Syrett, K., Lu, J., & Parrish K. (2024). Perceptual benefits of linguistic diversity and language background: Evidence from auditory free classification of English dialect accents and Asian accented English. *Glossa Psycholinguistics*, 3(1): 22, pp. 1–46. DOI: https://doi.org/10.5070/G60111672

Perceptual benefits of linguistic diversity and language background: Evidence from auditory free classification of English dialect accents and Asian accented English

Kristen Syrett, Rutgers, The State University of New Jersey – New Brunswick, US, kristen.syrett@rutgers.edu
Joy Lu, Rutgers, The State University of New Jersey – New Brunswick, US, lujoy.slp@gmail.com
Kyle Parrish, Goethe University Frankfurt, DE, kparrish.linguistics@gmail.com

Non-linguistic factors leave a distinct thumbprint on our speech production that is perceptible to listeners. A steadily growing line of research demonstrates that listeners can perceive a contrast between native and non-native (L2) speakers based on accents, and further classify these speakers according to dialectal variation, even when they are not native speakers of a language. Most of these studies have focused on dialectal variation *within* US English speakers, a combination of US and International English dialects, or L2 speakers representing a wide range of languages. Most have also mostly featured listeners who are monolingual native speakers of the target language coming from a homogenous background, or a contrast between these and a targeted set of L2 speakers. We therefore lack knowledge of how exposure to, or familiarity with, diverse accents and languages, or specific native language competence of the native language of L2 speakers, can guide listeners' accent perception and categorization. In this research, we employed a free classification task, presenting listeners with speech samples of native speakers with accents representing multiple English dialects, and L2 speakers of nine Asian languages across three geographic regions speaking Asian-accented English. There were six groups of listeners: monolingual US English listeners in a diverse linguistic context, monolingual US English listeners in a homogeneous linguistic context, native speakers of a non-Asian language and English (bilinguals), and native speakers of each of the three target Asian language groups who are L2 speakers of English. The results reveal that nearly all listeners are sensitive to accents capturing native/L2 contrasts and dialectal variation in English. While regular exposure to a diversity of accents results in increased classification accuracy, classification of Asian L2-accented English speakers is best performed when there is alignment of similar language family and geographic area, as demonstrated by South Asian listeners.

1. Introduction

Even without seeing a speaker, we can tell a great deal about *who* a speaker is based on *how* they speak. This is so, because our speech contains meaningful indexical information about us as a speaker – our age, gender, sexual orientation, social status, and so on (Harnsberger et al., 2008; Hunter et al., 2016; Lass et al., 1976; Leongómez et al., 2017; Merritt et al., 2024; Moyse, 2014; Munson, 2007; Munson et al., 2006). Speech accent, in particular, can serve as a window into a speaker's origins, the languages they know, and to what degree of proficiency they know them. As we listen to people speak, we actively recruit acoustic-phonetic information that not only allows us to infer more about the speaker, but also to decide how they align with, or differ from, other speakers along demographic indices. At the same time, our own language background and familiarity with languages and dialects can influence how we perceive a speaker's accent and how their accent compares with other speakers.

Previous research has shown that listeners are remarkably good at distinguishing between native and L2 speakers (Atagi & Bent, 2016; Bent et al., 2016; McCullough, 2015) and, further, can distinguish between accents reflecting dialects of their own language (Clopper & Pisoni, 2007; McCullough et al., 2019), and those accents reflecting groups of languages other than their own, although with diminished classification accuracy (Atagi & Bent, 2016). Most research in this domain has focused on monolingual English speakers' categorization of either US dialectal differences or differences between US and "foreign accented" English, based on speaker accent, and, for the most part, has recruited listeners of a fairly homogeneous American English background, and speakers representing a fairly narrow representation of English, or else a wide variety of the world's languages (see, e.g., Bradlow et al., 2010; Clopper, 2008; Clopper & Bradlow, 2008, 2009; Clopper & Pisoni, 2007; a.o.). We thus know very little about how exposure to accents representing a diversity of dialects and/or languages, being a multilingual speaker, or having language-specific competence in non-English languages can influence accent perception and categorization. At the same time, we also know little about how listeners might engage in making more fine-grained distinctions among accents reflecting languages from a similar geographic region in the world, which may or may not be members of the same language family. The goal of the current research is to address gaps in these aspects of speech accent perception, with the broader goal of addressing how we make inferences about speakers based on their accents in an ever-increasing multi-lingual and mobile global society.

We begin with an overview of reliable indicators of accent in production. We then turn to a comprehensive review of evidence concerning how listeners recruit segmental phonetic cues when perceiving differences in speaker status relative to geographic region and linguistic background, and how a listener's own linguistic background and/or familiarity with a language influences their identification of languages and/or dialects. The findings from the studies reviewed in the next section indicate that even listeners without fluency in a target language (or dialect) are able to detect L2 speaker status in production, and benefit from increased exposure to, and familiarity with, the language. We then identify a gap in the research to date: how membership in a diverse linguistic context (daily familiarity with multiple languages and dialects) benefits listeners' perception of accents, and how this compares with perception by listeners who are part of a homogeneous linguistic context, those who are multilingual speakers, and native speakers of these target languages within a broad and more specific geographic region. We present the results of an online free classification task that sheds light on the benefits of exposure to, and engagement in, linguistically diverse contexts, and the development of tailored linguistic knowledge. These findings reveal how qualitatively different kinds and levels of linguistic knowledge shape our perceptual organization of speaker accents.

2. Background

2.1 Linguistic and dialectal variation

It is by now well-known that speakers of the same language can pattern very differently in their speech production, and display different phonological systems and rules, based on the geographic region in which they live (Baranowski, 2008; Blake & Josey, 2003; Chambers, 2006; Clopper & Bradlow, 2009; Clopper & Pisoni, 2006; Clopper, Pisoni, & de Jong, 2005; Foster et al., 2017; Fridland, 1999; Hubbell, 1950; Labov, 1962, 1966, 1998; Labov et al., 2006; McCullough et al., 2019; Preston, 1993; Sankoff & Blondeau, 2007; Trudgill & Hannah, 2002). The phonemic and subphonemic acoustic cues that signal diverse dialects are perceptible to speakers of a target language, even without significant exposure to dialectal variation (Clopper, 2008; Clopper & Pisoni, 2007). Pronunciation differences can also signal level of proficiency and when a language may have been acquired. These differences may not only be perceptible to a listener and signal the second language (L2) status of the speaker, but may also allow them to be identified as a fluent or *native* speaker[1] of one specific language rather than another, even when they are speaking English (see, e.g., Hartshorne et al., 2018; Johnson & Newport, 1989; Lenneberg, 1967). Late learners of English who are influenced by their own native language display patterns that reflect morphosyntactic features (see the discussion and references in, e.g., Hopp, 2013),

[1] We are aware of the problematic status of "native language/speaker" terminology in psycholinguistics (Cheng et al., 2021), and yet we find the terminology unavoidable in the context of this article, given the distinctions in linguistic backgrounds that are central to the current research. Here, we define *native language* as a language (or languages) which a speaker acquired early in the critical period, which they would themselves identify as their first language(s). and in which they are most fluent. Further, when speaking their native language, a native speaker would be identified as such by someone else with a similar language background (i.e., would be identified as a native speaker or as someone who grew up speaking the language in question early in the critical period). We replace "non-native" with *L2*, even though this terminology also presupposes a particular sequential order of acquisition and number of languages.

or phonological and/or phonetic features (see, e.g., Baese-Berk et al., 2020; Bent et al., 2008; Davidson, 2006).

While L2 speaker status or regional dialect may be signaled at multiple levels of the speech signal, there is evidence that information encoded at the segmental level is weighted more heavily than information at the suprasegmental or intonational level (Alcorn et al., 2020; Carrie & McKenzie, 2018; Clopper & Pisoni, 2004a, 2007; Clopper, Pisoni, & de Jong, 2005; Flege, 1984; Leemann et al., 2018; Park, 2013; Ruch, 2018; Sereno et al., 2016; Van Bezooijen & Gooskens, 1999). At the same time, suprasegmental information may still play a role (Barkat et al., 1999; Munro, 1995; Munro et al., 2010). Research also indicates that listeners are highly sensitive to the presence of an L2 accent – even when they themselves are not native speakers of the target language (Major, 2007; see also Bradlow & Pisoni, 1999; Munro, 1998; Rogers et al., 2001, 2004; van Wijngaarden, 2001). Previous research has also demonstrated that L2 accents and accents of "non-standard" and/or unfamiliar dialects incur a processing penalty relative to those for native languages and familiar dialects (Adank et al., 2009; Clarke & Garrett, 2004; Floccia et al., 2006; Munro & Derwing, 1995; Wade et al., 2007).

Individual experience with languages and the environments or cultures in which these languages are spoken can improve a listener's performance (see, e.g., Clarke & Garrett, 2004; Xie et al., 2018). In fact, intensive exposure to a language is shown to result in more accurate production and perception of that language at the segmental level (Best & Strange, 1992; Flege et al., 1997). A further distinction may exist between highly proficient speakers and experienced L2 learners involved in regular contact with a specific language environment and culture, who experience pressure to acquire lexical content and "re-phonologize" the perception of phonological contrasts (Best & Tyler, 2007). While there is evidence that self-reported familiarity with a language may aid in identification of the language in accented speech, this still does not prevent it from being confused with another language in L2 accented speech (Atagi & Bent, 2015; Bent et al., 2016; Derwing & Munro, 1997; McKenzie, 2015).

It thus appears that perceived shared geographic or regional similarity plays a role in classification, in addition to perceived phonological similarity – a point that we leverage and expand upon in the current work. At the same time, while having some familiarity with particular languages may facilitate recognition of those languages via accented speech, *lack* of familiarity does not necessarily amount to inaccurate accent identification. Perceptible patterns of production by speakers of certain languages may be sufficient to allow listeners to identify a speaker's native language (see McCullough, 2015; Vieru et al., 2011).

Individual exposure to specific dialects can also improve listeners' performance. In their survey of attitudinal evaluations of New Zealand English, Australian English, and American English varieties, Bayard et al. (2001) reported that when asked to identify the nationality of a speaker, all native listeners were better at identifying their respective target dialects, as one would anticipate. However, Australian and New Zealand listeners were better at correctly categorizing American English than American English listeners were at categorizing New Zealand and Australian dialects, and New Zealand listeners were more likely to mis-categorize speakers of their dialect as Australian speakers than vice versa. Likewise, Adank et al.'s (2009) participants who were recruited from Glasgow showed degraded performance with Spanish-accented English relative to both standard Southern English and Glaswegian dialects, outperforming their standard English dialect counterparts with the latter dialect. In addition, individual speakers who have more experience with dialectal variety (e.g., through regional mobility) may show superior performance relative to those who have a more insular, region-specific experience (Clopper & Pisoni, 2004a, 2004b, 2006, 2007; Williams et al., 1999; although see Alcorn et al., 2020, and the possibility of effects due to age of exposure). These findings strongly suggest that increased exposure to, and familiarity with, dialects and languages outside of one's own generally results in enhanced perception and categorization of speakers of those languages and dialects, based on their accents. An open question that the current research seeks to answer is whether similar benefits of broader language exposure extend to listeners consistently exposed to multiple languages, and how specialized this additional linguistic experience must be to be beneficial.

2.2 Free classification task

The current study employs an *auditory free classification task*, in which participants are presented with audio samples, and must attend to acoustic phonetic attributes of the sound files, or sublexical and subphonemic cues in the speakers' productions, to perform comparisons across speakers and create their own categories without pre-specified labels or category guides, based on the features they perceive (or not). See Imai (1966), Clopper (2008), and Clopper and Bradlow (2009). Listeners, therefore, choose to create a larger or smaller number of *categories*, and larger or smaller *category sizes*, based on their decisions about category membership, which are based on their detection of relevant features. Previous researchers have successfully used free classification, sometimes paired with other tasks, to investigate how listeners from different language backgrounds categorize speakers of different dialects and native speaker backgrounds. A summary of recent auditory free classification tasks is presented in **Table 1**.[2] We include the present research in the last row of the table.

With the exception of Atagi and Bent (2016), most free classification tasks to date have not manipulated native and L2 listeners *and* native and L2 speakers simultaneously *within* a task. In the current research, we do so, in order to investigate the role of language exposure and

[2] There have also been other types of classification tasks employed that are not free classification tasks, as they have provided listeners with labels or categories, or explicit directions about how to create groupings (e.g., by region) or the number of groups to create. See, e.g., Clopper, Pisoni, & de Jong (2005), Clopper and Pisoni (2004a, 2004b).

Table 1: Summary of relevant previous studies and the current study (last row), using the free classification methodology.

	Speakers		Listeners	
Study	**Native**	**L2**	**Native**	**L2**
Alcorn et al. (2020)	6 regional US dialects	---	monolingual speakers of US English: 1: mostly Midland and Southern dialects; 2	---
Clopper & Pisoni (2007) experiments 1 and 2	6 regional US dialects	---	monolingual speakers of US English: 1: mostly Midland dialect; 2: non-mobile Midland, mobile Midland, non-mobile North, mobile North	---
Clopper (2008)	6 regional US dialects	---	monolingual speakers of US English: Midland dialect	---
Clopper & Bradlow (2008) experiments 3 and 4	4 regional US dialects	---	monolingual speakers of US English, 3 groups: Northern, New England + Midland + Western, mobile	---
Clopper & Bradlow (2009) experiments 1 and 2	4 regional US dialects	---	monolingual speakers of US English: mix of dialects	2 groups of speakers of non-English languages: mostly Mandarin, heterogeneous
Bradlow et al. (2010) experiment 1	17 languages	---	---	monolingual speakers of US English

(Contd.)

Study	Speakers		Listeners	
	Native	L2	Native	L2
Atagi & Bent (2013)	US midland	French, German, Spanish, Japanese, Korean, Mandarin	monolingual speakers of US English: primarily Midland	---
Atagi & Bent (2016)	US midland	French, German, Spanish, Japanese, Korean, Mandarin	monolingual speakers of US English: primarily Midland	2 groups of speakers fluent in English living in US: Korean, Spanish
McCullough & Clopper (2016)	---	Hindi, Korean, Mandarin, Spanish	monolingual speakers of US English	---
Bent et al. (2016)	6 US English dialects, 6 International English dialects	Arabic, French, German, Gujarati, Japanese, Korean, Mandarin, Russian, Spanish, Somali, Swahili, Thai	monolingual speakers of US English: primarily Midland	---
Current study	3 US English dialects, 3 International English dialects	speakers of 9 Asian languages from three regions: Southeast (Gujarati, Bengali, Urdu); South (Indonesian, Tagalog, Thai); East (Japanese, Korean, Mandarin)	monolingual speakers of US English: diverse linguistic context; monolingual speakers of US English: homogeneous linguistic context	bilingual speakers of non-Asian, non-English languages (mixed); speakers of 3 regional groups of Asian languages (East, South, Southeast)

linguistic background in perception and classification. Three previous studies cited in **Table 1** set the stage for the current work.

Clopper and Bradlow (2009) focused on the categorization of American English dialects. They targeted native and L2 listeners recruited from Northwestern University, a prestigious private Midwestern university in Evanston, Illinois. Across two experiments, their participants included native speakers of American English from diverse backgrounds and L2 speakers. Among the L2 group, the majority (38) were native speakers of Mandarin, and the remaining 30 represented 10–12 diverse languages and language families, thus constituting a very heterogeneous group. While most had spent little time in the US at the time of the study, they had been admitted to Northwestern and, therefore, demonstrated English proficiency in their high TOEFL scores. Clopper and Bradlow's participants listened to 20 speakers from the TIMIT Acoustic-Phonetic Continuous Speech Corpus (Garofolo et al., 1993), representing four dialect regions in the United States (New England, North, Midland, and South), with no international dialects or L2 speakers. The two experiments differed mainly in the sentence prompt from the TIMIT speakers that listeners heard to perform the classification.

Participants in Clopper and Bradlow's (2009) free classification task were asked to put all of the talkers from the same part of the country in a group together. Listeners across all three groups (native, L2 Mandarin, and L2 heterogeneous) generated the same number of groups: approximately 6 groups, on average, with three to four speakers per group. Perhaps unsurprisingly, native listeners were more accurate than the two L2 groups in their classifications, placing speakers from the same dialect together in a group and not placing speakers from different dialects together as often. Overall, native and L2 listeners patterned similarly within and across the two experiments. They were likely to identify three to four groups of speakers based on English dialects. Thus, both native and L2 listeners were able to classify native speakers of American English by regional dialect, but native listeners still outperformed their L2 counterparts.

Atagi and Bent (2016) focused on the categorization of a range of accents across diverse languages. They compared the performance of monolingual native listeners of US English, and native speakers of Spanish and Korean as L2 listeners, in a free classification task. Listeners were presented with two English sentences from the Hoosier Database of Native and Nonnative Speech for Children (Atagi & Bent, 2013; Bent, 2014) produced by speakers of various language backgrounds, including English. Native listeners of English were more accurate in their classification than L2 listeners, while listeners with a Spanish or Korean language background were more accurate in their categorizations of speakers of those languages than the native English listeners. What's more, for all three groups of listeners, the native US English speakers were more closely clustered together compared to speakers from the other language backgrounds, and were treated as more similar to speakers of German and French than to speakers of the Asian languages. The authors concluded that "linguistic experience plays a significant role in shaping listeners' classification of native and nonnative varieties of English" (p. 257). The authors noted that the largest population of international students at Indiana University (the home university of the native listeners) is Korean, and conjectured that “The high accuracy demonstrated by these native listeners, therefore, may be due to native listeners’ exposure to Korean-accented English on campus” (p. 257). This pattern led them to suggest that future researchers investigate native listeners with different experiences with L2 accents. The current research explicitly does so.

Finally, Bent et al. (2016) has had the most diverse sample of speakers and listeners to date. They employed a range of *speaker* productions of the script from the Speech Accent Archive (a segment of the exact script used in the present research), representing six U.S. regional dialects, six international English dialects, and twelve L2 accents within the same tasks. The U.S. regional dialects included Mid-Atlantic, Midland, New England, North, South, and West. The international English dialects included dialects spoken in Australia, England, Ireland, New Zealand, Scotland, and South Africa. The L2 accents were produced by speakers of Arabic, French, German, Gujarati, Japanese, Korean, Mandarin, Russian, Spanish, Somali, Swahili, and Thai. Thus, the geographic regions and language families were extremely wide-ranging. All told, there were 72 speakers representing these regions or language backgrounds. All of the *listeners* in their task were from a monolingual English background, and all were recruited from Indiana University in Bloomington, Indiana, a Midwestern state university with little ethnic and linguistic diversity. Forty-five of the 50 listeners grew up in the Midwest, and 30 grew up in Indiana. Most of the students had not studied abroad, and most had little experience with L2 speakers, some referring to their “foreign professors.”

Bent and colleagues paired together a *free classification task* (where listeners group speakers together) and *a ladder task* (where listeners rank talkers relative to their proximity to “standard American English”). Given the range of 24 possible groups, the listeners in Bent et al. (2016)’s study created, on average, 11 groups (range 5–19), with 7 speakers in a group, on average (range 4–14). Thus, despite their own lack of multilingual, multidialectal knowledge, listeners categorized beyond the native/L2 distinction, and created approximately five main groups: (a) International English, (b) a mix of American and International English, and within L2, three clusters including (c) French and German, (d) the Asian languages, and (e) a mixed bag, including Swahili, Russian, and Gujarati. It is these last two categories that led us to ask in the current research if exposure to diverse Asian languages and accents, in particular, and specialized knowledge of these languages within a geographic region, could result in more fine-grained groupings. Thus, we seek to probe further how a listener’s native background and regular exposure to other languages (whether at all, or relative to targeted speech samples) facilitates the categorization of accents.

2.3 Interim conclusions

The following picture emerges from the previous research. First, listeners of various backgrounds perceive a native/L2 contrast among speakers, sometimes even based on minimal segmental linguistic information. They are also able to distinguish among speakers of different dialects within a language. Second, experience with, and exposure to, speakers, languages, and dialects can positively impact – but does not necessarily ensure – accuracy in perception or classification. Third, both L2 and native listeners – even those from a homogeneous monolingual background – can successfully categorize native speakers of English according to regional dialects, and L2 speakers by general language groupings. However, native listeners are more accurate in their identification and classification than L2 listeners.

2.4 Current research

To date, no previous study has explicitly compared native listeners and L2 listeners with a linguistic background in *specific* target languages in a task involving both variation among native/regional English dialects *and* L2-accented English. In addition, no previous study has compared monolingual listeners with exposure to multiple languages, monolingual listeners with limited linguistic exposure, and multilingual speakers with or without knowledge of the target languages, to determine whether and how familiarity with linguistic diversity compares with specialized or multilingual competence. The current study not only aims to fill those gaps, it also builds on, and extends, previous research using an auditory free classification task methodology to shed light on the perceptual benefits of multilinguistic exposure and competence.

Our aims are twofold. First, we seek to investigate how accents capturing dialectal variation in English are perceived and classified, both by native English listeners with and without exposure to multiple languages and regional dialects, and by listeners with native proficiency in other languages. Second, we wish to determine how four specific listener groups perceive and classify a selection of Asian-language-accented speech: native speakers of these specific Asian languages, monolingual English speakers with exposure to multiple dialects and languages, including these Asian languages, monolingual English speakers with limited exposure to linguistic variation, and speakers who are multilingual, yet have no proficiency in these Asian languages. Three main questions guided our research and our choice of speakers and listeners.

First, *does exposure to linguistic diversity, including consistent exposure to specific languages and accented English spoken by native speakers of those languages, allow for fine-grained categorization of speakers by their accent, based on their dialect, and accurate categorization of speakers by accent, based on their native language-accented English?* To answer this question, we chose listeners who are native English speakers attending college in an extremely linguistically diverse context with a significant Asian population of speakers (where these English speakers can hear the Asian languages spoken, *and* hear Asian-language-accented English), and compared them to a group of similar-aged college students from a more homogeneous and non-/minimally linguistically diverse context who are not regularly exposed to these languages or related L2-accented English.

We presented both groups with speech samples from speakers of multiple English dialects (regional and international), and from speakers of nine Asian languages, most, if not all, of which are regularly spoken in the first listener group's context.

Second, *does multilingual status (rather than exposure to multiple accents or specific languages or accents) afford listeners an enhanced perceptual ability to discriminate among accents representing different dialects and language backgrounds, even those which the listeners have no proficiency in?* To answer this question, we chose listeners with knowledge of multiple European languages and some English proficiency, asking them to categorize the same speakers. We compare their performance to listeners who are native speakers of these languages, and the monolingual English speakers from a linguistically diverse setting.

Third, *does native speaker status result in listeners' increased ability to perceive and classify speakers not only from their own native language, but from languages spoken in the same general geographic region?* To answer this, we constructed our speaker sample of nine Asian languages out of three subgroups of three languages each, representing three geographic regions within Asia. We then asked how listeners representing these three subgroups of Asian languages compared with each other, and with the other listeners in our sample, in categorizing these speakers. We also sought to determine if this fine-grained knowledge of acoustic distinctions within Asian languages might generalize to a refined ability to categorize other accents, notably those representing dialects within English.

3. Experiment

3.1 Listeners

293 participants were recruited as listeners. Of these 293, 125 were recruited from a subject pool of introductory Linguistics and Cognitive Science students at Rutgers, The State University of New Jersey – New Brunswick, and were compensated with extra credit in their course. Represented in this population was our monolingual college-age population attending college in a highly ethnically and linguistically diverse setting (both in higher education and in the surrounding geographic location), in which they are exposed to multiple languages, accents, and dialogues on a daily basis, including the Asian languages and accented speech under investigation. The vast majority of these students grew up in the Mid Atlantic or Northeast region of the US. A second pool of 21 monolingual participants were recruited via targeted efforts from small liberal arts colleges (e.g., Swarthmore, Haverford, Bryn Mawr, Smith) or universities in the South (Texas Tech University) in which the participant population is highly homogeneous and majority monolingual, and were compensated with a $5 Amazon gift card for their participation. 147 participants were recruited from the Prolific platform online, and were compensated with a $5 Amazon gift card. No participant reported a history of hearing loss or a communication or speech

disorder. Each participant provided informed consent as part of an IRB-approved protocol, conforming to research regulatory guidelines.[3]

The classification task involved dragging emojis from the left side of the screen into a grid on the right, forming categories of at least three members each. Of the 292 participants, 52 participants were excluded across samples for the following reasons: they did not move any emojis from the left side of the screen into groups on the right side (n = 10); they omitted 8 (>10%) emojis from the task altogether (n = 1); they moved emojis from the left side of the screen into the grid on the right with no distinct grouping whatsoever (n = 14); they took less than 10 minutes to complete the task (n = 21); they did not complete the task (n = 3); they attempted to participate and submit answers twice (n = 2); or they misreported their native language (n = 1). After these exclusions, we were left with 241 participants (female: 189, male: 48, transgender: 1, non-binary: 1, no response: 2). 115 of these participants were recruited from the Rutgers University Linguistics and Cognitive Science subject pool; 21, from outside recruitment; and 105, via Prolific.

A separate group of monolingual English-speaking participants (n = 26) formed a baseline group that performed the free classification task based solely on sorting the emojis, with no sound files linked to images. The performance of this *emoji-sorting control group* allowed us to determine if there was anything in the images themselves that might have contributed to the formation of consistent categories that could have resembled accent-based groupings, and if there was any correlation between visual properties of the emojis and languages.[4] The rest of the participants (listeners) were divided into six groups.

The first group was *English monolingual listeners from a diverse linguistic context*,[5] recruited directly from Rutgers University – New Brunswick, which is a large state university in the US on the East Coast in central New Jersey (Middlesex County) and near New York, in which they were exposed to diverse languages and dialects) (n = 61). The second group was *English monolingual listeners* recruited from small liberal arts colleges in the US where the population is largely

[3] The use of these distinct participant pools was necessary in order to recruit participants from each of the target linguistic demographic backgrounds. Regardless of their recruitment pool, all participants completed the same online experiment. Based on previous experimental work reported elsewhere and conducted in our lab, we had no reason to anticipate any differences in performance between the two populations, other than a potentially higher attrition rate for participants recruited online, which we did not experience in this study. Moreover, independent studies comparing subject pools to online platforms such as Prolific and MTurk have reported better performance by the former, specifically in terms of attention checks (Douglas et al., 2023; Hauser & Schwarz, 2016), as well as following instructions, providing meaningful answers, and remembering previously presented information (Douglas et al., 2023; Hauser & Schwarz, 2016), none of which were relevant to the current free classification paradigm.

[4] We thank a reviewer for suggesting this control condition.

[5] We use the term *monolingual* here to mean that these participants self-report proficiency and fluency in *one* language and no others.

homogenous and heavily monolingual (n = 21).[6] The third group was fluent *L2 speakers of English and a non-target non-Asian language spoken in Western Europe* (i.e., Spanish, French, German, Greek) (n = 58). Measures of English proficiency, such as TOEFL scores, were not available for the L2 participants; they were recruited online on the Prolific platform for the sole purpose of the current study. All self-reported fluency and high proficiency in English.

The fourth, fifth, and sixth groups included fluent L2 speakers of English who are native speakers of at least one of the target Asian languages (n = 75). Within this set of *Asian language groups*, *three subgroups* were represented, which are characterized here by geographic region: (a) *South Asian* (e.g., Bengali, Gujarati, Hindi, Malayalam, Punjabi, Tamil, Telugu, Urdu) (n = 23); (b) *Southeast Asian* (e.g., Indonesian, Malay, Tagalog-Filipino, Thai, Vietnamese) (n = 27); and (c) *East Asian* (e.g., Cantonese, Korean, Mandarin, Japanese) (n = 25). Participant recruitment continued and was targeted towards a particular language and speaker samples until the groups and subgroups were comparable in size and robust enough for analysis.

3.2 Stimuli

3.2.1 Languages and speakers

We selected 45 speech samples from the Speech Accent Archive[7] online (Weinberger, 2015). This database contains recordings of speakers ranging from 18 to 65 years of age reading the same script. These particular speakers were selected from the larger set of speakers in the database for each dialect/language, based on the lack of background noise in the recording, as well as the lack of disfluencies, mispronunciations, excessive pauses, repetitions, and lexical substitutions, and because – based on the perceptible phonetic characteristics of their speech – they were assessed as representative of the target dialects and languages.[8] (See also Clopper & Bradlow, 2009; Clopper & Pisoni, 2004a, 2007).

[6] It was *much* easier to recruit the participants for the first group, who attend the home university of the authors, where we have a dedicated subject pool of students actively seeking extra credit opportunities and a much larger population of students to draw from in one location, than in the second monolingual group. We were persistent and creative in our recruitment of the second group, and stopped recruitment efforts when it was clear that we had exhausted our resources.

[7] We chose the Speech Accent Archive for the range of speakers and languages represented, and the length of the prompt we could target, which provided acoustic-phonetic cues in both consonants and vowels. It has also been successfully used in previous free classification tasks (e.g., Bent et al., 2016). By contrast, the TIMIT corpus only includes productions of dialects of American English, not International or L2 productions. Previous research has successfully used TIMIT to reveal differences in how these dialects are perceived and grouped (Clopper & Pisoni, 2004a, b; Clopper & Bradlow, 2009).

[8] Sound files were reviewed by the first two authors and lab members. The first author grew up in the American South until college, and has since lived in all three American dialect regions for extended periods of time, and traveled and lived internationally. The second author grew up in the Mid-Atlantic, is a native speaker of an East Asian language, and self-identifies as either a bilingual or a heritage speaker. Lab members represented a wide range of language backgrounds.

There are two main divisions among the languages represented among speakers: English and non-English/Asian. Within each, we divide the space further. Within US English, there were three regional dialects: New England, Midland, Southern. These were chosen based on the consistency of dialect identification in previous studies (see Clopper, 2008; Clopper & Pisoni, 2004b, 2006, 2007; McCullough et al., 2019).[9] We contrasted US English with an *International English* category: Australian, British, and South African English. The English dialects were contrasted with 9 Asian languages, spoken in three different regions: South Asian, Southeast Asian, East Asian. Within these Asian language subgroups, three specific languages were selected: South Asian: Bengali, Gujarati, Urdu; Southeast Asian: Indonesian, Tagalog, Thai; East Asian: Japanese, Korean, Mandarin.

Note that the grouping of these languages together does not presuppose a shared "genetic" origin, and we do not have reason to think that listeners will group them based on language families. See, for example, the relevant discussion in Bradlow et al. (2010), paired with findings from their ladder task in Experiment 3 with native speakers and listeners of various non-English languages, including Mandarin and Korean, which yielded similar patterns. Native US English listeners in Atagi and Bent (2013, 2016) also grouped Japanese, Korean, and Mandarin talkers within the same cluster, distinct from speakers of Romance languages and German.

While we are aware of the additive effect of gender on assessments of similarity among dialects (Clopper, Conrey, & Pisoni, 2005; Clopper, Levi, & Pisoni, 2006) and languages (Atagi & Bent, 2013), given our requirements on recording quality and speaker fluency and intelligibility, paired with the specific languages we targeted and the limited availability of some of those languages in the Speech Accent Archive (especially Southeast Asian languages), it was not possible to obtain speaker files in which speakers were all of the same gender. However, gender was otherwise carefully balanced throughout the samples, and the sound files were meticulously reviewed by the researchers for clarity, comprehensibility, and relative uniformity within sets. Other previous studies have also balanced gender; see, e.g., Clopper (2008) and McCullough and Clopper (2016). We also note that while Clopper et al. (2005) found some effects of gender on dialect categorization by region, consistent with the variationist literature on language change, they also replicated Clopper and Pisoni's (2004a) results using mixed-gender listener groups.

[9] Listeners in Clopper and Pisoni's (2004b) region identification task identified three main categories (New England; South and South Midland; North Midland and West), while listeners in Clopper's (2008) free classification task differentiated Mid-Atlantic/North/South from New England/Midland/West, and then Northern (New English, Mid-Atlantic, North) from non-Northern (Midland, South, West). Given these patterns, we targeted North/New England, South, and Midland.

3.2.2 Features of the linguistic stimuli produced

We targeted the first two sentences of each recording for presentation, since these contained segmental features that served to distinguish between the accents reflecting the relevant dialects and languages. Each speaker was heard producing the following passage in English.

(1) Please call Stella. Ask her to bring these things with her from the store: six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob.

Each truncated .wav file recording lasted approximately 8 to 13 seconds. Based on previous research, which targeted an even shorter segment of the productions for classification purposes (see, e.g., Atagi & Bent, 2013, 2016; Clopper, 2008; Clopper & Bradlow, 2009; McCullough & Clopper, 2016), we were confident that this length provided sufficient auditory data to guide participants' classifications. All sound files were normalized by amplitude to 67 dB.

3.3 Procedure

Individuals were invited to participate in a task called "Call Stella." They were directed to an online form where they completed an IRB-approved consent form and indicated their language background and proficiency, self-identified gender (male, female, transgender, binary, other), and the languages they speak with native or heritage fluency (and for each of these languages, the age at which they began learning it and where they learned it). For English speakers recruited from the subject pool, we also asked how long they had lived in the US, if they had lived in any other countries, and if so, for how long.

Participants then viewed a 2-minute instructional video to acclimate them to the task before proceeding on to the task proper. In the video, the narrator (a researcher and the second author) welcomed the participant and explained that the experiment would be accessed via Google slides available via a link after the video. The narrator then proceeded to explain the setup of the experiment, using an example with five flower emojis on the left side of the screen instead of the actual emojis used in the experiment. See **Figure 1**.

45 emojis were located on the left half of the screen, each arbitrarily representing a different speaker from one of the target languages/dialects. Care was taken not to have any visual features which were associated with the languages or dialects, or which invited the possibility that these superficial perceptual features could influence the classification in any way. The script read by all 45 speakers (see (1) above) was printed below the entire set of emojis. To the right of the emojis was a blank grid to be used as a background for classification (see **Figure 2**).

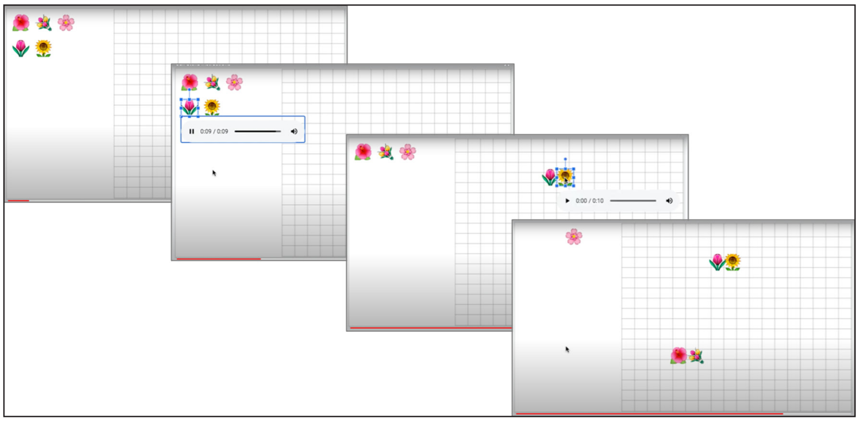
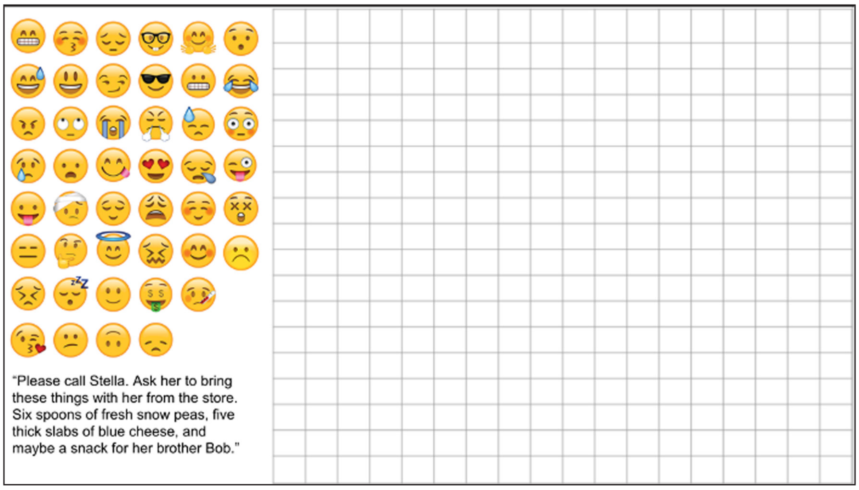


Figure 1: Sequence of images from the instructional training video preceding the experimental task demonstrating free classification.



"Please call Stella. Ask her to bring these things with her from the store. Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob."

Figure 2: Initial pre-classification scene with 45 emojis and script to the left of a blank grid.

Participants were instructed to click on each emoji to listen to the audio file linked to it. They were told that while the passage may seem strange, it was chosen because it highlights differences in the languages and dialects of different speakers. They were also explicitly told that the facial expressions of the emojis were irrelevant, and that they should listen to the audio files, and drag the emojis to the grid, arranging them into clusters based on how similar or different they sound, with similar ones close together. (The instructions for the baseline condition without audio were different, in that participants did not need to listen to the audio files.) The narrator demonstrated each of these steps in the video.

Participants were not required to situate the emojis relative to the grid or next to each other in any particular way. They were told that they could listen to the audio files as many times as they wanted, create as many clusters as they wanted, and put as many emojis in a cluster as they wanted, but that each cluster must have at least three speakers in it. They did not have to listen to the files in any particular order, and they were allowed to move the emojis around on the grid as many times as they wanted or needed to. Participants were not asked to provide labels for their groupings or hypothesize about why speakers/emojis belonged together. They simply had to group them into clusters based on how they sounded.

When the participants were satisfied that they had successfully grouped the emojis into clusters based on how they sounded, the task was complete, and they exited the task. Their work was autosaved on Google slides. Participants were not granted credit or given compensation if they spent less than 10 minutes completing the task, and they were informed of this requirement in advance. Typically, participants took between 15 to 25 minutes to complete the classification clustering. (However, the baseline no-audio condition took approximately 8 to 10 minutes, so the restrictions on timing were loosened.) Once the participants' classification was completed, a researcher then accessed the slides and took a screenshot of the final slide with the clusters, saving it with the participant number. A representative sample of clustering from a subgroup of participants is presented in **Figure 3**.

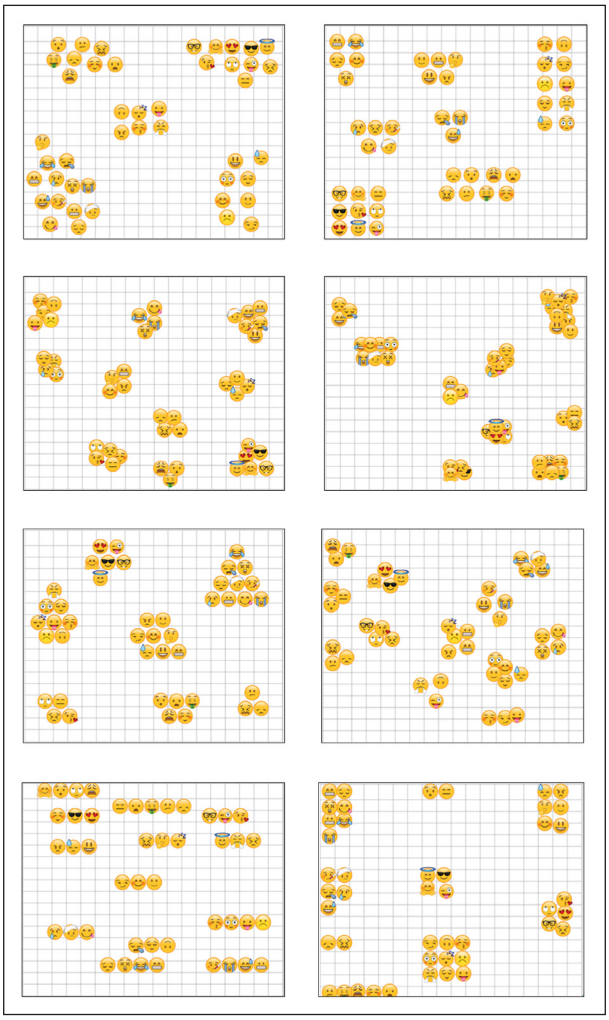


Figure 3: Examples of eight actual participant classifications/clusters of emojis/speakers.

Because our study was run online, not in a laboratory setting, we limited the timing of the experimental session to 30 minutes maximum.[10] Participants were instructed to complete the study in a quiet setting without distractions, in one sitting, and on a laptop, desktop, or tablet (*not* a smartphone). We recorded the time it took each participant to complete the task, and excluded those who took less than 10 minutes, which we determined via piloting would be the absolute minimum time necessary. (In other studies, we have also excluded participants whose time stamp exceeded a certain amount of time, indicating that they did not take the study in one sitting or were otherwise distracted. This step was unnecessary for this study.).

3.4 Predictions

3.4.1 Classifications based on speaker/language background

We generated the following predictions. First, we predict that listeners will be able to distinguish between accents for native and L2 speakers, and create at least two clusters based on these speaker profiles. It has been established that speaking rate is a reliable indicator of L2 status: L2 speakers produce speech at slower rates (Bent et al., 2016; Major, 2007; Munro & Derwing, 2001), pause more often (Bent et al., 2016), and are more variable in their rate of production, than native speakers (Baese-Berk & Morrill, 2015). Our auditory stimuli are long enough to allow for intonation and production rate to be factored into clustering. If listeners can use these two metrics (articulation rate and number of pauses) to classify speakers, then they should create at least two categories, differentiating between native and L2 speakers of English.

Beyond this distinction, we are confident that listeners will be able to make more fine-grained distinctions (and create a larger number of clusters), as previous research has also shown that both native and L2 listeners are able to sub-categorize dialects within a language. It is thus possible that native monolingual English listeners from the US – and perhaps also L2 listeners – will create three subcategories among the accents representing the US dialects, and may also differentiate among the non-US varieties of English. It is also possible that the implicit contrast between these two strands of English (native to the US and native to other countries), paired with the implicit contrast between English and non-English Asian languages, might also exert an influence on classification strategies. At a minimum, therefore, we anticipate observing three categories across listeners (US English, International English, and Asian), and additional subcategories within these groups, based on indexical features of the speakers and phonological features of the languages.

[10] In-person in-lab human subject collection was restricted by the University, due to COVID, when this study was first launched. For consistency's sake, we continued to administer the classification task online after this restriction was lifted.

While the non-English Asian languages were selected based on their geographic region, they do not all share the same language family membership, sound structure, or phonological inventory, and these differences may influence performance in a free classification task. Gujarati, Urdu, and Bengali (the South Asian languages) are members of the Indo-Aryan branch of the Indo-Iranian subfamily of the Indo-European language family. (English and many other European languages are also Indo-European languages.) The other languages, however, do not share language family membership. Within the Southeast Asian languages, both Indonesian (a standardized variety of Malay) and Tagalog are members of Malayo-Polynesian branch of the Austronesian family, while Thai is a Southern Tai language of the Kra-Dai language family. Tagalog and Indonesian (Malay) also share many cognates (see https://en.wiktionary.org/wiki/Appendix:Malay%E2%80%93Tagalog_relations).

Within the East Asian languages, Mandarin is a member of the Sino-Tibetan language family, Korean is a member of the Korean language family, and Japanese is a member of the Japanic language family. While Korean and Japanese are often anecdotally claimed to sound similar, Mandarin is a tone language. Indeed, McCullough and Clopper (2016) reported that while Korean and Mandarin speakers were perceptually similar, “they did not form a cohesive shared subcategory within the non-native space” (p. 30). However, Mandarin could be placed in the same category as Korean and Japanese through implicit comparison with English and other Asian languages, and by the indexical variable of the geographic region of the languages. At the same time, while both Thai and Mandarin are tonal languages and are anecdotally reported to sound alike when spoken by native speakers, the speakers in the current task were recorded producing utterances in English, not their native language, so these features are not expected to be a guiding factor to our listeners.

We, therefore, predict that the three South Asian languages (Gujarati, Urdu, and Bengali) have the best chance of being grouped together, since they share language family membership and have similar phonological inventories, and speakers of these languages may also speak Indian English. Further, we might expect that Korean and Japanese will cluster together, and that Tagalog and Indonesian (Malay) might cluster together. It is an open question where Thai and Mandarin will be grouped, based on their family membership, phonology, and geographic status. We do not, of course, expect listeners to know the family membership of these languages or to be guided by them at all in their classification strategy, and there appears to be no reason to think that the tonal status of these two languages will be detected in accented English.

The results from McCullough and Clopper’s (2016) free classification task are relevant here. Those results revealed that English monolingual listeners created tight clusters for both English and Hindi, grouping native speakers of these languages (all of whom were producing English sentences) together in their respective groups; however, Spanish speakers were often placed in a group with Hindi speakers. By contrast, listeners did not create tight clusters for Korean speakers or Mandarin speakers. Thus, perceptually similar productions were grouped together, but only for one language in particular other than English: Hindi. What is striking is that the Hindi talkers were not from the same region. By contrast, with other languages, even speakers from a particular region – even the same city – were not grouped together: Korean speakers from Seoul were not consistently in the same cluster. Thus, perceptual similarity appears to trump regional origin when listeners are forced to attend to the acoustic features in brief stimuli. It is an open question whether this same pattern will hold when listeners are presented with a contrast among nine different languages spoken on the same continent, with three possible regional geographic groupings.

3.4.2 Classifications based on listener background

While both native and L2 listeners may be able to differentiate among speakers and languages to some extent, a listener's *particular* language background may afford them an advantage, resulting in more accurate, tight-knit groupings. Based on previous research on the effects of mobility and exposure to dialectal variation, we predict that our native monolingual US English speakers, who attend a large, highly diverse state university (Rutgers; see 3.1) and are surrounded by a range of Asian accents on a daily basis, will be able to do more than just identify speakers of English as L2 speakers and accurately identify regional US English dialects. An open question is how much more fine-grained their classification will be beyond English, given their English background and the categories of speakers and languages represented. One might predict that familiarity with multiple accents and awareness of different phonological systems will result in better classification of accents representing dialectal and linguistic variation overall. A follow-up question is how these listeners compare with monolingual English speakers from a more homogeneous linguistic background, with multilingual listeners who are native speakers of non-Asian languages, and with listeners who are native speakers of those Asian languages.

We predict that the monolingual English speakers from a diverse context will be better at classification than those from a homogeneous context, given their exposure. However, we do not have an *a priori* prediction about how they will compare with speakers who know multiple languages. There is reason to think that listeners with specialized knowledge of Asian languages will be able to more accurately discriminate among speakers of Asian-language-accented English. Again, however, it is an open question how much these results would be realized for languages within and across different geographic regions. Given the results of previous free classification tasks, we predict that our listeners may create a comparable number of main clusters, and that no one group would emerge as choosing significantly more or fewer groups. Within these clusters, we then ask what further groupings we will observe, guided by language proficiency and exposure.

3.5 Preparation for analysis

The clusters created by each participant were reviewed manually by a researcher (the first or second author), with reliability coding conducted by at least one other researcher for each participant file. The researcher tracked categories by assigning the same number to each member of a category, with the numbers being chosen arbitrarily. To make this concrete, in the classification scheme in the top right corner of **Figure 3**, there would have been seven categories, and all ten emojis in the group in the top right corner of that grid would have been assigned the same number (e.g., any number between 1 and 7). This strategy provided a count of how many categories each participant created by identifying the maximum number used to indicate grouping. For example, the participant whose output is in the bottom right corner of **Figure 3** created nine categories, while the one in the upper left created five. When participants did not sort an emoji into a group, and it was either missing from the classification or remained to the left at the end of the task, this cell was indicated with n/a.

While we instructed participants (listeners) to create categories of *at least three* speakers, they did not always do so, and occasionally created categories with only one or two members. These 1- or 2-speaker groups were counted as errors and included in all analyses. The n/a classifications were replaced during the clustering analysis with a new category label that varied by participant. So, for example, if an individual made 14 groups of 3 and did not categorize the final 3 (or put them into individual groups), we counted this as three errors. If they grouped two speakers from the same category together and excluded one, this would be one error.

3.6 Results

We carried out multiple statistical analyses using R (R Core Team, 2022) to determine whether language background has an impact on the classification of accented English. We first present the description and rationale for each analysis, then present the results. Figures were generated using ggplot (Wickham, 2016).

3.6.1 Description and rationale for analysis

3.6.1.1 Analysis 1: Number of categories created and error rate analysis

First, we report descriptive statistics indicating *the total number of categories created by each of the 6 groups of listeners*, given the same 45 speakers. To determine how appropriate or correct these created categories were, three error rates were calculated for each group: 2-category creation, 5-category creation and 15-category creation. The 2-category error rate measured how often participants inappropriately grouped speakers from an Asian language category with speakers from an English language category, and vice versa. The 5-category error rate measured how often participants inappropriately grouped speakers from any of the five categories (English monolingual, International English, East Asian, South Asian, Southeast Asian) with speakers from another category. Finally, the 15-category error rate measured how often participants inappropriately grouped speakers from any of the 15 language categories with speakers from a different language category. Categories created with just one member (*single categories*) also counted as an error, since the minimum number of members in a category would be three (of the same language). Thus, in theory, the maximum number of errors was 45; this would occur if a participant created 45 single categories.

The error rate in each case was the total number of errors divided by the maximum number of errors (45). The label of the created category itself was based on whichever accent occurred most frequently. We considered errors to equal the total number of members with other accents relative to the total number of members in the category. We illustrate categories and errors with some hypothetical examples.

When a majority group within a created category did exist, the identity of the created category was determined by the most frequent accent in the group. If a participant grouped together two American English speakers and a British English speaker, then American English was considered correct and British English was considered incorrect and counted as one error. If a group had an equal number of accents (e.g., a 3-member group consisting of one speaker of Urdu, one speaker of Southern American English, and one speaker of British English), then no speakers were correctly grouped together (in the 15-category error rate), so they each counted as one error, for a total of three errors. In a group of four, with two Urdu speakers, one New Zealand English speaker, and one Southern American English speaker, any non-Urdu member was treated as an error, so in this case, there were two errors. If there was a group with 10 members, in which four had one accent and there was no other subgroup comparable in size, then the other six members were counted as errors, even if within those six, there were some correct groupings (e.g., 4 British English, 3 Southern American English, 3 other distinct accents). In the event that there was no majority group in the created category, this did not impact the total number of errors, and the overall category label was arbitrary. For example, if four speech samples were grouped into a category, and two of them were American English and the other two were British English, this would be counted as two errors, regardless of the overall category label.

Given that we have 5- and 2-category error rates as comparisons to the 15-category error rate, it seems appropriate not to consider this sort of nested correctness in the error rate. (In this last case, the 2-category error rate would capture the fact that the participants grouped English dialects together.) To have 45 errors, a participant would have to have *not* categorized any speakers into groups, or they could have created 15 categories of three each in which none of the group members matched in accent. A detailed example of error calculation can be found in the supplementary materials.

The total number of errors was analyzed using a Bayesian multilevel Poisson regression model in R. The outcome variable of the model was *the number of predicted errors*. The fixed effect predictors were language group (7 levels: English monolingual diverse, English monolingual homogeneous, South Asian, Southeast Asian, East Asian, non-Asian multilingual, and the emoji-sorting control group) and error classification type (3 levels: 2-category, 5-category, and 15-category). The random effects included a random intercept per participant to take into account the nested structure of the data. The model included the default brms function priors (Bürkner, 2017), a Student's T distribution with 3 degrees of freedom. The model was run using 2000 iterations of Hamiltonian Monte-Carlo sampling (1000 warm up), across 4 chains and 6 processing cores.

3.6.1.2 Analysis 2: Cluster analyses and additive trees

In addition to the total number of categories and error rate analysis, we conducted a *cluster analysis*, or *clustering*, which groups together objects or data points according to measured or perceived similarity. Rather than relying on pre-determined category labels and the tagging of objects with these identifiers, it allows for the identification of patterns based on several variables simultaneously, with the aim of finding structure in the data (Cheng et al., 2021; Jain, 2010). Clustering analyses entail an iterative pairwise calculation of distances, which can be represented either as a *dendrogram* or as points in a multi-dimensional space. In each case, similar objects or clusters of objects are nodes that get joined together. A similarity matrix is calculated, with each cluster serving as a single object that feeds into the comparison, and a similarity matrix is recalculated again.

The present work utilized *additive similarity trees*. Additive trees allow for the examination of how each listener group categorized specific languages together. In an additive similarity tree (Sattath & Tversky, 1977), clusters and the distances between these units are graphically represented in a *dendrogram*. This analysis was chosen as an alternative to k-means clustering, since that approach requires the experimenter to determine the number of clusters in advance, and hierarchical clustering implicitly assumes that the distance between objects within a cluster or category will be similar and shorter than across clusters/categories. Additive trees do not have these limitations.

3.6.2 Results of the analyses

3.6.2.1 Analysis 1: Number of categories created and error rate analysis

We begin by presenting the number of categories created by the listener groups. **Figure 4** shows the total number of categories created by each of our six listener groups and the emoji-sorting control group. Recall that the categories were calculated by the unique groups created by each participant, which could, in principle, range from 1 (if the participant grouped all 45 speakers together into one large category, which we did not expect), to 45 (if the participant decided that *none* of the 45 speakers should be grouped together, which we also did not expect). In reality, by design, the speakers represented 15 groups of three speakers each (as outlined in 3.2.1). Overall, the average number of groups created by the listeners ranged from 7.73 (SD: 2.68) to 10 (SD: 2.62). The English monolingual group from a linguistically diverse context created the *most* categories, on average, while the control group created the *lowest* number of categories, on average. This difference amounts to an effect size of Cohen's D = .87 (95% CI .39 – 1.34), which is typically considered to be a large effect (Cohen, 2013; Plonsky & Oswald, 2014).

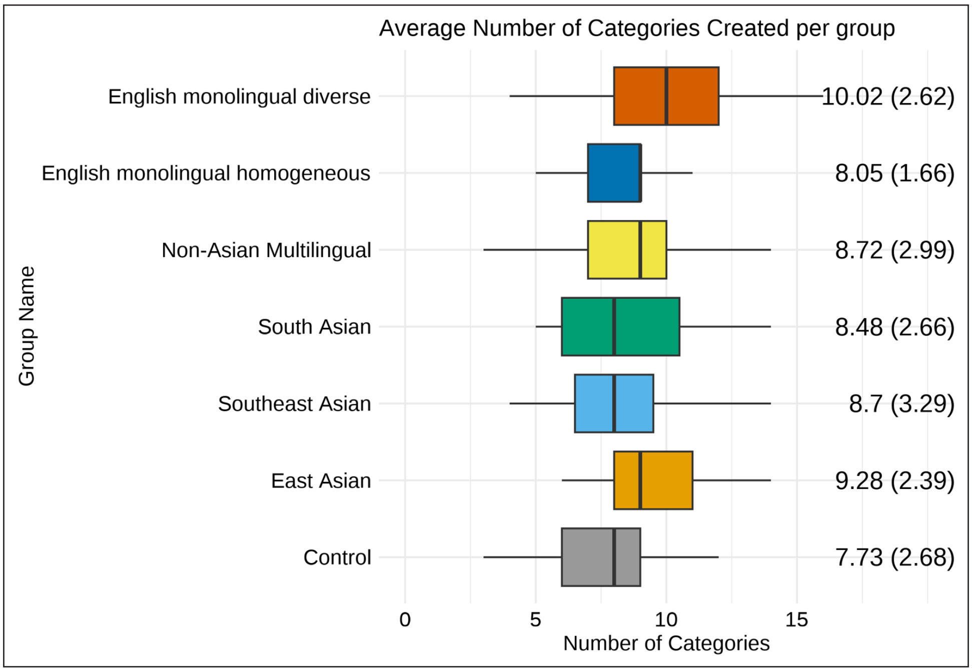


Figure 4: The average number of categories created by each of the six listener groups and the emoji-sorting control group.

3.6.3 Error rates

We then calculated the error rates for these groups. (More details on the process are included in the supplementary data site.) See **Table 2** below for the conditional effects for each group, linked to the Poisson regression model. **Figure 5** shows the error rates by each group (including the non-listener emoji-sorting control group) in the 2-, 5-, and 15-category classifications, where the vertical lines represent the standard deviation of the error percentage in each group. Single categories were counted as errors and included. As expected, the 2-category error rate was low for all groups (less than 5%), indicating that listeners committed few errors separating English native speakers from L2 speakers. By contrast, in all cases, the 15-category error rate was the highest; each group had a total error rate of over 50%. This result indicated that no listener group created 15 categories of three speakers each, corresponding to the pre-selected speaker profiles from the Speech Accent Archive. Finally, the 5-category error rate fell in between the 2- and 15-category error rate and ranged from 18% to 30%. This pattern reveals that the listener groups committed errors in determining the variety of Asian languages (East, South or Southeast) or in separating American and international English varieties, around 1 in 4 times. Note that the emoji-sorting control group deviated sharply in their error rate from the groups that listened to the sound files to classify them.

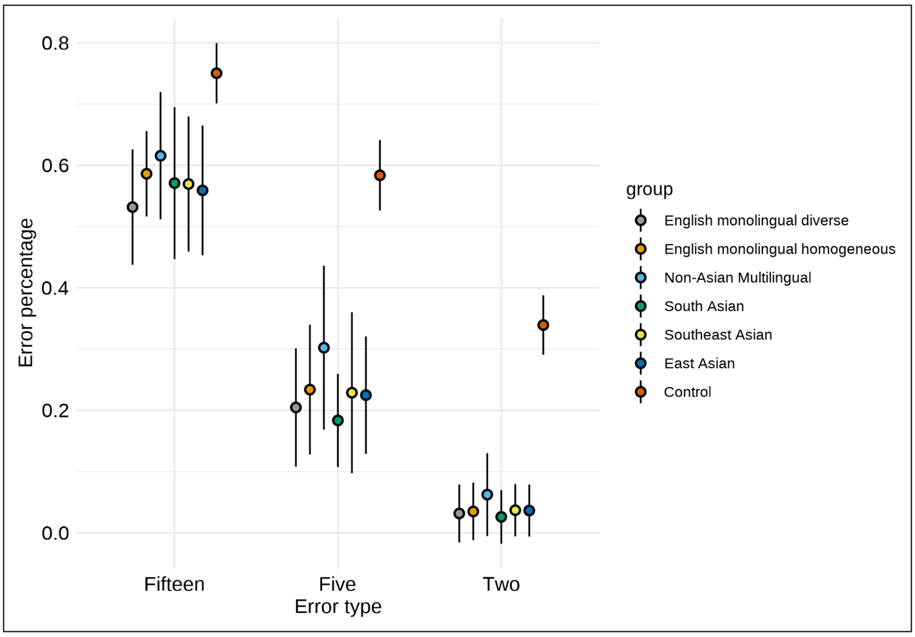


Figure 5: Percentage of errors by each group in 2-, 5-, and 15-category classifications.

We then took a closer look at the error rate for the listener groups, comparing the English monolinguals from a linguistically diverse context, the English monolinguals from a linguistically homogeneous context, the non-Asian multilinguals, and the three groups of Asian language speakers. **Figure 6** shows the posterior distributions of the predicted number of errors by each of these groups in each error classification. The distributions are made up of 4000 plausible estimates generated by the model (that is, those that are *possible* or *able* to occur, versus *likely* to occur; the posterior distribution is a distribution of outcomes that are able to occur, with some more likely than others). Each point represents the most plausible estimate (the mean of the posterior distribution), and the point interval represents the 66% and 95% of highest density intervals of the posterior distribution.

Overall, the emoji-sorting control group was predicted to produce the highest number of errors in all three error classification types. Their 2-category classification error number was predicted to be 15.16 [95% HDI 13.3–17.1], while their 5-category classification error number was 26.1 [95% HDI 23.4–29] and their 15-category classification error number was predicted to be 33.6% [95% HDI 30.4–37.2]. Accordingly, in **Table 2**, they also have the largest conditional predicted error rate. Of all of the listener groups, the non-Asian multilingual group had the second highest number of predicted errors: the model predicted an error total of 2.7 [95% HDI 2.3–3.2] in the 2-category error classification, 13.3 [95% HDI 12.2–14.4] in 5-category error classification, and 27.01 in 15-category error classification [95% HDI 25.2–29.1]. They also had a higher conditional predicted error rate than the other listener groups, who all seemed to have comparable error rates. Finally, the English monolingual diverse group had the lowest predicted number of 15 category errors (23.4 [95% HDI 21.8–25.1]) – a finding that may reflect the fact that listeners with native English proficiency who are exposed to multiple accents are best at discriminating among accents representing English dialects. In fact, the English monolingual diverse group consistently had lower predicted error rates than their English monolingual homogeneous counterparts, which comes across clearly in the 5- and 15-category error rates in **Figure 6** and **Table 2**.

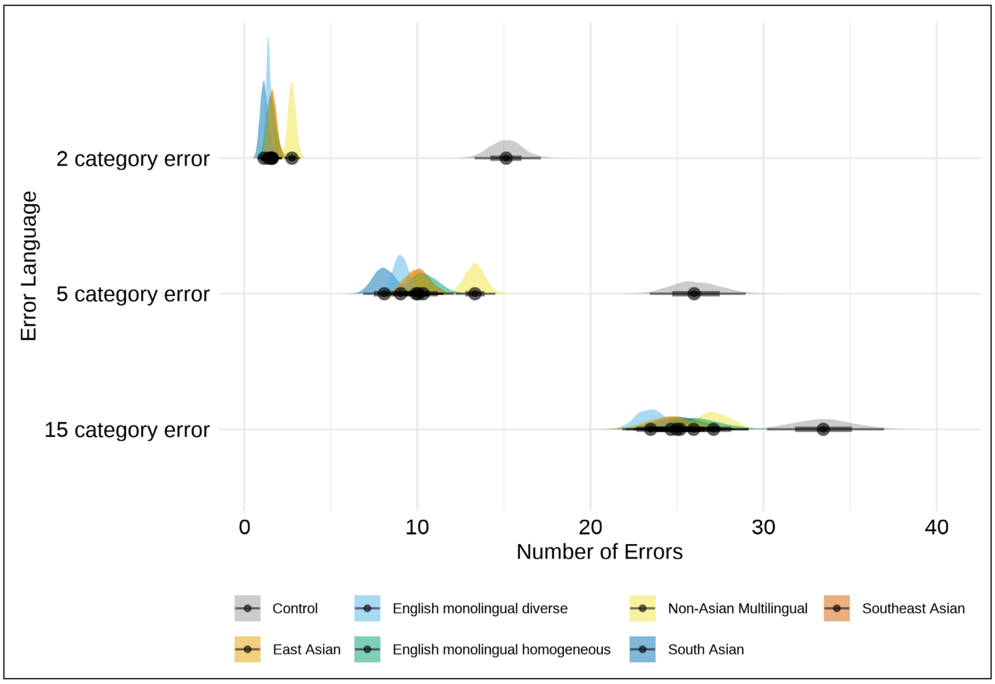


Figure 6: The predicted error rate per group in each error classification, including 2-category, 5-category, and 15-category error rates.

Table 2: The conditional effects of the number of errors by each group and error classification, based on the Poisson regression model.

Error Type	Group	Conditional Predicted Error Rate	Standard Error	Lower	Upper
2	Control	15.17	0.98	13.33	17.16
2	Non-Asian multilingual	2.74	0.24	2.32	3.23
2	South Asian	1.14	0.22	0.75	1.64
2	Southeast Asian	1.61	0.24	1.19	2.14
2	East Asian	1.60	0.25	1.15	2.17
2	English monolingual homogenous	1.54	0.28	1.06	2.16
2	English monolingual diverse	1.40	0.16	1.11	1.73
5	Control	26.11	1.43	23.42	29.08
5	Non-Asian multilingual	13.29	0.59	12.20	14.48
5	South Asian	8.08	0.68	6.87	9.51
5	Southeast Asian	10.04	0.71	8.73	11.46
5	East Asian	9.91	0.75	8.58	11.45
5	English monolingual homogenous	10.37	0.81	8.85	12.07
5	English monolingual diverse	9.03	0.44	8.19	9.92
15	Control	33.60	1.76	30.36	37.22
15	Non-Asian multilingual	27.10	0.99	25.23	29.17
15	South Asian	25.16	1.44	22.51	28.22
15	Southeast Asian	25.03	1.37	22.47	27.82
15	East Asian	24.69	1.35	22.12	27.32
15	English monolingual homogenous	26.03	1.55	22.97	29.28
15	English monolingual diverse	23.45	0.85	21.81	25.14

3.6.4 Analysis 2: Cluster analyses and additive trees

We now turn to our cluster analysis and additive trees, which provide an estimation of both the quantity of clusters created by each group and the group membership within each cluster. Neighbor-Joining trees (Saitou & Nei, 1987) were created in R, using the nj function. The x-axis of each of these plots represents *perceptual similarity*, such that the rightmost nodes are the most supported in clustering, while the leftmost nodes are the least supported. The distance of the languages on the vertical axis is irrelevant. The colors of the text represent the five pre-selected speaker groups mapping onto individual emoji images (American English dialects: purple; International English variants: blue;[11] East Asian: yellow; South Asian: red; Southeast Asian: orange). Below, we present an additive tree for each listener group, within which all five speaker groups are depicted. These trees continue to pull back the layers of the onion, revealing how language-specific exposure and knowledge benefits the perception of accents and the categorization of speakers.

Before turning to the results for the six listener groups, we want to be sure that participants in our free classification task were not influenced by the visual attributes of the emojis themselves when creating categories. While we had explicitly told participants verbally and in written instructions that the emojis were irrelevant, and they should sort by what they heard in the sound files, we could not be sure that the images did not influence their categorization. As we mentioned, we therefore ran a version of the task in which an independent group of monolingual English speakers was recruited from the same linguistically diverse context as the first monolingual group of participants. This group was simply instructed to "arrange the emojis into clusters any way you think they should be grouped. However you group them is up to you, but find some systematic way to group them!" Like the other groups, they were also told that each group should have at least three members.[12] The results of this control group are presented in **Figure 7**. As the figure shows, *no consistent classification pattern surfaced, and similar dialects and languages were generally not grouped together*.

Complementing the lack of any consistent pattern displayed in the additive tree, and the lack of groupings that could be perceived as based on the accents in the corresponding sound files, are the comments from a post-classification query that we included in this specific version of the task. We explicitly asked this group, unlike the other participant groups, to report on their classification strategy afterwards, thereby allowing us to determine if they were guided by a principled pattern (or, at least, were aware of one). While most participants were unsurprisingly influenced by the emotion or affect depicted, and by salient perceptual features, a wide variety of self-reported strategies of grouping and a wide variety in the resulting number of categories

[11] We included Afrikaans-accented English in this grouping, although admittedly Afrikaans is a language, not an English dialect.

[12] We thank an anonymous reviewer for the suggestion that we run this control condition.

surfaced, as captured in (2–5). Thus, we can be confident when we turn to the listener groups that any groupings reflecting English dialects or languages are driven by perceived accent.

(2) I have 5 groups of emojis. I grouped the top left with all emojis I thought showed happiness. Middle left is a bit more than happy. I would describe these emojis as silly. Bottom left is miscellaneous, but have personality. Top right shows negative emotions that are not sad. Middle right is all the emojis I perceived to show sadness.

(3) I grouped the emojis based on how I use them. The top left section is the shocked/annoyed/frustrated group. The middle left section is the shocked/confused group. The bottom left section is the sad/sick/burnt out group. The top middle section is the happy group. The middle section is the funny group. The top right section is the flirty/lovey group. The middle right section is the goofy/silly group. The bottom right section is the annoyed/mad/ frustrated group.

(4) Group 1: water/tear drops, Group 2: rosy cheeks, Group 3: distinct eyes, Group 4: showing teeth, Group 5: dot eyes, Group 6: line for mouth/eyes, Group 7: tongues, Group 8: expressions w/ eyebrows, Group 9: emojis with extra features/accessories

(5) From left to right: Happy, Sad, Distressed, Miscellaneous.

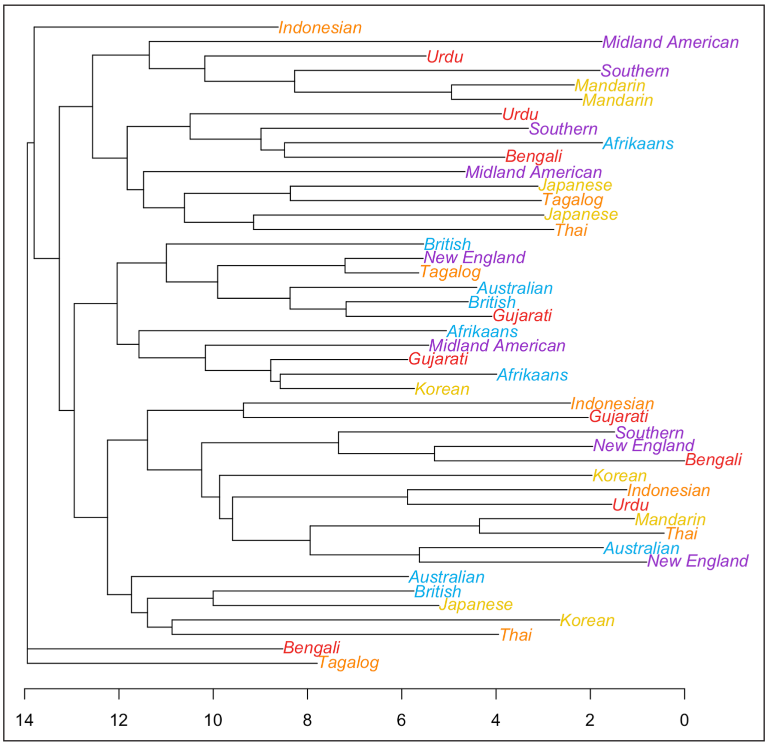


Figure 7: Non-listener emoji-sorting control group additive tree.

We now turn to our six listener groups. **Figure 8** shows the additive tree for the English monolingual listeners from a diverse linguistic context. The tree reveals two distinct clusters for the American English and International English categories (in the middle), which we predicted would happen. The listeners' ability to categorize the New England speakers together also reflects their familiarity with this dialect. They also grouped together the South Asian languages (at the top) and the East Asian languages (at the bottom), but struggled to group the Southeast Asian languages together, or consistently within an Asian group. These distinctions are entirely consistent with these listeners' exposure to specific Asian-language accents.

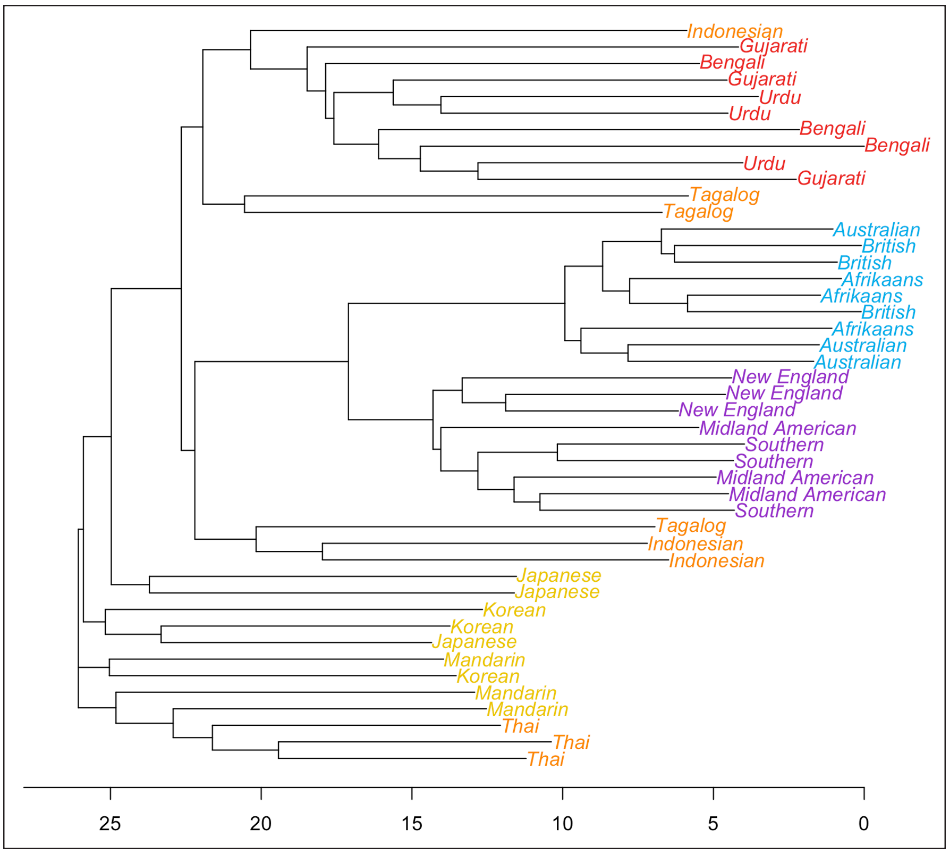


Figure 8: English monolingual group from diverse linguistic context additive tree.

Figure 9 shows the additive tree for the English monolingual listeners from a homogeneous context. The tree reveals a clear division between English and Asian-language-accented English, and within English, distinctions between American and International English dialects. However, these listeners failed to distinguish between Asian accents, and show no consistent subgroupings (though they do appear to diverge from the control group in **Figure 7**). In fact, even within American English dialects, these listeners struggled to make distinctions. Thus, the English monolingual group from a diverse linguistic context received an advantage in their ability to classify American English dialects *and* Asian-language-accented English through their consistent exposure to these accents.

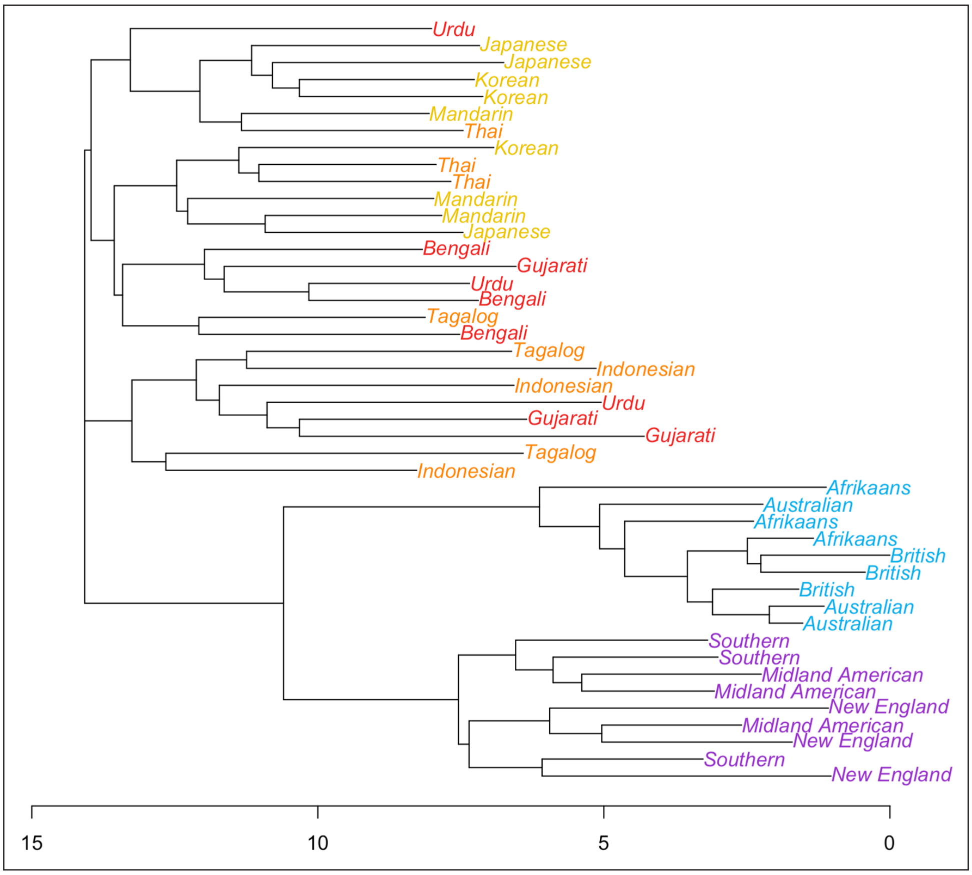


Figure 9: English monolingual group from homogeneous linguistic context additive tree.

We turn next to the tree for the non-Asian multilingual listeners in **Figure 10**. Like the first English monolingual group (from a diverse linguistic context), this group also successfully clustered American English, International English and South Asian language accents. However, they struggled to successfully categorize both the East and Southeast Asian language accents, and they also did not group together any American English dialects. Thus, in this direct comparison between **Figures 8–10** and their corresponding listener groups, we see clear and definite benefits of specific types of linguistic exposure.

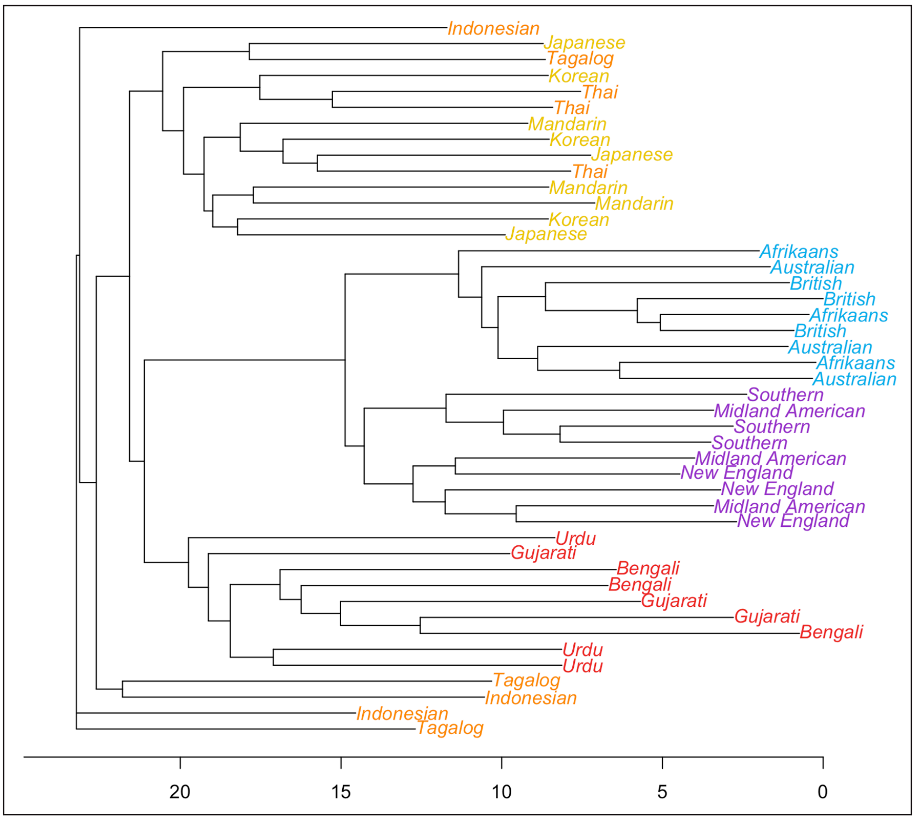


Figure 10: Non-Asian multilingual group additive tree.

We turn now to the listeners with an Asian language background. **Figure 11** presents the tree for the South Asian group. This group had a clear split between English language varieties (top) and Asian language accent varieties (bottom). In addition, both English groups belonged to distinct clusters (with the exception of one speaker of a New England dialect), although there was no grouping of dialects or varieties within either English group. Like the first English monolingual group, they grouped all of the South Asian speakers together, although they did not allow other languages to intrude in the cluster, and they grouped all of the Bengali speakers together, but unlike the English group, they mixed together the East and Southeast Asian speakers. Thus, their linguistic background afforded them the advantage of more accurately classifying South Asian language accents, but not classifying subgroups of Asian language accents in general, or perceiving distinctions within English dialects.

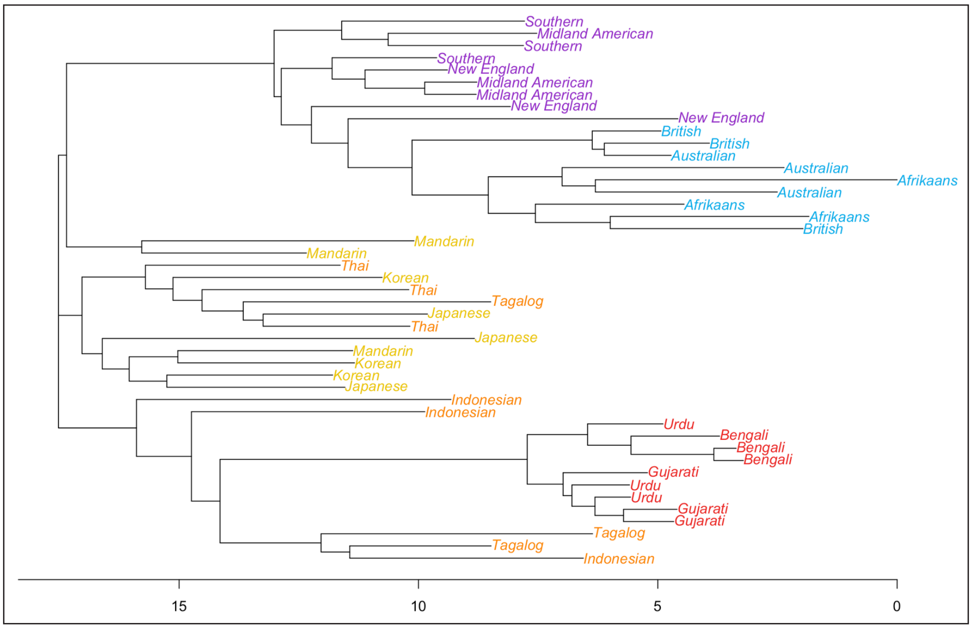


Figure 11: South Asian group additive tree.

Figure 12 presents the tree for the East Asian language group. They, too, successfully distinguished American English from International English. Additionally, the East Asian group showed some evidence of successful subgrouping clusters within the 15-category level. Perhaps unsurprisingly, they clustered all three Mandarin speakers together, and were somewhat able to distinguish between Japanese and Korean speakers. They also clustered six of the nine South Asian speakers together without distinguishing between the subgroups of speakers, and also successfully grouped together the Thai speakers. Rather surprisingly, they grouped all Southern American English speakers together, with one New England speaker and all Midland speakers together.

Finally, we turn to the Southeast Asian group in **Figure 13**. This group separated the English native speakers from the Asian L2 speakers, but unlike the other groups, mixed together the American and International English speakers. This group showed evidence of some success with smaller clusters of all three Asian groups, and was perhaps the most successful of all the listener groups in grouping together speakers of Southeast Asian languages (although not at a comparable level to how the other Asian listeners did with speakers of their language groups), and also did so with many of the South Asian speakers.

The results and visualizations of a two-dimensional multidimensional scaling (MDS) analysis complementing these additive trees are included in the supplementary data site.

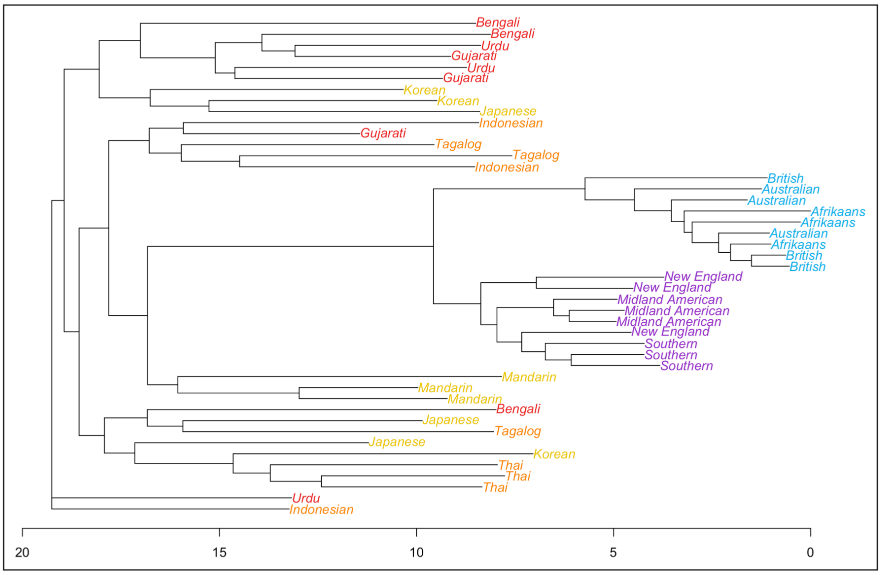


Figure 12: East Asian group additive tree.

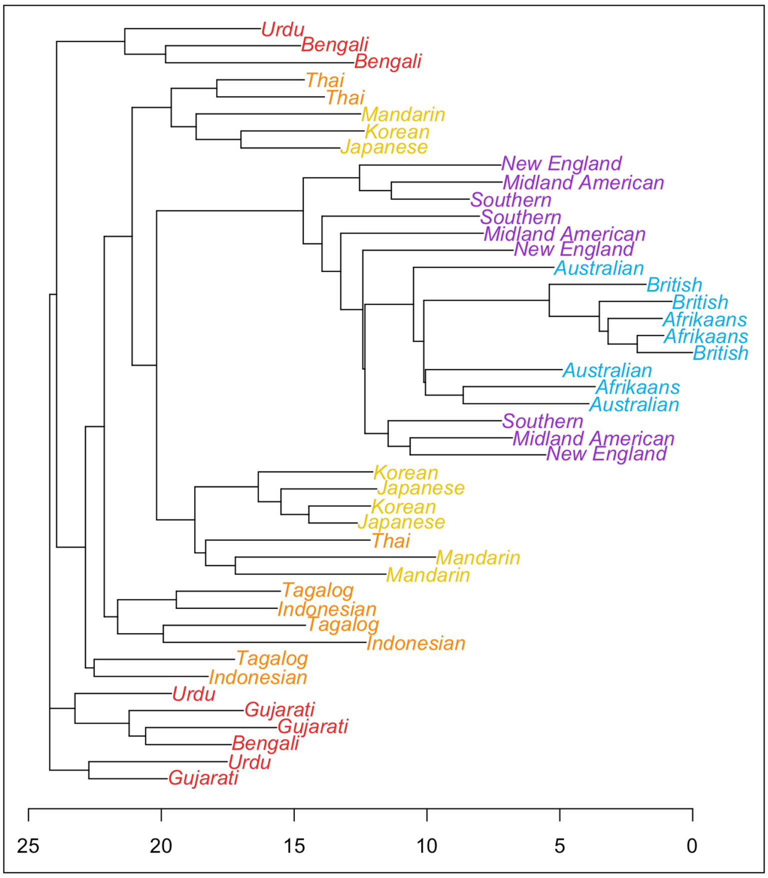


Figure 13: Southeast Asian group additive tree.

4. Discussion

We began with three main questions guiding our research, which our free classification experiment was designed to address. First, we asked if exposure to diverse languages and accents allows for an accurate, fine-grained categorization of speakers based on perception of their accents. Second, we asked if sheer multilingual status affords listeners enhanced perceptual skills and an ability to discriminate among, and categorize, different accents reflecting different dialects and language backgrounds, even for languages in which the listeners have no proficiency. Finally, we asked if status as a native speaker of a language allows listeners to successfully categorize not only speakers of that language, but also speakers of languages from the same geographic region (and, possibly, language family). To answer these questions, we enlisted listeners of six main groups: monolingual English speakers from a diverse linguistic context, monolingual English speakers from a more ethnically and linguistically homogeneous context, multilingual speakers of a Western European language, and speakers of Asian languages representing three regional subgroups (East Asian, South Asian, and Southeast Asian). We also included a control group that was asked to sort the corresponding images which the sound files were linked to, in order to verify that the classifications from our six listener groups were based on the auditory perception of speaker accents.

Our error rate and cluster analyses provide direct answers to each of these questions, at the same time as they expand our knowledge of how linguistic exposure and linguistic competence shape our perceptual organization of accents and speakers. First, all of the listener groups performed relatively comparably in the number of groups they created, and in the overall error rate (see **Figure 5**); there were no egregious differences. The one group that diverged from the others was the non-listener emoji-sorting control group. Their performance indicates that mere attention to the images alone did not lead to coherent classification strategies or clusters. Rather, dedicated categories emerged only when participants listened to the auditory stimuli linked to the images and created clusters based on perceived auditory similarity.

Second, the results from the English monolingual listeners from a linguistically diverse context, who are exposed to the South and East Asian accents featured in this experiment, reveal that such regular accent exposure does result in fewer errors and more accurate, fine-grained categorization. These listeners created, on average, more categories than the other listeners, and had the lowest predicted error rate for the 2-, 5-, and 15-category classifications (with the exception of the South Asian listeners for the 5-category error rate). This pattern stands in contrast to the English monolingual listeners from a linguistically homogeneous context and the multilingual listeners who are speakers of a non-Asian language. Moreover, they also outperformed the Southeast Asian listeners in many ways. Thus, concentrated exposure to diverse and particular accents supports more fine-grained categorization, whereas sheer knowledge of more languages, without specialized exposure to, or proficiency in, the target languages, does not support more accurate, fine-grained categorization. These results go beyond previous results (e.g., Atagi & Bent, 2015; Bayard et al., 2001; Clopper & Pisoni, 2004a, 2004b, 2006; a.o.) in showing that this pattern holds not only for the contrast between native English and L2 English speakers, and among English dialects, but also for distinctions *within* Asian-accented L2 speaker groups. Listeners need not be speakers of these languages to perform fine-grained classifications if they are familiar with that variety of accented English.

Third, the contrast between the non-Asian, Western European multilingual listeners and the other groups reveals that it is not enough to simply know multiple languages; this group was predicted to commit more errors in categorization than the Asian listeners, and did worse in creating subgroups than the English monolinguals from a diverse context, indicating that they lacked the requisite exposure and proficiency to achieve a rate of success comparable to these other groups. Still, they performed fairly well overall, and did not depart from the overall trend in the number and general types of categories created. All listener groups were able to differentiate between native and L2 English speakers, and also reliably distinguished between the American English and International English speakers to varying degrees – both as anticipated.

Finally, the comparison of the three groups of Asian language-speaking listeners revealed that language-specific knowledge appears to benefit listeners in categorization (as it did with the monolingual English speakers from a diverse context), and that geographic proximity and perceptual similarity appear to be influential drivers of category cohesiveness. However, it was not the case that each of the three groups did equally well in categorizing speakers across the three geographic regions. Rather, the listeners from the South Asian group were most successful in creating clusters of speakers from their own language group, which shares both geographic region and language family. These results raise questions about how languages in the other two Asian groups are perceived, and how they are assessed as clustering together.

There are two possible explanations for the differences among the Asian listeners. The first is that it is not uncommon for a speaker of a South Asian language to be familiar with, or speak, multiple South Asian languages, granting them additional currency with accent perception and discrimination, which can lead to more fine-grained and robust categorization. For example, a given speaker may know or be regularly exposed to Malayalam, Tamil, and Hindi, or to Urdu, Hindi, and Punjabi, or Hindi and Bengali. This was, in fact, the case for some of our listeners from this group, and for our monolingual English speakers from a diverse context exposed to South-Asian-accented English. Second, within the Southeast Asian group, there were more Tagalog-Filipino speakers than speakers of other languages, and only a few Thai speakers. This skew may explain why the Thai speakers were harder to classify, but it does not explain why the Indonesian and Tagalog speakers were not differentiated. None of this, however, speaks to why the East Asian listener group did not show a clear "own group" clustering signal. Future research should attempt to disentangle geographic proximity from the perceptual similarity of accent features.

Taken together, the results of this free classification study strongly suggest that listeners' own linguistic profile, including not only native language proficiency but also tailored and diverse exposure to dialects of the listeners' own language and to other specific languages, guides their perception and categorization of speakers and their accents. While listeners from a wide range of backgrounds successfully distinguished between native and L2 English speakers, their ability to perform more fine-grained categorization of subgroups beyond these two main groups depended heavily on their specialized exposure to, and proficiency in, specific accents and languages. We witnessed this difference within the two English groups, and within the Asian language groups. In the former, only those from a diverse linguistic background, exposed regularly to different English dialects and certain Asian-accented varieties of English, were able to successfully create corresponding subgroups. In the latter, only the South Asian listeners successfully distinguished those accents representing languages from the same language family spoken in the same geographical region, creating clusters reflecting this status.

Recall that our speech samples included two sentences and multiple intonational phrases, thereby providing listeners with segmental and suprasegmental information with which they could perform their free classification. We did not provide them with specific instructions about which aspects of speech to attend to, or how to group these speech samples – they were instructed only to group by similarity. It is, therefore, not possible to know *how*, exactly, speakers created their categories, or which aspects of the accented speech were more salient than others. An impressionistic analysis of the sound files (see the supplementary data site online) indicated that their classification could have been influenced by a number of factors: the overall fluency of native/L2 production, as manifested in rate of production and pauses, the presence or absence of plural morphology, epenthesis, clusters, and rhoticity. Together, these surface-level indicators of speaker status and language background may have contributed to the perception of groupings that resulted in our listeners' clusters. An exciting avenue for future research would be to disentangle these perceptual features and determine their weighting in classification strategies, perhaps even for listeners of different linguistic profiles.

While we did not ask our listeners to provide feedback about their classification process, or track their sorting in real time, we might ask what this process looks like, and whether all groups displayed the same strategy. One possibility is that listeners begin with broad brush-strokes, sorting speakers based on a native/L2 accent contrast, and then proceed to iteratively stipple in smaller clusters, based on factors such as perceived phonological similarity, geographic proximity, social factors and indexicality, and so on (see, e.g., Gnevsheva, 2018; McKenzie et al., 2019). Another is that they start by distinguishing one or more contrasting accents, and then expand and (re)organize these groups to create clusters based on perceived similarity, distance of accents, and sociocultural information about the speakers. Among these clusters, there could be a 'best exemplar' or (proto)typical member helping to anchor judgments. Online processing of cluster creation might reveal correlations between listener background and speaker/accent characteristics in cluster creation (see, e.g., Ruch, 2018).

One way in which our findings appear to deviate from previous research is that our listeners consistently distinguished between native English and L2-English Asian speakers. This result stands in contrast to that of Atagi and Bent (2013), who found that when native speakers were included, listeners did not reliably distinguish between native and L2 speakers. Their listeners frequently included L2 speakers within the native speaker group. However, this is where we may see an influence of the actual non-English languages that were chosen. All of our non-English L2 speakers represented an Asian language, whereas those in Atagi and Bent's (2013) study represented a range of languages (French, German, Spanish, Japanese, Korean, Mandarin). Thus, the contrast between accents representing English dialects and accents reflecting a more cohesive set of Asian language groups may have been distinct enough for all of our listeners to consistently produce two distinct groups, whereas in their study, the inclusion of non-English accents representing a wide range of language families and backgrounds may have blurred the boundaries between native and L2 speakers. Our findings may therefore align better with those of Flege (1984) and Park (2008), who contrasted native speakers with L2 speakers of one background (French and Korean, respectively).

At the same time, Atagi and Bent found that the inclusion of native speakers did not affect the classification of L2 speaker groups, and, like us, they also found that listeners grouped the accents of Japanese, Korean, and Mandarin speakers together, despite the fact that these languages do not have a common origin or phonological profile. Atagi and Bent suggested that more restrictive syllable structure may have been a key factor in the grouping of these languages together, and separate from the Western languages. Future research should probe further what guides listeners to group these three languages together, and what makes such a grouping distinct from other Asian languages, as we found in some categorization strategies in our study.

Our research findings also align with, and extend, those of Bent et al. (2016), who presented native American English listeners with longer speech samples representing a wide range of American English and International English accents, as well as a range of L2 accents. In their study, listeners recognized a distinct perceptual divide between native and L2 speakers. Further, listeners recognized the difference between American and International English speakers, and within the L2 speakers, between speakers of French and German, four Asian languages (Korean, Japanese, Mandarin, Thai), and a mixed bag of other non-English languages (Russian, Somali, Arabic, Spanish, Gujarati, and Swahili). Future research might consider presenting monolingual listeners of different linguistic backgrounds with L2 speakers of Asian languages alongside L2 speakers of non-Asian languages, as in Bent et al.'s (2016) study. Such comparisons will shed further light on the dual influence of a listener's personal linguistic profile and the implicit contrasts between perceptually distinct L2 speech accents.

Data accessibility statement

Supplemental data files associated with the manuscript may be accessed via https://osf.io/q52ys/?view_only=cdf78dc98d3449c382b51920c1bc5adf.

Ethics and consent

This research was conducted under the approved IRB protocol with the first author as PI. All authors/researchers have participated in CITI human subject training. All participants provided informed consent to participate, and were compensated either monetarily at a suggested rate, or granted course credit (1 point for each half hour of experimental participation).

Acknowledgements

We are grateful to the members of the Rutgers Laboratory for Developmental Language studies for their support in designing, piloting, and discussing this study. Tessa Bent provided helpful feedback in the early stages of the project. We are grateful to three anonymous reviewers for their substantive and thoughtful comments, which we appreciate immensely. All remaining errors are our own.

Competing interests

The authors have no competing interests to declare.

Authors’ contributions

A1 was responsible for project oversight, research activity and planning, mentorship of the research team, and overall execution. Participants were compensated through a University-internal research funds assigned to A1. A1 and A2 conceived of the study, identified and prepared the stimuli, developed the methodology, and conducted the review of the background literature. A2 collected the data with research assistants from A1’s lab, reviewed the data in consultation with A1, and wrote an initial summary of the data as part of their undergraduate honors thesis advised by A1. A1 and A3 conceptualized and planned the data analysis. A3 implemented the data analysis plan, performed the formal data analysis, wrote the experiment results section, and created the data visualizations, with consultation with A1. A1 was responsible for writing the original draft, and writing, reviewing, and editing of subsequent drafts.

ORCiD IDs

Kristen Syrett: https://orcid.org/0000-0002-3773-3035

Joy Lu

Kyle Parrish: https://orcid.org/0000-0001-8227-1370

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. DOI: https://doi.org/10.1037/a0013552

Alcorn, S., Meemann, K., Clopper, C. G., & Smiljanic, R. (2020). Acoustic cues and linguistic experience as factors in regional dialect classification. *Journal of the Acoustic Society of America*, *147*, 657–670. DOI: https://doi.org/10.1121/10.0000551

Atagi, E., & Bent, T. (2013). Auditory free classification of nonnative speech. *Journal of Phonetics*, *41*(6), 509–519. DOI: https://doi.org/10.1016/j.wocn.2013.09.003

Atagi, E., & Bent, T. (2015). Relationship between listeners' nonnative speech recognition and categorization abilities. *Journal of the Acoustical Society of America*, *137*(1), El44–El50. DOI: https://doi.org/10.1121/1.4903916

Atagi, E., & Bent, T. (2016). Auditory free classification of native and nonnative speech by nonnative listeners. *Applied Psycholinguistics*, *37*(2), 241–263. DOI: https://doi.org/10.1017/S014271641400054X

Baese-Berk, M. M., McLaughlin, D. J., & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass*, *14*(7), e12375. DOI: https://doi.org/10.1111/lnc3.12375

Baese-Berk, M. M., & Morrill, T. H. (2015). Speaking rate consistency in native and non-native speakers of English. *Journal of the Acoustical Society of America*, *138*(3), EL223–EL228. DOI: https://doi.org/10.1121/1.4929622

Baranowski, M. (2008). The fronting of the back upgliding vowels in Charleston, South Carolina. *Language Variation and Change, 20*(3), 527–551. DOI: https://doi.org/10.1017/S0954394508000136

Barkat, M., Ohala, J., & Pellegrino, F. (1999). Prosody as a distinctive feature for the discrimination of Arabic dialects. In *Proceedings of the 6th European Speech Communication and Technology (EUROSPEECH)*. https://www.isca-archive.org/eurospeech_1999/index.html. DOI: https://doi.org/10.21437/Eurospeech.1999-102

Bayard, D., Weatherall, A., Gallois, C., & Pittam, J. (2001). Pax Americana? Accent attitudinal evaluations in New Zealand, Australia and America. *Journal of Sociolinguistics*, *5*(1), 22–49. DOI: https://doi.org/10.1111/1467-9481.00136

Bent, T. (2014). Children's perception of foreign-accented words. *Journal of Child Language*, *41*(6), 1334–1355. DOI: https://doi.org/10.1017/S0305000913000457

Bent, T., Atagi, E., Akbik, A., & Bonifield, E. (2016). Classification of regional dialects, international dialects, and nonnative accents. *Journal of Phonetics*, *58*, 104–117. DOI: https://doi.org/10.1016/j.wocn.2016.08.004

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica*, *65*(3), 131–147. DOI: https://doi.org/10.1159/000144077

Best, C. T., & Strange, W. (1992). Effects of phonological and phonetic factors on cross-language perception of approximants. *Journal of Phonetics*, *20*(3), 305–330. DOI: https://doi.org/10.1016/S0095-4470(19)30637-0

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O. S. Bohn & M. J. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins. DOI: https://doi.org/10.1075/lllt.17.07bes

Blake, R., & Josey, M. (2003). The /ay/ diphthong in a Martha's Vineyard community: What can we say 40 years after Labov? *Language in Society*, *32*(4), 451–485. DOI: https://doi.org/10.1017/S0047404503324017

Bradlow, A. R., Clopper, C. G., Smiljanic, R., & Walter, M. A. (2010). A perceptual phonetic similarity space for languages: Evidence from five native language listener groups. *Speech Communication*, *52*(11–12), 930–942. DOI: https://doi.org/10.1016/j.specom.2010.06.003

Bradlow, A. R., & Pisoni, D. B. (1999). Recognition of spoken words by native and non-native listeners: Talker-, listener-, and item-related factors. *The Journal of the Acoustical Society of America*, *106*(4), 2074–2085. DOI: https://doi.org/10.1121/1.427952

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28. DOI: https://doi.org/10.18637/jss.v080.i01

Carrie, E., & McKenzie, R. M. (2018). American or British? L2 speakers' recognition and evaluations of accent features in English. *Journal of Multilingual and Multicultural Development*, *39*(4), 313–328. DOI: https://doi.org/10.1080/01434632.2017.1389946

Chambers, J. K. (2006). Canadian raising retrospect and prospect. *Canadian Journal of Linguistics-Revue Canadienne De Linguistique*, *51*(2–3), 105–118. DOI: https://doi.org/10.1017/S000841310000400X

Cheng, L. S., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*, 715843. DOI: https://doi.org/10.3389/fpsyg.2021.715843

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America*, *116*(6), 3647–3658. DOI: https://doi.org/10.1121/1.1815131

Clopper, C. G. (2008). Auditory free classification: Methods and analysis. *Behavior Research Methods*, *40*(2), 575–581. DOI: https://doi.org/10.3758/BRM.40.2.575

Clopper, C. G., & Bradlow, A. R. (2008). Perception of dialect variation in noise: Intelligibility and classification. *Language and Speech*, *51*(3), 175–198. DOI: https://doi.org/10.1177/0023830908098539

Clopper, C. G., & Bradlow, A. R. (2009). Free classification of American English dialects by native and nonnative listeners. *Journal of Phonetics, 37*(4), 436–451. DOI: https://doi.org/10.1016/j.wocn.2009.07.004

Clopper, C. G., Conrey, B., & Pisoni, D. B. (2005). Effects of talker gender on dialect categorization. *Journal of Language and Social Psychology*, *24*(2), 182–206. DOI: https://doi.org/10.1177/0261927X05275741

Clopper, C. G., Levi, S. V., & Pisoni, D. B. (2006). Perceptual similarity of regional dialects of American English. *Journal of the Acoustical Society of America*, *119*(1), 566–574. DOI: https://doi.org/10.1121/1.2141171

Clopper, C. G., & Pisoni, D. B. (2004a). Some acoustic cues for the perceptual categorization of American English regional dialects. *Journal of Phonetics*, *32*(1), 111–140. DOI: https://doi.org/10.1016/S0095-4470(03)00009-3

Clopper, C. G., & Pisoni, D. B. (2004b). Homebodies and army brats: Some effects of early linguistic experience and residential history on dialect categorization. *Language Variation and Change*, *16*, 31–48. DOI: https://doi.org/10.1017/S0954394504161036

Clopper, C. G., & Pisoni, D. B. (2006). Effects of region of origin and geographic mobility on perceptual dialect categorization. *Language Variation and Change*, *18*, 193–221. DOI: https://doi.org/10.1017/S0954394506060091

Clopper, C. G., & Pisoni, D. B. (2007). Free classification of regional dialects of American English. *Journal of Phonetics*, *35*(3), 421–438. DOI: https://doi.org/10.1016/j.wocn.2006.06.001

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America*, *118*(3), 1661–1676. DOI: https://doi.org/10.1121/1.2000774

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic Press. DOI: https://doi.org/10.4324/9780203771587

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, *34*(1), 104–137. DOI: https://doi.org/10.1016/j.wocn.2005.03.004

Derwing, T. M., & Munro, M. J. (1997). Accent, intelligibility, and comprehensibility: Evidence from four L1s. *Studies in Second Language Acquisition*, *19*, 1–16. DOI: https://doi.org/10.1017/S0272263197001010

Douglas, B. D., Ewell, P. J., & Brauer, M. (2023). Data quality in online human-subjects research: Comparisons between MTurk, Prolific, CloudResearch, Qualtrics, and SONA. *PLoS One*, *18*(3), e0279720. DOI: https://doi.org/10.1371/journal.pone.0279720

Flege, J. E. (1984). The detection of French accent by American listeners. *Journal of the Acoustical Society of America*, *76*(3), 692–707. DOI: https://doi.org/10.1121/1.391256

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*(4), 437–470. DOI: https://doi.org/10.1006/jpho.1997.0052

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1276–1293. DOI: https://doi.org/10.1037/0096-1523.32.5.1276

Foster, S. C., Stanley, J. A., & Renwick, M. E. (2017). Vowel mergers in the American South. *The Journal of the Acoustical Society of America*, *142*(4), 2540–2540. DOI: https://doi.org/10.1121/1.5014282

Fridland, V. (1999). The Southern shift in Memphis, Tennessee. *Language Variation and Change, 11*(3), 267–285. DOI: https://doi.org/10.1017/S0954394599113024

Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus. Linguistic Data Consortium. DOI: https://doi.org/10.35111/17gk-bn40

Gnevsheva, K. (2018). Variation in foreign accent identification. *Journal of Multilingual and Multicultural Development*, *39*(8), 688–702. DOI: https://doi.org/10.1080/01434632.2018.1427756

Harnsberger, J. D., Shrivastav, R., Brown, W. S., Jr., Rothman, H., & Hollien, H. (2008). Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of Voice*, *22*(1), 58–69. DOI: https://doi.org/10.1016/j.jvoice.2006.07.004

Hartshorne, J. K., Tenenbaum, J. B., & Pinker, S. (2018). A critical period for second language acquisition: Evidence from 2/3 million English speakers. *Cognition*, *177*, 263–277. DOI: https://doi.org/10.1016/j.cognition.2018.04.007

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods*, *48*, 400–407. DOI: https://doi.org/10.3758/s13428-015-0578-z

Hopp, H. (2013). The development of L2 morphology. *Second Language Research*, *29*(1), 3–6. DOI: https://doi.org/10.1177/0267658312465304

Hubbell, A. F. (1950). *The pronunciation of English in New York City: Consonants and vowels*. King's Crown Press. DOI: https://doi.org/10.7312/hubb94024

Hunter, E. J., Ferguson, S. H., & Newman, C. A. (2016). Listener estimations of talker age: A meta-analysis of the literature. *Logopedics Phoniatrics Vocology*, *41*(3), 101–105. DOI: https://doi.org/10.3109/14015439.2015.1009160

Imai, S. (1966). Classification of sets of stimuli with different stimulus characteristics and numerical properties. *Perception & Psychophysics*, *1*, 48–54. DOI: https://doi.org/10.3758/BF03207821

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. DOI: https://doi.org/10.1016/j.patrec.2009.09.011

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*(1), 60–99. DOI: https://doi.org/10.1016/0010-0285(89)90003-0

Labov, W. (1962). *The social history of a sound change on the island of Martha's Vineyard, Massachusetts* [Masters essay]. Columbia University.

Labov, W. (1966). *The social stratification of English in New York City*. Center for Applied Linguistics.

Labov, W. (1998). The three dialects of English. In M. D. Linn (Ed.), *Handbook of dialects and language variation* (pp. 39–81). Academic Press.

Labov, W., Ash, S., & Boberg, C. (2006) *The atlas of North American English*. Mouton de Gruyter. DOI: https://doi.org/10.1515/9783110167467

Lass, N. J., Hughes, K. R., Bowyer, M. D., Waters, L. T., & Bourne, V. T. (1976). Speaker sex identification from voiced, whispered, and filtered isolated vowels. *Journal of the Acoustical Society of America*, *59*(3), 675–678. DOI: https://doi.org/10.1121/1.380917

Leemann, A., Kolly, M.-J., Nolan, F., & Li, Y. (2018). The role of segments and prosody in the identification of a speaker's dialect. *Journal of Phonetics*, *68*, 69–84. DOI: https://doi.org/10.1016/j.wocn.2018.02.001

Lenneberg, E. (1967). *Biological foundations of language*. Wiley. DOI: https://doi.org/10.1080/21548331.1967.11707799

Leongómez, J. D., Mileva, V. R., Little, A. C., & Roberts, S. C. (2017). Perceived differences in social status between speaker and listener affect the speaker's vocal characteristics. *PloS One*, *12*(6), e0179407. DOI: https://doi.org/10.1371/journal.pone.0179407

Major, R. C. (2007). Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition*, *29*, 539–556. DOI: https://doi.org/10.1017/S0272263107070428

McCullough, E. A. (2015). Open-set identification of non-native talkers' language backgrounds. In *The Scottish Consortium for ICPhS 2015*, *Proceedings of the 18th International Congress of Phonetic Sciences*. University of Glasgow. Retrieved from https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0917.pdf

McCullough, E. A., & Clopper, C. G. (2016). Perceptual subcategories within non-native English. *Journal of Phonetics*, *55*, 19–37. DOI: https://doi.org/10.1016/j.wocn.2015.11.002

McCullough, E. A., Clopper, C. G., & Wagner, L. (2019). Regional dialect perception across the lifespan: Identification and discrimination. *Language and Speech*, *62*(1), 115–136. DOI: https://doi.org/10.1177/0023830917743277

McKenzie, R. M. (2015). The sociolinguistics of variety identification and categorisation: Free classification of varieties of spoken English amongst non-linguist listeners. *Language Awareness*, *24*(2), 150–168. DOI: https://doi.org/10.1080/09658416.2014.998232

McKenzie, R. M., Huang, M., Ong, T. T., & Snodin, N. (2019). Socio-psychological salience and categorisation accuracy of speaker place of origin. *Lingua*, *228*, 102705. DOI: https://doi.org/10.1016/j.lingua.2019.06.006

Merritt, B., Bent, T., Kilgore, R., & Eads, C. (2024). Auditory classification of gender diverse speakers. *Journal of the Acoustic Society of America*, *155*, 1422–1436. DOI: https://doi.org/10.1121/10.0024521

Moyse, E. (2014). Age estimation from faces and voices: A review. *Psychologica Belgica*, *54*(3), 255–265. DOI: https://doi.org/10.5334/pb.aq

Munro, M. J. (1995). Nonsegmental factors in foreign accent: Ratings of filtered speech. *Studies in Second Language Acquisition*, *17*(1), 17–34. DOI: https://doi.org/10.1017/S0272263100013735

Munro, M. J. (1998). The effects of noise on the intelligibility of foreign-accented speech. *Studies in Second Language Acquisition*, *20*(2), 139–154. DOI: https://doi.org/10.1017/S0272263198002022

Munro, M. J., & Derwing, T. M. (1995). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306. DOI: https://doi.org/10.1177/002383099503800305

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468. DOI: https://doi.org/10.1017/S0272263101004016

Munro, M. J., Derwing, T. M., & Burgess, C. S. (2010). Detection of nonnative speaker status from content-masked speech. *Speech Communication*, *52*, 626–637. DOI: https://doi.org/10.1016/j.specom.2010.02.013

Munson, B. (2007). The acoustic correlates of perceived masculinity, perceived femininity, and perceived sexual orientation. *Language and Speech*, *50*(1), 125–142. DOI: https://doi.org/10.1177/00238309070500010601

Munson, B., McDonald, E. C., Deboe, N. L., & White, A. R. (2006). The acoustic and perceptual bases of judgments of women and men's sexual orientation from read speech. *Journal of Phonetics*, *34*(2), 202–240. DOI: https://doi.org/10.1016/j.wocn.2005.05.003

Park, H. (2008). *Phonological information and linguistic experience in foreign accent detection* [Doctoral dissertation]. Indiana University.

Park, H. (2013). Detecting foreign accent in monosyllables: The role of L1 phonotactics. *Journal of Phonetics*, *41*(2), 78–87. DOI: https://doi.org/10.1016/j.wocn.2012.11.001

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. DOI: https://doi.org/10.1111/lang.12079

Preston, D. (1993). *American dialect research*. John Benjamins. DOI: https://doi.org/10.1075/z.68

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. URL https://www.R-project.org/

Rogers, C. L., Dalby, J. M., & Nishi, K. (2001). Effects of noise and proficiency level on intelligibility of Chinese-accented English. *The Journal of the Acoustical Society of America*, *109*(5), 2473–2473. DOI: https://doi.org/10.1121/1.4744783

Rogers, C. L., Dalby, J., & Nishi, K. (2004). Effects of noise and proficiency level on intelligibility of Chinese-accented English. *Language and Speech, 47*, 139–154. DOI: https://doi.org/10.1177/00238309040470020201

Ruch, Hanna. (2018). The role of acoustic distance and sociolinguistic knowledge in dialect identification. *Frontiers in Psychology*, *9*. DOI: https://doi.org/10.3389/fpsyg.2018.00818

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425

Sankoff, G., & Blondeau, H. (2007). Language change across the lifespan: /r/ in Montreal French. *Language*, *83*(3), 560–588. DOI: https://doi.org/10.1353/lan.2007.0106

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345. DOI: https://doi.org/10.1007/BF02293654

Sereno, J., Lammers, L., & Jongman, A. (2016). The relative contribution of segments and intonation to the perception of foreign-accented speech. *Applied Psycholinguistics, 37*(2), 303–322. DOI: https://doi.org/10.1017/S0142716414000575

Trudgill, P. W., & Hannah, J. (2002). *International English: A guide to the varieties of standard English*. Arnold.

Van Bezooijen, R., & Gooskens, C. (1999). Identification of language varieties: The contribution of different linguistic levels. *Journal of Language and Social Psychology*, *18*(1), 31–48. DOI: https://doi.org/10.1177/0261927X99018001003

van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication*, *35*(1–2), 103–113. DOI: https://doi.org/10.1016/S0167-6393(00)00098-4

Vieru, B., Boula de Mareueil, P., & Adda-Decker, M. (2011). Characterisation and identification of non-native French accents. *Speech Communication*, *53*(3), 292–310. DOI: https://doi.org/10.1016/j.specom.2010.10.002

Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds. *Phonetica*, *64*(2–3), 122–144. DOI: https://doi.org/10.1159/000107913

Weinberger, S. (2015). Speech Accent Archive. Retrieved from http://accent.gmu.edu

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Use R! Series. Springer. DOI: https://doi.org/10.1007/978-3-319-24277-4

Williams, A., Garrett, P., & Coupland, N. (1999). Dialect recognition. In D. R. Preston (Ed.), *Handbook of perceptual dialectology* (pp. 345–358). John Benjamins. DOI: https://doi.org/10.1075/z.hpd1.29wil

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., & Jaeger, T. F. (2018). Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker. *The Journal of the Acoustical Society of America*, *143*(4), 2013–2031. DOI: https://doi.org/10.1121/1.5027410

Typesetting queries

1. If possible, could you please provide ORCID id of the author "Joy Lu"?