ACOUSTIC CORRELATES OF PERCEIVED FOREIGN ACCENT IN NON-NATIVE ENGLISH

DISSERTATION

Presented in Partial Fulfillment of the Requirements

for the Degree Doctor of Philosophy

in the Graduate School of The Ohio State University

by

Elizabeth A. McCullough, B.A., M.A.

Graduate Program in Linguistics

The Ohio State University

2013

Dissertation Committee:

Mary Beckman, Advisor

Kathryn Campbell-Kibler

Cynthia Clopper

Shari Speer

ABSTRACT

Skilled perception of speech involves not just recognizing the words and sentences that a talker produces, but also perceiving properties imputed to the talker, such as being a foreigner. Because being perceived as foreign can have social consequences, it is important to understand the characteristics that contribute to this percept. Foreign accent perception is often studied in relation to talker characteristics, such as age of learning the second language (L2) or age of arrival in the L2-speaking country. However, listeners generally do not have direct access to such information. In order for the perception of foreign accent to be fully understood, it must be studied in relation to physical characteristics of the speech signal.

This dissertation reports a series of six experiments that elicited American English monolinguals' ratings of various properties of productions from native talkers of American English, Hindi, Korean, Mandarin, and Spanish, as well as their free classifications of talker language background. In the first four experiments, listeners heard samples of English and rated foreign accentedness or non-nativeness for each production. In the final two experiments, listeners rated non-Englishness for samples of each talker producing his or her native language (L1). Linear mixed effects regression models revealed that VOT, F1 frequency, and F2 frequency correlated with ratings of accentedness and non-nativeness for syllable- and word-length stimuli. In addition, F3 frequency and F2 tilt correlated with the ratings for syllable-length stimuli, and vowel duration with the ratings for word-length

stimuli. Non-Englishness ratings for word-length stimuli were closely related to listeners' ability to recognize the stimuli as known English words. Free classification results revealed that across listeners, grouping patterns for native talkers were more consistent than for most non-native talkers. The correlates of multidimensional scaling analyses of the free classification responses were similar to the correlates of the ratings.

The results of this investigation reveal, for a varied sample of non-native English, which characteristics of the speech signal may lead American English monolinguals to identify a talker as foreign. When perceiving syllable-length stimuli, listeners seem to attend to phonetic details resulting from transfer from the non-native talker's L1, while indications of the talker's L2 fluency may begin to influence perception in units as small as disyllabic words. Such information may effectively identify priorities for L2 learners of English interested in accent reduction.

ACKNOWLEDGMENTS

I am gratefully indebted to many people for helping to make this dissertation a reality. First among them is my advisor, Mary Beckman, who has been unfailingly supportive as my research has taken shape—and sometimes transformed—during my time in Columbus. Mary's talent for keeping her eye on the big picture when I was lost in the details made many harried afternoons more bearable, and her perpetual enthusiasm about my questions reignited my own on numerous occasions. I also extend my earnest appreciation to Cynthia Clopper, who did a great deal of advising for someone who was never formally my advisor. It was mostly in Cynthia's office that questions about topics that interested me were molded into actual experiments. I once considered myself a "production person"; it was only through Cynthia's guidance and generosity that I began to investigate speech perception at all. Thanks, too, to the remaining members of my committee, Kathryn Campbell-Kibler and Shari Speer, who prodded me in directions I might otherwise have ignored, reeled me in when I went too far, and were patient while I began to learn how to communicate across subfields.

It felt, at times, overwhelmingly ambitious to design controlled materials in languages I didn't know and often couldn't read. Manas Agrawal, Jeff Holliday, Ila Nagar, Lorena Sainz-Maza Lecanda, Jeonghwa Shin, Ritu Singh, and Qingyang Yan were vital in helping me over these hurdles. Charlie Ann, John Grinstead, Bridget Smith, Seth Wiener, and

especially Jeff Holliday and Dahee Kim went above and beyond to help me recruit talkers, as did a number of generous participants.

My officemates Jeff Holliday, Andy Plummer, Pat Reidy, and Rory Turnbull have for years assisted me with technical issues in R, Praat, and LaTeX. For their vast technical knowledge, and their non-technical chatter, I am grateful. Many others deserve credit for generally helping me to maintain my sanity during grad school, including Clint Bruce, Katie Carmichael, Lara Downing, Kathleen Hall, Dahee Kim, Eunjong Kong, Marivic Lesho, Kevin McCullough, Deborah Morton, Jen Phelan, Mike Phelan, Kayleigh Scalzo, Oxana Skorniakova, Talia Turnbull, Abby Walker, Dani Weatherholtz, Kodi Weatherholtz, Chantal White, and Chris Worth.

A small number of people from before the graduate school era of my life have nonetheless contributed to this accomplishment. Thanks to Sean O'Dell and Matt Kanefsky, who first led me to linguistics, and to Katherine Demuth, who persuaded me to pursue it. Thanks also to my parents, Carol and Ray McCullough, who taught me to believe I could do anything, even things that they knew nothing about. They probably know very little about this dissertation, as they were kind enough not to force me to discuss my research when I called them to take a break from it.

Finally, I offer my gratitude and my love to Greg Kierstead for the countless incarnations of academic and personal support through the years (especially the milkshakes).

This work could not have been completed without a mosaic of generous funding, including a Dean's Distinguished University Fellowship from the Graduate School and a Targeted Investment in Excellence grant from my department.

VITA

May, 2006 ..... B.A., Linguistics
Brown University
Providence, RI, USA

August, 2011 ..... M.A., Linguistics
The Ohio State University
Columbus, OH, USA

PUBLICATIONS

Journal Articles

McCullough, E. A. (2013). Perceived foreign accent in three varieties of non-native English. *Ohio State University Working Papers in Linguistics*, *60*, 51-66.

Demuth, K., & McCullough, E. (2009). The longitudinal development of clusters in French. *Journal of Child Language*, *36*, 425-448.

Demuth, K., & McCullough, E. (2009). The prosodic (re)organization of children's early English articles. *Journal of Child Language*, *36*, 173-200.

Conference Papers

McCullough, E. (2012). Acoustic correlates of perceived foreign accent in non-native English. Poster presented at the 13th Conference on Laboratory Phonology, Stuttgart, Germany, 27 July.

McCullough, E. (2011). Acoustic correlates of perceptual ratings of foreign-accented English. Poster presented at the 161st Meeting of the Acoustical Society of America, Seattle, WA, 24 May. Abstract published in *Journal of the Acoustical Society of America*, *129*(4), 2453.

Hardman, J. B., & McCullough, E. (2010). Applications of the Buckeye GTA Corpus for L2 teaching and research. In *Proceedings of the INTERSPEECH 2010 Satellite Workshop on Second Language Studies: Acquisition, Learning, Education, and Technology* (P2-14).

Demuth, K., McCullough, E., & Adamo, M. (2007). The prosodic (re)organization of determiners. In H. Caunt-Nulton, S. Kulatilake, and I. Woo (Eds.), *Proceedings of the 31st Annual Boston University Conference on Language Development* (pp. 196-205). Somerville, MA: Cascadilla Press.

Demuth, K., McCullough, E., & Kehoe, M. (2005). Representational issues in the acquisition of word-initial and word-final consonant sequences in French. Paper presented at the International Congress for the Study of Child Language, Berlin, Germany, 27 July.

FIELDS OF STUDY

Major Field: Linguistics

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure **Page**

CHAPTER 1: INTRODUCTION

Scovel (1981, 389) asserts that "Wherever languages come into contact . . . it is readily apparent that speech serves not only to unite, but also to divide. In Montreal, in Europe, in Africa, and in nation after nation, province after province, and even village after village, human speech serves to signal the difference between in-group and out-group." The present work considers the acoustic details that cause listeners to perceive differences in the ways people speak. Such details matter because the ways people speak can reveal clues about their geographic origins and social connections, and can influence how listeners think of them and behave toward them.

Specifically, this dissertation focuses on differences in speech that arise from talkers having different native languages from one another. Evidence from child development suggests that early in life, such differences can cue group membership. Kinzler et al. (2009) found that the monolingual English-speaking 5-year-olds they tested preferred to be friends with native speakers of their language over non-native speakers of their language, and over speakers of a foreign language. The preference for native speakers was equally strong regardless of whether the competitor was a different accent or a different language. Moreover, when native speakers of their language were not presented as an option, the children chose to be friends with non-native speakers of their language and speakers of a foreign language equally often. An additional experiment explored the role of race in social preferences. The white children included in this study preferred to be friends with a person of

their own race when no language was involved; however, when speech and race were pitted against one another directly, such that "the person who looked like an ingroup member sounded like an outgroup member" (629) by virtue of having speech that was perceived to be foreign-accented, the children preferred to be friends with people who looked different but sounded like them.

Adults, too, are influenced by the ways people speak. A large body of literature examines adult listeners' attitudes about non-native speech, and generally highlights issues of prejudice, discrimination, and/or negative stereotyping (Gluszek and Dovidio, 2010). Indeed, native listeners rate non-native talkers lower than native ones on measures of status (Brennan and Brennan, 1981; Lindemann, 2003; Ryan et al., 1977; Tsurutani, 2012) and solidarity (Ryan et al., 1977). Native listeners judge non-native talkers to be less suitable than native talkers for high-status jobs (Kalin et al., 1980), and foreign accentedness has been implicated in many employment-related legal cases. In the United States, Title VII of the Civil Rights Act of 1964 prohibits discrimination based on race and national origin, but it offers no protection from discrimination based on accent (Nguyen, 1993).

Foreign accentedness, then, affects the experiences of non-native talkers as well as the listeners who interact with them. However, as "foreign accent" is rarely defined precisely, clarification regarding the meaning of the term in the present work is merited. Some previous explanations include:

- "a set of pronunciation patterns, at both segmental and suprasegmental levels, which differ from pronunciation patterns found in the speech of native speakers" (Volín and Skarnitzl, 2010, 1010)
- "how different a pattern of speech sounds compared to the local variety" (Derwing and Munro, 2009, 476)

- "speech which differs acoustically from the native phonetic norm, and is auditorily detectable by native speakers" (Wayland, 1997, 346)

- "all potential deviation from speech that a native speaker would consider normal" (Calla McDermott, 1986, 34)

- "any deviations from [the] L2 phonetic norm perceived by native L1 informants as unnatural, unlikely, but definitely not regional realizations" (Tomaszczyk, 1981, 131)

- "a deviation from the generally accepted norm of pronunciation of a language that is reminiscent of another language, i.e. the speaker's native language. It has to be emphasized that such a deviation must be defined in terms of its perception by listeners who are native speakers of the respective language and not in terms of differences in articulation that may be instrumentally measurable. Only those deviations that are perceived as such can be considered instances of foreign accent" (Jilka, 2000, 9)

Taking into account these descriptions, as well as the literature more generally, the following definition is proposed for "foreign accent":

- the percept of deviations from a pronunciation norm that a listener attributes to the talker not speaking the target language natively

According to this definition, as well as many of the others cited above, foreign-accented speech differs in some way from speech produced by native talkers. However, identifying who counts as a native talker, where the boundaries of the "target language" are drawn, and what constitutes the relevant "norm" is not trivial. In the case of American English listeners, individuals from other countries may speak English natively, but sound extremely different from American English talkers (Scovel, 1995). Whether the speech produced

by such talkers is judged to sound foreign-accented by American English listeners is an underexamined empirical question.

The definition proposed above limits foreign accent to the realm of pronunciation, which is quite common, although not universal (Calla McDermott, 1986). Additionally, it does not refer to just any deviations in pronunciation, but only those which are detected by listeners and believed to arise from a particular characteristic of the talker: his or her status as a non-native. The listener's attribution of the source of the deviation is missing overtly from many definitions, but is clearly implied; "foreign accent" is not used to describe what an American English listener from Ohio thinks about the speech of a native American English talker from Alabama, or of a very young native talker from his or her own region, although productions from these talkers are likely to sound different from the local adult norm in perceptible ways.

Crucially, although it is commonly associated with talkers of non-native backgrounds, "foreign accent" itself involves a judgment from a listener about a talker's speech. Often this listener is a native speaker of the target language, although this is not required (Major, 2007). Researchers often make the role of perception obvious by collecting responses from large numbers of listeners, and sometimes by explicitly stating it: Derwing and Munro (2009, 478), for instance, assert outright that "listeners' judgments are the only meaningful window into accentedness." Even if a study is not framed as a speech perception experiment, however, listeners are required at some level of the analysis. For instance, native speakers of the target language (Brière, 1966) or trained linguists (Brennan and Brennan, 1981) may be called upon to listen to productions in order to make qualitative decisions about the pronunciations of non-native talkers. Because foreign accent reflects perception, listeners' judgments about foreign accentedness need not align with the actual language

backgrounds of talkers. That is, by the definition of "foreign accent" above, listeners may perceive no foreign accent in the productions of talkers who actually speak the target language non-natively, and they may perceive a foreign accent in the productions of native talkers of the target language.

Some researchers use terms relating to accentedness and non-nativeness interchangeably, in that studies that have purported to address "accentedness" actually demanded judgments from listeners about "non-nativeness," or even mixed the terms on a single rating scale. For the purposes of the present work, "non-nativeness" is defined perceptually as:

- a listener's belief that a talker is not a native speaker of the target language, based on the percept of deviations from a pronunciation norm

The main difference between accentedness and non-nativeness is the target of the judgment: accentedness is an evaluation of the talker's speech, while non-nativeness is an evaluation of the talker. By the definitions proposed above, it is expected that judgments on these two scales should not differ substantially. Nonetheless, listeners in this work were asked about accentedness and non-nativeness separately so that this prediction could be evaluated.

Another way that a person's speech can differ from the native norm is by actually being a foreign language. "Foreignness" has an obvious cue that foreign accentedness lacks, in that the linguistic content of a foreign language is generally unintelligible to the listener. However, there may also be acoustic cues to foreignness, including sounds or subsegmental phonetic patterns that are not present in the language(s) known to the listener. Such cues likely come into play when dealing with small units of language for which semantic interpretation is difficult or impossible. Additionally, the segmental and subsegmental characteristics which cue that a short sample of speech is produced in a foreign language

may also be characteristics that are transferred into a talker's non-native productions of other languages, indicating then that the talker is not a native speaker.

Judgments of foreignness, like those of foreign accentedness, are based on only those properties of the signal that are detected by listeners: Weinrich (1986, 188) clearly states that "la xénité ne résulte pas forcément de l'altérité . . . . La xénité . . . est une interprétation de l'altérité" ["foreignness does not necessarily result from otherness . . . . Foreignness . . . is an interpretation of otherness"]. Focusing on the acoustic rather than the semantic cues, "foreignness" of language may be defined as follows:

- the percept of deviations from a pronunciation norm that a listener attributes to the talker targeting a different language

As a talker's non-native productions are often influenced by pronunciation patterns from his or her native language (Brière, 1966; Flege, 1987), it is possible that listeners may use similar acoustic properties to evaluate accentedness (when the talker's non-native language is targeted) and foreignness (when the talker's native language is targeted).

Not all foreign languages are necessarily equal in foreignness. Some differences in pronunciation may be more salient than others, depending on the languages and sounds involved. Additionally, a listener's beliefs and experiences may impact his or her evaluations of foreignness, in that some languages may be personally and/or culturally considered to be quite exotic (more foreign) and others more familiar (less foreign). The term "foreign language" is used in the present work to refer to languages other than English, the native language of the listeners, while "foreignness" is reserved for the judgments of listeners regarding whether productions sound like English.

Foreign accent matters because "people use it to make social evaluations, and these evaluations clearly affect both listeners and speakers" (Derwing and Munro, 2009, 488).

Listeners are central to evaluations about accentedness, non-nativeness, and foreignness, as defined above, but it is not clear how listeners arrive at their judgments. While in many cases "listeners" also have visual information about their interlocutors, in some situations, such as on the telephone, the acoustic signal provides the only direct information by which one individual may evaluate another (see Purnell et al., 1999). However, the acoustic properties that lead listeners to perceive such characteristics in speech have not been explored in great detail. The research presented here investigates the acoustic correlates of foreign accentedness, non-nativeness, and foreignness, as well as relationships between these percepts. The remainder of this chapter provides an overview of the research design (Section 1.2) after positioning it in the context of prior work (Section 1.1).

1.1 Previous studies

While many well-known studies of foreign accent perception have related listeners' responses to details about non-native talkers' language experiences (Flege et al., 1995; Oyama, 1976), to which listeners do not have direct access, some previous investigations have included detailed analyses of the speech signal itself. In this section, such studies are reviewed, with a focus, where possible, on subsegmental aspects of the signal that are evaluated using acoustic measures.

1.1.1 Accentedness

One common experimental design in the study of perceived foreign accent is to have participants rate the degree of accentedness in various auditory stimuli, and then to relate these ratings to properties measured in the stimuli. For instance, Major (1987) used this approach to study the relationship between perceived foreign accent and VOT. Isolated words with initial voiceless stops were produced by 53 L1 Brazilian Portuguese talkers and

7 L1 American English talkers, and rated on a continuous scale from "no foreign accent at all" to "very heavy foreign accent" by 10 L1 American English listeners. The correlation between perceived foreign accent ratings and VOT values was significant at each of the three places of articulation. Voiceless stops in Brazilian Portguese exhibit short-lag VOT values, rather than the long-lag values of American English. L2 English productions with shorter VOT values—presumably more strongly influenced by the non-native talkers' L1 patterns—were judged as sounding more accented.

Perceived foreign accent was also found to be related to VOT in the L2 English of L1 Japanese talkers by Riney and Takagi (1999). In this study, VOT was measured for the initial voiceless stops in isolated words produced by 11 L1 Japanese and 5 L1 American English talkers. To allow for the study of longitudinal development, each group was recorded twice, with 42 months between sessions for the L1 Japanese talkers and 2 weeks between sessions for the L1 American English talkers. Perceived foreign accent ratings for these talkers were taken from a previous study (Riney and Flege, 1998), which had used 5 L1 American English listeners and a 9-point rating scale from "strong foreign accent" to "no foreign accent." Although the ratings were based on different samples of speech from these 16 talkers—sentences rather than isolated words—the correlation between perceived foreign accent and VOT was significant for /p/ in the later recordings, and for /t/ in both sets of recordings. As in Major's (1987) findings, talkers with shorter VOTs were judged to sound more accented, as English VOT targets are longer than those for Japanese.

Of course, VOT is not the only acoustic property that has been linked to foreign accent perception. In an investigation by Shah (2002), 10 L1 American English listeners rated accentedness in multisyllabic word productions from 22 L1 Dominican Spanish talkers and 5 L1 American English talkers on a scale from 1 ("least accented") to 9 ("most

accented"). Separate analyses for each independent variable revealed evidence for relationships between accentedness and word duration, stressed to unstressed vowel duration ratio, and duration of flapped /t/, although the effects were often limited to a small subset of lexical items. Somewhat notably, no relationship was found between accentedness and VOT, perhaps because Shah's (2002) approach was nonparametric, while most researchers, including Major (1987) and Riney and Flege (1998), opt for parametric statistics.

Munro (1993) focused on the role of vowels in foreign accent perception, and considered multiple acoustic properties simultaneously. In his experiment, 5 trained L1 English linguists rated the degree of foreign accent in the front vowels /i, ɪ, eɪ, ɛ, aɐ/ on a 100-point scale. The vowels were presented in /bVt/ contexts and produced by 21 L1 Arabic and 2 L1 English talkers. Vowels in Syrian and Sudanese Arabic, the dialects represented by the talkers in this study, differ from English vowels in two main ways: many contrasts are cued primarily by length rather than quality, and diphthongization is relatively minimal. Each acoustic variable was quantified as the squared difference between the value for a particular stimulus and the mean over the values for 12 L1 American English productions of the same vowel, in order "to characterize how much each rated token differed acoustically from a good exemplar of the English vowel category which it represented" (58). When data for all vowels were pooled, stepwise multiple linear regression revealed that 43% of the variance in accentedness ratings was accounted for by F1 at the 30% timepoint of each vowel, F1 movement between the 30% and 70% timepoints, F2 movement between the same timepoints, and vowel identity; greater deviations from L1 American English acoustic values corresponded to higher degrees of perceived accentedness. Further analyses for each vowel separately found that for /i/, there were no significant predictors; for /ɪ, ɛ, æ/, only F1 played a role, accounting for between 28% and 34% of the variance in responses;

and for /eɪ/, F2 movement and vowel duration explained 78% of the variance. Another set of analyses included unsquared, signed differences for the acoustic measures as well as the squared differences used earlier. Although the optimal regression model for /eɪ/ did not change, models for other vowels were improved, and accounted for between 36% and 57% of variance in ratings. Significant predictors included both squared (/ɪ/) and signed (/ɛ, æ/) values of F1, both squared (/i/) and signed (/æ/) values of F1 movement, signed values of F2 movement (/ɪ/), and signed differences between F2 and F1 (/i/). Munro (1993) suggested that perhaps F1 was so often significant because it was the primary dimension that distinguished the target vowels from one another, such that deviations in F1 might have been perceived as segmental substitutions. He also highlighted that formant movement, not just static formant values, seemed to relate to listeners' accentedness ratings, presumably due to greater diphthongization in the L1 American English productions.

While most research in this vein involves non-native English, Wayland (1997) investigated multiple acoustic correlates of perceived foreign accent in Thai. The stimuli were quite constrained, consisting of the sequences /k^{h}aːu/ and /naː/ with each of Thai's five lexical tones. Three L1 Thai listeners rated productions from 6 L1 English and 2 L1 Thai talkers on a 5-point scale. The analysis was modeled on the one adopted by Munro (1993), and used squared difference values for the acoustic measures to capture the degree of deviation from native speech. For /k^{h}aːu/, stepwise multiple linear regression identified the f0 valley and F2 of the /u/ portion of the diphthong as significant predictors, together accounting for 38% of the variance in accentedness ratings. For /naː/, only the f0 valley was significant, with an r^2 value of 63%. Additional analyses for individual words (that is, particular tones on each of the two segmental sequences) revealed varying results, but

overall, spectral properties rather than temporal properties dominated the models. This pattern was supported by production data that indicated that while VOT and vowel duration differences between the native and non-native talkers did not reach significance, measures of tone realization and vowel quality did.

In each of the studies described so far, only one foreign accent was tested. By contrast, Munro and Derwing (2001) examined the effect of speaking rate on perceived foreign accent ratings for L2 English talkers of 12 different L1s. Sentences produced by 48 L2 English talkers and 4 L1 Canadian English talkers were rated on a 9-point scale by 44 L1 Canadian English listeners. Speaking rate accounted for 15% of the variance in perceived foreign accent ratings. An additional experiment was then performed to ensure that speaking rate was indeed responsible for, rather than coincidentally correlated with, the rating differences found. English sentences from 10 L1 Mandarin talkers, as well as 7 L1 Canadian English talkers, served as stimuli. These original sentences, as well as a set with a speaking rate increased by 10% and an additional set with a speaking rate decreased by 10%, were rated as in the first experiment by 26 L1 Canadian English listeners. Manipulations of speaking rate accounted for 6% of the variance in perceived foreign accent ratings, with accelerated speech rated as less accented than natural and slowed speech. As the effect of speaking rate remained significant when all other factors were held constant, it was concluded that speaking rate contributed causally, though modestly, to foreign accent perception. With sentence-length stimuli, however, speaking rate seems to indicate fluency, which reflects "the degree of fluidity in speech" (Derwing et al., 2009, 534), more than the deviations from language-specific norms focused on in the studies reviewed above.

McCullough (2013), in a small pilot version of the present dissertation, investigated the relationship between multiple acoustic properties and perceived foreign accent ratings of

multiple varieties of non-native English. Measures of VOT and vowel quality correlated with the degree of foreign accentedness perceived by 28 L1 American English listeners in the English productions of 16 talkers with L1 backgrounds of American English, Hindi, Korean, and Mandarin. Higher degrees of accentedness were associated with greater acoustic deviation from L1 American English productions. The stimuli were stop-vowel sequences, and as such had relatively few possible acoustic correlates. Despite these short stimuli, though, vowel quality and VOT accounted for only 38% of the variance in ratings of stimuli containing voiceless stops, and 26% of the variance in ratings of stimuli containing voiced stops, suggesting that additional acoustic properties may have affected listeners' evaluations.

Clearly, there is little consensus in the literature as to exactly which elements of the acoustic signal might influence the perception of foreign accent. A variety of properties have been associated with listeners' responses to some degree. In particular, VOT has shown a relationship to perceived foreign accent ratings in some studies (Major, 1987; McCullough, 2013; Riney and Takagi, 1999), but not in others (Shah, 2002; Wayland, 1997). A larger problem is that with the exceptions of McCullough (2013) and Munro and Derwing (2001), it is not clear that these studies investigated "foreign accent" generally as opposed to the more specific scales of "Brazilian Portuguese accent," "Japanese accent," "Spanish accent," or "Arabic accent". If the task demanded only comparisons between native talkers and L2 talkers from the same L1 background, listeners might have implemented the more specific scale regardless of the instructions given.

1.1.2 Non-nativeness

Another common experimental design involves asking participants for binary or scalar judgments about the native speaker status of talkers heard in auditory stimuli. In discussions of these data, such judgments are often assumed to be equivalent to judgments of accentedness of the talkers' speech. Indeed, some of the most widely-cited papers in the perceived foreign accent literature use this approach. Flege (1984), for instance, had listeners identify each production as "native" or "non-native." The endpoints of Flege et al.'s (1995) continuous rating scale were labeled with "native speaker of English—no foreign accent" at one end and "native speaker of Italian—strongest foreign accent" at the other, drawing on perceptions of accentedness and non-nativeness simultaneously. However, Cheong (2007) collected ratings of "accentedness" and "nativeness" on 10-point scales and found that listeners "judged the concept of 'nativeness' more strictly as opposed to 'accentedness"' (158), in that for non-native talkers their ratings of nativeness were closer to the endpoint than those for accentedness. If non-nativeness shows different response patterns from accentedness, then it may also show different relationships to acoustic properties. In this section, studies investigating non-nativeness are reviewed, although it should be noted that many of them actually use "accentedness" or a similar term in discussing their research, and only a careful reading of the experimental design details lands them here.

The series of experiments described by Baker et al. (2011) included a rating task, with a 9-point scale from "native" to "foreign." One instance of this task involved paragraphs read by 13 L1 English talkers and 52 non-native talkers, primarily with L1 Chinese and L1 Korean backgrounds, rated by 50 L1 English listeners. In another instance, 15 L1 English listeners rated single intonational phrases extracted from spontaneous speech by 8 L1 English talkers, 18 L1 Chinese talkers, and 16 L1 Korean talkers. For both sets of

responses, within-speaker word duration variance was correlated with nativeness ratings, with greater degrees of variance sounding more native. For responses to stimuli from spontaneous speech, two additional characteristics correlated with greater perceived nativeness: greater similarity to mean durations by native talkers, and shorter function words as compared to content words. The relatively long stimuli in these experiments allowed for the use of rather complex acoustic measures, which again seemed to capture more about overall fluency in the L2 than about transfer from a talker's native language into the L2.

In Alba-Salas's (2004) study, English stop tokens were excised from word-initial contexts, from the release of the burst to the onset of periodicity of the following vowel. Listeners identified each production as "definitely native", "possibly native", "possibly foreign", or "definitely foreign." The L1 backgrounds of the 6 non-native talkers included Venezuelan, Puerto Rican, and Ecuadorian Spanish, and 6 L1 American English talkers were also included. There were 8 L1 American English listeners, half of whom had no knowledge of Spanish and half of whom were L2 speakers of Spanish. In the analyses, responses were made binary, with "definitely native" and "possibly native" collapsed to a single category, and likewise for "possibly foreign" and "definitely foreign." VOT correlated with responses from the listeners who spoke Spanish, accounting for 18% of the variance, with "native" responses increasing as VOT increased. VOT was unrelated to responses from the monolingual listeners. Thus, experience with another language, and specifically the native language of the non-native English talkers, seemed to influence the way listeners perceived non-nativeness.

Tsukada (1998) used as stimuli isolated CVt and CVd words produced by 6 L1 Australian English talkers and 14 L1 Japanese talkers. The words were presented in pairs consisting of one L1 Australian English production and one L1 Japanese production matched

for midpoint F1, midpoint F2, and vowel duration. The task of the 26 listeners, most of whom were L1 Australian English speakers, was to identify which stimulus in each pair had been produced by a native English talker. Listeners were correct 78% of the time, suggesting that there are acoustic indicators of nativeness beyond the three controlled for in this study. The author hypothesized that these indicators might involve coarticulation between consonants and vowels, and perhaps f0. Another possibility could be dynamic rather than static information about vowels (Munro, 1993).

Again, most studies of non-nativeness use English as the target language. Bond et al. (2008), however, collected responses to Latvian sentences from three groups: 28 native speakers of Latvian, 12 native speakers of Russian with some knowledge of Latvian, and 31 monolingual English-speaking Americans. The sentences had been extracted from a passage read by 10 native talkers and 10 L1 Russian talkers, and thus had consistent targets. The responses were of two types: categorical judgments of the talker as "native" or "non-native" (for Americans) or "Latvian" or "Russian" (for the other groups), and ratings on a 7-point scale from "definitely a native speaker" to "definitely not a native speaker." Generally, all groups of listeners distinguished between native talkers and low-proficiency non-native talkers in both types of response; listeners with knowledge of Latvian also distinguished between native talkers and high-proficiency non-native talkers. The relationship between acoustics and perception was explored for the categorical responses from American listeners. It was assumed that because they had no specific knowledge about the target language, Americans must have used some general cue to fluency. Utterance duration, which indexes hesitation, repetition, pausing, and other effects of low fluency, accounted for 42% of the variance in the responses, with longer utterances more likely to be judged as non-native. Again, however, a measure of fluency is somewhat different from acoustic

properties that can directly quantify phonetic transfer from the L1, such as the consonant- and vowel-related properties used in other studies.

Explicit decisions about the native speaker status of talkers, presented as continuous (Baker et al., 2011; Bond et al., 2008) or binary (Alba-Salas, 2004; Bond et al., 2008; Tsukada, 1998) responses, are common in previous work. Additionally, listeners' responses regarding talker non-nativeness have been shown to relate to a varied set of acoustic properties, on the basis of both rather short (Alba-Salas, 2004; Tsukada, 1998) and rather long (Baker et al., 2011; Bond et al., 2008) stimuli. Potential differences between responses about accentedness and responses about non-nativeness, though, are rarely considered.

1.1.3 Foreignness

The elicitation of the rating responses discussed above generally assumes that the target language is known and understood by the listener (but see Bond et al., 2008; Major, 2007). However, comprehension is not required for some types of judgments. Even in the absence of meaning, some sequences—perhaps containing particularly unfamiliar sounds and/or acoustic patterns—may sound more foreign to listeners than others. Another type of investigation, then, involves ratings of foreignness, generally operationalized as perceived distance from a specified language or language variety.

Two of the four experiments performed by Flege and Munro (1994) are particularly relevant to this topic. In the first of these experiments, recordings of the word *taco* in English and Spanish from 14 English-Spanish bilinguals, 7 English-speaking monolinguals, and 7 Spanish-speaking monolinguals were played for a small group of 3 phonetically trained listeners with L1s of English and German. These listeners first identified whether they thought the target language of each production was English or Spanish, and also used a

response lever to indicate a rating from "good example of English *taco*" to "good example of Spanish *taco*." For simplicity, only the rating data are addressed in the present summary. A variety of temporal and spectral cues were measured in all four segments of the word. It was found using stepwise linear regression that the VOT of /t/ alone accounted for 87% of the variance in ratings, with VOT of /k/, midpoint F3 of the first vowel, F2 at 20% of the way through the second vowel, F1 at 80% of the way through the second vowel, and intensity of the second vowel contributing small amounts to the ultimate model, which had an overall r^2 value of 0.97. Another experiment with manipulated productions of *taco* varied VOT of /t/, spectral quality of the first vowel, and duration of the second vowel to test the contributions of these acoustic properties independently of one another. One group of 15 monolingual American English speakers identified the target language of each stimulus as English or Spanish, while another group rated each stimulus on a 9-point scale from "least English-like" to "most English-like." Again, for simplicity, only the rating results are presently considered. Stepwise linear regression identified spectral quality of the first vowel as the best correlate, accounting for 25% of the variance in ratings. VOT of /t/ increased the r^2 value to 0.41, and duration of the second vowel further increased it to 0.57. Results from the identification tasks, although not reviewed in detail here, were largely consistent with those from the rating tasks.

Bradlow et al. (2010) had 23 L1 American English listeners rate the overall phonetic similarity of 17 languages to English based on roughly 2-second excerpts from read speech. Rather than explicitly assigning a numerical rating, listeners positioned stimuli on a "ladder" with English situated at the bottom, such that a larger number of "rungs" up the ladder reflected a greater distance from English. Multiple stimuli could be positioned on a single "rung" to indicate identical distances. Analysis revealed that the language rated as closest

to English was Dutch, while the most distant language from English was Cantonese. In a free classification experiment using the same 2-second stimuli, a different set of 25 L1 American English listeners had grouped together languages that sounded similar. Listeners' ratings of distance from English were found to correlate with one of the two dimensions of a multidimensional scaling analysis of the free classification data, indicating that rating and free classification responses were related to some degree. This dimension seemed to separate "Eastern" from "Western" languages, although the mapping between this geographical distinction and the speech signal was evidently complex, as no single direct phonetic correlate could be identified.

Magen (1998) examined the effect of a variety of acoustic cues on the perception of natural and manipulated L2 English. A single L1 Spanish talker recorded 32 English sentences. Three copies of each sentence were digitally manipulated by one factor each in order to be more native-like, as determined by acoustic comparison to a native production of the same sentence. 10 L1 American English listeners rated these 128 stimuli on a 7-point scale from "closer to native English" to "less close to native English." While the target language of the stimuli was not in question, this was not defined as a rating of talker nativeness, but as the proximity of each stimulus to a particular language variety. The process was repeated with a second talker and a second set of 10 listeners, using 24 sentences and 2 manipulated repetitions of each for a total of 72 stimuli. A total of 10 factors were manipulated. Manipulations which deleted an epenthetic schwa, deaffricated a target fricative, and inserted a deleted word-final /s/ were found to shift ratings for both talkers toward the "closer to native English" end of the scale.

While the studies discussed above demanded comparisons to English, one experiment described by Bond and Stockmal (2002) asked L1 American English listeners to rate the

degree of similarity between auditory stimuli and an unknown language, Korean. After 10 minutes of exposure to spoken Korean, listeners heard 5-second samples of Korean and other languages and rated the degree of similarity to Korean on a 7-point scale from "very different" to "very similar or identical." A group of 41 listeners heard languages described as "rhythm class competitors," with syllable-based timing. Novel samples of Korean were judged to be most similar to Korean, with Japanese, another East Asian language, also rated quite highly; Tagalog, Latvian, and Mbawa were judged to sound more different. A separate group of 20 listeners heard "geographical region competitors." Korean again received the highest ratings, with Indonesian and Mandarin judged to be more similar to Korean than were Japanese and Tagalog. Two languages, Japanese and Tagalog, were included as competitors for both groups. While it is clear that Tagalog was not perceived as similar to Korean, the conflicting results regarding Japanese were not explained. However, the similar results of these two languages in the "geographical region competitors" condition were suggested to be talker-specific, as the Tagalog and Japanese talkers were both relatively dramatic. It was not clear whether the Tagalog and Japanese talkers in the "rhythm class competitors" condition were also dramatic; different 5-second samples of each language were used in the two conditions, but the authors did not specify whether they were taken from recordings of the same talkers.

Overall, in addition to accentedness and non-nativeness, listeners are able to rate the foreignness of speech samples, especially foreignness relative to their own native language. Such ratings have previously been related to segmental (Flege and Munro, 1994; Magen, 1998) and structural (Magen, 1998) properties. Additionally, listeners' responses

suggested that they may have perceived some characteristic of the acoustic signal that differentiated "Eastern" languages from "Western" ones (Bond and Stockmal, 2002; Bradlow et al., 2010), although the details of this characteristic are unknown.

1.1.4 Classification by (native) language

Production studies have shown that foreign language pronunciation is influenced greatly, and in somewhat predictable ways, by a talker's native language (Brière, 1966; Flege, 1987). It might be expected that such influence carries over to perception. In this vein, Bond et al. (2003) argued that the use of a label such as "Japanese accent" indicates some perceptual commonality between foreign-accented speech and a foreign-accented speaker's native language. That is, more than just knowing what a "Japanese accent" is, listeners perceive some "acoustic signature" that allows them to link "Japanese accent" and "Japanese" directly. However, no evidence was found for this proposal. English-speaking listeners failed to match the L2 English of an L1 Japanese talker to spoken Japanese, and in three separate versions of the experiment, failed to match the L2 English of an L1 Latvian talker to spoken Latvian. While it is true that labels like "Japanese accent" are often used, these results seem to suggest that these labels might derive not from a perceptible "acoustic signature," but from real-life experience with people and language. In other words, perhaps English listeners learn explicitly that some speakers have Japanese as a native language, and then learn to call the English productions of these speakers "Japanese-accented."

A related question is the generalizability of such labels: whether listeners can match a non-native production from one speaker to a non-native production from another speaker with the same native language. For instance, while there is no evidence for some perceptibly "Japanese" quality that characterizes both the L2 English of L1 Japanese speakers and

Japanese itself, there may be some perceptibly "Japanese-accented" quality that characterizes all L2 English of L1 Japanese speakers to the exclusion of other varieties of English. If such a quality exists, English listeners might be able to match L2 English samples from different L1 Japanese speakers, with or without the explicit label of "Japanese accent." As in the rating tasks discussed above, listeners must use properties of the acoustic signal in such classification. Regardless of whether listeners' decisions match speakers' actual language backgrounds, a task that demands abstraction over multiple talkers of different non-native backgrounds might provide information about the acoustic properties that listeners use to classify different varieties of foreign accent.

Some previous studies have focused on the accuracy of such classification with the use of explicit labels. For instance, Derwing and Munro (1997) asked 26 L1 Canadian English listeners to choose whether phrase-length stimuli exhibited a Cantonese, Japanese, Polish, or Spanish accent. The stimuli were taken from recordings of guided storytelling by 12 talkers from each of the 4 L1 backgrounds. Listener performance, while consistently above chance, varied by the talkers' L1 background, ranging from 41% correct for L1 Japanese talkers to 63% correct for L1 Cantonese talkers. Examination of errors indicated confusion between the two Asian languages (Cantonese and Japanese), as well as confusion between the two European languages (Polish and Spanish).

Vieru et al. (2011) used talkers who targeted French as a foreign language, rather than English. In one experiment, 25 L1 French listeners heard 10-second excerpts of spontaneous speech from native speakers of Arabic, English, German, Italian, Portuguese, and Spanish, and had to choose each talker's L1 from these six possibilities. Overall performance was 52%, with accuracy rates ranging from 25% for L1 Portuguese talkers to 77%

for L1 Arabic talkers. In a second experiment, a similar group of 25 listeners heard 1-minute excerpts of read speech from the same non-native language backgrounds as well as from native French talkers, and chose each talker's L1 from these seven possibilities. Accuracy for all groups combined was 60%, largely because of near-perfect identification of L1 French talkers. Performance on only the non-native talkers was 54% accurate, ranging from a low of 34%, again for L1 Portuguese talkers, up to 73% for L1 English talkers. In both experiments, L1 Spanish and L1 Italian talkers were often confused with one another, as were L1 English and L1 German talkers.

Other studies have explored the accuracy of listeners' perceptual abilities when dealing with foreign languages rather than foreign accents. For instance, Bond and Fokes (1991) played 2-second samples of Arabic, Chinese, English, Japanese, and Spanish, produced by 2 native talkers of each language, in quiet and in noise for various groups of listeners. Similar to the studies described above, the task was forced-choice identification of the target language from the five possibilities. Performance by 14 L1 English listeners ranged from 73% correct (Spanish) to 100% correct (English) in quiet, and 31% correct (Japanese) to 79% correct (English) in noise. A separate group of 13 L1 English instructors in the Ohio Intensive English Program, presumably with a reasonable amount of exposure to foreign languages via their profession, had generally higher rates of correct identification.

Vasilescu et al. (2005) extracted filler words (such as English *uh* and *um*) from native recordings of Arabic, English, French, German, Italian, Mandarin, Portuguese, and Spanish. In a two-alternative forced-choice task, 20 L1 French listeners were asked whether each production was extracted from a recording of French or from a recording of one of the other languages. Similarly, 22 L1 French listeners were asked whether each production was extracted from a recording of Portuguese or from a recording of one of the other languages.

The first group of listeners were 75% accurate overall, and were above chance performance for each language. The second group of listeners were 70% accurate, and failed to perform above chance level for Arabic, German, and Italian stimuli. These results were taken to reflect a "mother tongue bias," in that performance was better when one response option was the native language of the listeners.

Besides the rating experiment described above, other experiments by Bond and Stockmal (2002) explored the ability of L1 American English listeners to correctly identify languages as "Korean" or "not Korean." As mentioned above, listeners in a test group heard spoken Korean for 10 minutes prior to completing the task, while listeners in a control group did not. The control listeners were basing judgments merely on the label "Korean," and in fact may have never heard Korean spoken. For the "rhythm class competitors" experiment, listeners in the control group frequently mistook Japanese and Tagalog (but not Latvian or Mbawa) for Korean, suggesting that some aspect of the acoustic signal may have distinguished Asian languages from the others. As further evidence of this possibility, in the "geographical region competitors" experiment, control listeners performed poorly, as all the languages were from Asia and thus the unidentified "Asian" property of the acoustic signal was of no use in completing the task.

Some related studies use discrimination tasks rather than explicit classification. The 16 monolingual British English-speaking listeners in Lorch and Meara's (1995) experiment heard 26 pairs of 2-second samples of speech produced by 2 Farsi talkers and 2 Greek talkers. The listeners indicated whether the languages in each pair of stimuli were the same or different, and after hearing all 26 pairs they were asked to identify the target languages. The entire procedure was then repeated. Mean accuracy was 63% accuracy for the first block and 65% for the second, although not all listeners performed above chance, and not all

listeners showed improved performance in the second block. The 5 listeners who correctly identified Greek as one of the stimulus languages did not show more accurate discrimination than other listeners. Although there were no correct identifications of Farsi, 7 listeners guessed that it was Arabic, which was characterized as a "closely related [language] from the cultural and geographic perspective" (68).

Stockmal et al. (1994) investigated foreign language discrimination abilities in children, but the present summary focuses primarily on the performance of their adult control group. As in the study by Bond and Fokes (1991) described above, the stimuli were 2-second samples of Arabic, Chinese, English, Japanese, and Spanish, spoken by 2 native talkers of each language. The stimuli were played in pairs for 38 adult listeners, who stated whether the two samples were in the same language or different ones. Some same-language pairs involved differences in the talker and/or the target phrase. Pairs involving English were excluded from the analysis due to listeners' nearly perfect performance. On the remaining pairs, adults answered correctly between 58% of the time (for same language, different talker, different phrase) and 97% of the time (for different languages, and thus different talkers and different phrases). Same-language trials involving Arabic and Spanish tended to challenge listeners. Additionally, listeners found it difficult to discriminate between Chinese and Japanese. A group of 12 7- and 8-year old children performed similarly to the adult listeners, but 13 4- and 5-year-olds only approached adult performance on different-language trials.

In studies like Stockmal et al.'s (1994), language discrimination may be confounded with talker discrimination, as samples of different languages are typically produced by different talkers. Stockmal et al. (2000) removed this confound by using 5-second excerpts of both languages spoken by bilingual talkers. The 4 male talkers were Arabic-French,

Hebrew-German, Akan-Swahili, and Latvian-Russian bilinguals, while the 4 female talkers were Korean-Japanese, Ombawa-French, Latvian-Russian, and Ilocano-Tagalog bilinguals. In two experiments, 131 L1 American English listeners heard pairs of stimuli produced by the same talker and determined whether the same language was targeted in both cases. Performance was generally above chance, indicating that listeners can discriminate between foreign languages even when they are produced by the same talker. However, some language combinations were more difficult than others, and stimuli produced by the female Latvian-Russian and Ilocano-Tagalog bilinguals were not successfully discriminated.

Relatively little attention has been given to the phonetic properties that listeners may use in tasks with foreign language stimuli, and relevant accounts tend to be based on impressionistic observations rather than on acoustic measures. The adult listeners in Stockmal et al.'s (1994) study, when asked how they completed the discrimination task, often mentioned segments. However, Stockmal et al. (2000) had an additional group of 52 listeners rate the degree of similarity between different-language samples from 4 of their bilingual talkers, and interpreted these data as suggesting that listeners' judgments were influenced by patterns of pitch and rhythm.

In production studies, the influence of a talker's native language on his or her non-native speech is often quite clear. It may seem surprising, then, that the listeners described by Bond et al. (2003) were unable to link "Japanese accent" to Japanese itself, especially as listeners in other investigations have correctly classified foreign languages (Bond and Fokes, 1991; Vasilescu et al., 2005) and the native language backgrounds of non-native talkers (Derwing and Munro, 1997; Vieru et al., 2011). What is not apparent, however, is whether listeners use the same properties of the speech signal to evaluate foreign and

non-native speech. The phonetic influence of a talker's native language on his or her non-native productions, though often noted in production-based investigations, can only serve as the perceptual link that Bond et al. (2003) imagine if listeners attend to similar cues in both types of speech. It is possible that the percepts of accentedness and non-nativeness might be based largely on general characteristics associated with low fluency, as in studies by Baker et al. (2011), Bond et al. (2008), and Munro and Derwing (2001), rather than on characteristics related in any specific way to a non-native talker's L1.

In sum, there is only preliminary evidence regarding which acoustic properties might influence perceived foreign accent ratings when multiple varieties of non-native speech are used, and how framing the phenomenon of interest as non-nativeness versus accentedness influences listeners' rating responses. Additionally, it is not clear whether listeners can accurately classify multiple varieties of foreign accent, which acoustic properties might influence such classification, or whether listeners generally perceive foreign accents and foreign languages in similar ways. This final point helps to focus the selection of acoustic properties in this dissertation to those that might reflect phonetic transfer from a talker's native language. As lack of fluency cannot be clearly linked to the influence of a particular language, it is not investigated in detail here; the stimuli are designed to focus on acoustic properties that might reflect transfer from the L1 rather than fluency in the L2.

1.2 The present study

The primary research questions of the present work are as follows:

1. Rating

    (a) What acoustic properties correlate with ratings of the degrees of foreign accentedness, non-nativeness, and foreignness?

(b) Are there relationships between these ratings, and/or between their acoustic correlates?

2. Free classification

   (a) Do free classifications of talkers' native languages align with their actual language backgrounds?

   (b) What acoustic properties correlate with free classifications of talkers' language backgrounds?

   (c) Is there a relationship between the acoustic correlates of free classifications and the acoustic correlates of ratings?

The non-native English talkers in the present investigation were native talkers of Mandarin, Hindi, Korean, and Spanish. Nearly half of the international students enrolled at United States institutions are from China, India, or South Korea (Institute of International Education, 2012), so undergraduate listeners are likely to have interacted with individuals of the first three language backgrounds. Substantially fewer international students are from Spanish-speaking countries, but as Spanish is commonly spoken in the United States, undergraduate listeners are likely to have interacted with native Spanish speakers in a variety of settings inside and outside the university. Pilot research (McCullough, 2013) indicated that American listeners, many of whom had themselves studied Spanish in academic settings, often reported hearing stimuli from L1 Spanish talkers when the only language backgrounds actually represented were Mandarin, Hindi, and Korean. Including stimuli from L1 Spanish talkers thus allowed for exploration of a variety of non-native English that listeners seemed to think they knew something about. Additionally, while Spanish may be stigmatized by some Americans due to social and political issues within the United States,

Spanish is "Western" rather than "Eastern," an important contrast in light of some of the results discussed above (Bond and Stockmal, 2002; Bradlow et al., 2010).

In addition to the largely segmental acoustic properties suggested by the findings of previous studies, as described above, there is evidence that prosodic and global temporal properties also influence perception of non-native speech (Anderson-Hsieh et al., 1992; Boula de Mareuil and Vieru-Dimilescu, 2006; Kang, 2010; Munro, 1995; Munro et al., 2010; van Els and de Bot, 1987). However, longer stimuli, as compared to shorter ones, have more acoustic details available to potentially influence listeners' responses, including some relating to fluency. In the present work, the use of rather short stimuli—syllables and words—allowed for relatively thorough investigation of segmental cues in the acoustic signal, and for closer focus on the aspects of foreign accent perception that may be directly linked to foreign language perception.

Each of the six experiments in this work involved both a rating and a classification task. Differences among the experiments are summarized in Table 1.1. Each pair of experiments explored listeners' ratings on a different scale, with English-language stimuli for scales of foreign accentedness and non-nativeness, and stimuli in multiple languages for the foreignness scale. Stimuli in some experiments were syllables extracted from words, while in other experiments they were whole words. While the classification task itself did not change across experiments, the classification stimuli did, as they were a subset of the stimuli used in the rating task.

For the sake of brevity, the three scales will generally be referred to, respectively, as "accentedness," "non-nativeness" and "non-Englishness." "Englishness" is chosen rather than "foreignness" to reflect the specific instructions given to the listeners, and "non-nativeness"

Experiment	Stimulus language	Stimulus length	Rating scale
1	English	syllables	foreign accentedness
2	English	words	foreign accentedness
3	English	syllables	certainty that talker is native
4	English	words	certainty that talker is native
5	native languages	syllables	certainty that stimulus is English
6	native languages	words	certainty that stimulus is English

Table 1.1: Overview of experiments

and "non-Englishness" (rather than the simpler "nativeness" and "Englishness") so that ratings described as "higher" align with speech from the same types of talkers on all three scales (i.e., non-native talkers of English). Results from the six experiments will be discussed in Chapters 4 through 7, after a review of previously reported (Chapter 2) and currently observed (Chapter 3) production patterns, which guided choices about the acoustic properties used in the analyses.

CHAPTER 2: LANGUAGE VARIETIES

The stimuli that will be described in Chapter 3 involve combinations of stops and vowels. In this chapter, to better inform the selection of acoustic measures examined in the present analyses, the realizations of stops and vowels in the relevant language varieties are examined. Section 2.1 reviews the stops and vowels of American English, Hindi, Korean, Mandarin, and Spanish, with particular focus on the phonetic realization of the stop contrast(s). Allophonic variation in the pronunciation of these sounds is also discussed. Section 2.2 addresses stop- and vowel-related pronunciation patterns noted in previous studies of the relevant varieties of non-native English. In Section 2.3, the acoustic properties that seem to capture differences among these language varieties are summarized.

2.1 Languages

2.1.1 American English

Phonologically, American English is generally described as having a voiced versus voiceless stop contrast at each of three places of articulation: bilabial, alveolar, and velar (Ladefoged, 1999). In word-initial position, this contrast tends to be realized phonetically as unaspirated versus aspirated. That is, phonologically voiceless stops are produced with long lag VOT, and phonologically voiced stops are produced with short lag VOT, although some speakers do exhibit lead voicing for the latter category (Lisker and Abramson, 1964). "Voiced" and "voiceless" tend to be better descriptors of the contrast in word-medial position. For instance, between voiced sounds, /b, d, g/ often exhibit voicing during closure,

and word-medial /p, t, k/ generally exhibit short lag VOT unless they begin a stressed syllable. Intervocalic alveolar stops are typically realized as voiced flaps before an unstressed vowel (Ladefoged, 1999).

The vowels of American English are /i, ɪ, eɪ, ɛ, æ, ʌ, ə, u, ʊ, oʊ, ɔ, ɑ, ɚ, aɪ, aʊ, ɔɪ/ (Hillenbrand et al., 1995; Ladefoged, 1999). In many dialects, /ɔ, ɑ/ have merged (Ash, 2003).

2.1.2 Hindi

The sixteen stops of Hindi are described phonologically as all combinations of voicing and aspiration at four places of articulation: bilabial, dental, retroflex, and velar (Ohala, 1999). Lisker and Abramson (1964) found that word-initially, unaspirated and aspirated voiced stops exhibited lead voicing, while VOT values for voiceless unaspirated stops were in the short lag range, and for voiceless aspirated stops they were in the long lag range. Kagaya and Hirose (1975) observed similar patterns for both word-initial and word-medial productions. Dutta (2007) showed that the four-way stop contrast involves a number of acoustic differences in addition to voicing, including f0 and spectral tilt measured at the beginning of the following vowel. Fundamental frequency revealed a three-way contrast, with the lowest values for voiced aspirated stops, higher values for voiced unaspirated stops, and higher values still for unaspirated and aspirated voiceless stops. Spectral tilt showed a two-way contrast, with voiced aspirated stops being breathier than voiced and voiceless unaspirated stops. /p^h/ may be pronounced as [f] (Sandahl, 2000).

Most sources claim that Hindi has ten oral vowels: /i, ɪ, e, u, ʊ, o, ɔ, ɑ, ə/, and /ɛ/ (Khan et al., 1994) or /æ/ (Bansal, 1981; Sandahl, 2000). The differing labels are presumably due to varying pronunciations across dialects (see Shapiro, 2003). Because /æ/ can be used to

represent an additional vowel that appears in loanwords from English (Ohala, 1999), /ɛ/ was chosen to represent the tenth vowel category in the present work. There are also nasal versions of all vowels except /æ/ (Ohala, 1999). The distinction between /i/ and /ɪ/, and between /u/ and /ʊ/, is neutralized to the long vowel word-finally (Shapiro, 2003). Ohala (1999) suggested that /ə/ is often pronounced as [ɐ].

2.1.3 Korean

Phonologically, Korean exhibits three types of stops at each of three places of articulation: bilabial, alveolar, and velar (Lee, 1999). These stops are commonly described as tense/fortis (/p*, t*, k*/), lax/lenis (/p, t, k/), and aspirated (/p^h, t^h, k^h/). The phonetic details of this contrast have been the target of many investigations. In the past, VOT served as the primary distinction among these categories. Word-initially, tense stops were clearly characterized by short lag VOT values, and aspirated stops by long lag VOT values; lax stops had intermediate VOT values of 17-62ms (Han and Weitzman, 1970; Lisker and Abramson, 1964) that did not clearly fit into the short or long lag categories. Recently, for younger speakers, VOT values for lax and aspirated stops have nearly merged in the long lag range. However, f0 following a lax stop is lower than for a tense or aspirated stop (Cho et al., 2002; Han and Weitzman, 1970; Kang and Guion, 2008; Silva, 2006). Thus, the current stop contrast simultaneously involves two acoustic properties: VOT differentiates tense stops from the others, and f0 differentiates lax stops from the others. Silva (2006) suggests that this f0 cue had long existed, but was redundant when the VOT differences were clear. As VOT became less reliable, in speakers born after 1965, f0 became more crucial. Spectral tilt seems to serve as another indicator of the contrast, in that the vowel following a tense stop is characterized by pressed voice, while the vowel following a lax stop is characterized by

breathy voice (Cho et al., 2002). Lax stops are voiced intervocalically. All coda stops are unreleased (Sohn, 1999).

Modern Korean has the monophthongs /i, e, ɛ, ɨ, ʌ, a, u, o/, and a number of diphthongs with onglides (Lee, 1999). For many speakers, however, /e, ɛ/ are no longer distinct, and appear in free variation (Sohn, 1999), more often as [e] (Choo and O'Grady, 2003). In the present work, no distinction is assumed, and /e/ is used to represent the combined vowel group. While /y, ø/ may be listed as monophthongs, they are produced as diphthongs [wi, we], respectively, in many dialects, including the Seoul-based standard (Lee, 1999; Sohn, 1999). Traditional accounts of Korean also mention a vowel length contrast, but this distinction is being lost in speakers born after the mid-20th century (Magen and Blumstein, 1993; Sohn, 1999).

2.1.4 Mandarin

Mandarin contrasts unaspirated and aspirated stops at each of three places of articulation: bilabial, denti-alveolar, and velar (Lee and Zee, 2003). Phonetic support of this description has been reported by Liu et al. (2000), who found that /p, t, k/ were produced with short lag VOT and /p^h, t^h, k^h/ with long lag VOT. Liu et al. (2007) reported similar results for /p, t/ and /p^h, t^h, k^h/.

Mandarin uncontroversially has the monophthongs /i, y, a, u/, although there is some disagreement about the remainder of the inventory. Lee and Zee (2003) identified two additional monophthongs, /ə, ɤ/, while Chin (2006) listed only /ə/, with [ɤ] as an allophone thereof in free variation with [ə]; the latter view is adopted in this work. /i/ has allophones [i, ɪ] in free variation. Like English, Mandarin also contains diphthongs /eɪ, oʊ, aɪ, aʊ/

(Chin, 2006), and a number of additional diphthongs with onglides rather than offglides, as well as several triphthongs (Lee and Zee, 2003).

Mandarin differs notably from the other languages considered in this work in that it employs lexical tone. The four tones of Mandarin are, respectively, high level (˥), high rising (˧˥), low rising or "dipping" (˨˩˦), and falling (˥˩). Syllables may also be characterized by unmarked neutral tone, and unaspirated voiceless stops may be voiced as the onsets of such syllables (Norman, 1988). Tone 1 (high level) is realized as tone 1 in isolation, tone 2 (high rising) if followed by tone 4, and tone 4 (falling) if followed by tone 1, 2, or 3. Tone 4 (falling) is produced as tone 2 (high rising) if followed by another instance of tone 4. Similarly, tone 3 (low rising) is pronounced as tone 2 (high rising) if followed by another instance of tone 3 (Sun, 2006). These realization patterns are referred to as "tone sandhi," and except for the highly phonologized tone 3 conventions, generally result from phonetic coarticulation. Related patterns pertaining to sequences of three syllables are not described here, as the longest stimuli in the present investigation were disyllabic.

2.1.5 Spanish

The present work involves several American varieties of Spanish. Spanish contrasts voiced and voiceless stops at three places of articulation: bilabial, dental, and velar (Dalbor, 1969). Lisker and Abramson (1964) found lead VOT values for voiced stops and short lag VOT values for voiceless stops in the productions of two speakers of Puerto Rican Spanish. Traditional accounts claim that the voiced stops /b, d, ɡ/ are generally realized as fricatives (spirants) [β, ð, ɣ], except following pauses and nasals, where they are [b, d, ɡ]; [d] also occurs after /l/. The reality of the situation, however, is somewhat muddier. Macken and Barton (1980) reported that adult Mexican Spanish speakers produced the

fricative allophones for 30% to 40% of stop targets following a pause. Furthermore, the allophones [b, d, ɡ] can occur more widely in careful speech, especially in word-initial position (Dalbor, 1969; Harris, 1969). Lewis (2001) found that word-medial voiceless stops were phonetically reduced in various ways, although these effects were more evident for relatively casual speaking styles than for words read in isolation.

Spanish has five monophthongs: /i, e, a, u, o/. [ɛ] appears as an allophone of /e/, often in closed syllables. Although the language is pronounced with numerous diphthongs and triphthongs, they are analyzed as fused sequences of monophthongs rather than as separate vowels, and occur across word boundaries as well as within words (Dalbor, 1969).

2.2 Non-native varieties of English

The discussion in this section focuses on pronunciation patterns that are relevant for the stimuli described in Chapter 3, and most notably omits details about /ɑ, ɔ/, which were not included among the stimuli.

2.2.1 L1 Hindi/L2 English

The status of English in India is different from that in the other countries of origin of the non-native talkers, as English is an official language of India, and educated Indians regularly use English to communicate with one another (Gargesh, 2004). English is typically acquired in a school context, beginning in primary school (Wiltshire and Harnsberger, 2006), and is the principal language of higher education in India (Gargesh, 2004). Indian English is also different from most second language learning situations in that the target is not the native language of some group of individuals, but a "nativized variety of the second language system" (Wiltshire and Harnsberger, 2006, 91); that is, Indian English is itself the target language. As in any second language situation, the pronunciation of Indian English

depends on a speaker's L1 background (Gargesh, 2004; Wiltshire and Harnsberger, 2006). The discussion below focuses on previous studies about L1 Hindi speakers whenever possible.

Awan and Stine (2011) found that in word-initial position, speakers of Indian English from a variety of L1 backgrounds including Hindi had VOT values of 33ms for /p/ and 40ms for /t/, as opposed to 69ms and 77ms, respectively, for American English speakers. Word-medially, Indian English speakers had VOT values of 34ms for /p/ and 39ms for /t/, compared to 67ms and 87ms for American English speakers. Similarly, Davis and Beckman (1983) reported that L1 Hindi speakers produced English voiceless stop targets with short lag VOT, and additionally found that most English voiced stop targets were produced with lead voicing. In Indian English, /d, t/ tend to be retroflex rather than alveolar (Vidyalankar, 2002).

Maxwell and Fletcher (2009) found that L1 Hindi speakers produced /eɪ, oʊ/ as monophthongs [e, o]. They tended to produce [ɐ] for /ʌ, ə/, the latter of which was noted by Ohala (1999) for Hindi. /ɚ/ overlapped spectrally with /ʌ/ for some speakers.

2.2.2 L1 Korean/L2 English

Kang and Guion (2006) reported acoustic details of "late" Korean-English bilinguals, who began learning English between ages 15 and 34. In word-initial position, they had mean VOT values of 19ms for voiced stops and 86ms for voiceless stops in English, compared to 14ms and 72ms, respectively, for American English monolinguals. Similarly, Schirra (2012) found longer VOTs for voiceless stops produced by two L1 Korean speakers than for those produced by a native speaker. Kang and Guion (2006) also measured

additional cues, and found that late Korean-English bilinguals differed from native speakers on voice quality for both categories of stops, and on f0 measures in the following vowel.

Schirra (2012) reported that when considering F1 and F2, L1 Korean speakers' productions of /i/ overlapped almost entirely with those of /ɪ/, while native productions of these vowels were somewhat more distinct. Furthermore, the degree of overlap was found to correlate with subjective accentedness judgments provided by native English-speaking listeners. Similarity between /i/ and /ɪ/ was also found by Tsukada et al. (2005), where a native English-speaking judge misclassified 37% of /i/ productions as /ɪ/ and 17% of /ɪ/ productions as /i/. Additionally, 66% of /ɛ/ tokens were heard as /æ/, and 82% of /ʌ/ tokens were heard as /ɑ/. Flege et al. (1997) provided further evidence of the bidirectional confusion of L1 Korean productions of /i, ɪ/, as well as /ɛ, æ/, based on classification by native English-speaking listeners. Finally, Schirra (2012) also revealed that /oʊ, u/ underwent the same degree of F1 and F2 movement over the course of the vowel in non-native and native productions.

2.2.3 L1 Mandarin/L2 English

According to Kim (2011) and Shimizu (2011), L1 Mandarin speakers clearly distinguish word-initial English stops based on VOT, with short lag values for voiced targets and long lag values for voiceless ones. However, in transcribing L1 Mandarin productions of English, Rogers and Dalby (2005) noted bidirectional confusion between voiced and voiceless stops at all places of articulation word-initially and word-finally. Word-medially, /p, k/ were sometimes perceived as their voiced counterparts. Thus, while both Mandarin and English contrast short lag VOT values with long lag VOT values word-initially, the English productions of L1 Mandarin speakers are not necessarily perceived as error-free.

Chen et al. (2001) measured F1 and F2 values of 11 American English vowels, and found that L1 Mandarin speakers had smaller vowel quadrilaterals than did native speakers. Rogers and Dalby (2005) reported many details about L1 Mandarin productions of American English vowels, including issues involving confusion between tense and lax vowels (/ɪ/ was perceived as /i/, and /ʊ/ as /u/, but /eɪ/ as /ɛ/), diphthongs produced as monophthongs (/aɪ/ as /ɑ, ɛ/), and backing of mid vowels (/ɚ/ as /ɔɚ/, /ʌ/ as /oʊ/). They also found many issues with vowel height (/u/ as /oʊ/, /ɛ/ as /æ/, /æ/ as /ɛ/, /ʌ/ as /ɑ/). Flege et al. (1997) confirmed some of these patterns, such as /ɪ/ as /i/, /ɛ/ as /æ/, and /æ/ as /ɛ/, and found others, such as /i/ as /ɪ/, /ɛ/ as /eɪ/, and /æ/ as /eɪ/. Wang and van Heuven (2006) found that L1 Mandarin speakers did not differentiate spectrally between tense and lax vowel productions in English, but did have generally native-like durational differences.

2.2.4 L1 Spanish/L2 English

L1 Puerto Rican Spanish speakers studied by Flege and Eefting (1987) produced over 70% of word-initial /b, d, g/ targets with lead VOT, and the remainder with short lag VOT. Word-initial /p, t, k/ productions had long lag VOTs, with means of 48ms for /p/, 56ms for /t/, and 67ms for /k/, as compared to 78ms, 89ms, and 94ms, respectively, for native English speakers. These L1 Spanish speakers were "later childhood bilinguals" who began learning English in school at age 5 or 6 but had never lived in an English-speaking environment. Nathan (1987) described a somewhat different population of L1 Spanish speakers from Colombia, Venezuela, and Costa Rica, who were first exposed to English in high school but felt that their meaningful experience with English was limited to the six months they had been enrolled in an intensive language class in the United States. These speakers were

re-recorded 18 months after an initial test to evaluate changes in their English pronunciation. Initially, their productions of English word-initial /p, t, k/ had a mixture of short lag and long lag VOT values; at retest, these values tended to be slightly longer. While most productions of /b, d/ exhibited lead voicing at both testing times, roughly 20% and later 50% of /g/ productions had short lag VOT values, indicating that the speakers seemed to be moving toward typical VOT patterns in the target language. Native American English listeners in Ortega-Llebaria (1997) misheard some L1 Spanish productions of voiceless stops as voiced, suggesting that these VOT patterns can have consequences for perception.

Zampini (1996), Ortega-Llebaria (1997), and Donadio (2002) reported that L1 Spanish speakers from a variety of dialects produced some English voiced stop targets as fricatives, often in environments where these targets would be produced as fricatives in Spanish, suggesting that this pattern resulted from L1 transfer. Flege and Davidian (1984) found the same pattern word-finally for L1 speakers of Mexican and Salvadoran Spanish, but not for L1 speakers of Chinese or Polish, languages which lack the spirantization pattern.

Ortega-Llebaria (1997) found that while American English listeners generally heard L1 Spanish speakers' productions of English tense vowels and diphthongs accurately, they commonly misheard L1 Spanish productions of English lax vowels as tense (/ɪ/ as /i/, /æ/ as /ɑ/, /ʌ/ as /ɑ, ɔ/). The speakers in Donadio (2002) showed the same tense-for-lax substitutions. In Flege et al. (1997), /æ/ was heard as /ɑ, ʌ/ and /ɪ/ as /i/, although /i/ was also frequently heard as /ɪ/. L1 Spanish speakers sometimes produce longer vowels than do native American English speakers, although Shah (2002) found that this was true more often for unstressed vowels than for stressed vowels, indicating that the English of L1 Spanish speakers exhibits less vowel reduction than that of native speakers.

2.3 Summary of acoustic properties

The non-native English talkers in the present investigation were native talkers of Hindi, Korean, Mandarin, and Spanish. Based on patterns observed in non-native English speech from talkers of these backgrounds, as well as in these four languages themselves, it seems that non-native stop productions may differ from native productions in VOT, and at least for L1 Hindi and L1 Korean talkers, in f0 and spectral tilt at the beginning of the following vowel. Non-native vowel productions differ from native productions in a variety of ways that are often revealed as misperceptions of the target category. Such differences might be captured acoustically by measuring vowel formants, including formant changes over the course of the vowel, and vowel duration. Measurements of these consonant- and vowel-related acoustic properties from the stimuli used in the present investigation will be presented in Chapter 3, following more detailed discussion of the stimuli and the talkers.

McCullough (2013) noted that English productions from L1 Hindi talkers received significantly higher accentedness ratings than those from L1 Korean and L1 Mandarin talkers. As noted above, L1 Hindi speakers tend to produce English alveolar stops as retroflexes (Vidyalankar, 2002), which might contribute to the perception of a strong foreign accent. As there is no widely accepted vowel-independent acoustic measure of retroflexion, this property was not measured in the present investigation. However, Chapters 4 and 5 include comparisons of the perceptual responses to L1 Hindi talkers' productions of alveolar stop targets with perceptual responses to their productions of stop targets at other places of articulation, to evaluate whether such a measure, should one be determined, might merit inclusion in future work.

CHAPTER 3: RECORDINGS

While the preliminary work by McCullough (2013) used recordings from the Buckeye GTA Corpus (Hardman, 2010) as acoustic stimuli, the use of these preexisting recordings allowed for little control of the linguistic context surrounding the segments of interest. Thus, for this expanded investigation, new talkers were recorded reading highly controlled materials, as described below.

3.1 Methods

3.1.1 Materials

Word lists were constructed for English, Hindi, Korean, Mandarin, and Spanish. The words on each list began with generally unique combinations of stops and vowels in each language,[1] and were real, although not uniformly familiar, lexical items. All words were disyllabic, and those on the English list were trochees. The structure of all words was stop-vowel-stop-vowel(-consonant). Real words were found for all target stop-vowel combinations in English. In Hindi, Korean, Mandarin, and Spanish, stop-vowel combinations for which no real word existed were omitted from the list, such that the total number of words for these lists was less than the total number of possible stop-vowel combinations.

The 60 words on the English list are presented in Table 3.1. This list crossed stops /b, d, ɡ, p, t, k/ with vowels /i, ɪ, ɛ, æ, ʌ, u, ɚ, eɪ, oʊ, aɪ/. Vowels /ɑ, ɔ/ were omitted due to

[1]The Hindi word list contained some pairs of words with the same initial stop-vowel sequence, but dental versus retroflex medial stops, for an investigation not pursued in the present work.

their inconsistent merger status across speakers of American English (Ash, 2003), and real words did not exist for all combinations of stops with /ʊ, aʊ, ɔɪ/. Additionally, /ə/ does not appear as a class in the word list because vowels were manipulated in the stressed initial syllable. As all entries on this list were characterized by initial stress, it was expected that native talkers of American English would flap the word-medial alveolar stops /d, t/.

The 60 Hindi words displayed in black in Table 3.2 were used as stimuli in Experiments 5 and 6. Additional words recorded but not used in the perception experiments are displayed in gray. This list crossed stops /b, d, ɖ, g, p, t, ʈ, k, p^h, t^h, $ʈ^h$, k^h/ with vowels /i, ɪ, e, ɛ, ə, u, ʊ, o, ɔ, ɑ/. The voiced aspirated stops /$b^ɦ$, $d^ɦ$, $ɖ^ɦ$, $g^ɦ$/ and nasal vowels were omitted in order to reduce the number of possible stop-vowel combinations. /æ/ was also omitted, as it occurs only in loanwords and was infrequent in words with the desired properties.

The Korean word list crossed stops /p*, t*, k*, p, t, k, p^h, t^h, k^h/ with vowels /i, ɛ, ɨ, ʌ, a, u, o/, as evidenced by the 55 words shown in Table 3.3. Diphthongs with onglides were avoided in the interest of limiting the number of possible stop-vowel combinations.

The 38 words on the Mandarin list, shown in Table 3.4, resulted from crossing stops /p, t, k, p^h, t^h, k^h/ with vowels /i, a, u, ə, eɪ, oʊ, aɪ, aʊ/. The monophthong /y/ does not appear after stops (Chin, 2006), and diphthongs with onglides and triphthongs were omitted because they are often analyzed as sequences of vowels rather than as individual vowels themselves (Chin, 2006; Sun, 2006). To ensure that Mandarin productions were prosodically comparable to productions in the other languages included in this work, the word list contained only words with tone 3 (low rising), tone 4 (falling), or neutral tone in the second syllable. Words with tone 1 (high level) and tone 2 (high rising) in the second syllable may have resembled productions with list intonation, which was discouraged in recordings of the other languages.

Table 3.5 shows the list of 27 Spanish words, for which stops /b, d, g, p, t, k/ were crossed with vowels /i, e, a, u, o/. Diphthongs were not included because they are not generally considered single phonemes in Spanish (Dalbor, 1969). As the words were elicited in isolation, the word-initial voiced stops should be realized as stops, and the word-medial voiced stops should be realized as fricatives, per the typical distribution of allophones.

A short list of sentences, List 7 from the Bamford-Kowal-Bench sentences revised for American English (Bamford and Wilson, 1979), was also recorded by each participant. Non-native English speakers also recorded translations of these sentences in their native languages. These productions were not used as stimuli in the perception experiments and will not be discussed further.

	b	d	g	p	t	k
i	beagle /bigəl/	deeding /didɪŋ/	geeky /giki/	Peter /pitɚ/	teepee /tipi/	keeper /kipɚ/
ɪ	bidder /bɪdɚ/	dipper /dɪpɚ/	giggle /gɪgəl/	pity /pɪti/	tickle /tɪkəl/	kibble /kɪbəl/
ɛ	bedding /bɛdɪŋ/	Debbie /dɛbi/	getting /gɛtɪŋ/	pepper /pɛpɚ/	techie /tɛki/	kegger /kɛgɚ/
æ	batter /bætɚ/	dapper /dæpɚ/	Gabby /gæbi/	paddle /pædəl/	tagging /tægɪŋ/	cackle /kækəl/
ʌ	buddy /bʌdi/	double /dʌbəl/	gutter /gʌtɚ/	pucker /pʌkɚ/	tugging /tʌgɪŋ/	couple /kʌpəl/
u	bootie /buti/	duping /dupɪŋ/	Google /gugəl/	poodle /pudəl/	tubing /tubɪŋ/	kooky /kuki/
ɚ	burger /bɚgɚ/	dirty /dɚti/	girdle /gɚdəl/	purple /pɚpəl/	turkey /tɚki/	curbing /kɚbɪŋ/
eɪ	baby /beɪbi/	dating /deɪtɪŋ/	gable /geɪbəl/	paper /peɪpɚ/	taking /teɪkɪŋ/	cable /keɪbəl/
oʊ	Boagie /boʊgi/	dopey /doʊpi/	goading /goʊdɪŋ/	poker /poʊkɚ/	Toby /toʊbi/	coating /koʊtɪŋ/
ɑɪ	Bible /bɑɪbəl/	diaper /dɑɪpɚ/	guiding /gɑɪdɪŋ/	piking /pɑɪkɪŋ/	tiger /tɑɪgɚ/	kiting /kɑɪtɪŋ/

Table 3.1: English stimuli

	b	d	ɖ	g	p	t
i	/bikər/ ‘beaker’	/dipək/ ‘light’	/ɖikɑ/ ‘Deeka’ (name)	/gitɑ/ ‘holy book’	pitəl/ ‘brass’	/tikʰɑ/ ‘tangy’
ɪ	/bɪgʊl/ ‘bugle’	/dɪkʰɑ/ ‘see’	/ɖɪpo/ ‘depot’		/pɪtɑ/ ‘father’	/tɪtʰɪ/ ‘date’
e	/bebi/ ‘baby’	/dekʰɑ/ ‘see’			/peʈi/ ‘box’	
					/pepər/ ‘paper’	
ɛ	/bɛbɪl/ ‘Bible’				/pɛdɑ/ ‘to be born’ /pɛɖəl/ ‘pedal’	
ə	/bəgəl/ ‘side’	/dəpʰən/ ‘funeral’	/ɖəbəl/ ‘double’	/gədɑ/ ‘mace’ /gəʈʰən/ ‘frame’	/pətɑ/ ‘address’	/tətʰɑ/ ‘and’
u	/butɑ/ ‘stamina’		/ɖubɑ/ ‘intent’	/gudɑ/ ‘pulp’	/putɪk/ ‘septic’	/tuti/ ‘yellow-’ hammer’
	/buʈɑ/ ‘plant’					
ʊ	/bʊki/ ‘bookie’					
o	/botəl/ ‘bottle’ /boʈi/ ‘chop’			/godi/ ‘godi’	/poti/ ‘granddaughter’	/totɑ/ ‘parrot’
ɔ		/dɔgi/ ‘doggie’		/gɔtəm/ ‘Gautama’	/pɔdʱɑ/ ‘plant’	/tɔbɑ/ ‘God forbid’
ɑ	/bɑbu/ (title)	/dɑdɑ/ ‘grandfather’	/ɖɑku/ ‘bandit’	/gɑtʰɑ/ ‘saga’	/pɑgəl/ ‘crazy’	/tɑpək/ ‘stove’

Table 3.2: Hindi stimuli (*continued on next page*)

Table 3.2 (*continued from previous page*)

	ʈ	k	pʰ	tʰ	ʈʰ	kʰ
i	/ʈibi/ ‘T.B.’		/pʰitɑ/ ‘tape’			
ɪ	/ʈɪkəʈ/ ‘ticket’					
e		/kebəl/ ‘cable’	/pʰeti/ ‘stir’		/ʈʰekɑ/ ‘contract’	/kʰeti/ ‘farm’
ɛ		/kɛdi/ ‘prisoner’ /kɛɖi/ ‘caddie’				
ə	/ʈəpər/ ‘topper’	/kətʰən/ ‘statement’	/pʰəʈɑ/ ‘burst’	/tʰəkɑ/ ‘tired’	/ʈʰəgi/ ‘trickery’	/kʰəbər/ ‘news’ /kʰəɖɑ/ ‘stand’
u		/kupən/ ‘coupon’	/pʰupʰɑ/ ‘uncle’			/kʰubi/ ‘good quality in a person’
ʊ		/kʊpɪt/ ‘wrathful’				
o	/ʈopi/ ‘cap’	/koko/ ‘cocoa’	/pʰokəs/ ‘focus’		/ʈʰokər/ ‘kick’	/kʰokɑ/ ‘hole in a wall’ /kʰoʈɑ/ ‘bad’
ɔ		/kɔpi/ ‘copy’				
ɑ	/ʈɑpu/ ‘isle’	/kɑpʰi/ ‘pretty’ /kɑʈɑ/ ‘bite’		/tʰɑpi/ ‘pallet’	/ʈʰɑkʊr/ ‘title’	/kʰɑtɑ/ ‘account’

	p*	t*	k*	p	t	k	p^h	t^h	k^h
i	/p*ita/ ‘sprain’	/t*ita/ ‘wear’	/k*ita/ ‘embrace’	/pip^{h}ɨ/ ‘beef’	/tipa/ ‘diva’	/kip*ɨm/ ‘joy’	/p^{h}ik^{h}el/ ‘pickle’	/t^{h}ik*ɨl/ ‘dust’	
e	/p*ep*e/ ‘gaunt’	/t*eta/ ‘burn’	/k*et*ʌk/ ‘rice cake with sesame’	/pek*op/ ‘navel’	/tep^{h}i/ ‘take shelter’	/keke/ ‘each one’	/p^{h}et^{h}ʌn/ ‘pattern’	/t^{h}ek^{h}ɨl/ ‘tackle’	/k^{h}epin/ ‘cabin’
ɨ		/t*ɨta/ ‘float’	/k*ɨta/ ‘extinguish’			/kɨt*e/ ‘then’		/t^{h}ɨta/ ‘sprout’	/k^{h}ɨki/ ‘size’
ʌ	/p*ʌk*uk/ ‘cuckoo’	/t*ʌkkuk/ ‘rice cake soup’		/pʌt^{h}in/ ‘button’	/tʌti/ ‘late’	/kʌki/ ‘there’	/p^{h}ʌt*ik/ ‘suddenly’	/t^{h}ʌpu/ ‘taboo’	/k^{h}ʌp^{h}i/ ‘coffee’
a		/t*ata/ ‘pick’	/k*at*ak/ ‘nod’	/pap*i/ ‘busily’	/tapaŋ/ ‘teahouse’	/kak*ɨm/ ‘sometimes’	/p^{h}at^{h}ɨ/ ‘part’	/t^{h}ake/ ‘break’	/k^{h}ap^{h}i/ ‘copy’
u		/t*uk*ʌŋ/ ‘lid’	/k*uta/ ‘dream’	/put^{h}ʌ/ ‘from’	/tupu/ ‘tofu’	/kup^{h}an/ ‘old edition’	/p^{h}utɨl/ ‘poodle’	/t^{h}uko/ ‘contribution’	/k^{h}uk^{h}i/ ‘cookie’
o	/p*op*o/ ‘kiss’	/t*opak/ ‘clearly’	/k*ok*o/ chicken sound	/pot^{h}em/ ‘help’	/toku/ ‘tool’	/kotiŋ/ ‘high class’	/p^{h}ok^{h}ʌ/ ‘poker’	/t^{h}ok*i/ ‘rabbit’	/k^{h}op^{h}i/ ‘nosebleed’

Table 3.3: Korean stimuli

	p	t	k	p^h	t^h	k^h
i	/pi↗k^hʊŋ↓/ ‘nostril’	/ti↘ti/ ‘younger brother’		/p^hi→kaɪ↓/ ‘correct an exam’	/t^hi↓t^hʊŋ↓/ ‘propriety’	
ə		/tə↗t^hi↓/ ‘appropriate’	/kə→ta/ ‘lump’	/ɥ/ ‘4’	/t^hə↘kaʊ↓/ ‘special feature’	/k^hə↓k^hoʊ↓/ ‘tasty’
a	/pa↘pa/ ‘father’	/ta↓ti↓/ ‘underpainting’	/ka→ka/ ‘quack’	/p^ha→p^ha/ (onomatopoeia)	/t^ha↓tiŋ↓/ ‘tower top’	/k^ha↓t^hə↘/ ‘Carter’
u	/pukoʊ↓/ ‘careful’	/tuk^{h}oʊ↓/ ‘ferry’	/kut^{h}i↓/ ‘solid’	/p^hu↗t^haʊ/ ‘grape’	/t^hu↗p^hu↓/ ‘atlas’	/k^hu↓təŋ↓/ ‘wait for a long time’
eɪ	/peɪ→taɪ↘/ ‘suspenders’			/p^heɪ↗t^hu↓/ ‘earth up’		
oʊ		/toʊ→toʊ/ ‘undergarment’	/koʊpən↓/ ‘break even’		/t^hoʊ↗kaʊ↓/ ‘contribute’	/k^hoʊ↓k^hə↓/ ‘thirst’
aɪ	/paɪ↗t^ha↓/ ‘White Pagoda’	/taɪ→pan↓/ ‘inflexible’	/kaɪ↓k^hoʊ↓/ ‘correct oneself’	/p^haɪ↗ku↓/ ‘pork chop’	/t^haɪ↘tu/ ‘manner’	/k^haɪ→t^həŋ/ ‘open-minded’
aʊ	/paʊ→ku↓/ ‘corn’	/taʊ↘ti↓/ ‘finally’	/kaʊ→t^ha↓/ ‘tower’	/p^haʊ↘t^ha↓/ ‘turret’	/t^haʊ↗p^haʊ↓/ ‘run away’	/k^haʊ↓k^haʊ/ ‘quiz’

Table 3.4: Mandarin stimuli

	b	d	g	p	t	k
i	/biga/ ‘column’	/dike/ ‘dam’		/pide/ ‘request’	/tipo/ ‘type’	/kitan/ ‘take away’
e	/beka/ ‘grant’	/debil/ ‘weak’	/geto/ ‘ghetto’	/pega/ ‘difficulty’		/kedo/ ‘stay’
a	/baba/ ‘saliva’	/daga/ ‘dagger’	/gato/ ‘cat’	/padel/ ‘paddle’	/tapa/ ‘lid’	/kaki/ ‘khaki’
u	/buke/ ‘ship’	/dudo/ ‘doubt’		/puber/ ‘adolescent’	/tute/ ‘card game’	/kupo/ ‘quota’
o	/bobo/ ‘stupid’	/dopan/ ‘drug’	/gota/ ‘drop’	/poker/ ‘poker’	/toga/ ‘toga’	/kodo/ ‘elbow’

Table 3.5: Spanish stimuli

3.1.2 Talkers

Native and non-native speakers of American English were recorded; the latter were native speakers of Hindi, Korean, Mandarin, and Spanish. Self-reported demographic details about the 6 native speakers of American English who were chosen as talkers for the perception experiments are provided in Table 3.6. The entries in the “code” column use letters to represent each talker’s native language and sex, and numbers 1 through 6 to uniquely identify each talker within a language background. In the present work, all talkers are referred to by these codes.

Although the L1 American English talkers described in Table 3.6 were from a variety of dialect areas in the United States, the author and another linguist experienced with phonetic variation in the United States independently judged the productions selected for use as stimuli as lacking any particular manifestation of stigmatized regional or ethnic accent. This attempt at “standardness” was made due to the fact that during piloting of an earlier

Code	Sex	Residence through age 18		Age
E1f	F	Lima, OH	0-18	31
E2f	F	Cincinnati, OH	0-3	33
		& Columbus, OH	3-18	
E3f	F	Evanston, IL	0-3	20
		& Plymouth, MA	3-15	
		& Powell, OH	15-18	
E4m	M	Wichita, KS	0-18	26
E5m	M	Cleveland, OH	0-18	31
E6m	M	Douglasville, GA	0	25
		& Claremore, OK	0-18	

Table 3.6: L1 American English talkers

perception experiment (McCullough, 2013), several listeners reported uncertainty regarding how to deal with productions that sounded "Southern" in the context of a task about foreign accentedness. An additional 33 native speakers of American English (25 females) were recorded, but were not selected as talkers for this investigation, generally because their productions more often sounded clearly regionally accented. Many of these speakers, including talker E3f, were recruited through the linguistics department subject pool and were compensated by partial course credit. In an effort to find individuals more similar in age to the non-native talkers discussed below, additional speakers, including the remaining 5 chosen as talkers, were recruited through personal contacts and received $10 for their participation.

Self-reported demographic details about the non-native speakers of American English who were chosen as talkers for the perception experiments are provided in Tables 3.7 through 3.10. Each talker's age of first exposure to English (FE) and age of arrival in the

United States (AoA) are reported. No objective criterion for English proficiency was imposed upon these talkers. However, all were living in central Ohio at the time of recording and were readily able to communicate with the experimenter in English. Speaking subsection scores from the TOEFL iBT (TOEFL-S) are included in the tables for the talkers who reported them. Some participants did not remember their scores or declined to disclose them, or had never taken the TOEFL iBT. This was especially true of the L1 Spanish talkers, who tended to be employees of the university or friends of students, rather than being students themselves. To the extent that differences in English proficiency levels across L1 backgrounds existed, they were assumed to represent actual differences in the proficiency levels of these local populations, and differences to which listeners in the perception experiments were likely accustomed. Because of the status of English in India and the resulting high proficiency levels of L1 Hindi speakers, local recruitment of speakers with equal proficiency levels across L1 backgrounds would have been difficult or impossible.

An additional 6 L1 Hindi speakers (1 female), 10 L1 Korean speakers (8 females), 8 L1 Mandarin speakers (5 females), and 6 L1 Spanish speakers (5 females) were recorded, but were not selected as talkers for this investigation. In general, talkers were selected so that the dialects represented within each L1 background were relatively comparable, although there remained some unavoidable variation within each group. Additionally, speakers who mispronounced relatively few targets, as judged perceptually by the author, were preferred; fewer mispronounciations facilitated the selection of stimuli for perception experiments, which is discussed in more detail in Section 3.1.4 below. All non-native speakers were recruited through personal contacts and received $10 for their participation.

Table 3.7 contains information about the 6 L1 Hindi talkers. One talker, H3f, had moved to central Ohio at age 15, and was judged by the author to be targeting American rather than

Code	Sex	Residence through age 18		Age	FE	AoA	TOEFL(S)
H1f	F	Pune, Maharashtra, India	0-18	22	4	22	28/30
H2f	F	Haryana, India	0-18	24	4	22	22/30
H3f	F	Patiala, Punjab, India & Dayton, OH	0-15 15-18	21	5	15	n.r.
H4m	M	New Delhi, Delhi, India	0-18	23	4	22	28/30
H5m	M	Jabalpur, Madhya Pradesh, India & Kota, Rajasthan, India	0-17 17-18	27	4	26	n.r.
H6m	M	Pune, Maharashtra, India	0-18	26	3	24	27/30

Table 3.7: L1 Hindi talkers

Indian English. None of these talkers reported English as a native language, although all were first exposed to English as young children.

The linguistic situation in India is complex, with many language varieties subsumed under the label "Hindi," and with varieties of Hindi present in many areas where another language dominates. Standard Hindi is based on speech from New Delhi, where talker H4m grew up. Talker H2f was from the nearby region of Haryana, where the local language is a variety of Hindi called Haryanvi. Haryanvi pronunciation involves numerous diphthongs, and free variation of /e/ with /ɑ/ and of /i/ with /e/ (Mishra and Bali, 2011). Haryanvi also exhibits contrastive tone (Masica, 1991). Talker H5m spent most of his childhood in an area characterized by another Hindi dialect, Bagheli, which also has diphthongs and variation between /i/ and /e/ and between /u/ and /o/ (Mishra and Bali, 2011). He also spent time in a Harauti-speaking area (Masica, 1991). Harauti is a variety of Rajasthani, which is itself inconsistently classified as a dialect of Hindi or as a separate language (Shapiro, 2003). Harauti lacks length distinctions in the front vowels (Masica, 1991), at least phonologically, and has certain restrictions on the distribution of aspiration and voicing within

a word (Allen, 1957). Punjabi, an Indo-Aryan language like Hindi, is the dominant language in the area where talker H3f lived. Punjabi has numerous diphthongs and a relatively low /ɛ/ vowel, and its historical voiced aspirates were replaced by position-dependent tonal patterns (Shackle, 2003). Talkers H1f and H6m were from an area dominated by Marathi, another Indo-Aryan language. In fact, talker H6m reported both Hindi and Marathi as native languages, but this bilingualism was not considered problematic for the current investigation because the stop inventories of Hindi and Marathi are identical and the VOT values are nearly so (Lisker and Abramson, 1964), and because the vowel inventories of these two languages are quite similar (Ohala, 1999; Pandharipande, 2003).

Information about the 6 L1 Korean talkers is presented in Table 3.8. Most of these talkers were from Seoul, in the central dialect region, which is considered the standard for the language. Talker K3f lived in Gumi and Daegu, both in the Gyeongsang region, as well as in China briefly at age 10 and for 2 years as a teenager. Speakers in the Gyeongsang region palatalize velar stops before high vowels, neutralize /ɨ, ʌ/ to [ʌ], and use tone phonemically (Sohn, 1999). Talker K5m lived in two cities in the central dialect region, Suwon and Incheon, as well as Gwangju, in the Jeolla region. Characteristics of the Jeolla dialect include tensification of word-initial lax stops, palatalization of velar stops before high vowels, backing of high vowels, and fronting and/or raising of /ɨ, a, e/ (Sohn, 1999). As talker K5m only lived in this region for several years as an adolescent, the extent to which these patterns might be evident in his speech is uncertain. Because all the talkers in this work were born after 1965 (that is, were under the age of 47 when recorded in 2012), they are expected to use both VOT and f0 to distinguish Korean stops. Although talkers K2f and K6m moved to the United States as teenagers, neither was judged by the author to sound like a native speaker of American English.

Code	Sex	Residence through age 18		Age	FE	AoA	TOEFL(S)
K1f	F	Seoul, South Korea	0-18	26	15	20	28/30
K2f	F	Seoul, South Korea	0-16	20	13	16	21/30
		& Seattle, WA	16-18				
K3f	F	Gumi, South Korea	0-10	26	9	22	26/30
		& Shanghai, China	10				
		& Daegu, South Korea	11-16				
		& China	16-18				
K4m	M	Seoul, South Korea	0-18	24	13	22	18/30
K5m	M	Suwon, South Korea	0-8	26	13	26	n.r.
		& Gwangju, South Korea	9-12				
		& Incheon, South Korea	13-18				
K6m	M	Seoul, South Korea	0-16	26	13	16	n.r.
		& Rocky Mount, NC	16-17				
		& Fork Union, VA	17-18				

Table 3.8: L1 Korean talkers

Demographic information for the 6 L1 Mandarin talkers is shown in Table 3.9. Because of substantial variation among the different language varieties spoken in China, care was taken to ensure that the talkers chosen were from Mandarin dialect areas. Standard Mandarin is based on the northern dialect spoken in Beijing. Two talkers, M1f and M4m, were from Beijing itself, while M5m and M6m were from other northern dialect areas. Talker M2f was from the southwestern dialect area, and talker M3f from the eastern dialect area. The Mandarin dialects are characterized by "general uniformity" (Norman, 1988, 192), and most pronunciation differences involve segments that do not appear in the stimuli discussed above. The realization of tone, however, is subject to regional variation.

Table 3.10 summarizes information about the 6 L1 Spanish talkers, who were from a variety of dialect areas in the Americas. Specifically, talkers S1f, S4m, and S5m were from the Caribbean, and talker S2f from the Latin American Highlands, while talker S6m had

Code	Sex	Residence through age 18		Age	FE	AoA	TOEFL(S)
M1f	F	Beijing, China	0-18	23	15	18	n.r.
M2f	F	Wuhan, Hubei, China	0-18	26	12	23	22/30
M3f	F	Hefei, Anhui, China	0-18	34	12	33	23/30
M4m	M	Beijing, China	0-18	21	8	20	27/30
M5m	M	Lankao, Henan, China	0-18	27	10	23	21/30
M6m	M	Chaoyang, Liaoning, China	0-18	28	14	27	n.r.

Table 3.9: L1 Mandarin talkers

Code	Sex	Residence through age 18		Age	FE	AoA	TOEFL(S)
S1f	F	Carolina, Puerto Rico	0-18	27	4	23	n.r.
S2f	F	Cochabamba, Bolivia	0-18	33	29	30	n.r.
S3f	F	Concepción, Chile	0-18	41	14	36	n.r.
S4m	M	Caracas, Venezuela	0-18	20	5	18	n.r.
S5m	M	Santo Domingo, Dominican Republic	0-18	34	15	25	n.r.
S6m	M	Cochabamba, Bolivia & Caracas, Venezuela & Cochabamba, Bolivia	0 0-13 13-18	38	13	36	n.r.

Table 3.10: L1 Spanish talkers

lived in both these regions. Chile, where talker S3f grew up, is itself a dialect area (Dalbor, 1969). While traditional accounts detail numerous pronunciation differences among these varieties of American Spanish, none of these differences involve stops, and only two involve vowels: Dalbor (1969) claimed that speakers of Caribbean dialects produce /e/ as [ɛ] in open as well as closed syllables, and Canfield (1981) reported that Bolivian Spanish speakers reduce vowels in unstressed syllables.

3.1.3 Recording procedure

Each talker began by filling out a brief language background questionnaire, which is included in Appendix A. He or she was then seated in a sound-attenuated booth wearing a Shure SM10A head-mounted microphone, and read aloud English words and sentences presented sequentially on a computer screen. These productions were recorded, via an ART Tube MP Project Series preamplifier, at 22.5kHz in Audacity 1.2 on a Dell XPS M1210 computer running Windows XP. Each word and sentence was produced twice consecutively, and each target was displayed until the talker clicked the mouse to proceed. The block containing words preceded the block containing sentences. Within each block, targets were presented in a unique random order for each talker, except that the first 4 items always consisted of practice items. Practice words (*boating*, *cougar*, *kettle*, *piper*) duplicated initial consonant-vowel combinations of other target words, but were not themselves included in the word list above. Talkers were instructed to either guess or ask the experimenter if they were unsure about the pronunciation of a target. The experimenter did not offer pronunciation assistance unless it was explicitly requested by the talker for a particular target.

Each talker who was not a native speaker of American English then immediately completed a second recording in his or her native language. The target words and sentences were presented on a computer screen in the native orthography of the language: Devanagari for Hindi, Hangul for Korean, and simplified Chinese characters for Mandarin. Except for the language of the words and sentences elicited, this procedure was identical to the procedure described for the English materials.

3.1.4 Selection of potential stimuli

The second production of each word was extracted from the recording, except in cases where this production was obscured by non-speech sounds, when the first production was taken instead. No production which was first modeled by the experimenter was considered for inclusion among the perception experiment stimuli. English productions which were perceived by the author to be produced with complete segmental substitutions (except as related to stop voicing), or insertions or deletions of segments, were also excluded from the pool of potential stimuli in order to ensure comparable acoustic measurements across all stimuli; such productions were often the result of misreading the target word, or guessing incorrectly at the pronunciation of an unknown word. The stimuli actually used in the perception experiments are discussed in detail in Chapters 4 through 7.

3.2 Acoustic patterns

Based on the expectations about L1 interference summarized in Section 2.3, VOT, f0, spectral tilt, vowel duration, and vowel quality were measured in the first syllable of each of the productions used in the perception experiments. Measurements were limited to the first syllable so that syllable- and word-length stimuli would be associated with exactly the same set of acoustic values, and because the content of the second syllable was not as highly controlled as the content of the first. A total of 540 tokens were measured, consisting of 300 productions in English (60 from talkers of each of the 5 language backgrounds) and 240 productions in other languages (60 from L1 talkers of each of the 4 other languages). Each talker was represented by 10 tokens in English and an additional 10 tokens in his or her native language (if it was not English). Stop voicing, stop place of articulation, and vowel identity were balanced as well as possible within each talker and across talkers

from the same language background. Further details are provided in Chapter 4 for the 300 productions in English, and in Chapter 6 for the 240 productions in other languages.

Statistical tests were not performed on the comparisons presented below, as production differences are not the primary focus of this dissertation, and as differences among talkers of the same language background, and not only among sets of talkers of different language backgrounds, may be relevant to analyses of perception. Nonetheless, examination of talkers' productions can confirm that the acoustic properties measured might reasonably be dimensions of the signal that listeners employ in these experiments.

3.2.1 VOT

VOT was measured as the time between the onset of the rapid change in amplitude indicative of stop release, and the upward-going zero crossing indicating the onset of periodicity. Negative VOT values represent cases in which the onset of periodicity preceded the stop release (i.e., lead voicing). VOT measurements are shown in Figure 3.1 for talkers' L1 productions, and in Figure 3.2 for talkers' English productions. In these and all subsequent figures, measurements for female and male talkers are displayed separately due to the method by which acoustics were incorporated into the statistical models described in later chapters (see Section 4.1.1). Each figure contains a total of 30 data points per sex for each of the 5 language backgrounds. In these figures, *d* represents data points for voiced stop targets, regardless of place of articulation, and *t* represents data points for voiceless stop targets. Data points for aspirated stop targets appear as *h*. Because the labels are based on phonological rather than phonetic descriptions, English stops are represented as *d* and *t*. In Korean, where the terminology differs somewhat from the other languages considered, data

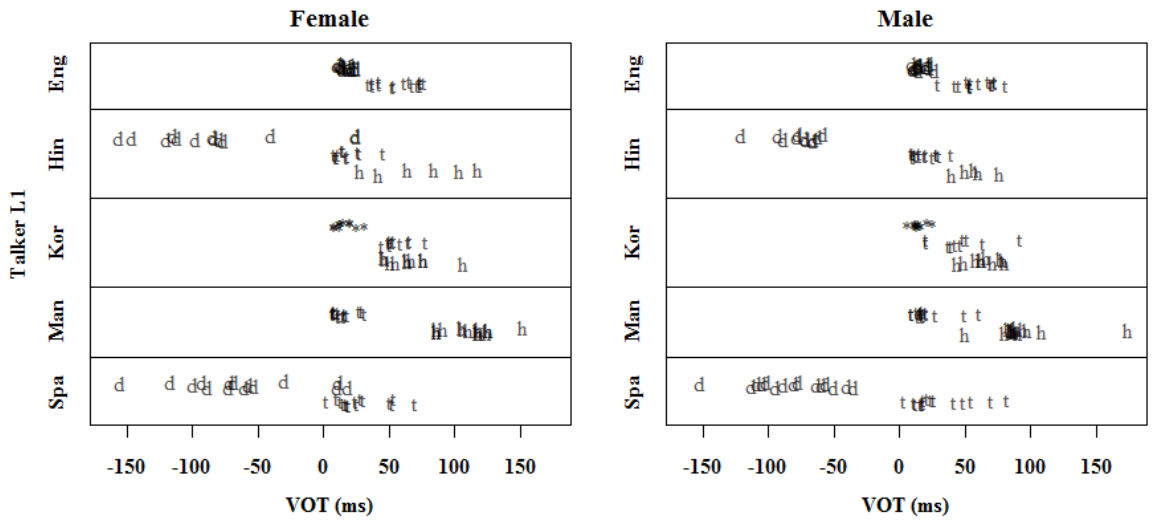

Figure 3.1: VOT in L1 productions

points for tense/fortis stop targets appear as *; for lax/lenis stop targets, *t*; and for aspirated stop targets, *h*.

The patterns in Figures 3.1 and 3.2 generally align with those described in Chapter 2. In English, differences between the productions of native and non-native talkers, particularly the many instances of lead voicing for L1 Hindi and L1 Spanish talkers and the short lag values for voiceless stop targets produced by L1 Hindi talkers and male L1 Spanish talkers, can reasonably be argued to arise from influence of the talkers' L1s.

3.2.2 Fundamental frequency (f0)

Fundamental frequency was measured as the mean over the first 25ms of the vowel; the onset of the vowel was taken to be the upward-going zero crossing indicating the onset of periodicity following the stop release. Values were extracted automatically using a Praat script that displayed a spectrogram with an overlaid fundamental frequency track, such that the author could check token-by-token for tracking errors. A 25ms window was chosen

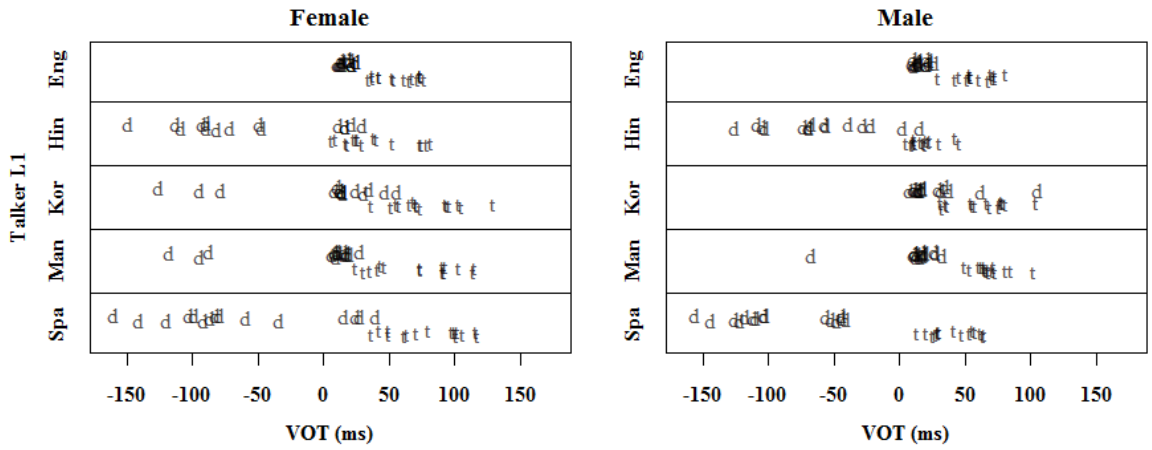

Figure 3.2: VOT in English productions

because measurement at the exact onset of the vowel, or over smaller windows at the beginning of the vowel, was not possible for a relatively high proportion of tokens. In 13 of the 540 stimuli, fundamental frequency was not measurable over the first 25ms, and the mean over the measurable remainder of the vowel was substituted, as missing data would have complicated the acoustic aspects of the perception analyses. Measurements of f0 are shown in Figure 3.3 for talkers' L1 productions and in Figure 3.4 for their English productions. Each figure contains a total of 30 data points per sex for each of the 5 language backgrounds. In these figures, the symbols represent the stop that preceded the vowel measured. The expected pattern of lower f0 following a lax stop than a tense or aspirated stop is confirmed by the Korean talkers, and seems to persist in L1 Korean productions of English as lower f0 following a voiced stop than a voiceless stop. Contra Dutta (2007), f0 values in Hindi are not clearly lower for unaspirated voiced stops than for unaspirated and aspirated voiceless stops, at least for male talkers.

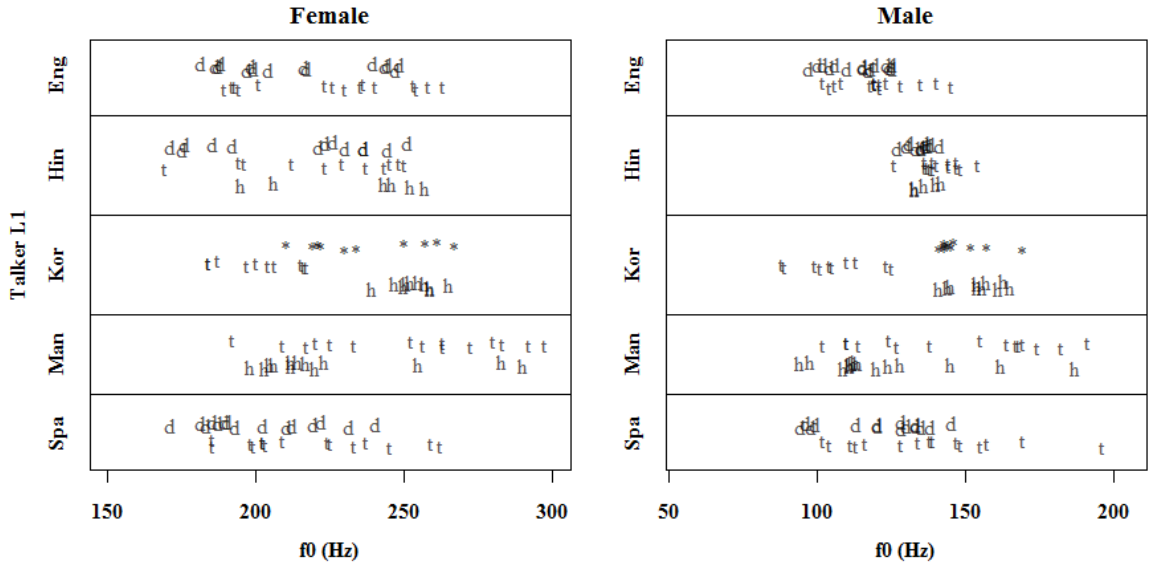

Figure 3.3: f0 in L1 productions

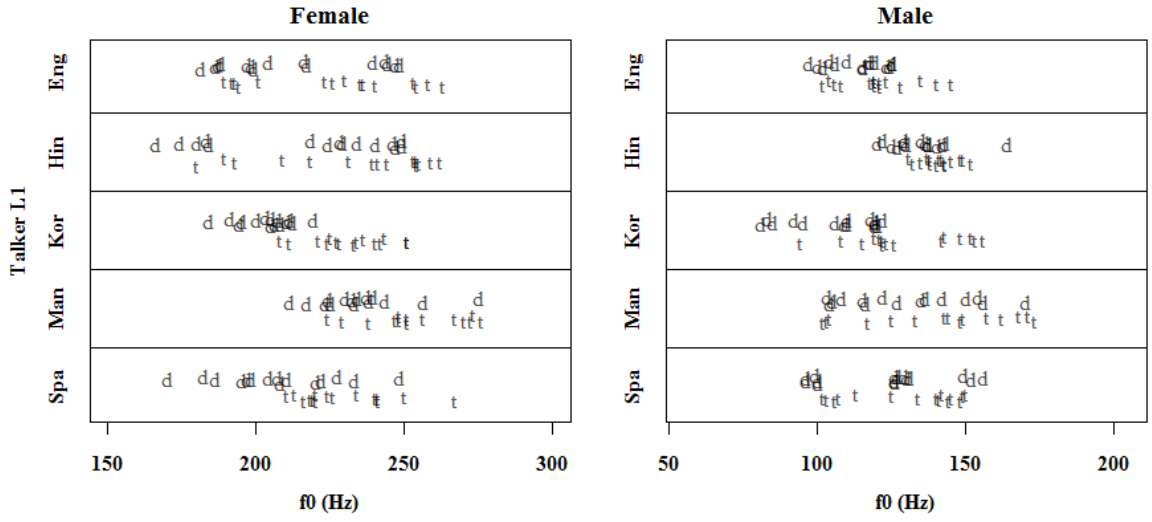

Figure 3.4: f0 in English productions

3.2.3 Spectral tilt (H1-H2)

Spectral tilt was measured as the difference in decibels between the first and second harmonics (H1-H2) over the first 25ms of the vowel. The 25ms window was chosen to match the window over which f0 was measured, and H1-H2 was the common measure of spectral tilt used by Dutta (2007) for Hindi and by Cho et al. (2002) for Korean. The values were extracted automatically using a Praat script that displayed an FFT spectrum with peaks highlighted for the first and second harmonics, such that the author could check token-by-token for errors. Spectral tilt values are shown for talkers' L1 productions in Figure 3.5, and for their English productions in Figure 3.6. Each figure contains a total of 30 data points per sex for each of the 5 language backgrounds. In these figures, the symbols represent the stop that preceded the vowel measured. The voice quality differences that have been noted for Korean, with lower H1-H2 values (pressed voice) for vowels following tense stops and higher values (breathy voice) for those following lax stops (Cho et al., 2002), are not especially striking in this set of data, especially for female talkers. Spectral tilt in Hindi distinguishes only voiced aspirated stops from other categories (Dutta, 2007), so differences cannot be seen in this investigation, as voiced aspirated stop targets were not recorded. In English productions, female talkers from most L1 backgrounds have generally breathier productions for voiceless than for voiced stop targets, while male talkers show a minimal difference at best. For L1 Korean talkers, however, this pattern is more evident in productions by male talkers than in those by female talkers.

3.2.4 Vowel duration

Vowel duration was measured as the time between the beginning and the end of the vowel. Again, the beginning of the vowel was taken to be the upward-going zero crossing

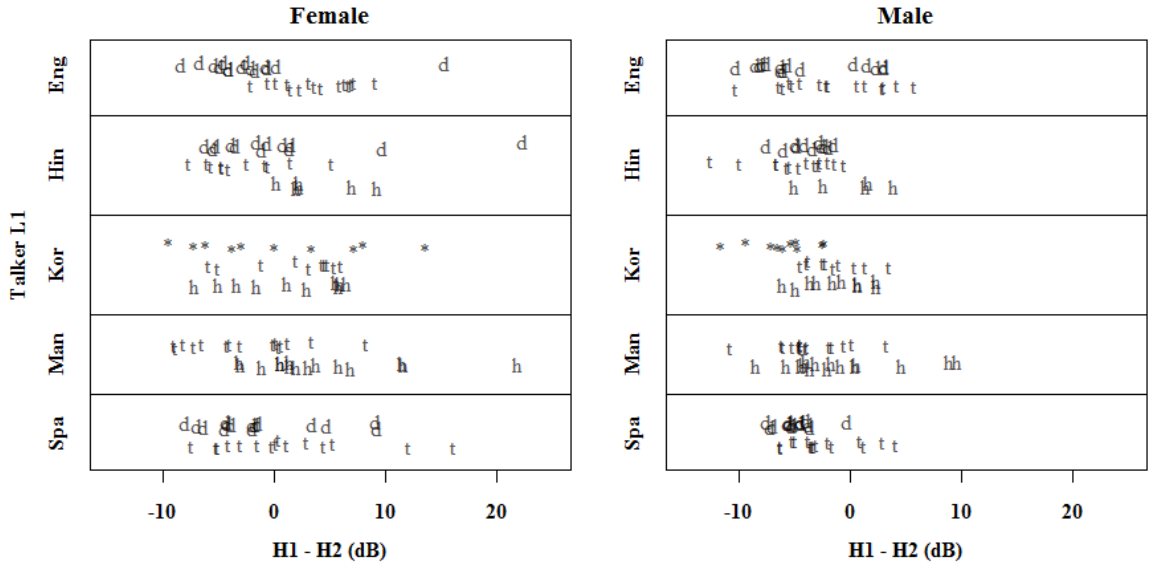

Figure 3.5: H1-H2 in L1 productions

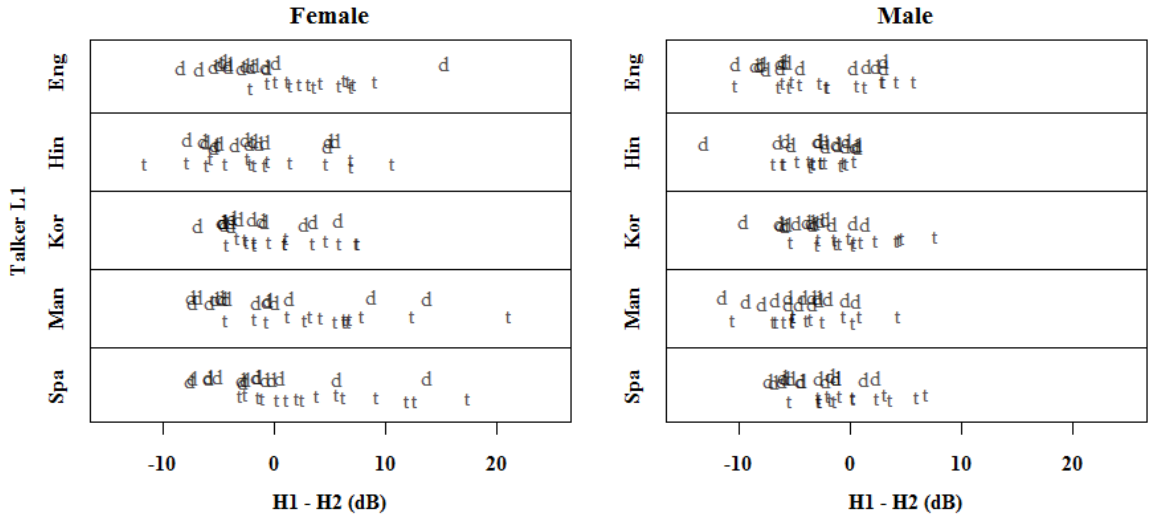

Figure 3.6: H1-H2 in English productions

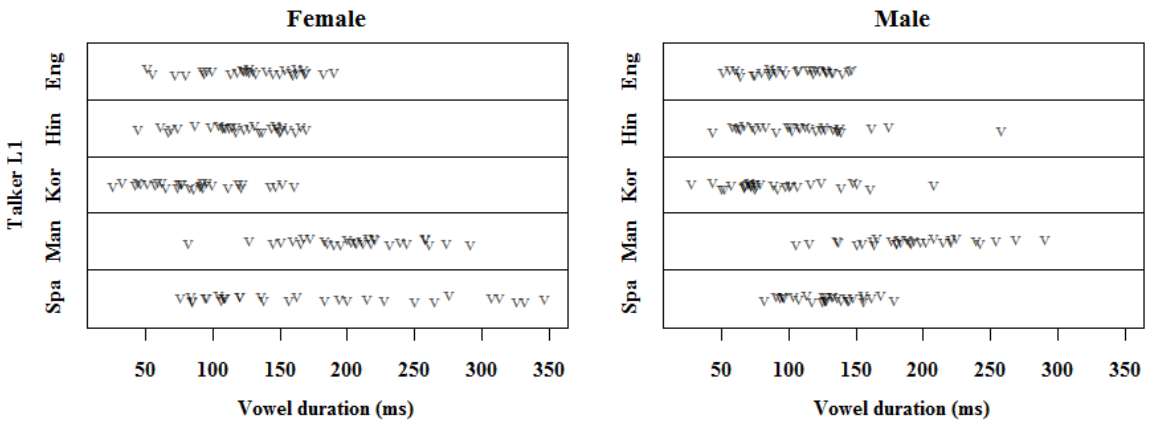

Figure 3.7: Vowel duration in L1 productions

indicating the onset of periodicity following the stop release. The end of the vowel was taken to be the final upward-going zero crossing in the periodic portion of the waveform. In cases where voicing continued through the beginning of the medial stop closure, the end of the vowel was taken to be the final upward-going zero crossing in the non-sinusoidal portion of periodicity. Vowel duration measurements are shown in Figures 3.7 and 3.8 for productions in talkers' L1s and in English, respectively. Each figure contains 30 data points per sex for each of the 5 language backgrounds. Across languages, measurements of vowel duration capture inherent durational attributes of the vowel targets themselves (e.g., /i/ is long, but /ɪ/ is not) as well as speaking rate. For readability, however, the different vowel targets are not displayed in these figures. Vowels in Spanish and especially Mandarin appear to have generally long durations, and Korean vowels are relatively short. In English, where the vowel targets for each L1 group were identical, it seems that non-native talkers generally spoke slightly more slowly than native talkers, although the differences are minimal.

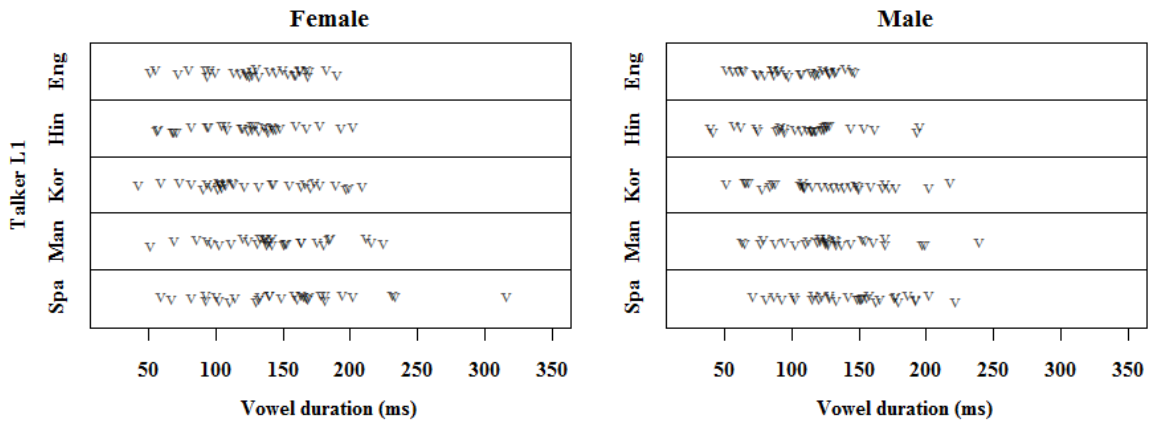

Figure 3.8: Vowel duration in English productions

3.2.5 Vowel quality

Formants 1 through 3 were measured for the middle 60% of the duration of each vowel, as defined above, in order to minimize the effects of adjacent consonants. Values were extracted automatically using a Praat script that displayed a spectrogram with overlaid formant tracks, such that the author could check token-by-token for tracking errors. While most errors were resolved by targeting a different number of formants, measurements for 11 of the 540 stimuli required hand correction. Figures 3.9 through 3.17 show the midpoint values of each of the first 2 formants. Although these values were not directly used in the analyses, such depictions of formants are substantially more familiar than those in the following section, and are provided for reference. Each of these figures contains 30 data points per sex. Some influence of talkers' L1s can be seen in their L2 English vowel productions. For instance, /i, ɪ/ tend to overlap in both F1 and F2 for L1 Korean, L1 Mandarin, and L1 Spanish talkers, none of whom have this contrast natively.

The measurements used in the analyses were the first 3 coefficients of a discrete cosine transform (DCT) for each formant. A DCT models the track of each formant as a

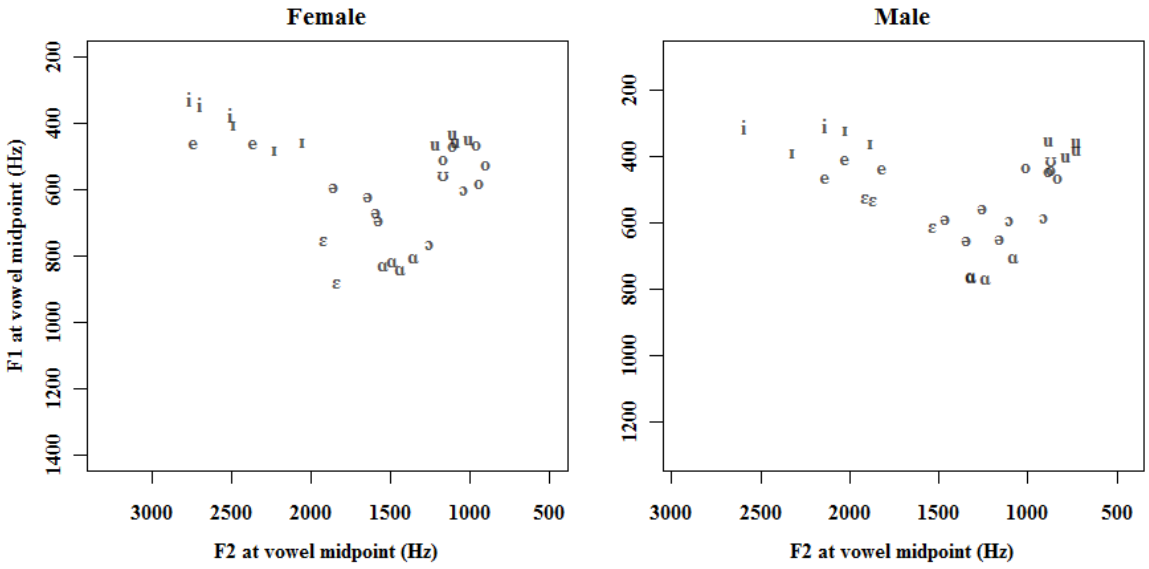

Figure 3.9: F1 and F2 in L1 Hindi productions

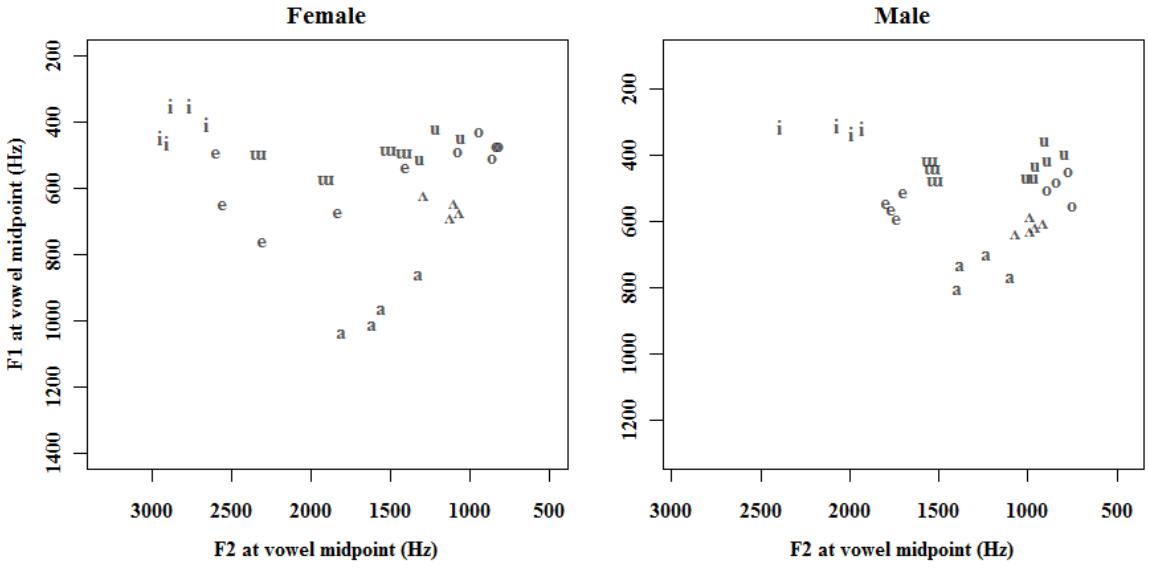

Figure 3.10: F1 and F2 in L1 Korean productions

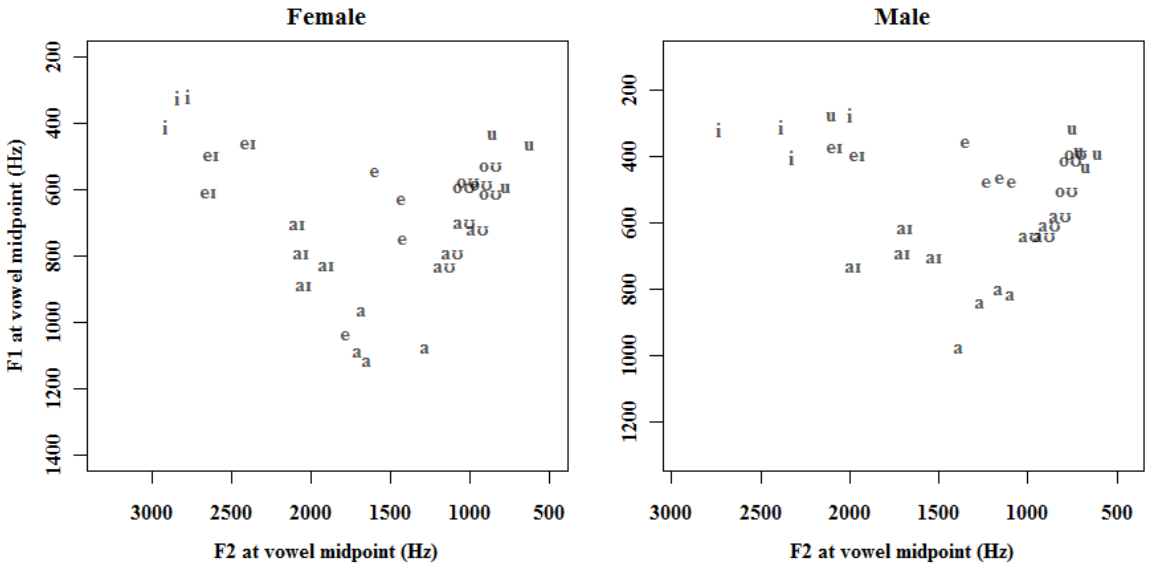

Figure 3.11: F1 and F2 in L1 Mandarin productions

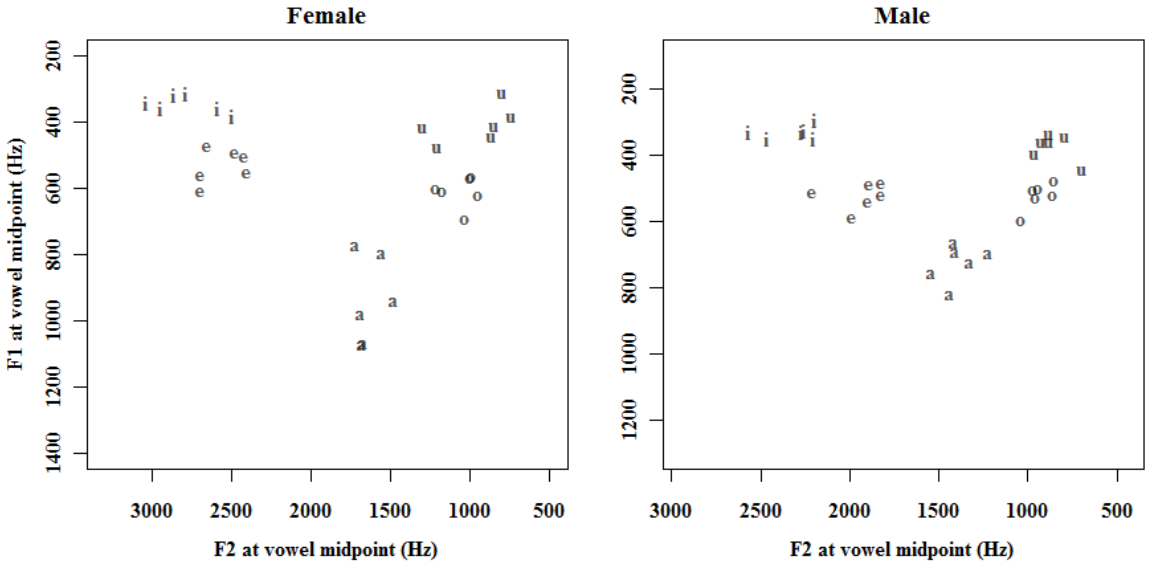

Figure 3.12: F1 and F2 in L1 Spanish productions

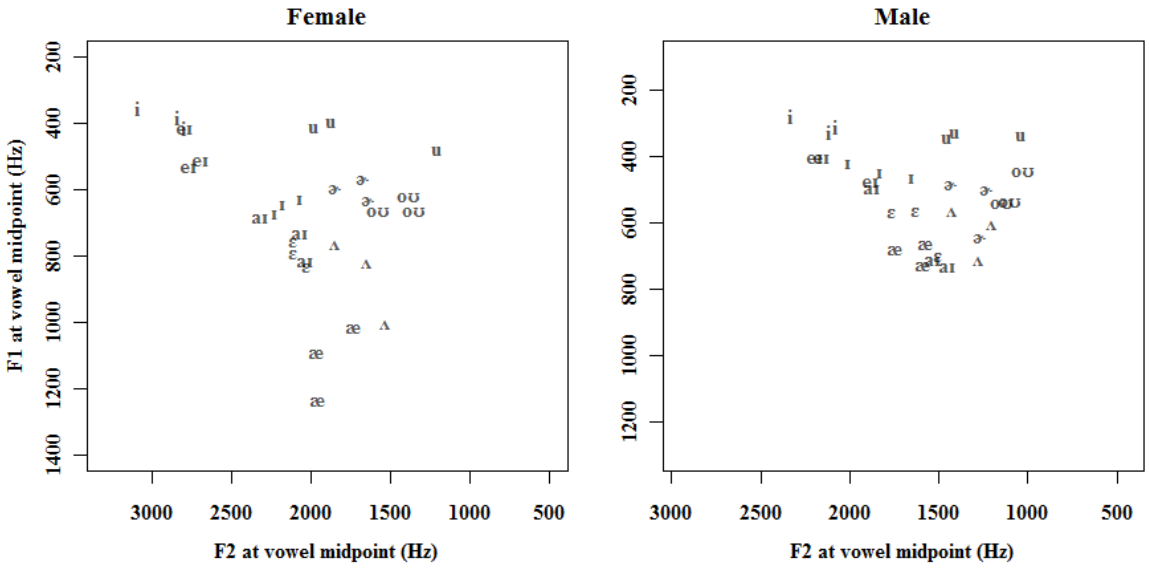

Figure 3.13: F1 and F2 in L1 English productions

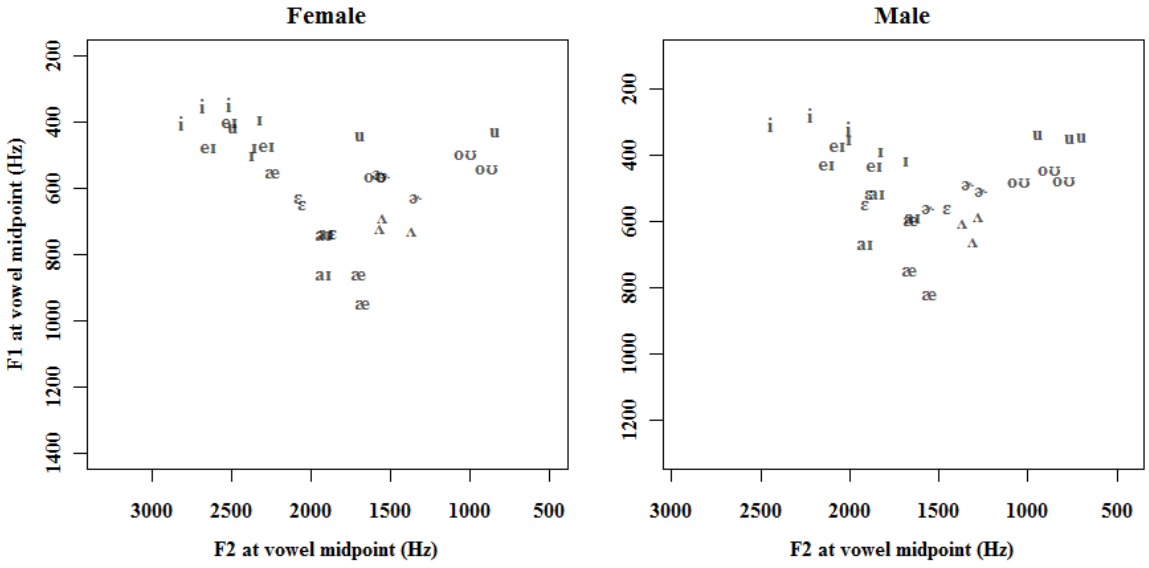

Figure 3.14: F1 and F2 in L1 Hindi talkers' English productions

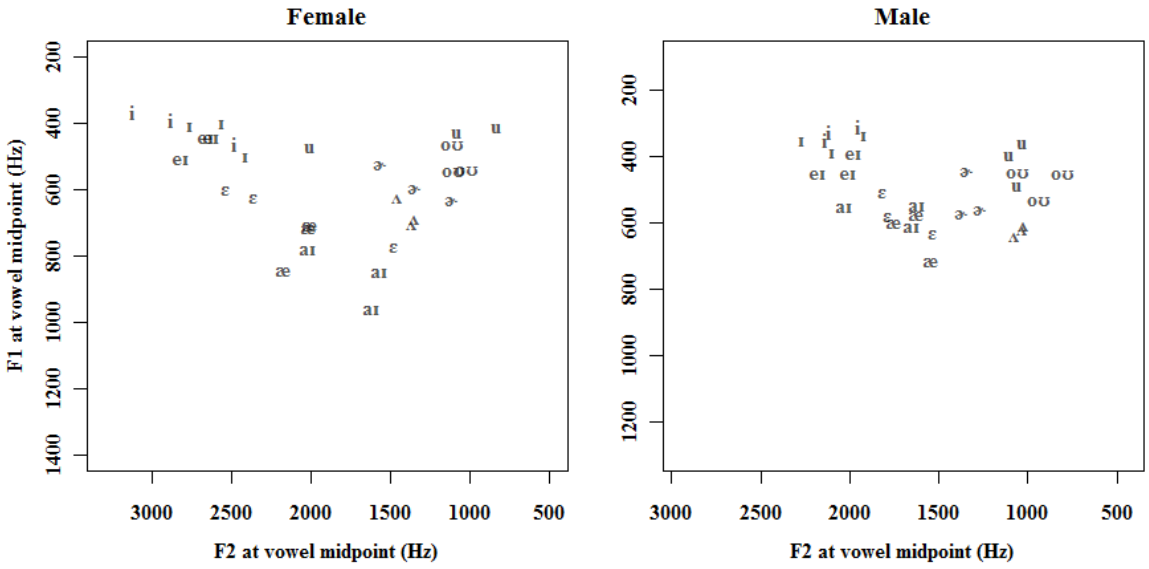

Figure 3.15: F1 and F2 in L1 Korean talkers' English productions

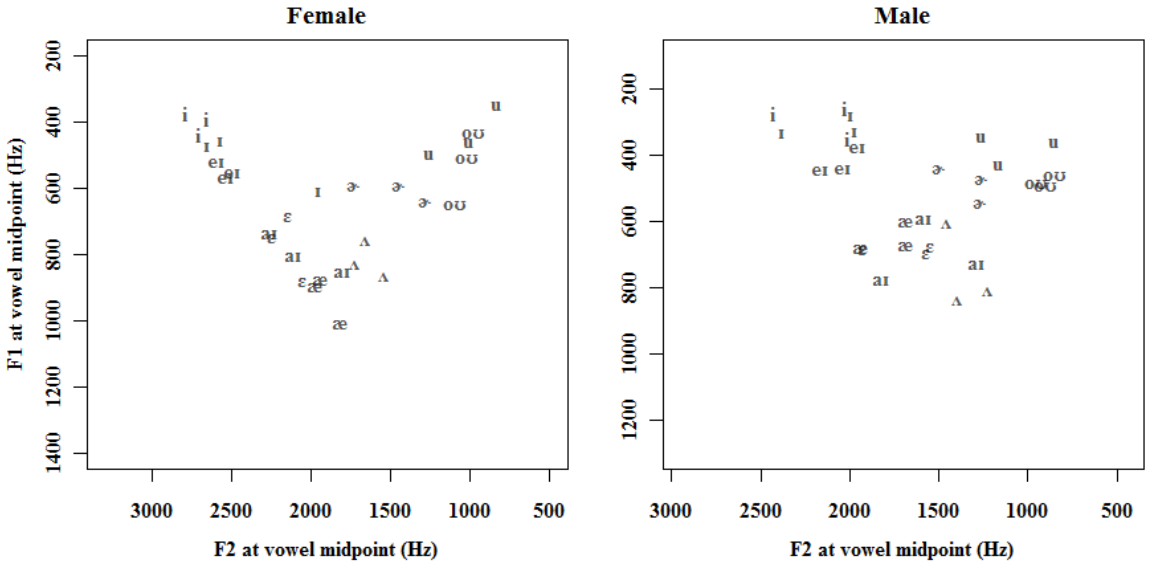

Figure 3.16: F1 and F2 in L1 Mandarin talkers' English productions

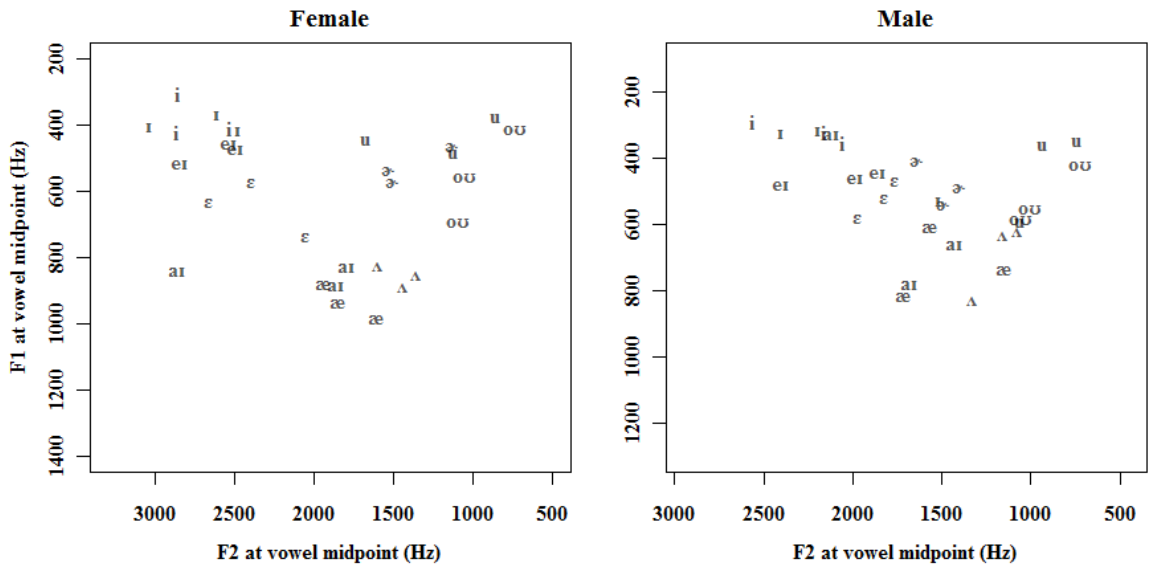

Figure 3.17: F1 and F2 in L1 Spanish talkers' English productions

sum of cosine functions with different periods. The zeroth coefficient, the DC offset, is proportional to the mean value of the formant. The first coefficient modifies a half-cycle cosine wave, and reflects the direction and magnitude of the formant track's tilt. The second coefficient modifies a full-cycle cosine wave, and relates to the formant track's curvature (Watson and Harrington, 1999). DCTs were calculated with code replicating Watson and Harrington (1999) (Plichta, 2012). Values for the zeroth coefficients are shown in Figures 3.18 and 3.19, for the first coefficients, Figures 3.20 and 3.21, and for the second coefficients, Figures 3.22 and 3.23. In these figures, the symbols indicate the identity of the formant plotted (first, second, or third). The first figure in each pair shows values for talkers' L1 productions, while the second shows values for their English productions. Each of these figures contains 30 data points per formant per sex. More extreme values for the first coefficient reflect higher degrees of tilt, while the sign indicates the direction of tilt.

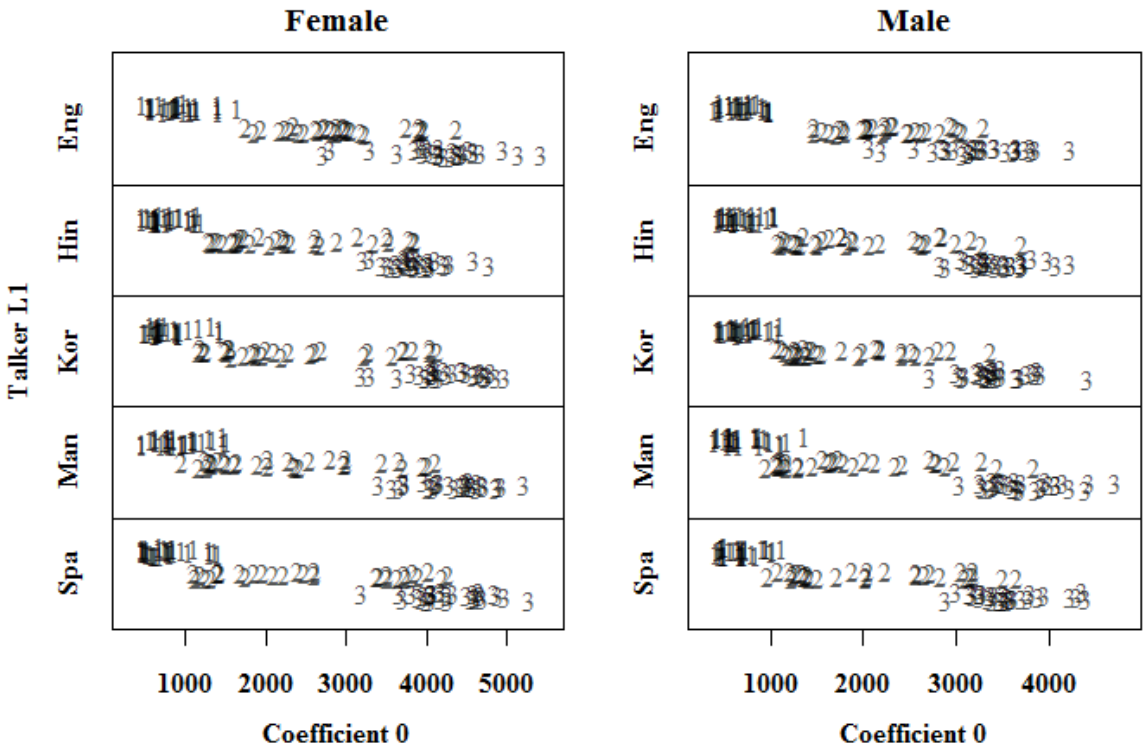

Figure 3.18: DCT coefficient 0 (mean frequency) in L1 productions

Similarly, more extreme values for the second coefficient reflect higher degrees of curvature, while the sign indicates the direction of curvature. Although vowel-specific patterns are not recoverable from Figures 3.18 through 3.23 and are not discussed here, these data provide some evidence that formant dynamics can differ across language varieties. For instance, from Figure 3.23 it is clear that the curvature of the second and third formants in English productions is somewhat more extreme for male L1 Spanish talkers than for male L1 American English talkers.

The various acoustic properties plotted in this chapter are used as predictor variables in the perception analyses detailed in Chapters 4 through 7. On the whole, the properties selected show reasonable amounts of variability in their values, which may allow them

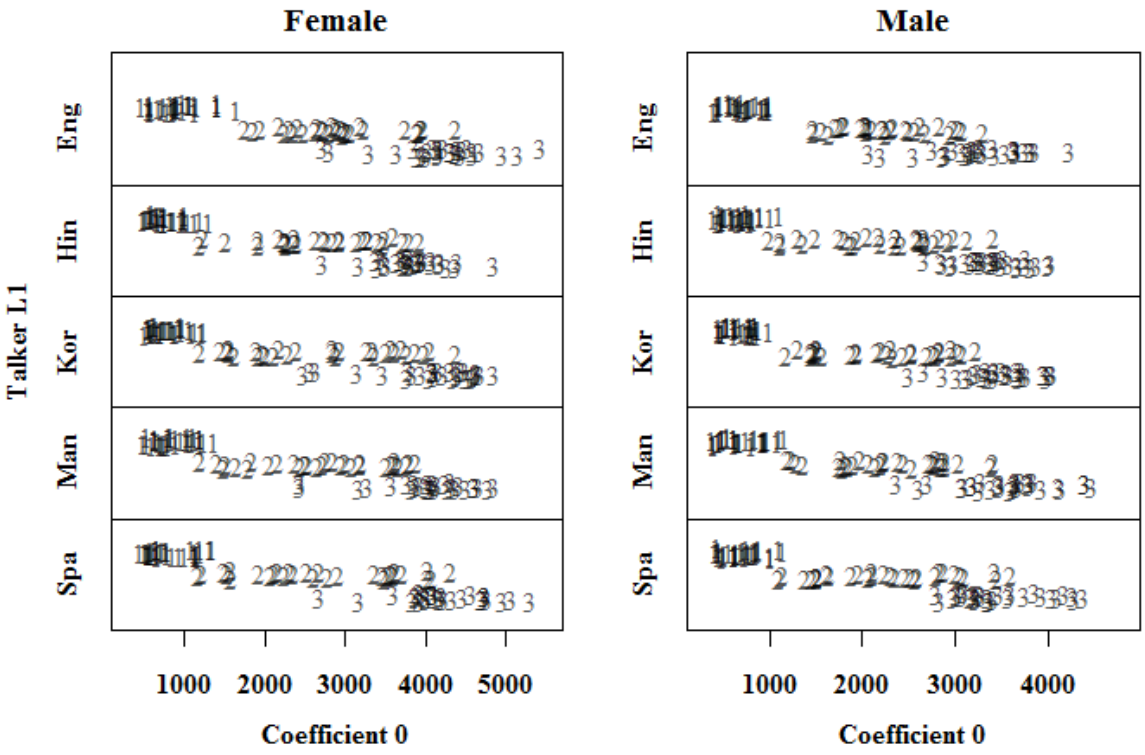

Figure 3.19: DCT coefficient 0 (mean frequency) in English productions

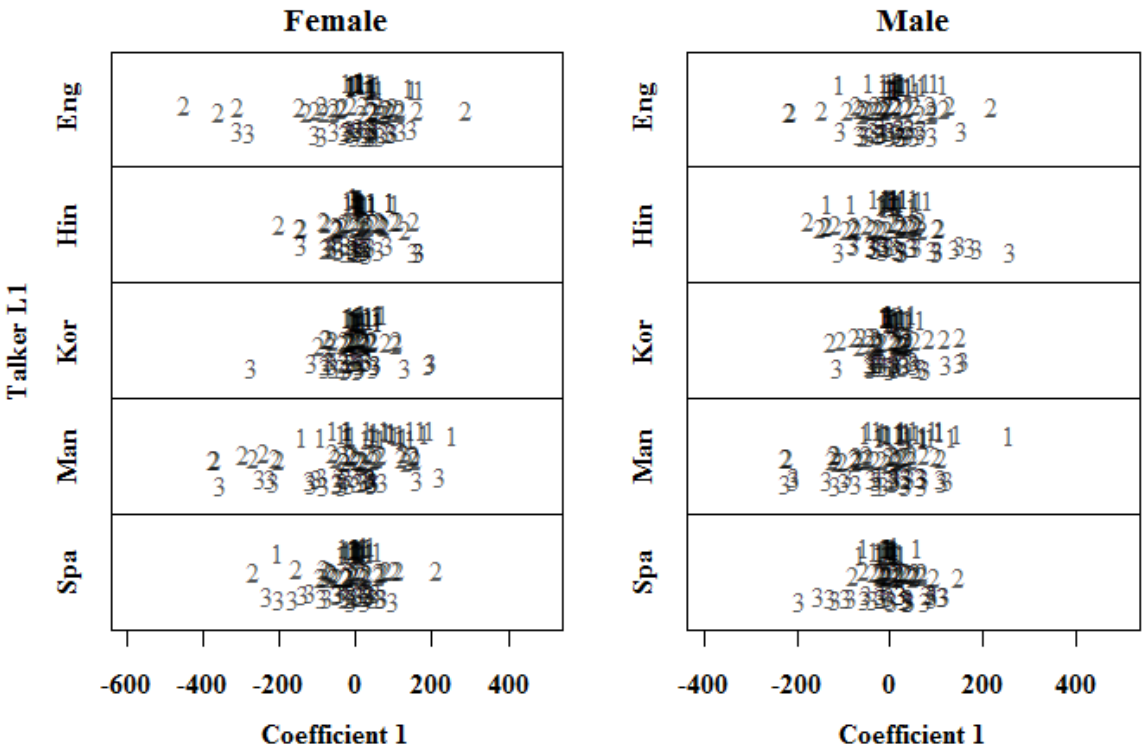

Figure 3.20: DCT coefficient 1 (tilt) in L1 productions

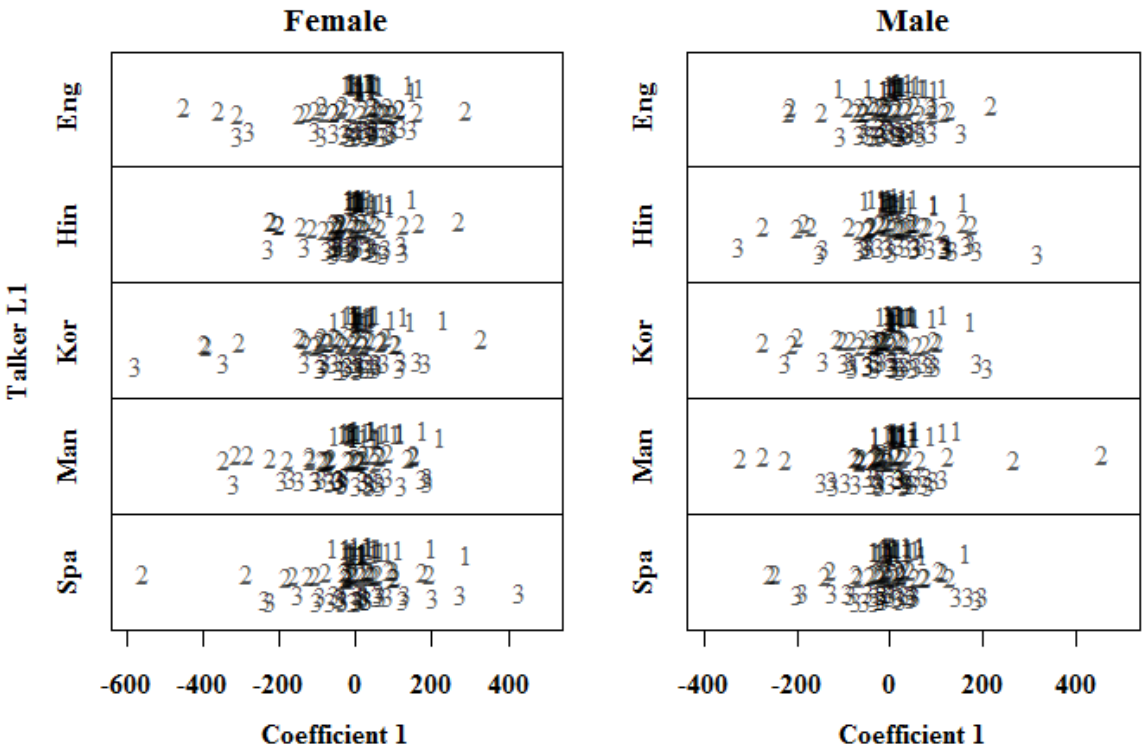

Figure 3.21: DCT coefficient 1 (tilt) in English productions

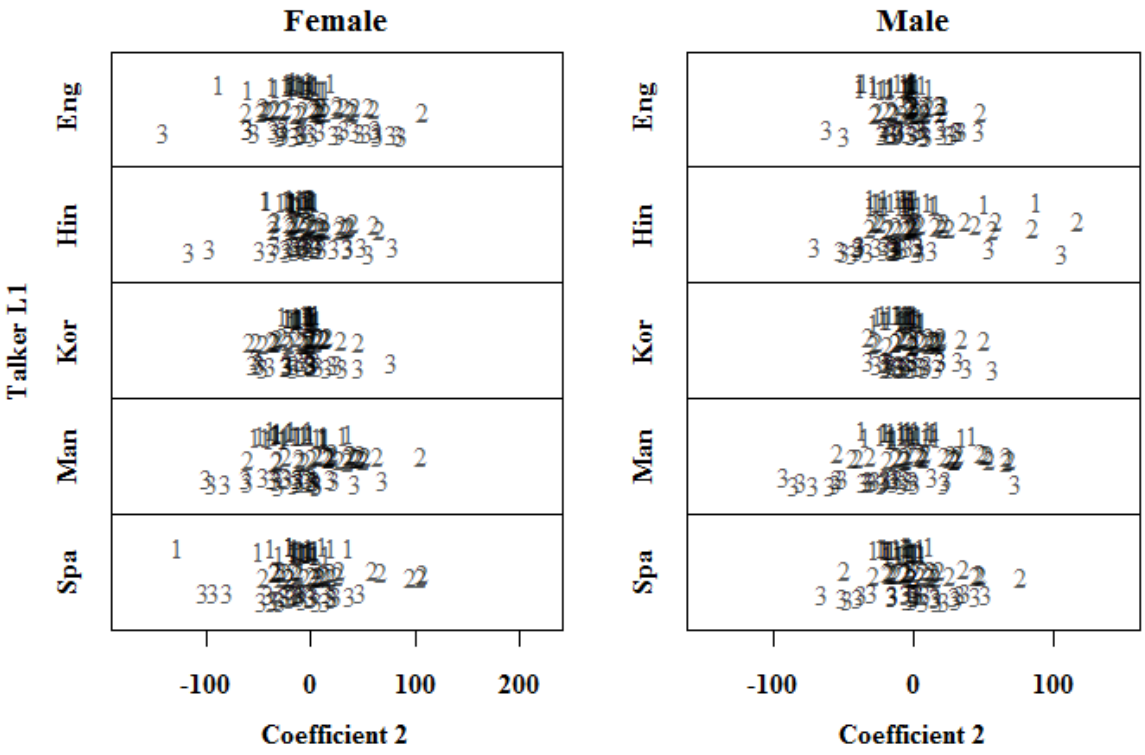

Figure 3.22: DCT coefficient 2 (curvature) in L1 productions

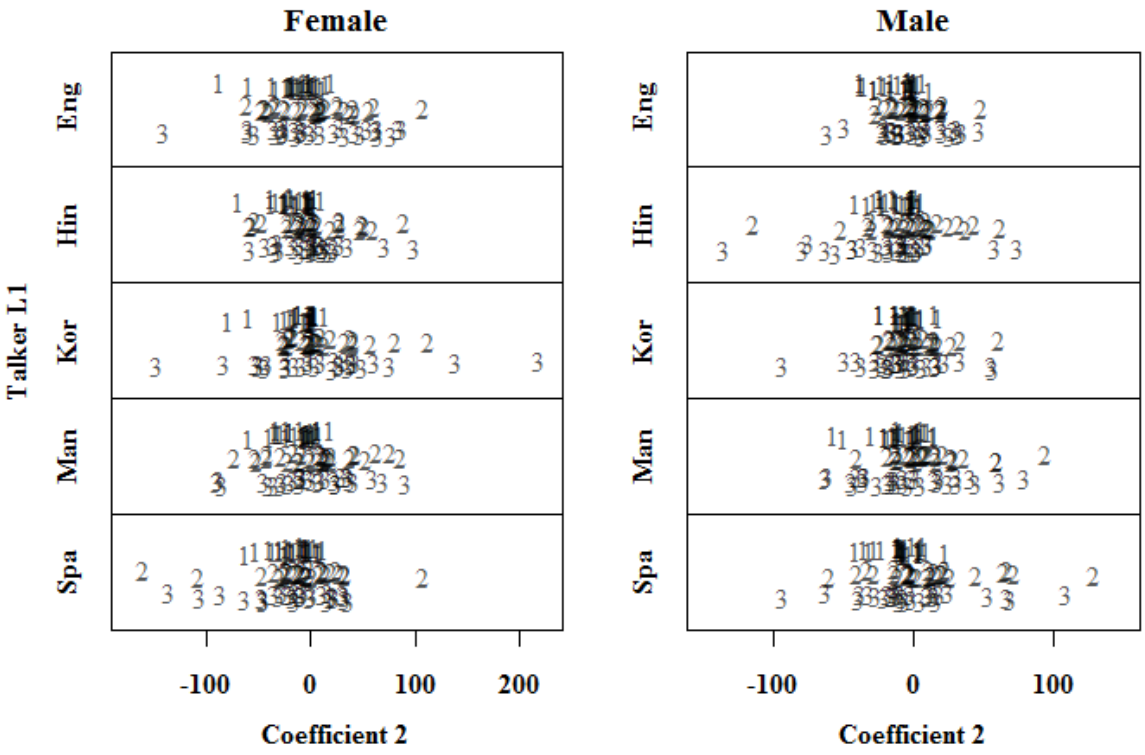

Figure 3.23: DCT coefficient 2 (curvature) in English productions

to correlate with listeners' responses. Exploration of the potential relationships between acoustics and the perceptions of foreign accentedness, non-nativeness, and non-Englishness is begun in Chapter 4, following further discussion of the experimental design.

CHAPTER 4: EXPERIMENTS 1 AND 2: RATING OF FOREIGN ACCENTEDNESS

Although previous studies have investigated the acoustic correlates of foreign accentedness, no clear patterns have emerged. The role of VOT has been especially controversial, showing a relationship to judgments about accentedness in some cases (Major, 1987; McCullough, 2013; Riney and Takagi, 1999), but not in others (Shah, 2002; Wayland, 1997). In Experiments 1 and 2, listeners heard samples of English produced by native talkers of American English, Hindi, Korean, Mandarin, and Spanish, and rated the degree of accentedness perceived in each stimulus. The full set of acoustic properties detailed in Chapter 3 was included in the analysis.

4.1 Experiment 1: CVs

4.1.1 Methods

Procedure

Listeners began by filling out the native speaker version of the language background questionnaire presented in Appendix A. The experimental procedure consisted of a rating task and a free classification task, which were separated by a brief sentence completion task as a filler. For half the listeners, the rating task was administered first and the free classification task third, while the other half of listeners performed the experiment in the opposite order. Because equal numbers of participants completed the tasks in each order, no effect of order was explored. Listeners generally completed the three tasks in 35 to

40 minutes. All tasks were performed on computers running Windows XP with listeners wearing Sennheiser HD 280 Pro headphones. The rating and sentence completion tasks were run in EPrime 1.1. Details about the free classification task are given in Chapter 7.

In the rating task, listeners saw a word displayed orthographically on the computer screen and then heard a CV syllable extracted from the beginning of the word. They were asked to make a judgment about the talker's degree of accent by sliding a bar along a continuous rating line labeled "no foreign accent" on the left and "strong foreign accent" on the right, as in Figure 4.1. A continuous rating line was chosen because listeners have fairly sensitive responses in rating tasks with non-native speech (see Flege et al., 1995), and a continuous rating scale allowed for the possibility of better correlation with continuous acoustic measures. The sliding bar began at the left end of the rating line for each trial, on the assumption that this label would be inappropriate for many of the stimuli, thus encouraging listeners to use the rest of the rating line.[2] The final location of the bar was recorded as an integer from 0 to 100, which represented the distance from the left end of the rating line. There were 20 practice trials, so that listeners could become comfortable with the task, and 300 test trials. During the testing phase, listeners were permitted to rest briefly, if desired, after each set of 50 trials. The stimuli were randomly ordered by the experiment presentation software, and appeared in a unique order for each listener. After the rating task, listeners were asked what accents they thought they heard and what they thought they based their ratings on, and typed their responses into text boxes.[3]

To ensure that listeners understood the task, the instructions preceding the rating task required them to slide the bar sequentially to four clearly indicated areas of the rating line.

[2]The labels were not counterbalanced due to this decision to begin on the "no foreign accent" side, as sliding right-to-left to indicate greater accentedness was expected to be counterintuitive for monolingual speakers of a language written left-to-right.

[3]These responses are not analyzed in the present work.

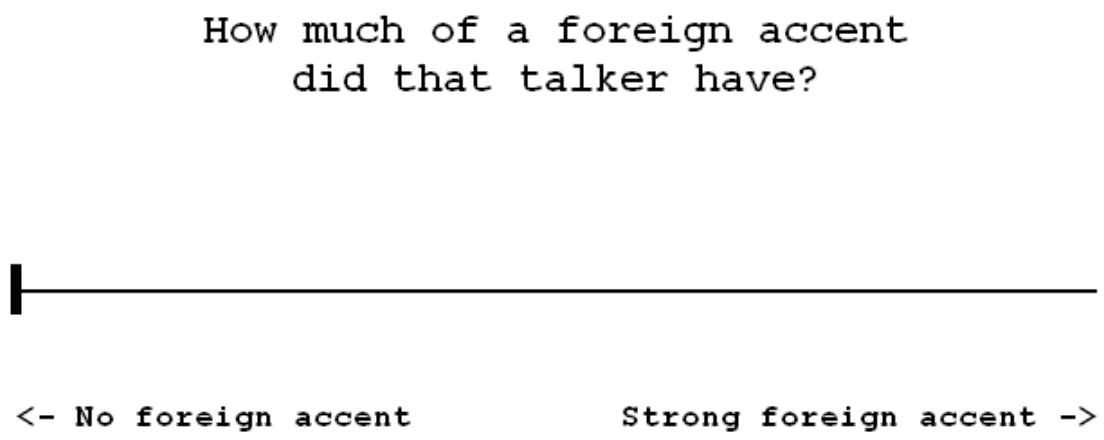

Figure 4.1: Rating screen

If a listener failed to slide the bar to the area indicated on more than one of these four trials, this was considered "failure to follow instructions" regarding the use of the rating line, and the listener's responses were not included in the rating or free classification analyses. The counts of listeners eliminated for "failure to follow instructions" reported in this chapter and in Chapters 5 and 6 also include those who failed to follow instructions for the free classification task, about which more detail is provided in Chapter 7.

The rating and free classification tasks shared some auditory stimuli, and required listeners to make judgments about related characteristics. In an effort to minimize the effect of rating on free classification or vice versa, these tasks were separated by an unrelated filler task which directed listeners' attention away from the details of speech and toward language itself. In this sentence completion task, listeners saw and heard a short sentence that was missing its final word, and typed a guess at the missing word in a text box on the computer screen. There were 4 practice trials, so that listeners could become comfortable

with the task, and 16 test trials. The stimuli were randomly ordered by the experiment presentation software, and appeared in a unique order for each listener.

Stimuli

The 300 auditory stimuli in the rating task included 5 repetitions of CVs extracted from the beginning of each target word shown in Table 3.1, produced once by a talker from each of the 5 L1 backgrounds, with no target produced by more than 3 female or 3 male talkers. Each of the 30 talkers provided 10 of the 300 stimuli. Stimuli were selected such that the 10 CVs produced by each talker included each of the 10 vowels exactly once and equal numbers of voiced and voiceless consonants, with no more than 2 of any single consonant. The 20 practice stimuli included 5 repetitions of CVs extracted from the beginning of the 4 additional target words recorded as practice words by the talkers, produced once by a talker from each of the 5 L1 backgrounds. Within each L1 background, the 4 words were produced by 2 female and 2 male talkers; 1 randomly chosen female and 1 randomly chosen male talker were omitted to limit the number of repetitions of each practice word as well as the amount of time required to complete the practice session. The final 25ms of each CV stimulus gradually decreased in intensity to reduce the audibility of any coarticulatory cues with the following consonant.

The 16 auditory stimuli in the sentence completion task were taken from recordings of BKB-R List 7 (Bamford and Wilson, 1979) by 16 native speakers of American English in the Buckeye GTA Corpus (Hardman, 2010), with the last word excised and the final 50ms of the resulting stimulus gradually decreasing in intensity to reduce the audibility of any coarticulatory cues with the missing word. As in the other tasks, half the voices were female and half were male, but the talkers were different from the L1 American English talkers in the other tasks. The 4 practice stimuli were taken from recordings of BKB-R List

8 (Bamford and Wilson, 1979) in the same corpus, read by 2 of the female talkers and 2 of the male talkers represented in the test stimuli.

Participants

Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual American English-speaking listeners (13 females) were considered here. In the perception studies in this work, "American English monolingual" was strictly defined as an individual who reported only English as their native language and only English spoken to them by childhood caretaker(s), who had never lived outside the United States, and whose childhood caretaker(s) had not grown up outside the United States. Data from 14 additional listeners were excluded for the following reasons: non-native English speaker (2), early exposure to another language (5), lived abroad (1), caretaker(s) grew up abroad (1), self-reported hearing loss (1), failure to follow instructions (3), and technical problems with the experiment presentation software (1).

Analyses

As mentioned in Chapter 2, retroflexion was not measured acoustically. However, if some L1 Hindi productions were characterized by retroflexion, and if this property influenced the perceptual responses, listeners should have assigned higher ratings for L1 Hindi talkers' productions of stimuli with alveolar stop targets than for their productions of stimuli with stop targets at other places of articulation. This possibility was explored in the ratings from Experiment 1. First, for each production, 20 accentedness ratings, one from each of the 20 listeners, were combined into a single mean rating. Next, from the mean

rating for each L1 Hindi production was subtracted the mean rating for the L1 American English production of the same sequence, so that sequences which received generally higher ratings across the board would not skew the results. The mean of these "corrected" ratings over each L1 Hindi talker's productions of stimuli with alveolar stop targets was not different from the mean of the corrected ratings over his or her productions of stimuli with labial and velar stop targets, as revealed by a paired t-test ($t(5) = 0.7640$, $p > 0.05$), indicating that an acoustic measure of retroflexion would not necessarily improve the analysis below.

Because participants saw the target word displayed prior to hearing and rating each stimulus, it is assumed that they were not basing their ratings on the absolute measures of the stimuli, but on how much the stimuli diverged from their expectations. As in Munro (1993) and Wayland (1997), this is captured in the analysis by using difference values, rather than raw measurements, for each acoustic property. The listener's "expectation" was represented by the mean over measurements of that acoustic property in productions of the same target word by the 3 L1 American English talkers of the same sex as the talker in question. For instance, the VOT difference value for an L1 Hindi female talker's production of *pity* was calculated by subtracting from the raw measurement the mean VOT for the 3 L1 American English female talkers' productions of *pity*. "Expectations" were based only on the productions of same-sex talkers because the values of spectral acoustic parameters generally differ between men and women. While most of the L1 American English measurements were based on recordings not selected for use in the perception experiment, the stimuli did contain one L1 American English production of each of the 60 unique targets; thus, difference values for 60 productions were calculated with the value for that production itself having contributed to the "expectation." Because VOT values are

often not averaged across lead voicing and lag voicing categories (see Lisker and Abramson, 1964), 12 instances of lead voicing were dropped from the 360 L1 American English measurements, such that the values for some VOT "expectations" were means over fewer than three productions.

Three parameterizations of each difference value were calculated: signed and squared differences, as in Munro (1993), as well as absolute values of differences, as in McCullough (2013). Like squared differences, absolute differences capture the magnitude of deviation from the "expectation" while abstracting over the direction of deviation. Unlike squared differences, absolute differences do not exaggerate the effects of larger deviations. As 3 parameterizations of each of the 13 acoustic variables discussed in Chapter 3 yielded a total of 39 possible independent variables before considering any non-acoustic properties or any interactions, a factor analysis was performed on the acoustic measures to explore how these 39 potential variables could sensibly be reduced.

In an analysis with 13 factors, each factor loaded with high values on the absolute and squared difference parameterizations of a single acoustic property, as shown in Table 4.1. Loading values for the signed differences were somewhat smaller in magnitude, and loading values for other acoustic properties were much smaller in magnitude; the only case in which the latter exceeded 0.3 was for Factor 7, which loaded on signed and squared differences of F1 tilt with values less than 0.4. Additionally, each factor loaded highly on a unique acoustic property, such that all 13 acoustic properties measured seemed to be captured by the 13 factors. This suggested that there was sufficiently little multicollinearity among the 13 acoustic properties to include all of them in the analysis. Examination of the correlation coefficients of all possible pairings of the 13 acoustic variables within each parameterization (excluding correlations of identity) revealed a maximum of 0.39 for absolute

Factor	Property	Loading for absolute difference	Loading for squared difference
1	VOT	0.965	0.966
2	vowel duration	0.938	0.942
3	F1 frequency	0.902	0.960
4	F2 tilt	0.960	0.884
5	f0	0.954	0.956
6	F3 frequency	0.932	0.950
7	F2 curvature	0.826	0.925
8	H1-H2	0.896	0.967
9	F2 frequency	0.953	0.897
10	F1 curvature	0.941	0.886
11	F3 curvature	0.830	0.967
12	F1 tilt	0.867	0.883
13	F3 tilt	0.917	0.907

Table 4.1: Factor analysis loading values for best acoustic property matches

differences, and a maximum of 0.63 for squared differences. As the 13 orthogonal factors identified by the factor analysis were readily interpretable as either absolute or squared differences, and the highest pairwise correlations were in the set of squared differences, only absolute differences were used for all acoustic analyses of rating responses.

Again, each stimulus was rated 20 times, once by each of the 20 listeners. In the models discussed below, these 20 ratings were combined into a single mean accentedness rating for each item. These mean rating values were bimodally distributed, as shown in the left panel of Figure 4.2; vertical ticks along the x-axis indicate the 300 mean rating values. This bimodality seemed to arise from a general bias to choose ratings near the labeled ends of the rating line. Examination of the ratings from each individual listener confirmed that listeners generally did use the entire rating line, although not uniformly.

Rating	Logit
0 (adjusted to 1)	-4.60
25	-1.10
50	0.00
75	1.10
100 (adjusted to 99)	4.60

Table 4.2: Mapping between ratings and their logit-transformed values

To make the distribution more appropriate for linear regression analyses, the logit values of the mean ratings were calculated. This approach treats the ratings as probabilities—specifically, listeners' estimates of the probability that each item was produced with a strong foreign accent.[4] The logit, or "log-odds" (that is, the logarithm of the odds) is a transformation of probability values that reduces crowding near the bounds at 0 and 1. Several representative ratings from the 100-point rating scale and their corresponding logit-transformed values are shown in Table 4.2. As the table suggests, this approach requires ratings at the extremes to be adjusted, as the logit transformations of 0 and 1 are infinite. The effect of the logit transformation on the mean rating values for Experiment 1 is shown in the right panel of Figure 4.2. For these data, no values were extreme enough to need adjustment.

By using mean rating values, a production which all listeners rated as having a moderate degree of foreign accent is indistinguishable from a production which some listeners rated as having "no foreign accent" and other listeners rated as having a "strong foreign

[4]Admittedly, the inclusion of the scalar term "strong" makes it somewhat complex to conceptually map this to a probability. The approach is more straightforward for the rating scales used in Experiments 3 through 6, however, and is used here both for consistency and for its effectiveness in dealing with the bounded rating data.

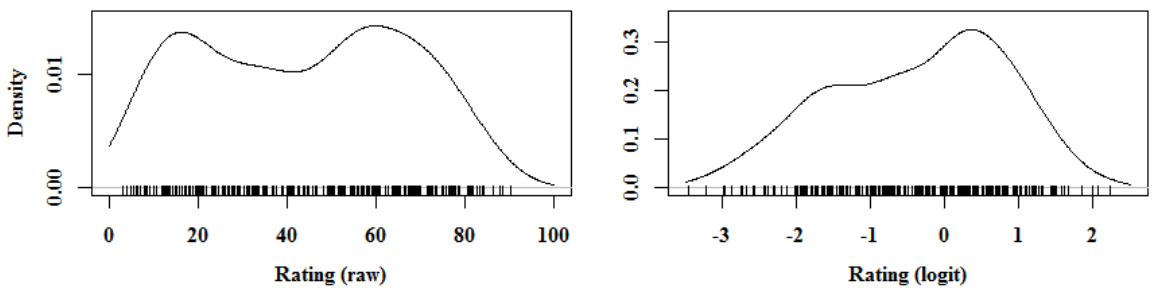

Figure 4.2: Raw and logit-transformed ratings from Experiment 1

accent." The use of each listener's individual ratings was also considered. However, the bimodality of these individual responses was not sufficiently reduced by the logit transform. As individual differences between listeners are not at issue in the present work, and means seem to reasonably reflect the perception of non-native speech by L1 American English listeners as a community, mean ratings were used instead.

4.1.2 Results

For this experiment and subsequent ones, linear mixed effects regression models were used to evaluate the acoustic correlates of ratings. In Experiment 1, the dependent variable in all models was accentedness rating, quantified as the logit of the mean rating across listeners for each item. The grand model included random intercepts for talker and word, and fixed effects of the 13 acoustic variables, the interaction of each of these acoustic variables with talker sex (female or male), and the interaction of VOT, f0, and H1-H2—that is, the consonant-related variables—with target voicing category (voiced or voiceless). All acoustic variables were centered. Interactions with talker sex were included because the acoustic variables were not explicitly normalized for talker sex, although the use of same-sex reference groups in the calculation of difference values, as described in Section 4.1.1, somewhat

attenuated any sex-based acoustic disparity. Interactions with target voicing category were considered because foreign accentedness ratings of voiced and voiceless stops have previously been found to have slightly different acoustic correlates (McCullough, 2013). While different acoustic correlates have also been found for individual vowels (Munro, 1993), the large numbers of both vowel-related acoustic variables (10) and target vowels (10) in the present investigation made these interactions impossible to consider. Random intercepts for talker were included because some talkers may sound more accented than others in ways that are not accounted for by other variables, and random intercepts for word were included because some words may serve as better vehicles of foreign accent than others in ways that are not accounted for by other variables.

Multiple models were considered, beginning with the largest and stepping down. Random effects and the interactions between acoustic properties and stop voicing categories were considered for removal, and were eliminated if log-likelihood ratio testing indicated no reduction in model fit. The simple fixed effects of acoustic properties were not considered for removal, as one purpose of this experiment was to test which acoustic properties were correlated with perceptual ratings when many acoustic properties were evaluated simultaneously, as was done by Munro (1993) and Wayland (1997). Likewise, the interactions between acoustic properties and talker sex were not considered for removal, as these were included to identify any problems created by the lack of explicit normalization of the acoustic measures. Significance of the fixed effects was evaluated using Markov Chain Monte Carlo (MCMC) sampling.

Log-likelihood ratio testing (α = 0.05) indicated that the random intercept for word did not contribute significantly to the model's fit, so it was removed. Subsequent elimination of the interactions of f0 and H1-H2 with target voicing category was similarly justified.

Property	Coefficient	t value	Significance
VOT	0.0117	5.610	$p < 0.001$
F1 frequency	0.0026	4.177	$p < 0.001$
F2 frequency	0.0008	3.368	$p < 0.01$
F3 frequency	0.0005	2.508	$p < 0.01$
F2 tilt	0.0034	3.083	$p < 0.01$
VOT:voicing	0.0049	2.554	$p < 0.01$
H1-H2:sex	-0.0439	-2.225	$p < 0.05$

Table 4.3: Significant fixed effects for Experiment 1

However, the interaction of VOT with target voicing category was found to contribute significantly to model fit, and thus was retained. Target voicing category was coded as a sum contrast, with voiced as -1 and voiceless as 1. Talker sex was also coded as a sum contrast, with female as -1 and male as 1. If the role of VOT differs for voiced as opposed to voiceless stops, or if the role of any acoustic property differs for female as opposed to male talkers, such effects should be revealed as interactions. Significant fixed effects for Experiment 1 are shown in Table 4.3, and their relationships to ratings are plotted in Figures 4.3 (simple effects) and 4.4 (interactions).

The simple fixed effects of VOT, F1 frequency, F2 frequency, F3 frequency, and F2 tilt were found to be significant. Positive coefficients indicated that as these acoustic measures deviated more from native talker norms, accentedness ratings increased. A significant interaction between VOT and target voicing category reflected a higher slope for the relationship between ratings and VOT for voiceless as compared to voiced stop targets, as shown in Figure 4.4, with lines plotted from the minimum to the maximum difference value for each category. This interaction resulted from the fact that the large variation in the extent of lead VOT for voiced stop targets was associated with only moderate variation

in accentedness rating, while for voiceless stop targets, the smaller variation in the extent of deviation from the American English norm of long lag VOT was associated with a relatively large variation in accentedness rating. The significant interaction between H1-H2 and talker sex is also shown in Figure 4.4. As H1-H2 differed more from the native talker norm, accentedness ratings increased for stimuli produced by female talkers, but decreased for stimuli produced by male talkers. However, this interaction seemed to result from the disproportionate influence of a small number of female talker stimuli with especially high H1-H2 difference values. Without these stimuli, the slopes for female and male talkers would have been more similar, and the interaction may not have reached significance.

The random intercepts for each talker are provided in Table B.1 in Appendix B. The intercepts for L1 American English talkers were negative, while they were positive for all L1 Hindi talkers except H3f, who was judged to be targeting American English. For the other L1 backgrounds, the patterns were somewhat more varied, indicating individual differences in perceived foreign accentedness among talkers from the same language backgrounds.

The results of Experiment 1 confirmed a relationship between perceived foreign accent and VOT that has been widely (Major, 1987; McCullough, 2013; Riney and Takagi, 1999), although not universally (Shah, 2002; Wayland, 1997), reported previously. McCullough (2013) noted a difference in this relationship for voiceless as compared to voiced stop targets, captured in the present data as a significant interaction. This difference probably reflected that voiceless stops produced with short lag VOT, as by L1 Hindi and some L1 Spanish talkers, are phonetically what L1 American English listeners might expect for voiced stop targets. Because the listeners knew that the target was meant to be voiceless, these productions crossed a category boundary and sound highly accented to them. In contrast, while extreme instances of prevoicing may not have sounded like good productions of

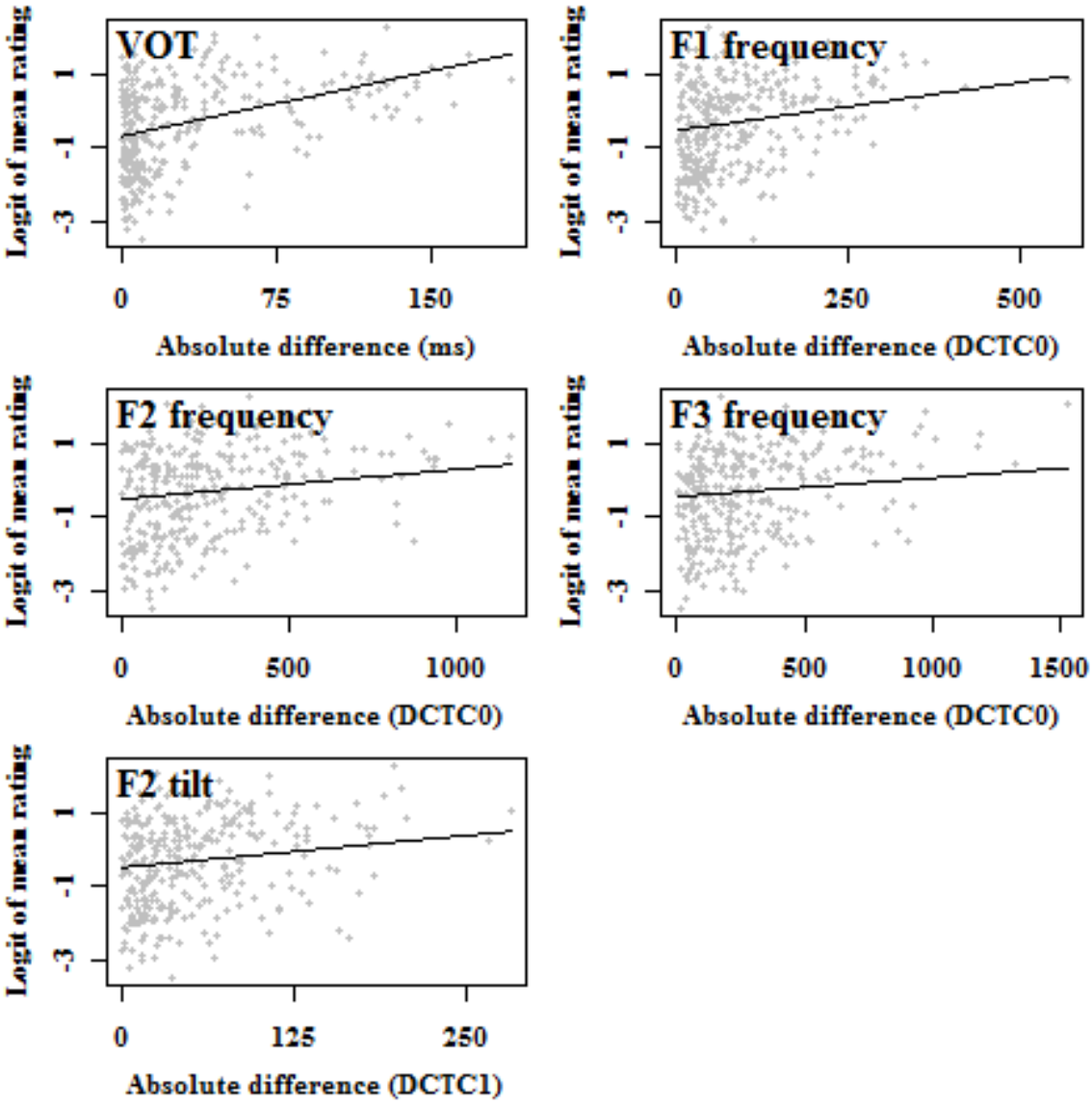

Figure 4.3: Significant simple fixed effects from Experiment 1

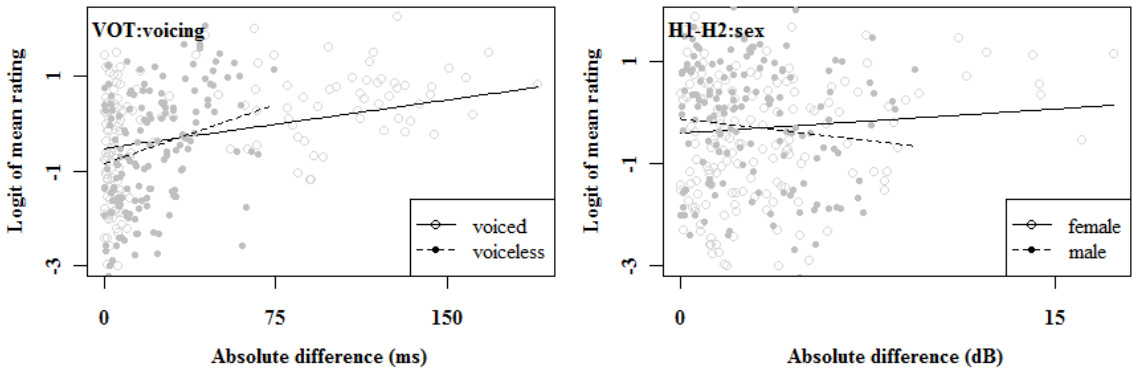

Figure 4.4: Significant fixed interactions from Experiment 1

voiced stop targets, they also did not map to a different segment. Many of the vowel-related results also supported earlier accounts: Munro (1993) found that static measures of F1 and dynamic measures of F2 were correlated with accentedness ratings, and a similar study by Wayland (1997) showed the importance of static F2 values. In short, accentedness ratings of CV-length stimuli showed relationships to acoustic properties of both consonants and vowels, and to both temporal and spectral aspects of speech, in ways consistent with prior findings.

4.2 Experiment 2: Words

4.2.1 Methods

The procedure and materials for Experiment 2 were identical to those used in Experiment 1, except that the full word was played rather than the initial CV. Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual American English-speaking listeners (15 females) were considered in this analysis. Data from 9 additional listeners were excluded

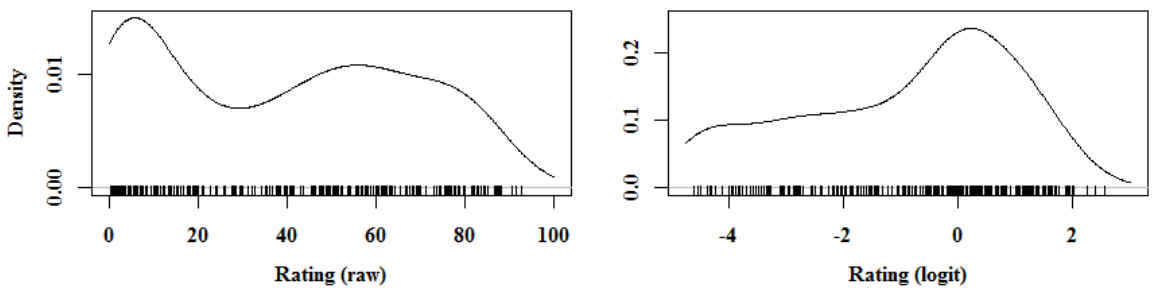

Figure 4.5: Raw and logit-transformed ratings from Experiment 2

for the following reasons: non-native speaker (3), native bilingual (1), early exposure to another language (2), caretaker(s) grew up abroad (1), and failure to follow instructions (2). Data from 12 additional and otherwise eligible listeners were excluded because their responses to a follow-up question were not recorded due to a coding error.

The influence of potential retroflexion in L1 Hindi productions on rating patterns was evaluated for Experiment 2 as it was for Experiment 1. Again, a paired t-test revealed that the mean of the corrected ratings over each L1 Hindi talker's productions of stimuli with alveolar stop targets was not different from the mean of the corrected ratings over his or her productions of stimuli with labial and velar stop targets ($t(5) = -1.3026$, $p > 0.05$), offering little motivation to add an acoustic measure of retroflexion to future models.

The analysis for Experiment 2 was identical to the analysis used for Experiment 1, except that 16 mean ratings of less than 1 on the 100-point scale were adjusted to 1 before the logit transformation was performed. Logit values of small probabilities are relatively large negative numbers, and omitting this step would have skewed the distribution of ratings further. Raw and logit-transformed values of the accentedness ratings from Experiment 2 are shown in Figure 4.5.

Property	Coefficient	t value	Significance
VOT	0.0022	1.257	$p < 0.05$
vowel duration	0.0059	1.843	$p < 0.05$
F1 frequency	0.0037	4.804	$p < 0.001$
F2 frequency	0.0008	2.656	$p < 0.01$
f0:sex	0.0106	1.673	$p < 0.05$
F2 curvature:sex	0.0074	2.522	$p < 0.05$

Table 4.4: Significant fixed effects for Experiment 2

4.2.2 Results

As in Experiment 1, the dependent variable in all models for Experiment 2 was accentedness rating, quantified as the logit of the mean rating across listeners for each item. The grand model was the same as the grand model for Experiment 1. Log-likelihood ratio testing (α = 0.05) revealed no significant reduction of model fit when the random intercept for word, the interactions of f0 and H1-H2 with target voicing category, and the interaction of VOT with target voicing category were eliminated, in three separate steps, from the model. Thus, the final model included only the 13 acoustic variables and the interaction of each acoustic variable with talker sex, as well as a random intercept for each talker. As above, talker sex was coded as a sum contrast with female as -1 and male as 1. Significant fixed effects for Experiment 2 are shown in Table 4.4, and their relationships to ratings are plotted in Figure 4.6.

VOT, vowel duration, F1 frequency, and F2 frequency were revealed as significant simple fixed effects. Again, greater deviation from native talker norms led to higher ratings of accentedness. Interactions of f0 and F2 curvature with talker sex were also found to be significant, and are shown in Figure 4.6, with lines ranging from the minimum to the

maximum difference values for female and male talkers separately. As f0 and F2 curvature differed more from native talker norms, accentedness ratings increased for stimuli produced by male talkers, but decreased for stimuli produced by female talkers. Again, these interactions seemed to result from the disproportionate influence of a small number of female talker stimuli with particularly high difference values, without which the interactions may not have reached significance.

The random intercepts for each talker are provided in Table B.1 in Appendix B. These random intercepts showed the same general patterns as the random intercepts from Experiment 1, although the values tended to be more extreme.

Experiment 2's findings again supported those of Major (1987), McCullough (2013), and Riney and Takagi (1999) for VOT, Munro (1993) for F1, and Wayland (1997) for F2. Additionally, Munro (1993) found that vowel duration was one of several acoustic properties that correlated with accentedness ratings on /eɪ/, and Wayland (1997) reported the same for /k^haːu/ with mid tone. The results of Experiment 2 confirmed that correlates of foreign accentedness occur in multiple parts of the acoustic signal, and showed that temporal correlates of accentedness ratings may characterize vowels as well as consonants.

4.3 Discussion

The results of Experiments 1 and 2 were not surprising, given previous reports in the literature, but they were also not the same. In the following section, the findings from these experiments are compared to one another, beginning with the ratings themselves and proceeding to the acoustic correlates.

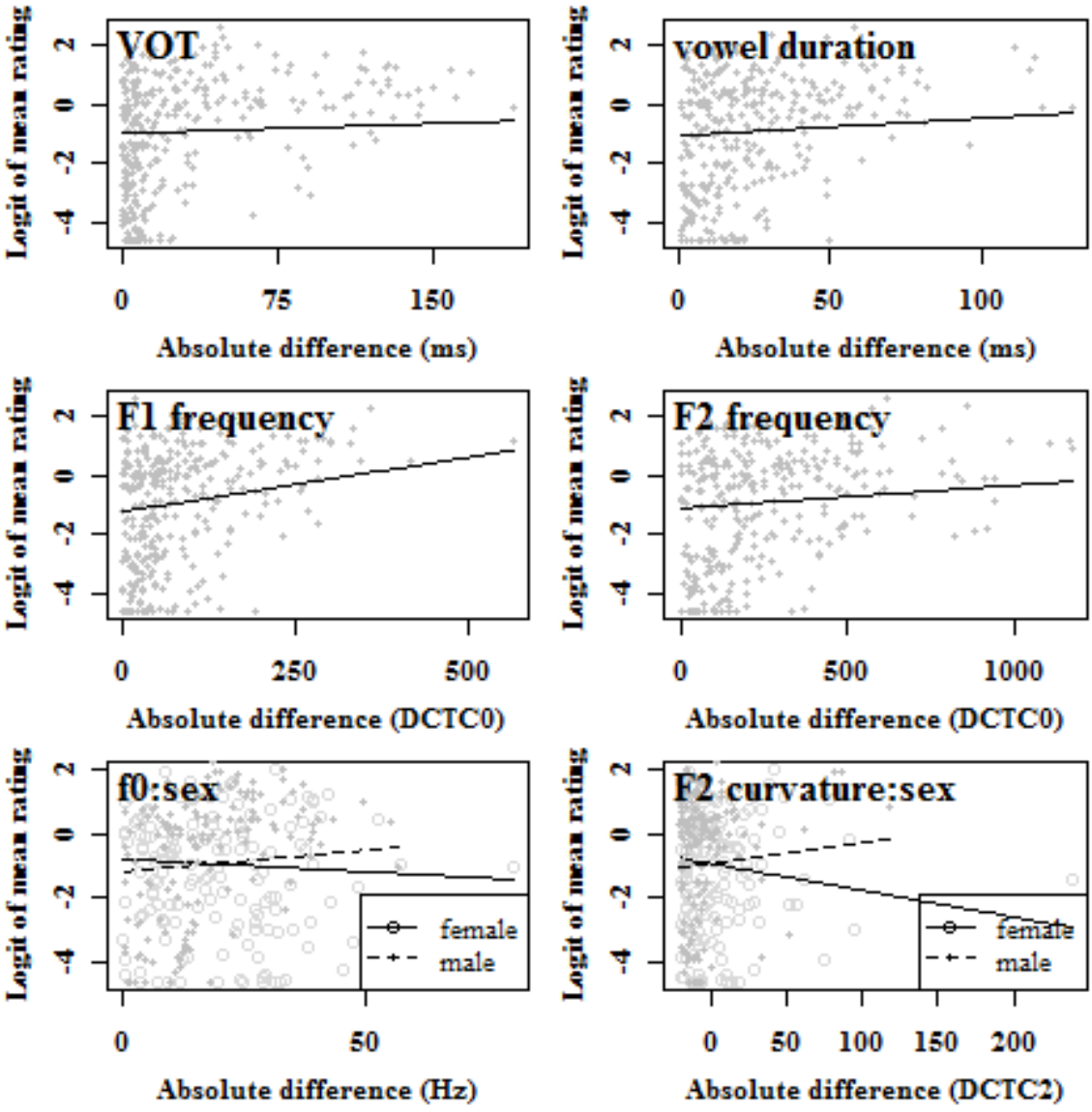

Figure 4.6: Significant fixed effects from Experiment 2

4.3.1 Effect of stimulus length (Experiments 1 and 2)

The relationship between ratings from Experiments 1 and 2 is shown in Figure 4.7, in which each point represents a single talker and is labeled with that talker's code. Each axis displays logit-transformed mean ratings over all listener responses to all productions by each talker. Linear regression showed a strong correlation between the two sets of ratings (b = 1.7833, r^2 = 0.92, $p < 0.001$). Although related, these sets of ratings were not identical, as the solid regression line is clearly distinct from the dashed y = x line. Some data points, including all those for L1 American English talkers, are below the dashed line, indicating that the talkers received lower accentedness scores on words than on CVs. If a talker was perceived as having a relatively low degree of accentedness, listeners felt more strongly or more confident about this when they heard longer stimuli. Points for many of the moderately accented talkers fall more or less on the y = x line, such that stimulus length did not seem to influence listeners' ratings for these talkers. Finally, the points representing a small number of highly accented talkers are above the dashed line, as their ratings on words were higher than on CVs. If a talker was perceived as having a very high degree of accentedness, listeners felt more strongly or more confident about this when they heard longer stimuli. Overall, to the extent that hearing word- rather than CV-length stimuli impacted accentedness ratings, it pushed them closer to the ends of the scale.

VOT, F1 frequency, and F2 frequency were correlated with accentedness ratings for both CV- and word-length stimuli. Additional correlates of ratings for CV-length stimuli included F3 frequency and F2 tilt, and an interaction between VOT and target voicing that reflected steeper changes in accentedness ratings for stimuli with voiceless as compared to voiced stops. Interactions between various spectral properties and talker sex seemed to result from the disproportionate influence of small numbers of stimuli with particularly

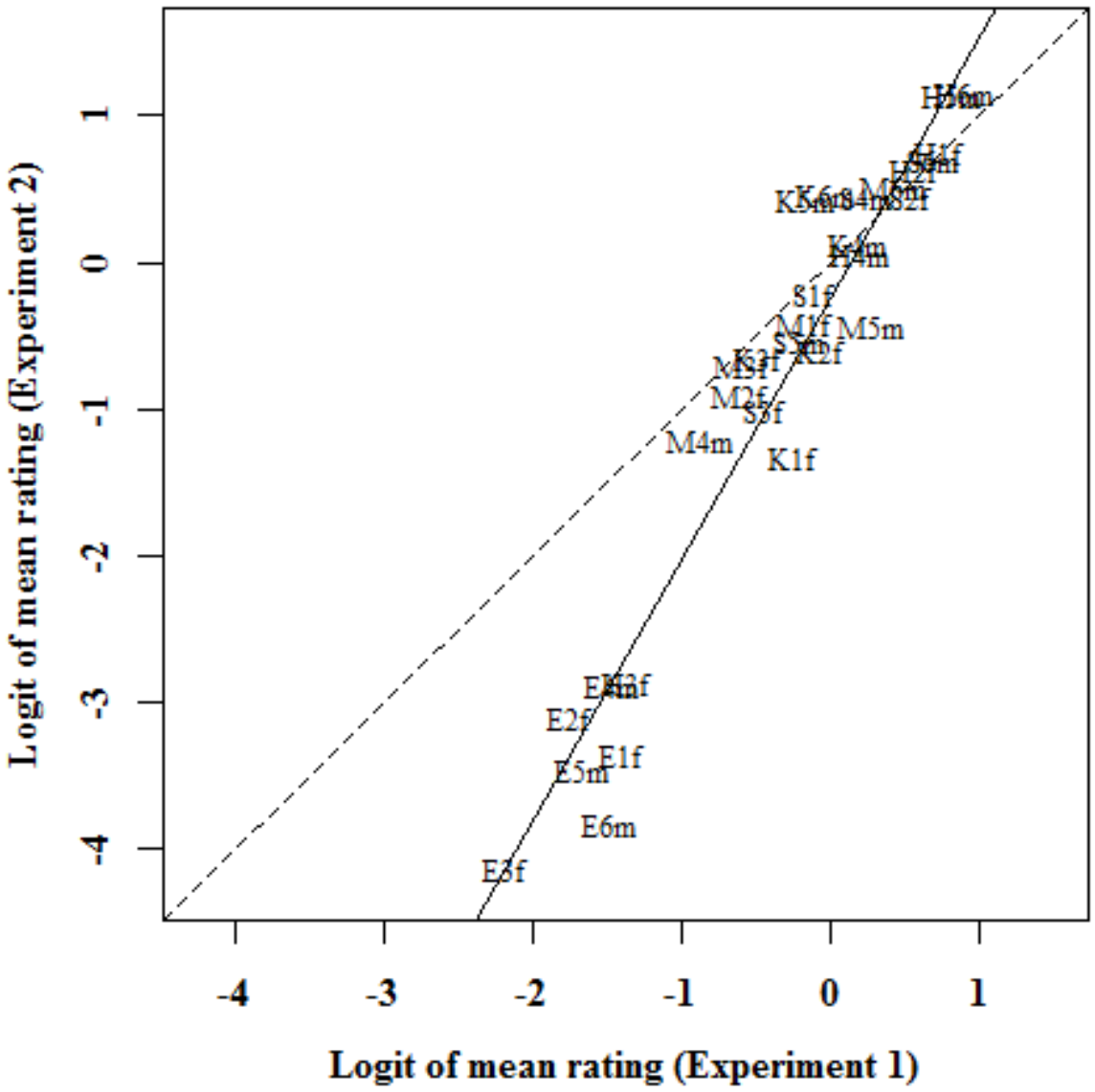

Figure 4.7: Accentedness ratings by talker on CVs (Experiment 1) and words (Experiment 2)

high difference values rather than from true differences between groups. Finally, vowel duration correlated with ratings for Experiment 2, but not for Experiment 1.

This difference in the relationships between accentedness ratings and vowel duration may have been an artifact of the method of stimulus preparation, as each CV-length stimulus was excised from a disyllabic word production and the intensity of the last 25ms was gradually reduced, thus potentially altering the perceived vowel duration from its natural state. Formant measures were made during the middle 60% of each vowel and were likely unaffected by this method of preparation. Alternatively, this result might be interpreted as the effect of a more profound difference between the two types of stimuli. Specifically, if vowel duration is interpreted as a reflection of more global temporal properties such as overall speaking rate or word-level prosodic rhythm, this result suggests that listeners can perceive such global temporal properties even in utterances as short as two syllables. In turn, this possibility highlights the importance of stimulus design in identifying components of the percept of foreign accent that are related to phonetic differences between the L2 and the L1, as opposed to components that are related to fluency in the L2.

Many studies of foreign accent perception use words (Major, 1987; Shah, 2002; Wayland, 1997) or even sentences (Munro and Derwing, 2001; Riney and Flege, 1998) as stimuli, but Experiment 1 shows, as did Munro (1993), that ratings on the basis of shorter units are possible. In addition, in the present work, ratings on CV-length stimuli were highly consistent with ratings on word-length stimuli. However, the acoustic correlates of the two sets of accentedness ratings differed slightly, suggesting that while foreign accent perception for single syllables seems to be based largely on phonetic characteristics carried over from a non-native talker's L1, for words—still relatively short samples of speech—both phonetic characteristics of L1 transfer and fluency in the L2 may play a role.

CHAPTER 5: EXPERIMENTS 3 AND 4: RATING OF CERTAINTY THAT TALKER IS NATIVE

As defined in Chapter 1, non-nativeness differs from accentedness in that it describes a perceived property of a talker, rather than a perceived property of a talker's speech. Many studies (Alba-Salas, 2004; Bond et al., 2008; Tsukada, 1998) have collected binary responses about native speaker status, perhaps because listeners may view clearly "native" and clearly "non-native" as the only possible responses. However, some investigations of nativeness have successfully used rating scales with more than two options (Baker et al., 2011; Bond et al., 2008). In Experiments 3 and 4, listeners heard samples of English produced by native talkers of American English, Hindi, Korean, Mandarin, and Spanish, and rated the degree of certainty that each stimulus was produced by a native English speaker. The idea of certainty was invoked to encourage listeners to use as much of the rating line as possible.

5.1 Experiment 3: CVs

5.1.1 Methods

The procedure and materials for Experiment 3 were identical to those described for Experiment 1 in Section 4.1.1, except that the rating scale ranged from "yes, definitely" on the left to "no, definitely not" on the right, in response to the question "Was that talker a native speaker of English?" Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual

American English-speaking listeners (14 females) were considered here. Data from 11 additional listeners were excluded for the following reasons: non-native English speaker (4), native bilingual (2), early exposure to another language (1), caretaker(s) grew up abroad (1), failure to follow instructions (2), and technical problems with the experiment presentation software (1). Data from a 12^{th} additional listener were excluded because she was one of the L1 American English female talkers. Although she did not appear to recognize her own voice, her responses were excluded as a precaution, as she clearly would have known her own native speaker status.

The impact of possible retroflexion in L1 Hindi productions on listeners' ratings was evaluated for Experiment 3 as it was for Experiment 1. A paired t-test again showed that the mean of the corrected ratings over each L1 Hindi talker's productions of stimuli with alveolar stop targets was not different from the mean of the corrected ratings over his or her productions of stimuli with labial and velar stop targets ($t(5) = 1.2536$, $p > 0.05$), suggesting that any future inclusion of an acoustic measure of retroflexion may not substantially improve the analysis.

The ratings considered in the analysis below were the means over all listeners for each item, for a total of 300 mean ratings. Raw and logit-transformed values of the nonnativeness ratings from Experiment 3 are shown in Figure 5.1. For these data, no values were extreme enough to require adjustment.

5.1.2 Results

As in Chapter 4, linear mixed effects regression models were used to explore the relationships of interest. For Experiment 3, the dependent variable in all models was nonnativeness rating, quantified as the logit of the mean rating across listeners for each item.

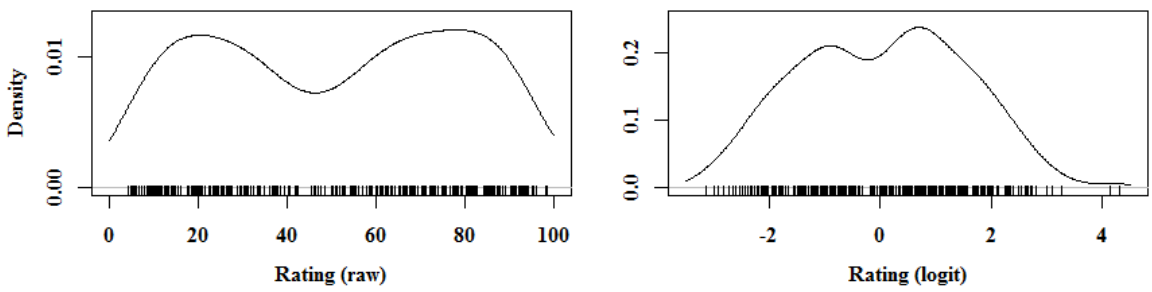

Figure 5.1: Raw and logit-transformed ratings from Experiment 3

The grand model, as before, included random intercepts for talker and word, and fixed effects of the 13 acoustic variables, the interaction of each of these acoustic variables with talker sex (female or male), and the interaction of VOT, f0, and H1-H2 with target voicing category (voiced or voiceless). All acoustic variables were centered. The simple fixed effects of talker sex and target voicing category were not considered. Log-likelihood ratio testing ($\alpha = 0.05$) revealed no significant reduction of model fit when the random intercept for word, the interactions of f0 and H1-H2 with target voicing category, and the interaction of VOT with target voicing category were eliminated, in three separate steps, from the model. The final model therefore included only the 13 acoustic variables and the interaction of each acoustic variable with talker sex, as well as a random intercept for each talker. Talker sex was coded as a sum contrast with female as -1 and male as 1. Significant fixed effects for Experiment 3 are shown in Table 5.1, and their relationships to ratings are plotted in Figure 5.2.

VOT, F1 frequency, F2 frequency, F3 frequency, and F2 tilt were identified as significant simple fixed effects. Greater deviation from native talker norms led to higher ratings of nonnativeness. The interaction between H1-H2 and talker sex was also found to be significant, although Figure 5.2 suggests that, as in Experiment 1, this interaction resulted from the

Property	Coefficient	t value	Significance
VOT	0.0087	4.902	$p < 0.001$
F1 frequency	0.0032	4.088	$p < 0.001$
F2 frequency	0.0010	3.442	$p < 0.001$
F3 frequency	0.0007	2.872	$p < 0.01$
F2 tilt	0.0045	3.183	$p < 0.01$
H1-H2:sex	-0.0462	-1.845	$p < 0.05$

Table 5.1: Significant fixed effects for Experiment 3

influence of several female talker stimuli with especially high H1-H2 difference values. Indeed, in general, the acoustic correlates for Experiment 3 were strikingly similar to those for Experiment 1.

The random intercepts for each talker are shown in Table B.1 in Appendix B. Again, these random intercepts were similar to the random intercepts from Experiment 1, with negative values for L1 American English talkers, positive values for all L1 Hindi talkers except H3f, and variability within each of the other language backgrounds.

Unfortunately, there are relatively few studies of the perception of non-nativeness by which to evaluate the results of Experiment 3. Baker et al. (2011) and Bond et al. (2008) measured acoustic properties of relatively long stimuli which are not directly comparable to the acoustic properties included in the present investigation. However, VOT was correlated with perceptual responses about non-nativeness from one group of listeners described by Alba-Salas (2004). In general, the results of Experiment 3 are reminiscent of the acoustic correlates of foreign accentedness previously reported in the literature, including VOT (Major, 1987; McCullough, 2013; Riney and Takagi, 1999), F1 (Munro, 1993), and static (Wayland, 1997) and dynamic (Munro, 1993) measures of F2. Explicit comparison of the

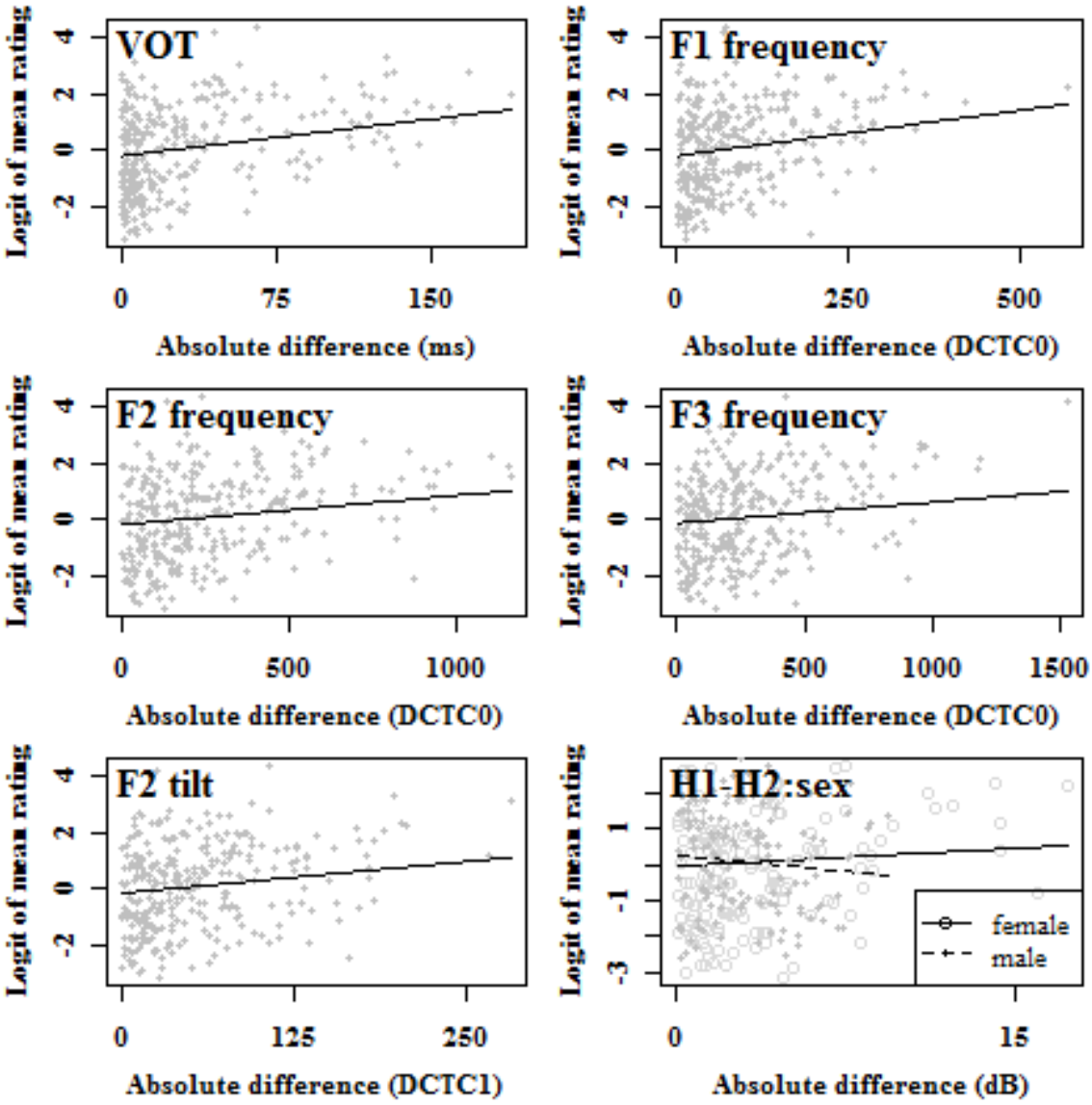

Figure 5.2: Significant fixed effects from Experiment 3

findings from Experiments 1 and 3 is reserved for the discussion section below, following the description of Experiment 4.

5.2 Experiment 4: Words

5.2.1 Methods

The procedure and materials for Experiment 4 were identical to those used in Experiment 3, except that the full word was played rather than the initial CV. Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual American English-speaking listeners (14 females) were considered here. Data from 6 additional listeners were excluded for the following reasons: non-native English speaker (1), early exposure to another language (2), lived abroad (1), self-reported hearing loss (1), and failure to follow instructions (1).

The importance of possible retroflexion in L1 Hindi productions was evaluated for Experiment 4 as it was for Experiment 3. Again, the mean of the corrected ratings over each L1 Hindi talker's productions of stimuli with alveolar stop targets was not different from the mean of the corrected ratings over his or her productions of stimuli with labial and velar stop targets ($t(5)$ = -1.4239, $p > 0.05$), as revealed by a paired t-test. The rating patterns in Experiment 4 provided no justification for adding an acoustic measure of retroflexion to future investigations.

The analysis for Experiment 4 was identical to the analysis used for Experiment 3. Raw and logit-transformed values of the non-nativeness ratings from Experiment 4 are shown in Figure 5.3. For these data, no values were extreme enough to require adjustment.

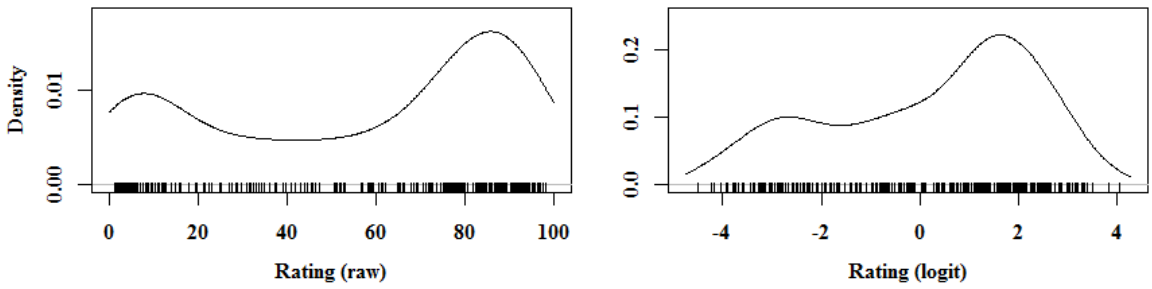

Figure 5.3: Raw and logit-transformed ratings from Experiment 4

5.2.2 Results

As in Experiment 3, the dependent variable in all models for Experiment 4 was non-nativeness rating, quantified as the logit of the mean rating across listeners for each item. The grand model was the same as the grand model for Experiment 3. Log-likelihood ratio testing (α = 0.05) revealed no significant reduction of model fit when the interactions of f0 and H1-H2 with target voicing category and the interaction of VOT with target voicing category were eliminated, in two separate steps, from the model. Thus, the final model included only the 13 acoustic variables and the interaction of each acoustic variable with talker sex, as well as random intercepts for talker and word. As above, talker sex was coded as a sum contrast with female as -1 and male as 1. Significant fixed effects for Experiment 4 are shown in Table 5.2, and their relationships to ratings are plotted in Figure 5.4.

In Experiment 4, the simple fixed effects of VOT, vowel duration, F1 frequency, and F2 frequency were found to be significant. Again, higher non-nativeness ratings were associated with stimuli characterized by acoustic measurements farther from native talker norms. No interactions were significant.

Property	Coefficient	t value	Significance
VOT	0.0027	1.385	$p < 0.01$
vowel duration	0.0089	2.619	$p < 0.01$
F1 frequency	0.0032	3.808	$p < 0.01$
F2 frequency	0.0006	1.938	$p < 0.05$

Table 5.2: Significant fixed effects for Experiment 4

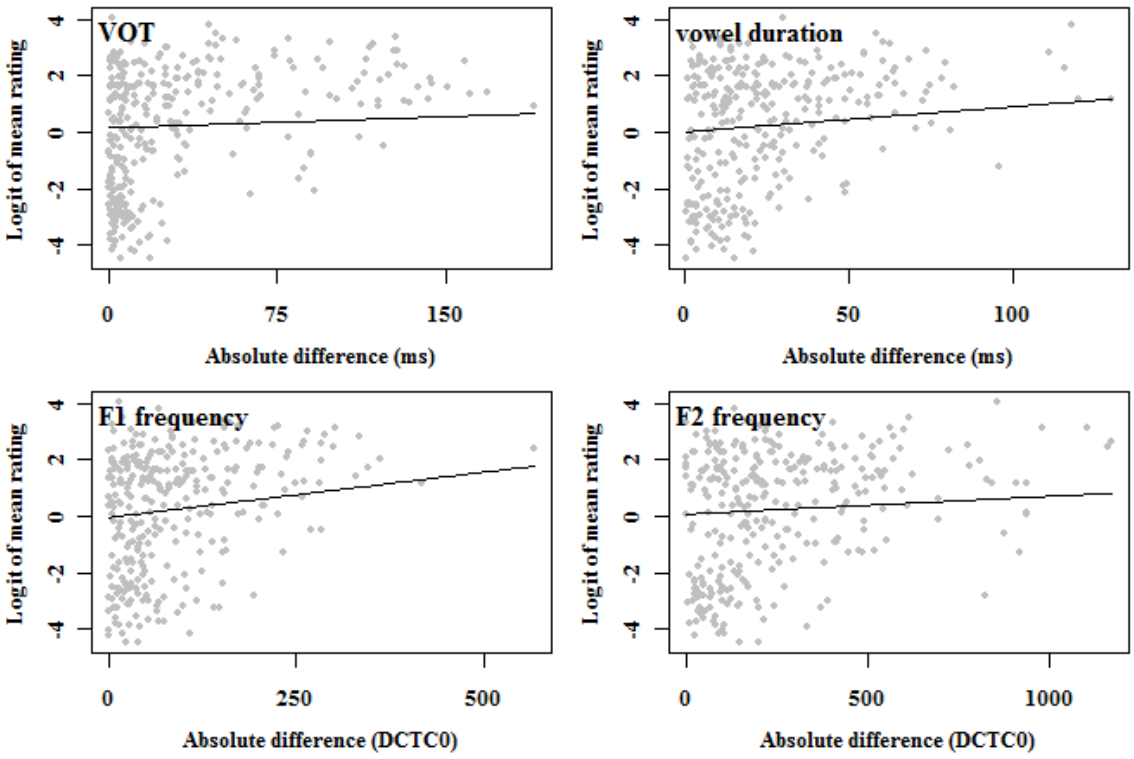

Figure 5.4: Significant fixed effects from Experiment 4

The random intercepts for each talker are shown in Table B.1 in Appendix B. Their values were similar to those of the random intercepts from previous experiments, with particularly extreme negative values for L1 American English talkers and talker H3f, and particularly extreme positive values for the other L1 Hindi talkers. The random intercepts for words are not presented here, but were generally smaller in magnitude than the random intercepts for talkers, ranging from -0.4656 for *gable* to 0.4363 for *dipper*.

Again, as was noted for Experiment 3, it is difficult to directly compare the findings from Experiment 4 to findings from earlier studies of non-nativeness perception, although they do confirm Alba-Salas's (2004) significant result for VOT. However, all the significant correlates of non-nativeness in Experiment 4 have been identified in previous studies as correlates of foreign accentedness. The discussion section below includes a comparison of the present investigation's results regarding accentedness with its results regarding non-nativeness.

5.3 Discussion

The four experiments described thus far were designed to differ as minimally as possible from one another in order to facilitate comparison across separate sets of responses. With the findings from Experiments 3 and 4, two interesting types of comparisons can be made. As with Experiments 1 and 2, the findings from Experiments 3 and 4 may be compared to each other to evaluate the role of stimulus length on the perception of non-nativeness. Additionally, the perception of accentedness explored in Experiments 1 and 2 may be compared to the perception of non-nativeness explored in Experiments 3 and 4 to see whether listeners treated these two scales differently. Each of these comparisons is

addressed in turn below, again beginning with the ratings themselves and continuing on to the acoustic correlates.

5.3.1 Effect of stimulus length (Experiments 3 and 4)

The relationship between non-nativeness ratings from Experiments 3 and 4 is shown in Figure 5.3.1, in which each point represents a single talker and is labeled with that talker's code. Each axis displays logit-transformed mean ratings over all listener responses to all productions by each talker. Linear regression revealed a strong correlation between the two sets of ratings ($b = 1.5884$, $r^2 = 0.91$, $p < 0.001$). Again, although they were related to one another, these two sets of ratings were not identical, as shown by the difference between the solid regression line and the dashed y = x line. The points representing L1 American English talkers, as well as the L1 Hindi talker thought to be targeting American English (H3f), appear below the y = x line, reflecting greater perceived nativeness for words than for CVs. In contrast, nearly all the points representing non-native talkers appear above the y = x line, indicating higher non-nativeness ratings for words than for CVs. Listeners who heard word- rather than CV-length stimuli were more certain of their judgments, as reflected by ratings closer to the ends of the scale.

VOT, F1 frequency, and F2 frequency correlated with non-nativeness ratings for both CV- and word-length stimuli. F3 frequency and F2 tilt also correlated with ratings for CVs. The interaction between H1-H2 and talker sex was significant for Experiment 3, but did not seem to reflect true differences between the groups. Finally, vowel duration correlated with ratings for word-length stimuli, but not with those for CV-length stimuli. Again, while this may have been a consequence of the method of CV-length stimulus preparation, it may

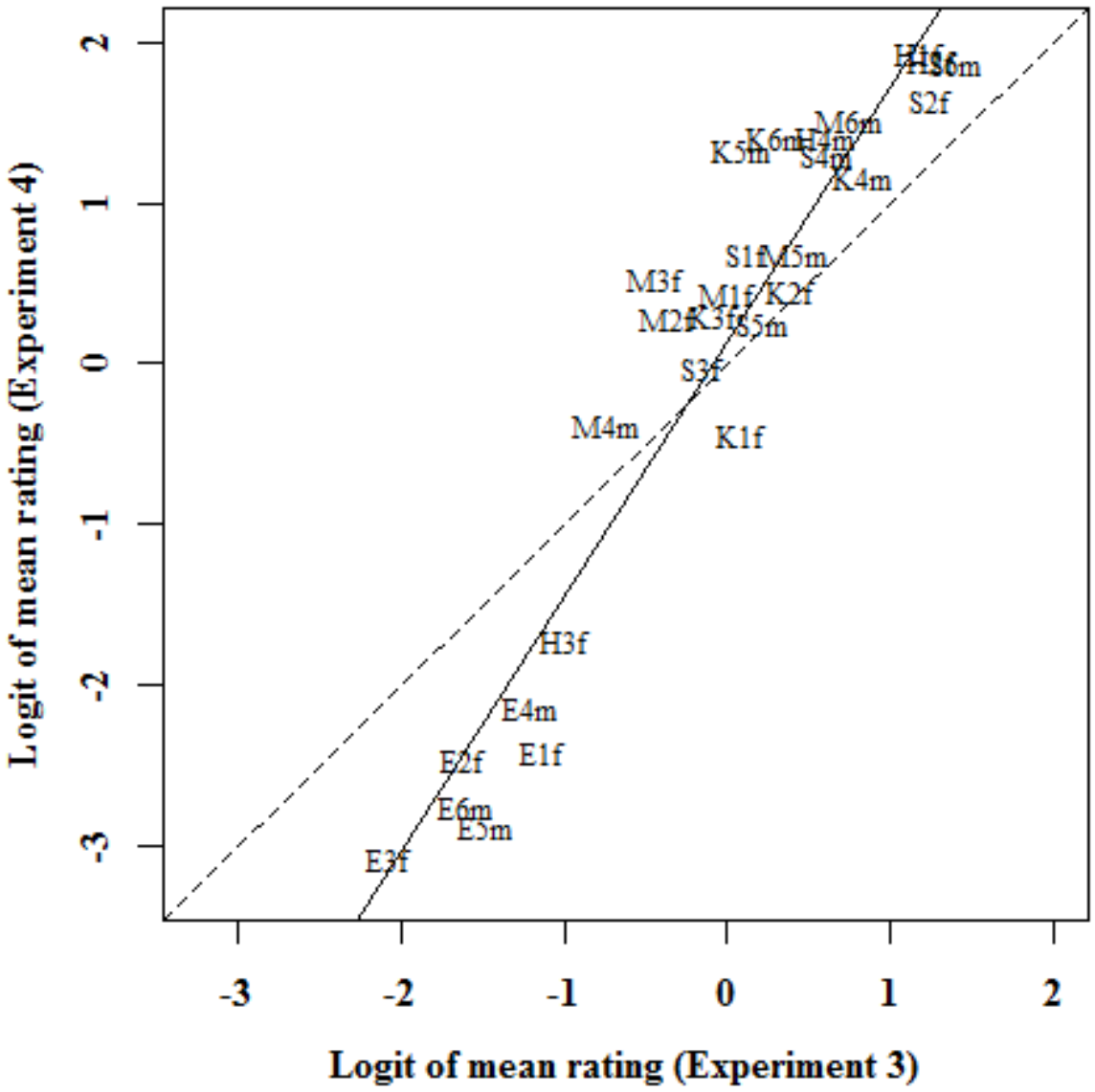

Figure 5.5: Non-nativeness ratings by talker on CVs (Experiment 3) and words (Experiment 4)

also suggest that listeners' attention turns to more global properties of speech with longer stimuli.

In general, ratings of non-nativeness were correlated with a variety of acoustic properties from multiple portions of the speech signal. While there was little basis for comparison to prior studies of non-nativeness, the acoustic correlates of non-nativeness identified in Experiments 3 and 4 resembled those from prior studies of foreign accentedness. In light of this apparent similarity, the present results for accentedness and non-nativeness are compared directly in the following section.

5.3.2 Accentedness versus non-nativeness (Experiments 1/2 and 3/4)

The relationships between ratings from Experiments 1 and 3 and between ratings from Experiments 2 and 4 are shown in Figure 5.6, in which each point represents a single talker and is labeled with that talker's code. Each axis displays logit-transformed mean ratings over all listener responses to all productions by each talker. Linear regression confirmed that the two types of ratings were highly correlated for both CVs (b = 1.1797, r^2 = 0.98, $p < 0.001$) and words (b = 1.0615, r^2 = 0.99, $p < 0.001$). Again, it is clear from the figure that these sets of ratings, while nearly perfectly correlated, were not identical. For CVs, the two ratings were very similar for stimuli from L1 American English talkers, but diverged progressively as accentedness and non-nativeness ratings increased, with non-nativeness ratings exceeding ratings of accentedness. That is, the certainty of the talker attribute increased more quickly than the degree of the speech attribute. For words, higher degrees of accentedness also corresponded to higher ratings of non-nativeness, although this was true for all talkers rather than for only the non-native ones. Indeed, the slopes of the solid

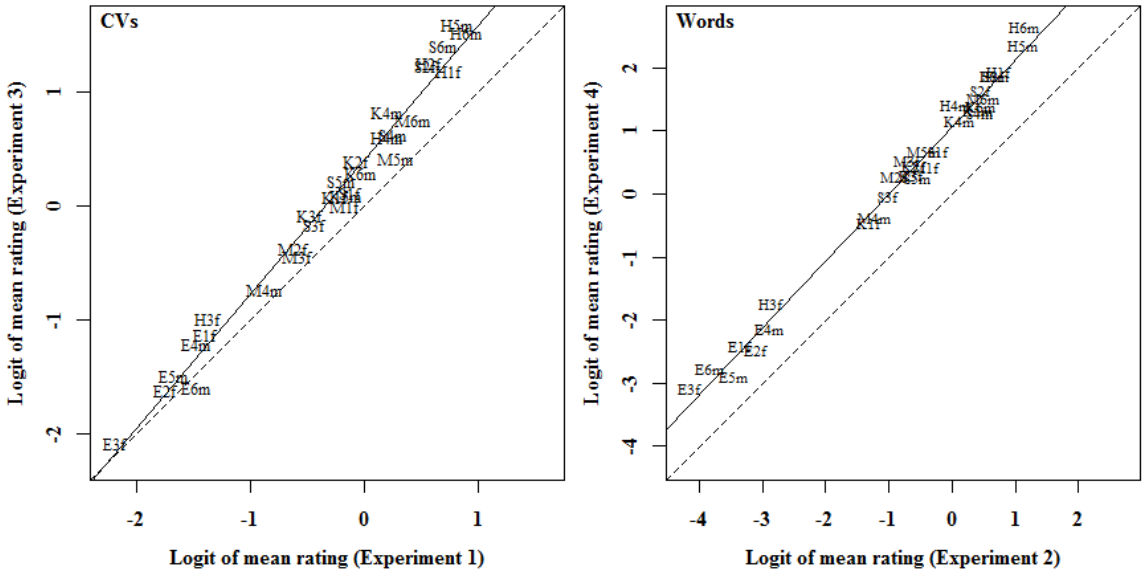

Figure 5.6: Ratings by talker for accentedness (Experiments 1/2) and non-nativeness (Experiments 3/4)

regression line and the dashed y = x line are nearly parallel, indicating little to no interaction of this effect with talker background. This pattern seemed to result from differing distributions of ratings. The "no foreign accent" peak for mean ratings in Experiment 2 (Figure 4.5) was around 5 on the 100-point scale (logit value: -2.94), while the "native" peak in Experiment 4 (Figure 4.5) was around 10 (logit value: -2.20). Further investigation of the responses revealed that like the mean ratings, ratings from individual listeners tended to be closer to the right end of the scale in Experiment 4 than in Experiment 2. Even if most ratings for a particular native talker were near the left end of the scale, several ratings of 100 would increase the mean rating more than several ratings of 75. Thus, small numbers of very high ratings increased the mean non-nativeness ratings for L1 American English talkers in Experiment 4.

With the exception of a small number of fixed interactions, most of which appeared spurious, the acoustic correlates of accentedness and non-nativeness ratings were identical for both CV-length stimuli (Experiments 1 and 3) and word-length stimuli (Experiments 2 and 4). Also, the difference between correlates of accentedness ratings for CV-length and word-length stimuli (Experiments 1 and 2) was mirrored in the difference between correlates of non-nativeness ratings for the two stimulus lengths (Experiments 3 and 4). For these reasons, and because ratings of accentedness and non-nativeness were nearly perfectly predictable from one another, the interpretation of non-nativeness ratings as reflecting accentedness, as is widespread in the literature, is perhaps not problematic. Listeners did appear to use the scale somewhat differently when rating non-nativeness rather than accentedness, reminiscent of Cheong's (2007) finding that nativeness ratings were more extreme than accentedness ratings. Nonetheless, the patterns of relationships to acoustic cues suggested that listeners equate the speech property of "accentedness" with the talker property of "non-nativeness."

CHAPTER 6: EXPERIMENTS 5 AND 6: RATING OF CERTAINTY THAT STIMULUS IS ENGLISH

Previous work has shown that listeners can rate the similarity of acoustic stimuli to a particular reference language (Bond and Stockmal, 2002; Bradlow et al., 2010) or to multiple reference languages simultaneously (Flege and Munro, 1994). In Experiments 5 and 6, listeners heard samples of English, Hindi, Korean, Mandarin, and Spanish produced by native talkers of each language, and rated the degree of certainty that each stimulus was extracted from a recording of English. As was done for the non-nativeness rating task in Experiments 3 and 4, the idea of certainty was invoked to encourage listeners to use as much of the rating line as possible.

6.1 Experiment 5: CVs

6.1.1 Methods

Procedure

The procedure was similar to the procedure described for Experiment 1 in Section 4.1.1, except that the rating scale ranged from "yes, definitely" on the left to "no, definitely not" on the right, in response to the question "Was that sample taken from a recording of someone speaking English?" It was not specified whether the question referred to someone speaking English natively.[5] Because some of the stimuli were in languages other than English, as described below, no target word was displayed on the computer screen.

[5]The sole participant who asked for clarification on this point was instructed that the language, rather than the talker's background, was at issue.

Stimuli

The 300 auditory stimuli in the rating task included CVs extracted from the beginning of each target word shown in Tables 3.1 through 3.5, with 60 productions from each language. Each talker was represented only by words in his or her native language; thus, unlike Experiments 1 through 4, no non-native English productions were included. The English stimuli were identical to those produced by the L1 American English talkers in Experiment 1. For languages with fewer than 60 unique target words, the stimuli included multiple examples of some sequences produced by different talkers. Except for Hindi, stimuli were selected such that the 10 CVs produced by each talker included 1-2 instances of each consonant and each vowel. In English, Mandarin, and Spanish, which have two-way stop contrasts at each place of articulation, the 10 CVs produced by each talker included equal numbers of consonants from each target voicing category. Korean has a three-way stop contrast at each place of articulation, which cannot be evenly represented by 10 stimuli. The tenth stimulus contained a tense stop for 2 talkers, a lax stop for 2 talkers, and an aspirated stop for 2 talkers. In Hindi, because the list of target words involved 11 stops (targets with initial /p^h/ were omitted due to frequent pronunciation as [f] (Sandahl, 2000)) and 10 vowels, it was not possible for the 10 CVs produced by each talker to contain all target sounds. Hindi stimuli were selected such that the 10 CVs produced by each talker included 1-2 instances of most consonants and most vowels, with 1-2 consonants and 0-3 vowels per talker not represented.[6] The 20 practice stimuli included 20 CVs extracted from the beginning of additional target words recorded as practice words by the talkers, with 4 stimuli from talkers of each of the 5 L1 backgrounds. These CVs contained a variety of consonants and

[6]In theory, the 10 CVs for each talker might have included all 10 vowels. However, lexical gaps and the simultaneous attempt to vary the initial consonants made this impossible in practice.

vowels. Within each L1 background, the 4 words were produced by 2 female and 2 male talkers; 1 randomly chosen female and 1 randomly chosen male talker were omitted to limit the amount of time required to complete the practice session. The final 25ms of each CV gradually decreased in intensity to reduce the audibility of any coarticulatory cues with the following consonant.

The auditory stimuli in the sentence completion task were identical to those used in Experiments 1 through 4.

Participants

Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual American-English speaking participants (14 females) were considered here. Data from 9 additional participants were excluded for the following reasons: native bilingual (1), early exposure to another language (6), and failure to follow instructions (2).

Analysis

As in earlier experiments, the ratings considered were the means over all listeners for each item, for a total of 300 mean ratings. Raw and logit-transformed values of the accentedness ratings from Experiment 5 are shown in Figure 6.1.

Figure 6.1 shows that the ratings from this experiment differed substantially from the others: the raw rating values were unimodally rather than bimodally distributed. The labels for this scale were clear opposites, as in Experiments 3 and 4, and the unimodality of the ratings suggested that listeners may have been generally unsure of their responses. Moreover, further consideration of the listeners' approach to this task revealed problems with initial assumptions about how to model their performance. Although the listeners

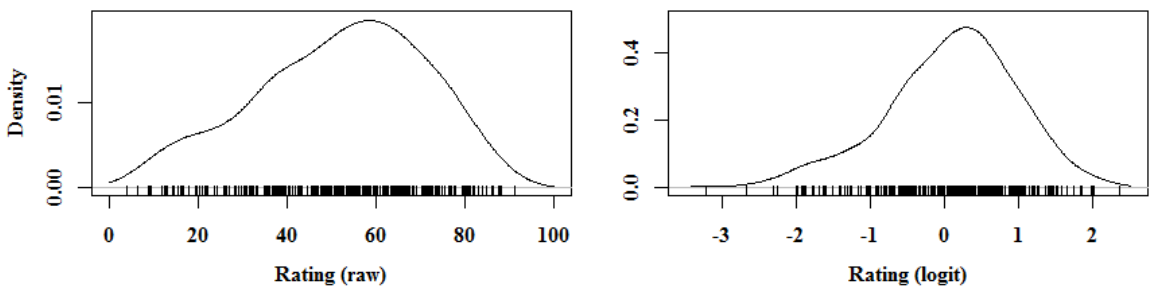

Figure 6.1: Raw and logit-transformed ratings from Experiment 5

did not speak most of the languages represented in the stimuli, the sequences were quite constrained, consisting only of stops and vowels. It is not unlikely that listeners might have assimilated these sounds to those of their own language, especially given that the rating task specifically invoked English. Additionally, many of the CVs produced by L1 American English talkers may have themselves mapped to words in English (e.g., the initial syllables of *baby*, *dopey*, *keeper*), and CVs produced by other talkers may have, as well, depending on how the sounds were assimilated. Recognition of a stimulus as a lexical item in English could have important consequences for a rating of whether it was "taken from a recording of someone speaking English." In previous rating experiments with stimuli in multiple languages, this problem was avoided because listeners understood either all the stimuli (English and Spanish *taco* in Flege and Munro's (1994) experiments) or none of them (multi-second excerpts from only unfamiliar languages in studies by Bond and Stockmal (2002) and Bradlow et al. (2010)).

If listeners assimilated all the stimuli to English sound categories, then the acoustic measures in the analysis should be quantified as differences from their "expectations," as in Experiments 1 through 4. However, such differences are impossible to calculate because

the categories to which the Hindi, Korean, Mandarin, and Spanish sounds were assimilated are unknown. Additionally, if lexical recognition influenced listeners' responses, then acoustic properties may not predict ratings of non-Englishness at all. This latter issue is explored below for Experiment 6, where the collection of additional perceptual data allowed for further analysis.

6.2 Experiment 6: Words

6.2.1 Methods

The procedure and materials were identical to those used in Experiment 5, except that the full word was played rather than the initial CV. Listeners were recruited from the linguistics department subject pool and received partial course credit for their participation. Data from 20 monolingual American English-speaking listeners (13 females) were considered here. Data from 12 additional listeners were excluded for the following reasons: non-native speaker (1), native bilingual (1), early exposure to another language (2), lived abroad (2), caretaker(s) grew up abroad (2), and failure to follow instructions (4). Data from a 13th additional listener were excluded because one week prior to testing she had participated in the additional experiment described below, which featured identical acoustic stimuli.

As suggested above, an obvious complication with this experiment was that some sequences clearly mapped to known words of English, especially the actual English words produced by native talkers of English. The word lists for other languages also included some words which had cognates in English, often as a result of lexical borrowing. Participants were reminded that languages borrow words from one another and encouraged to

consider whether each production was "taken from a recording of someone speaking English" rather than only whether they recognized the word. Nonetheless, lexical cues might be expected to have played a role in listeners' ratings. To quantify the effect of lexical cues, an additional experiment was run to explore how readily each production was interpreted as an English word. In this experiment, listeners heard the same stimuli as in the rating portion of Experiment 6, but were simply asked to indicate whether they recognized each production as a word of English by typing "y" (yes) or "n" (no). If "yes," they then typed the word into a text box on the screen.[7] Data from 10 listeners, again recruited from the linguistics department subject pool and compensated with partial course credit, were included. Four of these listeners had participated in another of the experiments discussed in this work, but none in Experiment 6. Data from 10 listeners were excluded for the following reasons: non-native speaker (2), native bilingual (1), early exposure to another language (3), lived abroad (2), caretaker grew up abroad (1), self-reported hearing loss (1). Data from an 11th additional listener were excluded because he reported that he recognized one of the L1 American English talkers, a graduate student who had administered another linguistics experiment. The rate of "yes" responses from this experiment is the independent variable in the analysis for Experiment 6. As recognition rates are bounded at 0 and 1, these rates, like the ratings from Experiments 1 through 6, were logit-transformed.

Again, the ratings considered were the means over all listeners for each item, for a total of 300 mean ratings. 20 mean ratings of less than 1 on the 100-point scale were adjusted to 1 before the logit transformation was performed. Raw and logit-transformed values of the accentedness ratings from Experiment 6 are shown in Figure 6.2.

[7]As a word was only typed if the listener recognized it as English, this task did not provide information about assimilation for all non-English stimuli.

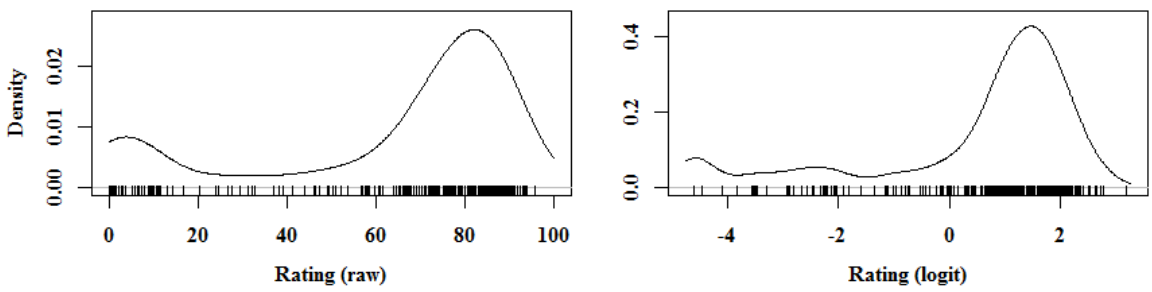

Figure 6.2: Raw and logit-transformed ratings from Experiment 6

6.2.2 Results

Linear regression revealed a significant relationship between rate of recognition as a word of English and rating of non-Englishness, with stimuli as items (b = -0.5938, r^2 = 0.76, $p < 0.001$). As the rate of recognition as a word of English increased, the rating of non-Englishness decreased–that is, the item was rated as more likely to have been taken from a recording of someone speaking English. This lexical recognition accounted for 76% of the variance in the non-Englishness ratings, suggesting that listeners were highly influenced by lexical cues in this rating task. Thus, although it turned out to be impossible to appropriately model relationships between acoustic properties and ratings of non-Englishness, due to not knowing the relevant perceptual assimilation patterns, listeners may have relied only modestly on acoustic information for these evaluations in the first place.

6.3 Discussion

While Experiments 1 through 4 involved nearly identical stimuli, differing only in whether the second syllable of each production was played, Experiments 5 and 6 included stimuli in four languages besides English. That the non-native talkers produced entirely

different sequences in Experiments 5 and 6 as opposed to Experiments 1 through 4 makes it difficult to directly compare ratings across these sets of experiments. Additionally, the inability to identify acoustic correlates of the non-Englishness ratings in Experiments 5 and 6 precludes even high-level comparison of the findings across these sets of experiments. However, as the stimuli in Experiments 5 and 6 were quite similar to one another, again differing only in whether the second syllable of each production was played, the distribution of ratings may be examined to determine whether longer stimuli again resulted in more extreme, and thus more certain, ratings from listeners.

6.3.1 Effect of stimulus length (Experiments 5 and 6)

The relationship between ratings from Experiments 5 and 6 is shown in Figure 6.3, in which each point represents a single talker and is labeled with that talker's code. Each axis displays logit-transformed mean ratings over all listener responses to all productions by each talker. As in previous cases, there was a significant relationship between the ratings in these experiments (b = 2.3877, r^2 = 0.76, $p < 0.001$), and the extremely positive slope of the regression line indicates that ratings were nearer to the ends of the scale in Experiment 6 than in Experiment 5. When they heard full words, listeners were more certain that stimuli produced by L1 American English talkers were taken from recordings of English, and more certain that stimuli produced by other talkers were not taken from recordings of English, than when they heard CVs.

Ratings of non-Englishness resembled ratings of accentedness and non-nativeness in that more extreme responses were assigned to longer stimuli. Due to complications of the experimental design, acoustic correlates of non-Englishness could not be explored from

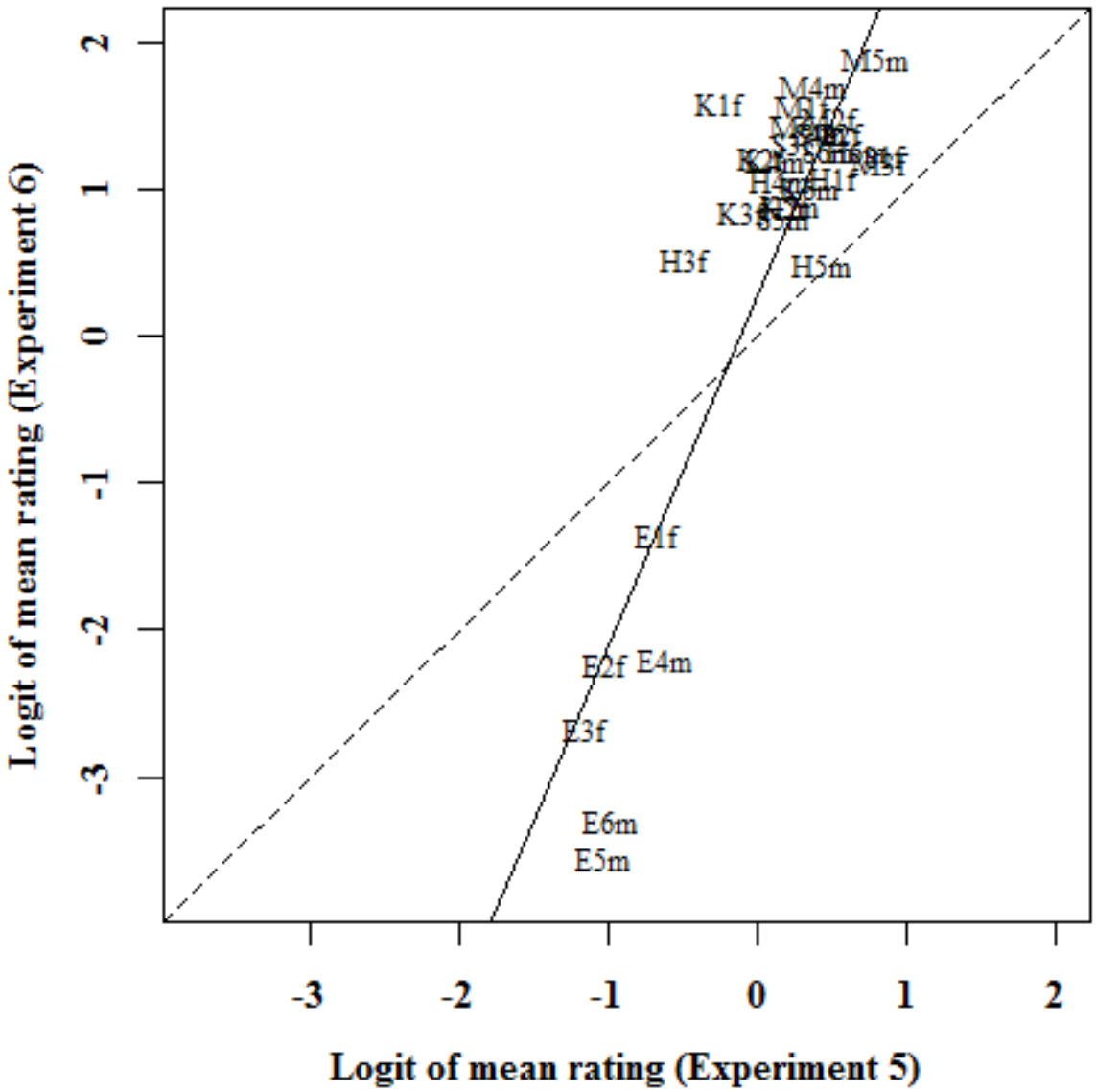

Figure 6.3: Non-Englishness ratings by talker on CVs (Experiment 5) and words (Experiment 6)

the perceptual data collected in Experiments 5 and 6. In future investigations of non-Englishness, listeners' judgments about the closest English sounds should be collected, so that the acoustic measurements for these productions may be expressed as difference values and so that the acoustic correlates of non-Englishness may be directly compared to those of accentedness and non-nativeness. As phonetic transfer from a talker's native language to his or her non-native productions is not uncommon, some similarity across these sets of acoustic correlates is expected.

For word-length stimuli, lexical knowledge related to evaluations of non-Englishness, as evidenced by the relatively strong correlation between rate of recognition as a word of English and non-Englishness rating. Thus, even in units as small as words, some important cues to perception are not necessarily acoustic.

CHAPTER 7: EXPERIMENTS 1 THROUGH 6: FREE CLASSIFICATION OF NATIVE LANGUAGE

Listeners in previous studies have been reasonably successful in classifying the native languages of non-native talkers (Derwing and Munro, 1997; Vieru et al., 2011), as well as languages themselves (Bond and Fokes, 1991; Vasilescu et al., 2005), in tasks that have offered explicit labels as response options. However, these labels may influence the way listeners complete such tasks, as they fix the number and identities of possible groups. Experiments 1 through 6 involved a free classification task in which listeners grouped speech from native talkers of American English, Hindi, Korean, Mandarin, and Spanish without the restriction of predetermined labels. Presentation of methods and results is divided into two parts: free classification of stimuli in English (Experiments 1 through 4), and free classification of stimuli in the talkers' L1s (Experiments 5 and 6).

7.1 Experiments 1 through 4: English

7.1.1 Methods

Procedure

The free classification task, run in Microsoft PowerPoint 2003, was the first task of the experiment for half the listeners and the third task for the remainder. In this task, listeners saw a screen with 30 icons arranged to the left of a 16x16 grid, as in Figure 7.1. Double-clicking on an icon played a short sample of speech from the talker whose initials appeared on the icon. Single-clicking on an icon allowed it to be dragged around the screen. Listeners

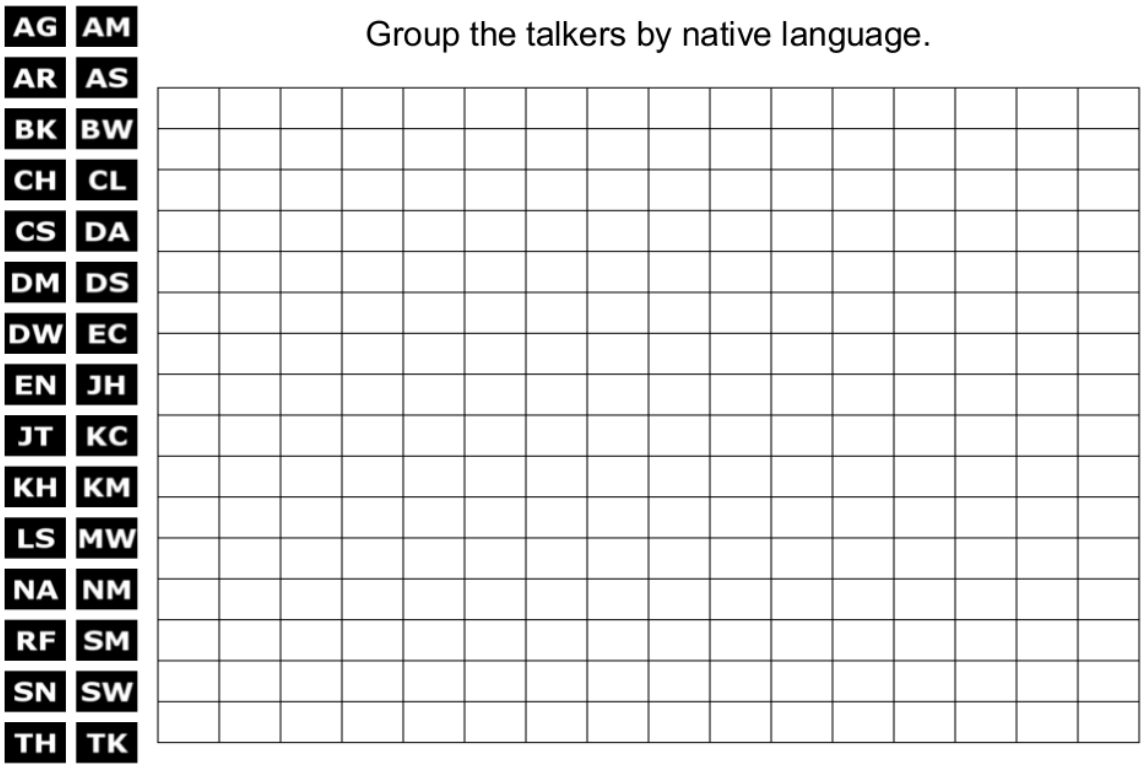

Figure 7.1: Free classification screen

were instructed to listen to the talkers and arrange them into native language groups, such that all talkers with the same native language were grouped together. Group membership was indicated by adjacency on the 16x16 grid. There were no restrictions on the number of groups or the number of talkers per group, and listeners were allowed to hear each talker as many times as desired and to rearrange the icons until they were satisfied. Afterward, listeners were asked to identify the native language of the talkers in each of the groups that they had made, and recorded their responses on paper.[8]

If a listener did not clearly place all the icons on the grid, generally because some icons were no longer visible on the screen due to unintentional scrolling, this was considered

[8]These labels are not analyzed in the present work.

"failure to follow instructions" and the listener's responses were not included in the free classification or rating analyses. In the rare event that a listener's identifications of talkers' native languages made it clear that he or she had not understood the task properly (for instance, when the labels were segments rather than languages), this was also considered "failure to follow instructions" and the listener's responses were excluded.

Transliterations of some talkers' names seemed likely to bias listeners against thinking that their initials could be those of native speakers. For instance, *X* and *Y* are unusual initials for native speakers of American English, but were not uncommon among the Mandarin and Korean speakers recorded, respectively. Thus, the initials displayed on the icons were not the talkers' actual initials, but initials taken from the larger set of native speakers of American English who were recorded for this investigation (discussed in Section 3.1.2). Assignment of initials to talkers was pseudorandom to ensure that the talkers within an L1 group exhibited a variety of first initials, last initials, and locations on the computer screen when the icons were arranged in alphabetical order.

Stimuli

The 30 auditory stimuli in the classification task for each experiment were a subset of those included in the rating task, each produced by a different talker. Each talker produced a unique sequence in English. Stimuli were selected such that the 6 talkers within an L1 background produced unique initial consonants and a variety of vowels. Experiments 2 and 4 used productions of entire words as stimuli, while Experiments 1 and 3 used only the initial CV extracted from each of these productions. The stimuli in Experiments 2 and 4 were identical, and the stimuli in Experiments 1 and 3 were identical.

Participants

In each experiment, 20 listeners completed the free classification task. These listeners also completed the rating task, and thus were previously described in Chapters 4 and 5.

Analysis

A 30x30 talker similarity matrix was constructed for each listener on the basis of the groups they created on the 16x16 grid. If a listener placed two talkers in the same group, the similarity between those two talkers was coded as 1. If a listener placed two talkers in different groups, their similarity was coded as 0. These matrices were averaged over all 20 listeners in a given experiment, such that a similarity of 1 indicated a pair of talkers who had been grouped together by all listeners, and a similarity of 0 indicated a pair of talkers who had never been grouped together by any listeners. These similarities were then converted to dissimilarities by subtracting each averaged value from 1.

Two types of analyses were performed with the dissimilarity matrices. The first was a clustering analysis, in order to reveal general patterns of grouping by listeners. The GTREE program (Corter, 1998) was used to create additive similarity trees, which represent perceptual distances between each pair of talkers as well as arranging the talkers into clusters. However, as response patterns for 30 data points were quite complex, the discussions below focus only on talker membership in the highest-level clusters displayed in each tree. Although interpretation at this level of the clustering solution may not capture all the nuances of listeners' responses, it allows for the comparison of major patterns across experiments. The full clustering solution for each experiment is provided in Appendix B, labeled with each talker's identity and the word targeted in his or her speech sample.[9]

[9]The design of this free classification experiment differs from many others in that each talker produced a unique CV or word, rather than all talkers sharing a common target. A full investigation of clustering patterns

The second analysis used multidimensional scaling (MDS) to explore specific dimensions of perceptual similarity and relate them to acoustic properties. To facilitate comparison across experiments, which is not directly possible with most MDS methods, the INDSCAL algorithm was used (Carroll and Chang, 1970). INDSCAL solutions cannot be rotated and can be performed on multiple dissimilarity matrices at once, with weights for each dimension calculated for each input matrix. In the present work, dissimilarity matrices for multiple experiments involving similar auditory stimuli were entered into the same INDSCAL analyses, as discussed in more detail below.

7.1.2 Results

Clustering

Main cluster membership from the GTREE solution for Experiment 1 is shown in the left column of Figure 7.2. Cluster 1 might be considered the "L1 American English" cluster, as it contained all six talkers from that language background. Also included in this cluster were the L1 Hindi talker who was thought to be targeting American rather than Indian English (H3f), most of the L1 Korean talkers (K2f, K4m, K5m, K6m), three L1 Mandarin talkers (M2f, M4m, M5m), and a single L1 Spanish talker (S4m). Cluster 3 contained almost all the remaining talkers, including all the L1 Hindi talkers who were judged to be targeting Indian English (H1f, H2f, H4m, H5m, H6m), most of the L1 Spanish talkers (S1f, S2f, S3f, S5m), three L1 Mandarin talkers (M1f, M3f, M6m), and one L1 Korean talker (K1f). Only two talkers, K3f and S6m, were included in Cluster 2.

in these data should take into account the segments present in each sample of speech, as some segments may reveal information about a talker's native language more clearly than others. However, segmental effects seemed minimal at the high level of clustering discussed in the present work.

Overall, L1 American English talkers were consistently grouped together by listeners. Similarities also seemed to be perceptible among multiple talkers with language backgrounds of Korean, Hindi, and Spanish. L1 Hindi and L1 Spanish talkers were generally classified together, while L1 Korean talkers were grouped with L1 American English talkers. There was no clear pattern for L1 Mandarin talkers, who were evenly divided across two clusters.

The stimuli in Experiment 2 were like those in Experiment 1, but included the entire word rather than only the initial CV. Main cluster membership from the GTREE solution for Experiment 2 is shown in the right column of Figure 7.2. Cluster 1 was similar to the "L1 American English" cluster from Experiment 1, except that it did not include talkers K4m or S4m and did include talker K1f. Cluster 3 bore some resemblance to Experiment 1's Cluster 3, although some of the members were different. Nonetheless, it included the L1 Hindi talkers who were thought to be targeting Indian English (H1f, H2f, H4m, H5m, H6m), most of the L1 Spanish talkers (S1f, S4m, S5m, S6m), two L1 Mandarin talkers (M3f, M6m), and a single L1 Korean talker (K4m). Cluster 2, with two L1 Spanish talkers (S2f, S3f) as well as talkers K3f and M1f, was both small and diverse.

These results were extremely similar to those of Experiment 1. Classification based on L1 group was reasonably successful for talkers with American English, Hindi, Korean, and Spanish language backgrounds. L1 American English and L1 Korean talkers tended to be grouped together, as were L1 Hindi and L1 Spanish talkers. L1 Mandarin talkers, however, appeared in all three clusters of the analysis.

While ordering and thus numbering of clusters in GTREE solutions is arbitrary, clusters with the same numbers happened to be quite comparable across Experiments 1 and 2. The lines shown in Figure 7.2 highlight the talkers whose cluster membership differed in

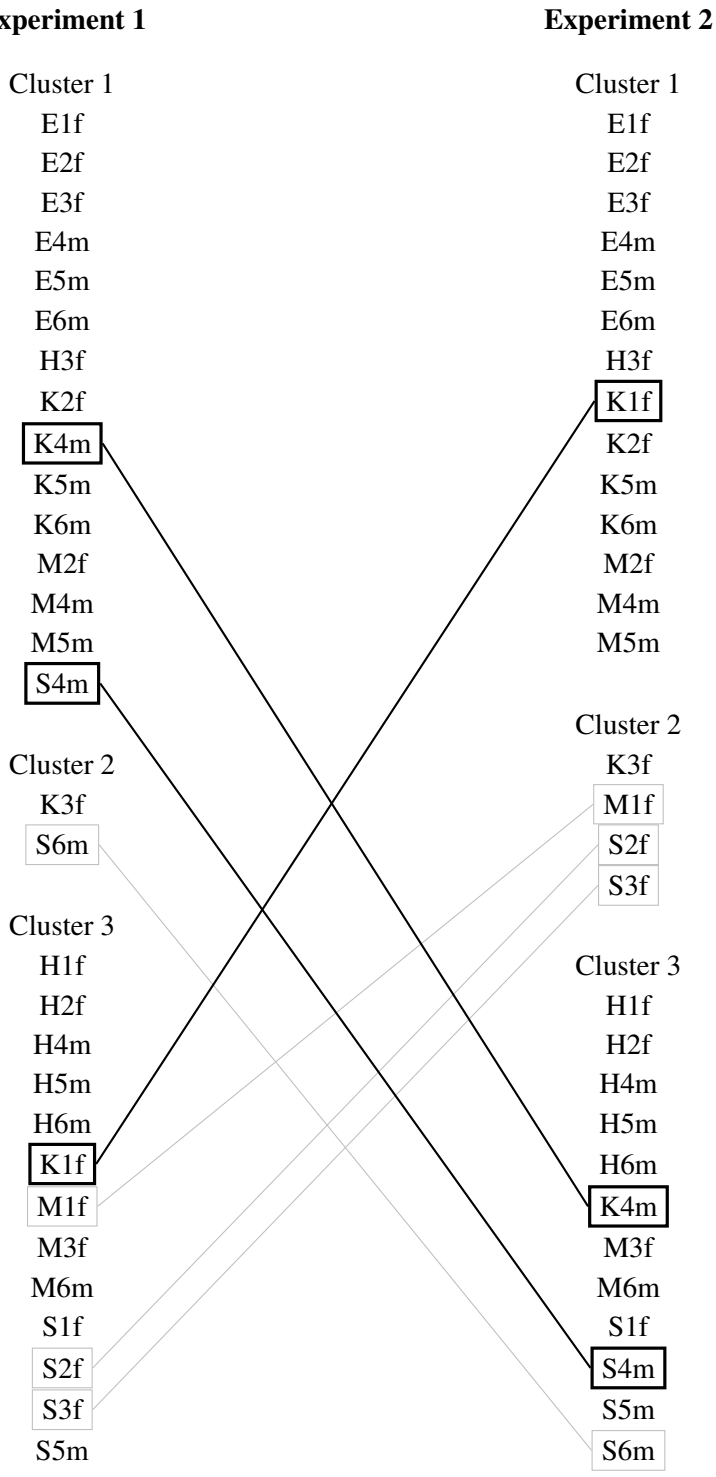

Figure 7.2: Main cluster membership of GTREE solutions for Experiments 1 and 2

Experiment 2 as compared to Experiment 1. Dark lines indicate talkers who belonged to the "L1 American English" cluster in one of the two experiments, while light lines are used for other talkers. On the whole, hearing a word rather than a CV did not substantially influence listeners' free classification responses, as revealed by the relatively small number of lines. That is, only 7 of the 30 talkers belonged to a different cluster in Experiment 2 as compared to Experiment 1.

Experiment 3 involved the same stimuli as Experiment 1. The procedure for the free classification task was also identical, except that Experiment 3 participants were told they were listening to "talkers from different language backgrounds" rather than "talkers with different accents," and prior to the free classification task, half the participants in Experiment 3 had completed a rating task about non-nativeness rather than accentedness. Main cluster membership from the GTREE solution for Experiment 3 is shown in the left column of Figure 7.3.[10] Cluster 1 was strikingly similar to the "L1 American English" cluster from Experiment 1. In fact, the only differences were the absence of talker K4m and the presence of two additional L1 Hindi talkers (H4m, H6m) and one additional L1 Spanish talker (S3f). In the remaining clusters, the results for Experiments 1 and 3 diverged somewhat, despite having been based on identical stimuli. Cluster 3 contained three L1 Hindi talkers (H1f, H2f, H5m), two L1 Spanish talkers (S5m, S6m), and two L1 Korean talkers (K1f, K3f). Cluster 2 for Experiment 3 included two L1 Mandarin talkers (M3f, M6m) and two L1 Spanish talkers (S1f, S2f). Only talkers K4m and M1f were in Cluster 4.

Again, listeners in Experiment 3 consistently grouped together the L1 American English talkers. The interpretation of the rest of these results, however, was much less clear

[10] As mentioned previously, ordering and thus numbering of clusters in a GTREE analysis is arbitrary. In Figure 7.3, Cluster 3 from Experiment 3 precedes Cluster 2 for a clearer visual comparison to Cluster 2 from Experiment 4.

than for the results from Experiment 1. Talkers from Korean, Mandarin, and Spanish backgrounds appeared in three of the four clusters, and L1 Hindi talkers were evenly split across two clusters. As the stimuli involved here were only CV sequences, with minimal repetition of the consonants and vowels contained therein, listeners had few direct comparisons upon which to base their decisions. Perhaps the clear patterning of clustering in Experiment 1 should be more surprising than the apparent lack of patterning in the present results.

Experiment 4 involved the same stimuli as Experiment 2. The procedure for the free classification task was also identical, except that Experiment 4 participants were told they were listening to "talkers from different language backgrounds" rather than "talkers with different accents," and prior to the free classification task, half the participants in Experiment 4 had completed a rating task about non-nativeness rather than accentedness. Main cluster membership from the GTREE solution for Experiment 4 is shown in the right column of Figure 7.3. Cluster 1 contained all six L1 American English talkers, the L1 Hindi talker who was judged to be targeting American rather than Indian English (H3f), and three L1 Mandarin talkers (M2f, M4m, M5m). Also included in this cluster were nearly all the L1 Korean talkers (K1f, K2f, K4m, K5m, K6m) and three L1 Spanish talkers (S1f, S2f, S5m). Cluster 2 contained all the L1 Hindi talkers who were thought to be targeting Indian English (H1f, H2f, H4m, H5m, H6m), the remaining three L1 Mandarin talkers (M1f, M3f, M6m), the remaining three L1 Spanish talkers (S3f, S4m, S6m), and a single L1 Korean talker (K3f).

In Experiment 4, classification of multiple talkers from the same language background was generally accurate for L1 American English and L1 Korean talkers, who tended to be in the same group, as well as L1 Hindi talkers, who comprised a separate group. L1 Mandarin and L1 Spanish talkers were split evenly across the two clusters.

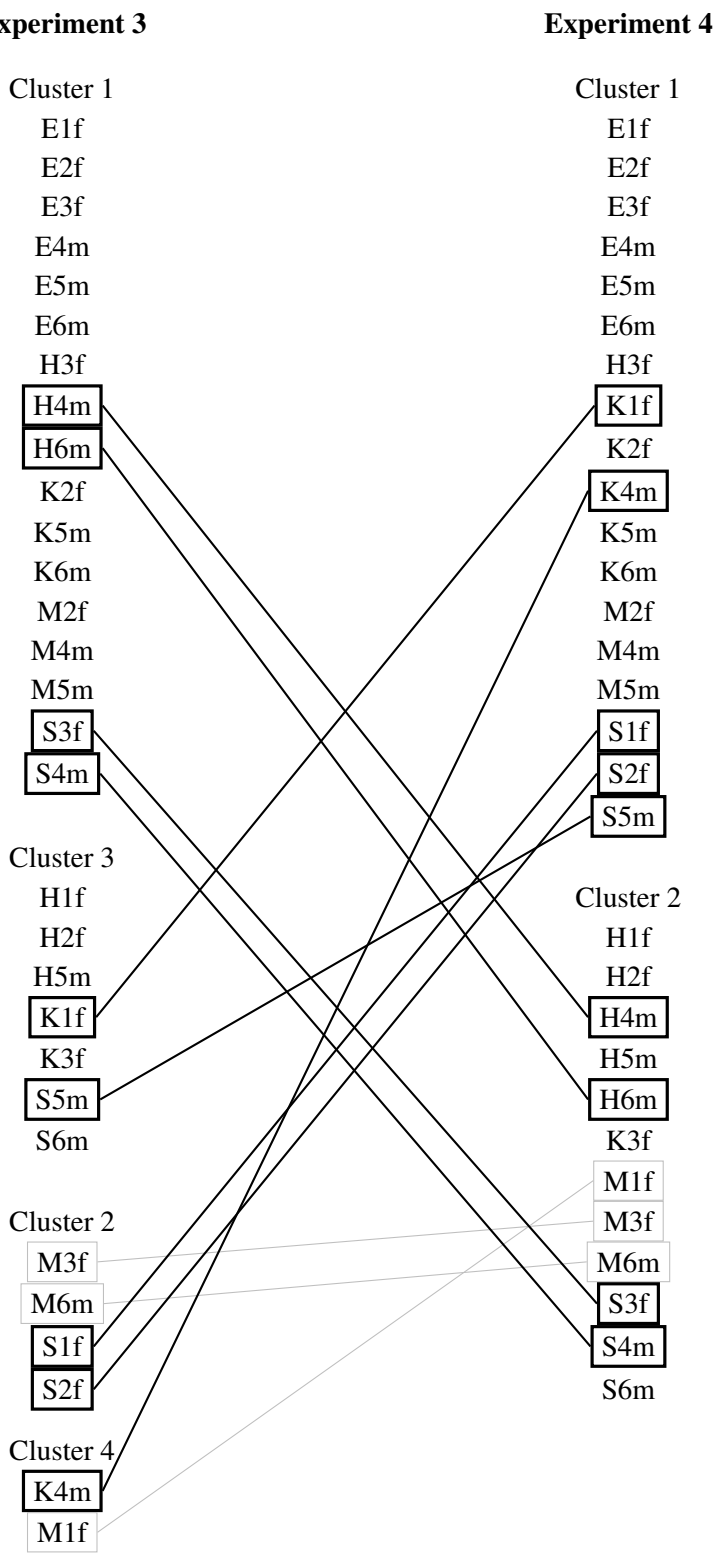

Figure 7.3: Main cluster membership of GTREE solutions for Experiments 3 and 4

The clusters from Experiments 3 and 4 were generally less comparable than were those from Experiments 1 and 2, largely because Experiment 3 involved four clusters while Experiment 4 had only two. Nonetheless, the "L1 American English" clusters were quite similar, and Cluster 3 from Experiment 3 shared many members with Cluster 2 from Experiment 4. The lines in Figure 7.3 signify talkers who were classified differently in Experiment 4 as compared to Experiment 3. As before, dark lines highlight talkers who belonged to the "L1 American English" cluster in one of the two experiments, and light lines are used for other talkers. Generally, while there were numerous non-native talkers in the "L1 American English" cluster for each experiment, the identities of many of those talkers differed. Hearing words rather than CVs caused listeners to classify some talkers differently, although the differences did not pattern consistently.

Experiments 1 through 4 as a whole showed that listeners were talented at grouping together native talkers of American English. They were also fairly successful with L1 Korean talkers, who tended to pattern with the L1 American English talkers, and with L1 Hindi talkers, who formed a different group entirely. L1 Spanish talkers were only sometimes classified with one another, and no more than half of the L1 Mandarin talkers were ever placed in the same group. Thus, L1 Hindi talkers are generally believed not to share a native language with L1 American English talkers, and foreign accent in the productions of L1 Spanish and L1 Mandarin talkers may be perceived less consistently than in those of other non-native talkers, at least in these particular stimuli. Additionally, the "L1 American English" group tended to be quite similar across experiments, while the other groups varied considerably, suggesting that listeners' judgments about native and native-sounding talkers were more consistent than their judgments about foreign-sounding talkers.

Multidimensional scaling

To explore the relationship between acoustic properties and free classification responses, the dissimilarity matrices from Experiments 1 through 4 were submitted to an INDSCAL analysis. In addition to the benefit of allowing direct comparison across multiple sets of data, this approach was deemed appropriate because these experiments involved shorter and longer versions of the same stimuli, and the clustering solutions discussed above revealed many similarities across the groupings made in these experiments. A plot of INDSCAL's model fit value (r^2) for models with 1 through 5 dimensions revealed no clear "elbow," and models with more dimensions were not possible with only 30 data points. A model with 4 dimensions ($r^2 = 0.50$) was selected, as one dimension in the model with 5 dimensions was uninterpretable as an acoustic property. The dimensions are plotted in Figure 7.4.

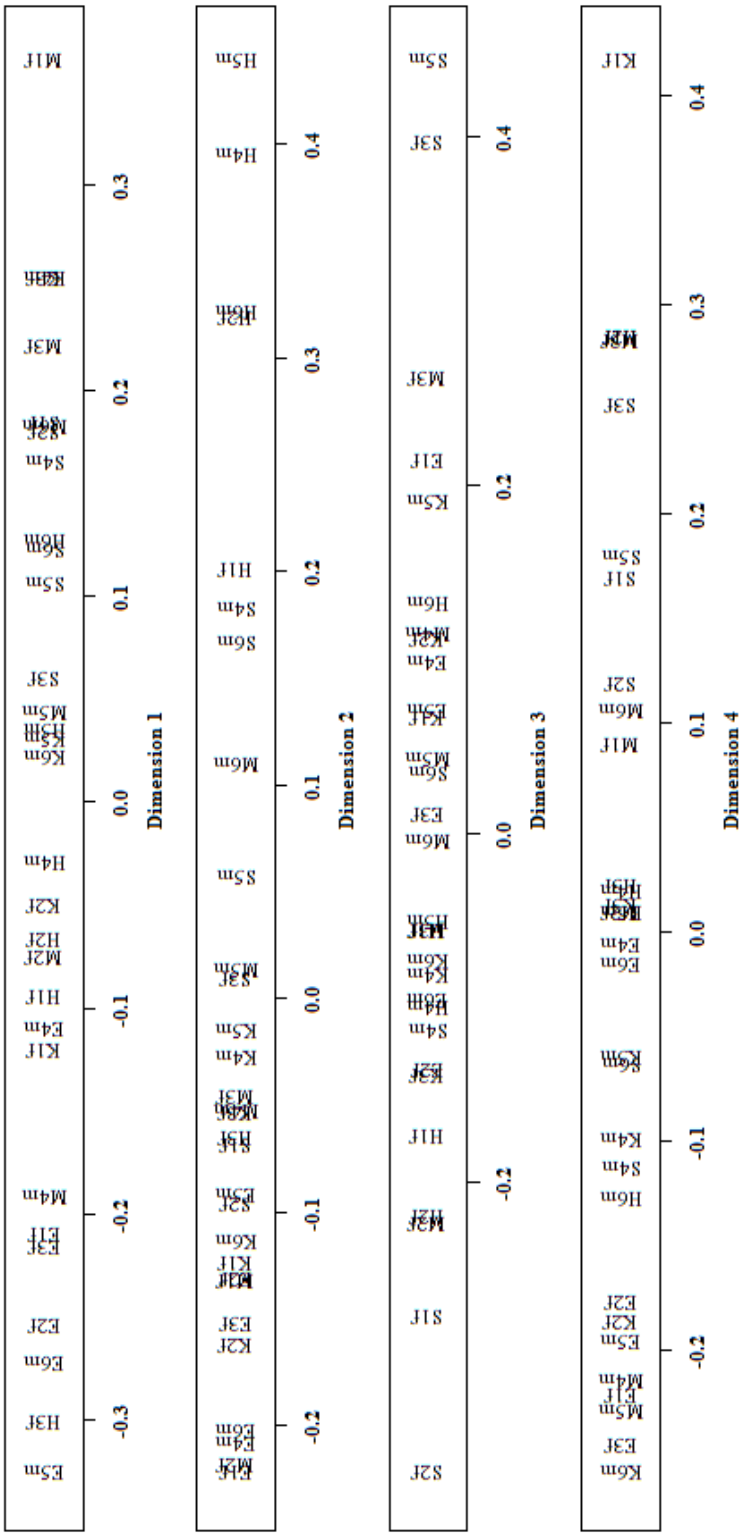

Figure 7.4: MDS dimensions 1 through 4 for Experiments 1 through 4

The R function glmulti() was used to select the single acoustic property, or two-way interaction between acoustic property and talker sex, that best correlated with each dimension. In fact, there were three separate glmulti() runs for each dimension, using the signed, absolute, and squared difference values of the acoustic properties. The r^2 values of the "best" variable from each of the three runs were compared, and the variable with the highest r^2 value was selected as the correlate for that dimension. In many cases the same acoustic property was selected on multiple runs, although one parameterization, typically the signed difference, had the best fit. The acoustic properties selected in this way, as well as the details of a linear regression between each acoustic property and the relevant MDS dimension, are shown in Table 7.1. Dimensions 1 through 4 correlated most strongly with vowel duration, VOT, F2 frequency, and F1 frequency, respectively. However, r^2 values ranging from 0.15 to 0.27 suggest that these relationships were not especially strong, possibly indicating that the relationships between the MDS dimensions of the classification responses and the properties of the signal to which listeners seem to attend in making their classification responses are more complex than can be captured by the current analysis.

In Chapters 4 and 5, it was found that VOT, F1 frequency, and F2 frequency correlated with all sets of ratings, and that vowel duration correlated with ratings for word-length stimuli. The fact that many acoustic properties correlated with perceptual data from both tasks strongly supports the idea that listeners attend to similar aspects of speech whether classifying talkers by language background or rating accentedness and non-nativeness.

The weights calculated by the INDSCAL algorithm are given in Table 7.2. In Experiments 1 and 3, Dimension 4 (F1 frequency) was weighted most heavily, while in Experiments 2 and 4, the most influential dimensions were 1 (vowel duration) and 2 (VOT). It is

Dimension	Property	Type	Coefficient	r^2	Significance
1	vowel duration	signed difference	0.0028	0.27	$p < 0.01$
2	VOT	signed difference	-0.0018	0.23	$p < 0.01$
3	F2 frequency	absolute difference	-0.0003	0.15	$p < 0.05$
4	F1 frequency	signed difference	-0.0007	0.21	$p < 0.01$

Table 7.1: Acoustic correlates of MDS dimensions for Experiments 1 through 4

notable that all dimensions were weighted similarly in the pairs of experiments with identical stimuli. In Chapters 4 and 5, vowel duration correlated with accentedness and nonnativeness ratings for word-length stimuli only, and likewise, the weights for Dimension 1 suggest that it was more influential in the classification of words than in the classification of CVs.

It is not immediately obvious why F1 frequency should matter more to the perception of CVs than words, and VOT more to the perception of words than CVs. One important difference between the rating and free classification tasks was that for free classification, the word targeted, or the word from which the CV had been extracted, was not made known to the listener. In non-native speech, acoustic deviations reflecting L1 transfer, unlike general measures of fluency, can only be determined if the relevant native phonetic norm is known. The words that served as stimuli (included in the labels in Figures B.1 through B.4) were fairly common, but lexical knowledge could not have been an advantage for listeners hearing only the initial CVs. Thus, perhaps the differing weights for F1 frequency and VOT were related to listeners in Experiments 1 and 3 making some incorrect phonetic comparisons.

Dimension	Experiment 1	Experiment 2	Experiment 3	Experiment 4
1	0.63	1.42	0.83	1.30
2	0.61	1.50	0.60	1.40
3	0.60	0.94	0.76	0.86
4	1.31	0.69	1.48	0.58

Table 7.2: Weights of MDS dimensions for Experiments 1 through 4

7.2 Experiments 5 and 6: Talkers' L1s

7.2.1 Methods

The procedure for Experiments 5 and 6 was identical to the procedure used in Experiments 1 through 4. The 30 auditory stimuli in the classification task for each experiment were a subset of those included in the rating task, each produced by a different talker. Each talker produced a unique sequence in his or her native language. Stimuli were selected such that the 6 talkers within an L1 background produced unique initial consonants and a variety of vowels. Experiment 6 used productions of entire words as stimuli, while Experiment 5 used only the initial CV extracted from each of these productions. In each experiment, 20 listeners completed the free classification task. These listeners also completed the rating task, and thus were previously described in Chapter 6.

Listeners were again instructed to "group the talkers by native language," although the target languages as well as the native languages differed. However, the same instructions were given for the free classification portion of all six experiments in order to facilitate comparison of the results.

7.2.2 Results

Clustering

Main cluster membership from the GTREE solution for Experiment 5 is shown in the left column of Figure 7.5.[11] Cluster 1 might be considered the "L1 American English" cluster, as it contained the six L1 American English talkers. However, it also included most of the L1 Hindi talkers (H3f, H4m, H5m, H6m), two L1 Korean talkers (K4m, K6m), and talkers M6m and S4m. Cluster 3 included the remaining five L1 Mandarin talkers (M1f, M2f, M3f, M4m, M5m), as well as two L1 Spanish talkers (S1f, S2f). The members of Cluster 2 were quite varied, with four L1 Korean talkers (K1f, K2f, K3f, K5m), three L1 Spanish talkers (S3f, S5m, S6m), and two L1 Hindi talkers (H1f, H2f).

In Experiment 5, listeners successfully grouped together the L1 American English talkers, and were also rather adept at grouping talkers with language backgrounds of Mandarin, Korean, and Hindi. Surprisingly, L1 Hindi talkers were classified with L1 American English ones, while L1 Mandarin talkers and L1 Korean talkers formed additional groups. L1 Spanish talkers did not pattern consistently, appearing in all three clusters.

Main cluster membership from the GTREE solution for Experiment 6 is shown in the right column of Figure 7.5. Cluster 1 included only the six L1 American English talkers. Cluster 2 contained all six L1 Mandarin talkers, four L1 Korean talkers (K1f, K2f, K3f, K6m), three L1 Hindi talkers (H1f, H3f, H6m), and talker S1f. Cluster 3 included the remaining five L1 Spanish talkers (S2f, S3f, S4m, S5m, S6m), three L1 Hindi talkers (H2f, H4m, H5m), and two L1 Korean talkers (K4m, K5m).

With word-length stimuli, listeners in Experiment 6 were able to classify the L1 American English talkers perfectly accurately into their own distinct group. They were also

[11] Again, ordering and thus numbering of clusters in a GTREE analysis is arbitrary. In Figure 7.5, Cluster 3 of Experiment 5 precedes Cluster 2 for a clearer visual comparison to Cluster 2 from Experiment 6.

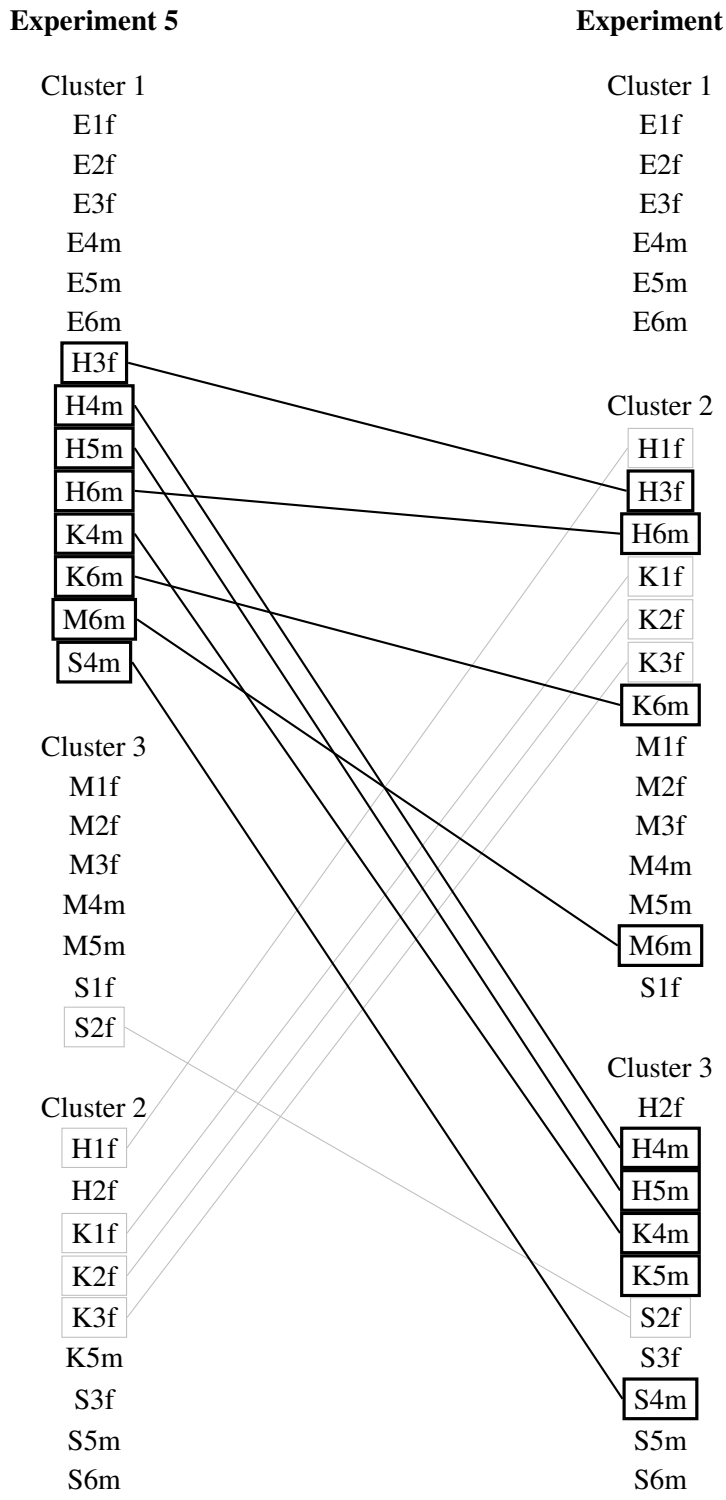

Figure 7.5: Main cluster membership of GTREE solutions for Experiments 5 and 6

successful in grouping together multiple talkers from Mandarin, Korean, and Spanish language backgrounds. L1 Hindi talkers, however, were evenly divided across two clusters.

Cluster 1 from Experiment 5 was like Cluster 1 from Experiment 6 in that it included all six L1 American English talkers, although in Experiment 5 it also included eight talkers of other language backgrounds. Cluster 3 from Experiment 5 and Cluster 2 from Experiment 6 could be compared on the basis of containing most of the L1 Mandarin talkers, and Cluster 2 from Experiment 5 and Cluster 3 from Experiment 6 could be compared as they both included a number of the L1 Spanish talkers. The lines in Figure 7.5 indicate talkers whose cluster membership differed in Experiment 6 as compared to Experiment 5. Again, dark lines signify talkers who belong to the "L1 American English" cluster in one of the two experiments, while light lines are used for other talkers. Two important differences are evident. No non-native talker was included in the "L1 American English" cluster based on responses to word-length stimuli, although many were in the clustering solution based on responses to shorter stimuli. Additionally, L1 Korean talkers and L1 Mandarin talkers were robustly clustered together for the word-length stimuli only. Thus, hearing words rather than CVs had clear effects on listeners' free classification responses.

Generally, listeners were generally able to group together talkers from the same language background (that is, talkers speaking the same language), with somewhat less accurate performance for L1 Spanish talkers (with CVs) and L1 Hindi talkers (with words). With CV-length stimuli, L1 Korean talkers were grouped separately from L1 Mandarin talkers, while with word-length stimuli, these talkers were combined. This pattern is particularly notable given the difficulty that other English-speaking listeners have experienced in distinguishing between East Asian languages in a variety of tasks (Bond and Stockmal, 2002; Derwing and Munro, 1997; Stockmal et al., 1994). This result seems to indicate that

there are shared characteristics of Korean and Mandarin that are not salient when initial CV sequences are spliced out of their larger whole-word contexts.

Multidimensional scaling

In the rating task, listeners had to respond to one stimulus before hearing another, and thus could only compare each production to a mental representation or memory of speech. In the free classification task, listeners could play the stimuli repeatedly and in any order, allowing for easier comparison among the stimuli themselves. Although raw measurements of acoustic properties were not appropriate for the analysis of the ratings in Chapter 6, as the targets against which listeners were evaluating the stimuli were not collected, they are somewhat more suitable for the analysis of results from the free classification task, in which all stimuli could be compared to one another. Multidimensional scaling was again used to link listeners' classification responses to acoustic characteristics of the stimuli.

The dissimilarity matrices from Experiments 5 and 6 were submitted to an INDSCAL analysis so that the results could be compared directly. Again, a plot of INDSCAL's model fit value (r^2) for models with 1 through 5 dimensions revealed no clear "elbow," and models with more dimensions were not possible with only 30 data points. A model with 3 dimensions ($r^2 = 0.42$) was selected, as some dimensions in the more complex models were not interpretable as any of the acoustic properties measured. The dimensions are plotted in Figure 7.6.

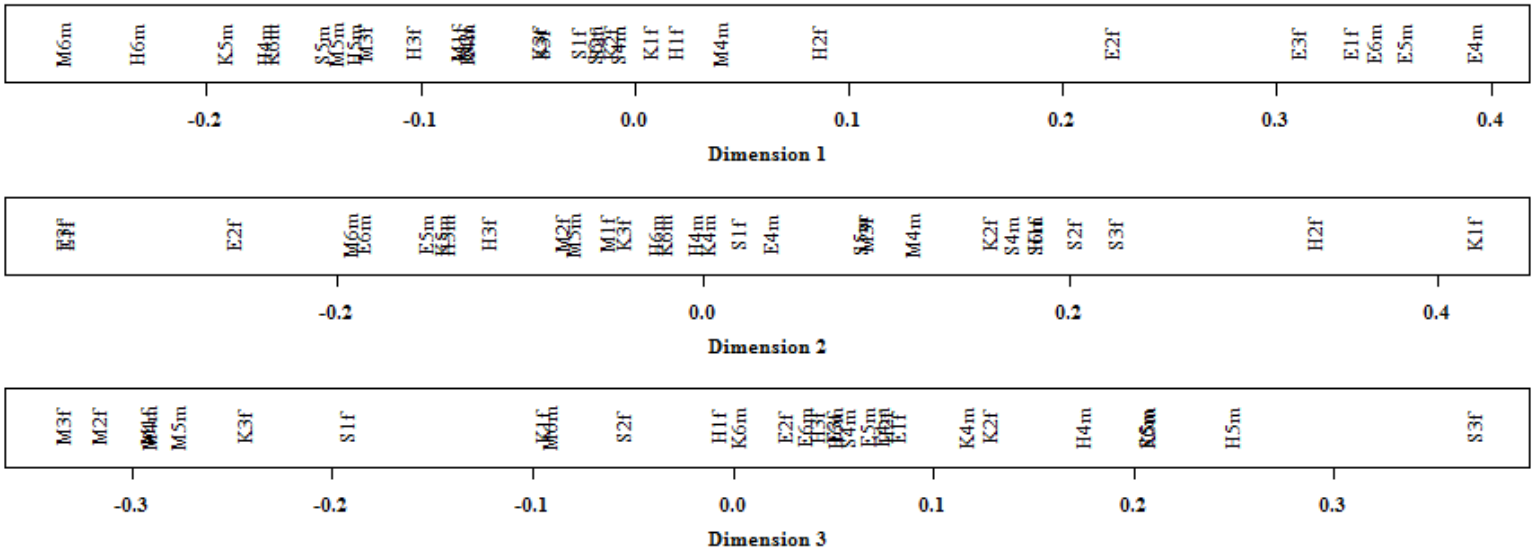

Figure 7.6: MDS dimensions 1 through 3 for Experiments 5 and 6

Dimension	Property	Type	Coefficient	r^2	Significance
1	F2 tilt	raw	0.0007	0.11	$p < 0.05$
2	VOT	raw	-0.0015	0.17	$p < 0.05$
3	vowel duration	raw	-0.0015	0.25	$p < 0.01$

Table 7.3: Acoustic correlates of MDS dimensions for Experiments 5 and 6

The procedure described in Section 7.1.2 was used to investigate the relationship between these dimensions and the various acoustic properties measured. Because difference values were not calculable for these stimuli, as discussed in Chapter 6, only one run was performed for each dimension, using raw measurements of the acoustic properties. The results of this analysis are shown in Table 7.3. Dimensions 1 through 3 correlated most strongly with F2 tilt, VOT, and vowel duration, respectively, although relatively low r^2 values ranging from 0.11 to 0.25 again suggest that this analysis could potentially be improved upon. When the rate of recognition as a word of English with included in the model, in light of its strong relationship to the ratings from Experiment 6, Dimension 1 correlated most strongly with this property (b = 0.0481, r^2 = 0.52, $p < 0.001$), revealing another similarity between the rating and free classification data.

The INDSCAL weights for these dimensions are given in Table 7.4. In Experiment 5, Dimension 2 (VOT) was weighted most heavily, while in Experiment 6, Dimensions 1 (F2 tilt/rate of recognition as a word of English) and 3 (vowel duration) were most important. Again, as observed in the free classification results of Experiments 1 through 4, vowel duration was more heavily weighted for word-length than for CV-length stimuli.

Dimension	Experiment 5	Experiment 6
1	0.51	1.50
2	1.11	1.01
3	0.72	1.28

Table 7.4: Weights of MDS dimensions for Experiments 5 and 6

7.3 Discussion

Overall patterns in the free classification responses showed that L1 American English listeners were most consistent at grouping stimuli from L1 American English talkers. They also tended to group together stimuli from L1 Korean talkers, and L1 Spanish talkers were grouped together in some experiments, but not in others. Results for talkers from the other language backgrounds depended on the language of the stimuli. L1 Mandarin talkers speaking English were split across groups, but they were consistently grouped together when speaking Mandarin. Conversely, L1 Hindi talkers were generally grouped together when speaking English, but less consistently when speaking Hindi. While previous studies have shown that not all talkers or productions are matched equally well to (native) language labels (Bond and Fokes, 1991; Derwing and Munro, 1997; Vasilescu et al., 2005; Vieru et al., 2011), the present work demonstrates that instances of speech from the same language variety may be classified separately even when explicit labels are removed from the task.

Investigation of acoustic properties revealed that vowel duration, VOT, F1 frequency, and F2 frequency and tilt were correlated with various dimensions of listeners' perceptual spaces in the free classification tasks, with some differences depending on the length and language of the stimuli. Notably, the dimensions that correlated with vowel duration

were weighted more heavily in experiments with word-length rather than CV-length stimuli, bearing some resemblance to patterns observed regarding the acoustic correlates of accentedness and non-nativeness ratings. On the whole, while the rating and free classification tasks demanded very different types of responses, the acoustic correlates of their responses were quite similar. Furthermore, although acoustic correlates of non-Englishness ratings could not be calculated, as discussed in Chapter 6, rate of recognition as a word of English correlated with both the ratings and the free classification data. In general, similar properties of the signal seemed to guide listeners' responses across these different tasks.

CHAPTER 8: CONCLUSION

While many investigations have explored listeners' attitudes toward different varieties of speech, it has not been clear what characteristics of the signal trigger those attitudes. As the number of non-native English speakers grows, and interaction between native and non-native English speakers increases, the perception of foreign accent and related characteristics becomes more relevant to English speakers and merits more attention from language researchers. The experiments in this dissertation contribute to the overall understanding of accentedness, non-nativeness, and foreignness in speech by identifying the acoustic properties that might contribute to their perception and exploring listeners' abilities to abstract over different productions.

Of course, the research discussed here studied these perceptions with only English as a target language (for accentedness and non-nativeness) or a reference language (for foreignness), and only by L1 American English listeners enrolled in linguistics courses at a single university. While foreign accent is not unique to this particular English-speaking community, it may be manifested differently in other cultures. Moreover, it is surely manifested differently in other target languages, such that the acoustic correlates of accentedness and non-nativeness in English, or of foreignness with reference to English, are not necessarily relevant for other contexts. However, this investigation could serve as a model for similar investigations in other varieties of English and in other languages.

As with most studies of foreign accent, the results of these experiments have implications for foreign language instruction. For the many L2 learners of English who express

accent reduction as a high priority (Derwing and Munro, 2009), language instructors would be wise to focus attention on those details of the signal that contribute most to the perceptions of accentedness and non-nativeness, which might not overlap entirely with sounds that are articulatorily difficult. Knowing which details these were would be of particular value in classrooms where the learners do not share a common native language, and thus make a variety of different types of errors in production of the target language. A talker perceived to be less accented may be afforded more employment opportunities, and be socially more accepted, than a talker with more accented speech (see Gluszek and Dovidio, 2010; Munro, 2003).

8.1 Summary of results

Listeners provided ratings closer to the labeled ends of the rating scales when hearing word-length as compared to CV-length stimuli. As in Flege and Munro (1994), it seemed that listeners were attending to acoustic details throughout the stimulus, such that their judgments were more certain for longer stimuli. Additionally, ratings of accentedness and non-nativeness were nearly perfectly correlated with one another, suggesting that listeners may not have viewed these as distinct concepts. However, similar to Cheong's (2007) finding, listeners' ratings tended to be closer to "no, definitely not [native]" than to "strong foreign accent" for the very same talker. Such a pattern may reflect a more binary interpretation of non-nativeness than of accentedness—with fewer ratings between the ends of the scale—which would not be surprising given that the labels on the non-nativeness scale (but not the accentedness scale) were clear opposites. Nonetheless, that this detail affects the distribution of responses should encourage experimenters to consider their labeling choices carefully.

Acoustic property	Experiment 1	Experiment 2	Experiment 3	Experiment 4
VOT	✓	✓	✓	✓
vowel duration		✓		✓
F1 frequency	✓	✓	✓	✓
F2 frequency	✓	✓	✓	✓
F3 frequency	✓		✓	
F2 tilt	✓		✓	
VOT:voicing	✓	n/a	n/a	n/a
f0:sex		✓		
H1-H2:sex	✓		✓	
F2 curvature:sex		✓		

Table 8.1: Summary of acoustic correlates in Experiments 1 through 4

The relationships between 13 acoustic properties and ratings of accentedness were tested simultaneously, as were the relationships between these acoustic properties and ratings of non-nativeness. Acoustic correlates of the ratings from Experiments 1 through 4 are shown in Table 8.1. The present summary focuses on the simple effects, as it was less clear how to interpret many of the interactions.

The results for Experiments 1 and 3, and for Experiments 2 and 4, show that listeners perceive accentedness and non-nativeness quite similarly, again supporting the possibility that they do not distinguish these concepts. That vowel duration correlated with ratings for words, but not for CVs, may possibly be explained as listeners attending to more global properties in longer stimuli. In the absence of information about the assimilation patterns of foreign sounds by L1 American English listeners, acoustic correlates of non-Englishness ratings in Experiments 5 and 6 were not evaluated, but ratings for Experiment 6 correlated with the rate of recognition of each stimulus as a word of English.

The acoustic properties that correlated with multidimensional scaling (MDS) analyses of listeners' responses in the free classification task included, for English stimuli, vowel duration, VOT, F2 frequency, and F1 frequency, and for stimuli in other languages, F2 tilt, VOT, and vowel duration. Although the relationships between MDS dimensions and these acoustic properties were not especially strong, it is striking that each of these acoustic correlates also related to accentedness and non-nativeness ratings for at least one version of the English stimuli. Additionally, the non-acoustic property that correlated with ratings of non-Englishness, rate of recognition as a word of English, also correlated with one MDS dimension of the free classification responses. Listeners seemed to attend consistently to the same properties in the speech signal, suggesting that perception of these language varieties was not task-dependent but rather was guided by the salient characteristics of the samples.

Discrete cosine transforms (DCTs) were used in this investigation to model formant tracks as curves, so that the role of dynamic as well as static formant information could be tested. Despite this, the only dynamic formant information found to be significant in any analysis was F2 tilt. Overall, the mean frequencies of F1 and F2, the formants that most strongly influence vowel perception generally (Pols et al., 1969), were consistently correlated with perceptions of accentedness and non-nativeness. Indeed, an equally consistent correlate, VOT, is generally important in stop perception (Abramson and Lisker, 1970). Socioindexical information is sometimes associated with acoustic properties that are not needed for linguistic contrasts, such as spectral tilt in English (Hanson and Chuang, 1999), but in this instance some acoustic properties seem to be communicating both socioindexical and linguistic information.

8.2 Discussion and future directions

Of course, the work presented here could be further developed in a variety of ways. An obvious direction would be to attempt to improve the performance of the models. Some of the random intercepts for talkers in Table B.1 have relatively extreme values, indicating that the acoustic properties included in these models were not fully capturing the rating patterns; indeed, if the acoustic properties accounted sufficiently well for the ratings observed, the random intercepts would not be needed at all. Likewise, some of the correlations between acoustic properties and the MDS dimensions of the free classification responses had rather low r^2 values, suggesting only weak relationships. It is highly likely that the acoustic properties measured in the present investigation did not capture all the relevant perceptible dimensions of the stimuli. For instance, many non-native speakers of English produce stops rather than flaps for word-medial /d, t/, but the acoustic measures for word-length stimuli did not include any information beyond the initial CV.

One acoustic property that was not quantified was otherwise considered in this work: the possible retroflexion of alveolar stop targets that might have contributed to the high accentedness ratings for L1 Hindi productions in a study by McCullough (2013). Although many of the L1 Hindi productions of the alveolar stop targets of English in the present investigation were informally judged by the author to sound retroflexed, in no experiment did an L1 Hindi talker's productions of stimuli with alveolar stop targets receive higher ratings than his or her productions of stimuli with labial and velar stop targets. If the listeners perceived this acoustic characteristic, they did not seem to make use of it in their evaluations of accentedness and non-nativeness, even in very short stimuli. Thus, it is not necessarily the case that every instance of L1 transfer is relevant to the perception of non-native speech.

It is possible that certain acoustic properties relevant to perception are not traditionally thought of as characterizing short units of speech, and thus were left out of the present analysis. Munro et al. (2010, 636) suggest that among other cues, voice quality resulting from "long-term configurations of the vocal tract" may help listeners to determine the native speaker status of talkers. "Articulatory setting" may well have influenced the perception of the present stimuli, as it is known to differ crosslinguistically (Wilson, 2006), but as it also lacks an accepted method of acoustic quantification, it was not included here.

Overall, the experiments analyzed in this dissertation contribute to the foreign accent literature by investigating the importance of 13 acoustic characteristics to the perception of short samples of English produced by L1 talkers of American English, Hindi, Korean, Mandarin, and Spanish, as well as to the perception of short samples of the talkers' native languages. When perceiving syllable-length stimuli, listeners seem to attend to phonetic details resulting from transfer from the non-native talker's L1, while indications of the talker's L2 fluency may begin to influence perception in units as small as disyllabic words.

REFERENCES

Abramson, A. S. and Lisker, L. (1970). Discriminability along the voicing continuum: Cross-language tests. In *Proceedings of the Sixth International Congress of Phonetic Sciences*, pages 569–573.

Alba-Salas, J. (2004). Voice Onset Time and foreign accent detection: Are L2 learners better than monolinguals? *Revista Alicantina de Estudios Ingleses*, pages 9–30.

Allen, W. S. (1957). Some phonological characteristics of Rajasthani. *Bulletin of the School of Oriental and African Studies*, pages 5–11.

Anderson-Hsieh, J., Johnson, R., and Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmental, prosody, and syllable structure. *Language Learning*, pages 529–555.

Ash, S. (2003). A national survey of North American dialects. *Publication of the American Dialect Society*, pages 57–73.

Awan, S. N. and Stine, C. L. (2011). Voice onset time in Indian English-accented speech. *Clinical Linguistics & Phonetics*, pages 998–1003.

Baker, R. E., Baese-Berk, M., Bonnasse-Gahot, L., Kim, M., Van Engen, K. J., and Bradlow, A. R. (2011). Word durations in non-native English. *Journal of Phonetics*, pages 1–17.

Bamford, J. and Wilson, I. (1979). Methodological considerations and practical aspects of the BKB sentence lists. In Bench, J. and Bamford, J., editors, *Speech-Hearing Tests and the Spoken Language of Hearing-Impaired Children*, pages 148–187. Academic Press, London.

Bansal, R. K. (1981). English and Hindi: A contrastive phonological study. *CIEFL Bulletin*, pages 51–60.

Bond, Z. S. and Fokes, J. (1991). Identifying foreign languages. In *Proceedings of the XII International Congress of Phonetic Science, Aix-en-Provence*, pages 198–201.

Bond, Z. S. and Stockmal, V. (2002). Distinguishing samples of spoken Korean from rhythmic and regional competitors. *Language Sciences*, pages 175–185.

Bond, Z. S., Stockmal, V., and Markus, D. (2008). A note on native and non-native accentedness judgments. *Ohio University Working Papers in Linguistics and Language Teaching*, pages 1–8.

Bond, Z. S., Stockmal, V., and Moates, D. R. (2003). Searching for foreign accent. *Vigo International Journal of Applied Linguistics*, pages 13–24.

Boula de Mareuil, P. and Vieru-Dimilescu, B. (2006). The contribution of prosody to the perception of foreign accent. *Phonetica*, pages 247–267.

Bradlow, A., Clopper, C., Smiljanic, R., and Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, pages 930–942.

Brennan, E. S. and Brennan, J. S. (1981). Measurements of accent and attitude toward Mexican-American speech. *Journal of Psycholinguistic Research*, pages 487–501.

Brière, E. J. (1966). An investigation of phonological interference. *Language*, pages 768–796.

Calla McDermott, W. L. (1986). *The scalability of degrees of foreign accent*. Dissertation, Cornell University.

Canfield, D. L. (1981). *Spanish pronunciation in the Americas*. University of Chicago Press, Chicago.

Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an n-way generalization of "Eckart=Young" decomposition. *Psychometrika*, pages 283–319.

Chen, Y., Robb, M., Gilbert, H., and Lerman, J. (2001). Vowel production by Mandarin speakers of English. *Clinical Linguistics & Phonetics*, pages 427–440.

Cheong, S. H. (2007). *The role of listener affiliated socio-cultural factors in perceiving native accented versus foreign accented speech*. Dissertation, Ohio State University.

Chin, S. (2006). *Sound Systems of Mandarin Chinese and English: A Comparison*. Lincom Europa.

Cho, T., Jun, S.-A., and Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, pages 193–228.

Choo, M. and O'Grady, W. (2003). *The Sounds of Korean: A Pronunciation Guide*. University of Hawai'i Press.

Corter, J. E. (1998). An efficient metric combinatorial algorithm for fitting additive trees. *Multivariate Behavioral Research*, pages 249–272.

Dalbor, J. B. (1969). *Spanish Pronunciation: Theory and Practice*. Holt, Rhinehart and Winston.

Davis, K. and Beckman, M. (1983). Production and perception of the voicing contrast in Indian and American English. *Working Papers of the Cornell Phonetics Laboratory*, pages 77–90.

Derwing, T. M. and Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, pages 1–16.

Derwing, T. M. and Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, pages 476–490.

Derwing, T. M., Munro, M. J., Thomson, R. I., and Rossiter, M. J. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, pages 533–557.

Donadio, A. (2002). *Spanish accented English: Pronunciation accuracy and factors affecting L2 acquisition*. Dissertation, Florida Atlantic University.

Dutta, I. (2007). *Four-way stop contrasts in Hindi: An acoustic study of voicing, fundamental frequency and spectral tilt*. Dissertation, University of Illinois at Urbana-Champaign.

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, pages 692–707.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics*, pages 47–65.

Flege, J. E., Bohn, O. S., and Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, pages 437–470.

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

Flege, J. E. and Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, pages 67–83.

Flege, J. E., Munro, M., and MacKay, I. (1995). Factors affecting strength of perceived foreign accent in a second language. *Journal of the Acoustical Society of America*, pages 3125–3134.

Flege, J. E. and Munro, M. J. (1994). The word unit in second language speech production and perception. *Studies in Second Language Acquisition*, pages 381–411.

Gargesh, R. (2004). Indian English: Phonology. In Schneider, E. W., Burridge, K., Kortmann, B., and Mesthrie, R., editors, *A Handbook of Varieties of English*, pages 992–1002. Mouton de Gruyter, Berlin.

Gluszek, A. and Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review*, pages 214–237.

Han, M. S. and Weitzman, R. S. (1970). Acoustic features of Korean /P,T,K/, /p,t,k/ and /p^h,t^h,k^h/. *Phonetica*, pages 112–128.

Hanson, H. M. and Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, pages 1064–1077.

Hardman, J. B. (2010). *The intelligibility of Chinese-accented English to international and American students at a US university*. Dissertation, Ohio State University.

Harris, J. W. (1969). *Spanish phonology*. MIT Press, Cambridge, MA.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, pages 3099–3111.

Institute of International Education (2012). Open Doors report on international educational exchange.

Jilka, M. (2000). *The contribution of intonation to the perception of foreign accent: Identifying intonational deviations by means of F0 generation and resynthesis*. Dissertation, Universität Stuttgart.

Kagaya, R. and Hirose, H. (1975). Fiberoptic, electromyographic, and acoustic analyses of Hindi stop consonants. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*, pages 27–46.

Kalin, R., Rayko, D. S., and Love, N. (1980). The perception and evaluation of job candidates with four different ethnic accents. In Giles, H., Robinson, W. P., and Smith, P., editors, *Language: Social Psychological Perspectives*, pages 197–202. Pergamon Press, Oxford.

Kang, K.-H. and Guion, S. G. (2006). Phonological systems in bilinguals: Age of learning effects on the stop consonant systems of Korean-English bilinguals. *Journal of the Acoustical Society of America*, pages 1672–1683.

Kang, K.-H. and Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of the Acoustical Society of America*, pages 3909–3917.

Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, pages 301–315.

Khan, I., Gupta, S. K., and Rizvi, S. H. S. (1994). Formant frequencies of Hindi vowels in /hvḋ/ and C1VC2 contexts. *Journal of the Acoustical Society of America*, pages 2580–2582.

Kim, M. R. (2011). Native and non-native English speakers' VOT productions of stops. *The Linguistic Association of Korea Journal*, pages 97–116.

Kinzler, K. D., Shutts, K., DeJesus, J., and Spelke, E. S. (2009). Accent trumps race in guiding children's social preferences. *Social Cognition*, pages 623–634.

Ladefoged, P. (1999). American English. In *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, pages 41–44. Cambridge University Press, Cambridge.

Lee, H. B. (1999). Korean. In *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, pages 120–123. Cambridge University Press, Cambridge.

Lee, W.-S. and Zee, E. (2003). Illustrations of the IPA: Standard Chinese (Beijing). *Journal of the International Phonetic Association*, pages 109–112.

Lewis, A. M. (2001). *Weakening of intervocalic /p,t,k/ in two Spanish dialects: Toward the quantification of lenition processes*. Dissertation, University of Illinois at Urbana-Champaign.

Lindemann, S. (2003). Koreans, Chinese or Indians? Attitudes and ideologies about non-native English speakers in the United States. *Journal of Sociolinguistics*, pages 348–364.

Lisker, L. and Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, pages 384–422.

Liu, H., Ng, M. L., Wan, M., Wang, S., and Zhang, Y. (2007). Effects of place of articulation and aspiration on voice onset time in Mandarin esophageal speech. *Folia Phoniatrica et Logopaedica*, pages 147–154.

Liu, H.-M., Tseng, C.-H., and Tsao, F.-M. (2000). Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *Clinical Linguistics and Phonetics*, pages 447–464.

Lorch, M. P. and Meara, P. (1995). Can people discriminate languages they don't know? *Language Sciences*, pages 65–71.

Macken, M. A. and Barton, D. (1980). The acquisition of the voicing contrast in Spanish: A phonetic and phonological study of word-initial stop consonants. *Journal of Child Language*, pages 433–458.

Magen, H. S. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, pages 381–400.

Magen, H. S. and Blumstein, S. E. (1993). Effects of speaking rate on the vowel length distinction in Korean. *Journal of Phonetics*, pages 387–409.

Major, R. C. (1987). English voiceless stop production by speakers of Brazilian Portuguese. *Journal of Phonetics*, pages 197–202.

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, pages 539–556.

Masica, C. P. (1991). *The Indo-Aryan languages*. Cambridge University Press, Cambridge.

Maxwell, O. and Fletcher, J. (2009). Acoustic and durational properties of Indian English vowels. *World Englishes*, pages 52–69.

McCullough, E. (2013). Perceived foreign accent in three varieties of non-native english. *Ohio State University Working Papers in Linguistics*, pages 51–66.

Mishra, D. and Bali, K. (2011). A comparative phonological study of the dialects of Hindi. *International Congress of Phonetic Sciences*, pages 1390–1393.

Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, pages 39–66.

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, pages 17–34.

Munro, M. J. (2003). A primer on accent discrimination in the Canadian context. *TESL Canada Journal*, pages 38–51.

Munro, M. J. and Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, pages 451–468.

Munro, M. J., Derwing, T. M., and Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, pages 626–637.

Nathan, G. S. (1987). On second-language acquisition of voiced stops. *Journal of Phonetics*, pages 313–322.

Nguyen, B. B.-D. (1993). Accent discrimination and the Test of Spoken English: A call for an objective assessment of the comprehensibility of nonnative speakers. *California Law Review*, pages 1325–1361.

Norman, J. (1988). *Chinese*. Cambride University Press, Cambridge.

Ohala, M. (1999). Hindi. In *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, pages 100–103. Cambridge University Press, Cambridge.

Ortega-Llebaria, M. (1997). *An explanatory intelligibility test for Spanish accented English*. Dissertation, Indiana University.

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, pages 261–285.

Pandharipande, R. (2003). Marathi. In Cardona, G. and Jain, D., editors, *The Indo-Aryan languages*, pages 698–728. Routledge, London.

Plichta, B. (2012). Akustyk blog: Discrete Cosine Transform. `http://bartus.us/blog/?p=531`.

Pols, L. C. W., van der Kamp, L. J. T., and Plomp, R. (1969). Perceptual and physical space of vowel sounds. *Journal of the Acoustical Society of America*, pages 458–467.

Purnell, T., Idsardi, W., and Baugh, J. (1999). Perceptual and phonetic experiments on American English dialect identification. *Journal of Language and Social Psychology*, pages 10–30.

Riney, T. J. and Flege, J. E. (1998). Changes over time in global foreign accent and liquid identifiability and accuracy. *Studies in Second Language Acquisition*, pages 213–243.

Riney, T. J. and Takagi, N. (1999). Global foreign accent and voice onset time among Japanese EFL speakers. *Language Learning*, pages 275–302.

Rogers, C. L. and Dalby, J. (2005). Forced-choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research*, pages 306–322.

Ryan, E. B., Carranza, M. A., and Moffie, R. W. (1977). Reactions toward varying degrees of accentedness in the speech of Spanish-English bilinguals. *Language and Speech*, pages 267–273.

Sandahl, S. (2000). *A Hindi reference grammar*. Peeters, Leuven, Belgium.

Schirra, R. (2012). Attitudes toward Korean-accented and Korean American English. Master's thesis, University of Washington.

Scovel, T. (1981). The recognition of foreign accents in English and its implications for psycholinguistic theories of language acquisition. In Savard, J. G. and Laforge, L., editors, *Proceedings of the 5th Congress of AILA*, pages 389–401, Laval. University of Laval Press.

Scovel, T. (1995). Differentiation, recognition, and identification in discrimination of foreign accents. In Archibald, J., editor, *Phonological Acquisition and Phonological Theory*, pages 169–182. Lawrence Erlbaum, Hillsdale, NJ.

Shackle, C. (2003). Panjabi. In Cardona, G. and Jain, D., editors, *The Indo-Aryan languages*, pages 581–621. Routledge, London.

Shah, A. (2002). *Temporal characteristics of Spanish-accented English: Acoustic measures and their correlation with accentedness ratings*. Dissertation, City University of New York.

Shapiro, M. C. (2003). Hindi. In Cardona, G. and Jain, D., editors, *The Indo-Aryan languages*, pages 250–285. Routledge, London.

Shimizu, K. (2011). A study on VOT of initial stops in English produced by Korean, Thai and Chinese speakers as L2 learners. In *Proceedings of the 17th International Congress of Phonetic Sciences*, pages 1818–1821.

Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, pages 287–308.

Sohn, H.-M. (1999). *The Korean language*. Cambridge University Press, Cambridge.

Stockmal, V., Moates, D., and Bond, Z. S. (2000). Same talker, different language. *Applied Psycholinguistics*, pages 383–393.

Stockmal, V., Muljani, D., and Bond, Z. (1994). Can children identify samples of foreign languages as same or different? *Language Sciences*, pages 237–252.

Sun, C. (2006). *Chinese: A Linguistic Introduction*. Cambridge University Press.

Tomaszczyk, J. (1981). Some thoughts on accented speech: The English of Polish Americans. *Studia Anglica Posnaniensia*, pages 131–147.

Tsukada, K. (1998). Japanese-accented English vowels: A perception study. *Asia Pacific Journal of Speech, Language, and Hearing*, pages 43–65.

Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., and Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics*, pages 263–290.

Tsurutani, C. (2012). Evaluation of speakers with foreign-accented speech in Japan: The effect of accent produced by native English speakers. *Journal of Multilingual and Multicultural Development*, pages 589–603.

van Els, T. and de Bot, K. (1987). The role of intonation in foreign accent. *The Modern Language Journal*, pages 147–155.

Vasilescu, I., Candea, M., and Adda-Decker, M. (2005). Perceptual salience of language-specific acoustic differences in autonomous fillers across eight languages. In *Proceedings of INTERSPEECH 2005*, pages 1773–1776.

Vidyalankar, J. (2002). The treatment of English [t] and [d] in the Indian English. *Osmania Papers in Linguistics*, pages 1–4.

Vieru, B., Boula de Mareuil, P., and Adda-Decker, M. (2011). Characterisation and identification of non-native French accents. *Speech Communication*, pages 292–310.

Volín, J. and Skarnitzl, R. (2010). The strength of foreign accent in Czech English under adverse listening conditions. *Speech Communication*, pages 1010–1021.

Wang, H. and van Heuven, V. J. (2006). Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. *Linguistics in the Netherlands*, pages 237–248.

Watson, C. I. and Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, pages 458–468.

Wayland, R. (1997). Non-native production of Thai: Acoustic measurements and accentedness ratings. *Applied Linguistics*, pages 345–373.

Weinrich, H. (1986). Petite xénologie des langues étrangères ["Minor foreignness of foreign languages"]. *Communications*, pages 187–203.

Wilson, I. L. (2006). *Articulatory settings of French and English monolingual and bilingual speakers*. Dissertation, University of British Columbia.

Wiltshire, C. R. and Harnsberger, J. D. (2006). The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. *World Englishes*, pages 91–104.

Zampini, M. (1996). Voiced stop spirantization in the ESL speech of native speakers of Spanish. *Applied Psycholinguistics*, pages 335–354.

APPENDIX A: LANGUAGE BACKGROUND QUESTIONNAIRE

1. Date of birth Current age
2. Sex MALE FEMALE
3. Where have you lived, and what age were you at the time (starting with place of birth)?
4. Where did your parents or other caretakers grow up? What languages do/did they speak?
5. What language(s) did your parents or other caretakers speak to you at home?
6. What do you consider to be your native language(s)?
7. Including your native language(s), what languages do you know? At what age did you begin learning each language? How well can you write, read, speak, and understand each language?
8. For each language you mentioned in the question above, please estimate the percentage of your current language use that takes place in each language.
9. *(on questionnaires for non-native speakers only)*
   What were your English teachers' native languages?
10. *(on questionnaires for non-native speakers only)*
    Circle the dialect(s) of English that you studied in school.
    AMERICAN ENGLISH BRITISH ENGLISH INDIAN ENGLISH OTHER
    If you circled more than one option, or if you circled *other*, please describe these educational experiences in detail.
11. *(on questionnaires for non-native speakers only)*
    Please provide your best TOEFL exam score and circle its format. CBT PBT iBT
12. *(on questionnaires for non-native speakers only)*
    Please provide your best iBT TOEFL speaking section score.
13. Do you interact with native speakers of English from other parts of the world (India, Great Britain, Australia, New Zealand, etc.)? YES NO
    If you circled *yes*, please describe who, how often, and where the speakers are from.
14. Do you interact with non-native speakers of English? YES NO
    If you circled *yes*, please describe who, how often, and the speakers' native language(s).
15. Circle the highest level of education you have completed so far.
    PRIMARY JUNIOR HIGH HIGH SCHOOL COLLEGE POST-GRADUATE
16. What is your profession? (If academic or student, please indicate field of study.)
17. What is the highest level of education your parents or other caretakers completed?
18. What are/were the professions of your parents or other caretakers?
19. Do you have any speech, language, or hearing disorders? YES NO
    If you circled *yes*, please provide details.

APPENDIX B: ADDITIONAL TABLES AND FIGURES

Talker	Experiment 1	Experiment 2	Experiment 3	Experiment 4
E1f	-0.5429	-2.1236	-0.7048	-2.4722
E2f	-0.8306	-1.9082	-0.9573	-2.4424
E3f	-1.2136	-2.6504	-1.5086	-2.9442
E4m	-0.3243	-1.7441	-0.2688	-2.1148
E5m	-0.4337	-2.0199	-0.4407	-2.5987
E6m	-0.7214	-2.3100	-0.9849	-2.6694
H1f	0.5305	1.2629	0.5913	1.4826
H2f	0.0738	1.0154	0.4584	1.5273
H3f	-0.6519	-2.1910	-0.8206	-2.0346
H4m	0.4812	1.2228	0.6598	1.5136
H5m	1.0567	2.2160	1.6537	2.4647
H6m	0.7831	2.3412	1.2330	2.5257
K1f	0.0161	-0.9038	-0.1018	-0.8859
K2f	-0.1165	-0.3943	-0.1560	-0.0561
K3f	-0.4034	-0.2551	-0.4277	-0.1415
K4m	0.6999	1.0106	1.0475	1.0222
K5m	0.4209	1.7114	0.4020	1.5413
K6m	0.2850	1.4634	0.3952	1.5020
M1f	0.1673	0.3638	-0.0422	0.3879
M2f	-0.3557	-0.0298	-0.4231	0.1844
M3f	-0.2844	0.0059	-0.5409	0.2917
M4m	-0.4071	-0.4932	-0.5441	-0.6051
M5m	0.5625	0.0775	0.3215	0.5153
M6m	0.7771	1.1867	0.7449	1.1726
S1f	-0.6381	-0.1620	-0.8092	-0.3288
S2f	0.3374	1.0898	0.4546	1.2817
S3f	-0.1999	-0.4740	-0.4758	-0.4814
S4m	0.2241	0.9779	0.0343	0.7819
S5m	0.1364	0.2908	0.2745	0.1808
S6m	0.5715	1.4232	0.9358	1.3990

Table B.1: Random intercepts for talkers for Experiments 1 through 4

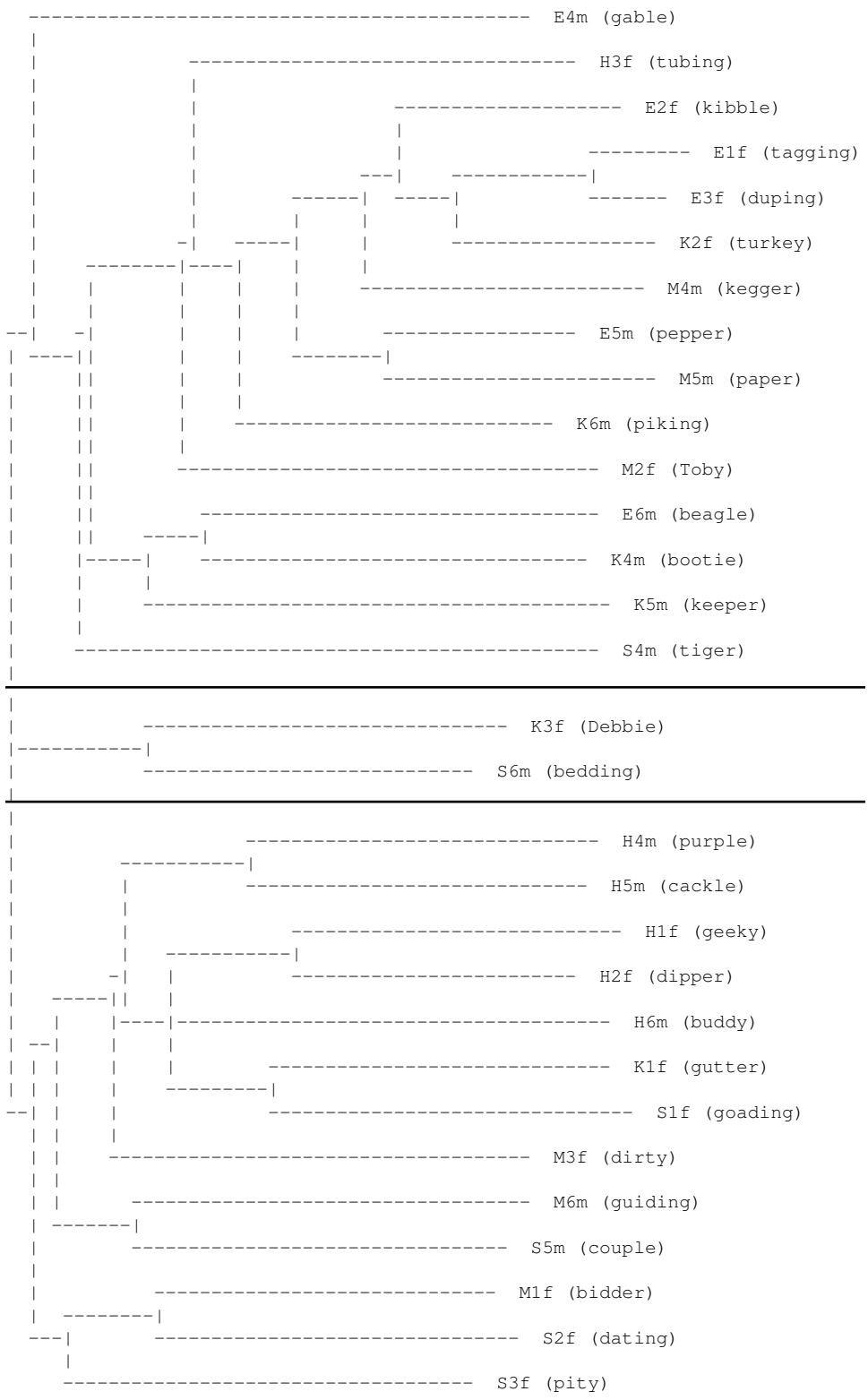

Figure B.1: Clustering solution for free classification in Experiment 1 (CVs)

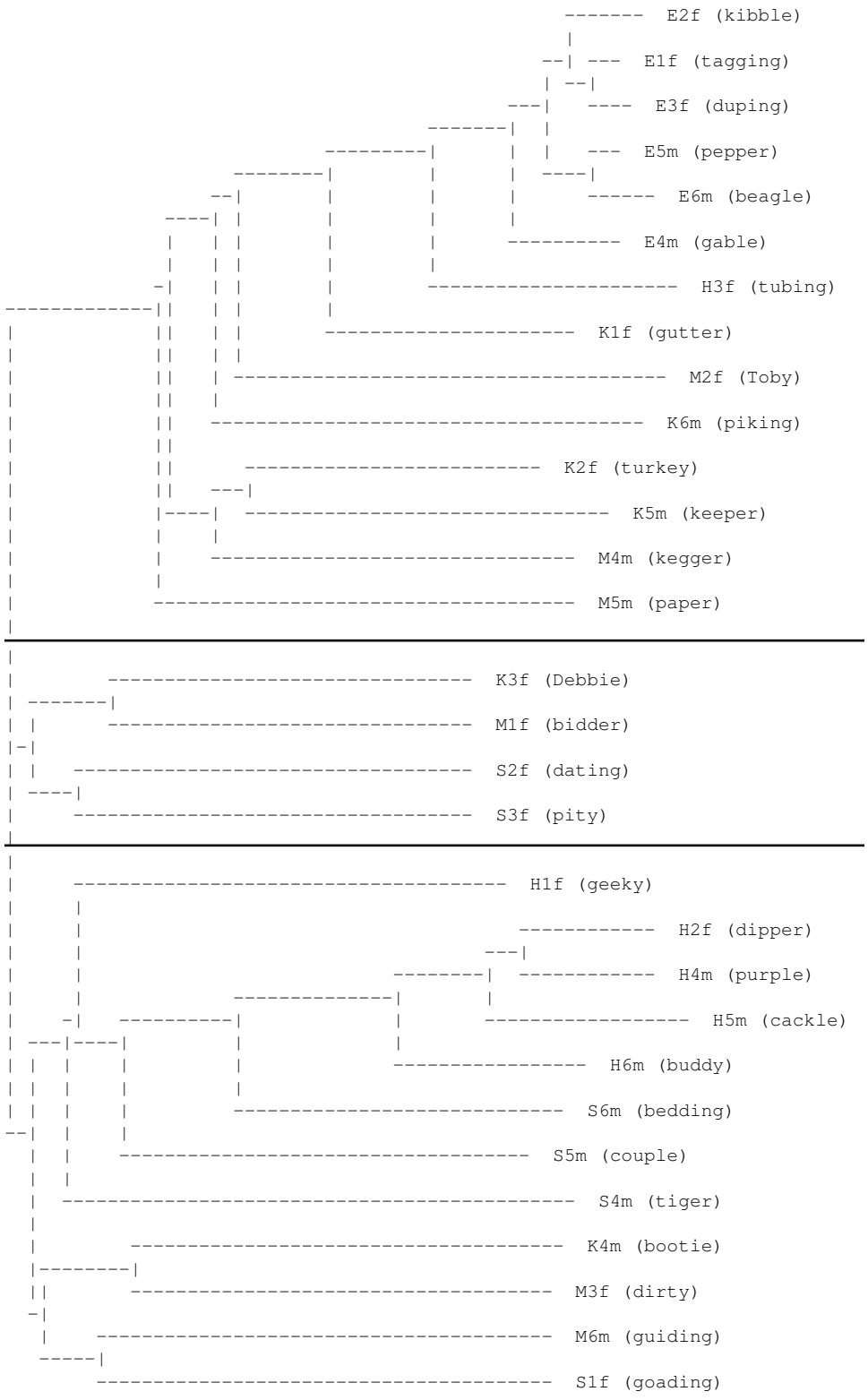

Figure B.2: Clustering solution for free classification in Experiment 2 (words)

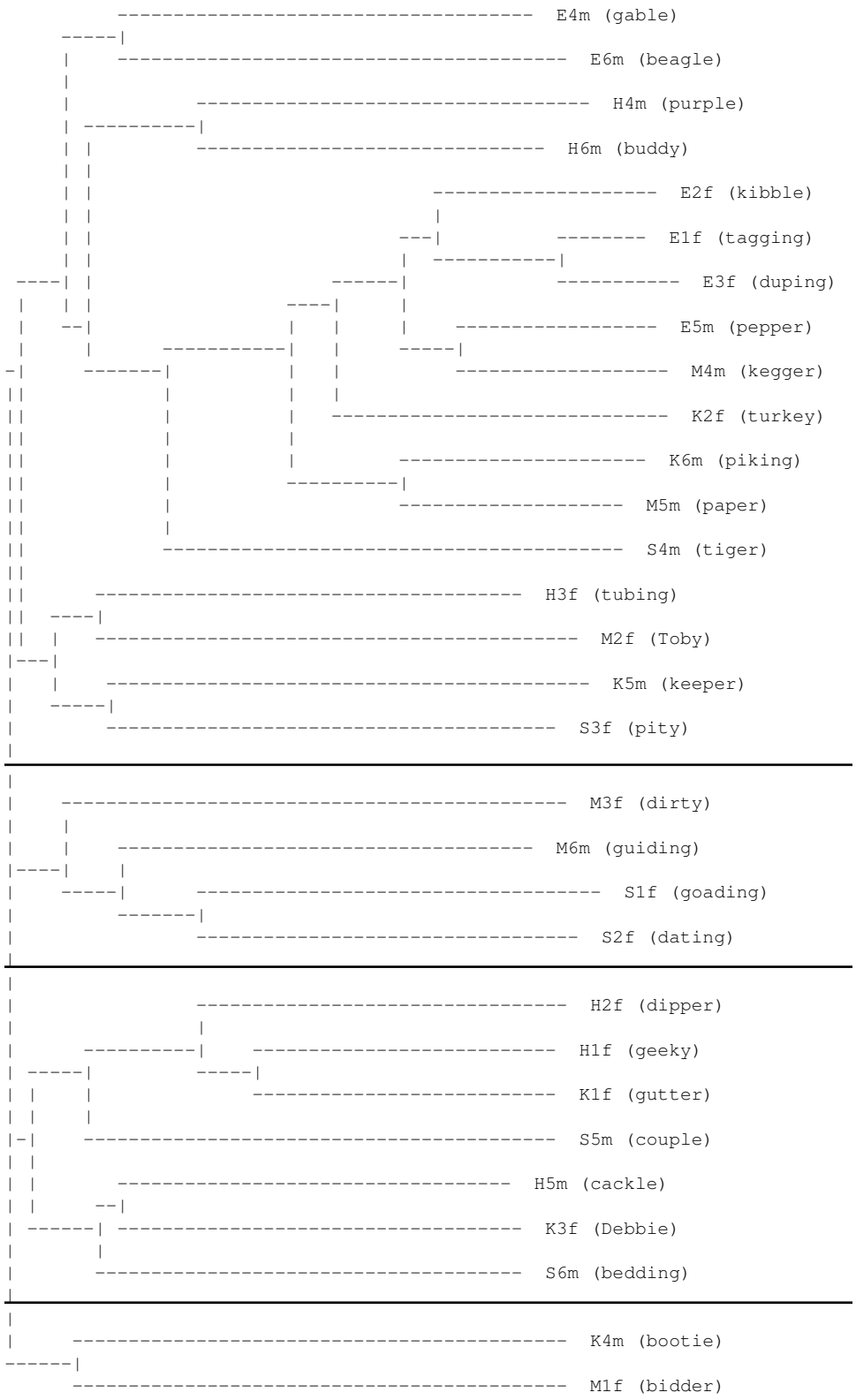

Figure B.3: Clustering solution for free classification in Experiment 3 (CVs)

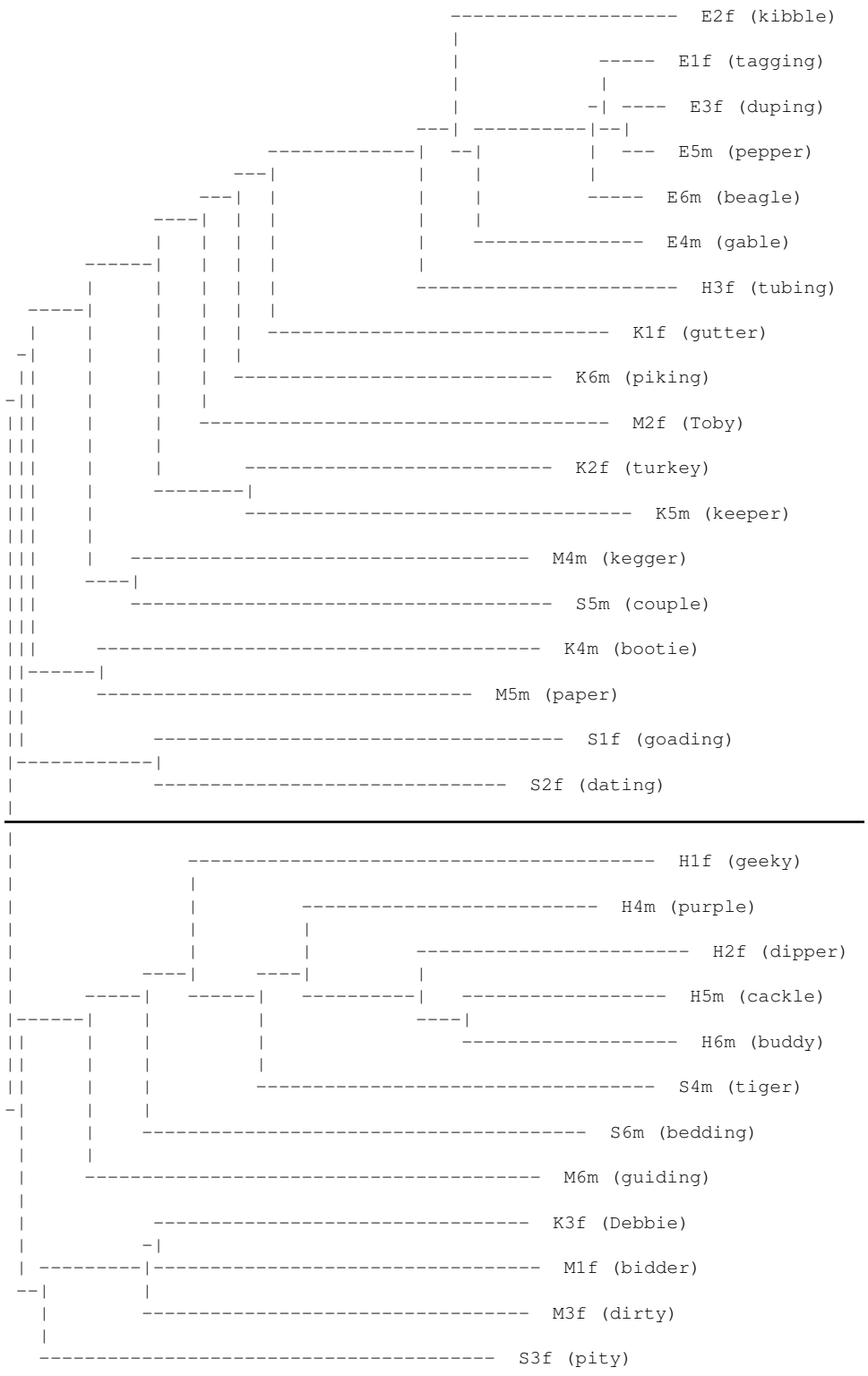

Figure B.4: Clustering solution for free classification in Experiment 4 (words)

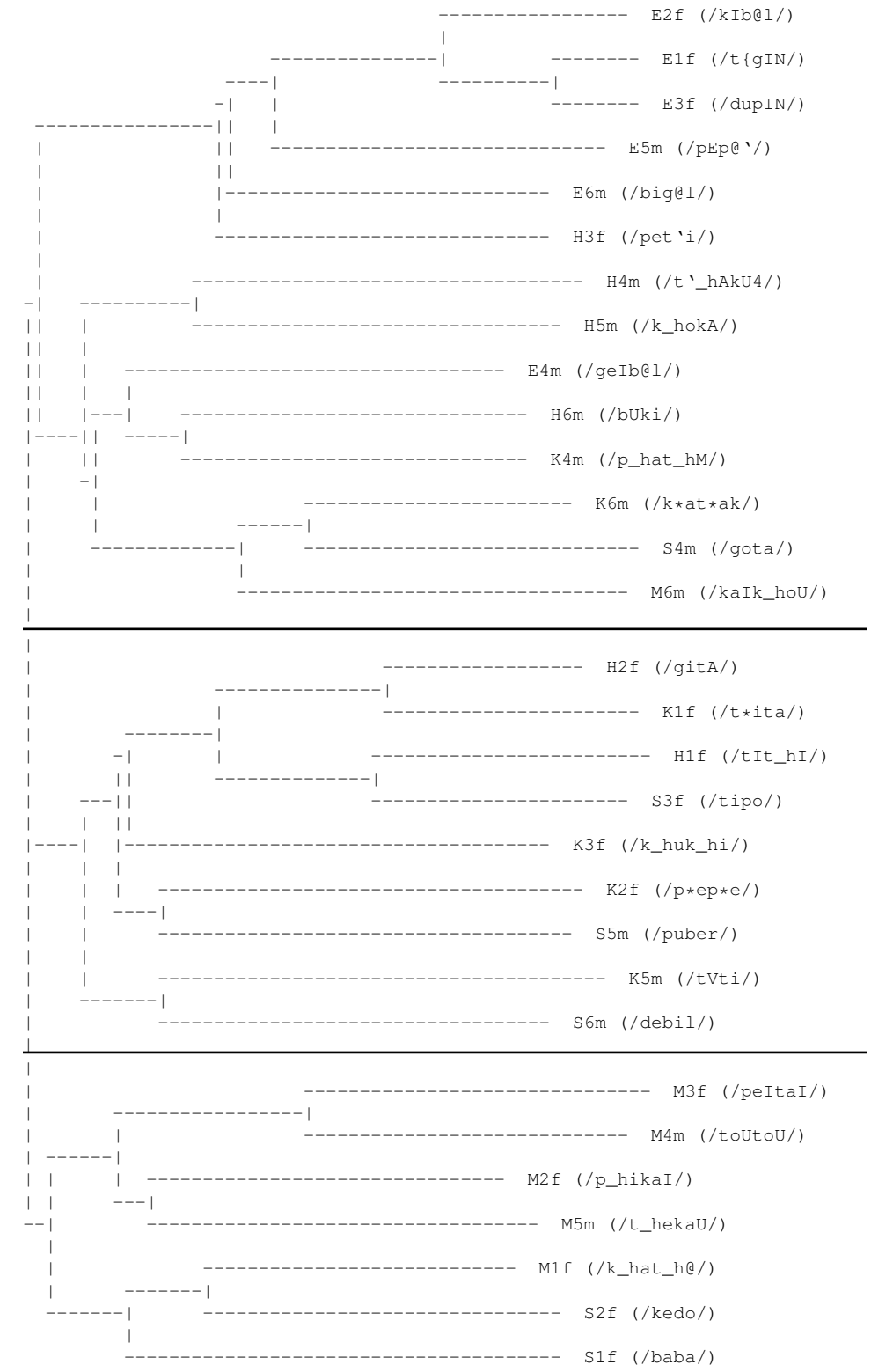

Figure B.5: Clustering solution for free classification in Experiment 5 (CVs)

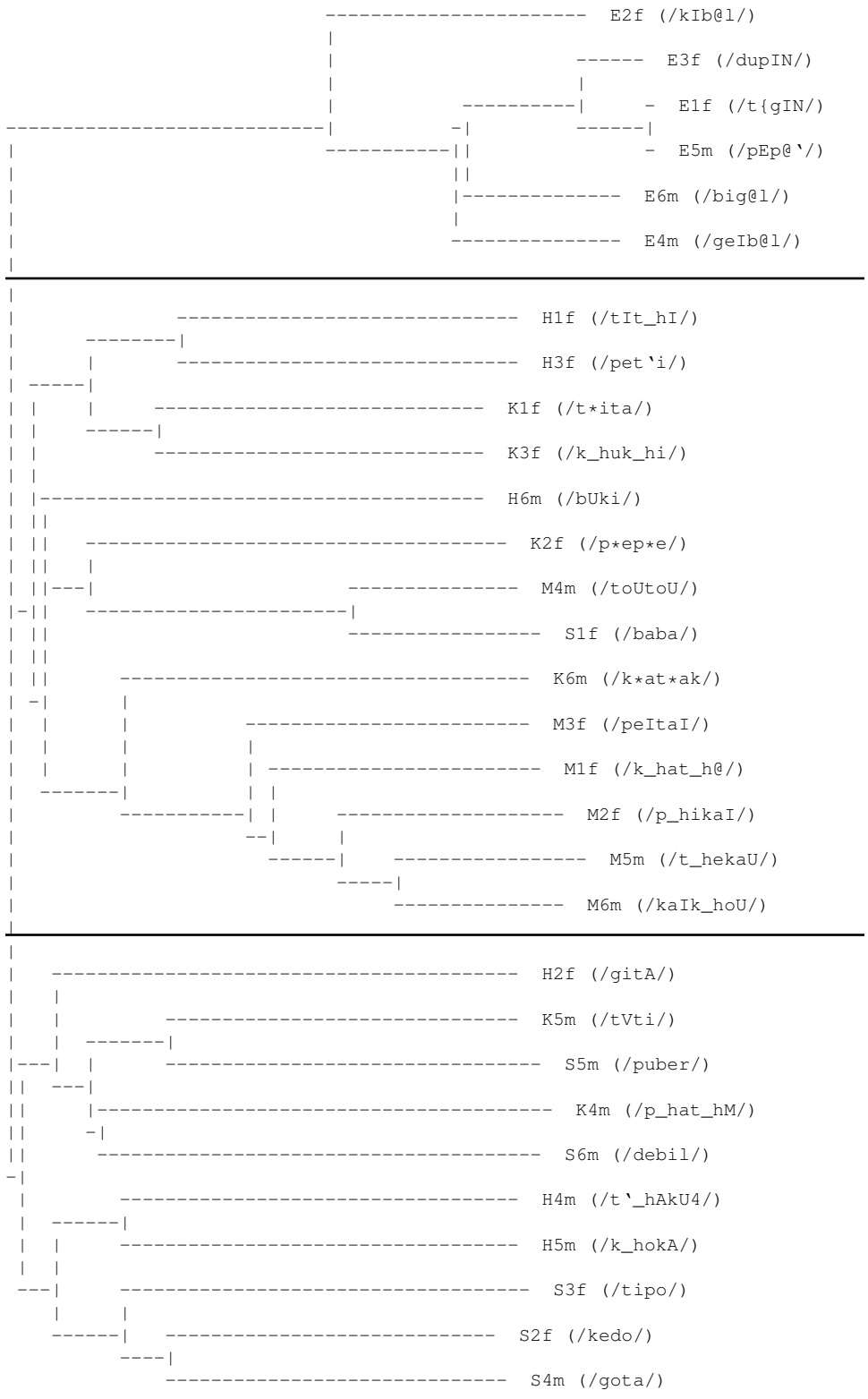

Figure B.6: Clustering solution for free classification in Experiment 6 (words)