



Research Article

Perceptual subcategories within non-native English

Elizabeth A. McCullough^{a,b,*}, Cynthia G. Clopper^b^a Department of Psychology, Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, USA^b Department of Linguistics, Ohio State University, 1961 Tuttle Park Place, Columbus, OH 43210, USA

ARTICLE INFO

Article history:

Received 3 October 2014

Received in revised form

9 November 2015

Accepted 17 November 2015

Available online 17 December 2015

Keywords:

Free classification

L2 English

Foreign accent

Speech perception

VOT

ABSTRACT

Listeners' ability to distinguish native from non-native speech has been robustly attested, but less is known about the nature of perceptual subcategories within non-native speech. This study examined American English-speaking listeners' abilities to perform auditory free classification by talker native language (L1) background on the basis of CV- and word-length excerpts of English productions by L1 American English, L1 Hindi, L1 Korean, L1 Mandarin, and L1 Spanish talkers. Results are examined in terms of classification accuracy and perceptual similarity, and the phonetic properties predicting listeners' classification responses are explored. Overall, L1 American English talkers were grouped together, as were L1 Hindi talkers. Talkers from L1 Korean and L1 Mandarin were perceptually similar to one another, although they did not comprise a single shared group. Deviation from native norms in VOT and in spectral and temporal properties of vowels predicted listeners' classification responses. These results suggest that listeners can use specific phonetic properties of the speech signal to group talkers on the basis of very little input, and that L1 Hindi talkers form a clear and consistent subcategory of non-native talkers for American English-speaking listeners.

© 2015 Elsevier Ltd. All rights reserved.

1. Introduction

When people speak, they communicate not only a linguistic message, but also information about themselves. Indexical information in the speech signal reveals personal characteristics including sex, age, and dialect membership (Abercrombie, 1967). Another attribute that may be exposed by indexical information is a talker's native speaker status. Sensitivity to the binary distinction between native and non-native speech has been evaluated by asking listeners whether a particular talker is a native or non-native speaker of the target language (Bond, Stockmal, & Markus, 2008; Flege, 1984; Park, 2013) or which of a pair of productions was said by a native speaker of the target language (Tsukada, 1998). Listeners are quite skilled at such judgments, and even short samples of speech, such as monosyllabic words (Park, 2013; Tsukada, 1998), segments (Flege, 1984), and 30-ms excerpts from stop consonants (Flege, 1984) convey indexical information relevant to this two-way division between native and non-native speech.

Whereas listeners' abilities to differentiate between the broad categories of native and non-native speakers are well-established in the literature, the perception of L1-based subcategories within the non-native speech category is less understood. That is, although listeners readily identify talkers as non-native speakers of a particular target language (Bond et al., 2008; Flege, 1984; Park, 2013; Tsukada, 1998), it is less clear whether they can distinguish, for example, Spanish-accented from Mandarin-accented varieties of the same language. Previous studies suggest that while listeners are to some extent aware of talker L1-based subcategories within non-native speech, certain L1 backgrounds may be difficult for them to tell apart. For instance, Derwing and Munro (1997) played English phrases for L1 Canadian English listeners and asked them to identify each accent as Cantonese, Japanese, Polish, or Spanish. On average, the listeners were 51.5% correct, but their responses revealed confusion between L1 Cantonese and L1 Japanese talkers, as well as between L1 Polish and L1 Spanish talkers. Similarly, Vieru, Boula de Mareuil, and Adda-Decker's (2011) L1 French listeners were 52% accurate in a forced-choice task in which they identified accents in 10-s excerpts of French speech as Arabic, English, German, Italian, Portuguese, or Spanish. Listeners' responses again exposed common confusions, in this case between L1 Italian and L1 Spanish talkers and between L1 English and L1 German talkers. In general, listeners can make somewhat accurate

* Corresponding author at: Department of Linguistics, University of Washington, Box 352425, Seattle, WA 98195-2425, USA. Tel.: +1 206 543 9027; fax: +1 206 685 7978.
E-mail address: emccullo@uw.edu (E.A. McCullough).

judgments about non-native talkers' L1 backgrounds, although they exhibit perceptual confusions among historically related and/or geographically close L1s.

Talker-based subcategories within non-native speech, to the extent that they are perceived, may considerably impact the process of communication. For instance, [Sumner, Kim, King, and McGowan \(2014\)](#) and [Kleinschmidt and Jaeger \(2015\)](#) proposed models of speech perception that involve simultaneous processing of linguistic information and social information about the talker. An open question about these and similar models pertains to the level of detail in the social information perceived, encoded, and used by the listener. Thus, in [Kleinschmidt and Jaeger's \(2015\)](#) model, the listener draws upon experience with “similar” talkers to interpret speech from an unfamiliar talker. However, whether all non-native talkers are “similar” to one another by virtue of being perceived as non-native, or whether listeners employ subcategories of non-native talkers (e.g., Spanish-accented, Mandarin-accented, etc.) to draw upon smaller, more relevant sets of prior experiences remains an open question.

Studies of adaptation to non-native speech provide indirect evidence about the role of non-native talker similarity in speech perception. [Bradlow and Bent \(2008\)](#) showed that exposure to English productions by multiple L1 Mandarin talkers helped listeners better understand a new L1 Mandarin talker, but not a new L1 Slovak talker. Further, [Baese-Berk, Bradlow, and Wright \(2013\)](#) showed that exposure to English productions by talkers from a variety of L1s (Hindi, Korean, Mandarin, Romanian, and Thai) helped listeners better understand both a new L1 Mandarin talker and a new L1 Slovak talker. The results obtained by [Baese-Berk et al. \(2013\)](#) suggest that the “non-native speech” category is important to listeners, in that experience with a variety of non-native talkers benefits the comprehension of other non-native talkers, regardless of whether a talker's specific L1 background has been encountered previously. However, the results described by [Bradlow and Bent \(2008\)](#) suggest that listeners may also rely on L1-based subcategories within this non-native category, because exposure to a single subcategory does not benefit the comprehension of talkers from other subcategories. Overall, listeners may evaluate similarity among talkers at either the more general non-native category level or at the more specific L1-based subcategory level, depending on the task presented and the talkers encountered.

Crucially for this interpretation of [Bradlow and Bent's \(2008\)](#) findings, for L1 Mandarin and L1 Slovak talkers to fall into different L1-based perceptual subcategories, listeners must be skilled at discriminating talkers' L1s. However, from the work by [Derwing and Munro \(1997\)](#) and [Vieru et al. \(2011\)](#), it is known that listeners misidentify talkers' L1s roughly half the time when choosing from a relatively short list of possible responses. Thus, perceptual subcategories within non-native speech require further investigation before their impact on speech processing can be fully understood. The overarching goal of the present work was to explore the nature of listeners' subcategories within non-native speech, including the perceptual similarity structure among talkers' L1 backgrounds and the phonetic properties that define the L1-based subcategories.

1.1. Objectives of the current study

The first objective of the current study was to explore the perceptual similarity structure of different L1 backgrounds within non-native speech. While forced-choice tasks like those conducted by [Derwing and Munro \(1997\)](#) and [Vieru et al. \(2011\)](#) represent one approach to addressing this objective, the labels offered by the experimenters in these tasks impose an inherent limitation on listeners' responses, which may not optimally reflect their perceptual subcategories. Listeners can only choose the correct label if they are at least somewhat familiar with the named L1 background, which may underestimate their perceptual skills; they may be capable of recognizing that two talkers belong to the same L1-based subcategory without being able to explicitly identify that background, or even without having heard such speech previously. Thus, with a forced-choice design, when performance for a particular L1 background is low, it is not clear whether the talkers did not meet listeners' expectations for the correct label, or whether the listeners were unfamiliar with speakers characterized by the correct label and thus unable to reliably assign the label to talkers. In addition, the presence of labels may induce listeners to attend primarily to more stereotyped properties of some accents, and ignore sources of phonetic variation which are less familiar or less salient.

To obtain perceptual responses about the subcategories of non-native speech without introducing these biases, a free classification task can be used. In free classification tasks, listeners create their own perceptual categories, which may be more or less fine-grained than the divisions known to the experimenter ([Clopper, 2008](#)). Auditory free classification tasks have been used to investigate listeners' perceptions of regional dialect variation in the United States ([Clopper & Bradlow, 2008, 2009](#); [Clopper & Pisoni, 2007](#)) and perceptions of similarity across different languages ([Bradlow, Clopper, Smiljanic, & Walter, 2010](#)). Auditory free classification was also used by [Atagi and Bent \(2013\)](#) to study perceptual subcategories within non-native English. In their study, L1 American English listeners were instructed to group talkers by general similarity or by perceived L1 background on the basis of sentences. With the latter set of instructions, listeners were nearly perfect at grouping together L1 American English talkers, and were generally least accurate for L1 French talkers; performance on L1 German, L1 Japanese, L1 Korean, L1 Mandarin, and L1 Spanish talkers was intermediate. These results thus provided strong evidence for the general non-native category, and weaker evidence for specific L1-based subcategories. Degree of talker accentedness was found to correlate with one dimension of a multi-dimensional scaling (MDS) model of the perceptual responses, suggesting that listeners were comparing the speech samples they heard against their expectations for native speech.

In the present work, as the subcategories themselves were the objects of study, a free classification task rather than a forced-choice task was used to provide the listeners with full control over the representation of the perceptual space. With a free classification task, perceived similarity among individual talkers and among L1 backgrounds can be explored by evaluating listeners' grouping responses. Although this approach is similar to the approach used by [Atagi and Bent \(2013\)](#), the present study differed substantially from theirs in two ways. First, the present study included non-native English productions from L1 Hindi talkers, an L1

background not tested by [Atagi and Bent \(2013\)](#). In the empirical investigation of listeners' perceptual subcategories within non-native speech, the inclusion of a new L1 background may extend previously attested perceptual subcategories or reveal new ones. Second, the stimulus materials in the present work were quite brief. When grouping by perceived native language, [Atagi and Bent's \(2013\)](#) listeners heard the same short sentence from all talkers. While sentence-length stimuli are common in free classification studies ([Clopper & Bradlow, 2008, 2009; Clopper & Pisoni, 2007](#)), [Flege \(1984\)](#) and [Park \(2013\)](#) showed that listeners can identify non-native talkers after hearing one syllable or less. Syllable- and word-length stimuli were therefore used in the present investigation of L1-based subcategories to allow for stronger comparisons to these previous investigations of native and non-native (sub)categories.

The second objective of the current study was to identify the phonetic properties that predict listeners' perceptual subcategories within non-native speech. This approach often relies upon fine-grained perceptual responses, which may be framed as ratings of the strength of the accent ([Major, 1987; Munro, 1993; Munro & Derwing, 2001; Riney & Takagi, 1999; Wayland, 1997](#)) or the degree of nativeness of the talker ([Baker et al., 2011; Bond et al., 2008](#)). The relationships between such ratings and acoustic properties or other production-based measures are then evaluated. In English, this approach has identified VOT ([Major, 1987; Riney & Takagi, 1999](#)), vowel formants ([Munro, 1993](#)), speaking rate ([Munro & Derwing, 2001](#)), and several measures based on word duration ([Baker et al., 2011](#)) as potential cues to non-nativeness. Many of these studies use monosyllabic words as stimuli and reveal correlations between listeners' responses and segment-specific acoustic properties ([Major, 1987; Munro, 1993; Riney & Takagi, 1999; Wayland, 1997](#)). Larger units of speech such as phrases, sentences, or paragraphs are also examined, although responses to such stimuli are generally correlated with more global acoustic properties such as speaking rate ([Munro & Derwing, 2001](#)).

By using syllable- and word-length stimuli, the current study can reveal the specific phonetic properties that define listeners' subcategories of non-native talkers, similar to previous investigations involving ratings of accentedness ([Major, 1987; Munro, 1993; Munro & Derwing, 2001; Riney & Takagi, 1999; Wayland, 1997](#)) and nativeness ([Baker et al., 2011; Bond et al., 2008](#)). In a departure from some free classification tasks, the talkers in the present work were not targeting the same sequences as one another. Due to the diverse stimuli, listeners were unable to directly compare productions of the same sequences across talkers, as many pairs of stimuli had no segments in common. This design was chosen because the question of interest was not, for instance, whether listeners could use particular VOT ranges for a given stop consonant to subcategorize non-native talkers by L1 background, but instead whether listeners could use judgments about the phonetic variation in stop consonants that is captured by VOT to accomplish this goal. Thus, by including multiple target sequences, listeners had to base their judgments on phonetic abstractions rather than direct acoustic comparisons. Variation in VOT, f_0 , spectral tilt, vowel duration, and F1 through F3 were measured as potential correlates of listeners' subcategory judgments.

1.2. Predictions for the current study

The four non-native L1 backgrounds in this work, Hindi, Korean, Mandarin, and Spanish, were chosen for their likely familiarity to listeners: students from India, South Korea, and China are commonly enrolled in United States universities ([Institute of International Education, 2014](#)) and Spanish is the most common non-English language spoken at home in the United States ([Ryan, 2013](#)). The listeners' open-ended self-reports of regular interactions with non-native English speakers supported this presumption: "Chinese," "Spanish," "Indian," and "Korean" were 4 of the 5 L1 backgrounds most commonly listed, with 33% (Chinese), 18% (Spanish), 14% (Indian), and 13% (Korean) of listeners mentioning each one. "Japanese" (15%) was also included among the top 5. An additional 10 non-native backgrounds were also listed, but by only 1–4% of listeners each.

Based on previous findings, it was expected that listeners would be able to separate native from non-native talkers ([Flege, 1984; Park, 2013](#)), and that listeners would be able to differentiate among at least some of the non-native varieties included in the current study ([Atagi & Bent, 2013; Derwing & Munro, 1997; Vieru et al., 2011](#)). However, the auditory stimuli chosen for the present investigation were substantially shorter than in many previous studies ([Atagi & Bent, 2013; Derwing & Munro, 1997; Vieru et al., 2011](#)), raising the question of whether they can even convey phonetic differences needed for listeners to perceive talker L1-based subcategories within non-native speech. Select pronunciation patterns in the four relevant varieties of non-native English are summarized below. It is presumed that listeners are relatively familiar with these patterns based on the prevalence of these particular L1 backgrounds in their prior experience, and could therefore use their knowledge of these patterns to successfully complete the free classification task.

One way in which the four selected non-native varieties of English differ from one another is in the production of stop consonants. Mandarin has a two-way stop contrast based on aspiration ([Liu, Tseng, & Tsao, 2000](#)), and L1 Mandarin speakers, like L1 American English speakers, tend to realize the English contrast as short lag versus long lag voicing ([Kim, 2011; Shimizu, 2011](#)). L1 Korean speakers also produce these VOT categories in English ([Kang & Guion, 2006](#)), although Korean itself has a three-way stop contrast that in recent decades has involved f_0 as well as VOT ([Kang & Guion, 2008; Silva, 2006](#)). Like English, Spanish has a two-way stop contrast, but one that is based on voicing rather than aspiration ([Lisker & Abramson, 1964](#)). In English, L1 Spanish speakers may have short or long lag VOT for voiceless stop targets, and often exhibit lead voicing for word-initial voiced stop targets ([Flege & Eefting, 1987; Nathan, 1987](#)). Finally, Hindi has a 4-way stop contrast at each of 4 places of articulation which relies on multiple acoustic cues ([Dutta, 2007](#)). In English, L1 Hindi speakers produce lead voicing for many voiced stop targets, and short lag voicing for voiceless stop targets ([Davis & Beckman, 1983](#)). Thus, stop production patterns in English are very different for L1 Mandarin and L1 Korean talkers on one hand, and L1 Spanish and L1 Hindi talkers on the other, with the former group realizing the stop voicing contrast more similarly to native American English talkers.

The vowel spaces of these languages also differ, which may impact talkers' English vowel productions in different ways. American English includes the monophthongs /i, ɪ, e, ε, æ, ʌ, ə, u, ʊ, o, ɔ, ɑ, ə-/ (Hillenbrand, Getty, Clark, & Wheeler, 1995; Ladefoged, 1999), though in many dialects /ɔ, ɑ/ have merged (Ash, 2003) and /e, o/ are diphthongal. Hindi, too, has a fairly large vowel inventory, with the oral monophthongs /i, ɪ, e, ə, u, ʊ, o, ɔ, ɑ/ and either /ɛ/ (Khan, Gupta, & Rizvi, 1994) or /æ/ (Bansal, 1981; Sandahl, 2000) depending on the dialect. With a comparable number of divisions in the vowel space and largely similar category labels, L1 Hindi talkers' productions of American English vowels may be reasonably similar to native productions. Korean includes the short vowels /i, e, ε, ʌ, u, o, ø, i, a/ (Lee, 1999), although /e, ε/ have merged for many speakers (Choo & O'Grady, 2003; Sohn, 1999). Mandarin has the vowels /i, ə, u, y, a/; sometimes /ɤ/ is included separately (Lee & Zee, 2003) and sometimes it is treated as an allophone of /ə/ (Chin, 2006). Spanish includes the vowels /i, e, u, o, a/. Mandarin and Spanish have many fewer categories in the vowel space than English, and these languages as well as Korean lack many vowels included in the American English inventory, perhaps leading to less accurate production of American English vowel targets by talkers with these L1s than by L1 Hindi talkers.

As even the shortest stimuli in the current study contain a stop and a vowel, there are multiple means by which listeners may be able to distinguish these L1 backgrounds, and multiple possible patterns of categorization. Given the predicted variation in English production across L1 backgrounds discussed above, the free classification results were expected to reveal which phonetic properties are most relevant to listeners' L1-based non-native subcategories. If listeners attend to phonetic differences in stop production, they should create a mixed group of L1 Hindi and L1 Spanish talkers and a separate mixed group of L1 American English, L1 Korean, and L1 Mandarin talkers. If listeners instead attend to phonetic characteristics of vowels, then talkers with Korean, Mandarin, and Spanish L1 backgrounds should be grouped separately from talkers with American English and Hindi L1 backgrounds. Of course, it is quite possible that listeners base their perceptual classification responses on multiple cues in the input, and that the classification results will thus reflect multiple phonetic motivations for listeners' judgments.

2. Methods

2.1. Listeners

Eighty monolingual native speakers of American English were recruited from Ohio State University linguistics courses to participate in this experiment. 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 40 additional listeners were excluded due to failure to meet this strict definition of "American English monolingual" (N=28), failure to follow instructions in any portion of the experiment (N=8), self-reported hearing loss (N=2), and technical problems with the experiment presentation software (N=2). Data from one additional listener were excluded because her voice was included among the auditory stimuli, although she did not appear to recognize her own voice. Listeners received partial course credit for their participation.

2.2. Stimulus materials

An English word list of 60 disyllabic trochees, each beginning with a unique combination of a stop /b, d, g, p, t, k/ and a vowel /i, ɪ, ε, æ, ʌ, u, ə, eɪ, ou, aɪ/, was constructed. The second consonant in each word was also a stop, and the rhyme of the second syllable was selected from the set /i, ə, ɪ, ɪŋ/. All items in the list were real English words. However, the words were not uniformly familiar, due to the limited availability of words with the desired structure for some initial stop-vowel combinations.

Each word was produced in isolation by 3 female and 3 male talkers from each of 5 L1 backgrounds: American English, Hindi, Korean, Mandarin, and Spanish. Detailed information about all 30 talkers is reported in Table A.1 in Appendix A. Although all talkers were recorded at a single location in the United States, the L1 American English, L1 Hindi, and L1 Mandarin talkers were originally from a variety of regions in the United States, India, and China, respectively. In contrast, most of the L1 Korean talkers (2 female and 2 male) were from the same city, Seoul, while the 2 remaining talkers were from other regions in South Korea. The L1 Spanish talkers were raised in numerous countries in Latin America. While regional varieties of Spanish exhibit some pronunciation differences, such differences are unlikely to impact pronunciations of the English targets represented in these stimuli, as stops and vowels are produced relatively consistently among the talkers' regions of origin (Canfield, 1981; Dalbor, 1969). The target words were recorded with a head-mounted microphone in a sound-attenuated booth at a sampling rate of 22.05 kHz.

Table 1
Words selected for use as stimuli in the free classification task, each produced by a single talker from the L1 group.

Group	Words
L1 American English	beagle, duping, gable, pepper, tagging, kibble
L1 Hindi	buddy, dipper, geeky, purple, tubing, cackle
L1 Korean	bootie, Debbie, gutter, piking, turkey, keeper
L1 Mandarin	bidder, dirty, guiding, paper, Toby, kegger
L1 Spanish	bedding, dating, goading, pity, tiger, couple

Thirty tokens were selected for use as stimulus materials in the free classification task, with each of the 30 talkers producing exactly one unique token. Within each L1 background, the 6 talkers produced all 6 initial stop consonants and a variety of vowels in the first syllable. The target words for these 30 tokens are shown in Table 1. Some listeners heard the full word, while other listeners heard only the consonant–vowel (CV) sequence extracted from the beginning of each word. The final 25 ms of each CV stimulus gradually decreased in intensity to reduce the audibility of any coarticulatory cues with the following consonant.

As the experimental procedure required participants to manipulate icons on the computer screen, the auditory stimuli were converted to digital movies with a static visual display of the talker's initials. However, transliterations of some talkers' names seemed likely to bias listeners against thinking that the 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 L1 Mandarin and L1 Korean speakers recorded, respectively. Thus, the initials displayed on the icons were not the talkers' actual initials, but the initials of 30 native speakers of American English who had been recorded for other purposes. The assignment of initials to talkers was pseudorandom to ensure that the talkers within each L1 background exhibited a variety of first initials, last initials, and locations on the computer screen when the icons were arranged in alphabetical order.

2.3. Procedure

In the free classification task, listeners saw a screen with the 30 talker icons arranged to the left of a 16×16 grid, as shown in the left panel of Fig. 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 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 16×16 grid, as shown in the right panel of Fig. 1. There were no restrictions on the number of groups or the number of talkers per group, and listeners were allowed to listen to each talker as many times as desired and to rearrange the icons until they were satisfied with their groups.

This free classification task was one of three tasks administered in a larger experiment, which also included a rating task and a brief filler task involving sentence completion. In the rating task, listeners heard 10 tokens from each of the 30 talkers in the free classification task (including the single token produced by the talker in the free classification task, as well as 9 other tokens that were similar in structure to the free classification token) and rated either the strength of the talker's foreign accent on a continuous scale from “no foreign accent” to “strong foreign accent” or the certainty that the talker was a non-native speaker of English on a continuous scale from “definitely a native speaker of English” to “definitely not a native speaker of English.” As the rating task provided experience with the talkers' speech, completing it prior to the free classification task might have allowed listeners to form clearer ideas about the talkers' L1 backgrounds. The free classification results are therefore presented as separate conditions, determined in part by whether the free classification task was completed first or last in the experimental session, and if last, by the scale used in the earlier rating task (i.e., strength of the talker's foreign accent or certainty that the talker was a non-native speaker of English), in case having focused on a particular characteristic in the rating task influenced free classification performance. Stimulus length (word or initial CV syllable) was also manipulated between conditions. The six conditions are summarized in Table 2.

2.4. Acoustic measurements

To investigate which aspects of the signal predict listeners' perceptual subcategories within non-native speech, various acoustic properties of the stimuli were measured. These properties were the VOT of the initial stop consonant, and six properties of the first or only vowel: duration, mean f_0 over the first 25 ms, spectral tilt (H_1 – H_2) over the first 25 ms, and F_1 , F_2 , and F_3 over the middle 60% of the duration. VOT was selected because instrumental and perceptual evaluations have shown that the stop consonants of L2 English speakers from the target L1 backgrounds differ from those of L1 English speakers (L1 Hindi: Davis & Beckman, 1983; L1

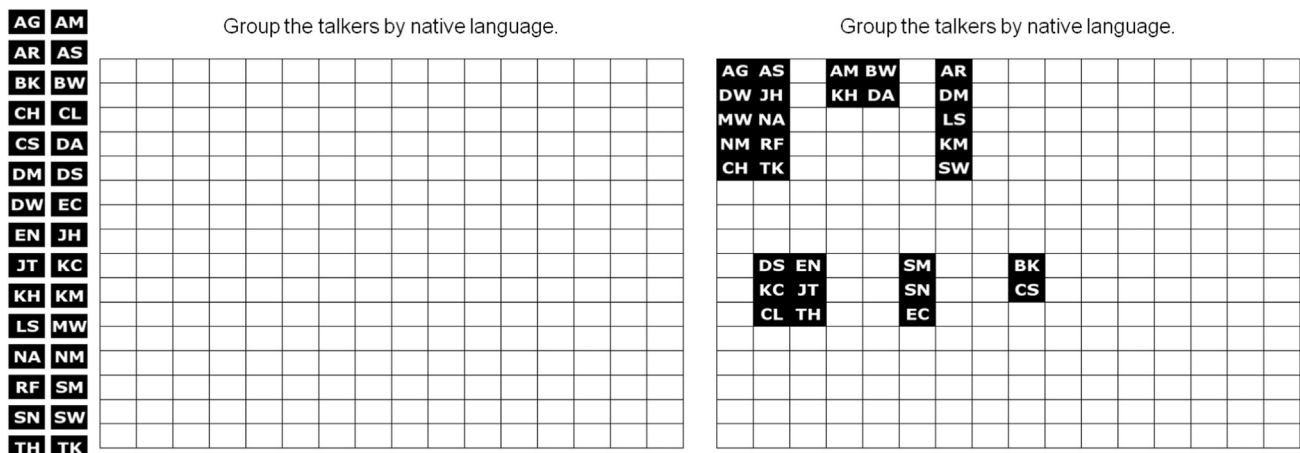


Fig. 1. Free classification screen before (left) and after (right) task completion.

Table 2
Overview of conditions in the free classification task.

Condition	Stimuli	Order	Earlier rating task	Number of listeners
CVsFirst	CVs	first	n/a	20
CVsPostFA	CVs	last	strength of foreign accent	10
CVsPostNN	CVs	last	certainty that talker is non-native	10
WordsFirst	words	first	n/a	20
WordsPostFA	words	last	strength of foreign accent	10
WordsPostNN	words	last	certainty that talker is non-native	10

Korean: Kang & Guion, 2006; Schirra, 2012; L1 Mandarin: Rogers & Dalby, 2005; L1 Spanish: Flege & Eefting, 1987; Nathan, 1987; Ortega-Llebaria, 1997). Similarly, vowel duration and formant measurements were included because L2 English vowel quality is often not native-like (L1 Hindi: Maxwell & Fletcher, 2009; L1 Korean: Flege, Bohn, & Jang, 1997; Schirra, 2012; Tsukada et al., 2005; L1 Mandarin: Chen, Robb, Gilbert, & Lerman, 2001; Flege et al., 1997; Rogers & Dalby, 2005; Wang & van Heuven, 2006; L1 Spanish: Donadio, 2002; Flege et al., 1997; Ortega-Llebaria, 1997). To thoroughly represent vowel quality, formant measures were parameterized as the first 3 coefficients of a discrete cosine transform, representing the mean value, tilt, and curvature of each formant (Watson & Harrington, 1999). Finally, spectral tilt and f_0 play important roles in the stop contrasts of Hindi (Dutta, 2007) and Korean (Cho, Jun, & Ladefoged, 2002; Han & Weitzman, 1970; Kang & Guion, 2008; Silva, 2006), which may influence L2 English productions of speakers from these backgrounds (Kang & Guion, 2006). Measures of spectral tilt and f_0^1 from the beginning of the vowel were therefore also included. All acoustic measurements were made from Praat TextGrids hand-aligned to the recordings by the first author, using periodicity of the waveform to identify the onset of voicing and the boundaries of the vowel, and a sharp increase in its amplitude to identify the stop burst.

As participants were generally unable to make direct acoustic comparisons among the stimuli because each talker produced a unique target, the participants may have attempted to evaluate the pronunciations more abstractly as sequences of known English sounds. Thus, the most appropriate phonetic parameterization may not be the raw acoustic values, but how much the stimuli diverged from listeners' expectations. This possibility is supported by Atagi and Bent's (2013) finding of the importance of perceived talker accentedness in a similar free classification task. The idea of divergence from expectations is captured in the current analysis by including difference values for each acoustic measure, following Munro (1993) and Wayland (1997). The listener's "expectation" was represented by the mean of the 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 *geeky* was calculated by subtracting from the raw measurement the mean VOT for the 3 L1 American English female talkers' productions of *geeky*. "Expectations" were based only on the productions of same-sex talkers because the values of spectral acoustic parameters generally differ between men and women, and spectral measurements were not normalized due to concerns about the appropriateness of any normalization method for talkers from a variety of L1 backgrounds (Disner, 1980). Smaller sex-based differences in temporal parameters, such as VOT and speaking rate, have also been attested (Robb, Gilbert, & Lerman, 2005; Swartz, 1992). Thus, same-sex comparisons were also used for VOT and vowel duration. Absolute differences from the L1 American English means rather than signed differences were used, capturing the degree of deviation but not the direction of deviation, as they yielded more interpretable results in preliminary analyses.

While listeners' actual expectations are presumably based upon their rich experiences with speech, the mean values of the L1 American English talkers' productions are used as approximations of American English speech norms. As evidence that these talkers are reasonable representatives of American English, Table A.2 in Appendix A shows the VOT means from this study alongside Robb et al.'s (2005) "laboratory setting" measurements and Lisker and Abramson's (1964) voicing lag measurements. The latter's voicing lead measurements are not included for comparison; voicing lead was rare among the L1 American English speakers in the present study, occurring in only 12 of their 180 voiced stop productions, and these values were thus omitted from the L1 American English means. For additional context, Table A.3 shows the mean formant measurements for the 7 vowels included both in this work and in the work by Peterson and Barney (1952).

2.5. Analysis of the free classification responses

In exploring the structure of listeners' perceptual subcategories of non-native speech, one approach is to examine to what extent listeners' free classification judgments align with talkers' actual L1 backgrounds. Following Clopper and Bradlow (2009) and Atagi and Bent (2013), three measures of overall accuracy were used to evaluate listeners' free classification performance. First, the correct talker pairing rate was calculated for each listener as the number of times the listener placed talkers from the same L1 background in the same group divided by the total possible number of same-L1 pairings; 75 same-L1 pairings were possible (5 same-L1 talkers $\times$ 30 talkers/2 as talker order was irrelevant). Next, the pairwise talker error rate was calculated for each listener as the

¹ While f_0 may differ among talkers for physical reasons unrelated to L1, the f_0 values in this work were comparable among same-sex talkers from different language backgrounds. A two-way ANOVA with talker sex and language background as between-subjects variables revealed a significant main effect of talker sex ($F(1,20)=135.88$, $p<0.001$), but neither a significant main effect of L1 nor a significant interaction between talker sex and L1.

number of times the listener placed talkers from different L1 backgrounds in the same group divided by the total possible number of different-L1 pairings; 360 different-L1 pairings were possible (24 different-L1 talkers $\times$ 30 talkers/2 as talker order was irrelevant). Correct talker pairings and pairwise talker errors are similar to “hits” and “false alarms,” respectively, in signal detection theory (Macmillan, 1993). As these calculations were not corrected for the number of groups created, and larger groups resulted in both more correct and more incorrect talker pairings, a difference score was calculated for each listener by subtracting the pairwise talker error rate from the correct talker pairing rate. This difference score resembles d' in signal detection theory. Additionally, correct talker pairing rates were calculated for each of the 5 L1 backgrounds, and pairwise talker error rates were calculated for each of the 10 pairs of different L1 backgrounds, to reveal the particular varieties with which listeners' subcategories aligned best.

Another important aspect of the structure of listeners' non-native subcategories is the perceptual similarity between L1 backgrounds. That is, when listeners' subcategory judgments do not align with talkers' actual backgrounds, it is important to know which pairs of L1 backgrounds are confused with one another. While the L1-specific pairwise talker error rates reflect these judgments generally, examining the particular talkers comprising each group offers a far more detailed representation of listeners' perceptions. To explore classification patterns at the individual talker level, a 30×30 talker similarity matrix was constructed for each listener on the basis of the groups they created in the free classification task. 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 listeners in a given condition, 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. A clustering analysis was performed with the dissimilarity matrices to reveal patterns of individual talker groupings by listeners. The GTREE program (Cortier, 1998) was used to create additive similarity trees, which represent perceptual distances between each pair of talkers as well as perceptual clusters of talkers. In these trees, perceptual distances are represented by the shortest horizontal distance needed to connect two items; vertical distances and ordering are irrelevant. As clustering solutions for 30 items are quite complex, the analysis below focuses only on talker membership in the highest-level clusters displayed in each tree to allow for the comparison of the major perceptual talker clusters across conditions.

2.6. Analysis of the phonetic correlates of free classification responses

To explore the phonetic information that predicted listeners' perceptual subcategories within non-native speech, a logistic regression analysis was performed for each of the high-level clusters identified as described in Section 2.5. For each cluster, the possible independent variables were the 13 acoustic absolute difference measures described in Section 2.4, and the dependent variable was membership in the given cluster (1 for members and 0 for non-members). In these analyses, the single acoustic property that best predicted membership in each cluster was identified². The independent variable was limited to a single acoustic property due to the relatively small number of observations in each model ($N=30$). The R function `glmulti` was used to generate all models from the set of possible predictor variables and select the best-fitting model.

3. Results

3.1. Classification strategy

In this section, listeners' free classification responses are examined to determine whether overall performance differed across conditions. Table 3 includes the number and size of the groups created by each of the 6 sets of listeners, as well as the three measures reflecting the accuracy of their performance: correct talker pairing rate, pairwise talker error rate, and difference score, as described in Section 2.5. Across the different conditions, listeners made 5–7 talker groups with approximately 5–6 talkers per group, and ranged widely in correct talker pairing rates and difference scores. In contrast, pairwise talker error rates were fairly consistent, ranging from 15% to 23%.

One reason to compare overall performance across conditions is to determine whether certain conditions may be combined. As described in Section 2.3, some conditions differed only in an aspect external to the free classification task: the nature of the scale used in the preceding rating task. As the free classification task itself was identical for these groups of listeners, their results might have been comparable. Indeed, for the listeners who heard CV-length stimuli, t -tests revealed no significant difference between the PostFA (post foreign accentedness rating scale) and PostNN (post non-nativeness rating scale) conditions on number of talker groups, number of talkers per group, or any of the three accuracy measures (all $p>0.05$). Similarly, for the listeners who heard word-length stimuli, performance on these five variables did not differ between the PostFA and PostNN conditions (all $p>0.05$). These results suggest that the scale used in the rating task (i.e., foreign accentedness vs. non-nativeness) had no effect on listeners' overall performance in the free classification task. Thus, in further analyses, the CVsPostFA and CVsPostNN conditions were combined into a condition called CVsLast, and the WordsPostFA and WordsPostNN conditions were combined into a condition called WordsLast.

² Interactions with talker sex were also examined because the acoustic variables were not explicitly normalized for talker sex, as described in Section 2.4, although the use of same-sex reference groups in the calculation of difference values somewhat attenuated this disparity. In no instance was a two-way interaction with talker sex a better predictor than a single acoustic property.

Although comparisons between CVsFirst and CVsLast and between WordsFirst and WordsLast revealed no differences on four of the five variables summarized in Table 3, *t*-tests showed a significant effect of task order on difference scores, with greater difference scores when free classification was last (CVs: $t(38) = -2.30$, $p < 0.05$; words: $t(38) = -2.19$, $p < 0.05$). This superior performance likely resulted from the participants' additional exposure to the talkers prior to the free classification task. Comparisons between CVsFirst and WordsFirst and between CVsLast and WordsLast exposed no differences on three of the five variables, but *t*-tests revealed a significant effect of stimulus length on correct talker pairing rates (first: $t(38) = -4.20$, $p < 0.001$; last: $t(38) = -3.79$, $p < 0.001$) and on difference scores (first: $t(38) = -4.86$, $p < 0.001$; last: $t(38) = -6.71$, $p < 0.001$). In both cases, performance was better with the longer, word-length stimuli. Given the differences found in listeners' responses, CVsFirst, CVsLast, WordsFirst, and WordsLast were all maintained as separate conditions.

Different task orders and stimulus lengths resulted from the inclusion of the free classification task in a larger experiment. Another reason to compare performance across conditions is to evaluate the general impact of each of these manipulations on listeners' free classification responses. Although the *t*-tests reported above suggest what the impacts are likely to be, simultaneous testing of the effects of task order and stimulus length offers a more thorough understanding of these manipulations. A series of two-way ANOVAs investigating the effects of task order, stimulus length, and their interaction on each of the five variables was conducted. No significant effects were found for number of talker groups, number of talkers per group, or pairwise talker error rates. However, the main effects of task order and stimulus length were significant for correct talker pairing rates (task order: $F(1,76) = 6.19$, $p < 0.05$; stimulus length: $F(1,76) = 31.91$, $p < 0.001$) and for the difference scores (task order: $F(1,76) = 9.69$, $p < 0.01$; stimulus length: $F(1,76) = 64.00$, $p < 0.001$). Performance on both variables was better when free classification occurred last rather than first, and for word-length than CV-length stimuli. That is, listeners were more accurate at classifying talkers by native language when they had prior experience with the talkers and when they heard longer utterances. These findings are not surprising, given the preliminary analyses above as well as the

Table 3

Summary of free classification strategies by listeners in different conditions, with standard deviations in parentheses.

	Condition					
	CVs			Words		
	First	PostFA	PostNN	First	PostFA	PostNN
Mean number of talker groups produced	7.3 (3.8)	5.1 (1.9)	7.4 (4.3)	6.2 (2.5)	7.2 (2.9)	6.9 (3.1)
Mean number of talkers per group	5.1 (2.4)	6.4 (1.5)	5.1 (2.1)	5.4 (1.6)	4.8 (1.8)	5.0 (1.8)
Percent correct talker pairings	20 (13)	32 (13)	22 (9)	35 (9)	39 (10)	42 (11)
Percent pairwise talker errors	18 (11)	23 (10)	16 (7)	19 (8)	15 (6)	21 (13)
Difference score (% correct–% errors)	2 (7)	9 (7)	6 (4)	16 (10)	24 (8)	21 (8)

Table 4

Percentages of correct talker pairings (on the diagonals) and pairwise talker errors (below the diagonals) in each condition.

	English	Hindi	Korean	Mandarin	Spanish
Condition: CVsFirst ($N=20$)					
English	27				
Hindi	14	24			
Korean	19	18	15		
Mandarin	17	17	19	15	
Spanish	14	19	19	19	18
Condition: CVsLast ($N=20$)					
English	51				
Hindi	17	28			
Korean	26	18	16		
Mandarin	29	19	21	19	
Spanish	14	16	20	18	22
Condition: WordsFirst ($N=20$)					
English	77				
Hindi	15	29			
Korean	29	13	24		
Mandarin	23	11	26	21	
Spanish	14	18	20	19	23
Condition: WordsLast ($N=20$)					
English	90				
Hindi	14	42			
Korean	33	11	24		
Mandarin	20	12	26	23	
Spanish	7	20	18	19	25

general advantage afforded by including more input in a speech perception task. The interaction between task order and stimulus length was not significant for correct talker pairing rates or difference scores.

3.2. Classification performance

The accuracy measures in Table 3 show listeners' overall successes, but do not reveal any L1-specific detail. To understand the structure of listeners' perceptual subcategories within non-native speech, as is the first objective of this study, more detailed information about their responses is needed. Table 4 breaks down the rates of correct talker pairings and pairwise talker errors by L1 combination. For each condition, the cells along the diagonal are the correct talker pairings; if listeners performed perfectly, and placed all talkers within each L1 in the same group, all of the cells along the diagonal would contain values of 100%. The cells below the diagonal are the incorrect talker pairings; if listeners performed perfectly, and did not place any talkers from different L1s in the same group, all of the cells below the diagonal would contain values of 0%. Each percentage represents the number of pairings that were actually made divided by the total possible number of pairings of talkers from the L1 background(s) in question. For each correct talker pairing cell, there were 15 possible pairings (of 6 talkers); the value in the cell is the percentage of these 15 pairings that were actually made, averaged across listeners. For each pairwise talker error cell, there were 36 possible pairings (of 6 talkers from one background with 6 talkers of another); the value in the cell is the percentage of these 36 pairings that were actually made, averaged across listeners. In the following sections, the results in Table 4 are discussed separately for each condition.

As described in Section 2.5, similarity between L1 backgrounds can be further explored by using a clustering analysis to show free classification patterns at the individual talker level. The clustering solutions for the CV and word conditions are shown in Figs. 2 and 3, respectively. The code representing each stimulus indicates the talker's L1 (E, H, K, M, or S), a unique talker identification number (1–6), and the talker's sex (f or m). The target word follows the talker code in parentheses. The high-level clusters are separated in the figures by horizontal lines. Each high-level cluster includes targets with a variety of consonants and vowels, confirming that listeners

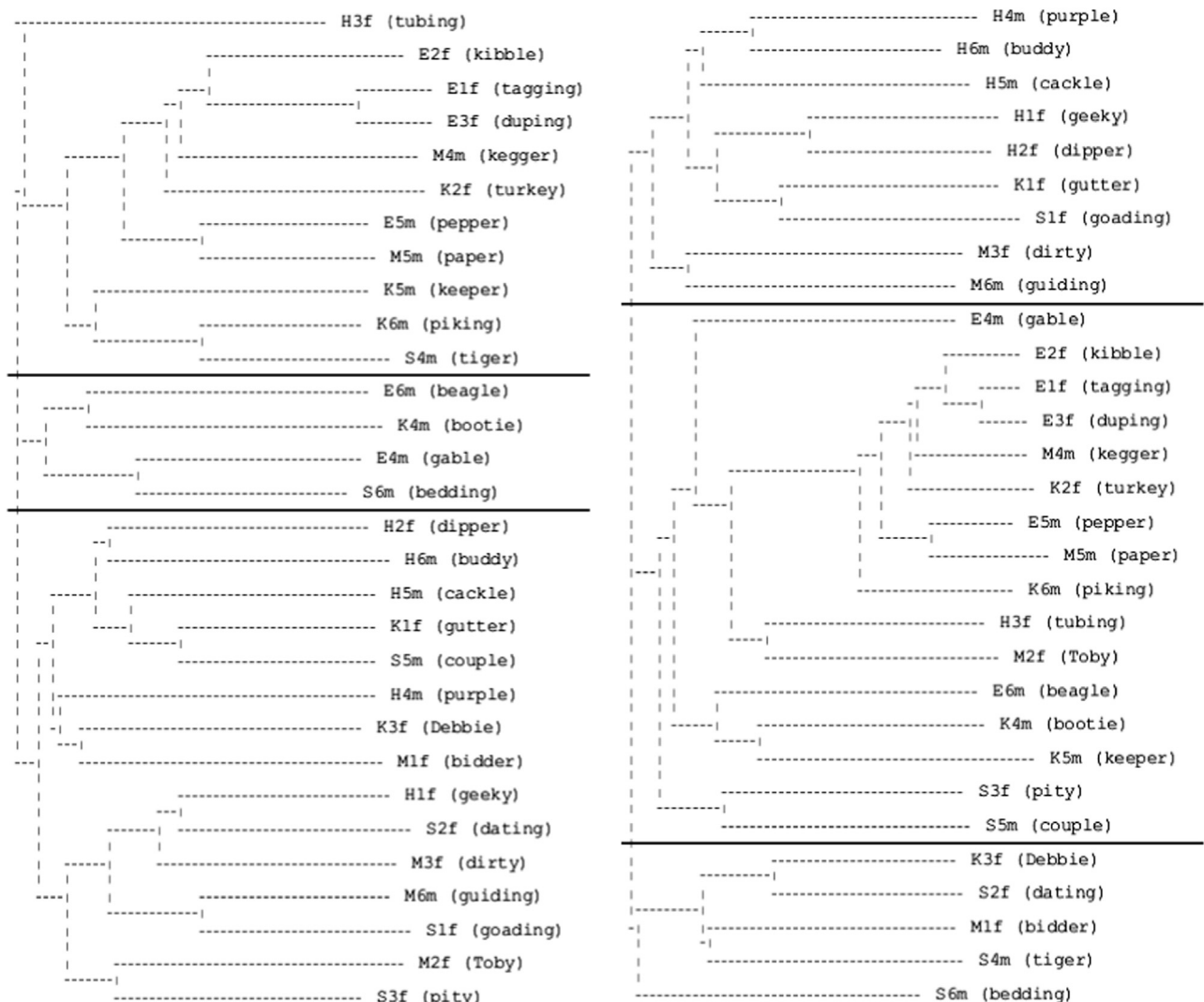


Fig. 2. Additive clustering solutions for the CVsFirst condition (left) and the CVsLast condition (right).

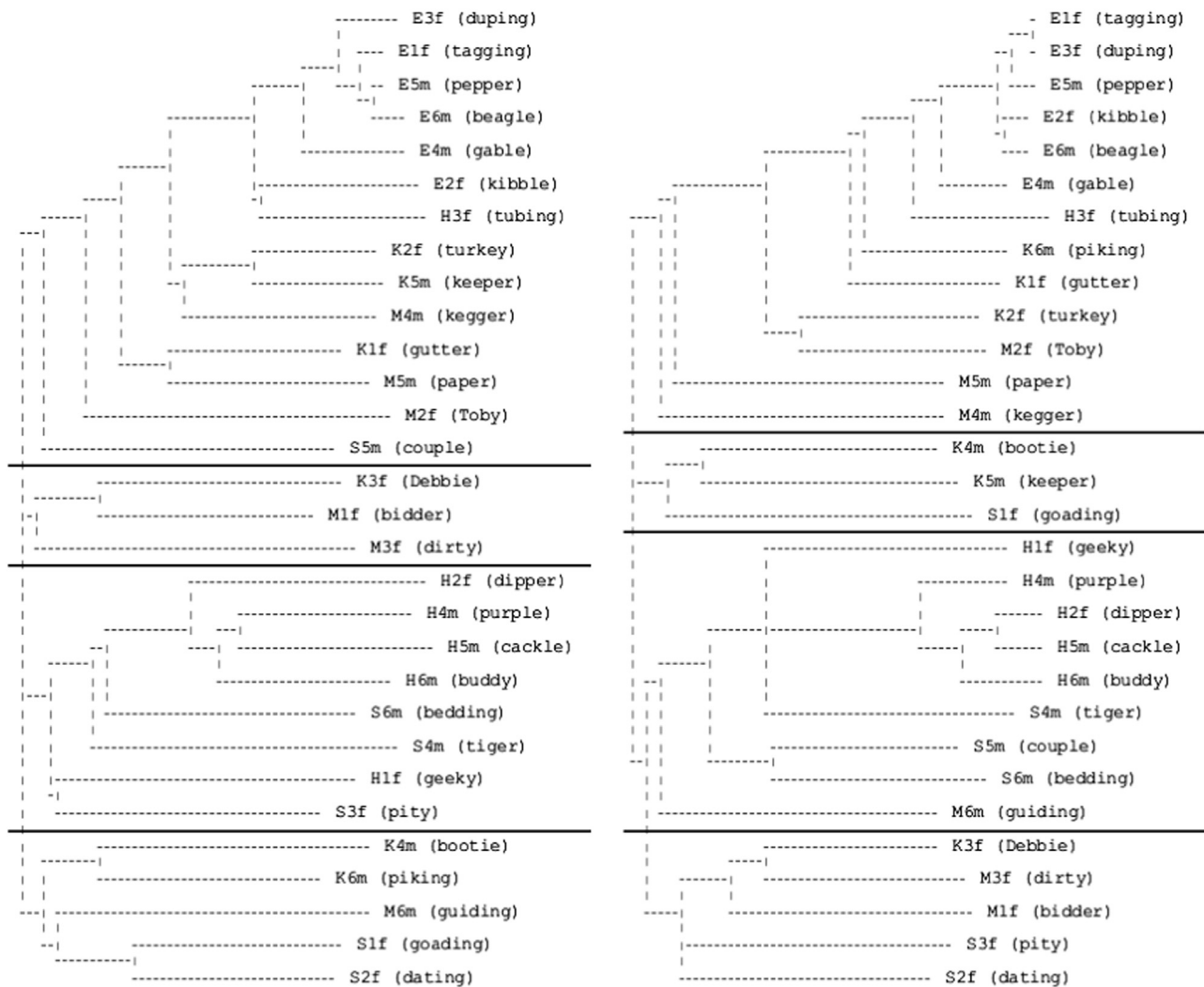


Fig. 3. Additive clustering solutions for the WordsFirst condition (left) and the WordsLast condition (right).

did not create groups based on linguistic content, but were instead abstracting over the specific consonants and vowels heard. The following sections also include discussion of the clustering solution for each condition.

3.2.1. CVsFirst

Table 4 shows that in the CVsFirst condition, correct talker pairing rates were lowest for L1 Korean talkers (15%) and L1 Mandarin talkers (15%), and highest for L1 American English talkers (27%) and L1 Hindi talkers (24%), although even these latter values reflect far from ideal performance. Pairwise talker error rates were generally similar across L1 combinations, with listeners least often grouping together L1 American English and L1 Hindi talkers (14%), and L1 American English and L1 Spanish talkers (14%). Although this condition seemed generally challenging—indeed, the overall difference score reported in Table 3 was only 2%—listeners were somewhat able to group L1 Hindi talkers with one another, and to place L1 American English talkers together in a group that did not include talkers from L1 Hindi or L1 Spanish backgrounds.

The clustering solution for the CVsFirst condition, which accounted for only 57% of the variance in responses, is shown in the left panel of Fig. 2. There were 3 main clusters. The first cluster included 4 L1 American English talkers (E1f, E2f, E3f, E5m), 3 L1 Korean talkers (K2f, K5m, K6m), 2 L1 Mandarin talkers (M4m, M5m), and talkers H3f and S4m. The second cluster was small, with only talkers E4m, E6m, K4m, and S6m. All remaining talkers were in the third cluster: 5 L1 Hindi (H1f, H2f, H4m, H5m, H6m), 4 L1 Mandarin (M1f, M2f, M3f, M6m), 4 L1 Spanish (S1f, S2f, S3f, S5m), and 2 L1 Korean (K1f, K3f). As suggested by the correct talker pairing and pairwise talker error rates reported in Table 4, L1 Hindi talkers were mostly clustered together, and many of the L1 American English talkers appeared in a cluster separate from the L1 Hindi and L1 Spanish talkers (except H3f and S4m). The low correct talker pairing rate for L1 Korean talkers in this condition is reflected in the presence of L1 Korean talkers in all 3 major clusters. Talkers K1f, K2f, K4m, and K6m, all from Seoul, made Korean the most regionally consistent L1 background. Nonetheless, of these 4 Seoul talkers, only K2f and K6m appeared in a cluster together. As correct talker pairing rates in Table 4 were also low for

L1 Mandarin talkers and L1 Spanish talkers, it is somewhat surprising that the third cluster included 4 talkers from each of these backgrounds. However, the diversity of these clusters, especially the third cluster, highlights listeners' general difficulty in identifying coherent non-native subcategories in this condition, which involved CV-length stimuli and no prior exposure to the talkers.

3.2.2. CVsLast

As shown in Table 4, listeners in the CVsLast condition also had correct talker pairing rates that were highest for L1 American English talkers (51%) and L1 Hindi talkers (28%), and lowest for L1 Korean talkers (16%) and L1 Mandarin talkers (19%). Based on pairwise talker error rates, L1 American English talkers generally were not perceived to share a native language with L1 Hindi (17%) or L1 Spanish talkers (14%), nor were L1 Hindi and L1 Spanish talkers themselves perceived to have a common L1 background (16%). On the other hand, L1 American English talkers were often incorrectly grouped with L1 Korean (26%) and L1 Mandarin talkers (29%).

The right panel of Fig. 2 shows the clustering solution for the CVsLast condition, which accounted for 80% of the variance in listeners' responses. Similar to the CVsFirst condition, the solution for CVsLast had 3 main clusters. The first cluster was dominated by L1 Hindi talkers (H1f, H2f, H4m, H5m, H6m), but also included 2 L1 Mandarin talkers (M3f, M6m), K1f, and S1f. All 6 L1 American English talkers were in the second cluster, along with a variety of other talkers: 4 L1 Korean (K2f, K4m, K5m, K6m), 3 L1 Mandarin (M2f, M4m, M5m), 2 L1 Spanish (S3f, S5m), and H3f. The third cluster was relatively small, and included 3 L1 Spanish talkers (S2f, S4m, S6m), K3f, and M1f. Consistent with the correct talker pairing and pairwise talker error rates shown in Table 4, most of the L1 Hindi talkers appeared in the same cluster, which was largely without L1 Spanish talkers (except S1f). L1 American English talkers were clustered together, separately from most of the L1 Hindi and L1 Spanish talkers (except H3f, S3f, and S5m), but with 4 L1 Korean and 3 L1 Mandarin talkers. L1 Korean, L1 Mandarin, and L1 Spanish talkers each appeared in all 3 clusters, resulting in the low correct talker pairing rates for these language groups. Three of the L1 Korean talkers from Seoul (K2f, K4m, K6m) were included in the same cluster, but the fourth (K1f) appeared in a cluster with no other L1 Korean talkers. On the whole, the clustering solution for CVsLast confirmed the general patterns expected based on Table 4. All 6 L1 American English talkers were clustered together in the CVsLast condition, and there was a clear non-native subcategory of L1 Hindi talkers.

3.2.3. WordsFirst

Table 4 shows that in the WordsFirst condition, the correct talker pairing rate for L1 American English talkers was fairly high (77%), although performance for talkers from other backgrounds was much less accurate, ranging from 21% (L1 Mandarin) to 29% (L1 Hindi). Pairwise talker error rates indicate that listeners did not often group L1 Hindi with L1 American English (15%), L1 Korean (13%), or L1 Mandarin (11%) talkers, nor L1 American English with L1 Spanish (14%) talkers. However, L1 Korean talkers were sometimes perceived as sharing a native language with L1 American English (29%) and L1 Mandarin (26%) talkers.

The clustering solution for the WordsFirst condition, which is shown in the left panel of Fig. 3, accounted for 87% of the variance in listeners' responses. In this case, the solution had 4 main clusters. The first cluster included all 6 L1 American English talkers, as well as 3 L1 Korean talkers (K1f, K2f, K5m), 3 L1 Mandarin talkers (M2f, M4m, M5m), H3f, and S5m. Only talkers K3f, M1f, and M3f were in the second cluster. The third cluster was made up of 5 L1 Hindi talkers (H1f, H2f, H4m, H5m, H6m) and 3 L1 Spanish talkers (S3f, S4m, S6m), while the last cluster included K4m, K6m, M6m, S1f, and S2f. Based on the correct talker pairing rates in Table 4, it was expected that L1 American English talkers would cluster together, while talkers from each of the other backgrounds would not. This pattern was generally observed, although the presence of 5 L1 Hindi talkers in the same cluster was surprising, given that the correct talker pairing rate for the L1 Hindi talkers was only 29%. As suggested by the pairwise talker error rates in Table 4, the L1 American English cluster included several L1 Korean talkers, but almost no L1 Hindi or L1 Spanish talkers. The exceptions were talkers H3f and S5m, who were also clustered with the L1 American English talkers in the CVsLast condition. Showing further consistency with the pairwise talker error rates detailed above, the L1 Hindi cluster contained no L1 American English, L1 Korean, or L1 Mandarin talkers, and L1 Korean and L1 Mandarin talkers often appeared in the same clusters. Of the 4 L1 Korean talkers from Seoul, K1f and K2f were together in one cluster, while K4m and K6m were together in a separate cluster. In the WordsFirst condition, as in the CVsLast condition, listeners perceived a category of L1 American English talkers and a non-native subcategory of L1 Hindi talkers. While L1 Korean and L1 Mandarin talkers were scattered among clusters, thus not forming a cohesive shared subcategory, there was substantial perceptual similarity between these L1 backgrounds.

3.2.4. WordsLast

Accuracy patterns for the WordsLast condition were similar to those for the WordsFirst condition, as shown in Table 4. L1 American English talkers were frequently grouped together (90%). While correct talker pairing rates were relatively low for L1 Korean (24%), L1 Mandarin (23%), and L1 Spanish (25%) talkers, performance for L1 Hindi talkers was somewhat more accurate (42%). Again, L1 Hindi talkers were not placed into groups with L1 American English (14%), L1 Korean (11%), or L1 Mandarin (12%) talkers, and L1 American English talkers were not often classified with L1 Spanish talkers (7%), but groups with L1 Korean talkers also included talkers from L1 American English (33%) and L1 Mandarin (26%) backgrounds.

The clustering solution for the WordsLast condition, shown in the right panel of Fig. 3, accounted for 91% of the variance in the responses. This solution also involved 4 clusters, the first of which included all 6 L1 American English talkers, 3 L1 Korean talkers (K1f, K2f, K6m), 3 L1 Mandarin talkers (M2f, M4m, M5m), and talker H3f. The second cluster contained only talkers K4m, K5m, and S1f. Five L1 Hindi talkers (H1f, H2f, H4m, H5m, H6m), 3 L1 Spanish talkers (S4m, S5m, S6m), and talker M6m were in the third cluster. The fourth cluster included talkers M1f, M3f, S2f, S3f, and K3f. Again, the backgrounds with the highest correct talker pairing

rates in Table 4, L1 American English and L1 Hindi, each had at least 5 of their members in a single cluster, suggesting robust perceptual categories for these two L1 backgrounds. Talkers from each of the other backgrounds were spread across 3 clusters each. L1 Hindi talkers were again generally distinct from L1 American English, L1 Korean, and L1 Mandarin talkers; although the L1 Hindi cluster for this condition included talker M6m, the only other members were L1 Spanish talkers (S4m, S5m, S6m). In accordance with the pairwise talker error rates in Table 4, the L1 American English cluster included several L1 Korean talkers, but no L1 Spanish talkers and just one L1 Hindi talker, H3f, who was also clustered with the L1 American English talkers in the CVsLast and WordsFirst conditions. Additionally, many of the L1 Korean and L1 Mandarin talkers appeared in the same two clusters, which is consistent with the high pairwise talker error rate noted above. Listeners again seemed to perceive similarity between these two L1 backgrounds. As in the CVsLast and WordsFirst conditions, listeners in the WordsLast condition grouped all the native talkers together and had a clear non-native subcategory for L1 Hindi talkers.

3.2.5. Summary of classification performance

Several generalizations emerge across conditions. First, listeners were able to correctly group L1 American English talkers with one another and to avoid including L1 Hindi and L1 Spanish talkers in these groups. Thus, the perceived difference between native and non-native talkers was especially robust for these two L1 backgrounds. Second, L1 Hindi talkers were themselves often grouped together, and in the case of word-length stimuli, separately from L1 Korean and L1 Mandarin talkers, although not separately from L1 Spanish talkers. L1 Hindi talkers formed a clear, consistent subcategory within the non-native talkers; for word-length stimuli, this subcategory included L1 Spanish talkers but was perceptually distinct from L1 Korean and L1 Mandarin talkers. Third, L1 Korean talkers were classified with L1 American English talkers, and for word-length stimuli, also with L1 Mandarin talkers. That is, with longer stimuli, L1 Korean and L1 Mandarin talkers were perceptually similar to one another, although they did not form a cohesive shared subcategory within the non-native space. Collectively, these three generalizations, which reflect perceptual similarity between L1 Hindi and L1 Spanish talkers and between L1 Korean and L1 Mandarin talkers, suggest that VOT might influence listeners' responses, a possibility considered further in Section 3.3.

3.3. Phonetic cues to classification

The results presented in Section 3.2 show which talkers were generally perceived as sharing L1 backgrounds. However, knowing which talkers were grouped together does not explain how listeners came to make these groups. Because listeners were responding to auditory stimuli, information in the speech signal must have guided their classification decisions. In this section, the acoustic properties that predict listeners' perceptual responses are identified, as these properties may define listeners' non-native subcategories. Although the relatively short auditory stimuli in this experiment made the free classification task challenging for listeners, one advantage that words and CVs have over sentences is that with fewer sounds per stimulus, there are fewer phonetic properties that might influence listeners' responses.

The acoustic measure best predicting membership in each cluster and the details of each logistic regression are shown in Table 5. Clusters are numbered beginning from the top of each panel in Figs. 2 and 3. In the following sections, the acoustic correlates for each cluster are discussed separately for each condition.

Table 5
Best predictors for logistic models of cluster membership.

Cluster	Acoustic measure	β	z	Significance
Condition: CVsFirst				
1	VOT	−0.03	−1.80	$p=0.07$
2	vowel F1 curvature	−0.91	−1.89	$p=0.06$
3	VOT	0.02	1.93	$p=0.05$
Condition: CVsLast				
1	VOT	0.02	2.41	$p=0.02$
2	VOT	−0.05	−2.28	$p=0.02$
3	vowel F1 tilt	0.09	2.19	$p=0.03$
Condition: WordsFirst				
1	VOT	−0.06	−2.14	$p=0.03$
2	vowel F1 curvature	0.16	2.11	$p=0.04$
3	VOT	0.03	2.54	$p=0.01$
4	vowel duration	0.05	2.17	$p=0.03$
Condition: WordsLast				
1	VOT	−0.06	−2.07	$p=0.04$
2	vowel duration	0.05	2.03	$p=0.04$
3	VOT	0.02	2.06	$p=0.04$
4	vowel F1 curvature	0.14	2.32	$p=0.02$

3.3.1. CVsFirst

While no significant predictor was found for any of the 3 clusters in the CVsFirst condition, the best predictors approached significance ($p=0.05$ to $p=0.07$). VOT was the best predictor for clusters 1 ($\beta=-0.03$) and 3 ($\beta=0.02$). As the phonetic properties were represented as absolute acoustic differences from L1 American English means, the positive coefficient for cluster 3 indicates that as a token's VOT value deviated more from the L1 American English mean, the talker who produced that token was more likely to be included in cluster 3. Given the stop production patterns discussed in Section 3, it is sensible that 5 of the L1 Hindi talkers and 4 of the L1 Spanish talkers appeared in this cluster. In contrast, the negative coefficient for cluster 1 suggests that as a token's VOT value deviated more from the L1 American English mean, the talker who produced that token was less likely to be included in the cluster. Because membership in cluster 1 was more probable when VOT was similar to the L1 American English mean, it is not surprising that this cluster contained 4 of the L1 American English talkers. Vowel F1 curvature was the best predictor for cluster 2 ($\beta=-0.91$). The negative coefficient again suggests that as a token's degree of F1 curvature deviated more from the L1 American English mean, the talker who produced that token was less likely to be included in cluster 2. This relationship made the presence of 2 L1 American English talkers in cluster 2 unsurprising. With only 4 members, however, cluster 2 was quite small, suggesting that listeners relied mainly on VOT when classifying talkers in this condition. As expected, this strategy resulted more or less in the separation of L1 Hindi and L1 Spanish talkers from talkers with other L1 backgrounds.

3.3.2. CVsLast

Each of the 3 clusters in the CVsLast condition had a significant acoustic predictor ($p<0.05$). Membership in cluster 2 was best predicted by VOT ($\beta=-0.05$), with greater deviation from the L1 American English means reducing the likelihood of cluster membership. Accordingly, cluster 2 included all 6 L1 American English talkers. Membership in clusters 1 and 3 grew more likely as acoustic measurements deviated more from L1 American English means. VOT was the relevant dimension of deviation for cluster 1 ($\beta=0.02$), which included 5 L1 Hindi talkers and a single L1 Spanish talker. For cluster 3, the relevant dimension was vowel F1 tilt ($\beta=0.09$). Cluster 3 was the smallest cluster, consisting of only 5 talkers from Korean, Mandarin, and Spanish L1 backgrounds. Thus, as expected, this property isolated only talkers whose L1s had substantially different vowel spaces from American English. In general, VOT again explained most of the perceptual judgments made by listeners. Consistent with the predictions described in Section 3, attention to VOT resulted in L1 Hindi talkers (although not L1 Spanish talkers) being grouped separately from the others.

3.3.3. WordsFirst

A significant acoustic predictor was found for each of the 4 clusters in the WordsFirst condition ($p<0.05$). This predictor was VOT for clusters 1 ($\beta=-0.06$) and 3 ($\beta=0.03$). Talkers who produced tokens with VOT values similar to the L1 American English means were likely to appear in cluster 1, while those who produced tokens with VOT values different from the L1 American English means were likely to appear in cluster 3. Accordingly, all 6 L1 American English talkers were included in cluster 1, while 5 L1 Hindi talkers and 3 L1 Spanish talkers made up cluster 3. Thus, as expected, a VOT-based approach generally separated L1 Hindi and L1 Spanish talkers from talkers with other L1 backgrounds. Greater deviation on the dimensions of vowel F1 curvature and vowel duration predicted higher likelihoods of membership in clusters 2 ($\beta=0.16$) and 4 ($\beta=0.05$), respectively. Again, these clusters were smaller than the ones for which membership was predicted by VOT, and as expected, they included only talkers whose L1 vowel spaces differed considerably from American English (i.e., Korean, Mandarin, and Spanish).

3.3.4. WordsLast

Each of the 4 clusters in the WordsLast condition had a significant acoustic predictor ($p<0.05$). Again, membership in the two largest clusters was predicted by VOT, with greater deviation from L1 American English means increasing the likelihood of membership in cluster 3 ($\beta=0.02$), but decreasing the likelihood of membership in cluster 1 ($\beta=-0.06$). Unsurprisingly, cluster

Table 6
VOT deviation (ms) for all free classification stimuli.

Talker	Word	VOT deviation (ms)	Talker	Word	VOT deviation (ms)
E3f	duping	0	K6m	piking	15
K4m	bootie	1	S3f	pity	30
E6m	beagle	2	K2f	turkey	31
E1f	tagging	2	K5m	keeper	31
S5m	couple	2	M2f	Toby	35
M6m	guiding	2	H4m	purple	36
E4m	gable	4	H5m	cackle	44
M4m	kegger	5	S4m	tiger	45
H3f	tubing	5	H6m	buddy	82
K1f	gutter	6	S2f	dating	99
M1f	bidder	8	S1f	goading	102
E2f	kibble	9	H1f	geeky	112
E5m	pepper	10	H2f	dipper	128
K3f	Debbie	12	M3f	dirty	132
M5m	paper	14	S6m	bedding	138

1 included all 6 L1 American English talkers, while cluster 3 included 5 of the L1 Hindi talkers and 3 of the L1 Spanish talkers. Thus, the basic pattern expected in the case of VOT-based classification was again upheld. As in the WordsFirst condition, deviations on the dimensions of vowel F1 curvature and vowel duration predicted membership in clusters 4 ($\beta=0.14$) and 2 ($\beta=0.05$), respectively. These clusters included L1 Korean, L1 Mandarin, and L1 Spanish talkers, exactly the groups of talkers who were expected to deviate from American English vowel norms. There was some overlap between WordsFirst and WordsLast in the talkers who made up these clusters: specifically, 3 of the 5 talkers whose cluster membership was predicted by vowel F1 curvature appeared in this cluster for both conditions (K3f, M1f, M3f), and 2 of the 6 talkers whose cluster membership was predicted by vowel duration appeared in this cluster for both conditions (K4m, S1f). For these talkers especially, deviation from the typical L1 American English values for vowel F1 curvature or vowel duration might have been so salient to listeners that these properties overrode the common strategy of classifying by VOT.

3.3.5. Summary of the phonetic cues to classification

Overall, listeners seemed to attend to properties of both stops and vowels in classifying talkers by perceived L1 background. VOT predicted clustering patterns for many of the talkers, while other talkers were distinguished based on spectral properties of F1, and for word-length stimuli, vowel duration. Notably, in all four conditions, 5 of the L1 Hindi talkers (H1f, H2f, H4m, H5m, H6m) appeared in the cluster characterized by VOT values deviating from the L1 American English means. Because VOT values in English differ substantially between L1 American English and L1 Hindi talkers (Davis & Beckman, 1983), it is possible that VOT alone explains why these L1 Hindi talkers formed a consistent perceptual subcategory while talkers from each of the other non-native backgrounds did not. The VOT deviation values for all stimuli in the free classification task are shown in Table 6, with values from L1 Hindi talkers displayed in bold. All 5 of the L1 Hindi talkers who were consistently clustered together had VOT deviation values larger than at least two-thirds of the talkers. The most extreme deviation values were for words beginning with voiced stop targets, which were produced with lead voicing by the L1 Hindi talkers but generally not by the L1 American English talkers on which the reference means were based. L1 Spanish talkers also exhibited lead voicing and thus extreme deviation values for these targets, but neither the full set of L1 Spanish talkers nor the subset who targeted word-initial voiced stops (S1f, S2f, S6m) were consistently clustered together. Possible explanations for listeners' success with L1 Hindi talkers, as well as the distinctiveness of talker H3f, are addressed further in Section 4.2.

Clearly, natural speech is often longer than the CV- and word-length samples used in this investigation, and listeners interacting with non-native talkers in natural settings likely attend to more than the small number of phonetic properties examined in this study. Remarkably, in the CVsLast condition, listeners' responses to CV-length stimuli could be explained by two acoustic measures. With the addition of a single syllable in the WordsFirst and WordsLast conditions, resulting in stimuli that were meaningful words of only two syllables, three acoustic measures were needed to account for the resulting clusters. Sentence-length stimuli, with many more syllables and greater room for global acoustic properties such as speaking rate, may be too complex to analyze using the present segmental approach. Nonetheless, this approach permitted a test of Atagi and Bent's (2013) claim that the listeners in their free classification task compared the auditory samples they heard to their expectations for native speech. Specifically, the current investigation revealed that listeners can use deviation from the native norm on particular phonetic dimensions to categorize talkers by perceived L1 background.

4. Discussion

The present study explored the structure of American English-speaking listeners' L1-based perceptual categories, as well as the acoustic properties predicting listeners' classification patterns. The well-attested perceived difference between native and non-native speech was confirmed, although it was found to be more robust for some non-native varieties (L1 Hindi and L1 Spanish) than for others (L1 Korean and L1 Mandarin). L1 Hindi talkers formed a clear and consistent subcategory within the set of non-native talkers. L1 Korean and L1 Mandarin talkers did not form a single cohesive subcategory, but were perceptually similar to one another. The acoustic property that most often predicted listeners' categorizations was VOT, although spectral information about F1 was also used, as was vowel duration. Overall, the study demonstrated that even with extremely short auditory stimuli, listeners used VOT deviation from the native norm to recognize a reliable L1 Hindi subcategory within non-native speech.

4.1. Comparison to previous work

The present study extends Flege's (1984) and Park's (2013) findings that listeners can evaluate talker L1 backgrounds somewhat accurately based on very brief speech samples. More detailed aspects of the results can be compared to the work of Atagi and Bent (2013), who used the free classification methodology and many of the same L1 backgrounds that were included in the present study to explore perceptual subcategories of non-native speech. When listeners were instructed to group talkers by perceived native language based on sentence-length materials, Atagi and Bent (2013) found correct talker pairing rates of 98–100% for L1 American English talkers, 34–39% for L1 Korean talkers, 17–42% for L1 Mandarin talkers, and 12–31% for L1 Spanish talkers. The correct talker pairing rates reported in the current study in Table 4 for L1 Mandarin (15–23%) and L1 Spanish (18–25%) fall generally within these ranges. However, the correct talker pairing rates for L1 Korean talkers (15–24%) and especially L1 American English talkers (27–90%) are lower than the ranges reported by Atagi and Bent (2013). The present study was quite different from Atagi and Bent's

(2013) in that participants heard short, diverse stimuli, rather than the same sentence from each talker. Acoustic information not present in these CV- and word-length stimuli, such as sentence-level intonation or particular segments, may contribute considerably to the perception of L1 background, especially for L1 American English and L1 Korean talkers, and thus explain the apparent discrepancies in correct talker pairing rates across the two studies.

Comparison of these two studies also clearly reveals that listeners who hear relatively long stimuli are more accurate at free classification by perceived native language than listeners who hear relatively short stimuli. Thus, it is not surprising that of the present listeners, those in the conditions with words performed better than those in the conditions with CVs. [Flege \(1984\)](#) similarly found that listeners' foreign accent detection abilities were generally better with longer stimuli. The benefit of prior exposure to the talkers is also consistent with previous research, as even small amounts of exposure allow for faster processing of non-native speech ([Clarke, 2002](#); [Clarke & Garrett, 2004](#)), and the intelligibility of non-native speech improves after listeners adapt to a particular L1 background ([Bradlow & Bent, 2008](#); [Sidasas, Alexander, & Nygaard, 2009](#)) or to non-native speech in general ([Baese-Berk et al., 2013](#)).

In the results described by [Atagi and Bent \(2013\)](#), as well as those reported here, listeners were better at grouping together L1 American English talkers than talkers from any non-native background. This effect speaks to the listeners' keen awareness of perceived deviation from their native language norms and their ability to accurately discriminate native from non-native talkers ([Flege, 1984](#); [Park, 2013](#)). Perception of L1-based subcategories within non-native speech was less robust, with some degree of confusion among L1 Korean, L1 Mandarin, and L1 Spanish talkers found in both studies. However, in the present study, listeners showed superior performance on one non-native background, L1 Hindi, as compared to the others, suggesting a robust L1 Hindi subcategory within non-native speech. With short, diverse stimuli that made the task especially difficult, what stands out is not that the listeners did not form clear groups of talkers from the L1 Korean, L1 Mandarin, and L1 Spanish backgrounds, but that they were so successful at doing so for L1 Hindi talkers.

4.2. Exceptionality of L1 Hindi talkers

The findings raise the question of what made the L1 Hindi talkers, but not the talkers from each of the other non-native backgrounds, perceptually similar to one another. One possibility is that regional variability within each L1 background may carry over into talkers' English productions. While diverse regions of origin may explain the lack of L1-based cohesiveness among L1 Mandarin and L1 Spanish talkers, the 4 L1 Korean talkers from the same city, Seoul, were not consistently clustered with one another. In addition, the L1 Hindi talkers were also from numerous regions, but 5 of them consistently appeared in a single cluster together across conditions. Thus, regional variability within each L1 background does not sufficiently explain the observed results.

A second possible explanation for listeners' superior performance with the L1 Hindi talkers is that the status of English in India is different from the status of English in the other non-native talkers' countries of origin. English is an official language of India, and educated Indians regularly use English to communicate with one another ([Gargesh, 2004](#)). For Indian learners of English, the target is not the native speech of a distinct population, such as American or British English, but a "nativized variety of the second language system" ([Wiltshire & Harnsberger, 2006:91](#)): that is, Indian English is itself the target. In contrast, talkers from the other L1 backgrounds generally reported that their education had focused on American English. While other talkers may have deviated from American English pronunciation norms by virtue of inexperience with American English and/or interference from another language, L1 Hindi talkers may have deviated from American English pronunciation norms simply because Indian English is pronounced differently. Talker H3f, who patterned with the L1 American English talkers rather than with the other L1 Hindi talkers in the clustering analyses, moved to the United States during high school and was judged by the first author to be targeting American English rather than Indian English, which may explain her consistently different treatment by listeners.

Although it is important that the non-native talkers were targeting distinct language varieties, this fact alone is not an explanation for listeners' perceptual responses. Rather, it shifts the underlying question to what makes Indian English sound more similar among talkers than the English of L1 Korean, L1 Mandarin, or L1 Spanish talkers. Given that the present stimuli were stop-initial, it is certainly relevant that VOT values are generally shorter in Indian English than in American English, with voiced stops often exhibiting lead voicing ([Davis & Beckman, 1983](#)). Per the stimulus design, there was no repetition in the segments contained in the initial or only syllables of all stimuli within an L1 background. Correct talker pairings could not be achieved by direct acoustic comparison of the same segments; whatever similarity was noted among L1 Hindi talkers must have come from phonetic abstractions made by the listeners, such as noting the talkers' common deviations from L1 American English pronunciation norms. Because the target sequences were not independently made known to the listeners, such judgments required listeners to guess at the intended segments. Performance on word-length stimuli may have been more accurate than on CV-length stimuli in part because lexical knowledge made listeners better able to deduce the word-length targets, and thus better able to make the appropriate phonetic evaluations.

One abstraction that listeners seemed to make was how the tokens they heard compared to their expectations for L1 American English speech in terms of VOT. VOT may have been especially important in this investigation because all of the stimuli were stop-initial, and because across the stimuli there were fewer stop types than vowel types. For both voiced and voiceless stops, VOT consistently predicted membership in the largest high-level clusters and explained the clear differentiation between L1 American English talkers and L1 Hindi talkers, as well as the unusual treatment of talker H3f, who exhibited a VOT value more typical of L1 American English than Indian English. That the L1 Spanish talkers were not consistently grouped together, despite many of them also deviating considerably from L1 American English VOT norms, may be explained by considering the other acoustic measures that were identified as significant predictors of cluster membership. Talkers S1f, S2f, and S6m had VOT deviation values as large as L1

Hindi talkers; however, examination of Figs. 2 and 3 reveals that most of the small clusters for which membership was predicted by deviation on properties other than VOT contained 1 or 2 of these talkers. If L1 Hindi talkers differed from native norms perceptibly on VOT, but L1 Spanish talkers differed from native norms perceptibly on vowel F1 tilt, vowel F1 curvature, and/or vowel duration as well as on VOT, then the L1 Spanish talkers may have been more difficult to group together. This possibility is reminiscent of suggestions that variation in L2 productions is “broader and less predictable” (Girard, Floccia, & Goslin, 2008:427) than variation in regional L1 varieties. Although the L1 Hindi talkers in this investigation self-identified as L2 speakers of English, the different role of English in India as compared to China, Korea, and Latin America may have led to greater between-talker variability for learners in the latter locations than those in India.

4.3. Methodological extensions

A number of methodological improvements could be made in future research. Levels of English proficiency varied among the talkers used here, and to some extent among the L1 backgrounds, as shown by the TOEFL speaking subsection scores in Table A.1. As a group, the L1 Hindi talkers were clearly the most proficient. If the L1 Hindi talkers were also the most intelligible, listeners may have been able to focus more closely on indexical rather than linguistic aspects of their speech. Alternatively, listeners may have based their classification responses largely on proficiency and/or intelligibility. Including talkers with more comparable proficiency and/or intelligibility levels both within and across L1 background groups would make the results more compelling.

Individual variation in many other dimensions pertaining both to talkers and to listeners also impacted this investigation. There are idiosyncratic properties of voices that may make them sound dissimilar, and individual listeners' different experiences with non-native speakers may cause them to structure their perceptual subcategories within non-native speech differently. Future work including more talkers and more detailed questions about listeners' experiences with non-native talkers could reveal more about these sorts of influences.

Investigations that involve multiple non-native varieties and relatively open-ended tasks (e.g., Atagi & Bent, 2013; the present study) may seem quite distant from investigations that involve a single non-native variety and a simple forced-choice task (e.g., Flege, 1984; Park, 2013). To more closely link these types of investigations, each of the multiple non-native varieties could be tested against native speech in an L1 background discrimination task. Of the four non-native varieties in the current study, listeners should be best at discriminating L1 Hindi talkers from native talkers. A similar L1 background discrimination task could also be performed with pairs of non-native varieties. It is possible that, for instance, listeners might successfully discriminate L1 Korean productions in English from L1 Mandarin ones if there were no other speech varieties involved in the comparison, although they often grouped them together when L1 American English, L1 Hindi, and L1 Spanish productions were also presented.

Another extension involves explicit identification of talkers' L1 backgrounds, but with open-ended rather than forced-choice responses to avoid limiting or biasing listeners' judgments. In the present free classification task, listeners did not need to explicitly identify talkers' L1 backgrounds; however, listeners might be able to offer labels that reveal more detail about their perceived non-native subcategories. In a free classification task it can only be known, for instance, that some L1 Hindi and L1 Spanish talkers are perceived as sharing an L1 background. An open-ended identification task can reveal whether the listener believes the L1 background of the talkers to be Hindi or Spanish (or neither).

4.4. Theoretical significance

Listeners draw upon their previous experiences when interpreting the speech of unfamiliar talkers (Goldinger, 1996). However, in a given instance, some previous experiences may be more pertinent than others. This study helps to identify listeners' knowledge about talker similarity on one dimension, L1 background. Despite being explicitly incorporated into some models of speech perception (Kleinschmidt & Jaeger, 2015), the mechanisms underlying such knowledge are currently poorly understood. The present results suggest that when encountering English speech produced by talkers from some L1 backgrounds, such as Hindi, listeners are likely to rely on past experiences with speakers from the same background in interpreting the speech of new talkers. In contrast, other talkers' L1 backgrounds, such as Korean, Mandarin, and Spanish, do not form clear perceptual subcategories, and it cannot be assumed that the previous experiences deemed relevant by listeners necessarily involve talkers described by that L1 label. For instance, a listener hearing English produced by an unfamiliar L1 Mandarin talker may consider some prior experiences with L1 Korean talkers to be highly relevant, and may ignore some past experiences with L1 Mandarin talkers.

The findings from the present study also contribute to the interpretation of earlier work in perceptual adaptation to non-native speech. For example, as discussed in Section 1, Baese-Berk et al.'s (2013) listeners were exposed to talkers from a variety of L1 backgrounds, including Korean, Mandarin, and Hindi, and then tested on a new L1 Mandarin talker and a new L1 Slovak talker. For both test talkers, the previous exposure facilitated listeners' comprehension, perhaps because the talkers' L1 backgrounds (at least Korean/Mandarin on one hand and Hindi on the other) were perceived as quite different, thus leading listeners to activate the more general non-native category rather than any specific subcategory of non-native speech (e.g., Mandarin-accented or Hindi-accented). Bradlow and Bent's (2008) listeners were exposed to a variety of L1 Mandarin talkers, which facilitated comprehension of a new L1 Mandarin talker but not a new L1 Slovak talker. These results are somewhat more difficult to explain, given that L1 Mandarin talkers did not form a single coherent subcategory in the present work. However, they suggest that Mandarin and Slovak are not perceptually similar non-native accents, and that listeners in a free classification task with L1 Mandarin and L1 Slovak talkers would not group them together. Overall, understanding listeners' perceptions of talkers'

L1 backgrounds, rather than relying solely on demographic labels such as “L1 Mandarin,” can better inform work addressing the processing of non-native speech.

Acknowledgments

This work was supported by an Ohio State University Dean's Distinguished University Fellowship to the first author. Thanks to Jeff Holliday and three anonymous reviewers for valuable comments on this manuscript.

Appendix A

See [Tables A.1–A.3](#).

Table A.1

Demographic information about the stimulus talkers, including age at recording, age of arrival in the United States (AoA), and score on the speaking subsection of the TOEFL (TOEFL(S)), if available.

Talker	Residence through age 18	Age	AoA	TOEFL(S)
E1f	Lima, OH, USA	31	–	–
E2f	Cincinnati, OH, USA & Columbus, OH, USA	33	–	–
E3f	Evanston, IL, USA & Plymouth, MA, USA & Powell, OH, USA	20	–	–
E4m	Wichita, KS, USA	26	–	–
E5m	Cleveland, OH, USA	31	–	–
E6m	Douglasville, GA, USA & Claremore, OK, USA	25	–	–
H1f	Pune, Maharashtra, India	22	22	28/30
H2f	Haryana, India	24	22	22/30
H3f	Patiala, Punjab, India & Dayton, OH, USA	21	15	n.r.
H4m	New Delhi, Delhi, India	23	22	28/30
H5m	Jabalpur, Madhya Pradesh, India & Kota, Rajasthan, India	27	26	n.r.
H6m	Pune, Maharashtra, India	26	24	27/30
K1f	Seoul, South Korea	26	20	28/30
K2f	Seoul, South Korea & Seattle, WA, USA	20	16	21/30
K3f	Gumi, South Korea & Daegu, South Korea & Shanghai, China	26	22	26/30
K4m	Seoul, South Korea	24	22	18/30
K5m	Suwon, South Korea & Gwangju, South Korea & Incheon, South Korea	26	26	n.r.
K6m	Seoul, South Korea & Rocky Mount, NC, USA & Fork Union, VA, USA	26	16	n.r.
M1f	Beijing, China	23	18	n.r.
M2f	Wuhan, Hubei, China	26	23	22/30
M3f	Hefei, Anhui, China	34	33	23/30
M4m	Beijing, China	21	20	27/30
M5m	Lankao, Henan, China	27	23	21/30
M6m	Chaoyang, Liaoning, China	28	27	n.r.
S1f	Carolina, Puerto Rico	27	23	n.r.
S2f	Cochabamba, Bolivia	33	30	n.r.
S3f	Concepción, Chile	41	36	n.r.
S4m	Caracas, Venezuela	20	18	n.r.
S5m	Santo Domingo, Dominican Republic	34	25	n.r.
S6m	Cochabamba, Bolivia & Caracas, Venezuela	38	36	n.r.

Table A.2

Mean lag VOT values (ms) from the current study, Robb et al. (2005), and Lisker and Abramson (1964).

		Current study	RG&L (2005)	L&A (1964)
/b/	male	11	6	1
	female	10	4	
/d/	male	16	15	5
	female	16	12	
/g/	male	26	22	21
	female	22	18	
/p/	male	53	58	58
	female	52	73	
/t/	male	65	76	70
	female	71	83	
/k/	male	69	78	80
	female	70	88	

Table A.3

Mean midpoint F1, F2, and F3 values (Hz) from the current study and Peterson and Barney (1952).

		F1		F2		F3	
		Current study	P&B (1952)	Current study	P&B (1952)	Current study	P&B (1952)
/i/	male	301	270	2173	2290	2844	3010
	female	392	310	2855	2790	3298	3310
/ɪ/	male	444	390	1794	1990	2330	2550
	female	601	430	2260	2480	3067	3070
/e/	male	559	530	1754	1840	2362	2480
	female	776	610	2094	2330	3010	2990
/æ/	male	687	660	1628	1720	2317	2410
	female	1022	860	1853	2050	2878	2850
/ʌ/	male	625	640	1263	1190	2294	2390
	female	826	760	1596	1400	2872	2780
/u/	male	345	300	1332	870	2164	2240
	female	440	370	1676	950	2847	2670
/ə/	male	489	490	1330	1350	1589	1690
	female	611	500	1602	1640	1918	1960

References

- Abercrombie, D. (1967). *Elements of general phonetics*. Chicago: Aldine.
- Ash, S. (2003). A national survey of North American dialects. *Publication of the American Dialect Society*, 88, 57–73.
- Atagi, E., & Bent, T. (2013). Auditory free classification of nonnative speech. *Journal of Phonetics*, 41, 509–519.
- Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America*, 133(3), EL174–EL180.
- Baker, R. E., Baese-Berk, M., Bonnasse-Gahot, L., Kim, M., Van Engen, K. J., & Bradlow, A. R. (2011). Word durations in non-native English. *Journal of Phonetics*, 39, 1–17.
- Bansal, R. K. (1981). English and Hindi: A contrastive phonological study. *CIEFL Bulletin*, 17(2), 51–60.
- Bond, Z. S., Stockmal, V., & Markus, D. (2008). A note on native and non-native accentedness judgments. *Ohio University Working Papers in Linguistics and Language Teaching*, 16, 1–8.
- Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106, 707–729.
- Bradlow, A., Clopper, C., Smiljanic, R., & Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, 52, 930–942.
- Canfield, D. L. (1981). *Spanish pronunciation in the Americas*. Chicago: University of Chicago Press.
- Chen, Y., Robb, M., Gilbert, H., & Lerman, J. (2001). Vowel production by Mandarin speakers of English. *Clinical Linguistics & Phonetics*, 15(6), 427–440.
- Chin, T. (2006). *Sound systems of Mandarin Chinese and English: A comparison*. Munich: Lincom Europa.
- Cho, T., Jun, S.-A., & Ladefoged, P. (2002). Acoustic and aerodynamic correlates of Korean stops and fricatives. *Journal of Phonetics*, 30, 193–228.
- Choo, M., & O'Grady, W. (2003). *The sounds of Korean: A pronunciation guide*. Honolulu: University of Hawai'i Press.
- Clarke, C. M. (2002). Perceptual adjustment to foreign-accented English with short term exposure. In *ICSLP-2002*, pp. 253–256.
- Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, 116(6), 3647–3658.
- Clopper, C. G. (2008). Auditory free classification: Methods and analysis. *Behavior Research Methods*, 40, 575–581.
- Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, 51, 175–198.
- Clopper, C. G., & Bradlow, A. R. (2009). Free classification of American English dialects by native and non-native listeners. *Journal of Phonetics*, 37, 436–451.
- Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, 35, 421–438.
- Cortez, J. E. (1998). An efficient metric combinatorial algorithm for fitting additive trees. *Multivariate Behavioral Research*, 33, 249–272.
- Dalbor, J. B. (1969). *Spanish pronunciation: Theory and practice*. New York: Holt, Rinehart and Winston.
- Davis, K., & Beckman, M. (1983). Production and perception of the voicing contrast in Indian and American English. *Working Papers of the Cornell Phonetics Laboratory*, 1, 77–90.
- Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, 20, 1–16.
- Disner, S. F. (1980). Evaluation of vowel normalization procedures. *Journal of the Acoustical Society of America*, 67(1), 253–261.
- 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*, 76, 692–707.
- Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers' production and perception of English vowels. *Journal of Phonetics*, 25, 437–470.
- Flege, J. E., & Eefting, W. (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, 15, 67–83.
- Gargesh, R. (2004). Indian English: Phonology. In E. W. Schneider, K. Burridge, B. Kortmann, & R. Mesthrie (Eds.), *A handbook of varieties of English* (pp. 992–1002). Berlin: Mouton de Gruyter.

- Girard, F., Floccia, C., & Goslin, J. (2008). Perception and awareness of accents in young children. *British Journal of Developmental Psychology*, 26, 409–433.
- Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22(5), 1166–1183.
- Han, M. S., & Weitzman, R. S. (1970). Acoustic features of Korean /P,T,K/, /p,t,k/ and /pʰ,tʰ,kʰ/. *Phonetica*, 22, 112–128.
- Hillenbrand, J., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, 97(5), 3099–3111.
- Institute of International Education (2014). *Open Doors Report on International Educational Exchange* Retrieved from (<http://www.iie.org/opendoors>).
- Kang, K.-H., & 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*, 119(3), 1672–1683.
- Kang, K.-H., & Guion, S. G. (2008). Clear speech production of Korean stops: Changing phonetic targets and enhancement strategies. *Journal of the Acoustical Society of America*, 124(6), 3909–3917.
- Khan, I., Gupta, S. K., & Rizvi, S. H. S. (1994). Formant frequencies of Hindi vowels in /hVd/ and C1VC2 contexts. *Journal of the Acoustical Society of America*, 96(4), 2580–2582.
- Kim, M.-R. (2011). Native and non-native English speakers' VOT productions of stops. *The Linguistic Association of Korea Journal*, 97–116.
- Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, 122(2), 148–203.
- Ladefoged, P. (1999). *American English. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* (pp. 41–44). Cambridge: Cambridge University Press.
- Lee, H. B. (1999). *Korean. Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet* (pp. 120–123). Cambridge: Cambridge University Press.
- Lee, W.-S., & Zee, E. (2003). Illustrations of the IPA: Standard Chinese (Beijing). *Journal of the International Phonetic Association*, 33(1), 109–112.
- Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, 20, 384–422.
- Liu, H.-M., Tseng, C.-H., & Tsao, F.-M. (2000). Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *Clinical Linguistics & Phonetics*, 14(6), 447–464.
- Macmillan, N. A. (1993). Signal detection theory as data analysis method and psychological decision model. In G. Keren, & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 21–57). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Major, R. C. (1987). English voiceless stop production by speakers of Brazilian Portuguese. *Journal of Phonetics*, 15, 197–202.
- Maxwell, O., & Fletcher, J. (2009). Acoustic and durational properties of Indian English vowels. *World Englishes*, 28(1), 52–69.
- Munro, M. J. (1993). Productions of English vowels by native speakers of Arabic: Acoustic measurements and accentedness ratings. *Language and Speech*, 36(1), 39–66.
- Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, 23, 451–468.
- Nathan, G. S. (1987). On second-language acquisition of voiced stops. *Journal of Phonetics*, 15, 313–322.
- Ortega-Llebaria, M. (1997). *An explanatory intelligibility test for Spanish accented English (Dissertation)*. Indiana University.
- Park, H. (2013). Detecting foreign accent in monosyllables: The role of L1 phonotactics. *Journal of Phonetics*, 41, 78–87.
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America*, 24(2), 175–184.
- Riney, T. J., & Takagi, N. (1999). Global foreign accent and voice onset time among Japanese EFL speakers. *Language Learning*, 49, 275–302.
- Robb, M., Gilbert, H., & Lerman, J. (2005). Influence of gender and environmental setting on voice onset time. *Folia Phoniatrica et Logopaedica*, 57(3), 125–133.
- Rogers, C. L., & Dalby, J. (2005). Forced-choice analysis of segmental production by Chinese-accented English speakers. *Journal of Speech, Language, and Hearing Research*, 48, 306–322.
- Ryan, C. (2013). *Language use in the United States: 2011. American Community Survey Reports, U.S. Census Bureau*.
- Sandahl, S. (2000). *A Hindi reference grammar*. Leuven, Belgium: Peeters.
- Schirra, R. (2012). *Attitudes toward Korean-accented and Korean American English (Master's thesis)*. University of Washington.
- Shimizu, K. (2011). A study on VOT of initial stops in English produced by Korean, Thai and Chinese speakers as L2 learners. *Proceedings of the 17th International Congress of Phonetic Sciences*, 1818–1821.
- Sidaras, S. K., Alexander, J. D., & Nygaard, L. C. (2009). Perceptual learning of systematic variation in Spanish-accented speech. *Journal of the Acoustical Society of America*, 125(5), 3306–3316.
- Silva, D. J. (2006). Acoustic evidence for the emergence of tonal contrast in contemporary Korean. *Phonology*, 23, 287–308.
- Sumner, M., Kim, S. K., King, E., & McGowan, K. (2014). The socially weighted encoding of spoken words: A dual-route approach to speech perception. *Frontiers in Psychology*, 4, 1–13.
- Sohn, H.-M. (1999). *The Korean language*. Cambridge: Cambridge University Press.
- Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, 75, 983–992.
- Tsukada, K. (1998). Japanese-accented English vowels: A perception study. *Asia Pacific Journal of Speech, Language, and Hearing*, 3, 43–65.
- Tsukada, K., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Flege, J. (2005). A developmental study of English vowel production and perception by native Korean adults and children. *Journal of Phonetics*, 33, 263–290.
- Vieru, B., Boula de Mareüil, P., & Adda-Decker, M. (2011). Characterisation and identification of non-native French accents. *Speech Communication*, 53, 292–310.
- Wang, H., & van Heuven, V. J. (2006). Acoustical analysis of English vowels produced by Chinese, Dutch and American speakers. *Linguistics in the Netherlands*, 237–248.
- Watson, C. I., & Harrington, J. (1999). Acoustic evidence for dynamic formant trajectories in Australian English vowels. *Journal of the Acoustical Society of America*, 106(1), 458–468.
- Wayland, R. (1997). Non-native production of Thai: Acoustic measurements and accentedness ratings. *Applied Linguistics*, 18(3), 345–373.
- Wiltshire, C. R., & Harnsberger, J. D. (2006). The influence of Gujarati and Tamil L1s on Indian English: A preliminary study. *World Englishes*, 25(1), 91–104.