japanese [e] and therefore cannot be due to directional interference.

In approximating the English phone [E] the subjects seemed to again surpass even more convenient targets of similar Japanese vowels. Many of the transfer theories working from inventory approaches to IL description would predict that the learner

Table 8.5 Formant Values for English and Japanese-English two scales (Hz, Barks)

	ENGLISH				**JAPANESE-ENGLISH**			
	F1		**F2**		**F1**		**F2**	
	Hz	**Barks**	**Hz**	**Barks**	**Hz**	**Barks**	**Hz**	**Barks**
i:	310	244	2790	2062	313	246	2648	1958
I	430	332	2480	1835	445	343	2648	1958
E	610	464	2330	1725	506	388	2361	1748
@	860	647	2050	1520	765	578	1711	1271
A	850	640	1220	911	814	614	1359	1013
U	470	361	1160	867	456	351	1477	1100
u	370	288	950	713	543	415	1315	981
^	760	574	1400	1043	864	650	1616	1202
r	500	383	1640	1219	701	530	1534	1141
oU	504	386	803	606	719	544	1180	882
eI	469	361	2102	1558	428	331	2443	180
*Values following were computed. For method and source of computed values, see Chapter 7)								

substitute a Japanese [e] for the English [E] since it is closest. What occurred, in fact was the substitution of a vowel approximately half-way between the English [I] and the Japanese [e]. In fact, this IL substitution for [E] very closely approximates the mid-point region of the English [eI].

The low vowels are both more mid-central and slightly higher than their English targets. In approximating the English [@], the Japanese subjects seemed to be influenced by their L1 [A], which lies near the center of the vowel space. This appears to be a clear instance of transfer. The subjects were quite accurate in approximating the English [A], considering the distance between the L1 base and the target. The IL form, however appears to be slightly closer to the [^] target than the [A] target. Because of the small distance that English allows for this vowel distinction, even the fairly accurate approximation of the Japanese may not be sufficient to achieve identifiable distinction as being an [A] unless tested in controlled phonological contexts where coarticulation and phonotactics could not disambiguate.

The back vowels were substantially more central than their targets. The Japanese IL [oU] lies unusually low and is anterior to (more central than) the target form. In fact, this form lies closer to the target [A] and [@] than it does to the target [oU]. The IL [U] lies toward the center of the vowel space, very close to the placement of [3r]. The IL [u] lies lower and more central than the target [u]. Again, the IL form lies nowhere near the Japanese analogue [u]. It is unlikely, therefore that the variability of this IL form is due to directional interference.

The central vowels were both much lower than their targets. The IL [3r] lies quite low relative to the target and is slightly back. The English target has no analogue in Japanese. Observation of Japanese speaking this phone show that it is often more lateral than retroflexed. The English [^] also has no analogue in Japanese. The IL form of this

vowel is lower than the target and closely approximates the target [A]. This could be considered an IL substitution, the result of transfer.

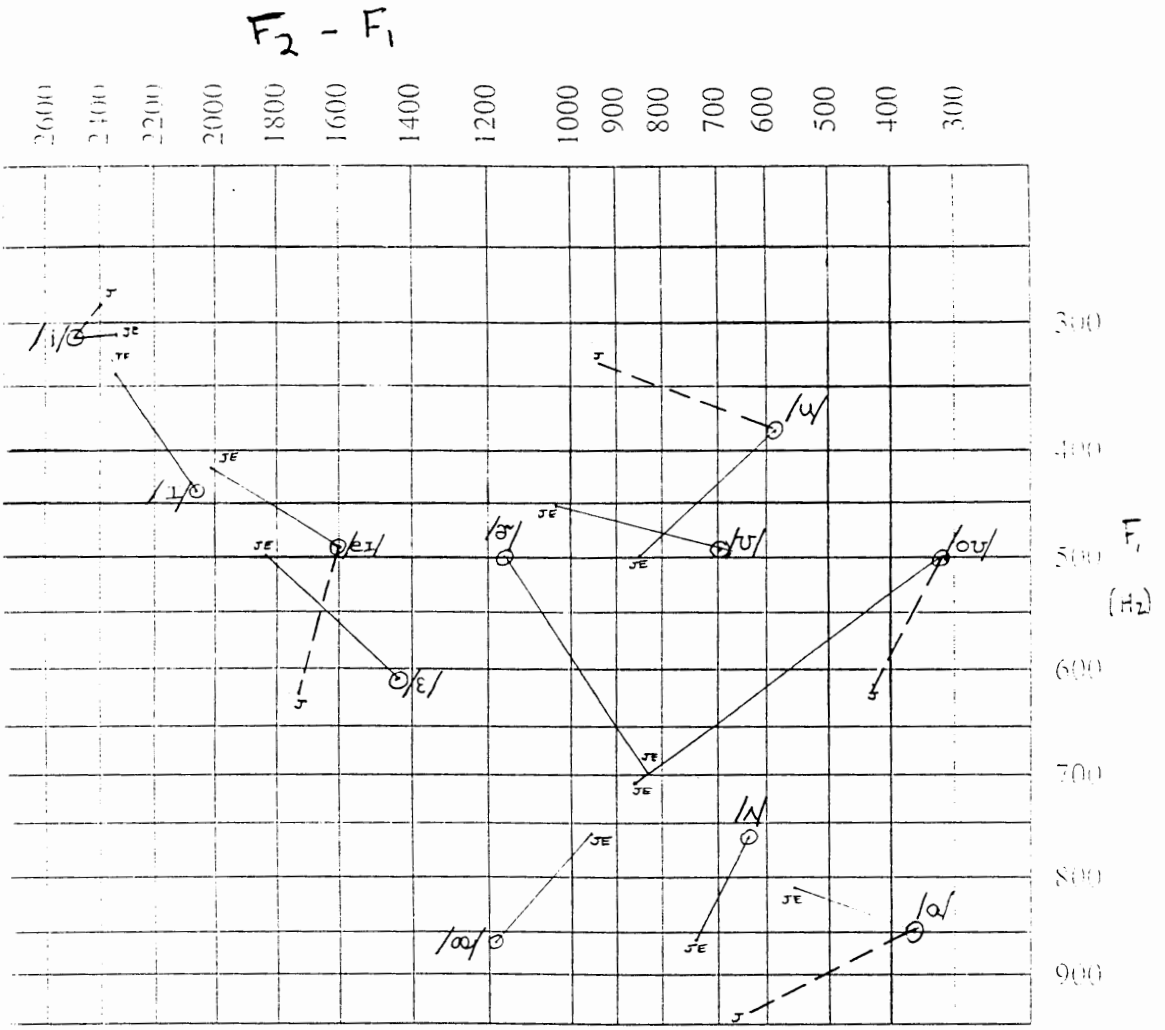
Most of the IL Japanese-English forms were quite centralized. The only exception to this is the [I] which actually lies beyond the English [I]. The Japanese seem to offer a single IL substitute for the [@], [3r], and [oU], none of which are present in Japanese. (The Japanese [o] is static whereas the English [oU] is a dipthong. Also, the Japanese [o] is lower than the midpoint of the English [oU]). The substitution that the subjects give is mid-low central.

2. Comparisons with Mixed Group Data

Not having data on the acoustics of the various L1's represented in the Mixed group, it is impossible to say which trends seem indicative of transfer and which trends can only be explained as developmental. However, comparisons between the Japanese group and the English group can be helpful in illuminating differences related to L1 and experience. The chief difference between the Japanese group and the Mixed group is that, without exception, each of the Mixed group prototypes is more compact (central) than its target. The Japanese, on the other hand, have a series of front vowels which are all more fronted than (peripheral to) their targets.

Some of the Japanese-English vowels give evidence of transfer while others do not. Common to all of these is a tendency toward the center of the vowel space. The only exception to this are the front vowels. Perhaps the reason that the learners form more central vowels is that they would rather be "vague" than "ambiguous". The further one places a vowel towards the periphery of the vowel space the more likely that vowel is to be mis-interpreted. If the learner forms a vowel that is somewhat more central the listener can, with little effort, trace what was intended.

Figure 8.3 VSD's Japanese-English and English



This mutual tendency for the English learners to produce centralized vowels raises questions about the nature of IL vowel development. It seems that the Japanese learners share more commonalities with other ESL students than they share differences.

Table 8.6 Formant Values for English and Mixed Group (Hz)

Vowel	Formant	English	Mixed ESL
i:	F1	310	317
	F2	2790	2391
I	F1	430	363
	F2	2480	2344
eI*	F1	469	421
	F2	2102	2187
E	F1	610	655
	F2	2330	2056
@	F1	860	830
	F2	2050	1788
A	F1	850	867
	F2	1220	13266
oU*	F1	504	576
	F2	803	1158
U	F1	470	408
	F2	1160	1263
u	F1	370	343
	F2	950	1299
^	F1	760	721
	F2	1400	1417
3r	F1	500	528
	F2	1640	1482

*Values following were computed rather than measured

Table 8.7 Mean Error Distance by Articulatory Position (Barks)

Height parameter		**Front-back parameter**	
high (4)	164	front (5)	164
mid (5)	197	central (2)	191
low (2)	161	back (4)	190

3. Effect of Target[1] Location on Error Distance

"Error Distances" of IL forms from their targets were calculated using the Pythagorean theorem ($a^2+b^2=c^2$) explained in section 6. The vowel error distances were ranked in order to see if there were any vowel regions which had more error. No such regional effect was found (Table 8.7). It is interesting that [i] had the smallest error distance. According to the Peterson and Barney (1952) data this vowel had the highest rate of unanimous listening identifications. Apparently, this high front corner of the acoustic space is very distinct for perception.

Student errors seem to be uniformly weighted across the vowel space with no apparent effect of height or backness. The most significant factor, therefore, seems to be whether the vowels are similar or different from L1 vowels, and even that does not make a substantial difference.

4. Effects of L1 Inventory Size on IL Vowel Quality

It is popularly observed that learners' whose L1 contains a larger inventory of vowels will find it easier to approximate the vowels of English since English has a relatively large set (11). In order to test this, subjects from the Mixed ESL group were assigned to one of two groups based on the size of their L1 inventory. The groupings according to subjects' native language are shown in Table 8.9.

Table 8.8 Vowel Inventories of Languages Represented in the "Mixed Group"

Small Vowel Inventories			**Large Vowel Inventories**		
	# Subjects	# Vowels		# Subjects	# Vowels
Spanish	(1)	5	Vietnamese	(2)	11
Indonesian	(2)	5	French	(1)	12
			Taiwanese	(1)	9

A two-way ANOVA test measuring the relative effects of L1 inventory size and the particular target vowel being approximated on F1 and F2 values revealed disparate results. (Table 8.10) While there was no main effect of inventory size on F1, F2 yielded a significant effect ($p<.001$). Apparently vowels and inventory size interacted on F1. As predicted, individual vowels were sufficiently distinct from each other to be significant statistically (main effect has probability of $p<.001$) for both F1 and F2.

Just as Flege's research cast doubt on the effect of L1 inventory size on the utterance-to-utterance *precision* of IL vowels, so this data fails to support the notion that inventory size affects IL vowel *accuracy*. The results of the various factor analyses are not unanimous, however, because F2 shows a significant effect of inventory size.

Table 8.9 ANOVA tests for Effects of Inventory Size and Vowels Targeted (Mixed Group)

DEP VAR: F1 (N=70)

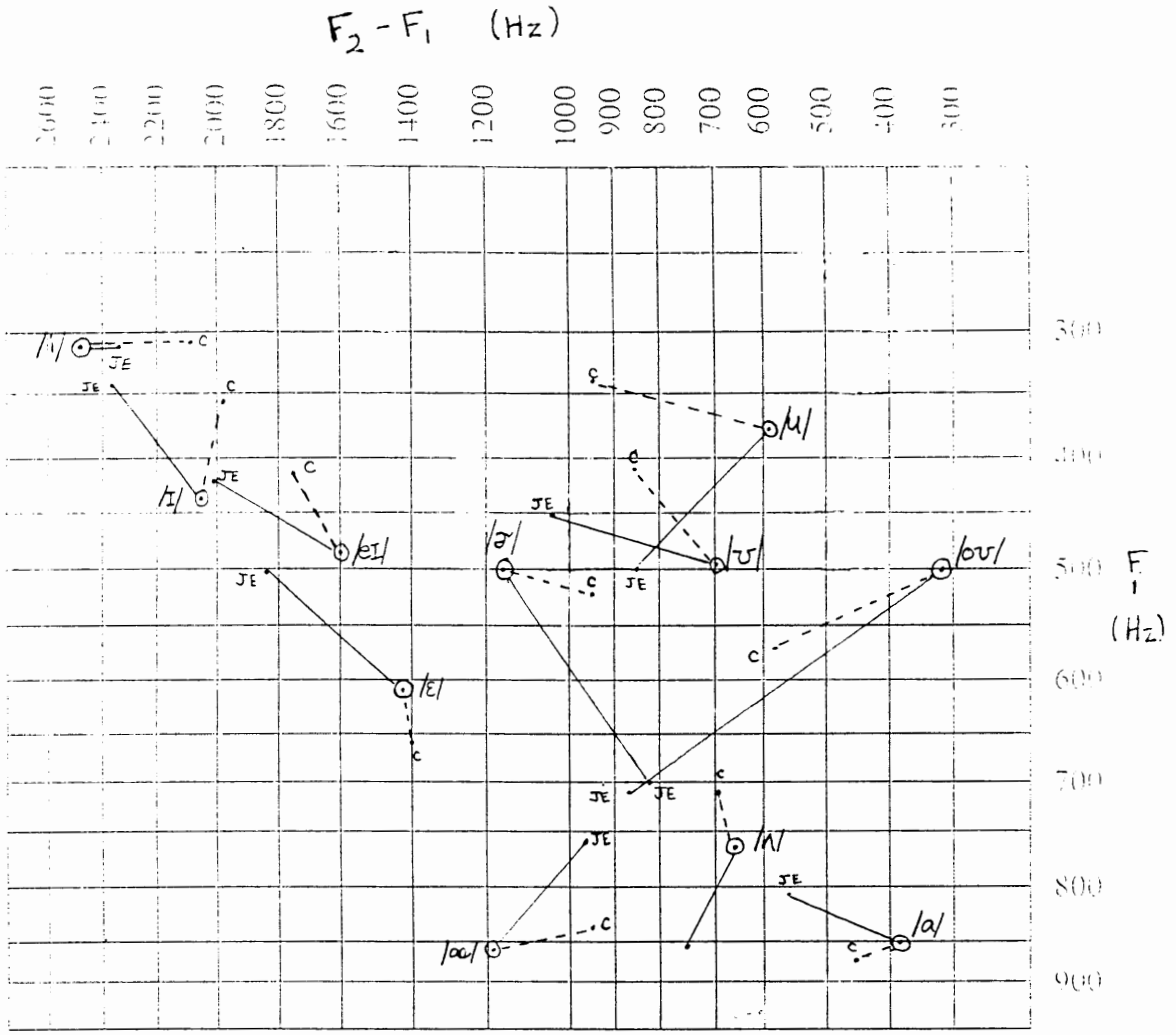
SOURCE	SS	DF	MS	F	P
VOWEL	2193188.350	9	243687.594	13.055	0.000*
INVENT	55453.125	1	55453.125	2.971	0.091
VOWEL x INVENT	229128.607	9	25458.734	1.364	0.230
ERROR	933334.125	50	18666.683		

DEP VAR: F2 (N=70)

SOURCE	SS	DF	MS	F	P
VOWEL	.132840E+08	9	1475996.060	23.879	0.000*
INVENT	1169429.719	1	1169429.719	18.919	0.000*
VOWEL x INVENT	926828.340	9	102980.927	1.666	0.122
ERROR	3090559.583	50	61811.192		

*statistically significant

Figure 8.4 VSD of Three Groups: English, Japanese-English, and Mixed ESL



C. INTER-SPEAKER VARIABILITY: TWO HYPOTHESES

While Part A dealt with those characteristics of an individual's speech which vary from utterance to utterance, this section is concerned with the variability that takes place between speakers. Peterson and Barney (1952) state that while intra-speaker variability is not statistically significant, inter-speaker variability is. With this notion of individual differences established, certain questions arise:1) Do individuals vary more in their second language or in their first? 2) Does L1 constrain the acoustic variability to the extent that learners from a given L1 share a smaller range of variability than ESL learners in general? To address these questions I will present data from the two second language groups described throughout this study: Japanese learners of English and a heterogeneous group of ESL students.

HYPOTHESIS 1:

> *A group of language learners will demonstrate greater variability in the production of their L2 vowels than in the production of their L1 vowels.*

This hypothesis held true for the height but not for the front-back parameter. Inter-speaker variability as with intra-speaker variability was greater for F2 than F1 probably for the reason that F2 values are higher and differences are less significant at the higher ranges. It is counter-intuitive that the Japanese students would exhibit greater degrees of individual differences for their own native language than for English. It is almost as if Japanese learners were funnelling their productions into a smaller range of values (which were not necessarily "correct").

HYPOTHESIS 2:

One's mother tongue constrains the range of productions to the extent that learners from a particular L1 will have less variability than a group of learners from a variety of mother tongues.

Again, it seems intuitive that the Japanese group, which is linguistically and experientially homogeneous would have less speaker-to-speaker variability than the heterogeneous "Mixed ESL group". To test this, the data from both groups were subjected

Table 8.10 Standard Deviations of Japanese Speaking Japanese and Speaking English (Hz)

	F1		F2	
	English	**Japanese**	**English**	**Japanese**
i:	51.3	76.1	240.6	402.8
I	76.3		155.8	
eI	95.0	77.1	150.2	261.6
E	152.8		164.6	
@	192.4		211.3	
A	125.1	98.9	185.6	153.3
oU	128.2	126.7	153.9	117.7
U	120.3		345.8	
u	99.3	84.1	177.0	353.8
^	134.4		167.1	
3r	78.0		133.3	
Mean:	113.9	92.6	189.5	257.8

Table 8.11 Standard Deviations of Both Groups Speaking English (Hz)

	JAPANESE GROUP		MIXED GROUP	
VOWEL	F1	F2	F1	F2
i:	51.3	240.6	51.8	257.7
I	76.3	155.8	46.0	243.3
eI	95.0	150.2	56.0	490.7
E	152.8	164.6	141.4	281.6
@	192.4	211.3	170.0	267.5
A	125.1	185.6	184.9	215.9
oU	128.2	153.9	155.1	364.2
U	120.3	345.8	228.2	182.3
u	99.3	177.0	14.5*	250.3*
^	134.4	167.1	182.3	180.1
3r	78.0	133.3	90.1	145.6
Mean:	113.9	189.5	135.7	226.5

* The scores for [u] should be taken with caution since only two case were usable. These scores were not used to compute the mean.

to a two-way ANOVA test for the factors "group" and "vowels". The results can be found in Table 8.12. There were no significant differences of F1 variability between the groups, however F2 revealed an effect of grouping. Also, F1 revealed an interaction between the particular vowel spoken and the language group while F2 did not. The evidence is inconclusive for the construct "vowel quality". The hypothesis must specify behavior at the sub-phonetic (formant) level. As Table 8.13 shows, the Japanese did indeed produce a more confined range of variability.

Table 8.12 ANOVA tests for Effects of Language Background (Mixed Group)

DEP VAR: F1 (N=172)

SOURCE	SS	DF	MS	F	P
VOWEL	5443761.591	9	604862.399	35.675	0.000*
GROUP	21581.321	1	21581.321	1.273	0.261
VOWEL x GROUP	408862.774	9	45429.197	2.679	0.006*
ERROR	2577154.994	152	16954.967		

DEP VAR: F2 (N=172)

SOURCE	SS	DF	MS	F	P
VOWEL	.397608E+08	9	4417866.360	75.632	0.000*
GROUP	1014074.700	1	1014074.700	17.360	0.000*
VOWEL x GROUP	693115.770	9	77012.863	1.318	0.232
ERROR	8878762.486	152	58412.911		

*statistically significant

PART D: TESTING A CURRENT TRANSFER THEORY: FLEGE'S SPEECH LEARNING MODEL

1. The Equivalence Classification Hypothesis

Flege's Speech Learning Model predicts that learners will have better pronunciation for segments which are new to them (i.e., have no similar corresponding sound in their L1). This is because, Flege argues, learners are expected to transfer native forms where the target sounds are similar to the L1 forms ("equivalence classification").

2. Error Distance

If the SLM is correct, we should expect some of the Japanese-English vowels to be closer to their English targets than others. Specifically, the English vowels [i:], [E], [A], [oU], and [u], having analogous forms in Japanese should be more accented when spoken by the learners than the vowels [I], [eI], [@], [^], [3r], and [U]. (For an explanation of what "similar" and "different" mean and how they apply to Japanese/English comparisons see Chapter 3). Data are given in both linear and logarithmic scales (Table 8.15) in order to facilitate comparisons with the various charts in this work.

3. Variance

Not only is it important to establish the effect of "similar"and "different" vowels on the error distance, it can be helpful to understand behavior across the members of the group. The SLM would predict that learners classify sounds similar to sounds in their L1 as equivalent and therefore have more difficulty producing such sounds accurately.

4. Discussion

The data fail to support Flege's hypothesis. There is no indication that the "similar" vowels were more difficult to produce. On the contrary, the "different" vowels were indeed more distant from the target. Not only do these two subgroups differ in "error distance", the variability of "different" vowels is greater than that of "similar" vowels which casts further doubt on the SLM predictions.

Table 8.13 Japanese IL Phones Ranked by Error Distance (Barks)

Rank	Vowel	Error Distance	Position
1	i:	109	high, front
2	^	113	mid, central
3	I	113	high, front
4	E	125	mid, front
5	A	131	low, back
6	u	190	high, back
7	@	193	low, front
8	oU	197	mid, back
9	U	243	high, back
10	3r	269	mid, central
11	eI	282	mid, front

Table 8.14 Testing Flege's Model: Error Distances

SIMILAR		DIFFERENT	
Distance from Target		Distance from Target	
i:	109Barks 145Hz	I	113Barks 153Hz
E	125Barks 170Hz	^	113Barks 153Hz
A	131Barks 179Hz	@	193Barks 262Hz
u	190Barks 258Hz	U	243Barks 331Hz
oU	197Barks 269Hz	3r	269Barks 367Hz
		eI	282Barks 384Hz
Total:	752.0 Barks, 1022.0Hz		1213.0 Barks 1650.0Hz
Mean:	150.4 Barks, 204.4Hz		202.17 Barks 275.0Hz
SE:	30.4 Barks, 41.8Hz		62.5 Barks 85.7Hz

Table 8.15 Testing Flege's Model: Inter-Speaker Variability

SIMILAR			**DIFFERENT**		
Distance from Target			Distance from Target		
i:	51.3	240.6	I	76.3	155.8
E	152.8	164.6	^	134.4	167.1
A	125.1	185.6	@	192.4	211.3
u	99.3	177.0	U	120.3	345.8
oU	78.0	154.9	3r	78.0	133.3
			eI	95.0	150.2
Mean:	101.3	184.5		116.1	255.1

Apparently transfer is taking place at this point in the learners IL developmental. While the Japanese students had studied English for years, the entire group had an average of 2 months experience in "an English speaking country" which consisted of their present stay in the US. This interpretation is in line with Major's Ontogeny Model (1986) described in Chapter 3. I would expect the Japanese learners to demonstrate an increase in developmental processes by rapidly minimizing the error distances for the "different" column until an eventual "peak" improvement is made, triggering the onset of "fossilization" (see Chapter 3) Meanwhile, the Model predicts, vowels in the "similar" column will continue to gradually minimize their error distance until reaching their potential. Perhaps it is then that Flege's notion of "equivalence classification" becomes explanatory. It seems that the SLM is best suited for types of speech which have already fossilized. It would be interesting to track the students' progressions over time, not only to test Major's hypothesis, but additionally to investigate the path of the learner's development within the acoustic space.

E. SUMMARY

Part A considered the stability of L2 vowels from utterance to utterance. Data supported the hypothesis that the first formant would be most variable for vowels which contrast primarily in height. Thus, it was concluded that intra-speaker variability was not random for second language learners but rather involved knowledge-driven variation along known acoustic parameters. As expected, the Japanese group showed greater token-to-token variability than the Mixed group, most likely due to the differences in experience between the groups.

In Part B, acoustic measurements of the L2 vowels were compared with the native speaker English targets. The measurements of Japanese-English vowels were compared with data from the same individuals speaking their native language, Japanese and also with native speaker English targets. This data enabled comparisons about direction of approximation. It was found that while some vowels showed signs of transfer, many of the approximations were not linear approaches toward the target. Apparently, the "new vowels" arise independently of any targets in the L1. Furthermore, ANOVA tests between sub-groups of the Mixed sample failed to confirm the hypothesis that vowel inventory size affects accuracy of production.

Part C explored two hypotheses pertaining to the degrees of individual variation within the groups under investigation: one concerning the relative proportion of variability between a group speaking their L1 and the same group speaking their L2, the second concerning the influence of L1 on variability. Surprisingly, the Japanese group showed greater F2 variability while speaking Japanese than English. As discussed in Chapter 3, this is likely not related to inventory size of Japanese since Flege (1989) has demonstrated that vowel precision in a given language is independent of the number of vowels in that

language. It does suggest, however, that there are some fairly fixed influences on L2 vowel articulation which seem to constrain L2 approximations more than L1 products. These findings are in support of the notion that second language learning procedes as an interlanguage with characteristics common to all learners.

In Part D, Flege's Speech Learning Model was tested with the Japanese students data. Vowel error distances from the English targets proved greater for the "different" vowels than for the "similar" vowels which contradicts Flege's equivalence classification hypothesis. It may be that Flege's model applies best to experienced learners about to fossilize certain aspects of their speech. The vowel positions were interpreted as transfer and developmental based on data from Part B. Finally, I extended predictions from Major's Ontogeny model concerning the future development of the Japanese and Mixed groups.

FOOTNOTES:

[1]The term "target language" is ambiguous when the language being learned has many valid spoken varieties as has modern English. Due to prevailing sociolinguistic dynamics occuring in the home countries of the research subjects, it is quite defensible to say that the "target speech variety" for these English learners is Standard American English (SAE). It is beyond the scope of this paper to detail the specific social and economic factors leading to this tendency, nonetheless, it can be argued that for each of the English learners being investigated (except for the possible exception of the Taiwanese speakers) SAE is the predominant target speech variety.

9
Discussion

An attempt has been made to implement basic methods of acoustic phonetics in the study of second language vowels. Various levels of analysis yielded evidence for the idea that all language learners hold certain characteristics in common throughout their development of second language proficiency regardless of their native language backgrounds. This idea is most commonly referred to in the literature as the "Interlanguage". This conceptualization is in conflict with notions that second language speech variability is random where not driven by L1 interference.

One of the clearest patterns rising from the data showed that L2 vowels are spread out over a more compact vowel space regardless of the learner's L1. Less experienced learners produced vowels that were more central while more experienced learners produced vowels that were more peripheral in the vowel space approximating the target English forms. There was no evidence for "overshooting" the target as would be possible if the learners were simply transferring their native language vowels to their IL speech.

Further investigations of *intra*-speaker and *inter*-speaker variability were made in order to test hypotheses related to this issue. The idea that one produces more regular speech in her first language than she does in her second language was invalidated by the Japanese data. That is, utterance-to-utterance precision was no greater when the

Japanese spoke their native language than when they spoke English. This finding is conflict with the prevalent notion that IL speech is "random".

In part A, I made the statement that IL intra-speaker variation is not random but knowledge-driven. In order to test this, I hypothesized that intra-speaker variation would be greatest on F1 for those vowels which differed primarily in height. This hypothesis was only weakly supported since, in general, F2 showed greater token-to-token variability than F1. However both groups did exhibit greater F1 variances for the vowels in question ([i], [I], [u], and [U]) than for other vowels.

Target vowels differ from each other in difficulty. An understanding of which vowels in a given TL would be most difficult could be of use to language teachers. Flege (1980) and others have investigated the relative difficulty of certain target language vowels over other target language vowels. Though not uncontroversial, some have hypothesized that there exists a hierarchy of difficulty among the vowels which, it is said, may be based on the vowel's organic formation. According to this theory, vowels made at the terminal positions of the mouth [i:], [A], and [u] would be produced more precisely (and presumably more accurately) than those made mid-way along the height and front-back parameters (Peterson and Barney, 1952). The data from both groups showed the contrary. Neither inter-speaker variability, intra-speaker variability, or error distance showed a heightened degree of accuracy for this articulatorial terminal positions.

Flege proposed an explanation for varying degrees of success in approximating the target vowels in his Speech Learning Model. He posited that learners undergo a process of "Equivalence Classification" for vowels which are similar to vowels in the learner's L1, while those vowels which are substantially different would be perceived as different and would eventually gain greater accuracy than the "similar" vowels. The Japanese vowel

data did not support this prediction, however this may have been a result of the general lack of experience of the group.

If the notion of Interlanguage variability is accurate, it will have to account for individual variation. ANOVA tests of language grouping showed no significant differences for F1 variability, however F2 revealed an effect of grouping. Also, F1 revealed an interaction between the particular vowel spoken and the group while F2 did not. The evidence was inconclusive for the construct "vowel quality". Evidently, the hypothesis should have specified behavior at the sub-phonetic (formant) level.

Many of the comparisons performed revealed that F1 and F2 differences influence vowel production more than any of the other studied factors (i.e., language experience grouping, similar vs. different vowels, influence of vowel inventory sizes). Since height and frontedness can be quantified as F1 and F2 values respectively, the data demonstrates a greater level of accuracy on height parameters than on the front-back dimension. Universal language tendencies also make greater use of the height dimension in the distribution of vowels across the auditory space. Perceptual experiments also show a greater awareness of F1 differences than F2 differences. Apparently, second language vowel production is subject to the same tendencies that first language production exhibits.

NEEDS FOR FURTHER RESEARCH

Many of the issues raised in this study are far from being resolved. Most needed in a study of IL development is longitudinal data such as was presented by Major (1987). I have commented on the placement of vowels within the auditory space. The strength of this data lies in that measurements were taken by individuals speaking both their L1 and

their L2 and a control sample was taken in order to compare results. These, however, are synchronic data making it necessary to infer processes over time through grouping variables.

More robust conclusions, for example, could be drawn from data which chart the course of particular vowels across the vowel space from their origin to the place where they stabilize. It would be particularly interesting to know if such a course exists and if it could be tracked over time. As the data in this study suggest, I would expect such studies to yield a *gravitational* development of vowels rather than a linear development. It is likely that vowels would progress not in a linear path from their L1 forms (or from a place somewhere near the form when learning "new" vowels) to the TL forms, but rather in an elliptical fashion. Comparisons from the experienced ESL group and the less experienced Japanese group demonstrate that, unless native-proficiency pronunciation is attained, such endpoints will likely be located more centrally than the target. Whether or not the position of such endpoints is predictable will have to be determined by further research. If the Japanese ESL data in this study are representative of beginning learners, the origin of such a gravitational path will be independent of L1 forms.

One of the greatest difficulties in describing *first language* vowels stems from the vast amount of differences between speakers. Many attempts have been made to control for vocal tract differences yet no simple model has arisen. Advances in this area would prove very useful for second language research as well as the design of interactive speech training systems. Computational models used in speech processing rely on a vast number of inputs across the feature space which must be interpreted through stochastic methods (i.e., neural networks or Hidden Markov Models). Others have proposed a vocal tract modeling method of simply obtaining values from the speaker's L1 terminal vowels [i], [a], and [u] and inferring the length and area scaling of the speaker based on vowel resonator

models. This, however, makes the assumption that these vowels are produced in the same fashion across languages. This is probably too great an assumption. Even the auditory product demonstrates that these terminal vowels vary substantially from language to language as demonstrated here with measurements of Japanese and English.

One of the goals of this study has been to demonstrate the power of acoustic instrumentation for the analysis of L2 speech. While I have made the claim that the formants gathered and measured in this study represent vowel quality, I have merely touched on one kind of measurement for vowel quality. In reality, a number of different computational methods have been used successfully for digitally modeling speech. Beyond vowel quality, many other acoustic features have been analyzed acoustically including, duration, intensity, voice-onset time (VOT) and pitch.

Findings from acoustic phonetics have become valuable to applied linguists for more purposes than research. Speech science and engineering has seen the development of many interesting applications of acoustic signal processing including speech recognition, speaker recognition, speech synthesis, the development of human-computer interfaces for the handicapped, and speech training for people with communicative disorders. A number of software packages exist for training individuals to perceive and produce difficult sounds. The IBM Speechviewer series provides automated visual feedback on the user's articulations. Other pre-market systems have been developed and presented in major conferences such as ICASP and IEEE.

For the most part, these have been marketed to speech and audiology clinics but could also be of use by the worldwide markets of language learning and teaching. The recent investment boom in telephony industries has caused rapid increases in investments in speech research and developments. Software engineering companies and other science industries throughout the world are incorporating a focus on speech research. Given the

normal cycles of innovation and competition, soon I would expect to see the development of applications for personalized interactive speech training software packaged uniquely for language learning needs and affordable by the consumer.

While such packages are not yet available to language learners and teachers, there are quite a number of analysis software packages with spectrograms, formant displays and editing tools. The Canadian Speech Research Environment (CSRE), used in this study, is available for PC's and includes various analysis algorithms, a formant tracker, tools for speech synthesis, as well as a scripting language for the development of perceptual experiments. The package can be purchased for around $1000 by educational and research institutions. Though more expensive, the Kay Computerized Speech Lab (CSL) provides these functions as well as a library of analysis algorithms built for the detection of pathological speech components such as vocal jitter and hyper-nasalization. Familiarization with the capabilities of such tools can prove quite valuable for the teacher as well as the researcher. Learners taking special courses in speech training and articulatory phonetics could be taught principles of acoustic phonetics such as spectrogram reading which can be taught very quickly.

References

Abercrombie, D., *et al.* (1964) *In Honor of Daniel Jones*. London: Longman.

Abercrombie, David (1985) "Daniel Jones's Teaching". in Victoria Fromkin (ed.) *Phonetic Linguistics: Essays in Honor of Peter Ladefoged.* Academic Press, Inc., Orlando.

Altenberg and Vago, (1983) Theoretical implications of an error analysis of second language phonology production. *Language Learning*. 33:427-447

Berndt, T. (1992) *Child Development*. Fort Worth: Harcourt Brace Jovanovich, p.247.

Best, C.T. & Strange, W. (1991) Cross-language approximate perception. *Journal of Phonetics.* p304-330.

Bohn, O. and Flege, J. (1992) The production of new and similar vowels by adult German learners of English. *Studies in Second Language Acquisition* 14, 131-158.

Borden, G. and Harris, K. (1980) *Physiology, Acoustics, and Perception of Speech.* Williams and Williams, Baltimore.

Broselow, E. (1992) Transfer and universals in second language epenthesis. in Gass, S. and Selinker, L. eds, *Language Transfer in Language Learning.* J. Benjamins, Philadelphia, 71-86.

Cichocki, Wladyslaw, House, and Lister (1993) Cantonese speakers and the acquisition of French consonants. *Language Learning.* 43:1 Mar 43-68.

Dickerson, L. (1975) The learner's IL as a system of variable rules. *TESOL Quarterly* 9:401-407.

Eckman, F. (1977) Markedness and the contrastive analysis hypothesis. *Language Learning.* 27:315-330.

Eckman F. (1987) The reduction of word-final consonant clusters in the Interlanguage. In A. James. and J. Leather eds., *Sound Patterns in Second Language Acquisition.* 143-162, Dordrecht: Foris.

Flege, J. E. (1980) Phonetic Approximation in second language acquisition. *Language Learning* 30:117-134.

Flege, J. E. (1981) The phonological basis of foreign accent: a hypothesis. *TESOL Quarterly* (15) 4:443-55.

Flege, J. E. and Davidian, R. (1984) Transfer and developmental processes in adult foreign language speech production. *Applied Psycholinguistics* 5:323-347.

Flege, J. E. (1987a) A critical period for learning to pronounce foreign languages. *Applied Linguistics* 8:162-167.

Flege, J. E. (1987b) The production of 'new' and 'similar' phones in a foreign language: evidence from the effect of equivalence classification. *Journal of Phonetics* 15:47-65.

Flege, J. E. (1989) "Differences in Inventory Size Affect the Location but not the Precision of Tongue Positioning in Vowel Production" *Language and Speech.* 32:(2) 123- 147.

Fischer-Jorgensen, Eli (1985) "Some Basic Vowel Features" in Fromkin, Victoria ed. *Phonetic Linguistics: Essays in Honor of Peter Ladefoged.* Academic Press, Inc., Orlando.

Gass, S., Madden, C., Preston, D., and L. Selinker eds. (1989) *Variation in Second Language Acquisition.* Vol. 2: Psycholinguistic Issue. Philadelphia: Multilingual Matters.

Gombert, J. (1992) *Metalinguistic Development.* Univ. of Chicago Press, Chicago.

Gray, G. W. and Wise, C.M. (1946) *The Basis of Speech* (Harper Bros., New York) 217-302.

Hecht, B. and Mulford, R. (1982) The acquisition of a second language phonology: interaction of transfer and developmental factors. *Applied Psycholinguistics.*

Hieronymus, J. L. ASCII Phonetic Symbols for the World's Languages: Worldbet AT&T Bell Laboratories, Murray Hill, NJ.

Holbrook, Anthony and Fairbanks, Grant (1962) "Dipthong Formants and their Movements". *Journal of Speech and Hearing Research.* Vol. 5 . 38-58.

Honikman, B. (1964) Articulatory Settings. In Abercrombie, et al. (1964:73-84).

Johansson, F. (1973) *Immigrant Swedish Phonology: A study of multiple contact analysis.* Lund, Sweden: CWK Gleerup.

Joos, M. (1948) *Acoustic Phonetics*. Language Monographs No. 23., Linguistic Society of America. Baltimore, MD: Waverly Press

Kawakami, S. (1977) *Nihongo Onsei Gaisetsu.* Tokyo: Ofusha.

Krashen, S. and Terrel, T. (1983) *The natural approach.* Pergamon, NY.

Kuhl, P. K. (1987) Perception of speech in early infancy, in P. Salapatek and L. Cohen (eds.), *Handbook of infant perception. Vol2. From perception to cognition.* 275-382, Orlando, FL: Academic.

Kuhl, P. K. and Miller, J. D. (1978) Speech perception by the chinchilla: identification. *Journal of the Acoustic Society of America, 63, 905-917.*

Ladefoged, P (1962) The nature of vowel quality. *Revista do Labbatório de fonética experimental da faculdade de letras da universidade de Coimbra.* Also in Ladefoged, Peter (1967) Three areas of experimental phonetics. 50-142 London, Oxford UP.

Ladefoged, Peter (1971) *Preliminaries to Linguistic Phonetics.* Chicago, Univ. Chicago UP.

Ladefoged, Peter (1975) *A course in phonetics*. NY: Harcourt, Brace and Jovanovich.

Ladefoged, Peter (1976) The phonetic specifications of the languages of the world. *UCLA Working Papers in Phonetics*. 31, 3-21.

Ladefoged, Peter and DeClerk, J., Lindau, M., and Papcun, G. (1972). An auditor-motor theory of speech production. *UCLA Working Papers in Phonetics*, 22, 48-75.

Ladefoged, P. and Maddieson, I. (1990) Vowels of the world's languages. *Journal of Phonetics* 18, 93-122.

Ladefoged, P. (1993) *A Course in Phonetics,* Harcourt, Brace, Jovanovich, Orlando.

Lado, R. (1957) *Linguistics across cultures* Univ. of Michigan Press, Ann Arbor.

Larsen-Freeman, D. and Long, M. ed. (1991) *An Introduction to Second Language Acquisition Research*. Longman, NY.

Laver, J. (1978) The concept of articulatory settings: an historical overview. *Historiographia Linguistica*, 5, 1-14.

Leather, J. (1986) F_0 pattern inference in the perceptual acquisition of Chinese tone. In *Sound Patterns in Second Language Acquisition,* eds. James, A. and Leather J., Dordrecht: Foris, 59-80.

James, A. and Leather J. (1986) *Sound Patterns in Second Language Acquisition,* eds. Dordrecht: Foris, 59-80.

James, A. and Leather J. (1991) The acquisition of second language speech. *Studies in Second Language Acquisition* 13,3 p305-341.

Lehiste, I. ed. (1967) *Readings in Acoustic Phonetics.* MIT Press. Cambridge.

Lindau, M. (1978) "Vowel Features" *Language*, 55, 541-563. Also in *UCLA Working Papers in Phonetics*, 38, 49-81.

Maeda, M. (1971) *Kokugo On'inron no Koso.* Tokyo: Keibundo.

Macken, M. and Ferguson, C. (1981) Phonological universals in language acquisition. In *Native-Language and Foreign-Language Acquisition,* ed. H. Winitz. NY: NY Academy of Sciences, 110-129.

Major, R. (1986) The ontogeny model: evidence from L2 acquisition of Spanish /r/. *Language Learning.* 36 (4) 453-504.

Major, R. (1987) Foreign accent: recent research and theory. *IRAL* 25:185-202.

Mattingly, I. and Studdert-Kennedy, M. ed. (1991) *Modularity and the Motor Theory of Speech Perception.* Lawrence Erlbaum Associates, NJ.

Martin, Samuel. (1987) *The Japanese Language Through Time*. Yale University Press, New Haven.

Meyer E. A. (1910) Untersuchungen über Lautbildung. *Die nueren Sprachen*, 18, (Festschrift Viëtor), 966-248.

Mochizuki-Sudo, M. (1991) Production and perception of stress-related durational patterns in Japanese learners of English *Journal of Phonetics* (191) 231-48.

Munro, M. (1993) Productions of English vowels by native speakers of Arabic: acoustic measurements and accentedness ratings. *Language and Speech. 36:1 39-66.*

Nearey, T. M. (1978) *Phonetic feature systems of vowels.* Bloomington Indiana, Indiana University Linguistics Club.

O'Shaughnessy, Douglas O. (1987) *Speech Communication: Human and Machine.* Addison-Wesley, Reading, Mass.

Oller, J. W. and Ziahosseiny, S. M. (1970) The CAH and spelling errors. *Language Learning.* 20:183-9.

Paikeday, Thomas (1985) *The Native Speaker is Dead.* Paikeday Publishing, Inc. NY.

Penfield, W. and Roberts, L. (1959) *Speech and Brain Mechanisms.* Atheneum Press, NY.

Peterson, G. E. and Barney, H. L. (1952) Control methods used in a study of vowels. *Journal of the Acoustic Society of America.* 24:175-184.

Russel, G. O. (1928) *The Vowel.* Columbus: Ohio State UP.

Russel, G. O. (1936) Synchronized X-ray, oscillograph, sound and movie experiments showing the fallacy of the vowel triangle and open-closed theories. In D. Jones and D. B. Fry (Eds.), *Proceedings of the 2nd International Congress of Phonetic Sciences.* 198-205. Cambridge: Cambridge UP.

Sakuma, M. (1973) *Hyojun Nihongo no Hatsuon, Akesento*, expanded ed. Tokyo: Koseisha Koseikaku.

Sakow and McNutt (1991) Perception of /r/ by native speakers of Japanese and Korean: internal and external perception. *IRAL.*

Scovel, T. (1988) *A Time to Speak: a psycholinguistic inquiry into the critical period for human speech.* Newbury House/ Harper and Row, NY.

Stampe, D. (1969) The acquisition of phonetic representation. *Papers from the 5th Regional Meeting of the Chicago Linguistics Society,* eds. R. Binnick et al., 443-454.

Stevens, Kenneth N, and House, Arthur S. (1955) Development of a quantitative description of vowel Articulation. *Journal of the Acoustic Society of America* Vol. 27, No. 3.

Stevens, S. and Volkman, J., and Newman, E. B. (1937) A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustic Society of American* 8, 185-190.

Straight (1980) in Yeni-Komshian, Kavanagh, and Ferguson eds. *Child Phonology Vol. 1: Perception.* New York: Academic Press.

Tarone, E. (1980) Some influences on the syllable structure of interlanguage phonology. *IRAL.* 18:2, 139-152.

Tarone, E. (1989) Accounting for style-shifting in interlanguage. In Gass and Macken.

Vance, Timothy, J. (1987) *An Introduction to Japanese Phonology.* State of NY Press, NY.

Wardaugh, R. (1970) The Contrastive Analysis Hypothesis. *TESOL Quarterly.* 4:123-30.

Walsh, Len (1969) *Read Japanese Today.* Charles E. Tuttle Co., Tokyo.

Weinberger, S (1990) Minimal segments in L2 phonology. In Leather and James (eds) *New Sounds 90: Proceedings of the Amsterdam Symposium on the Acquisition of Second-Language Speech. 137-170,* Amsterdam: University of Amsterdam.

Whitman, R. and Jackson, K. L. (1972) *The unpredictability of Contrastive Analysis. Language Learning.* 22:19-41.

Whitman, R. (1970) Contrastive analysis: problems and procedures. *Language Learning.* 20:191-7.

Wode, H. (1983) *Papers on Language acquisition, Language Learning and Language Teaching.* Heidelberg: Julius Groos Verlag.

Wood, S. (1975) The weakness of the tongue arching model. Working papers, 11, 55-108. Lund: Lund University, Department of General Linguistics.

Wood, S. (1982) X-ray and model studies of vowel articulation. *Working papers,* 23, Lund: Lund University, Department of Linguistics.

Yeni-Komshian, Kavanagh, and Ferguson eds. (1980) *Child Phonology Vol. 1: Perception.* New York: Academic Press.

Zue, Victor (1985) *Acoustic Theory of Speech Production*, MIT, preliminary draft.

Appendix 1: Word Lists used for Elicitation---English

Please read each sentence two times:

Larry said "heat" yesterday

Larry said "hit" yesterday

Larry said "hate" yesterday

Larry said "head" yesterday

Larry said "had" yesterday

Larry said "hot" yesterday

Larry said "hut" yesterday

Larry said "hurt" yesterday

Larry said "hope" yesterday

Larry said "hood" yesterday

Larry said "hoot" yesterday

Japanese Hiragana Syllabary (From Walsh, 1969)

あ A	い I	う U	え E	お O
か KA	き KI	く KU	け KE	こ KO
さ SA	し SHI	す SU	せ SE	そ SO
た TA	ち CHI	つ TSU	て TE	と TO
な NA	に NI	ぬ NU	ね NE	の NO
は HA	ひ HI	ふ HU	へ HE	ほ HO
ま MA	み MI	む MU	め ME	も MO
や YA		ゆ YU		よ YO
ら RA	り RI	る RU	れ RE	ろ RO
わ WA		ん N		を O

No: _____

LINGUISTIC BACKGROUND QUESTIONNAIRE

1. What country are you from originally?_____
2. How many languages do you speak? Please list them._____
3. What language was your first language as a child?_____
4. How long have you spoken English?_____
5. When you began learning, were you:
   a. under 5 years old
   b. 6-9 years old
   c. 10-13 years old
   d. 14-15 years old
   e. older than 16
6. How long have you lived in the US?_____
7. Have you lived in any other English-speaking countries besides the US? If so, what countries?_____
8. Please circle one of the following: ENNR ESL neither
9. How many years of college have you completed?_____

Appendix 2: Conversion Formulae for Logarithmic Frequency Scales

Hertz to Mels (psychological magnitude pitch)

$$m = 2595 \log_{10}(1 + f/700) \tag{2.1}$$

Hertz to Barks (critical band rate)

$$z = 13\arctan\left(0.76\frac{f}{\text{kHz}}\right) + 3.5\arctan\left(\frac{f}{7.5\text{kHz}}\right)^2 \tag{2.2}$$

z represents the bark unit of measure while f, the frequency unit (cps/Hz)