2005

Vowel identification by monolingual and bilingual listeners: Use of spectral change and duration cues

Merete MÃ¸ller Glasbrenner
University of South Florida

Vowel Identification by Monolingual and Bilingual Listeners: Use of Spectral Change and Duration Cues

by

Merete Møller Glasbrenner

A thesis submitted in partial fulfillment
Of the requirements for the degree of
Master of Science
Department of Communication Sciences and Disorders
College of Arts and Sciences
University of South Florida

Major Professor: Catherine L. Rogers, Ph.D.
Stefan A. Frisch, Ph.D.
Jean Krause, Ph.D.
Joseph Constantine, Ph.D.

Date of Approval:
07.26.2005

Keywords: speech perception, synthesized vowels, formant flattening, duration neutralization

To my parents: You always encouraged me to pursue my dreams even if it meant to go abroad.

My husband, Andrew: You initially suggested pursuing this field. You motivated me not to reach for the closest apple but for the star farthest away. Most of all, thank you for all your love and support.

To all of my friends in the Speech-Language Pathology program: Thank you for keeping me real and putting up with me in both times of happiness and sadness. What a ride this has been!

To Patti and Katherine, my two externship supervisors: Thank you for all your support and understanding.

To Dr. Rogers: For being a wonderful mentor and role model. I would not have missed a minute of it.

Acknowledgments

I would like to express my gratitude to a number of people who have been of great help in the course of writing my thesis but also inspired and motivated me.

To Sylvie Wodzinski, Teresa DeMasi, and Michelle Bianchi: Thank you for all of your help in the lab and your friendship. I will truly miss you.

To my committee members, Dr. Jean Krause, Dr. Stefan Frisch, and Dr. Joseph Constantine: I’m honored to have had the opportunity to work with you. I appreciate your opinions and your flexibility. I still think this: Great defense questions! To Dr. Diane Kewley-Port for mentoring me, and reinforcing my interest within speech science.

Most importantly, I would like to thank Dr. Catherine Rogers. You not only inspired me and sparked my interest for speech science three years ago, but I had the honor and privilege of working closely with you and learn from you. Thank you for giving me this great opportunity. I am grateful for your support, confidence, and advice in both good and bad times. Thank you for your patience with me and for dedicating numerous hours of preparing data or proofreading my convoluted “Danish” thoughts. You opened up a door to a world of science that intrigues me so much. I hope the journey will continue.

Table of Contents

List of Tables

List of Figures

Vowel Identification by Monolingual and Bilingual Listeners: Use of Spectral Change and Duration Cues

Merete Møller Glasbrenner

ABSTRACT

Recent studies have shown that even highly-proficient Spanish-English bilinguals, who acquired their second language (L2) in childhood and have little or no foreign accent in English, may require more acoustic information than monolinguals in order to identify English vowels and may have more difficulty than monolinguals in understanding speech in noise or reverberation (Mayo, Florentine, & Buus, 1997). One explanation that may account for this difference is that bilingual listeners use acoustic cues for vowel identification differently from monolinguals (Flege, 1995).

In this study, we investigated this hypothesis by comparing bilingual listeners' use of acoustic cues to vowel identification to that of monolinguals for six American English vowels presented under listening conditions created to manipulate the acoustic cues of vowel formant dynamics and duration. Three listener groups were tested: monolinguals, highly proficient bilinguals, and less proficient bilinguals.

Stimulus creation included recording of six target vowels (/i, ɪ, e$^{\text{ɪ}}$, ɛ, æ, ɑ/) in /bVd/ context, spoken in a carrier phrase by four American monolinguals (two females, two males). Six listening conditions were created: 1) whole word, 2) isolated vowel, 3) resynthesized with no change, 4) resynthesized with neutralized duration, 5) resynthesized with flattened formants, and 6) resynthesized with flattened formats and neutralized duration. The resynthesized stimuli were created using high-fidelity synthesis procedures (Straight; Kawahara, Masuda-Katsuse, & Cheveigné, 1998) and digital manipulation. A six-alternative forced choice listening task was used. The main experiment was composed of 240 isolated vowel trials and 48 whole word trials.

Data from 17 monolinguals, 25 highly proficient bilinguals, and 18 less proficient bilingual listeners indicate a consistent but relatively small decrease in performance for the proficient bilinguals compared to the monolinguals, a substantially greater decrease in performance for the less proficient bilinguals compared to the proficient bilinguals, and a greater decrease in performance due to formant flattening than to duration neutralization for all groups. In support of the hypothesis of differing cue use by bilinguals, the data showed significantly different patterns of performance across vowels and listening conditions for the three listener groups.

Chapter 1

Introduction

Previous research has indicated that monolingual listeners use dynamic spectral information to identify vowels. While static spectral information is often sufficient to identify most vowels, monolingual listeners also benefit from the inherent spectral change that occurs during the course of vowel production (Hillenbrand & Nearey, 1999; Strange, Jenkins, & Johnson, 1983). Data suggest that removal or modification of inherent vowel information affects intelligibility negatively. Nevertheless, although we know that listeners rely on certain cues to identify sounds, their relative importance and acquisition order still remain somewhat unclear.

The field of speech perception encompasses many areas which all contribute to improved understanding of auditory processing. For instance aural rehabilitation treatment of a hard of hearing client builds on a hierarchy going from easy to difficult. In the easy approach, print, topic, visual and auditory cues may all be available to the client. These cues will additively help the listener to identify the spoken stimuli. A difficult therapy approach may exclude print, topic, or visual cues. Thus, the only factor the listener relies on is hearing the stimuli. Depending on the client's hearing and processing level, stimuli may be presented using many or few cues. Thus, speech perception not only relates to what is heard but also how well a person remembers a sequence: One three-word sentence is easier to memorize, word for word, than ten long embedded sentences.

The point with the hierarchies is that we rely upon our senses, experiences, and memory to hear, store, process, and extract meaning of what we hear (Major, 2001; Tye-Murray, 1998).

Normal-hearing listeners use similar strategies to identify sounds. Many redundant cues are used and all work together, however, the way in which bilingual listeners use these same cues is less clear. A valid reason why bilingual speech perception may be of importance is the increase of bilinguals or polyglots in the United States. This fact raises the importance of studies that examine production and perception in speakers who acquire English as their second language (L2).

The significance of the demographic changes in the U.S. is apparent in the U.S. Census data. The release of the U.S. Census 2000 revealed a significant growth in the non-English speaking population in only 10 years. Analysis of the data showed that 47 million persons or 18% of the American population aged 5 and above, were reported to speak a language other than English in their homes. Spanish was ranked the first among these languages, with Spanish speakers composing 10% of the total American population (approximately 28 million). In only ten years the Spanish speaking population increased from 17 to 28 million nationwide (Shin & Bruno, 2003).

Noteworthy for the 2000 Census was a detailed analysis of speakers' English skills. Speakers identified themselves as speaking English "very well," "well," "not well," or "not at all." About 20 million of the Spanish speaking population rated their English skills within the levels of "well" to "very well." However, nearly 8 million rated their abilities as "not well" or "not at all" (Shin & Bruno, 2003).

A noticeable increase of Spanish speakers was also seen within the state of Florida. Here the bilingual population increased from 12 to 15 million for people aged 5 and over from 1990 to 2000. Out of the 15 million, 10% reported that they spoke English less than “very well” (Shin & Bruno, 2003).

Returning to the issue of speech perception, a recent study performed by Febo (2003) found that even early Spanish-English bilingual listeners performed more poorly than monolingual English listeners when presented with monosyllabic English words in different levels of background noise and reverberation. Results showed that even bilingual listeners who acquired their second language before age 6 and were rated as having little or no foreign accent in English identified fewer words correctly than monolinguals when the words were presented in noisy and reverberant conditions, although both groups showed identical (perfect) performance in quiet. This study indicates that even early acquisition of L2 appears to impact speech recognition negatively in certain listening conditions (Febo, 2003).

The Febo study suggests that there is indeed a difference in speech processing abilities between monolingual and bilingual listeners. However, the data do not provide information for why these differences occur. In the present study, we compared bilingual listeners’ use of acoustic cues to American English vowel identification with that of monolinguals for vowels presented in various distorted listening conditions.

The next sections will be dedicated to explaining speech acoustics, speech perception, as well as current research within the field of second language acquisition.

Vowel Acoustics

To understand the perception of vowels, it is essential to first summarize the physics of speech. Vowels are best explained with what is known as the “source-filter model” of speech production. The production of a vowel begins at the level of the larynx at which a periodic or cyclic signal is generated. Thus, the vocal folds are referred to as the periodic source. The sound travels through the vocal tract. The vocal tract behaves like a variable filter that changes in response to tongue and jaw position. The fundamental frequency is typically modulated by increasing or decreasing tension of the vocal folds but it also dependent on the length of the vocal tract (Borden, Harris, & Raphael, 1994; Fry, 1979).

Vowels are formed by shaping the vocal tract in a manner in which certain frequencies are attenuated and others amplified. The amplified points are called formants or resonance points and those are the points that make vowels so salient. Vowels can be characterized by their first, second, and third formants (F1, F2, and F3) (Borden et al., 1994; Fry, 1979). However, when the vowel is isolated, meaning no consonant precedes or follows it, F1 and F2 are often sufficient to identify most vowels (Pickett, 1999; Strange, 1999). Height of jaw, position of tongue constriction (back vs. front), and lip rounding will generate vowel-specific formant patterns (Pickett, 1999; Strange, 1999). F1 is influenced most by jaw and tongue height. If the jaw is raised and thus the tongue also, then F1 tends to be low (Borden et al., 1994; Fry, 1979; Pickett, 1999; Strange, 1999). The second formant (F2) is influenced more by the location of tongue constriction. If the tongue constriction is towards the front, F2 increases (Borden et al., 1994; Fry, 1979). Thus, the vowel /i/, which is articulated with the tongue high and forward, has low F1 and high F2 values, whereas /ɑ/, which is articulated with the tongue low and back, has high F1 and low F2 values.

Two additional acoustic cues that vary across vowels are duration and spectral change. In terms of duration, English American vowels can be divided into tense and lax vowels. Lax vowels typically are shorter in length than tense vowels (e.g., lax /ɪ/ versus tense /i/) (Borden et al., 1994; Fry, 1979; Pickett, 1999; Strange, 1999).

Spectral change or vowel drift (Hillenbrand, Getty, Clark, & Wheeler, 1995; Hillenbrand & Nearey, 1999) becomes crucial when describing the dynamic formant change of diphthongs or even monophthongs. Unlike /i/ for which formant contours remain fairly static, according to Hillenbrand (1999), the formant frequencies of the diphthong /a^I/ move from /a/ formant values to /ɪ/ values. The physical movement of the jaw and lips illustrates the change of this diphthong; the jaw moves upwards as does the tongue while the location of target constriction moves forward in the mouth. This type of change also occurs for many American English vowels that are typically described as monophthongs, although to a lesser degree than for a diphthong such as /a^I/.

Consonantal environments also influence vowel formant patterns causing consonant-vowel (CV) or vowel-consonant (VC) formant transitions to differ in the onset and offset of the vowel depending on the consonant context. An example is the effect of stop consonants on vowel formant transitions. Stops or plosives are sounds formed by stopping air flow from the mouth at some point and suddenly releasing it. Vowels that precede or follow a stop will, depending on the place of articulation of the stop, have different formant transitions. For instance, a bilabial plosive elongates the vocal tract forming lower “burst” frequencies. The formant transition following the burst, specifically of F2 and F3 will have low to high or rising contour. Different for the velar produced plosive /k/, the vocal tract is shortened and the resonance patterns become more complex, producing falling F2 and rising F3 CV transitions (Strange, 1999).

Another effect of the consonant-vowel (CV) context is the elongating effect of word-final voicing: a vowel that is followed by a voiceless consonant has shorter duration than one followed by a voiced consonant (e.g., /bit/ versus /bid/) (Strange, 1999).

Synthesis

Synthesis and speech acoustics are related in the sense that traditional synthesis cannot be performed unless a thorough knowledge of the speech physics is available. To create artificial speech sounds requires an in-depth understanding of source generation, filter, and the transfer function to achieve the desired final output or signal. In speech perception studies the Klatt synthesis or formant synthesis method is mostly used. However, in recent years, use of a new approach is winning popularity. This is the high-fidelity resynthesis of speech created by Kawahara et al. (1998). Both synthesis methods will be discussed in the upcoming paragraphs. An advantage of using synthesis in perceptual studies, regardless of the method, is the ability to control variables. By the same token, the drawback of speech synthesis is that perception may be compromised due to lack of natural-like sound quality.

Formant Speech Synthesis. Formant speech synthesis uses the principles of source, filter and transfer function to compute and generate signals. A number of possible synthetic software programs exist, some more sophisticated than others. One of the more frequently used synthesis programs in perception studies was developed by Dennis Klatt.

Using Fant’s and Lawrence’s hardware principles, Klatt created a phoneme synthesis by rule program: a hybrid of cascade and parallel formant synthesis. Cascade and parallel synthesis were originally different types of hardware, each of which was better adapter to synthesis of different sounds (Klatt, 1987).

Parallel formant synthesis by Lawrence employed the principles of the vocal tract transfer function using anti-formants and formant resonators, employed separate independent filters to fabricate a specific resonant frequency (described in Klatt, 1987). The output signals of these filters were then added together to form one sound. Filters allowed buzzing (voicing), hissing (devoiced), or a combination (devoiced and voiced) of noise realizing obstruent sounds (e.g., /s/ versus /z/). Fant’s cascade approach allowed for output of one filter to feed into the next. The resulting relation of amplitude among the formants is more like real speech for vowels generated in this way. Thus, the cascade formant synthesis was fount to be better-adapted for vowel synthesis while parallel synthesis produced better quality consonants (Klatt, 1987).

Klatt combined these two principles creating a hybrid synthesis system which allowed for advantages of both cascade and parallel techniques to be exploited for synthesis of consonants and vowels (Klatt, 1987). To increase naturalness of speech, Klatt proposed a number of duration rules for different contexts (see Klatt, 1987). The advantage of Klatt synthesis is that variables can easily be shifted, meaning that signals can be shifted easily from a synthetic male speaker to female by software commands.

The disadvantage of creating signals by hand using the Klatt formant synthesis software is the cumbersome process of generating signals. Formant synthesis requires extensive knowledge of individual plosive voice onset time (VOT), locus of frication noise, locus of burst energy, and formant transition to obtain the desired voicing and noise features. Additional parameters needed to create synthetic sentences include intensity, duration, and F0 patterns. These factors aid in enhancing a number of linguistic attributes: syllabic structure, vocal effort, stress, speaking rate, syntactic structure, intonation, stress, and gender. Consequently, a single word requires precision work for quality to be optimal (Klatt, 1987).

Another dilemma with the Klatt formant synthesis is that some features are more difficult to generate than others. Klatt and Klatt (1990) noted that Klatt-synthesized female breathiness as well as female pitch contours sounded artificial. Producing natural-like plosives was also a difficult to achieve, which was complicated by the locus of the burst energy and the CV or VC formant transitions. Nevertheless, many speech perception studies have used the Klatt formant synthesis principles to gain understanding listeners' use of speech cues.

Speech Resynthesis. A method of high-fidelity resynthesis developed by Kawahara and colleagues has been recently introduced in speech perception studies (Kawahara, Masuda-Katsuse, & Cheveigné, 1998). The uniqueness of this type of resynthesis is its close resemblance to natural speech. The difference between formant speech synthesis and high-fidelity speech resynthesis developed by Kawahara and colleagues is that the latter one resynthesizes already recorded speech. One may immediately see the disadvantage of high-fidelity resynthesis; without a speech source or speaker, resynthesis is very difficult to perform.

The making of high-fidelity resynthesis can be accredited to Kawahara and colleagues who developed Straight as a speech manipulation tool. Straight uses input speech signals and decomposes them into source and spectral traits (Kawahara et al., 1998). Straight is also designed to enable spectral alterations of pitch, duration, and amplitude.

Straight builds on Dudley's vocoder (see Dudley, 1939) and Klatt's synthesis principles (see Klatt, 1980). In the user interface developed for Straight, the user typically begins with an existing speech signal, which is resynthesized in Straight; that is, the Straight parameters are set to model the existing speech sample. Broad parametric changes (such as shifting fundamental frequency, duration, or the frequency of all formants by a specified proportion) are accomplished relatively easily using the Straight user interface. More focused changes, such as changing the frequency of one formant without changing the others, are less easily accomplished. One way of accomplishing such a goal is to change by hand the spectral matrix in which the values for amplitude and frequency are stored. Another way is to set initial (source) and target frequency points at various times and to morph from source to target. However, initial testing of the method by the author and mentor for generation of static vowels resulted in trajectories that were flattened formants but not perfectly flat.

For optimal resynthesis, a filter shape or analyzing wavelet is superimposed on the formant spectrum or spectral envelope of a complex sound. The area of the spectrum covered by the wavelet determines both the signal to noise ratio and the frequency resolution with greater area covered resulting in better SNR and resolution (Kawahara et al., 1998). Thus, it is important to cover as great an area of the formant spectrum as possible to obtain high frequency resolution and high signal-to-noise ratio. A spectral envelope of the signal is extracted instantaneously and inserted into the speech manipulation system that converts the spectrum into amplitude, frequency, and times, respectively.

Speech Perception and Synthesis

Fry and colleagues were one of the early pioneers in speech perception using synthetic stimuli. In 1962, they conducted a study in which vowels were converted from [ɪ] to [ɛ] to [æ] in a 13-step continuum (Fry, Abramson, Eimas, & Liberman, 1962). By formant synthesis, F1 was incrementally changed from low to high frequencies while F2 was moved in the opposite direction. Results of identification tests indicated that the vowel categories were not clearly defined. Instead, the percent identification of the vowels gradually sloped from one to the other, which is also referred to as continuous perception (Fry et al., 1962). Similar studies have been conducted reaching comparable results. Although these studies have provided understanding of the continuous perception of vowels, most of these have dealt with vowel formants as static entities.

Silent Center. Strange, Jenkins and Johnson (1983) conducted a study testing lax and tense vowels in spoken /bVd/ words. A designated center portion of the vowel duration was silenced (50% for lax vowels and 65% for tense vowels). To test if listeners relied on temporal information, the duration of the silenced portion was equalized across stimuli in two ways, by shortening the silent interval to 57 ms for all stimuli and by lengthening the silent interval to 163 ms for all stimuli. Strange and colleagues also tested if listeners' performance was influenced by the amount and type of information provided in four conditions: initial CV only, final VC only, silent center, and center alone (or isolated vowel). American monolingual listeners were presented with the stimuli and asked to choose which word they heard from a list of 10 alternatives. Results indicated that listeners make increasingly more errors when more acoustic information is removed (e.g., CV only versus both CV and VC) (Strange et al., 1983).

Whole words were on average identified by only about 10 percentage points better than the silent-center stimuli, suggesting that listeners can in fact use CV and VC formant transitions for vowel identification because identification was not substantially reduced when only the CV and VC transitions were available. Temporal manipulations resulted in little change when shortening the silent portion. However, lengthening the silent portion resulted in an increase of errors. Isolated vowel centers (shorter than silenced portion) showed a significant increase in errors compared to the silent center syllables. The study concluded that listeners depend highly on dynamic spectral cues, somewhat on temporal cues, and greatly on completeness of information in that vowels in CV only VC only conditions were very poorly identified in that both initial and final information was needed for good identification of silent center vowels (Strange et al., 1983).

Formant Contours and Synthesis. Hillenbrand and colleagues (1999) conducted a study with the purpose to examine effects of formant contour movements of vowels on vowel identification. Twelve American English vowels were recorded in /hVd/ context, spoken by men, women, and children. Three listening conditions for each vowel were created: natural /hVd/ vowel (NAT), original-formant (OF) synthesized /hVd/ vowel, and flat-formant (FF) synthesized vowel. The OF and FF conditions were both created using Klatt synthesis (Hillenbrand & Nearey, 1999).

To create the synthetic stimuli, acoustic measurements of F0 and F1-F4 were made from LPC spectra extracted every 8 ms (Hillenbrand & Nearey, 1999). Frequency and temporal information were noted for F0 and F1 through F4 at the onset of the vowel, the offset of the vowel, and at the steady state portion. Formant frequencies were measured at the 20% and 80% vowel duration points to avoid measurements from the CV and VC transitions. Using the Klatt synthesis (formant synthesis) method, OF and FF stimuli were generated. For creation of the OF stimuli, synthesis parameters were simply set to match the F0 and F1-F4 values measured every 8 ms using the LPC formant extraction methods detailed above. The FF stimuli were created by identifying a vowel steady point, at which F1 and F2 were judged to be minimally changing and setting values for F1-F4 for the entire vowel to values measured at the steady point. CV and VC transitions were altered accordingly to match these steady point values. Monolingual listeners identified the stimuli (/hVd/) using a closed-set identification task, in which the 10 /hVd/ alternative words were presented and listeners were asked to choose which one they had heard. A total of 900 test signals were presented to each listener in random order (300 original signals, 300 OF signals, and 300 FF signals).

Results showed that listeners excelled when presented with the natural /hVd/ stimuli (see below). Listeners averaged more poorly when signals were synthesized. Performance decreased even further when formant contours were flattened (NAT: 95.4%, OF: 88.5%, FF: 73.8%) (Hillenbrand & Nearey, 1999). Analysis of the vowel categories revealed interesting patterns. For example, performance for the vowel /i/, which has little vowel drift, only decreased minimally with flattening of the formants. On the other hand, a significant decrease in identification was noted for /e^{I}/ which is known to have more spectral change due to its diphthongized features. Noteworthy was that identification patterns appeared to differ across vowel categories.

The second part of the experiment tested if listeners' performance changed with decreased information presented. Four listening conditions were generated: natural /hVd/ utterance, natural vowel only, original formant synthesized /hVd/ utterance, and original formant synthesized vowel only. Identification scores for naturally produced stimuli revealed that listeners performed slightly but significantly worse when consonant context was removed. The difference between synthesized OF /hVd/ utterances and OF isolated vowels was too small to be significant. Furthermore, naturally spoken stimuli (both /hVd/ and isolated vowels) were noted to be significantly more intelligible than the original formant synthetic signals (NAT /hVd/: 96.7%, NAT vowel alone: 94.4%, OF /hVd/: 91.0%, and OF vowel alone: 90.3%).

The findings of this study suggest that spectral change plays an important role in vowel identification. Vowel intelligibility also increased when in consonant context rather than in isolation. Maintaining original dynamic formant patterns of vowels in synthetic stimuli resulted some decrease in intelligibility compared to natural speech, suggesting that formant synthesis fails to reproduce naturally spoken vowels with complete accuracy. Lastly, listeners were noted to perform differently for individual vowels in formant flattened conditions, implying that listeners use formant dynamic information more for some vowels than others (e.g. /i/ versus /e^{I}/).

Although Hillenbrand and colleagues demonstrated that CV and VC information, synthesis, and flattening of formants all influence vowel identification negatively, this experiment was performed with monolingual listeners only and used formant synthesis. One question to be answered is whether high-fidelity resynthesis of speech differs from the Klatt synthesis. Another question is whether bilinguals would be affected differently by these changes than monolinguals.

Liu and Kewley-Port (2004) investigated the question of whether Straight-resynthesized speech is recognized differently than Klatt-synthesized speech by comparing listeners' ability to discriminate vowels using Klatt and Straight synthesis, respectively. Listeners were tested discriminating several vowels (/ɪ, ɛ, æ, ʌ/) in syllable, phrase, and sentence contexts. Shifts of formant peaks were performed for the target vowels by determining their spectral locations through analysis of matrices generated by Straight and copying and pasting the associated spectral peak intensity values to a higher frequency location and replicating values for the spectral troughs. Fourteen-step continua, from smaller to larger formant frequency changes (0.9 to 10% for most vowels) were generated separately for F1 and F2 using this method. Liu and Kewley-Port found similar thresholds for formant discrimination for Straight-synthesized stimuli as had been previously found for Klatt-synthesized stimuli, indicating that the authors' method of altering the more natural-sounding Straight-synthesized stimuli resulted in a valid collection of data, in that listeners responded similarly in a discrimination task as they had in previous studies with Klatt-synthesized stimuli. Nevertheless, the results for all three listening conditions showed a slight inflation of the vowel discrimination scores when synthesized by Straight rather than by Klatt synthesis.

Generation of a large number of stimuli using the methodology employed by Liu and Kewley-Port (2004) would be at least as time-consuming as using Klatt synthesis. The advantage of STRAIGHT is in the naturalness of the resulting stimuli, not in the time taken to create them. Because the stimuli generated using Straight are more similar to natural speech, it follows that the use of better quality synthesis may result in patterns of listener performance that are more representative of listeners' responses to natural speech.

Assman and Katz (2005) conducted a study to test perceptual differences between stimuli created using formant synthesis and stimuli created using Straight. Using the Hillenbrand & Nearey protocol (1999), Assman and Katz replicated the Hillenbrand & Nearey (1999) study to re-examine the role of formant contours for vowels synthesized using Straight. Results of the Assman and Katz (2005) study were compared to those of the Hillenbrand et al. (1999) study to evaluate the differences between Straight and Klatt formant synthesis. Twelve vowels were presented in /hVd/ context in three listening conditions: natural speaking condition, formant synthesized conditions, and Straight resynthesized conditions. Comparison of natural speech, Straight, and Klatt synthesis revealed that listeners showed significantly better vowel identification performance for Straight synthesized stimuli than for Klatt synthesized stimuli.

While these studies have shown that native listeners use the cues of duration and vowel formant dynamics, less is known about how bilinguals use these cues. That is, the cues used may be different or they may be used differently. It also remains to be answered whether more natural sounding stimuli result in less effect of synthesis for isolated vowels.

Second Language Acquisition and Foreign Accent

Proficiency in a second language is frequently associated with a speaker's reduction of a foreign accent. However, proficiency is truly determined by a second language (L2) speaker's linguistic abilities in the areas of L2 syntax, lexicon, pragmatics, phonology, and phonetics (Major, 2001). Nevertheless, L2 proficiency is often most apparent in a speaker's spoken or written language (Major, 2001). Additionally, an L2 learner's production is commonly characterized by a foreign accent, which can be defined as the difference between the pronunciation patterns of a non-native speaker of a language and those of a native speaker (Flege, 1995; Major, 2001). Thus, foreign accent is primarily influenced by phonetic and phonological factors, meaning vowels and consonants may be distorted, substituted or even omitted, as compared to standard production of native speakers of the target language (Flege, 1995; Major, 2001). Similar to production, differences may also exist in the way that native and non-nonnative speakers perceive the sounds of a language (Flege, 1995).

Possible factors affecting speakers' degree of accent or difficulty in correctly identifying second language speech sounds are differences in the age of onset of learning (AOL) the second language, differences between vowel and consonant inventories between the first language (L1) and L2, and differences in the degree of linguistic experience and exposure to L2 (Flege, 1995; Major, 2001).

Returning to the issue of foreign accent, it has been suggested that brain plasticity plays a crucial role in learning an L2. Older learners (i.e., adolescents and adults) may have decreased flexibility of sensorimotor neuro-wiring; i.e., speech movements become automated and harder to change with increasing age (Flege, 1995; Major, 2001). As a consequence, adults learning a second language will be more likely to speak with a foreign accent due to their previously wired neuro-motor patterns. Thus, with late L2 acquisition, previously learned L1 sound production patterns tend to influence and distort intended L2 sound production causing decreased speech intelligibility (Flege, 1995).

To what degree L1 sounds affect L2 sound perception and production skills, and if there are other aspects that may account for different proficiency levels, remain key questions. Two prominent explanations of the L1-L2 relationship are Best's Perceptual Assimilation Model (PAM) and Flege's Speech Learning Model (SLM). These two models each describe hypotheses regarding L2 learners' potential acquisition of non-native sounds.

Perceptual Assimilation Model

The Perceptual Assimilation Model (PAM) offers predictions of how an L2 learner will identify and discriminate non-native sounds. According to Best (1995), non-native sounds are assimilated into the already established L1 sound system. Best explains that patterns of identification of single non-native sounds can be narrowed down using three perceptual principles. Non-native or L2 sounds can either be assimilated to a *native category* or they can be perceived as *uncategorizable* speech sounds that are still within the native phonological space. Finally, L2 sounds may be perceived as *non-speech* and will not be assimilated into the native phonological space but fall outside.

If an L2 listener is presented with pairs of different L2 sounds, PAM predicts discrimination of L2 sounds to fall out into one the following categories: Two-Category Assimilation (TC Type), Category-Goodness Difference (CG Type), Single-Category Assimilation (SC Type), Both Uncategorizable (UU Type), Uncategorizable versus Categorized (UC Type), and Nonassimilable (NA Type) (Best, 1995).

Two-Category (TC) assimilation occurs when both L2 sounds are distinctively different within the L1 sound system. That is, the sounds are both perceived as speech sounds and each falls within a different native category. The listener is expected to perceive the presented L2 sounds as different and to assimilate them into already existing L1 sound categories. In the case of two L2 sounds being less distinct, meaning their proximity is close to a common L1 sound for both, the L2 sounds may in fact merge and be perceived as only one L1 sound which is also referred to as SC Type assimilation.

According to the PAM model, two-category type sounds are discriminated better than single-categorical sounds due to the difference factor. Category-Goodness (CG) difference is when two sounds are perceived to match one L1 sound but one is perceived as a better exemplar of the category than the other. The listener will perceive a slight difference in the sounds. Thus, the production of one L2 sound will be closer to the ideal while the other L2-sound is perceived as a relatively poor example of the intended L1 sound. Consequently, the listener will approximate the L2 sounds to one L1 sound, but will still perceive a slight difference between them.

Assuming the principles of PAM to be true, and given a bilingual learner of Spanish and English (Spanish L1), discrimination of English vowels is expected to be compromised since English contains 11 monophthongal vowels (Crystal, 1997), compared to 5 in Spanish (Dalbor, 1969). Furthermore, the English vowel quadrilateral is more densely populated with front and back vowels than Spanish. If an L2 listener is presented with two front vowels, both with raised jaw (e.g., /i/ vs. /ɪ/) the sounds may merge and be perceived as only one sound: /i/ (e.g., SC type or CG type). The Spanish listener may only perceive a temporal difference between /i/ and /ɪ/ but may produce these as a long and short /i/ (Dalbor, 1969).

Although Best’s assimilation model of non-native sounds suggests that perception depends on perceived dissimilarity between L2 sounds and their “goodness” in comparison to L1 sounds, PAM does not address the role of AOL on the bilingual learner’s speech perception and production patterns. The Speech Learning Model (SLM) by Flege addresses the relationship between production and perception as well as the impact of ongoing L2 exposure (Flege, 1995, 1996; Flege, Munro, & MacKay, 1995).

Speech Learning Model

Flege investigated the issue of why, after a certain age of onset of L2 acquisition (AOL), children demonstrate less native-like production in their L2 (Flege, 1981). Previous studies have indicated that in order to acquire an L2 with little to no accent, the L2 language had to be acquired before a certain age (Flege, 1995; Major, 2001). This reasoning links to thinking that plasticity of neuro-sensory wiring decreases in adolescence, also known as the Critical Period Hypothesis (CPH) (Flege, 1995; Major, 2001). The CPH was originally designed to explain L1 acquisition (Lenneberg described in Major, 2001). Evidence of the CPH hypothesis has been found in traumatic brain injury (TBI) victims. Comparisons of young children and adolescents with TBI found that children recovered nearly completely, whereas adolescents sustained some cognitive deficit (Major, 2001).

However, with the 1981 study, Flege found suggestions that CPH alone did not explain accentedness. The study tested children who presumably were within the CPH and had increased sensorimotor abilities compared to younger children (Flege, 1981). However, the data revealed that despite optimal conditions the children demonstrated difficulties learning both vowels and consonants (Flege, 1981). This led a belief that “learning” cutoff age was either ill-defined, nonexistent, or a third factor accounted for the learning barrier. Conversely, several perceptual studies indicated that there is a correlation between onset of learning an L2 and accentedness, which will be discussed in detail later.

To account for this paradox between increased sensorimotor skills, increased foreign accent and decreased perception performance, Flege and colleagues created a set of hypotheses called the Speech Learning Model (Flege, 1995). The objective of the Speech Learning Model (SLM) is to explain L2 speakers' learning process and to explain the importance of AOL on foreign accent. The four postulates of the SLM suggest that an L2 learner's L1 system remains adaptive over the life span and can be applied in learning an L2 (postulates 1 and 2). The L1 phonetic system reorganizes itself on introduction of new L2 sounds (Flege, 1995). Further, the SLM suggests that bilinguals strive to establish distinct categories separate L1 and L2 sounds (Flege, 1995). Flege and colleagues created the SLM as a result of a number of speech-perception and production studies that involved L2 learners. A collection of seven hypotheses comprises the SLM; all of these are believed to be important for second language learning.

The principles of the SLM and the PAM are similar to the degree that predictions regarding L2 perception and production depend on perceived similarities and differences of L2 phonemes from L1 phonemes. However, the SLM specifies that small differences between an L1 and an L2 sound can be discerned at the allophonic level; that is, duration may be what sets an 'L1 vs. L2' pair apart (Flege, 1995). For example, the English vowel /æ/ does not exist in the Spanish language. A Spanish listener may perceive the sound as the phoneme /ɛ/ or /ɑ/ because both share similar spectral properties with /æ/, but the sound may be perceived as a longer (or otherwise different) variant of one of these L1 phonemes.

Also, according to the SLM, early onset of L2 will result in improved phonetic realization or production, as well as improved perception; that is, the learner will be able to tell L1 and L2 sounds apart even if the cues differ minimally. However, with increasing age of onset of the L2, a bilingual listener-speaker will experience increased difficulties discerning the small differences between L1 and L2 sounds that may be perceived as allophonic differences (Flege, 1995). The SLM further asserts that L1 and L2 sounds are grouped together in a common phonetic space, although they have diaphonic realization; that is, L1 and L2 sounds will be produced in the appropriate language context (e.g., L1 sounds in L1 and L2 sounds in L2). The theory also asserts that this perceptual merging of L1 and L2 sound inventories will also lead to a merging of production of two L2 sounds that are both similar to a single L1 sound. Thus, when L1 and L2 categories are too similar, the realization of L2 categories, and thus perception and production, may be altered from native-speaker norms.

An example of sound merging can hypothetically be found in a Spanish speaker's production of the two American English vowels /e$^{\text{ɪ}}$/ and /ɛ/. A Spanish listener may be able to discern the differences between the two sounds; however, when integrating the sounds into the Spanish vowel system the Spanish speaker might encounter difficulties mapping the sounds to separate sound categories. Spanish has the five vowels /ɪ, e, u, o, ɑ/, whereas American English has 11 monophthongal vowels, not counting rhoticized vowels and diphthongs depending on dialect (/ɪ, i, e$^{\text{ɪ}}$, ɛ, æ, ʌ, u, ʊ, o, ɔ, ɑ /) (Crystal, 1997). In considering the front vowels, American English /e$^{\text{ɪ}}$/ and /ɛ/ may both be perceived as exemplars of Spanish /e/, although they may be heard as distinct from one another. Since they are both identified as members of a single category, the theory suggests that the two sounds will not be produced distinctively, even if the listener can perceptually discriminate between the two sounds when presented in pairs.

The SLM also stresses the dynamic nature of the learning process. Flege suggests that both L1 and L2 sound systems influence each other with time and that this bidirectional influence may actually contribute to shifts in perception in both L1 and L2 (Flege, 1995). Initially, this may mean that increased language experience may improve L2 perception as well as production. However, research has shown that L1 sounds may undergo spectral changes in production for highly experienced bilinguals (Flege, 1995).

Moreover, the model suggests that for L1 and L2 sounds that are very similar, an experienced bilingual may produce a single vowel that is intermediate between the native speaker norms for L1 and L2. That is, the experienced bilingual may produce the same, intermediate, sound in L1 and L2 and the sound's production will be different from that of native speakers of either language (Flege, 1995).

The complexity of perceptual assimilation patterns was demonstrated by Rochet (1995). Rochet (1995) demonstrated how foreign accent and perceptual biases may relate to each other. Rochet argued that foreign accent is rooted in biased perception and consequently that imitation of vowels would be distorted. The study comprised a production and a perception portion. For the production portion, Portuguese and English listeners were asked to repeat single-syllable words that contained the French vowels /i, y, u/, and /ɑ/. Rating of the productions was performed by native French listeners. Analysis of the rating data revealed that when Portuguese speakers' production of /y/ was inaccurate, the native French listeners mostly rated their productions as /i/. The English speakers’ production of /y/, however, was rated mostly as /u/. The perception portion of the experiment required the participants to identify /i/ and /u/. Using formant synthesis, the second formant of the vowel was incrementally changed in 100 Hz steps from 500 to 2500 Hz. Comparison of identification performance of French, Portuguese and English listeners revealed that the vowel boundaries were located at different frequencies for each. For English listeners the /u/-/i/ boundary was located at about 1900 Hz, while for Portuguese listeners, the boundary was located at about 1600 Hz (Rochet, 1995). Analysis of identification patterns for native listeners of French showed that the vowel boundary between /u/ and /y/ was centered around 1200 Hz, whereas the /y/-/i/ boundary was located at about 2100 Hz. Noteworthy is that native listeners of French (which has a relatively large vowel inventory) were able to reliably identify (i.e., with 100% identification performance) tokens as /i, y/ and /u/ within an F2 range that was only slightly larger that that needed for native listeners of Portuguese (which has a relatively small vowel inventory) to reliably identify only two phonemes (/u/ and /i/). That is, the Portuguese listeners showed larger regions of inconsistent performance than did the French, so that performance for the French listeners was more categorical. To illustrate, the decrease of vowel identification from 100% /u/ to 0% /u/ (100% /i/) for the Portuguese listeners stretched from 1200-2200 Hz; for the French listeners, however, the decrease for /u/ began at 900 Hz and 100% identification as /i/ began at 2300 Hz. For native listeners of English (which, like French, has a relatively dense vowel inventory), performance was also more categorical (i.e., inconsistent performance extended over a smaller frequency range). From these data, Rochet concluded that not only do different languages differ in frequency boundaries for the “same” phonemes, but perceptual vowel category boundaries extend to adjacent categories, leaving no uncommitted space (as evidenced by the broader range of inconsistent performance for Portuguese). Thus, the Rochet study shows that prediction of vowel production and perception patterns in a second language cannot be determined based on the phonemic level alone; rather, acoustic measurements of vowels in both languages may be necessary to predict vowel assimilation patterns.

This complexity of interaction between L1 and L2 phoneme inventories is also seen in consonants. A comparison of the number of consonants between the two languages (L1 and L2) has important implications for what phonemes will be produced with a foreign accent; however, of equal importance is a comparison of the languages’ use of spectral cues concerning place, manner, and voicing (Flege, 1995, 1996). These distinct spectral features not only make consonants unique in isolation, but an interaction between vowels and consonants is often seen at the point of transition (e.g., VOT of plosives or locus of frication noise) (Strange, 1999). Even at the level of syllable shapes or word-position, the variation of consonants across context appears to account for compromised perception or production skills e.g., syllable initial production and perception may be more native like than syllable-final production or perception (Flege, 1995; Major, 2001; Strange, 1999). The reasoning for this perceptual limitation may lay in coarticulation (e.g., “key” versus “cooh”), assimilation patterns (e.g., ‘would you’ becomes ‘wou cha’), or simply increased amount of new information that is processed and stored (e.g., phonological short-term memory) (Flege et al., 1995; Major, 2001).

Flege also suggests that listeners will be more prone to detect and reproduce an L2 phonetic difference when they encounter it in early life and when their encounter with that difference is lengthy. One study that supports this prediction is Flege et al. (1995), in which Italian-English bilinguals whose contact with English was early and lengthy were better able to produce word-initial consonants /r, ð, θ/ in English and were able to produce distinctive contrasts, in word final consonants /b/-/p/, /t/-/k/, and /k/-/g/ in English better than later bilinguals with less extensive contact with the L2 (Flege et al., 1995).

American Vowels. The Speech Learning Model has been the starting point for many cross-linguistic studies. Most studies using the SLM have, however, shown greater concern with consonant perception and production than with vowels (Flege, 1995). As mentioned earlier, data from previous studies suggest that monolinguals' ability to identify vowels depend on formant transitions, duration, and formant frequency and formant dynamic. Vowel perception by bilinguals or L2 speakers is assumed to be influenced by the vowel inventory. Cross-language comparisons of phoneme inventories have indicated that languages' vowel and consonant repertoires and the interactions between them have a crucial influence on L2 acquisition despite the importance of L1 and L2 use of acoustic cues noted above (Flege, 1995). One observation made is that languages with sparsely populated vowel spaces (i.e., languages with a relatively small number of vowels in their inventories) may give the learner limited examples of vowels for production and perception (Flege, 1995, 1996). Vowels that vary from the few vowels in the inventory may be perceived as poor examples of one of the small number of vowels in the inventory, rather than as separate vowels (cf. Best, 1995). An example of a language with a relatively sparse vowel inventory is Spanish, which is typically described as having five vowels (/i, e, a, u, o/) (Dalbor, 1969). American English, on the other hand, is typically described as having a relatively dense vowel inventory, with 11 monophthongal vowels (/i, ɪ, e, ɛ, æ, u, ʊ, o, ɔ, ɑ, ʌ/) (Crystal, 1997). Thus, a person whose first language is Spanish and who is learning English as a second language must create at least six new vowel categories in their vowel space.

Because American English vowels are, acoustically speaking, closely spaced, neighboring vowel categories for adults, women, and children may overlap with one another (cf. Peterson & Barney, 1952). Consequently, the number of vowels in the inventory is not the only essential factor for misinterpretation; the spectral properties of the sounds may also play a role. Although a two-dimensional formant diagram (depicting the first and the second formants) for a single talker or talker group may suggest that vowels are distinctively separated, a depiction of vowel productions across age and gender groups shows much more overlapping of the acoustic features of the sounds (e.g., /i/ produced by an adult female may overlap with a child's /ɪ/ production) (Peterson & Barney, 1952). If spectral change from the beginning to the end of the vowel is plotted as the third dimension, however, it then becomes apparent that English vowels bear a significant dynamic property which is also referred to as "vowel drift," which helps to separate vowels in the space (Hillenbrand et al., 1995). Whether vowel drift is a factor to be found in other languages has been less investigated. What has been found though is that second language learners are perceived better by native-English listeners if spectral and duration cues are mastered (Flege et al., 1995; Kewley-Port, Akahane-Yamada, & Aikawa, 1996). If a second language learner were unable to use the vowel drift information, however, American English vowels might seem much more difficult to discriminate than if the vowel drift information were available.

Summarizing the previous sections on SLM, PAM, and American vowels, L2 learners are faced with a number of potential obstacles that increase the difficulty of perception. In the following section, speech perception and production studies are discussed to delineate similarities and perceptual differences between bilinguals and monolinguals.

Studies on L2 Acquisition

As mentioned earlier, Strange and colleagues (1983) found evidence that monolingual listeners depend on dynamic vowel information (CV and VC transitions) and also that increased vowel information leads to better performance than less information (e.g., both CV and VC transitions versus CV transition alone).

Similar findings have been found using bilingual listeners. Mayo and colleagues (1997) examined whether L2 age of onset of learning is a factor influencing speech perception in different listening conditions. Monolingual and bilingual listeners were tested. Three bilingual groups were used: 1) bilingual since infancy (BSI), 2) bilingual since toddler (BST), and 3) bilingual-post-puberty (BPP). All participants listened to short sentences ending with a target word. One set of sentences contained high context predictability whereas the context of the second set was manipulated to reduce predictability of the target word. The practice portion of the sentences was presented without noise. Step two of the study was to determine the listeners' speech recognition threshold when stimuli were presented in noise. Once the threshold was established, several signal-to-noise ratios (SNR) were selected for the next portion encompassing

SNRs for which 15%-85% correct performance was obtained in part 2. Stimuli were then presented at each of these SNR (part 3) (Mayo et al., 1997). Data analysis from responses to part 3 included percent of correct responses, level of predictability of words, and noise level. Results of the study suggested that the bilingual post-puberty listeners performed more poorly on both high- and low-predictability targets. The early AOL bilingual groups (e.g., BSI and BST) performed similarly to monolinguals on predictable sentences; however, low-predictability targets were identified more poorly by early bilinguals than monolinguals. Z-scores for the noise levels indicated that late bilingual learners (BPP) depend on higher thresholds for good identification. Ultimately, late AOL of L2 appears to cause decreased perception skills in which the threshold is increased as well as when target words are unrelated to the context. The findings of this study were supported by Febo’s study (2003) in which bilinguals were shown to need more audible information for speech recognition as well.

Lopez (2004), using the methodology of Strange et al. (1983), wanted to see if silent center vowel perception of monolingual English speakers and Spanish-English bilinguals differed. The bilingual participants were divided into two proficiency groups: high and low. Results revealed that monolingual speakers excelled in identifying silent-center vowels. Highly proficient speakers performed better than less proficient bilinguals, but not as well as the monolingual speakers. These data suggest that even proficient bilinguals may need more information for vowel identification than monolinguals.

Sebastian-Galles and Soto-Faraco (1999) tested the question of whether AOL affects bilinguals in discriminating between two vowel pairs and two consonant pairs, presented in CV.CV (e.g., /gesi/-/gezi/) or CVC.CV (e.g., /nesku/-/nɛsku/) non-word context. Gating procedures were used, in which about 30 ms of the word was presented on the initial gate and 10 additional ms was presented on each successive gate. The entire non-word was presented on the tenth gate. Non-words were presented in a two-alternative forced-choice format; that is the listener was asked to choose which of the alternatives they had heard at each gate. Two groups of bilingual speakers were tested; both had acquired Catalan and Spanish at an early age. One group had been exposed only to Spanish up until 4 years old. Hereafter, both Catalan and Spanish were introduced. The other group had been exposed to both Spanish and Catalan from birth on. Comparisons of the two groups found that Spanish dominant speakers or "later" learners needed more information than those who were bilinguals from birth to discriminate between the phonemes presented.

Meador, Flege, and MacKay (2000) also examined the importance of age of arrival (AOA) in country for word identification skills. In addition, the effects of percentage of L1 usage and Length of Residence (LOR) on the recognition of English words by native speakers of Italian bilinguals were examined. Since Italian and English differ in both consonant and vowel inventories (see Meador et al., 2000), the SLM predicts that age of onset (AOL) as well as exposure to L2 may be integral factors for optimal perception. For comparison, bilinguals with different (AOL) and different percentage of L1 usage, as well as monolingual listeners were tested repeating English sentences with low semantic predictability presented in noise (Meador et al., 2000). Results showed that longer LOR correlated positively with correct word identification. Since identification performance was by having the listeners repeat the target sentences, foreign accentedness of the bilingual participants was also studied. Results showed that

L2 speakers with increased LOR demonstrated reduced accentedness as well as increased word identification abilities. A further finding indicated that early bilinguals who used their L1 less demonstrated increased word recognition ability. Thus, age of arrival might be taken as a guide of the learner's state of neurological development at the time L2 learning begins. This study suggests that accentedness is affected by both AOA and AOL, as well as by the continued use of L1.

MacKay, Meador, and Flege (2001) elaborated on the Meador et al. study by examining whether age of arrival (AOA) in country is important for L2 perception, whether use of L1 impairs L2 production and perception, and lastly whether differences in phonological short-term memory (PSTM) between L1 and L2 participants would differ. Using CVC words in semantically unpredictable sentences, participants were asked to repeat target words (last word of the sentence). Italian-English bilinguals (Italian being participants’ L1) were purposefully chosen by the authors due to the phonetic and phonotactic differences between the two languages. For instance, Italian does not have many words that end in consonants and the Italian consonant inventory is smaller than that of English (see MacKay et al., 2001). Referring to the SLM, this difference may either cause certain English acoustic cues to be ignored by an Italian listener, while other cues may be weighted more heavily. Bilingual participants were grouped according to AOA, length of residence (LOR), and percent use of Italian. Four groups were formed: early-low (early AOA mean = 7 years, LOR mean = 40 years, low percent use of Italian = 8%), early (early AOA mean = 7 years, LOR mean = 40 years, high percent use of Italian = 32%), mid (AOA mean =14 years, LOR mean = 34 years, percent use of Italian

= 20%), and late (AOA mean = 19 years , LOR mean = 28 years, percent use of Italian = 41%).

Data analysis for the identification task showed no overall AOA effect between the four native Italian speaker groups. However, an interaction between consonant position and L1 use was found: early Italian-English bilinguals using L1 seldom were noted to have fewer difficulties identifying initial consonants than final consonants. Early Italian learners with greater use of L1 demonstrated increased errors in both initial and final word position compared to early low L1 users. Furthermore, late bilinguals were noted to demonstrate increased errors in initial consonant identification compared to monolingual English speakers and early learners. Different from the Sebastian-Galles and Soto-Faraco study, MacKay and colleagues attributed compromised identification skills primarily to the degree of use of L1 and not age of onset of L2.

Bohn and Flege (1999) examined predictions of the SLM in adults learning new L2 vowels with regard to both production and perception of bilinguals. In the study, Bohn and Flege addressed two questions: 1) whether adults can learn to produce and perceive a second language vowel category for which no counterpart exists in their native language and 2) whether there is a relation between their production and perception skills for this new category. The vowel /æ/ was chosen because it is not in the German vowel inventory. To test the questions, monolingual speakers of English and bilingual speakers of German and English were asked to produce the two vowels /ɛ/ and /æ/ in bVt context in a sentence (e.g., “I will say ___ “) (see Bohn and Flege 1999 for further details). Two proficiency groups were recruited on the basis of their AOL, LOR in the U.S, and L2 training: experienced versus inexperienced bilinguals. Measurements of duration and the frequencies of F0 and F1-F3 were performed for each vowel and compared across speaker groups. Acoustic analysis for the three speaker groups showed that some of the experienced bilinguals' vowel patterns resembled the monolinguals' production, while inexperienced native German speakers' productions of /ɛ/ and /æ/ were noted to overlap (Bohn & Flege, 1999). Analysis of duration patterns revealed experienced L2 speakers' production to be similar to that of the native speakers. Inexperienced speakers, on the other hand, showed overall reduced duration for both vowels (Bohn & Flege, 1999).

Using Klatt formant synthesis, 33 stimuli were created with differing F1-F3 frequencies and duration (11 stimuli per duration - 3 durations: 150 ms, 200 ms, and 250 ms). The three listener groups identified the stimuli as "bet" or "bat." Analysis of the percent of words identified as "bet" for the three durations indicated that both groups of L2 speakers performed differently from the native listeners; greater differences from monolinguals were found for inexperienced than for the experienced L2 learners A spectral effect was found for the experienced listeners, in which lowering F1 and raising F2 resulted in more "bet" responses. For the inexperienced listeners' performance was affected more by duration information than by spectral information. Lengthening /ɛ/ from 200 ms to 250 ms resulted in more /æ/ responses regardless of spectral information. Further analysis was made comparing production and perception. Spectral and duration changes were arrayed to compare perception results to production details. The main pattern to be found was that all experienced L2 participants who showed more native-like production patterns also showed native-like perception patterns. The reverse, however, was not true; many experienced L2 participants who showed native-like perception patterns did not show native-like production patterns. None of the less experienced L2 participants showed native-like production patterns. A suggestion to be made from this study is that adult learners (after age 30) appear to be able form new sound categories when L2 exposure is intensive. However, as the SLM states, category formation is easiest when new L2 sounds are dissimilar to any of the L1 sounds. Further, these results appear to support Flege’s hypothesis that accurate perception tends to precede accurate production.

The studies described in this section provide evidence that even proficient bilinguals depend on more acoustic information than monolinguals in difficult listening conditions (Febo, 2003; Lopez, 2004; Mayo et al., 1997; Sebastian-Galles & Soto-Faraco, 1999). Furthermore, data indicate that increased accentedness and decreased perceptual abilities are associated with decreased L2 exposure and increased use of L1 (Bohn & Flege, 1999; Flege et al., 1995; MacKay et al., 2001; Meador et al., 2000). Furthermore, studies have shown that spectral and duration cues may be used differently by bilinguals than by monolinguals (Bohn & Flege, 1999). However, no studies have, as we are aware of, investigated the role of both duration and vowel formant dynamic cues on vowel identification, and how these may differ across different L2 proficiencies by bilinguals. It also remains to be answered if bilingual listeners’ performance is affected by Straight synthesis.

Purpose of Study

This study focuses on the perception of American English vowels. The performance of monolingual American speakers and Spanish bilingual speakers will be compared. The study compares perception of monolingual perception and bilingual listeners’ identification of the vowels /i, ɪ, e$^{\text{ɪ}}$, ɛ, æ, ɑ/ when eliminating consonantal environment from /bVd/ syllables, reducing temporal cues, and flattening formant contours using resynthesis. The performance of monolingual American English speakers, highly-proficiency Spanish-English bilinguals, and low-proficiency Spanish-English bilinguals will be compared. Thus, the present study will address five main research questions.

1. Are vowels perceived differently in isolation than in whole words for these groups?
2. Are vowels synthesized using high fidelity synthesis (Straight) perceived differently than naturally produced vowels?
3. Do high- and lower-proficiency bilinguals use vowel formant dynamic and duration cues differently than monolinguals?
4. Do the effects of vowel isolation, synthesis, formant dynamic cues, and duration cues differ across the six vowels studied?
5. Do patterns of confusions differ across the listener groups?

Chapter 2

Method

Participants

Speakers. Five monolingual native English speakers (3 males and 2 females) participated as speakers in this experiment. All were screened to exclude persons with a history of speech or hearing impairment. A trained phonetician and native English speaker determined that all five spoke English without a strong regional accent. Speakers ranged from 25 to 35 years of age. Two speakers originated from Florida, whereas the remaining three were from New York, Illinois, and Pennsylvania.

Potential Listener Screening and Language Background Questionnaire. Potential listeners, both monolingual and bilingual, were screened for accent and voice quality during a telephone interview prior to participating in the study. Persons with poor voice quality or a strong regional accent in English were excluded from the study. Potential listeners were also screened to exclude those with a history of speech or hearing impairment. During the telephone interview, potential bilingual participants were also given a preliminary classification as higher or lower proficiency based on the screener's perception of their degree of accentedness in English, but final classification was based on age of onset of learning English. The two telephone screeners were graduate students in speech-language pathology.

Upon their arrival at the study site, all participants completed a language background questionnaire. For monolinguals, demographic and basic language background data were collected (see Appendix A). Persons who indicated fluency in a second language were eliminated from the study. Data from the language background questionnaire were also used to further screen monolingual participants for dialect. Speakers of Southern varieties of English were excluded due to the vowel shifts and mergers typical of these dialects. Speakers from the Tampa area or other urban regions of south Florida were preferred. According to Labov (2005), monolingual English speakers from South Florida, specifically urban areas, do not typically demonstrate vowel mergers characteristic of some Southern dialects.

A more detailed language background questionnaire was used for the bilingual listeners (see Appendix B). In addition to basic demographic information, one set of questions probed the participants' languages spoken, dialect background in Spanish, age of onset of first learning English, the context in which English was first learned, number of years in the United States, and the age of onset of learning English in a context in which the language was used extensively, or age of onset of learning intensively (AOLI). AOLI is defined as the time at which L2 learners are exposed to English or their L2 intensively, typically indicated by immersion in an L2 culture. In an additional set of questions, participants were asked to state their percent of daily use of their two languages in work, home and other contexts, as applicable, and to compare their abilities and to indicate which of their two languages they felt most comfortable using in the domains of speaking, listening, reading, and writing. Answers to these questions were used to estimate self-rated language dominance for each domain.

Listeners. Sixty persons between the ages of 18 and 59 participated as listeners in this experiment. Seventeen monolingual listeners (females=15, males=2) were selected based on lack of a strong regional accent in English. Basic demographic and language background data are as follows. Fourteen of the participants originated from Florida, of which all but one resided in the Tampa Bay area. The remaining three participants were from New York, Ohio, and California, respectively. The mean age of the participants was 21.24 years and ranged from 18 to 38 years of age.

A total of 43 Spanish-English bilinguals were included as listeners for this study. Potential bilingual listeners were identified as persons who described themselves as speaking Spanish and English only; persons who indicated fluency in a third language were excluded.

Speakers of Castilian varieties of Spanish were also excluded from the study, but speakers of any American variety of Spanish were included. Although there are many differences in pronunciation patterns across American varieties of Spanish, they are typically viewed as being more similar to one another in pronunciation than to Castilian varieties of Spanish (Dalbor, 1969). A variety of dialect backgrounds was permitted because this diversity is representative of the population of Spanish-English bilingual persons in south Florida.

Based on the potential participants' age of onset of learning English intensively (AOLI) and preliminary accentedness classification by the two screeners, 25 participants were classified as highly-proficient early learners of English and 18 were classified as less-proficient later learners. Tables 1 and 2 illustrate selected demographic and language background data for each subject in the highly-proficient bilingual and less-proficient bilingual groups, respectively. As shown in Tables 1 and 2, all of the highly-proficient bilinguals began learning English intensively at age 12 or earlier, while all of the less-proficient bilinguals began learning English intensively at age 15 or later.

Of the 25 highly-proficient bilinguals, 24 rated themselves as equally proficient in Spanish and English or more proficient in English in at least three of the four domains (speaking, listening, reading and writing); the remaining participant rated herself as English dominant in two domains (speaking and writing) and Spanish dominant in two domains (listening and reading). Additionally, thirteen of these participants indicated that they were born and raised in the U.S. and began learning English when they started school or preschool and were educated exclusively in English. All were categorized by the screeners as having only a slight or no foreign accent in English.

As shown in Table 2, all of the less-proficient bilinguals began learning English intensively at age 15 or later. Of these 18 participants, 12 rated themselves as Spanish dominant in all four domains; two rated themselves as balanced or Spanish dominant in three domains; two rated themselves as Spanish dominant in two domains and English dominant in two domains; and two rated themselves as English dominant in three domains. All were categorized by the screeners as having a noticeable foreign accent in English.

Code	Age	Gender	Origin	AOLI	Time in U.S. (yrs)	% English used at:			Most comfortable language for:			
						Work	Home	Other	Speak	Listen	Read	Write
HI02	18	F	Dominican Republic	7	17	100	0	75	E	E	E	E
HI04	18	F	Mexico	6	13	50	50	--	E	E	E	E
HI05	19	F	Cuba	4	19	80	80	85	E	E	E	E
HI06	19	F	Mexico	5	16	--	75	80	B	E	E	E
HI07	21	M	Costa Rica	1	4	100	100	99	B	B	E	E
HI08	19	F	Nicaragua	8	11	100	50	70	E	E	E	E
HI09	18	M	Mexico	4	18	--	99	99	B	E	E	E
HI10	19	F	Nicaragua	6	19	40	20	60	B	B	B	B
HI11	20	F	Cuba	6	20	95	70	80	E	E	E	E
HI12	24	F	Puerto Rico	10	14	100	5	100	E	E	E	E
HI13	19	M	Peru & El Salvador	3	19	100	80	95	E	E	E	E
HI14	20	F	Cuba	4	20	100	0	100	E	E	E	E
HI16	19	F	Mexico	6	19	--	50	50	S	E	E	E
HI17	19	F	Cuba	4	19	--	95	98	E	E	E	E
HI18	35	F	Venezuela	9	5	100	10	100	E	S	S	E
HI19	18	F	Cuba	4	18	--	50	--	E	E	E	E
HI20	27	F	Venezuela	4	8	60	0	--	B	E	E	B
HI22	18	M	Peru / El Salvador	5	18	100	90	95	E	S	E	E
HI23	29	F	Puerto Rico	9	20	100	80	50	E	E	E	E
HI24	26	F	Colombia	5	26	95	40	100	E	E	E	E
Hi25	21	F	Colombia	11	10	100	10	90	E	E	E	E
HI26	26	F	Venezuela	12	14	98	0	90	B	B	E	E
HI29	19	F	Cuba	2	19	--	45	55	B	B	E	E
HI30	19	F	Venezuela	8	11	100	30	85	B	B	B	E
HI31	22	F	Mexico	6	22	80	80	100	E	E	E	E
Avg. / Summ.	21.3	21F 4 M	6 Cuba 5 Mexico 4 Venezuela 10 other	6.0	16.0	89.4	48.4	84.4	16 E 8 B 1 S	18 E 5 B 2 S	22 E 2 B 1 S	23 E 2 B 0 S

Table 1. Demographic and language background information for highly proficient bilingual group. Notes: AOLI = age of learning English intensively. E = English; S = Spanish; B=both English and Spanish rated equally. Origin = country of birth or country of birth of parents for participants born in the U.S.

Code	Age	Gender	Origin	AOLI	Time in U.S. (yrs)	% English used at:			Most comfortable language for:			
						Work	Home	Other	Speak	Listen	Read	Write
LO01	30	F	Panama	21	9	70	30	75	E	S	B	B
LO02	41	M	Peru	24	15	100	50	100	E	E	S	E
LO03	19	F	Colombia	15	4	--	2	80	S	S	S	S
LO04	23	M	Peru	17	5	50	99	50	S	S	S	S
LO06	19	F	Colombia	18	1	--	70	100	S	S	S	S
LO07	50	F	Colombia	45	5	98	100	100	S	S	S	S
LO09	21	F	Colombia	20	<1	--	20	40	S	S	E	S
LO10	28	F	Colombia	28	<1	--	0	100	S	S	S	S
LO11	22	F	Colombia	22	<1	100	60	40	S	S	S	S
LO12	35	F	Colombia	35	<1	--	15	90	S	S	S	S
LO13	19	F	Puerto Rico	16	3	70	10	30	S	S	S	S
LO14	25	M	Peru	22	3	100	0	100	S	S	S	S
LO15	22	F	Colombia	18	4	90	0	20	S	S	S	S
LO16	49	F	Colombia	46	10	100	0	20	S	S	S	S
LO17	26	M	Colombia	16	10	50	10	---	S	S	E	E
LO18	57	F	Colombia	30	27	85	0	100	S	S	S	S
LO19	22	F	Cuba	19	3	100	10	100	S	S	E	E
LO20	29	M	Colombia	23	6	100	0	80	S	E	E	E
Avg. / Summ	29.8	13 F 5 M	12 Colombia 3 Peru 3 Other	24.2	5.9	85.6	26.4	72.1	16 S 2 E	16 S 2 E	13 S 4 E 1 B	13 S 4 E 1 B

Table 2. Demographic and language background information for less proficient bilingual group. Notes: AOLI = age of learning English intensively. E = English; S = Spanish; B = both English and Spanish rated equally. Origin = country of birth or country of birth of parents for participants born in the U.S.

Materials and Instrumentation for Speaker Recording

Instrumentation. Audio recordings were made using the following equipment: 1) an Audio-Technica AT4033a Transformerless Capacitor Studio Microphone; 2) a Roland Digital Workstation model VS890EX; 3) an Applied Research and Technology Professional Tube Mic Pre-amplifier; 4) a Dell Opti-Plex CX110 personal computer; and

5) a sound-treated, single-wall booth. Speakers were seated individually in the sound-treated booth. The speech was processed through the pre-amplifier and digitized at a sample rate of 44.1 kHz by the digital workstation and transferred via digital input to the computer's M-Audio Audiophile 2496 sound card. CoolEdit (Johnston, 2000) was used to record the digital input to a sound file. Sound files for each speaker were saved for later stimulus preparation.

Speech materials. Each speaker was recorded speaking the six target vowels in /bVd/context in a carrier phrase ("Say ______ again"). The /bVd/ items were "bead, bid, bayed, bed, bad," and "bod." Speakers read a list with the target words aloud to the experimenter prior to recording to ensure they used the target pronunciations and to avoid reading errors. The carrier phrase was used to maintain the final /d/ release. Speakers read an 18-item list two times; each target word was represented on the list three times, for a total of six repetitions of each target word. Each speaker read the list from a sheet of paper presented on a reading stand at eye level.

Recording Procedure. The microphone was positioned at a 45-degree angle approximately ten centimeters from the speaker's mouth to avoid peak clipping and popping. Instructions were given to read the sentences avoiding flapping of the final /d/ release. To prevent glottal mode, speakers were instructed to take breaths in between phrases. Pausing to breathe between sentences also aided speakers in maintaining similar pitch and intensity levels across items. Before recording the stimulus items, a test recording was made to adjust the sound levels on the workstation and on the computer to rule out peak clipping.

Stimulus Creation Procedures. For each target word, one repetition was chosen for each speaker as the best exemplar. Next, the following six stimulus versions were created for each of 24 words (6 target items X 4 speakers):

1) whole word (WH), for which the word was simply extracted from the carrier phrase without modification;
2) original vowel (OV), for which the vowel was isolated from the word by removing CV and VC transitions without additional modification;
3) natural preserved (NP), for which the isolated vowel stimulus was resynthesized using Straight without modification of any parameters;
4) natural neutral (NN), for which the isolated vowel stimulus was resynthesized using Straight and its duration was adjusted within Straight to match the average vowel duration measured across the four speakers;
5) flat preserved (FP), for which pitch pulse replication and resynthesis were used to create stimuli with static formant frequency values across the entire duration of the vowel;
6) flat neutral (FN), for which the manipulations in 4) and 5), above, were combined to create neutral duration stimuli with static formant frequency values.

The process for creation of each of these versions of each word is detailed below.

Word Extraction. Each of the six tokens of each target word produced by the five speakers was isolated from the carrier phrase. First, separate sound files were created for each carrier phrase. Next, target words were isolated from the carrier phrase using CoolEdit2000 (Johnston, 2000) software by visually identifying and noting the time of the onset of the release of the initial /b/ and the end of the burst for the final /d/. Using CoolEdit2000 (Johnston, 2000), the surrounding carrier phrase (“say” and “again”) was silenced out leaving the target word and 10 ms of silence preceding the release of the initial /b/ and following the end of the burst for the final /d/. When pre-voicing was present, amplitude of pre-voicing of the initial /b/ was linearly ramped up (from 0% to 100% amplitude) for 3 ms following the transformation of the preceding 10 ms to silence. Similarly, energy following the release burst of the final /d/ was linearly ramped down (from 100% to 0% amplitude) for 3 ms following transformation of the following 10 ms to silence for each stimulus. For resynthesis, target words were down sampled in Praat to 11,050 Hz, the sample rate required for input to Straight. Using CoolEdit2000 (Johnston, 2000), all stimuli were amplitude adjusted so that the root-mean-square (RMS) amplitude would equal 25dB less than the maximum amplitude allowed without peak clipping (i.e., +/-32,768, or 90.3 dB) to avoid large differences in intensity across stimuli.

As stated above, one best exemplar was chosen from among the six repetitions of each target word for each speaker for use in the experiment. The best exemplar token for each target word was selected based on vocal quality, vowel quality, presence of the final /d/ release and similarity in voice quality to other target vowels in the set. Target words containing peak clipping, distortions, glottal fry, or nasalization were not considered for further use.

Using the software Praat 4.2 (Boersma & Weenink, 2003), the times of the onset of voicing for the vowel and the onset of closure for the /d/, were noted in an Excel file (Microsoft, 2000).

Target words produced by four of the speakers (two males, two females) were used to create stimuli for the main listening experiment. Target words produced by the remaining male speaker were used to create example and practice trial stimuli. Only the whole word (WH) and original vowel (OV) conditions were used in the example and practice trials, so only these conditions were created for the fifth talker.

The selected target words for each talker served as stimuli for the whole word (WH) condition, and additional modifications to these items were made to create the stimuli for the remaining conditions. Thus, 24 stimuli (six tokens X 4 talkers) were created for each condition.

Vowel Isolation. For the original vowel (OV) condition, the CV and VC transitions for the vowel were identified and isolated from the consonant context. The CV and VC transitions were identified as the first and last 40 ms of the vowel duration, following the onset of voicing for the vowel and preceding the onset of closure for the /d/, respectively. Thus, the time point corresponding to the vowel onset plus 40 ms of the vowel duration was identified in Praat 4.2 and all preceding portions of the signal were silenced. Similarly, the time point corresponding to the onset of closure for the /d/ minus 40 ms of the vowel duration was identified and all following portions of the signal were silenced. All but 10 ms of the resulting silence was deleted at beginning and end for each item. Next, the initial and final 3 ms of the signal were linearly ramped on and off, respectively, in CoolEdit2000 (Johnston, 2000) to avoid sudden changes in amplitude from the silencing of the CV and VC transitions resulting in audible clicks. Amplitude equalization to a RMS amplitude of negative 25dB, relative to the maximum, was again performed using CoolEdit2000 (Johnston, 2000).

Preparatory Acoustic Measurements for Formant Flattening. Formants 1 through 5 were identified using Linear Prediction Coding (LPC, Burg method), narrow band spectral slice, and wide band spectrogram. Using Praat 4.2 (Boersma & Weenink, 2003), a wide-band spectrogram (5 ms analysis window) was displayed for each isolated vowel token. The view range was set to 5,000 Hz with a dynamic range of 50.0 dB. Formant settings for a wide band spectrogram reading were set using a maximum formant frequency of 5512 Hz.

Automatic formant extraction procedures based on Linear Predictive Coding (LPC; Markel, 1972) were used to overlay formant tracks over the wide band spectrogram. By default, 10 poles were specified within the 5,000 Hz range analyzed. A 20-ms analysis window was used, updated at 5-ms intervals. Pre-emphasis was set for 50.0 Hz. The match between the tracks and the formants shown on the spectrogram was inspected and if formant tracks were not well aligned with the formants the number of poles used was increased or decreased as needed to achieve a better match. At each desired location, estimates of center frequency of F1-F5 were generated from the overlaid formant tracks using Praat query procedures.

To verify automatic formant estimates, narrow band spectral displays were generated at the desired time point and Monsen & Engebretson’s (1983) method of estimating location of spectral peaks was used. Spectrogram setting for the narrow band spectral slices was set with a window length of 0.029 s. In most cases, automatic estimates were used, but in cases of disagreement, the method described by Monsen & Engebretson (1983) for obtaining measurements from a narrow band spectral slice was used for formant frequency measurements.

In Monsen & Engebretson’s (1983) method for obtaining measurements from a narrow band spectral slice, the measurement is made by inferring the spectral peak from the intensity values of the harmonics in the vicinity of the formant in question. According to Monsen & Engebretson, “most researchers construct a type of pyramid or triangle above the group of harmonics which is thought to constitute a formant and then make a frequency measurement of the peak of that pyramid” (pg. 90). Using this method, three main scenarios of relative harmonic amplitudes in the vicinity of the formant are possible, resulting in different measurement decisions. Most simply, if there is one largest harmonic, with the two surrounding harmonics approximately equal in amplitude, then the formant frequency should be located at the frequency of the largest harmonic. Second, if the amplitude of two harmonics is approximately equal and greater than that of the surrounding harmonics, which are approximately equal in amplitude, then the formant frequency should be located between the frequencies of these two harmonics. The third scenario is the most complex. If the amplitudes of the two harmonics surrounding the largest harmonic are unequal, the formant frequency should be shifted away from the frequency of the largest and towards the frequency of the larger of the two surrounding harmonics; the shift should be greater when there is a greater difference between the amplitude of the two surrounding harmonics.

Measurements were made by two student raters. In cases where F1, F2, and F3 measurements of the two raters differed by more than 50 Hz, 150 Hz, or 250 Hz (Strange et al., 1998), respectively, a third rater (supervisor) also measured formant values. If the third rater’s measurements agreed within the specified amount with those of one of the two original rater's measurements then that rater's measurements were used. Otherwise, the measurements of the third rater were used.

Within the whole word file, vowel formant frequencies were measured at vowel onset plus 40 ms, vowel offset minus 40 ms, and at the vowel stable point, or the point at which absolute value of the slope of log(F1/F2) is minimized over several analysis frames (Hillenbrand et al., 1995; Hillenbrand & Nearey, 1999). The vowel stable point was computed by first transferring automatically generated formant tracks to an Excel file. Formant estimates were generated approximately every 4.7 ms and formant tracks were corrected if needed. Next, log(F1/F2) was computed for each frame. A seven-frame (approximately 33 ms) window was used to compute the slope of log(F1/F2) from the first frame to the last frame of the window; the slope was generated for each eligible window (seventh frame to the seventh from the end). As suggested by Hillenbrand (1995), the minimum slope that was not in the offglide section of the vowel was computed and compared with the plotted formant tracks to verify that the selected minimum coincided visually with a maximally stable portion of the vowel formants in the F1-F2 region.

Resynthesis Method for Natural Preserved Vowel Tokens. Using the extracted vowel, the natural preserved vowel tokens were generated by resynthesis, maintaining spectral information and duration cues. The Straight graphical user interface (GUI) (Kawahara et al., 1998) was initiated from within MatLab (The MathWorks, 2002). The stimulus file was accessed from within the GUI, the item was resynthesized by Straight with no changes specified and the synthesized item was saved to file. Sound quality and vowel duration were rechecked in Praat. Duration of silence prior to and after the vowel was checked as well as the RMS amplitude (-25dB). Ramping and amplitude equalization as described above were performed for vowels with temporal or amplitude deviations.

Resynthesis of Altered Vowel Conditions. The mean vowel duration for each word was used to resynthesize the natural neutral (NN) vowel condition. Mean vowel duration was computed as the average duration of the six isolated vowel stimuli across the four speakers used to create the experimental stimuli (i.e., average across 24 tokens). In the Straight-GUI, the file was loaded and the appropriate lengthening or shortening factor was specified to produce a vowel of the desired duration (approximately 193 ms). The output file was saved and the duration of each resynthesized vowel was rechecked in Praat. On some occasions, the Straight output contained small-amplitude voicing beyond the desired duration; in these cases, the stimulus was edited to match the target average duration. The Straight duration morphing routines automatically preserved the formant trajectories while either stretching or compressing the overall duration.

The two flat formant conditions (flat preserved [FP] and flat neutral [FN]) were both generated by replicating the measured stable portion of the vowel in Praat (Boersma & Weenink, 2003). The original vowel file was used to generate these two conditions. A total of approximately 30 ms around the measured stable (steady state) time was selected. To replicate the pitch periods, 5 or 6 pitch entire pitch periods were selected, whichever was closer to 30 ms. These pitch periods were copied and pasted into a file of 1 s of silence in Praat; this procedure was repeated carefully by zooming in and pasting the pitch pulses at a zero point until the total vowel duration matched the desired vowel duration for the original and neutral-duration vowels, respectively. Excess duration was deleted as needed until the duration of the target vowel version was matched. Vowel quality and waveform matching were carefully monitored to avoid unnatural jumps in amplitude at points where pitch pulses were posted. Visual inspection of spectrograms showed all formants to be flat. Below the formants of the vowel /æ/ after resynthesis and when both cues are neutralized (see Figures 1 and 2).

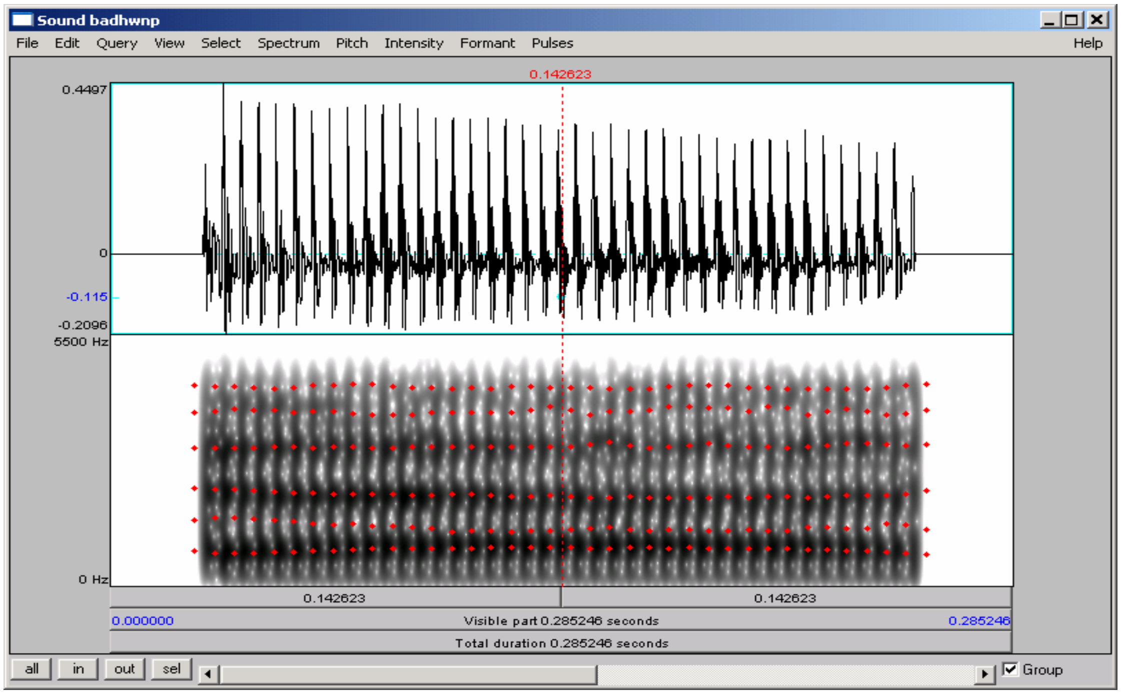


Figure 1. The natural vowel /æ/ when synthesized (NP) – spectral and duration cues were not modified.

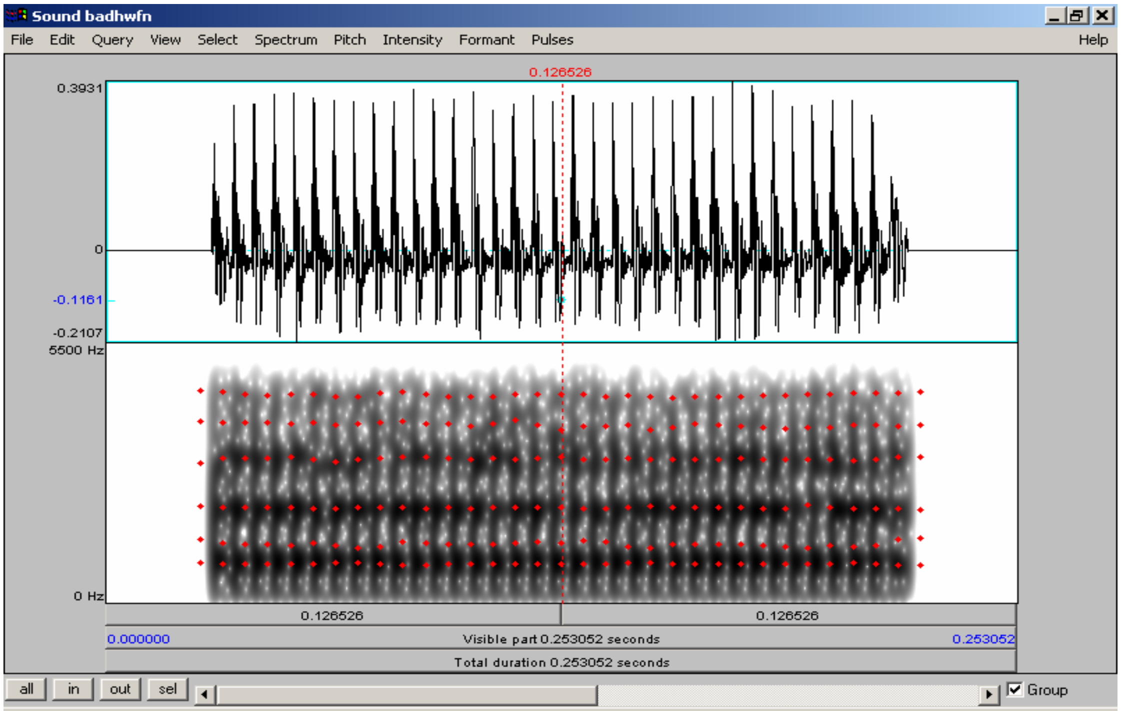


Figure 2. An example the vowel /æ/ with flattened formants and neutralized duration.

Both original-duration and duration-neutral flat formant vowels created by pitch pulse replication were loaded into the Straight-GUI (Kawahara et al., 1998) for resynthesis. Following synthesis, the duration of the vowels was checked in Praat assuring a pre- and post-vocalic silence of 0.03s (see Appendix C for detailed methodology).

Chapter 3

Procedure

Testing Procedure of Subjects

Calibration. The level for the presentation of stimuli was adjusted by playing a 10 second, 1000 Hz tone that was amplitude adjusted to –25dB from the maximum possible amplitude through the headphones to a sound level meter (Brüel Kjær Type 2235 Precision Sound Level Meter). Based on the readings of the sound level meter, the attenuation of the programmable attenuators (PA5) on the Tucker-Davis Technologies (TDT) ("TDT System III," 2001) Psychoacoustics System III (2001) hardware was adjusted until the measured intensity of the calibration tone was equal to 68 dB. These attenuation levels were then recorded and set within EcosWin (1999). Because the root mean square (RMS) amplitude level of the calibration tone was matched to that of the amplitude adjusted stimuli, this method assured 1) that the presentation level of the stimuli was approximately equal and 2) that the average presentation level for the stimuli was approximately 68 dB.

Testing. Each subject completed an informed consent form and language background form in a prior session to this experiment. A basic hearing screening on a Beltone AudioScout Audiometer calibrated to ANSI 1989 standards was administered to rule out hearing impairment (25 dB at 500, 1000, 2000, and 4000 Hz) in this prior session.

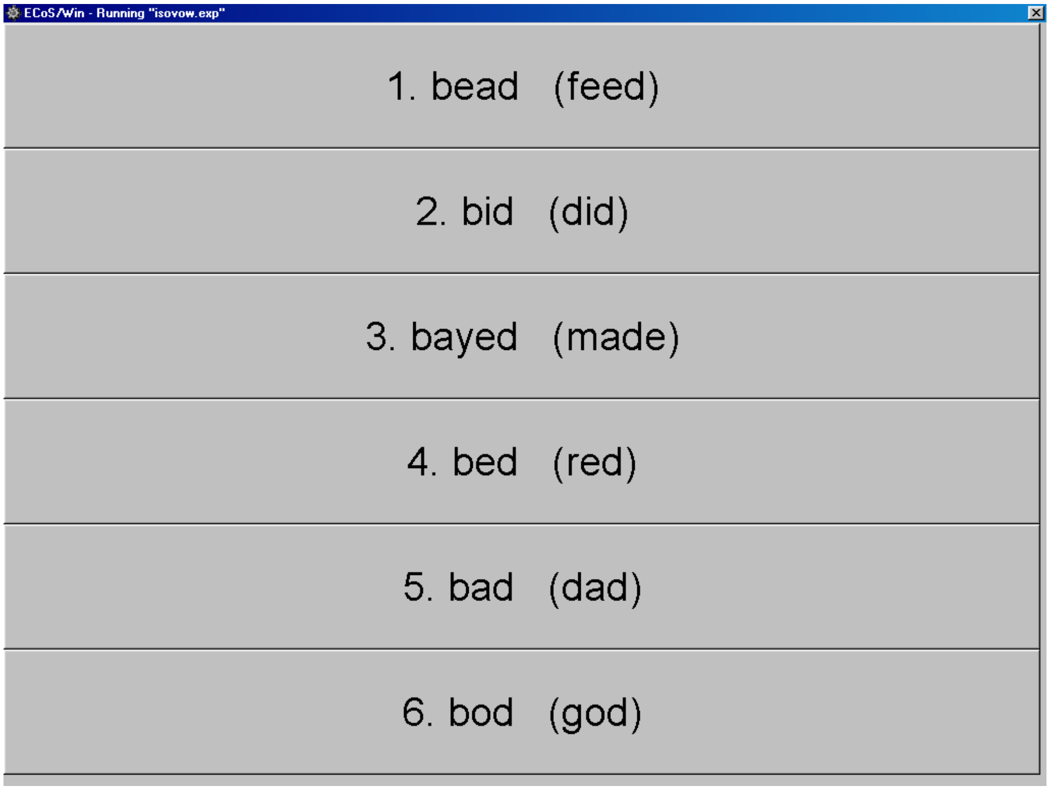
Seated in a quiet room, subjects were presented stimuli binaurally over headphones (Sennheiser HD 265). Prior to the main experiment, subjects completed several example and practice tasks. First were 12 whole-word example trials, followed by 12 isolated-vowel example trials. On the example trials, stimuli were presented to the listeners in a predetermined order and feedback was provided on each trial. For a correct response, the target lit up green, whereas an incorrect response was indicated by the color pink.

Next, each listener completed 3 blocks of 18 practice trials for the isolated-vowel stimuli. On the practice trials, stimuli were presented in random order and no feedback was provided. Stimuli for the example and practice tasks were the six whole word (WH) and six isolated vowel (OV) tokens created for the fifth (male) talker whose speech was not used for the main experimental task.

The main experiment tasks consisted of two 120-trial blocks for isolated vowels, followed by a 48-trial whole word task. Stimuli for the isolated vowel portion of the main experiment tasks were the stimuli created for the five isolated vowel conditions (OV, NP, NN, FP, and FN) for the four talkers and six target vowels (6 vowels X 5 conditions X 4 talkers = 120). The 120 stimuli were presented in random order in one trial block. Following a five-minute break, the same 120 stimuli were presented again in random order in a second trial block. This task lasted approximately 30 minutes.

Following another 5-minute break, the experiment was completed with a whole word task (1 block of 48 trials) lasting 10 minutes. The stimuli for this trial block were the whole word (WH) stimuli created for the four talkers for the six vowels, with each stimulus presented two times. Each subject was compensated $10 per hour of listening. The average duration for this experiment was less than one hour.

Trials. For each task (example, practice, and experimental), written directions on a computer screen and verbal instructions were provided. Subjects were instructed to respond as quickly and yet as accurately as possible. Upon presentation of the target word, the subject selected which word (or vowel) was heard from six alternatives presented on a computer screen (see Figure 3).



ECoS/Win - Running "isovow.exp"

1. bead (feed)

2. bid (did)

3. bayed (made)

4. bed (red)

5. bad (dad)

6. bod (god)

Figure 3. Stimulus screen.

Targets were presented in the individual /bVd/ words adjacent to a rhyming word (e.g. ‘beed’ and ‘feed’). Responses were indicated by the subjects using a mouse to select the target word they heard. Responses were recorded by EcosWin (1999) as correct or incorrect, generating an individual Excel file for each listening task. The files of the main task for both isolated vowels and whole words were saved to Excel files for data analysis.

Data Manipulation. Data from Ecos/Win (1999) (isolated vowel conditions and whole word) were loaded into Excel (Microsoft, 2000) for data analysis. A macro script was generated to sort the data separately for isolated vowel and whole word trials. The macro scripts were generated by two persons (supervisor and author) to avoid errors in the programming. The macro script for isolated vowel was used to sort correct and incorrect responses and sort these in alphabetical order (i.e., “bad, bayed, bead, bed, bid,” and “bod”) and by condition (original vowel, flat formant, etc.). Number correct for word and condition was automatically tallied and recorded for each subject. Percent-correct scores were computed from number correct for each condition. Within the macro, confusion matrices were generated for each listening condition (i.e., for the five isolated vowel conditions). For the confusion matrices, intended targets were depicted vertically (in rows) and actual responses horizontally (in columns). Raw scores were collected in a summary Excel file for all subjects for reliability purposes and statistical analyses. The macro script for the whole word was designed similarly; however, only one confusion matrix was generated for each listener (whole word).

Chapter 4

Results

Explanation of Percent Correct Analysis

Figure 4 below shows the mean percent correct for each listening condition by listening group: native (NA), highly-proficient (HP) bilinguals, and less-proficient (LP) bilinguals. The data will be discussed by conditions by group first, with reference to Figure 4. The six condition contrasts of interest will be discussed in turn: 1) whole word versus original vowel (WH-OV), representing the effect of vowel isolation; 2) original vowel versus natural preserved (OV-NP), representing the effect of resynthesis; 3) natural preserved versus natural neutral (NP-NN), representing the effect of duration neutralization alone; 4) natural preserved versus flat preserved (NP-FP), representing the effect of formant flattening alone; 5) flat preserved versus flat neutral (FP-FN), representing the effect of duration neutralization on already formant flattened stimuli; 6) natural preserved versus flat neutral (NP-FN), representing the effect of removing both formant dynamic and duration cues. Next, listening condition by vowel effects will be discussed. Following this section, confusion matrices showing each group's performance for each listening condition will be presented to elucidate patterns of error distribution for the listener groups.

Whole Word versus Original Vowel. Native listeners (NA) performed the best with a mean of 98.9% correct for the whole word condition. The highly-proficient listeners (HP) performed slightly lower by 2.8 percentage points (96.0% correct). The less-proficient bilinguals (LP) demonstrated difficulties identifying vowels at the word level. The LP mean was 81.9% correct or 16.6 percentage points lower than the NA listeners' performance. At this point, it may be suggested that even all listener groups perform best identifying whole words; however, less-proficient bilinguals appear somewhat challenged even at this whole word level.

Native listeners demonstrated a negligible decrease of 3.06 percentage points when the consonant context was removed (WH-OV condition). Highly proficient bilinguals demonstrated a 6.25% drop from WH to OV conditions, whereas less-proficient bilinguals' performance decreased by 8.91 percentage points from WH to OV conditions. The suggestion to be made from these results is that highly proficient listeners may rely more on consonant information than monolinguals, but less then the less-proficient bilinguals.

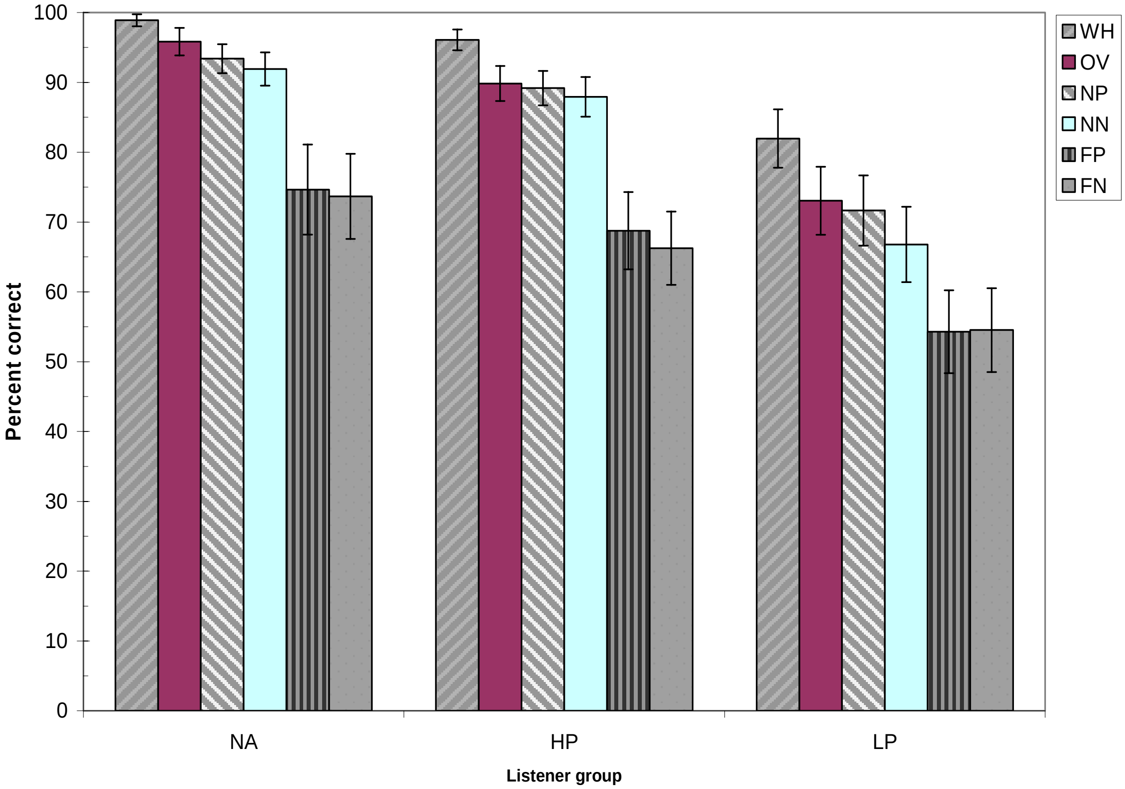


Figure 4. Mean percent-correct word-identification scores for each of the three speaker categories, monolingual native (NA), highly-proficiency bilingual (HP) and lower-proficiency bilingual (LP), at each of the six listening conditions. Error bars indicate two standard errors of the mean.

Original versus Natural Preserved Vowel.

The effect of resynthesis (OV-NP conditions) was noted to be small to negligible for all of the groups. Native listeners demonstrated the greatest decrease of the three groups (i.e., 2.45%) and HP listeners demonstrated the smallest decrease of 0.67%. However, native listeners still performed better than both bilingual groups.

Natural Preserved Vowel versus Natural Neutral Vowel .Neutralization of vowel duration information alone in the resynthesized stimuli (NP-NN) showed little decrease for native and highly proficient listeners (less than 2% for both). Less proficient listeners' performance was characterized by a 4.86 percent decrease from NP to NN conditions.

Natural Preserved versus Flat Preserved Vowel. Substantial decreases were noted for all of the listening groups between resynthesized vowels with all cues preserved (NP) and those with flattened formants (FP). Highly proficient bilinguals presented the largest decrease in percent correct identification performance (20.42%) from NP to FP conditions, whereas native and less proficient listeners averaged decreases of 18.75 and 17.36%, respectively.

The difference between dynamic preserved vowels and flattened preserved vowels represents the effect of removal of formant information alone. A similar pattern of decreases in performance was also seen between natural neutral vowels and flattened neutralized vowels (NA: 18.26%, HP: 21.67%, and LP: 12.27% for NN-FN).

Flat Preserved versus Flat Neutral. Minor drops were noted in percept correct identification performance when durations of flat preserved vowels were neutralized (FP-FN). The difference was greatest for the highly-proficient bilingual group (2.50%), with differences of less than one percent, in the positive and negative directions, respectively, for the native and less-proficient bilingual groups.

Overall Patterns. As shown in Figure 4, all three groups showed fairly similar patterns of performance across the listening conditions, meaning that performance for the whole word condition averages higher than that for the other listening conditions. The highly-proficient bilinguals performed better than the less-proficient bilingual listeners in all conditions. However, native listeners performed consistently higher than both highly- and less-proficient listeners. Additionally, all three listener groups performed best on whole word condition followed by original vowel. Noteworthy is the fact that resynthesis appears to have little effect on identification in regards to the original vowel. Only a slight decrease in performance is noted for all of the listening groups. Results for the duration neutralized vowel with preserved formant dynamics (NN) show a noticeable (about 5%) decrease in performance for the natural duration with preserved formant dynamic stimuli (NP) for less-proficient bilinguals only. It does not appear to substantially affect performance negatively to neutralize duration for the other two groups (see Figure 4). Similar for all three groups is the sizeable drop in performance when flattening the formants dynamic information is removed (NP-FP and NN-FN).

Nevertheless, some cross-group differences in performance across the different listening conditions are apparent. Most notably, the effect of vowel isolation (WH-OV) appears to be greater for the two bilingual groups than for the native listener groups. Second, the effect of duration neutralization alone appears to be greater for the LP bilingual group than for the HP bilingual and native listener groups.

Figure 4 also enables us to view the differences in performance and provides preliminary evidence that experience improves one’s perception (Flege, 1995, 1996; Flege et al., 1995; MacKay et al., 2001; Mayo et al., 1997) as can be seen in the highly proficient versus the less proficient bilinguals. However, this figure does not explain whether the removal of formant dynamic and duration cues affect different vowels differently.

Data from Figure 4 also suggest that native and highly-proficient bilingual listeners are able to identify Straight-synthesized isolated vowels quite well, even when duration information is removed (about 92 and 88% correct, respectively for the NP condition) and reasonably well (about 74 and 66 % correct, respectively) even when both formant dynamic and duration cues are removed. Performance is more than 20% lower for the LP group in the NP condition, however. Although the mean percent correct appears to give us an impression of cues used by the different listener groups, it needs to be stressed that data for all listening conditions were averaged across vowels. Consequently, it is necessary to analyze how performance of vowels by listening condition by group was distributed. Confusion matrices will also be used to clarify similar and different patterns between groups.

Analysis of Vowels by Condition

As stated previously, Figure 4 only provides information averaged across vowels. We cannot from these average patterns determine how different vowels were affected by changes in listening condition. Figure 5 depicts percent correct identification performance by condition and vowel, averaged across the listener groups. Of interest are the different patterns across vowels that indicate whether some vowels are affected more by some conditions than others. As can be seen, performance for /i/ stays relatively stable even after formant dynamic information is removed. This pattern may be attributable to its stable vowel drift (Hillenbrand et. al, 1995; Hillenbrand & Nearey, 1999). The lax vowel /ɪ/ is characterized by greater decreases for removing duration information than for removing formant dynamic information, in that the decreases for NP-NN and FP-FN are greater than that for NP-FP and NN-FN. As expected, the diphthong /e^{ɪ}/ stands out with the largest decrease in performance for all the vowels when formant dynamic information is removed (NP-FP). Before removing this information, /e^{ɪ}/ is the best perceived vowel (about 95% correct for NP), whereas after removing the formant dynamic information it is perceived least accurately of all (about 9% correct for FP). The vowel /ɛ/ is of interest because duration neutralization appears to cause reduced accuracy when formant dynamic information is also removed (approximately 7% decrease for FP-FN), but not when formant dynamic information is present (approximately 1% increase for NP-NN). Similarly, removing formant dynamic information has little effect when duration information is preserved (approximately 1% decreased for NP-FP). Thus, listeners appear to identify /ɛ/ similarly when one cue (either formant dynamic information or duration information) is removed, but perform less well when both cues are removed. The greatest decrease in performance for /ɛ/, however, is seen for the effect of vowel isolation (WH-OV), suggesting that listeners may rely more heavily on consonant context for this vowel; a similar but smaller effect of vowel isolation is shown for /ɑ/. The effect of formant flattening appears to be moderately strong for both /æ/ and /ɑ/. Removing vowel duration information appears to have little effect for /ɑ/, whereas /æ/ appears to show some effect of duration neutralization alone (about a 7% decrease for NP-NN). While these patterns highlight potential differences in the effects of the listening conditions across listener groups and vowels, they are only trends in the data. Statistical analyses are needed to confirm the significance of these trends. In the section below, parametric statistical analyses and post-hoc test results will be described for main effects and the two combinations of variables discussed thus far (listening condition by group and listening condition by vowel). The three-way interaction will also be examined and the results of post-hoc tests for the three-way interaction will be presented.. Following this, confusion matrices will be presented to explore any differences in effects within the three-way interaction (i.e., differences in the effects of listening condition across the vowels for the different listener groups).

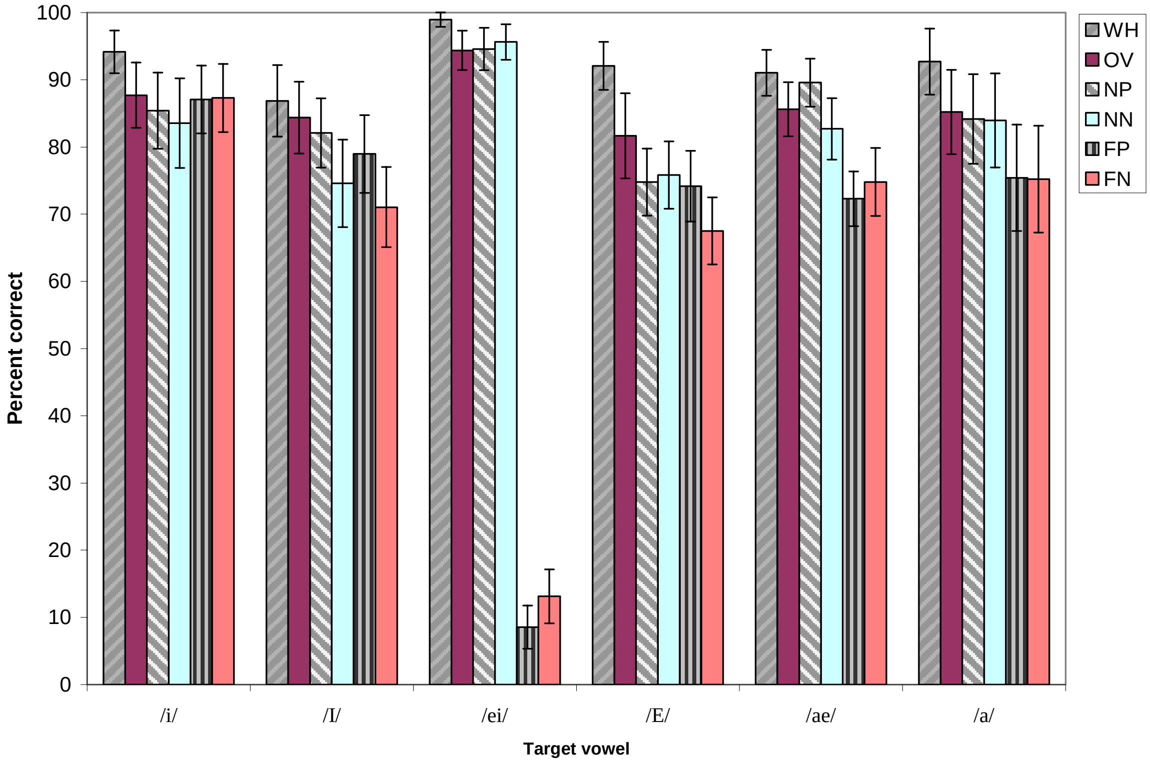


Figure 5. Mean percent-correct word-identification scores for each of the six target vowels and each of the six listening conditions. Error bars indicate two standard errors of the mean.

Statistical Analyses

Before performing parametric data analysis, number correct data were converted to rationalized arcsine transform units (RAUs) (Studebaker, 1985). According to

Studebaker (1985) converting raw data into RAUs is applied to proportional data to eliminate ceiling effects. In doing so, negative skewness of the data and correlation of the variance with the means is evaded (Studebaker, 1985).

Table 3. Statistical analysis table. The table depicts main effects, two-way interactions, and three-way interaction.

Effects	F value	Degrees of freedom	p value
Main			
Listener group	44.41	2, 57	<.001
Listening condition	247.55	5, 285	< .001
Vowel	16.29	5, 285	< .001
Two-way interactions			
Listener group by listening condition	2.38	10, 285	.01
Vowel by listening condition	90.43	25, 1425	< .001
Listener group by vowel	3.41	10, 285	< .001
Three-way interaction	1.27	50, 1425	.100

As can be seen from Table 3, all main effects and all of the two-way interactions were significant; most were highly significant (p<.001), with the exception of the listener group by listening condition interaction (p=.01).

Main Effects. For the between-subjects main effect of listener group, a Tukey HSD post-hoc test was performed. All listener groups were found to differ significantly from one another in vowel identification performance. For native vs. highly-proficiency bilingual, p = .045; for native vs. low-proficiency and high-proficiency vs. low-proficiency bilingual, p < .001.

For the main effect of listening condition, six comparisons probing six effects of interest were made: 1) effect of vowel isolation (WH-OV); 2) effect of Straight resynthesis (OV-NP); 3) effect of duration neutralization alone (NP-NN); 4) effect of formant flattening alone (NP-FP); 5) effect of duration neutralization for stimuli with already flattened formants (FP-FN); and 6) effect of removing both duration and formant dynamic cues (NP-FN). These effects were explored as simple main effects of condition. Bonferroni adjustment to the significance level of each comparison was made by dividing the alpha level obtained by six, for the six comparisons made (note that 15 comparisons were possible, but that significance level was not adjusted for this number because only six comparisons were explored). The effect of vowel isolation (WH-OV) was found to be significant (p<.006), as were the effects of formant flattening alone (NP-FP, p<.006) and removing both duration and formant dynamic cues (NP-FN, p<.006). The effect of Straight synthesis approached significance (OV-NP, p=.06). The effect of duration neutralization on flattened formants (FP-FN) did not approach significance. The main effect of vowel identity was significant but was not considered of primary interest in relation to the research questions and will not be discussed further. Similarly, the significant two-way interaction between group and vowel will not be discussed.

Two-way Interaction of Group and Listening Condition. For the significant two-way interaction between listener group and listening condition described in the text relating to Figure 4, the six contrasts of interest were compared at each level of the listener group variable. Bonferroni adjustment was again performed for the six comparisons made at each level of the group variable by multiplying the p value by six. For the native (monolingual) listener group, the effect of formant flattening alone was found to be significant (NP-FP, p<.006), as was the effect of removing both formant dynamic and duration cues (NP-FN, p<.006). The effect of Straight resynthesis approached significance (OV-NP, p=.084), while the effect of the other two contrasts (WH-OV and NP-NN) did not approach significance.

For both the highly-proficient and less-proficient bilingual listener groups, the following contrasts were found to be significant, all at p<.006: 1) the effect of vowel isolation (WH-OV); 2) the effect of formant flattening alone (NP-FP); and 3) the effect of removing both formant dynamic and duration cues (NP-FN). The remaining effects did not approach significance for the highly-proficient listener group; for the less-proficient bilingual listener group, the effect of duration neutralization alone (NP-NN) approached significance (p=.12). Considering the effects across groups, the post-hoc analyses confirmed some of the trends noted in patterns of performance across groups and listening conditions, but not others. That is, the effects of formant flattening alone (NP-FP) and that of removing both cues (NP-FN) were found to be large and highly significant for all three listener groups. Furthermore, the effect of vowel isolation (WH-OV) was found to be significant for the two bilingual groups but not for the monolingual group, confirming the relatively larger effect of vowel isolation for the two bilingual groups. The data lend only weak support to the increased effect of duration neutralization alone for the lower-proficiency bilingual group compared to the other groups, in that the effect approached significance for the lower-proficiency group but did not for the other two groups. Similarly, the post-hoc tests weakly support the greater effect of Straight synthesis (OV-NP) for the native group compared to the other groups, in that the effect approached significance for the native group, but did not for the other two groups.

Two-way Interaction of Vowel and Listening Condition. For the significant two-way interaction between vowel and listening condition described in the text relating to

Figure 5, the six contrasts of interest were compared at each level of the vowel variable. Bonferroni adjustment was again performed for the six comparisons made at each level of the vowel variable. The effects of vowel isolation and Straight resynthesis will be considered first; then the effects of the duration neutralization and formant flattening manipulations will be considered.

The effect of vowel isolation was significant for three of the six vowels (/e$^{\text{ɪ}}$, ɛ, ɑ/) at p=.006 or less for all. The effect of vowel isolation approached significance for two of the remaining vowels (/æ/, p=.108; /ɪ/, p=.06). These effects confirm the more substantial effect of vowel isolation for the vowels (/e$^{\text{ɪ}}$, ɛ, ɑ/) noted in the text relating to Figure 4 and suggest that the main effect of vowel isolation was primarily due to the effects for these three vowels.

The effect of Straight resynthesis (OV-NP) was significant only for the target vowel /ɛ/ (p=.006). The effect did not approach significance for any of the other vowels. This restriction of the effect to a single vowel is surprising, considering that the effect neared significance in the main effect of listening condition. However, it may help to explain why the effect did not reach significance for any single listener group, although it neared significance for the native listener group.

For /i/, the pattern of relatively little effect of listening condition was largely confirmed, none of the effects of cue manipulation approached significance. For the vowel /ɪ/, however, three effects reached significance: 1) the effect of duration neutralization alone (NP-NN, p=.03); 2) the effect of duration neutralization for already flattened stimuli (FP-FN, p=.006); and 3) the effect of removing both duration and formant dynamic cues (NP-FN, p<.006). Thus, these results confirm the patterns noted relating to Figure 4 of greater effects of duration neutralization than of formant flattening for this vowel.

For /e^{I}/, the effects of formant flattening alone (NP-FP) and the effect of removing both formant dynamic and duration cues (NP-FN) were both large and significant (p<.006). The effect of duration neutralization of already flattened stimuli (FP-FN) was also significant (p=.03) but was not in the expected direction, in that percent correct performance was higher for the FN than for the FP condition.

For the target vowel /ɛ/, the effect of removing both duration and formant dynamic cues (NP-FN) was significant (p=.024), as was the effect of duration neutralization of already flattened stimuli (p=.006). These results confirm the pattern noted in relation to Figure 4 of a duration neutralization effect mainly for the already flattened stimuli and little effect of formant flattening alone. For the target vowel /æ/, on the other hand, both duration neutralization and formant flattening effects were found to be significant (p=.03 for NP-NN and p<.006 for NP-FP). Similarly, the effect of removing both duration and formant dynamic cues was also significant (NP-FN, p<.006). Finally, for the vowel /ɑ/ the effect of formant flattening alone approached significance (NP-FP, p=.06), while the effect of removing both duration and formant dynamic cues just reached significance (NP-FN, p=.042).

Three-way Interaction. Although the three-way interaction was not significant, it did approach significance (p = .100). Furthermore, only six of the possible 15 comparisons in the listening condition factor were considered in the two-way interactions. Thus, it was determined that because Bonferroni adjustment in the post-hoc comparisons would be for six rather than 15 comparisons at each level of group and vowel, it was possible that significant three-way interaction effects might be found (i.e., significant differences in listening condition effects across the different vowels and listener groups). For this reason, the three-way interaction was explored in post-hoc analyses and significant three-way interaction effects were indeed found. These effects are presented in Table 4 and will be described below, with each of the six effects of interest considered in turn.

Table 4. Significant effects for six contrasts of interest in the three-way interaction. Each cell lists the vowels for which the contrast was found to be significant and the size (in RAUs) and direction of effect, with the associated p value in parentheses.

	Listener group					
Listening condition effect	Native		HP bilingual		LP bilingual	
Vowel isolation (WH-OV)			/ɑ/	11.61 (.018)	/e$^{\mathrm{I}}$/	10.20 (.030)
					/ɛ/	23.83 (<.006)
					/ɑ/	13.84 (.018)
Synthesis (OV-NP)	/ɛ/	17.24 (.012)				
Duration neutralization alone (NP-NN)					/ɪ/	15.62 (.018)
Formant flattening alone (NP-FP)	/e$^{\mathrm{I}}$/	94.83 (<.006)	/e$^{\mathrm{I}}$/	96.86 (<.006)	/e$^{\mathrm{I}}$/	88.07 (<.006)
	/æ/	22.36 (<.006)	/æ/	23.48 (<.006)	/æ/	16.17 (<.006)
			/ɑ/	15.05 (.018)		
Duration neutralization of flattened stimuli (FP-FN)			/ɪ/	12.93 (.006)		
Both formant flattening & duration neutralization (NP-FN)			/ɪ/	15.46 (<.006)		
	/e$^{\mathrm{I}}$/	85.08 (<.006)	/e$^{\mathrm{I}}$/	93.50 (<.006)	/e$^{\mathrm{I}}$/	82.98 (<.006)
	/ɛ/	11.71 (.048)	/ɛ/	10.17 (.036)		
	/æ/	15.09 (.018)	/æ/	21.43 (<.006)	/æ/	15.29 (.012)
			/ɑ/	17.38 (<.006)		

As can be seen from Table 4, different target words showed significant effects of vowel isolation (WH-OV) for the different listener groups. No target words showed a significant effect for the native (monolingual) listener group and only one vowel (/ɑ/) showed a significant effect of vowel isolation for the HP bilingual group. For the LP bilingual group, on the other hand, three vowels (/e^{I}, ɛ, ɑ/) showed a significant effect of vowel isolation. These effects are consistent with the evidence of greater vowel isolation effects for the HP and LP bilingual groups discussed with regard to the two-way interaction between group and listening condition. These data also demonstrate that the large effect of vowel isolation for the vowel /ɛ/ found in the two-way interaction of vowel by listening condition was most influenced by data for the LP bilingual group.

The effect of Straight resynthesis was significant only for the target vowel /ɛ/ for the native listener group. These data demonstrate that the nearly significant effect of Straight synthesis for both the main effect of listening condition and for the native group in the group by listening condition interaction was most strongly influenced by performance of the native listener group for the target vowel /ɛ/. This pattern is consistent with the sole significant effect of Straight resynthesis (OV-NP) for the vowel /ɛ/ in the vowel by listening condition interaction.

The effect of duration neutralization alone (NP-NN) was significant only for the vowel /ɪ/ for the LP bilingual group. The size of the effect was relatively large for this target vowel however (15.62 RAUs). These data demonstrate that the significant effect of duration neutralization alone in the vowel by listening condition interaction was largely due to performance for the LP bilingual group, also explaining the near-significant effect of duration neutralization alone for the LP group in the group by listening condition interaction. These data suggest that the LP group make more use of the duration cue in vowel identification than the other two groups (at least for the target vowel /ɪ/).

The effect of formant flattening alone (NP-FP) was significant for two vowels (/e^{I}, æ/) for the native group, three vowels for the HP bilingual group (/e^{I}, æ, ɑ/), and two vowels for the LP bilingual group (/e^{I}, æ/). The effect was, of course, largest for the target vowel /e^{I}/ (ranging from 88 to 97 RAUs), however substantial effects were obtained for the vowels /ɑ/ and /æ/, ranging from 16 to 23 RAUs. These data demonstrate that although the formant flattening effect was significant for all three groups in the group by listening condition interaction and was dominated by the effects for the vowel /e^{I}/, the effects did indeed vary across vowels for the different groups. It should also be noted that the size of the formant flattening effect was largest for the HP bilingual group for the two vowels (/e^{I}/ and /æ/) that showed a significant effect for all three groups. These data suggest that the HP bilingual group may rely more heavily on formant dynamic information than either the native or LP bilingual groups.

The effect of duration neutralization of already formant flattened stimuli (FP-FN) was only significant for the vowel /ɪ/ for the HP bilingual group. These data confirm the trend noted in the discussion of Figure 4, suggesting a larger effect of duration neutralization of already formant flattened stimuli for the HP group only. These data suggest that although both the HP bilingual and native listeners' performance appears to be robust to formant flattening effects for /ɪ/, the HP bilingual listeners may rely on duration information more heavily than native listeners for this vowel when formant dynamic information is removed. That is, it may be that the native listeners are able to use only static target formant frequencies for accurate vowel identification, while the HP bilinguals appear to perform similarly when either duration or formant dynamic information is removed, but cannot achieve peak performance when only static formant frequencies are available.

Finally, the effect of removing both formant dynamic and duration cues was significant for three vowel for the native listener group (/e^{I}, ε, æ/), five vowels for the HP bilingual listener group (all except /i/), and two vowels for the LP bilingual group (/e^{I}, æ/). For /e^{I}/ and /æ/ the effect was largest for the HP bilingual group. Again, these data suggest that the HP bilingual group is less able than the native group to reliably identify target English vowels based only on static vowel formant targets. The LP group shows fewer significant effects when both cues are removed (/e^{I}/ and /æ/), but show lower performance overall (about 15% lower on average than the HP group).

Confusion Matrices.

To examine differences in the use of cues and their order across groups and target vowels, it is useful to perform an analysis that provides not only information for each vowel by condition and listener group, but also information as to the identity of the misperceptions that occur. To illustrate listener-group specific performance across conditions and vowels, confusion matrices were created to show the average percent distribution of the intended vowel and the actual response for each target word. Tables 5-10 depict percent correct for the listening conditions for the native, highly-proficient, and less-proficient bilingual listeners. The target words are arrayed vertically in rows whereas the responses are listed horizontally in columns. Results for the native (NA), highly-proficient (HP), and less proficient bilingual listeners (LP) are arranged in separate rows within each table, in the mentioned order, to enable comparison within and between groups. The orange colored boxes signify the intended target. Boldfaced black numbers signify the highest score for each group out of all the words. Red boldfaced numbers indicate the lowest score of each group (see tables). Describing the data, this author established the following criterion: differences less than 5 percent points would not be discussed. Further, “confusion event” or “confusion” is defined as the vowel that is incorrectly identified (incorrect response) for a given target. Thus, if /i/ is the target five potential “confusions” are possible (i.e., /ɪ, e^{ɪ}, ɛ, æ/ and /ɑ/). For each listener group a total of 30 different confusion events is possible (five for each of the six target vowels).

Whole Word Confusion Matrix. As illustrated by Table 5, native listeners demonstrated perfect identification of the vowels /i/ and /e^{ɪ}/. Percent scores for the remaining vowels, with the exception of /æ/, indicate near to perfect identification score (error percent < 1%). The percent correct score for the vowel /æ/ reads 95.59%. Further analysis of /æ/ shows that native listeners identified target /e^{ɪ}/ as /æ/ 3.68% of the time.

Table 5. Confusion matrix for whole word condition – depicting performance of monolingual, highly-proficient bilingual, and less-proficient bilingual listeners.

		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		**100**	0	0	0	0	0
HP	Bead	97	3	0	0	0	0
LP		84.72	14.58	0	0.69	0	0
NA		0.74	99.26	0	0	0	0
HP	Bid	2	96.5	1	0	0.5	0
LP		24.31	**61.81**	0	13.89	0	0
NA		0	0	**100**	0	0	0
HP	Bayed	0.5	0	**99**	0.5	0	0
LP		1.39	0.69	**97.92**	0	0	0
NA		0	0	0.74	99.26	0	0
HP	Bed	0	4.5	0	95	0.5	0
LP		2.78	4.86	0.69	81.25	6.94	3.47
NA		0	0	0	3.68	**95.59**	0.74
HP	Bad	0	0	0.5	9.5	**90**	0
LP		0	0	0	3.47	88.19	8.33
NA		0	0	0	0	0.74	99.26
HP	Bod	0	0	0	0	1	**99**
LP		0	0	0	0	22.22	77.78

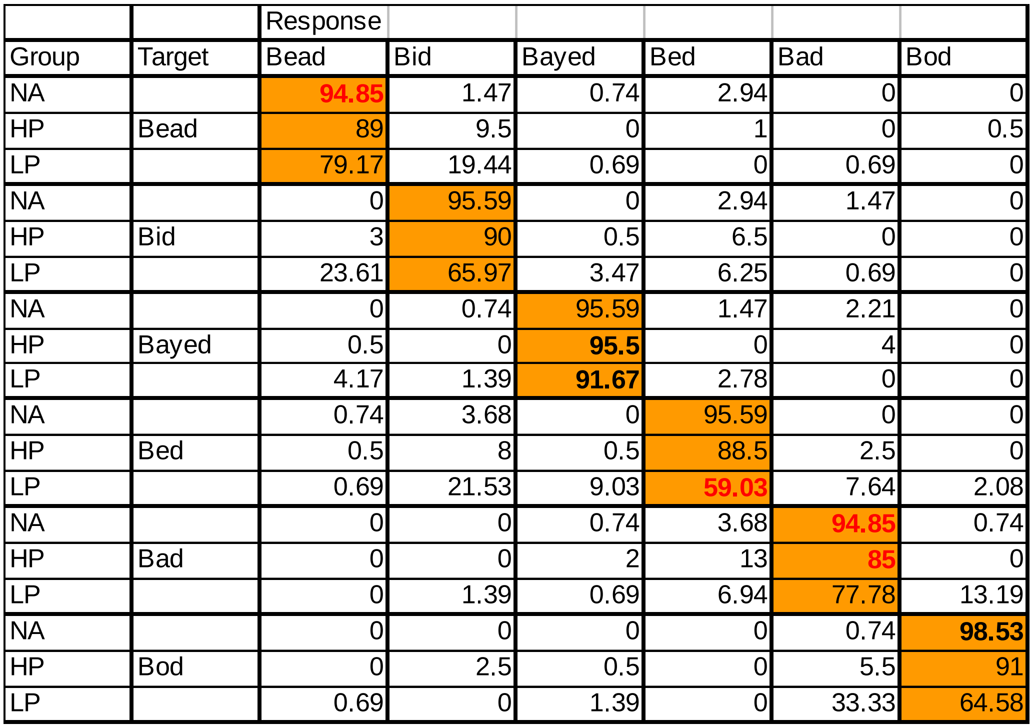
Highly-proficient listeners’ mean performance showed no perfect score for any of the vowels. The vowels /e^{ɪ}/ and /ɑ/ were noted to have the best percent correct scores (99% for both). A decrease of five percent points was noted for /ɛ/, with two confusions: /ɪ/ at 4.5% and /æ/ at 0.5%. A large decrease of 10 percent points was noted for the vowel /æ/ and two confusions were noted. HP listeners showed the greatest confusion for /ɛ/ (9.5%).

The response distribution of the less proficient (LP) bilingual listeners was as follows. Overall performance of the LP was noted to be lower than for the native and HP groups for all vowels. The range of percent correct for the vowels varied from 61.92% (/bɪd/) to 97.92% (/be^{ɪ}d/).

All three groups demonstrated strength identifying /e$^{\mathrm{I}}$/. NA and HP both performed the most poorly for target /æ/, with two confusions. Their greatest confusion for /æ/ was noted to be /e$^{\mathrm{I}}$/ (3.68% and 9.5% for the NA and HP groups, respectively). However, their second confusion differed, but both were less than one percent of responses (NA - /e$^{\mathrm{I}}$/ and HP - /ɑ/). The LP group demonstrated lower performance for /æ/ and as for NA and HP two confusions were noted. However, the LP listeners demonstrated greater identified target /æ/ as /ɑ/ (8.33%) most often, with target /æ/ identified as /ɛ/ less often (3.47%). LP listeners showed a generally consistent confusion pattern, meaning errors were noted for only one or two other vowels in the whole word condition. Noteworthy is the LP group's confusion pattern for target /ɪ/: /i/ was identified 24.31% for /ɪ/ and /ɛ/ 13.89%. Although all groups demonstrated the same first confusion for target /ɪ/, with much lower error percentages for the NA and HP groups than for the LP group, the HP group demonstrated an additional confusion that varied from LP's responses (i.e., HP's second confusion was /e$^{\mathrm{I}}$/ and third confusion was /æ/). LP listeners demonstrated a notably inconsistent confusion pattern for the target vowel /ɛ/; five confusions were noted of which /æ/ had the highest confusion percentage (6.94%) and /ɪ/ the second highest (4.86%).

Original Vowel Confusion Matrix. None of the listener groups demonstrated perfect identification for any of the vowels when consonant context information was removed (see Table 6). The vowel /ɑ/ was identified best by native listeners (98.53%). Highly and less proficient bilinguals' vowel performance was noted to be best for the vowel /e^{I}/ (HP: 95.5% and LP: 91.67%). The HP and LP groups achieved most poorly on the vowels /ɛ/ and /æ/, respectively, whereas native listeners demonstrated showed lowest performance for two target vowels (/i/ and /æ/ at 94.85%). Although HP and LP listener groups showed similar patterns for best vowel, their numbers and identity of confusions did not match up. An example is /ɛ/ where LP displayed five confusions, whereas HP only had four.

Table 6. Confusion matrix for the original vowel (OV) condition – depicting performance of the monolingual, highly proficient bilingual, and less proficient bilingual listeners.



		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		**94.85**	1.47	0.74	2.94	0	0
HP	Bead	89	9.5	0	1	0	0.5
LP		79.17	19.44	0.69	0	0.69	0
NA		0	95.59	0	2.94	1.47	0
HP	Bid	3	90	0.5	6.5	0	0
LP		23.61	65.97	3.47	6.25	0.69	0
NA		0	0.74	95.59	1.47	2.21	0
HP	Bayed	0.5	0	**95.5**	0	4	0
LP		4.17	1.39	**91.67**	2.78	0	0
NA		0.74	3.68	0	95.59	0	0
HP	Bed	0.5	8	0.5	88.5	2.5	0
LP		0.69	21.53	9.03	**59.03**	7.64	2.08
NA		0	0	0.74	3.68	**94.85**	0.74
HP	Bad	0	0	2	13	**85**	0
LP		0	1.39	0.69	6.94	77.78	13.19
NA		0	0	0	0	0.74	**98.53**
HP	Bod	0	2.5	0.5	0	5.5	91
LP		0.69	0	1.39	0	33.33	64.58

Natural Preserved Vowel Confusion Matrix. The percent correct for the targets were noted to be highest for the native listener group for all vowels except /e^{I}/ (range: 82.35% to 97.79%; see Table 7); for /e^{I}/, the performance of the HP bilingual group was slightly higher than that of the NA group (97% vs. 96.32%, respectively). Performance was lowest for the LP bilingual group for all vowels. Noteworthy for the NP condition is that all three listener groups performed most poorly for the target vowel /ɛ/ (NA: 82.35%, HP: 82%, and LP: 57.64%). The groups' largest confusion for target /ɛ/ was noted to be /ɪ/. Furthermore, the HP and LP groups both revealed five confusions rather than only two (NA) for target /ɛ/. The range for HP bilingual listeners was noted as 82% (/ɛ/) to 97% correct (/e^{I}/) while performance for the LP bilingual listeners ranged from 57.64 (/ɛ/) to 89.58% correct (/e^{I}/).

Table 7. Confusion matrix for the natural preserved vowel (NP) condition – depicting performance of the monolingual, highly proficient bilingual, and less proficient bilingual listeners.

		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		93.38	5.88	0	0.74	0	0
HP	Bead	86.5	11	0.5	2	0	0
LP		76.39	20.14	1.39	1.39	0.69	0
NA		0	94.12	0	5.88	0	0
HP	Bid	1	85	1	13	0	0
LP		18.75	66.67	2.78	11.11	0.69	0
NA		0	0	96.32	0	3.68	0
HP	Bayed	0	0.5	**97**	0.5	2	0
LP		1.39	2.08	**89.58**	6.25	0	0.69
NA		0	14.71	0	**82.35**	2.94	0
HP	Bed	1	13.5	0.5	**82**	2.5	0.5
LP		1.39	23.61	4.17	**57.64**	10.42	2.78
NA		0	0	1.47	2.21	96.32	0
HP	Bad	0	0.5	0	7	92.5	0
LP		0	0.69	1.39	4.86	79.17	13.89
NA		0	0	0	0.74	1.47	**97.79**
HP	Bod	0	0.5	1	0	6.5	92
LP		0	0	0	0	39.58	60.42

Natural Neutral Vowel Confusion Matrix. Patterns for the natural neutral vowel condition are characterized by native speakers performing best for all six vowels except /e^{I}/ (see Table 8). As for the NP condition, the performance of the HP bilingual group for target /e^{I}/ was slightly higher than that of the NA group (99.5% vs. 97.06%, respectively). Native speakers' best vowel was /ɑ/ (98.53%) followed by /e^{I}/ (97.06%). Native listeners demonstrated most difficulties identifying the target vowel /ɛ/ (86.76%), while the HP bilinguals experienced the most difficulty correctly identifying target /ɛ/ and /ɪ/ (82.5%). Further, the vowel /ɪ/ was most often confused for the target /ɛ/ by both native and HP bilingual listeners (NA: 10.29% and HP: 12%). Less proficient bilingual listeners exhibited an even lower percent correct (56.25%) than native and HP bilingual listeners for target /ɛ/, but the target vowel showing the lowest performance for this group /ɪ/ (50.69%). Nevertheless, LP bilingual listeners also demonstrated the highest confusion of the target /ɛ/ with /ɪ/ (19.44%). Highly and less proficient bilinguals performed best on the diphthongized target vowel /e^{I}/, (HP: 99.5% and LP: 88.89%), however both groups displayed different numbers and patterns of confusion: HP heard only /æ/ for target /e^{I}/, whereas LP demonstrated four confusions for target /e^{I}/, of which /ɛ/ was the most frequent (6.94%). The LP bilingual listeners exhibited greatest difficulties identifying /ɪ/ (50.69%). Moreover, four confusions were made for this target vowel, of which /i/ (21.53%) and /ɛ/ (22.92%) showed most errors.

Table 8. Confusion matrix for the natural neutral vowel condition – depicting performance of the monolingual, highly proficient bilingual, and less proficient bilingual listeners.

		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		92.65	3.68	0.74	2.94	0	0
HP	Bead	86.5	11	0	2.5	0	0
LP		70.83	26.39	2.08	0.69	0	0
NA		0.74	88.24	0.74	9.56	0.74	0
HP	Bid	1.5	**82.5**	0.5	15.5	0	0
LP		21.53	**50.69**	4.17	22.92	0	0.69
NA		0	0	97.06	0	2.94	0
HP	Bayed	0	0	**99.5**	0	0.5	0
LP		1.39	2.08	**88.89**	6.94	0.69	0
NA		0	10.29	0	**86.76**	2.94	0
HP	Bed	1	12	1.5	**82.5**	3	0
LP		1.39	19.44	6.94	56.25	11.81	4.17
NA		1.47	0	0	10.29	88.24	0
HP	Bad	0	0	0	14	86	0
LP		0.69	2.78	2.08	8.33	72.92	13.19
NA		0	0.74	0	0.74	0	**98.53**
HP	Bod	0	1	0.5	0	8	90.5
LP		0	0	0.69	0	38.19	61.11

Flat Preserved and Flat Neutral Vowel Confusion Matrix. The flat preserved vowel confusion matrix depicts a more uniform pattern of identification for all three groups (see Table 9). For the FP condition, native listeners performed better than the other two listener groups for all of the target vowels. Highly proficient listeners also performed better than the less-proficient listeners, and nearly as well as the native listeners for target vowels /e$^{\text{ɪ}}$/ and /ɛ/. The vowel /i/ was identified best by all three groups (NA: 96.32%, HP: 90.5%, LP: 73.61%). Also, the native and HP bilingual groups showed the next highest performance for target /ɪ/. The vowel /e$^{\text{ɪ}}$/ was perceived least accurately for all of the groups (NA: 9.56%, HP: 8%, LP: 8.33%). Other patterns included similar confusion responses; that is, all three groups heard /ɛ/ most often for target /æ/, and /æ/ most often for the target /ɑ/. Highly proficient bilingual listeners exhibited a substantial increase from the NN condition in the total number of vowels confused across all target vowels. HP listeners showed a total of 14 confusion events across all the vowels in the NN condition, compared to 23 confusions across all the vowels in the FP condition. This number of confusions by HP equals the figure misperceived by LP (i.e., 23) for the FP condition.

Table 9. Confusion matrix for the flattened preserved vowel (FP) condition – depicting performance of the monolingual, highly proficient bilingual, and less proficient bilingual listeners.

		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		**96.32**	2.94	0.74	0	0	0
HP	Bead	**90.5**	6.5	0.5	2	0	0.5
LP		**73.61**	22.92	2.78	0.69	0	0
NA		0	91.18	0.74	8.09	0	0
HP	Bid	1.5	83.5	0	15	0	0
LP		19.44	61.11	2.78	15.97	0	0.69
NA		1.47	49.26	**9.56**	36.03	3.68	0
HP	Bayed	3	39	**8**	49	1	0
LP		8.33	34.03	**8.33**	46.53	2.78	0
NA		0	13.97	0	80.88	5.15	0
HP	Bed	0.5	12.5	1	80	5.5	0.5
LP		2.78	16.67	2.08	59.72	15.28	3.47
NA		0	0	3.68	16.18	80.15	0
HP	Bad	0.5	0.5	1.5	24.5	72.5	0.5
LP		0	3.47	4.17	21.53	64.58	6.25
NA		0	0	0	0	10.29	89.71
HP	Bod	0	0	0.5	0.5	20.5	78
LP		0.69	0	0	0.69	40.28	58.33

The confusion patterns for the flattened neutral vowel condition are similar to those for the FP condition matrix, in that the strongest and weakest vowels were the same for all three groups for the two conditions (see Tables 9 and 10). Native, highly proficient bilingual, and less proficient bilingual listeners demonstrated most accurate identification performance for target /i/ (NA: 92.65%, HP: 92%, LP: 75.69%) and least accurate performance for target /e$^{\text{ɪ}}$/ (NA: 16.91%, HP: 11.5%, LP: 11.81%). The groups showed consistent confusion events for four vowels (/ɪ, e$^{\text{ɪ}}$, æ/, and /ɑ/ were mistaken mostly for /ɛ, ɛ, ɛ/, and /æ/, respectively). Further, the total numbers of vowels confused for targets made by native listeners increased dramatically from the FP to the FN listening condition (FP: 13 and FN: 20). However, native listeners continued to perform better than both groups of bilingual listeners for all target vowels, although their performance was only slightly better than that of the HP bilingual listeners for the target vowels /i/ and /ɛ/.

Table 10. Confusion matrix for the flattened neutral vowel (FN) condition – depicting performance of the monolingual, highly proficient bilingual, and less proficient bilingual listeners.

		Response					
Group	Target	Bead	Bid	Bayed	Bed	Bad	Bod
NA		**92.65**	2.94	0	4.41	0	0
HP	Bead	**92**	6.5	0.5	1	0	0
LP		**75.69**	20.83	2.78	0.69	0	0
NA		0.74	84.56	1.47	13.24	0	0
HP	Bid	4.5	72	1.5	21.5	0.5	0
LP		17.36	56.94	4.86	20.14	0	0.69
NA		0	34.56	**16.91**	44.12	4.41	0
HP	Bayed	1.5	35.5	**11.5**	48	3	0.5
LP		4.17	33.33	**11.81**	45.83	4.17	0.69
NA		0.74	17.65	0.74	72.79	7.35	0.74
HP	Bed	1	16.5	0	72	9.5	1
LP		2.78	13.19	2.78	56.25	16.67	8.33
NA		0	0	0.74	11.76	86.76	0.74
HP	Bad	0	0.5	1	23.5	74.5	0.5
LP		0	2.08	1.39	18.75	63.89	13.89
NA		0	0.74	0.74	0.74	9.56	88.24
HP	Bod	0	0	0.5	0.5	23.5	75.5
LP		0	0	0	0	37.5	62.5

Confusion Patterns. In discussing the confusion matrices, general patterns were noted for identification of the "strongest" and "weakest" vowel and the number of confusions made by condition. All three groups identified /e^{I}/ correctly most frequently for the WH condition. The vowel /e^{I}/ remained the strongest vowel for HP for the WH through NN listening conditions. All three listener groups heard /i/ most accurately for the FP and FN conditions. Similarly, all of the listener groups had most difficulties with /e^{I}/ for the FP and FN conditions.

Chapter 5

Discussion

In this study we explored monolingual and bilingual listeners' (highly proficient and less proficient bilinguals) use of perceptual cues for vowel identification. Six listening conditions were generated to examine use of consonant context, duration and formant cues in different listening conditions (i.e., whole word, isolated vowel, natural preserved vowel [resynthesized], natural neutral vowel [resynthesized], flattened formant vowels [resynthesized], and flattened neutral vowels [resynthesized]). The following questions were addressed: 1) Are vowels perceived differently in isolation than in whole words for these groups? 2) Are vowels synthesized using high fidelity synthesis (Straight) perceived differently than naturally produced vowels? 3) Do highly- and lower-proficiency bilinguals use vowel formant dynamic and duration cues differently than monolinguals? 4) Do the effects of isolation, synthesis and of formant dynamic and duration cues differ across the six vowels studied? 5) Do patterns of confusions differ across the listener groups?

Question 1: Effects of Vowel Isolation

Interpretations of the group by listening condition results (see Figure 4) imply that both highly and less proficient bilinguals rely more on consonant transitions (CV and VC) than native listeners (see performance for WH-OV conditions). Further, the three-way interaction analysis revealed a significant decrease in percent correct vowel identification performance for the vowel /ɑ/ for highly-proficient bilinguals and for /e^{I}, ɛ/ and /ɑ/ for the LP bilinguals when consonant transitions were removed (see Table 4). Although removing consonant transitions was not found to result in a significant decrease in performance for any of the vowels for native listeners (according to the three-way interaction analyses, see Table 4), a small but consistent decrease in performance (about 3% on average) was noted even for the native speakers.

An issue for further investigation may be why the LP bilinguals appear to rely so heavily on consonant context, especially for the target vowel /ɛ/, and, to a lesser extent, /ɑ/. The OV confusion matrix data suggest that the LP listeners hear /ɪ/ for target /ɛ/ (21.5% of target /ɛ/ perceived as /ɪ/, see Table 6). However, LP listeners also hear /e^{I}/ (9.0%) and /æ/ (7.6%) for /ɛ/. Negligible confusions were also noted for /i/ and /ɑ/ as response for the target word /ɛ/. According to Hillenbrand and Nearey (1999), the vowel drift of /ɛ/, i.e., measurements of F1 and F2 from the onset to the offset of the steady state portion of the vowel, was noted to have a decrease for both F1and F2 at the offglide relative to the onset. If this is taken into account, then the F1 offglide of /ɛ/, as perceived by Spanish-English bilinguals, should move towards /e^{I}/ and /ɪ/. Given this spectral change, LP bilinguals may, according to the PAM, assimilate /ɛ/ as a poor example of adjacent English vowels which in this case might be of /e^{I}/ and /ɪ/, although /æ/ would appear to be closer in the vowel space. As seen by the results of Table 6, this indeed seems to be the case in this study. Interestingly, both native and highly-proficient bilingual listeners appear to follow similar patterns; both listener groups confuse /ɪ/ for

/ɛ/ most frequently after the intended target vowel. One explanation for the particular difficulty experienced for the LP bilingual listeners may be the particular density of the vowel space in this region. Whereas no vowels occur between /e/ and /a/ in Spanish, two additional vowels (/ɛ/ and /æ/) occur between the vowels /e^{I}/ and /ɑ/ in English. Thus, two of the three apparently “new” vowels for Spanish speakers from among the six vowels studied in the present investigation are located between /e^{I}/ and /ɑ/. Confusions may occur for the LP bilinguals when consonant context is removed because categories for these vowels may be so poorly defined for the LP bilinguals that identification relies of specification of context-specific allophones.

Question 2: Effects of Straight Resynthesis

Effects of Straight resynthesis were only significant for the vowel /ɛ/ for the native listeners (see Table 4). This finding suggests that for most vowels identification performance for isolated vowels synthesized using high fidelity resynthesis (Straight) is comparable to that for naturally produced isolated vowels (OV), regardless of a listener’s proficiency. However, the fact that native listeners showed a significant decrease in identification performance for the vowel /ɛ/ raises the possibility that synthesis procedures may have been faulty for the vowel /ɛ/. It is interesting that the native listeners alone demonstrated this decrease for /ɛ/. Investigations of the resynthesis effects (OV to NP) for the remaining vowels show either equivalent scores (e.g., /ɑ/ for the native and HP bilingual listeners only), a slight increase in identification performance (e.g., /e^{I}, æ/ for all three groups) or only a slight decrease in identification performance (e.g., /ɪ/ for NA and HP only, /i/ for all three groups).

Further, comparisons with the results obtained by Assmann and Katz (2005) show a slightly larger overall decrease in performance from natural to Straight-synthesized stimuli by native listeners for the present study than for Assmann and Katz’s study. This difference may be related to the particularly poor performance for /ɛ/ in the present study, however percent correct performance is not broken down by vowel for natural vs. Straight-synthesized vowels in the Assmann and Katz (2005) data.

Question 3: Effects of Formant and Duration Cues for HP and LP Bilingual Listeners

As seen in Figure 4, native, highly proficient, and less proficient bilingual listeners showed similar and relatively small decreases in percent correct identification performance patterns for duration neutralization alone (NP-NN); all three groups’ NN performance was poorer than their NP performance, but only by approximately 1-5%. Overall, this effect was not significant for any of the listener groups, although it approached significance for the LP group. In the analysis of the three-way interaction, the effect of duration neutralization alone was found to be significant for the vowel /ɪ/ for the LP bilingual listener group only.

In the two-way interaction of listener group and listening condition, formant flattening (NP-FP) and the removing of both formant dynamic and duration cues (NP-FN) resulted in significant decreases in percent correct identification performance for all three listener groups. Inspection of the three-way interaction data, when post-hoc analyses were applied for only six comparisons (see Chapter 4), found that highly proficient bilingual listeners appear to use acoustic cues differently from both native and less proficient listeners. Highly proficient bilinguals demonstrated a significant effect of formant flattening alone (NP-FP) for the three vowels /e^{I}, æ/ and /ɑ/ whereas native and LP bilingual listeners only showed an effect for /e^{I}/ and /æ/. Of interest for the HP bilingual listener group, is that the size of effect (in RAUs) was noted to be greater for these two vowels for the HP bilinguals than for the native or the LP bilingual listeners.

The HP bilingual listeners demonstrated a significant decrease in performance for the five vowels /ɪ, e^{I}, ɛ, æ/ and /ɑ/ when both formant and duration cues were neutralized (NP-FN). Native listeners, on the other hand, showed a significant decrease in percent correct identification performance when both cues were for only three vowels (/e^{I}, ɛ, æ/). The increased size of this effect for the HP bilingual listeners for most vowels, in comparison to the native listeners (see Table 4) implies that HP bilinguals identify vowels with less consistently when static formant cues alone are available.

Less proficient bilingual listeners were noted to show significant decreases in performance when both formant dynamic and duration cues were removed for some vowels (i.e., /e^{I}/ and /æ/), the size of effect was not as great as for the HP bilinguals. Given that the less proficient bilingual listener group showed the greatest significant decreases in performance of all the groups for the effects of vowel isolation and duration neutralization alone and less extreme effects of formant flattening alone, it appears that LP bilinguals rely more heavily on CV and VC transitions and duration cues for vowel identification than the native and HP bilingual groups. The HP bilingual listener group, on the other hand, appears to rely more heavily than the native and LP bilingual groups on formant dynamic.

Question 4: Effects of Vowel and Listening Condition

As shown in Figure 5, average performance of all three groups by vowel and by condition reveal various perception patterns for the vowels across listening conditions. What is of interest is that each vowel carries at least one effect that is distinct. Common for all of the vowels, is the effect of consonant transition (WH-OV): all six vowels show considerable decreases. Although a decrease in performance is apparent to the eye for each vowel, only /e^{ɪ}, ɛ/ and /ɑ/ showed significant drops in performance from WH to OV conditions (according to the two-way interaction analysis). Noteworthy is that LP bilinguals showed significant decreases in performance from WH to OV conditions for all of these three vowels. For HP bilinguals, however, removing consonant transitions were found to significantly decrease percent correct identification performance only for the vowel /ɑ/, and the size of effect for this vowel was not as great as that for the LP bilingual listener group.

According to Figure 5, vowel isolation (WH-OV) appears to account for the greatest decrease in percent correct performance for the vowel /i/, although the effect was not found to be significant. The fact that no significant effect was evident for any of the listening condition comparisons helps to illustrate how different the effects of the various listening conditions are across the different vowels.

To illustrate, the target /ɪ/, on the other hand, showed significant decreases in percent correct identification performance for duration neutralization alone (NP-NN), duration neutralization of flattened stimuli (FP-FN), and when both duration and formant dynamic cues were removed (NP-FN). Note that the pattern of group by listening condition for this vowel is one of the most interesting (see Figure 5). It appears that the listener groups rely mostly on duration for /ɪ/. Tables 7 and 8 allows us to view the two bilingual listener groups' performance for /ɪ/ (see Appendices D.1-6). Confusion matrices for the NP and NN conditions reveal that LP bilinguals alone rely greatly on duration when no other vowel cue is modified. HP bilingual listeners, on the other hand, use duration cues more heavily when dynamic formant information is no longer present (FP-FN and NP-FN). Moreover, the confusion matrices reveal that HP and LP bilinguals hear /i/ and /ɛ/ for target /ɪ/ consistently for the NN, FP, and FN listening conditions. One reason for this observed pattern for the vowel /ɪ/ may be its close proximity to /i/ in terms of its formant frequencies at vowel onset. This proximity would suggest, according to Flege's SLM that /ɪ/, a new vowel for Spanish learners of English would be assimilated as a reasonably good exemplar of the Spanish /i/ category, making it a particularly difficult category to acquire for Spanish-English bilinguals. This hypothesis is also partially supported by the overall relatively poor performance of the LP bilingual group for this vowel (around 50-60% correct) and the LP group's apparently heavy reliance on duration cues for this vowel. Note that the LP group achieves only 60% correct performance for this vowel even in consonant context; for all other vowels the LP listeners achieve at least 75% in consonant context. Further investigate is needed to confirm or disconfirm this hypothesis.

Returning to Figure 5, the most obvious effect for the vowel /e^{I}/ is undoubtedly formant flattening (NP-FP and NP-FN): A very large decrease in percent correct identification performance was seen upon the introduction of formant flattening (NP-FP) for all three listener groups. No duration effect was evident for either of the groups. Of interest is the fact that native listeners perform as poorly as the HP and the LP listeners which suggest that both monolingual and bilingual listeners rely heavily on spectral information for the target vowel /e^{I}/. In the question 5 section, /e^{I}/ will be discussed in further detail with regards to FP-FN.

The resynthesis effect for /ɛ/ is of special interest, which relates to the previously discussed question. These results reveal that the only native listeners demonstrated a significant drop in percent correct identification from the OV to the NP condition and only for this vowel. That a significant effect of Straight resynthesis occurs only for the vowel /ɛ/ is intriguing and suggests further investigation, perhaps of the quality of synthesis for this vowel. Imprecise measurements may have caused too distorted quality causing native listeners to mistake this vowel for neighboring categories.

Percent correct identification of the vowel /æ/ decreased significantly for duration neutralization alone (NP-NN) and for formant flattening alone (NP-FP). However, of the two effects, that of formant flattening was larger and was the only one of the two that was found to be significant in the three-way interaction.

The results for the target vowel /ɑ/ showed significant effects of formant flattening alone (NP-FP) and removing both duration and formant dynamic cues (NP-FN) for HP bilingual listeners only.

The point to be made from Figure 5 is that all six vowels appear to be affected differently by the various listening conditions. However, to determine whether the listening groups show different confusion patterns for each vowel in the different conditions, it is necessary to review the confusion matrices.

Question 5: Confusion Patterns of Listener Groups

Analysis of the confusion matrices (Tables 5-10) and the three-way interaction results (Table 4) revealed that listener groups' performance differed substantially across vowels, listening conditions and listener groups. Tracking one vowel at a time through all of the conditions allows us to see that each group's performance fluctuates. For instance, for the target /ɑ/, HP bilinguals, but especially LP bilinguals, depend more heavily on consonant cues (CV and VC transitions) than native listeners. HP bilingual listeners appeared to use formant information (compare NP and FP condition performance) more profoundly than LP bilinguals and natives, but their performance deteriorated even more when both cues (NP to FN) were removed.

The confusion matrices enabled us to examine the breakdown and confusion events of each group by vowel and by listening conditions. Counting the number of confusion events performed by listener group by listening condition, revealed a larger number of events for the LP bilinguals than for the native and HP bilingual listeners from the whole word (WH) to the natural neutral (NN) listening condition. Upon removal of formant information (FP), the HP bilingual listeners increased their number of confusion events drastically. When both cues were removed, all three groups demonstrated a similar number of confusion events. Although this does not reveal differences in the identity of the confusions, the number of confusion events does imply that HP bilinguals become increasingly inconsistent with the removal of formant and duration cues (NP-FP and NP-FN).

The previously discussed vowel /e^{I}/ showed an increase in performance when both cues were removed for all the listener groups, in comparison with the FP condition. The target vowel /e^{I}/ remained the strongest item until the NN listening condition. The fact that all three listener groups demonstrated an increase of performance in the FN condition, relative to the flattened preserved condition (FP), is curious and gives reason for further investigation.

Of the remaining vowels, /ɛ/ and /ɪ/ are the only vowels for which all three listener groups demonstrated a decrease in performance from the FP to the FN condition. Other vowels displayed patterns in which both HP and LP bilingual listeners showed decreased performance only, or only HP and native listeners. However, native listeners and less proficient listeners did not share this pattern alone. i It is also of interest that the HP bilinguals either follow the native listeners' perception pattern or that of the LP bilinguals. Five out of six times, HP bilinguals followed the FP-FN tendency of the native listeners for /ɪ, ɛ, e^{I}, æ, ɑ/ (see Tables 9 and 10).

Summary

Parallels of the previously discussed results can be drawn to studies such as Sebastian-Galles & Soto-Faraco (1999), Lopez (2004), and Mayo et al. (1997), whose data suggested that bilinguals will be more challenged when 1) listening conditions are difficult, 2) fewer context cues are available, and 3) age of onset of learning the L2 is later. Our study supports these previous conclusions in that bilingual listeners appear to be more challenged when consonant, formant, and duration cues are manipulated more so than native listeners.

The data also suggest that increased L2 proficiency (or AOLI) appears to correlate positively with vowel identification. This can be seen in the less proficient listeners' consistently lower performance for all of the conditions, compared to that of the native and HP bilingual listeners (see Figure 4). The mean for the HP bilinguals was more accurate for all of the listening conditions than that the less proficient listeners, but it was lower than of the native listeners in most conditions. Nevertheless, it should be noted that the difference in performance between the native and LP bilingual listeners was, when averaged across vowels, between three and six times greater than the difference in performance between the native and HP bilingual listeners, depending on the listening condition.

In recapitulating Figure 4, the most striking tendencies are that 1) native listeners perform consistently better than the bilingual listeners whereas less-proficient bilinguals performed the poorest for all of the conditions, 2) highly but mostly less proficient bilinguals rely more heavily on consonant transitions than native listeners (WH-OV), 3) all listener groups demonstrate a significant drop in performance when both formant flattening and duration neutralization (NP-FN) are applied, 4) less proficient bilinguals appear to depend more heavily on duration cues than native and highly proficient bilinguals, 5) highly proficient bilinguals use dynamic formant cues more heavily for some vowels than do native listener.

Although the findings provide evidence that highly proficient bilinguals use vowel cues differently than both native and less proficient bilingual listeners, further investigation of production of the vowels is needed to determine whether improved perception of a vowel will reflect improved production also. A potential future study may include examining bilingual listeners' perception of L1, compared to that of monolinguals, when the same acoustic cues are altered as in the present study. Comparison between L1 and L2 perception may reveal whether the L2 listener demonstrates similar patterns for their L1 vowels.

One limitation of this study is that listeners were presented isolated words and vowels only. The ideal study would aim to examine listeners' perception of words in connected speech with the previously mentioned listening conditions. However, for this to materialize, more advanced speech resynthesis methodology would be required. The present study provides us only with a small frame of how listeners use acoustic cues in everyday life. Thus, conclusions of perception and production can for now be limited to one constituent: the vowel.

However, if adequate information regarding L2 listeners' use of vowel cues is available, then Straight resynthesis may be a likely software to be utilized for a number of therapy and education purposes: 1) improved second language acquisition training, 2) accent modification therapy, and 3) listening training for the hard of hearing population. Simplified resynthesis software may enable educators or therapists to record stimuli and target word of special interest. It is imaginable that modifying the cues through Straight may aid in strengthening an L2 learners' attention and perception of not-yet mastered vowel cues.

References

Assmann, P. F., & Katz, W. F. (2005). Synthesis Fidelity and Time-varying Spectral Change in Vowels. *Journal of Acoustical Society of America, 117*(2), 886-895.

Best, C. T. (1995). A Direct Realist View of Cross-Language Speech Perception. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 171-204). Timonium, MD: York Press, Inc.

Boersma, P., & Weenink, D. (2003). Praat [Computer software] (Version 4.2). Amsterdam, The Netherlands: Institute of Phonetic Sciences, University of Amsterdam.

Bohn, O. S., & Flege, J. E. (1999). Perception and production of a new vowel category by adult second language learners. In A. James & J. Leather (Eds.), *Second-Language Speech: Structure and Process* (pp. 53-74). Berlin: Mouton de Gruyter.

Borden, G. J., Harris, K. S., & Raphael, L. J. (1994). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech* (Third ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Crystal, D. (1997). *The Cambridge Encyclopedia of Language* (2 ed.). Cambridge, UK: Cambridge University Press.

Dalbor, J. B. (1969). *Spanish pronunciation: Theory and practice.* New York.: Holt, Rinehart & Winston, Inc.

Dudley, H. (1939). Remaking Speech. *The Journal of the Acoustical Society of America, 11*, 169-177.

ECoS/Win (Version 1.3) [Computer software] (1999). London, Ontario: AVAAZ Innovations, Inc.

Febo, D. M. (2003). *Effects of Bilingualism, Noise, and Reverberation on Speech Perception by Listeners with Normal Hearing.* Unpublished Doctor of Audiology, University of South Florida, Tampa, Florida.

Flege, J. E. (1981). The Phonological Basis of Foreign Accent: A Hypothesis. *TESOL Quarterly, 15*(4), 443-455.

Flege, J. E. (1995). Second language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-Language Research* (pp. 233-272). Timonium, MA: York Press.

Flege, J. E. (1996). English vowel production by Dutch talkers: more evidence for the "similar" vs "new" distinction. In A. James & J. Leather (Eds.), *Second-Language Speech: Structure and Process* (pp. 11-52). Berlin: Mouton de Gruyter.

Flege, J. E., Munro, M. J., & MacKay, I. R. A. (1995). Effects of age of second-language learning on the production of English consonants. *Speech Communication, 16*, 1-26.

Fry, D. B. (1979). *The Physics of Speech*. Cambridge, UK: Cambridge University Press.

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The Identification of and discrimination of synthetic vowels. *Language and Speech, 1*, 35-38.

Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *The Journal of the Acoustical Society of America, 97*(5), 3099-3111.

Hillenbrand, J. M., & Nearey, T. M. (1999). Identification of resynthesized /hVd/ utterances: Effects of formant contour. *Journal of Acoustical Society of America, 105*(6), 3509-3523.

Johnston, D. (2000). CoolEdit [Computer software] (Version 1.1). Phoenix, AZ: Syntrillium Inc.

Kawahara, H., Masuda-Katsuse, I., & Cheveigné, A. d. (1998). Restructuring speech respresentations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication, 27*, 187-207.

Kewley-Port, D., Akahane-Yamada, R., & Aikawa, K. (1996). *Intelligibility and acoustic correlates of Japanese accented English vowels.* Paper presented at the Proceedings of ICSLP 96, Philadelphia, PA.

Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of Acoustical Society of America, 67*(3), 971-995.

Klatt, D. H. (1987). Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America, 82*(3), 737-793.

Klatt, D. H., & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America, 87*(2), 820-856.

Labov, W. (2005). *Atlas of North American English*. Retrieved 6.21.2005, 2005, from http://www.ling.upenn.edu/phono_atlas/home.html

Liu, C., & Kewley-Port, D. (2004). Vowel formant discrimination for high-fidelity speech. *Journal of Acoustical Society of America, 116*(2), 1224-1233.

Lopez, A. S. (2004). *Silent-Center Vowel Perception by Spanish-English Bilinguals and Monolinguals English Speakers.* University of South Florida, Tampa.

MacKay, I. R. A., Meador, D., & Flege, J. E. (2001). The Identification of English Consonants by Native Speakers of Italian. *Phonetica, 58*, 103-125.

Major, R. C. (2001). *Foreign accent: the ontogeny and phylogeny of second language phonology*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Markel, J. D. (1972). Digital inverse filtering-a new tool for formant trajectory estimation. *IEEE Transaction on Audio and Electroacoustics, 20*(2), 129-137.

Mayo, L. H., Florentine, M., & Buus, S. (1997). Age of Second-Language Acquisition and Perception of Speech in Noise. *Journal of Speech-Language Hearing Research, 40*, 686-693.

Meador, D., Flege, J. E., & MacKay, I. R. A. (2000). Factors affecting the recognition of words in a second language. *Bilingualism: Language and Cognition, 3*(1), 55-67.

Microsoft (2000). Microsoft Excel 2000 [Computer software]: Microsoft.

Monsen, R. B., & Engebretson, A. M. (1983). The accuracy of formant frequency measurements: A comparison of spectrographic analysis and linear prediction. *Journal of Speech-Language Hearing Research, 26*, 89-97.

Peterson, G. E., & Barney, H. L. (1952). Control Methods Used in Study of the Vowels. *The Journal of the Acoustical Society of America, 24*(2), 175-184.

Pickett, J. M. (1999). *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology*. Boston, London, Toronto, Sydney, Tokyo, Singapore: Allyn and Bacon.

Rochet, B. L. (1995). Perception and Production of Second-Language Speech Sounds by Adults. In W. Strange (Ed.), *Issues in cross-language research* (pp. 379-410). Timonium, MD: York Press, Inc.

Sebastian-Galles, N., & Soto-Faraco, S. (1999). Online processing of native and non-native phonemic contrasts in early bilinguals. *Cognition*(72), 111-123.

Shin, H. B., & Bruno, R. (2003). *Language Use and English-Speaking Ability: 2000 - Census 2000 Brief*: United States Census Bureau.

Strange, W. (1999). Perception of vowels: dynamic constancy. In J. M. Pickett (Ed.), *The acoustics of speech communication: fundamentals, speech perception theory, and technology* (pp. 153-165). Surry, Maine: Allyn and Bacon.

Strange, W., Akahane-Yamada, R., Kubo, R., Trent, S. A., Nishi, K., & Jenkins, J. E. (1998). Perceptual Assimilation of American English Vowels by Japanese Listeners. *Journal of Phonetics*, *26*, 311-344.

Strange, W., Jenkins, J. J., & Johnson, T. L. (1983). Dynamic specification of coarticulated vowels. *Journal of Acoustical Society of America*, *74*(3), 695-705.

Studebaker, G. (1985). A “rationalized” arcsine transform. *Journal of Speech and Hearing Research*, *28*, 494-509.

TDT System III. [Computer hardware and software] (2001). Gainesville, FL: Tucker-Davis Technologies, Inc.

The MathWorks, I. (2002). MATLAB [Computer software] (Version 6.5.0): The MathWorks, Inc.

Tye-Murray, N. (1998). *Foundation of aural rehabilitation: Children, adults, and their family members*. San Diego, CA: Singular Publishing Group.

Appendices

Appendix A. Monolingual Language Questionnaire

Participant Background Questionnaire (Form A)

Name: _______________ Age: _______ Address (town & state): _________

Phone (optional). Home: _______ Office: _______

Email address (optional): _______________________

1. Is English your first (native) language? Circle one: Yes No

    a. If you answered "No" to (1) above, list your language here.

2. Did you speak any languages other than English while growing up (other than classroom instruction)? Circle one: Yes No

    a. If you answered "Yes" to (2) above, list those language here ___________ ________________________________

3. List any languages you speak other than English and rate your degree of proficiency on a scale from "1" to "5" for each (1=beginner, can't have a conversation; 5=like a native speaker):

    ____

4. Have you ever been diagnosed with a speech or hearing disorder or had speech or hearing difficulties? Circle one: Yes No

    a. If you answered "yes" to (4), above, please explain in the space provided below (or on back if you need more room): ____

5. How long have you lived in Florida (or current state)? ____________________

6. What state where you born in and how long did you live there? ____________

(Don't answer #'s 7 or 9 if you've lived all your life in 1 state)

7. What state have you lived in the longest in? _________________

    a. How many years did you live there? _________________

8. List any other states that you've lived in for over a year (if more than 3, list top three): ____

9. On a scale from "1" to "7", rate your experience with listening to speakers with a foreign accent (1=little or little experience; 7=every day or very frequent): ______

Appendix B. Bilingual Language Questionnaire

Participant Background Questionnaire (Form B)

Name: ________________ Age: _______ Address (town & state): _________

Phone (optional). Home: _____ Office: ______

Email address (optional): _____________________

1. How many years have you lived in your current area (town & state)? _______
2. Have ever been diagnosed with a speech or hearing disorder or had speech or hearing difficulties: Circle one: Yes No
    a. If you answered "yes" to (2), above, please explain in the space provided below (or on back if you need more room):
    ____
3. What language(s) did your parents speak with your? _______________
    a. If you answered with more than one language in (1), above, which language(s) did each parent speak with you?
    ____
4. Where were you born (give city, state, country) __________________________
    a. How many years did you live there? ______
    b. List other cities or regions you've lived in for more than one year and note number of years you lived there for each.
    ____
    c. What city and country are your parents from?

    Mother: _____________ Father: _____________
5. How old were you began learning English? ___________
    a. Why did you begin learning English? _______________________
    ____
6. If you moved to the United States from another country, how much did you speak English before moving here (describe years of study, if you learned English in a classroom & percent of time speaking English):
____

7. If you moved to the United States from another country, how long have you lived here? _______ years, ________ months.

8. On a typical day, what percent of your time do you spend speaking English at work? _____% At home? ______% Other (shopping, etc.)? ______%

9. On a typical day, what percent of your time do you spend speaking a language other than English at work? _____ % At home? ______%

Other (shopping, etc.)? ____% (if more than one, answer below for each language)

10. What percent of percent of day do you spend with people with people who speak both (or more) language that do? _______%

11. What language are you most comfortable speaking? ___________
    a. How much more comfortable are you in speaking that language on a scale of 1 to 5? (1=equal or nearly equal comfort; 5=much more comfortable) ______
12. What language are you most comfortable listening in? ___________
    a. How much more comfortable are you in listening in that language on a scale of 1 to 5? (1=equal or nearly equal comfort; 5=much more comfortable)
13. What language are you most comfortable reading in? ________
    a. How much more comfortable are you reading in that language on a scale of 1 to 5? (1=equal or nearly equal comfort; 5=much more comfortable) ________
14. What languages are you most comfortable writing in? _______
    a. How much more comfortable are you writing in that language on a scale of 1 to 5? (1=equal or nearly equal comfort; 5=much more comfortable) ______
15. Do you think your ability in the language you are less comfortable in is still improving for any of the skills in questions 9-12? Circle one: yes no

    a. If you answered yes in 13 above, indicate which abilities you believe are still improving.
    Circle any that apply: speaking listening reading writing

16. What academic degrees have you earned? (List language of education for each) ______________________________

17. For all languages that you speak, rate your level of ability on as scale of 1 to 5 (1=not proficient, like a child or beginner=very proficient, like a well-educated native speaker) for each of the following areas:
    a. Comprehension: ______________________________
    b. Fluency (ease of expression): ___________________
    c. Vocabulary: __________________________________
    d. Pronunciation: _______________________________
    e. Grammar: ____________________________________

Appendix C. Steady State Replication

1. Open Praat
    - a. Find original resampled sound file
    - b. Find the steady state time (steady state time can be found in vowedit_meas_goodCR.xls)
    - c. Zoom in around the steady state time
    - d. Select→Move cursor to→enter the steady state time
    - e. Zoom in total of 30 ms around the steady state time
        - - Select→Move begin of selection by-->enter -0.015s
        - - Select→Move end of selection by→ enter 0.015s
        - **Count the number of cycles (x) within the selected frame
            - -Find the first big negative or positive (whichever has more of a zero cross)
            - -Count the number of complete cycles plus one extra peak
            - -Enter the number of complete cycles in the excel spreadsheet
    - f. Select→Move begin of selection to nearest zero crossing
        - - Enter time (F5) in excel
    - g. Select→Move end of selection to nearest zero crossing
        - - Enter time (F7) in excel
2. Praat Objects
    - a. New→Sound→Create Sound
        - - Change name of file to "Silence"
        - - Change sample rate to 11050 Hz
        - - Formula: Delete the end of equation (+ randomGauss(0,0.1)) and change the amplitude from ½ to zero
        - ** This will provide one second of silence
3. Go back to opened Praat sound file and copy the selected measure into the "Silence" sound file
    - a. Paste in one set of selected measure
    - b. Zoom in and place cursor at the end of last complete cycle at the zero cross (before the extra peak)
    - c. Select→Move cursor to nearest zero crossing
    - d. Continue to paste in sets of the selected measure until have enough cycles to edit for the full vowel time (should have longer vowel duration than for the full vowel duration or for largest average)
        - -Paste in at least 12 complete cycles
    - ** In excel, want to look at time measures for the original word, average time of all words for that subject, and overall average time for all subjects
    - e. Find the last complete cycle and move cursor to the nearest zero crossing
    - f. Select remaining unnecessary extra information (extra peaks)
        - -Edit→Set selection to zero (this will silence the extra bit)
    - g. Remove extra silence before saving file (leave only 30ms of time at the beginning and end of the vowel)

4. Praat Objects window

    a. Write→Write to WAV file→Save in word edit folder as "name_FLRepLong.wav"

5. Repeat steps 3 f-g and 4 for the full vowel and natural vowel times

    a. Select→Move cursor to→ enter 0.03s

    b. Select→Move cursor by→ enter appropriate time form excel file

    c. Zoom in on new time

    d. Select→Move cursor to nearest zero crossing

    e. Select remaining extra peaks and silence

        -Edit→Set selection to zero

    f. Edit for 30ms of silence at the end

    g. Save the full vowel as "name_FLRepFulen.wav"

    h. Save the natural vowel as "name_FLRepEdlen.wav"

6. Copy files form the name edit directory to the Straight directory

7. Matlab Straight

    a. Start straight by entering the word "straight"

    b. GUI window will pop up

        - Select Initialize

        - Select Read from file-find "name_FLRepEdlen.wav"

    c. Continue through selecting prompts down the left hand side

    d. Play the original and resynthesized versions

    e. Save new resynthesized file as "name_FLRepEdlenResyn.aiff"

    f. Repeat steps b-e for "name_FLRepFulen.wav" and save the new resynthesized file as "name_FLRepFulenResyn.aiff"

8. Copy files form the Straight directory to the "name Edit" directory

9. Praat Objects

    a. Open the resynthesized sound files

        -Read→Read from file→select resynthesized files

    b. Select→Move cursor to→ enter 0.03s (zoom in on cursor)

    c. Select→Move cursor to nearest zero crossing

    c. Select waveform to the left of the cursor by using SHIFT and left clicking mouse simultaneously

    d. Edit→Set selection to zero

    e. Zoom in on beginning

    f. Select→Move cursor to→enter 0.03s

    g. Select→Move cursor by→enter appropriate vowel duration time (Fulen or Edlen as appropriate from excel)

    h. Hit F6 and copy that time (Ctrl C) that you moved to

    i. Zoom in and move cursor to the end vowel time

        -Select→Move cursor to→Ctrl V (will paste in time that was copied)

    j. Select→Move cursor to nearest zero crossing

    k. Use SHIFT and left click mouse to select remaining end time

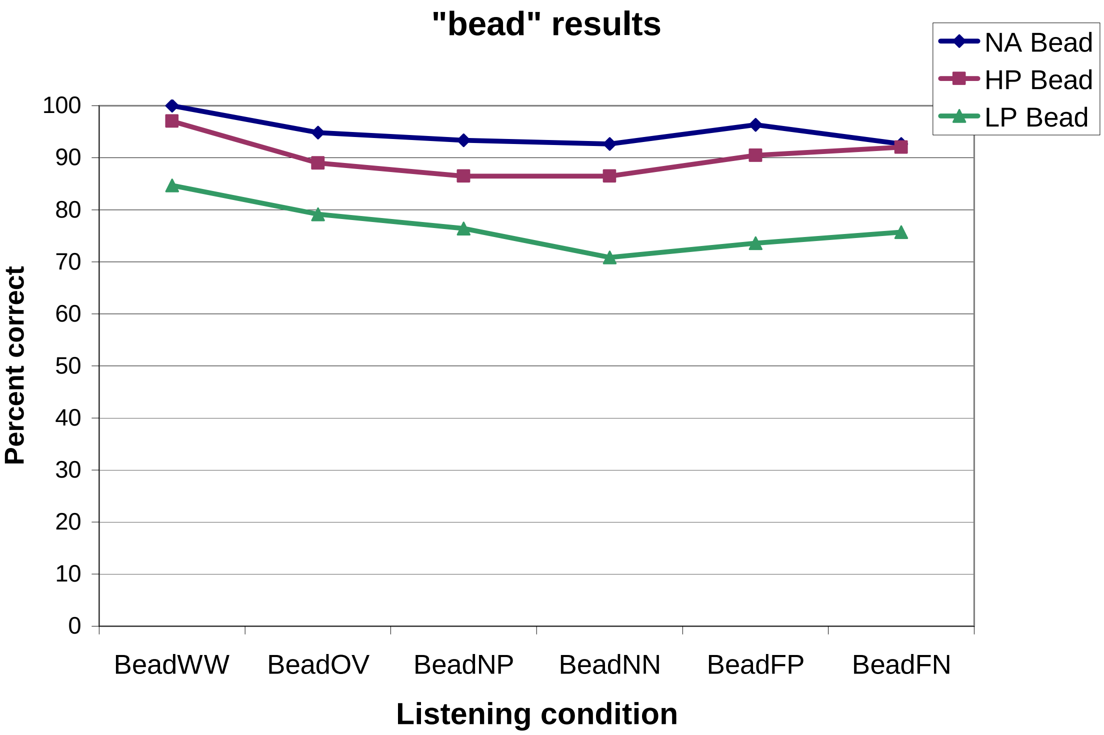
    l. Edit→Set selection to zero

-Make sure to leave 30ms of time at the end (if the difference from 30ms is greater than 1ms, then back into the file a little bit and edit at a new zero cross)

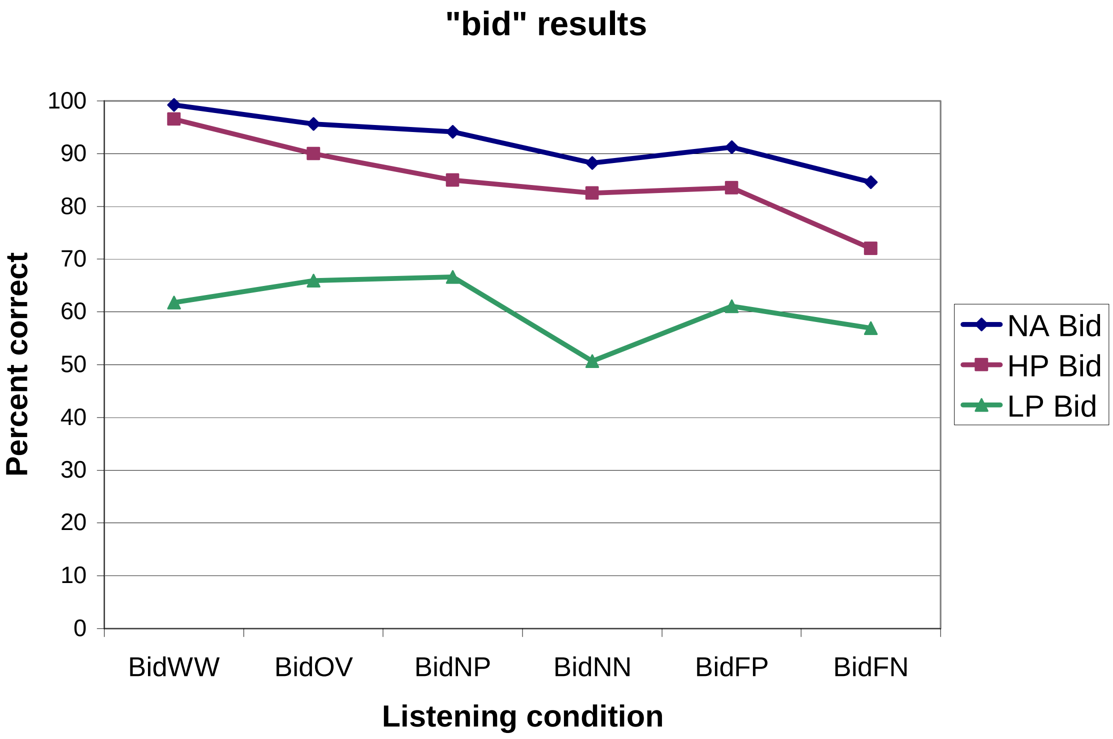
10. Praat Objects

a. Write→Write to WAV file→ save as "name_FLRepEdlenResynGdDur.wav" and "name_FLRepFulenResynGdDur.wav"

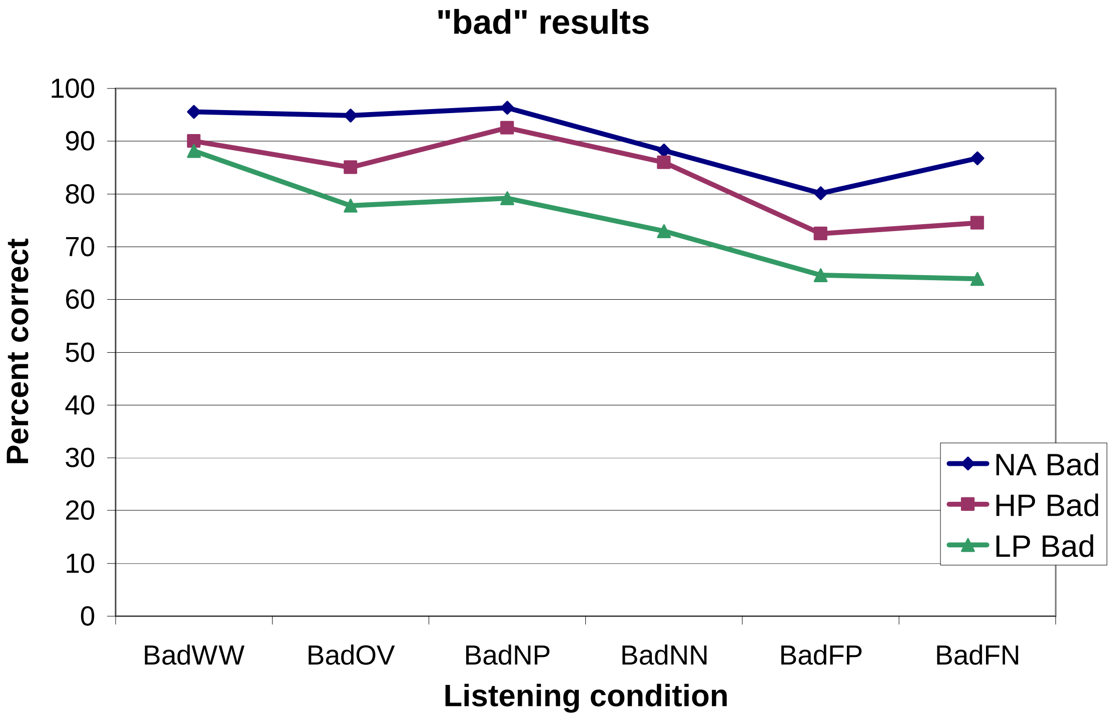
Appendix D.1. Results for the vowel /i/ by listening conditions by listener group.



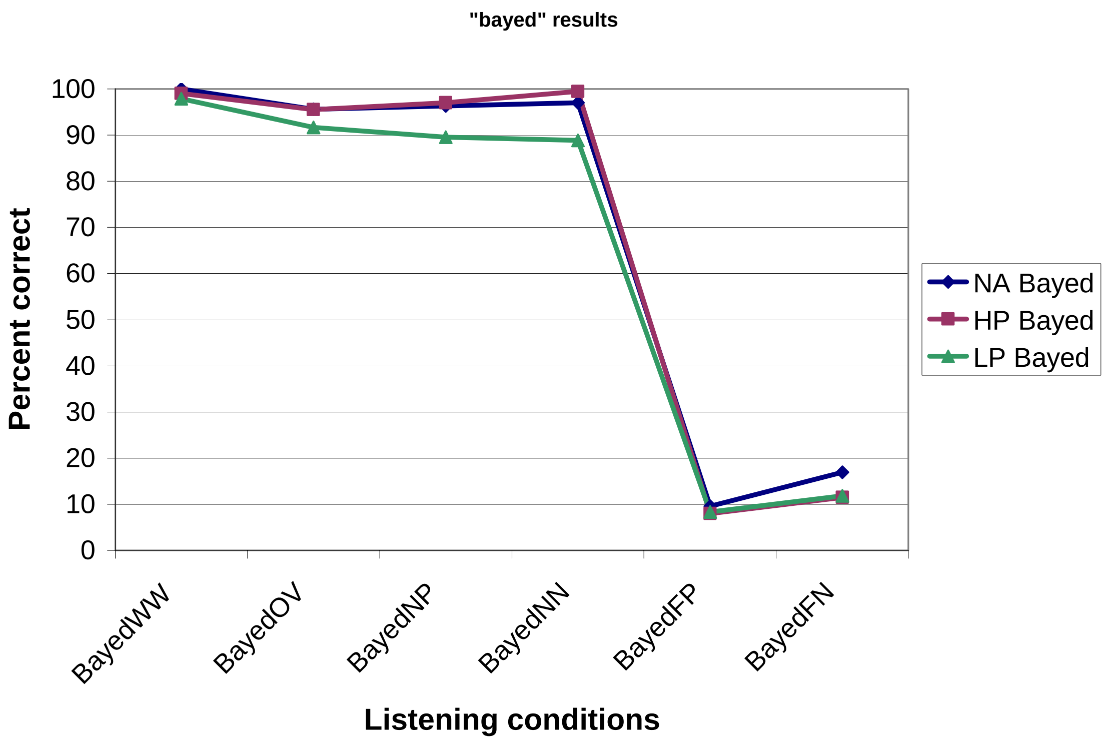
Appendix D.2. Results for the vowel /ɪ/ by listening conditions by listener group.



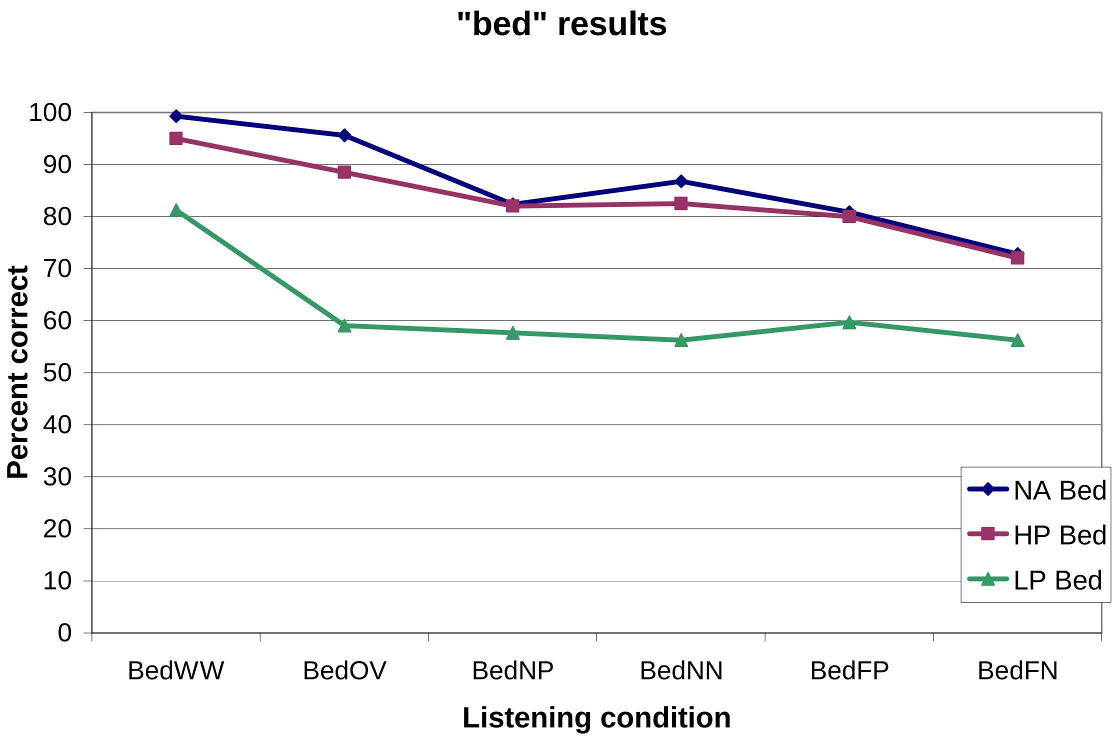
Appendix D.3. Results for the vowel /æ/ by listening conditions by listener group.



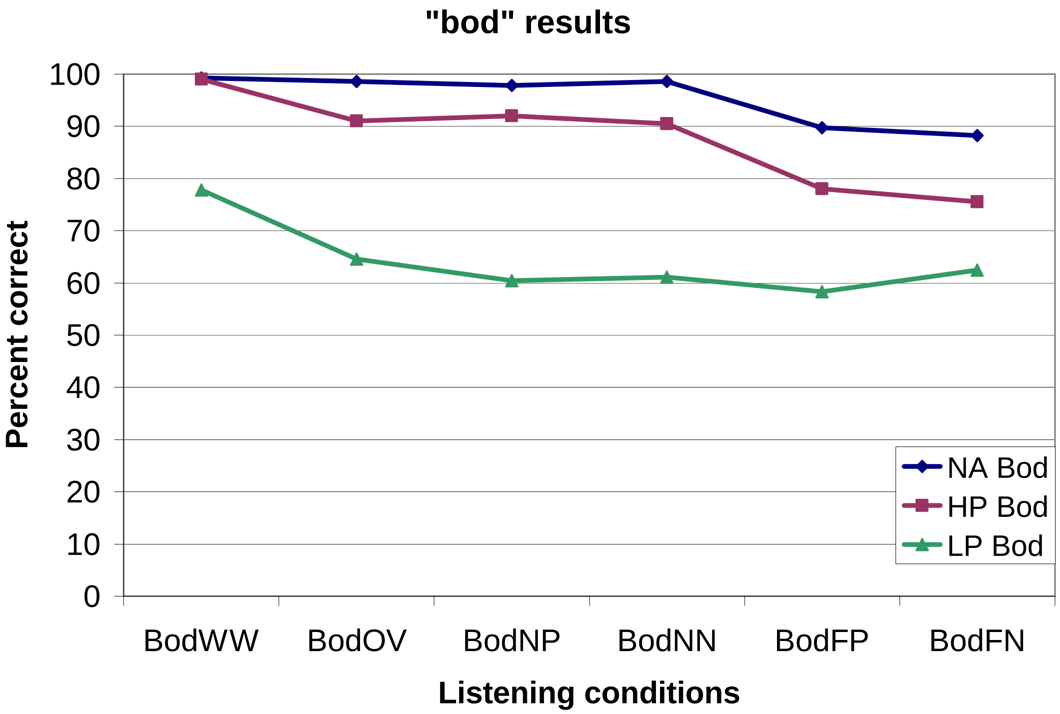
Appendix D.4. Results for the vowel /e^{I}/ by listening conditions by listener group.



Appendix D.5. Results for the vowel /ɛ/ by listening conditions by listener group.



Appendix D.6. Results for the vowel /ɑ/ by listening conditions by listener group.


"bod" results
Percent correct
BodWW
BodOV
BodNP
BodNN
BodFP
BodFN
Listening conditions
NA Bod
HP Bod
LP Bod