Comparing non-native and native speech: Are L2 productions more variable?

Xin Xie, and T. Florian Jaeger

Citation: The Journal of the Acoustical Society of America **147**, 3322 (2020); doi: 10.1121/10.0001141
View online: https://doi.org/10.1121/10.0001141
View Table of Contents: https://asa.scitation.org/toc/jas/147/5
Published by the Acoustical Society of America

JASA ARTICLE

Comparing non-native and native speech: Are L2 productions more variable?

Xin Xie[a] and T. Florian Jaeger
Department of Brain and Cognitive Sciences, University of Rochester, Rochester, New York 14627, USA

ABSTRACT:
Foreign-accented speech of second language learners is often difficult to understand for native listeners of that language. Part of this difficulty has been hypothesized to be caused by increased within-category variability of non-native speech. However, until recently, there have been few direct tests for this hypothesis. The realization of vowels and word-final stops in productions of native-English L1 speakers and native-Mandarin speakers of L2 English is compared. With the largest sample size to date, it is shown that at least proficient non-native speakers exhibit little or no difference in category variability compared to native speakers. This is shown while correcting for the effects of phonetic context. The same non-native speakers show substantial deviations from native speech in the central tendencies (means) of categories, as well as in the correlations among cues they produce. This relativizes a common and *a priori* plausible assumption that competition between first and second language representations necessarily leads to increased variability—or, equivalently, decreased precision, consistency, and stability—of non-native speech. Instead, effects of non-nativeness on category variability are category- and cue-specific.
 https://doi.org/10.1121/10.0001141
(Received 30 October 2019; revised 30 March 2020; accepted 6 April 2020; published online 8 May 2020)
[Editor: Benjamin V. Tucker] Pages: 3322–3347

I. INTRODUCTION

Non-native speech differs from native speech along many dimensions. These deviations from native speech are known to cause perception difficulty to native listeners. Previous work has compared non-native to native pronunciations mainly in terms of differences in the central tendencies (means) of phonological categories, i.e., *where* in the phonological space categories fall (e.g., Bohn and Flege, 1992; Darcy and Krüger, 2012; Fabra and Romero, 2012; Flege, 1987). More recent work has begun to investigate non-nativeness in the *distribution* of categories (e.g., Smith *et al.*, 2019; Vaughn *et al.*, 2019). These distributions—in particular categories' variabilities—affect, for example, how much neighboring categories overlap. For native speech, this is known to affect speech recognition (e.g., Clayards *et al.*, 2008; Nixon *et al.*, 2016): higher variability—or, equivalently, lower precision—in the realization of categories is associated with reduced intelligibility (e.g., Newman *et al.*, 2001; Mou *et al.*, 2018; Romeo *et al.*, 2013).

This raises the question of whether non-native speech exhibits increased category variability compared to native speech, contributing to its reduced intelligibility (for early discussion of this idea, see Jongman and Wade, 2007; Oh *et al.*, 2008). Specifically, it is uncontroversial that non-native speech tends to exhibit increased variability *across* talkers compared to native speech due to individual differences in L2 proficiency (Flege *et al.*, 1997; Wade *et al.*, 2007, among others). Less clear is whether non-native speech deviates from native speech *within*-talker, specifically in the distributional realization of categories, i.e., the orientation and magnitude of categories' dispersion in the phonetic space. Across different literatures, this type of within-talker variability is variously referred to as a speaker's "internal consistency" (Newman *et al.*, 2001), "intra-talker variability" (Romeo *et al.*, 2013; Smith *et al.*, 2019), "compactness" (Kartushina and Frauenfelder, 2014), "stability" (Kartushina *et al.*, 2016), "within-category dispersion" (Hazan *et al.*, 2013), or "group-level within-speaker variability" (Vaughn *et al.*, 2019).

This is the question we address here. There are several *a priori* reasons to consider it plausible that non-native speakers exhibit increased token-to-token variability in the production of a category. Non-native speakers inevitably face competition between L1 and L2, which may reduce stability of production compared to native speakers (e.g., Antoniou *et al.*, 2012; Goldrick *et al.*, 2014; Olson, 2013). Indeed, non-native speech tends to exhibit more speech errors, especially for L2 sounds that are not present in their native phonology (James, 1984), consistent with the notion that the degree of automaticity, as well as phonetic competition, differs between native and non-native speech. Unfamiliar L2 features that are not present in non-native speakers' L1 (or present but not in the same phonological context) may increase the difficulty of motoric control due to lack of practice, which in turn reduces production precision and increases variability. In addition, the speech input L2 speakers receive—often produced by other L2 learners, especially during the starting stages for classroom learners—may present greater perceptual confusion, which might further exacerbate the learning situation (e.g., Flege and Liu, 2001; Flege and MacKay, 2004). A lack of

[a]Electronic mail: xxie13@ur.rochester.edu

 0001-4966/2020/147(5)/3322/26/$30.00

https://doi.org/10.1121/10.0001141

perceptual sensitivity to fine phonetic differences between L2 categories that are absent in the learner's L1 may further contribute to expanded category representations, resulting in greater category variability for each of the categories (for discussion of these and related arguments, see Smith *et al.*, 2019; Vaughn *et al.*, 2019).

Whatever its cause, increased within-talker category dispersion would be expected to limit the intelligibility of a non-native talker *even for someone with perfect knowledge of that talker's speech characteristics* (Newman *et al.*, 2001; Kleinschmidt and Jaeger, 2015). Indeed, increased within-talker category dispersion has been evoked in explanations of non-native speech perception, foreign accent comprehension, and L1 acquisition in a multilingual environment (e.g., Wade *et al.*, 2007; Romero-Rivas *et al.*, 2015; Witteman *et al.*, 2014; Schmale *et al.*, 2011). It was not until recently, however, that studies began to systematically investigate differences in category dispersion between native and non-native talkers (Smith *et al.*, 2019; Vaughn *et al.*, 2019, summarized further below). Here, we build on these pioneering studies and assess how non-native speech differs from native speech in the realization of categories beyond their central tendencies. Assumptions about L2 learning—no matter how seemingly intuitive—are arguably particularly deserving of empirical evaluation given that they can come to affect preferences or even policies pertaining to pedagogical approaches to L2 instruction (for relevant discussion, see Berthele, 2019; Thomson and Derwing, 2014).

A. The present study

We present two studies based on the production data from a new database of native American English (L1 speech) and Mandarin-accented American English (L2 speech). Compared to previous works, the present studies increase the number of observations per speaker about two- to threefold (study 1) and tenfold (study 2). While common in research on speech production due to the inherently time-consuming nature of phonetic annotation, small sample sizes are particularly problematic for research on variability: robust estimation of category variability requires more data than estimation of category means.

Going beyond most previous work, we compare native and non-native pronunciations for both vowels (study 1) and consonants (word-final stop voicing, study 2) using data from the same talkers for both studies. Our decision to study both vowel and final stop categories (and to do so for L1 Mandarin L2 English) was motivated by their relevance to theories of L2 speech production. Specifically, we have three aims.

First, heeding a call by Vaughn *et al.* (2019), we draw on theories of L2 speech perception and production to derive predictions about the extent to which different categories are expected to differ between native and non-native speech. Previous work has largely asked whether non-native speech exhibits greater within-talker variability for *all* categories or across *all* categories (henceforth, the *across-the-board* hypothesis). This has led to mixed results. Some studies found evidence for greater within-talker variability in non-native compared to native speakers (Wade *et al.*, 2007; Baese-Berk and Morrill, 2015; Jongman and Wade, 2007). Others found no difference in the variability of native and non-native speech (Smith *et al.*, 2019; Vaughn *et al.*, 2019). These studies differed both in terms of the specific categories and the L1-L2 combinations investigated—differences that theories of L2 speech perception and production predict to matter. Specifically, L2 theories have long held that effects of non-nativeness are category specific or even cue specific. For instance, L2 categories and contrasts that do not exist in the speaker's L1 are predicted to be more strongly affected in non-native speech (Best, 1994; Escudero, 2005; Flege, 1995).

Vaughn and colleagues recognized this disconnect between the across-the-board hypothesis and theories of L2 production (Vaughn *et al.*, 2019, p. 28): "The relationship between the linguistic features in the native and target languages may have important consequences for the patterns of variability observed for native and non-native speakers. Carefully considering the factors contributing to variable productions of a given linguistic feature in a given language is an important piece of understanding variability in native and non-native speech." Following Vaughn and colleagues' (Vaughn *et al.*, 2019) call, we thus test both the across-the-board hypothesis (for comparability to previous work) and more nuanced hypotheses that the effects of non-native speech on variability are vowel specific or even cue specific, depending on the phonological inventory of both L1 and L2 categories.

Second, we extend previous work by focusing on phonetic features that exist in the L2 but not in the non-native speakers' L1. Previous work has exclusively examined features that are used in both L1 and L2 but differ in how they are used (e.g., formants of vowels). In study 2, we examine the production of word-final stop voicing among L1 Mandarin learners of English. The markers of word-final stop voicing involve acoustic features that are present in Mandarin as well as those that are absent. We test whether L1 influences on non-native speech are cue specific, differing between cues that non-native speakers are familiar with from their L1 and cues that are unfamiliar.

Our third aim is to more comprehensively characterize the distributions of the investigated phonetic categories in native and non-native speech. While our primary interest is in category variability, studies 1 and 2 compare several potential sources of differences between native and non-native productions: differences in the central tendency of categories (means), the magnitude of within-category variability, and—for the first time for non-native speech—*within*-category covariation among multiple phonetic cues.[1] As we demonstrate in studies 1 and 2, the consideration of within-category covariation allows us to better understand effects of non-nativeness on the *magnitude* of category variability in the context of the category's *orientation* in the multidimensional phonetic space. Indeed, one of the

take-home points of the present study is that an understanding of the former requires reference to the latter. (To avoid terminological confusion, we introduce below the term category *dispersion* as an umbrella term to refer to both the *magnitude* of the category's variability along any or all of its phonetic cues and the *orientation* of the category's expansion in that multidimensional cue space. We continue to use the term category variability to refer specifically to the magnitude of dispersion, which has constituted the focus of previous work). By comparing non-native to native speech along all three major distributional properties of categories (cue means, variance, and correlations), we aim to contribute to a fuller understanding of how these three factors might contribute to the intelligibility of non-native speech. This is an important prerequisite, in particular, for the future development and testing of computational models of accent adaptation, which link perception to cue distributions in native and non-native productions (for review, see Kleinschmidt and Jaeger, 2015).

II. STUDY 1

Study 1 compares American English vowel productions by native speakers of American English and L1 Mandarin non-native speakers of English. We ask whether non-native speakers exhibit greater within-category variability than native speakers, in general, and whether this increase, if any, varies as a function of differences in L1 and L2 vowel phonology. There is broad agreement that non-native speakers' phonetic representations are affected by interference from their native language (e.g., L2LP (Second Language Linguistic Perception), Escudero, 2005; PAM (Perceptual Assimilation Model), Best, 1994; SLM (Speech Learning Model), Flege, 1995).

Mandarin has a much smaller vowel inventory than English. In acquiring English, L1 Mandarin speakers, thus, must learn a number of new category contrasts that do not exist in their native language. The choice of L1 Mandarin L2 English thus facilitates comparison to previous work, which has similarly focused on the acquisition of more complex L2 vowel systems by native speakers of languages with less complex systems (e.g., L1 Spanish-L2 English, Wade *et al.*, 2007; Flege, 1989; L1 Mandarin-L2 English, Smith *et al.*, 2019). This creates the type of learning difficulty that is hypothesized to cause increased within-talker variability of non-native productions: vowel inventory size impacts the dispersion of vowel categories (e.g., due to adaptive dispersion, Liljencrants and Lindblom, 1972; although see Bradlow, 1995; Flege, 1989); therefore, non-native speakers whose L1 has a smaller vowel inventory are expected to have more variable L2 vowel productions (see Vaughn *et al.*, 2019, for a test in the opposite direction: from a crowed vowel system in L1 English to a more sparse system in L2 Japanese). We distinguish two hypotheses about L1-to-L2 influence on the within-category variability of non-native production.

First, we test a basic hypothesis that non-native speech is *generally* more variable, regardless of the specific category—for example, because of continued competition between L1 and L2 representations during production. We call this the *across-the-board* hypothesis. The across-the-board hypothesis would be compatible with findings that the degree of variability in non-native speech depends on the specific L1 and L2, but it does not make predictions for specific categories or cues. This hypothesis has been tested in recent work on speaking rate (Baese-Berk and Morrill, 2015), vowel production (Smith *et al.*, 2019; Vaughn *et al.*, 2019), and production of syllable-initial stops (Vaughn *et al.*, 2019).

Second, we consider a more nuanced hypothesis that predicts the effect of non-nativeness to be category specific. Under this hypothesis, we derive specific predictions from theories on L2 perception and production. This follows Vaughn and colleagues' (Vaughn *et al.*, 2019) call for category-specific predictions and analyses.

We begin with a brief summary of differences between English and Mandarin vowel inventories, focusing on nine English vowels that are well-represented in our database: /i, ɪ, æ, ɛ, ʌ, ɑ, ɔ, ʊ, u/. These vowels cover the entire American English vowel space with the exception of schwa-like center vowels and diphthongs. Figure 1 presents a comparison of simple vowels in English and Mandarin in the F1-F2 space. English has four front vowels /i, ɪ, ɛ, æ/, two mid vowels /ɝ, ʌ/, and four back vowels /u, ʊ, ɔ, ɑ/, all varying in height.[2] Researchers generally agree that Mandarin has a six-vowel system: three high vowels (front: /i, y/; back: /u/), one low vowel /a/, and two mid vowels. Comparing to English, which has only two high vowels /i, u/, Mandarin additionally has a high front rounded /y/. Comparing to English, which has two low vowels /æ, ɑ/, Mandarin has a single low vowel /a/ (which can be realized as allophonic variants [ɑ, a]). The exact description of mid vowels in Mandarin remains controversial. Some researchers suggest Mandarin has a mid-central vowel /ə/ and a mid-back vowel /ɤ/ (e.g., Jia *et al.*, 2006), while others adopt a distinction between unrounded mid vowel /ɤ/ and a rounded mid vowel /o/ (e.g., Mou *et al.*, 2018). The realization of Mandarin vowels is highly dependent on the phonetic contexts. For the current purpose, it is sufficient to note that Mandarin has a somewhat underspecified mid-central vowel (Wiese, 1997; we use /ɤ/ to indicate this vowel, following Mou *et al.*, 2018) whose phonetic realizations may include allophonic variants similar to English [e, æ, ɛ] and a rounded mid vowel whose variations are similar to English [ɔ].

Theories of L2 speech production and perception tend to agree that the production of an L2 category in non-native speech is affected by the category's relative placement within both the L2 and the L1 phonological inventory, as well as the mapping of phonological categories to phonetic cues in the L1 and L2 (Best, 1995; Best and Tyler, 2007; Escudero, 2005; Flege, 1995). The strongest tests of these theories are beyond the scope of the present work as they would require three critical components that are as of yet not available in the necessary

https://doi.org/10.1121/10.0001141

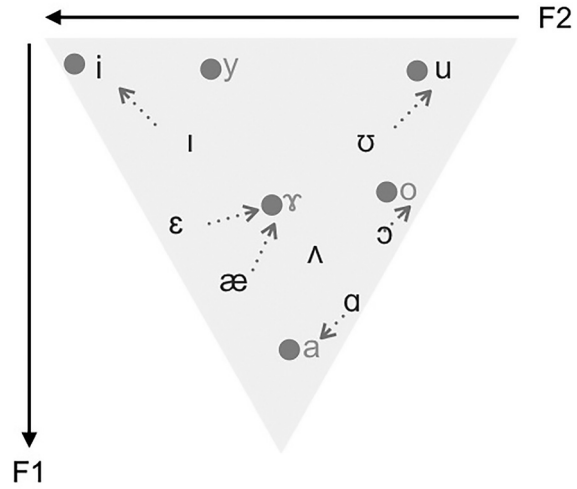

FIG. 1. A comparison of English (black) and Mandarin (grey; dotted) monophthong vowels. The arrows indicate predicted assimilation patterns of English vowels into Mandarin vowels.

combination: (1) computational *models* that integrate fully specified learning hypotheses with (2) articulatory constraints and (3) quantitative phonetic information about both the L1 and L2 (i.e., for the present study, data are from both English *and* Mandarin).

Instead, the present work aims to take a modest step toward such tests. To this end, we derive qualitative predictions about category-specific differences in the variability of non-native and native speech. We follow the common simplifying assumption that non-native pronunciations of L2 categories are primarily affected by the phonologically most similar category in the speakers' L1 (e.g., L2LP, Escudero, 2005; SLM, Flege, 1995; PAM-L2, Best and Tyler, 2007; we revisit this assumption later). Theories of L2 phonological acquisition predict learning difficulty—and thus delayed or unsuccessful category formation—for L2 categories that are "poor exemplars" of the L1 categories the learner knows (e.g., Best, 1995). Inherent in the notion of a poor exemplar is the notion of the closest neighbor: L2 categories are poor exemplars of an L1 category (their closest L1 neighbor) when, on the one hand, they are "close enough" to the L1 category in phonetic space and, yet, on the other hand, they only partially overlap with the L1 category. For native-Mandarin L2 learners of English, this applies to four vowels (/æ, ɛ, ʊ, ɪ/): /ʊ/ is a poor exemplar of Mandarin /u/; /ɪ/ is a poor exemplar of Mandarin /i/; /æ/ and /ɛ/ are both poor exemplars of the unrounded mid vowel Mandarin /ɤ/.[3] In contrast, L2 categories are predicted to lead to less learning difficulty in two scenarios: either they are good exemplars of the closest L1 neighbor or there is no nearby L1 neighbor to the L2 category (*uncategorized assimilation* in the terminology of PAM). For L1 Mandarin L2 learners of English, this applies to the remaining five vowels in our data /ɑ, ɔ, i, u, ʌ/: /ɑ/, /ɔ/, /i/, and /u/ are good exemplars of Mandarin /a/, /o/, /i/, and /u/, respectively; /ʌ/ has no nearby Mandarin neighbors.[4]

Following previous work (e.g., Flege, 1995; Vaughn *et al.*, 2019; Bosch and Ramon-Casas, 2011; Kartushina and Frauenfelder, 2014), we hypothesize that difficult-to-learn categories show increased variability in production—either because these categories take more learning to stabilize or because the same factors that impede learning also impede production (e.g., competition between L1 and L2 categories). As these predictions are generated based on assimilation patterns of L2 categories into the closest L1 categories, we refer to this hypothesis as the *closest-neighbor* hypothesis. Table I summarizes our predictions.

Before we describe our database, we emphasize again that the closest-neighbor hypothesis is best thought of as a first step toward testing the more general idea that the production of an L2 category is affected by its placement relative to surrounding L1 categories and their phonetic realizations. For example, SLM (Flege, 1995, 2007) predicts that all L1 and L2 sounds are represented in a shared phonetic space. As L2 learners acquire new categories, "their combined L1-L2 phonetic space becomes more crowded than that of monolingual speakers of either the L1 or the L2" (Flege, 2007, p. 359). To the extent that L2 learners aim to avoid overlap between L1 and L2 categories, this predicts that multiple L1 categories that are surrounding an L2 category—rather than just the closest neighbor—might come to affect the realization of that L2 category. Indeed, we find that several findings of studies 1 and 2 seem to be best understood in light of this more general hypothesis.

A. Caveat emptor: A note on phonetic context

An anonymous reviewer raised an important potential confound for the comparison of non-native and native speech: the phonetic realization of vowels in F1-F2 space is known to be affected by the surrounding phonetic context, and it is possible that non-native speakers do not exhibit these context effects to the same extent as native speakers. For the comparison of category means, a failure to take into account this possibility might obscure whether non-nativeness in the central tendencies of categories is general

TABLE I. Expected patterns of differences in category variability for vowel categories between native (N) and non-native (NN) speech [$\Delta\sigma$ (NN-N)]. For details, see the text. The last column indicates whether we expect greater or equal variability in NN speech compared to N speech.

L2 category	Closest L1 neighbor	Status of the target L2 category	Prediction $\Delta\sigma$ (NN-N)
ʌ		Not similar to any L1 categories	Equal variability
ɑ	a	Good exemplar	Equal variability
ɔ	o	Good exemplar	Equal variability
æ → ɤ	ɤ	**Poor exemplar**	**Greater variability**
ɛ → ɤ	ɤ	**Poor exemplar**	**Greater variability**
i → i	i	Good exemplar	Equal variability
ɪ → i	i	**Poor exemplar**	**Greater variability**
u → u	u	Good exemplar	Equal variability
ʊ → u	u	**Poor exemplar**	**Greater variability**

across contexts or driven by how non-native speakers realize phonetic context effects. For the comparison of category dispersion, a failure to take into account phonetic context might even confound the comparison of non-native to native speech: specifically, if non-native speakers exhibit smaller context effects than native speakers, a failure to account for this will artificially inflate the variability of native speech (as it conflates systematic variability caused by context effects with random variability caused by, e.g., motor noise). This would make it harder to detect increased variability of non-native speech or might even cause non-native speech to appear *less* variable than native speech.

Neither previous work (see Smith *et al.*, 2019; Vaughn *et al.*, 2019) nor the present study were designed to address this question. There are, however, at least two reasons that lend credence to the possibility that non-native speech exhibits reduced context effects. First, at least some effects of phonetic context are phonologized (e.g., Lahiri and Marslen-Wilson, 1991) and thus would have to be learned. It is plausible that (some) non-native speakers would not (yet) have learned these context-specific effects on the articulation of vowels. Second, non-native speakers might react differently to being recorded, choosing more careful speech registers. In more careful speech, coarticulatory effects would be reduced (e.g., Moon and Lindblom, 1994).

If not taken into account, differences in category means due to phonetic context can therefore lead to inflated estimates of category variability.[5] Unfortunately, there is no trivial way to control for phonetic context effects (a token-level predictor) in the analysis of category dispersion [i.e., the talker-level standard deviation (SD) of the category]. The analyses presented in this paper thus ignore potential effects of context. However, in the supplementary material, we present detailed *post hoc* analyses of phonetic context effects in native and non-native speech. These supplementary analyses leave our central findings unchanged. They do, however, suggest that two surprising results we obtain below are at least in part due to context effects. Wherever the supplementary analyses deviate from those in the main text, we explicitly note so in the discussion of study 1.

B. Materials

We recorded productions of 180 English words from 10 native-English male speakers and 10 native-Mandarin male speakers. All Mandarin speakers were native speakers of Mandarin (although they differ in their regional dialects; see Appendix B) and L2 speakers of English. All Mandarin and English speakers (ages 18–35 years old) were students at U.S. universities. Although we aimed to recruit a comparable sample of native and non-native talkers (gender and age-matched college students), we did not collect dialect information for our native speakers. In order to derive dialect-specific predictions, we would need an even larger sample of talkers *and* a phonological description of the vowel system of each dialect, neither of which were available to us. All Mandarin speakers acquired English as an L2 in classroom settings prior to coming to the U.S. At the time of the recording, they were students at a university in the northeastern U.S. (University of Connecticut). For all of the recorded speakers, this was also the first immersive English-speaking environment. The length of residence in the U.S. ranged from five months to five years ($M = 2.3$ yr). The age of arrival ranged from 15 to 26 years of age ($M = 19$ yr). A full list of stimuli is provided in Appendix A. These words were taken from a previous study (Weil, 2003). Each word contained at least one phoneme that is known to be difficult for L1 Mandarin speakers. Each word was recorded three times by each speaker.

For study 1, we selected all 125 words (each recorded 3 times for each speaker) that contained 1 of the 9 vowels. The first row of Table II shows the total number of tokens recorded for each of the nine vowel categories for each speaker.

The onset and offset of the vowels were first marked automatically by ForcedAligner (P2FA, Yuan and Liberman, 2008) and then manually corrected by phonetically trained experimenters. Vowel formant values were obtained using Praat (Boersma and Weenik, 2018) and GSU Praat tools (Orwen, 2011). The analysis window was 50 ms. For each token, we extracted the mean F1 and F2 over the middle 50% of the vowel. Following conventions (e.g., Gahl *et al.*, 2012), tokens with mean formant values more than 2.5 SDs away (~7% of all tokens) from the speaker- and vowel-specific means were manually checked for annotation mistakes: where possible, formants for such tokens were measured by hand. This was most commonly required for vowels produced in glide or nasal contexts (e.g., *glean*, *room*). Following past work, 33 tokens for which formant values could not be accurately obtained were removed (e.g., Gahl *et al.*, 2012). When the vowel of a word was clearly mispronounced by a speaker, we excluded all three recordings of that word from the analysis. For Mandarin speakers, these mispronunciations appeared to reflect lexical unfamiliarity with irregular spellings (e.g., *lose* pronounced as *loss* or *loz*; see Allen and Miller, 1999; Stibbard, 2004). Only a small number of tokens (63, 1.3%) were removed for this reason (all 3 tokens from 8 words across L1 speakers and 13 words across L2 speakers). Exclusion rates differed marginally between the two groups of speakers ($\chi^2 = 3.12$, $p = 0.078$). This left a total of 7404 tokens for analysis (3707 tokens from English speakers, 3697 tokens from Mandarin speakers).

TABLE II. Number of tokens for each vowel category available for analysis for N and NN speakers. The last two rows indicate the number of tokens included in the analysis after data exclusion.

Vowel	ʌ	ɑ	ɔ	æ	ɛ	i	ɪ	u	ʊ	Total
Tokens per speaker	48	48	21	60	30	63	60	27	18	375
Total tokens (N speech)	471	477	207	599	300	624	586	268	175	3707
Total tokens (NN speech)	477	480	210	589	297	623	584	263	174	3697

https://doi.org/10.1121/10.0001141

On average, there are 41.2 tokens per category per speaker from English speakers and 41.1 tokens per category per speaker from Mandarin speakers. The last two rows of Table II show the number of vowel tokens available for analysis: /i/ has the most tokens per speaker ($n = 63$) and /ʊ/ has the least ($n = 18$). Below, we present results using Lobanov-normalized F1 and F2, following previous work on category variability (Vaughn *et al.*, 2019). Lobanov-normalization is an effective (Escudero and Bion, 2007) and standard approach to remove talker-specific differences, facilitating the comparison of *groups* of talkers (such as native vs non-native speakers).[6]

C. Roadmap

Before we turn to the primary goals of the present study, we test whether we can replicate some of the hallmarks of Mandarin-accented English in the new database (following related previous works, Vaughn *et al.*, 2019; Wade *et al.*, 2007). We begin by examining L2 speakers' deviations in the category *centers* (means) relative to native-English speakers' productions. Previous work has found that differences between native and Mandarin-accented English in terms of category center location are widely present across vowel categories regardless of whether or not the category is present in Mandarin (Chen *et al.*, 2001; Flege *et al.*, 1997). We assess whether these findings replicate in our data.

Compared to native speech, non-native speech tends to be processed more slowly and less accurately by native listeners (Adank *et al.*, 2009; Munro and Derwing, 1995). Our second question is thus whether the non-native speech in our database would be predicted to cause reduced recognition accuracy. While we do not have perception data, we can use a talker's production data to approximate the *predicted* recognition accuracy for a listener familiar with the talker's speech. For example, everything else being equal, most major theories of speech perception would predict that a reduced distance between the means of neighboring categories will result in reduced recognition accuracy (see, e.g., exemplar theory, Todd *et al.*, 2019; Bayesian models, Feldman *et al.*, 2009; Kleinschmidt and Jaeger, 2015; Kronrod *et al.*, 2016; connectionist models, McClelland and Elman, 1986; neighborhood activation model, Luce and Pisoni, 1998). Similarly, increased category overlap due to increased variability would be predicted to cause reduced recognition accuracy (for evidence, see Clayards *et al.*, 2008; Newman *et al.*, 2001; Nixon *et al.*, 2016). The measure of category separability we employ provides a simple, albeit coarse-grained, nonparametric approximation of these effects (it is closely related to recognition rules used in exemplar models, cf. Todd *et al.*, 2019). Our analysis of separability therefore serves to test whether our Mandarin-accented L2 speakers are indeed predicted to be more difficult to comprehend for native listeners compared to native-English speakers. This approach is similar to the use of discriminant analysis in Wade *et al.* (2007).

Measures of category separability can capture—but do not disentangle—differences in both category means and category variability. Differences in separability between native and non-native speech could thus originate from either of these differences. The final parts of study 1 address this by turning to our primary question: does non-native speech exhibit greater category dispersion?

Our analyses of category dispersion extend previous work in two ways. First, previous studies have typically compared category variability separately for the different cue dimension—for example, conducting one comparison of native and non-native speech for variability along F1 and another comparison for F2. However, for analyses that seek to assess whether non-native speech is less *precise*, this is arguably problematic. Figure 2 illustrates hypothetical ways in which non-native speech might differ from native speech. This includes cases in which the overall variability in the F1-F2 space (area of the ellipses)—and hence the talkers' precision—is identical across native and non-native speech even when separate analyses of F1 and F2 would suggest differences [Fig. 2(A)]. Similarly, there are cases in which separate analyses of F1 and F2 would suggest no difference between native and non-native speech, whereas in reality there are stark differences in the overall variability [Fig. 2(B)]. We thus focus our tests of the across-the-board

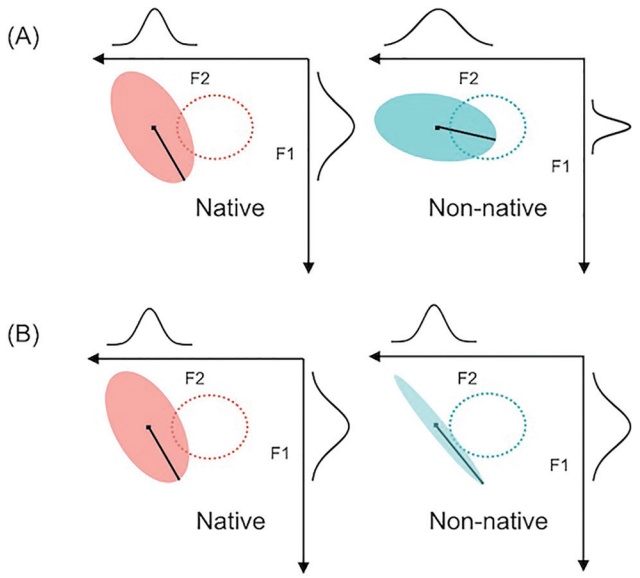

FIG. 2. Distribution of /i/ (solid ellipses) from a hypothetical native speaker and a hypothetical non-native speaker. (A) The native speaker has greater F1 variation and smaller F2 variation; the non-native speaker has smaller F1 variation and greater F2 variation. Black lines indicate the direction in which the category has the widest dispersion (covariance) and are of equal length for both speakers. The F1-F2 covariance is more negative for the native speaker. Thus, even for identical category means and an identical amount of dispersion, the orientation of dispersion can still differ. (B) Both speakers have equal variation in F1 and F2 alone but differ in terms of the overall dispersion. Black lines indicate that the orientation of dispersion again differs between the two speakers even though they have identical category means and identical cue-specific variation. Both panels show that a difference in the orientation of dispersion can result in different degrees of overlap with a neighboring category (dotted ellipses) and therefore impact category separability and perception difficulty.

hypothesis on the overall variability in the multidimensional (F1-F2) phonetic space, which we take to provide a more adequate measure of a talker's precision in realizing a category. To facilitate comparison to previous work, we *also* analyze dispersion along F1 and F2 separately.

Figure 2 also points to the second innovation of the present study. We identify two aspects of category dispersion, namely the *magnitude of dispersion* (what we have referred to so far and will continue to refer to as the within-category variability of cues) and the *orientation of dispersion* (the within-category correlation between cues). This contrasts with previous work, which has exclusively focused on the magnitude of dispersion (Wade *et al.*, 2007 report but do not analyze within-category cue correlations). As is visible in Fig. 2, differences in the orientation of category dispersion can affect the interpretation of differences in the magnitude of dispersion. Differences in the orientation of dispersion can also affect the degree of overlap between categories (also shown in Fig. 2). Since such overlap is known to affect the distinguishability of categories in perception (Feldman *et al.*, 2009; Kleinschmidt and Jaeger, 2015; Kronrod *et al.*, 2016), analyses of cue correlations thus promise to contribute to a fuller understanding of how the distributional properties of non-native speech affect its perception by native listeners. For all of these reasons, we expand previous work and compare category-specific differences in the orientation of dispersion between native and non-native speech.

Together, the different analyses we present assess all three potential sources of differences between non-native and native speakers that might contribute to the decreased intelligibility of even proficient foreign-accented speakers: differences in category means, the magnitude of category dispersion, or the orientation of category dispersion.

D. Results

1. Comparing native and non-native category means

We employed mixed-effects linear regressions (Baayen *et al.*, 2008) over the combined data from all nine vowels and both native and non-native speakers. The analysis was based on 180 data points (= 10 speaker * 2 accents * 9 vowel categories), where each data point was a speaker's category mean. Figure 3 shows the distribution of talker's category means for both native and non-native speakers. The analysis contained vowels, accent, and their interaction as fixed-effect predictors, as well as the maximal random effect structure (by-talker intercepts). To assess differences between native (N) and non-native (NN) speech, we report simple effects of accent (sum-coded, NN speakers = 1, N speakers = −1) at each level of vowel.[7] The results are presented in Table III. Except for /æ/ and /ɛ/, the means of all vowel categories differed in at least one formant dimension between native and non-native speech. These differences present primarily in terms of F1 (vowel height) rather than F2 (vowel backness).

Figure 3 summarizes the changes in category means from native to non-native speech. The differences in category means are largely consistent with common patterns from past work. In particular, Mandarin-accented English tends to shift the mean of /ɪ/ along both F1 and F2 and the mean of /ʊ/ along F1 (Chen *et al.*, 2001; Flege, 2003; Wang and van Heuven, 2006). These deviations from native pronunciations pull /ɪ/ and /ʊ/ closer to their competing categories (/i/ and /u/, respectively). The changes for /æ/ and /ɛ/—decreased F2 for /æ/ and increased F2 for /ɛ/—show a similar, albeit non-significant pattern.

a. Effects of phonetic context on category means. The SI presents *post hoc* analyses of phonetic context effects on category means. These analyses replicate well-known effects on vowel pronunciations—including, for instance, F1 lowering of /æ/ before nasals (Labov *et al.*, 2006), F2 drop of /u/ and /ʊ/ before laterals (Labov, 2006), and F1 lowering of multiple vowels (/æ, ɛ, ɑ, ɪ/) before voiced consonants (Moreton, 2004). Controlling for context did not change most of the effects reported in Table III (which essentially are the main effects of native vs non-native speech when averaging across phonetic contexts).[8] We did, however, find that non-native speakers in our sample often exhibited reduced context effects (although in the same direction as native speakers). This raises the possibility that differences in context effects could confound the comparison of category dispersion across native and non-native speech. We address this possibility below.

In summary, category means differ between native-English and Mandarin-accented English speech (as expected given previous work). They do so in ways that move categories closer together in Mandarin-accented English (also as expected). This is compatible with the hypothesis that non-native speech is characterized by a greater amount of phonetic overlap between neighboring vowel categories. The next analysis assesses this hypothesis more directly.

2. Comparing the separability of neighboring categories: The cases of /æ/-/ɛ/, /i/-/ɪ/, and /u/-/ʊ/

Decreased phonetic distance between tokens of neighboring categories reduces intelligibility (e.g., Bradlow, 1995; Wright, 2004). We therefore calculate the separability of vowel pairs that are close neighbors, and we compare this separability between native and non-native speech. We focused on three vowel contrasts hypothesized to be perceptually confusable for Mandarin speakers, namely /æ/-/ɛ/, /i/-/ɪ/, and /u/-/ʊ/.

Following past work (Wedel *et al.*, 2018), we operationalized each vowel category's separability from the neighboring category as the average distance of vowel tokens to the midpoint position of the neighboring category.[9] Figure 4 visualizes how this measure was calculated. Take, for example, the vowel pair /i/-/ɪ/. For each speaker, we determined the category centers as the average F1 and F2 values for /i/ and /ɪ/, respectively (marked by vowel labels in Fig. 4). Then, we calculated the distance between each /i/ token of that speaker and the center of /ɪ/ category for that speaker (dotted grey line). By averaging across all /i/ tokens, we

https://doi.org/10.1121/10.0001141

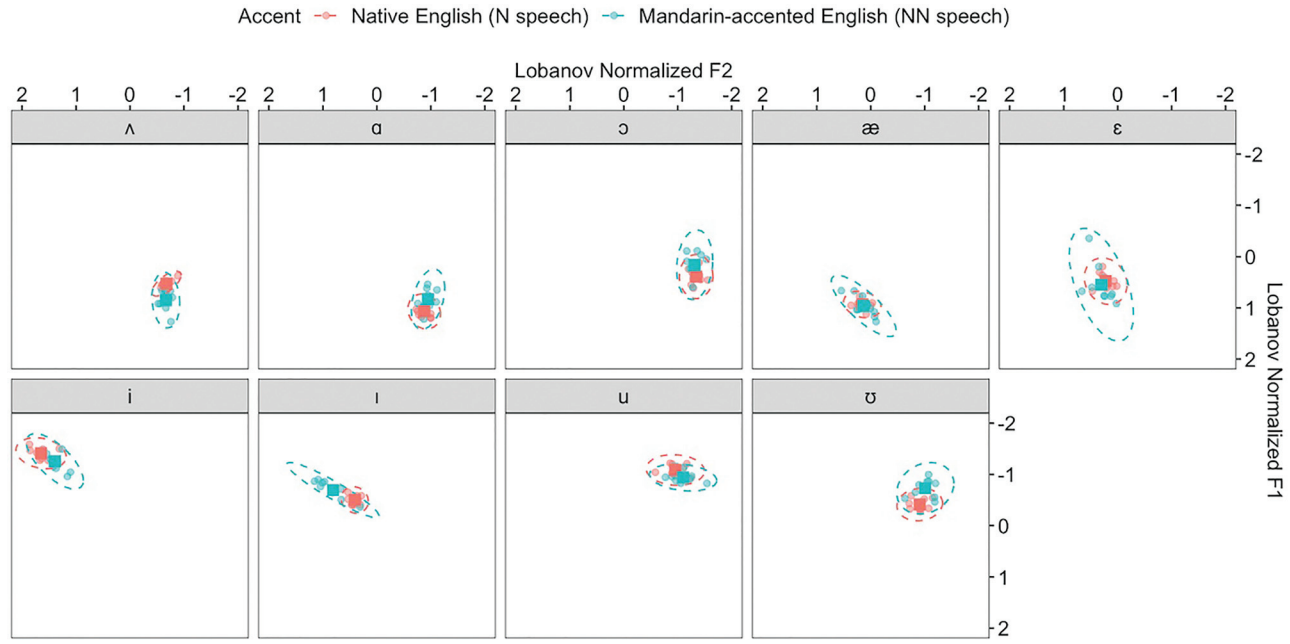

FIG. 3. Distribution of talker means (points) across native (N) and non-native (NN) speech is shown for all nine vowel categories. Squares indicate the category center for each accent (color-coded) averaged across all talkers within an accent. Ellipses indicate bivariate Gaussian 95% confidence interval of talker means. Note that this is not identical to visualizing the typical variability of categories in the two accents (no information about within-talker category dispersion is shown here). Also visible here is that many category means varied considerably more across non-native speakers than across native speakers. This replicates previous work (Wade *et al.*, 2007) and matches intuition: even a relatively homogenous sample of non-native talkers (as recruited here) is likely to be considerably more heterogenous in terms of proficiency than a group of native talkers. Crucially, this cross-talker variability is not to be confused with the within-talker category dispersion that constitutes the focus of the present work.

obtain a score of how separable /i/ is from /ɪ/ for that speaker. The separability of /i/ from /ɪ/ is thus calculated following this formula (n represents the number of tokens for the category):

$$\text{Separability of /i/ from /I/} = \frac{\sum_{k=1}^{n} \sqrt{(F1_{\text{token } k \text{ of /i/}} - F1_{\text{Center of /I/}})^2 + (F2_{\text{token } k \text{ of /i/}} - F2_{\text{Center of /I/}})^2}}{n}.$$

Similarly, we can obtain a score of how separable /ɪ/ is from /i/. This measure hence permits an asymmetry of separability between the two categories within each contrast—a phenomenon that has been documented in perceptual discriminability (e.g., Cutler *et al.*, 2006).

We employed the same mixed-effects regression approach used in Sec. II D 1 to analyze by-speaker by-category separability. All predictors and coding were identical except that we only included the vowels /æ/-/ɛ/, /i/-/ɪ/, and /u/-/ʊ/. Table IV summarizes the results. We found no L1-L2 difference for the /æ/-/ɛ/ contrast. Tokens of /i/-/ɪ/ and /u/-/ʊ/ were more separable from the neighboring category in native speech than in non-native speech. In other words, non-native speakers' realization of /i/-/ɪ/ and /u/-/ʊ/ were acoustically more confusable in F1-F2 space.

Replicating past work (e.g., Wade *et al.*, 2007), we found that non-native speech exhibits decreased separability for competing vowel categories. This decreased separability could be due solely to smaller distances between category means in non-native speech or it could additionally be exacerbated by greater category dispersion in non-native speech. The next analysis therefore addresses our primary question: how do native and non-native speech differ in terms of within-talker within-category variability?

3. Comparing the magnitude of category dispersion

We compare the degree of variability in native and non-native speech at three "grain-sizes," testing the different hypotheses we laid out earlier. First, we compare variability while pooling the data from all categories (without distinguishing between them). This lets us test the across-the-board hypothesis that L2 speech in our database is generally more variable than L1 speech regardless of the specific vowel category. This prediction—depicted in Fig. 5(a)—has

TABLE III. Comparison of vowel category means between N and NN speakers [$\Delta\mu$ (NN-N)] based on mixed-effects linear regression. Each row shows the simple effect of accent (NN-N). * represents $p < 0.05$, ** represents $p < 0.01$, *** represents $p < 0.001$, and † represents $p < 0.1$.

	Mixed-effects model: Lobanov-normalized F1 × F2 category means					
Vowel	Measure	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\mu$ (NN-N)
ʌ	F1	0.162	0.039	4.192	0.000***	Higher F1
	F2	0.006	0.036	0.158	0.874	
ɑ	F1	−0.120	0.039	−3.109	0.002**	Lower F1
	F2	−0.035	0.036	−0.978	0.330	
ɔ	F1	−0.113	0.039	−2.937	0.004**	Lower F1
	F2	0.014	0.036	0.398	0.691	
æ	F1	0.015	0.039	0.383	0.702	
	F2	−0.011	0.036	−0.305	0.761	
ɛ	F1	0.036	0.039	0.923	0.357	
	F2	0.040	0.036	1.125	0.262	
i	F1	0.079	0.039	2.043	0.043*	Higher F1
	F2	−0.128	0.036	−3.572	0.000***	Lower F2
ɪ	F1	−0.097	0.039	−2.516	0.013*	Lower F1
	F2	0.202	0.036	5.626	0.000***	Higher F2
u	F1	0.074	0.039	1.916	0.057†	Higher F1
	F2	−0.072	0.036	−2.021	0.045*	Lower F2
ʊ	F1	−0.159	0.039	−4.112	0.000***	Lower F1
	F2	−0.052	0.036	−1.438	0.152	

been the focus of previous studies, typically assessed through omnibus tests (analyses of variance, ANOVAs).

Second, we compare whether the vowel categories we *expect* to exhibit more variability in L2 speech—based on L2 sounds' assimilation patterns into closest L1 categories—indeed exhibit more variability compared to categories not expected to differ between L1 and L2 speech (see Table I). This prediction, according to the closest-neighbor hypothesis, is depicted in Fig. 5(b). It has not been tested in previous work.

Specifically, we predict that native and L2 speakers will exhibit comparatively little difference in their within-category variability for a number of L2 categories, including categories that are either equivalent counterparts

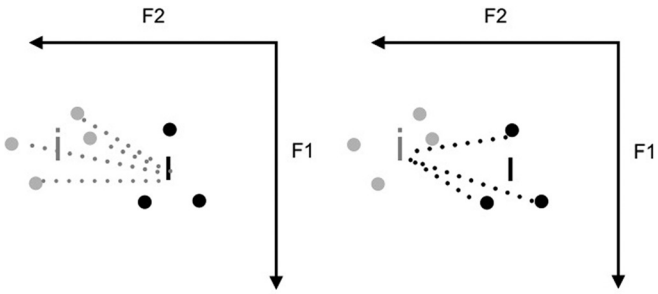

FIG. 4. Schematic representation of how we calculated separability. Vowel labels indicate category centers. Dotted lines show the distance of a token to the center of the neighboring category. For instance, the separability of /i/ from /ɪ/ is calculated as the mean length of all grey dotted lines (left), and the separability of /ɪ/ from /i/ is calculated as the mean length of all black dotted lines (right).

(/ɑ/ and /ɔ/) to or good exemplars of a closest L1 category (/i/, /u/) or entirely uncategorized (/ʌ/). The remaining four L2 categories are poor exemplars of an L1 category in a *single-category assimilation* type or they are the less good exemplar of the L1 category in *category-goodness assimilation* type (as summarized in Table I). We expect increased category variability for /æ/, /ɛ/, /ʊ/, and /ɪ/ in L2 speech.

Last, we compare variability in native and non-native speech for each vowel category. This final analysis follows our analyses of category means and separability; it lets us compare within-talker category-specific differences in variability between native and non-native speech. This final type of analysis parallels the *post hoc* category-specific comparisons conducted in previous work whenever omnibus (ANOVA) returned significant interactions between vowel and speaker group.

Deviating from most previous work, our primary measure of variability assesses the overall amount of a category's variability along both F1 and F2 (following Wade *et al.*, 2007). For each vowel and talker, we determined its category centers as its average F1 and F2 values. For all tokens of each category, we then calculated their Euclidean distances to the category's center in F1-F2 space. For example, for /i/,

$$\text{Overall variability of /i/} = \frac{\sum_{k=1}^{n} \sqrt{\left(\text{F1}_{\text{token } k \text{ of /i/}} - \hat{\mu}_{\text{F1 of /i/}}\right)^2 + \left(\text{F2}_{\text{token } k \text{ of /i/}} - \hat{\mu}_{\text{F2 of /i/}}\right)^2}}{n}.$$

The resulting measure is on a comparable scale with the SD of a category's dispersion along individual cue dimensions.[10] For the purpose of contrasting the across-the-board hypothesis and the closest-neighbor hypothesis, this measure provides an appropriate assessment: neither of these hypotheses makes cue-specific predictions; rather the question is whether non-native speakers are more variable (i.e., less precise) in their production and, if so, whether the increased variability is category specific or not.

Additionally, we present separate measures of variability along just F1 and F2. As outlined above, relying solely on

https://doi.org/10.1121/10.0001141

TABLE IV. Comparison of vowel category separability between N and NN speakers [Δseparability (NN-N)], based on mixed-effects linear regression. Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Lobanov-normalized F1 × F2 distance to contrastive category						
Vowel	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	Comparison	Δseparability (NN-N)
æ	−0.042	0.045	−0.928	0.364	æ → ɛ center	
ɛ	0.048	0.046	1.029	0.315	ɛ → æ center	
i	−0.330	0.045	−7.332	0.000***	i → ɪ center	N > NN
ɪ	−0.318	0.045	−7.055	0.000***	ɪ → i center	N > NN
u	−0.192	0.047	−4.108	0.000***	u → ʊ center	N > NN
ʊ	−0.147	0.048	−3.041	0.005**	ʊ → u center	N > NN

those separate measures can be misleading if the question is whether non-native speech is less precise. We present these measures here to facilitate comparison to previous work, and because—when accompanied by differences in the overall variability—they shed light on *how* the dispersion of categories in non-native speech differs from their dispersion in native speech. (In Sec. IV, we discuss more specific predictions, including cue-specific influence from L1 phonology.)

The analyses of variability are identical to the ones reported so far except that we employed *generalized* linear mixed-effects regression with a gamma-distributed outcome (log-link). This approach accounts for the expected distribution of SDs, which are always positive and tend to exhibit positive skewness.

a. Testing the across-the-board hypothesis: Is non-native speech generally more variable than native speech regardless of the category? Mixed-effects models were fitted with speaker-specific category SDs as the dependent variable. This first analysis included only accent (sum-coded, NN = 1, $N = -1$) as the fixed effect and random intercepts by speaker and vowel category.

There was no significant main effect of accent on overall variability in F1-F2 space ($\hat{\beta} = 0.010$, SE = 0.037, $t = 0.264$, $p = 0.79$). Neither was there a main effect on the SD of F1 ($\hat{\beta} = 0.039$, SE = 0.033, $t = 1.166$, $p = 0.24$) or SD of F2 ($\hat{\beta} = 0.023$, SE = 0.055, $t = 0.408$, $p = 0.68$). The present data thus lend no significant support for the hypothesis that non-native speech universally exhibits increased variability [cf. Fig. 5(a)].

b. Testing the closest-neighbor hypothesis: Are the categories predicted to be more variable in non-native speech indeed more variable than in native speech? Next, we tested whether non-native speech has increased variability for the four vowels that theories of L2 perception and production predict to be most affected (/æ/, /ɛ/, /ʊ/, and /ɪ/) compared to all other vowels. The vowels /æ/, /ɛ/, /ɪ/, /ʊ/ were coded as "expected increase" and the remaining categories were coded as "no expected increase" (sum-coded, expected increase = 1, no expected increase = −1). We then repeated the analysis presented in Sec. II D 3 a but included expectation and its interaction with accent as fixed effects.

Neither the main effect of accent nor the main effect of expectation was significant on overall variability in F1-F2 space ($ps > 0.69$). Crucially, there was a significant interaction between accent and expectation ($\hat{\beta} = 0.042$, SE = 0.018, $t = 2.319$, $p = 0.020$; see Fig. 6). This interaction is predicted by theories of L2 speech perception and production [see Fig. 5(b)]. Simple effects analysis revealed that for vowels expected to be more variable in non-native speech, there was indeed numerically greater overall variability in non-native speech ($\hat{\beta} = 0.057$, SE = 0.043, $t = 1.336$, $p = 0.18$); for the vowels expected to have no increased variability, there was numerically less variability in non-native speech ($\hat{\beta} = -0.027$, SE = 0.041, $t = -0.664$, $p = 0.51$). The direction of the former

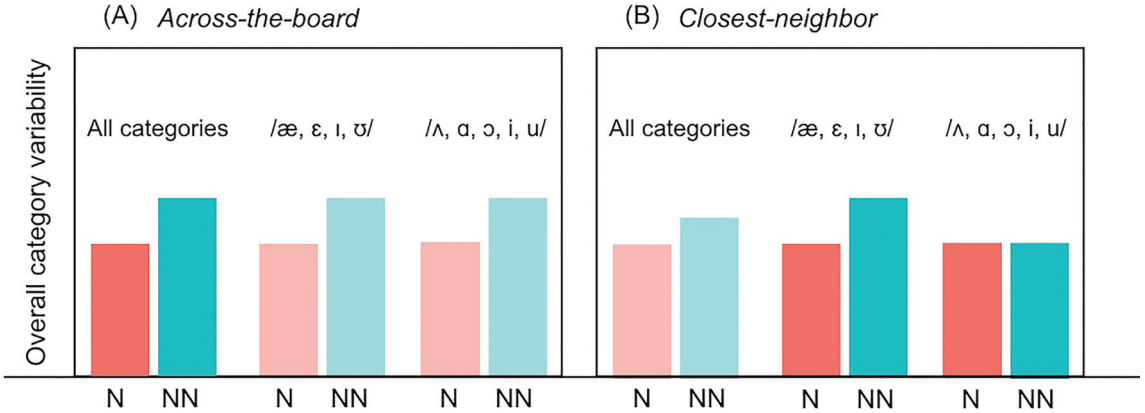

FIG. 5. Predictions of two different hypotheses about difference in category variability between native (N) and non-native (NN) speech. (a) The across-the-board hypothesis predicts that non-native speech is more variable than native speech regardless of category. (b) The closest-neighbor hypothesis predicts increased variability only for some specific vowels. The focus of each hypothesis is indicated by opaqueness.

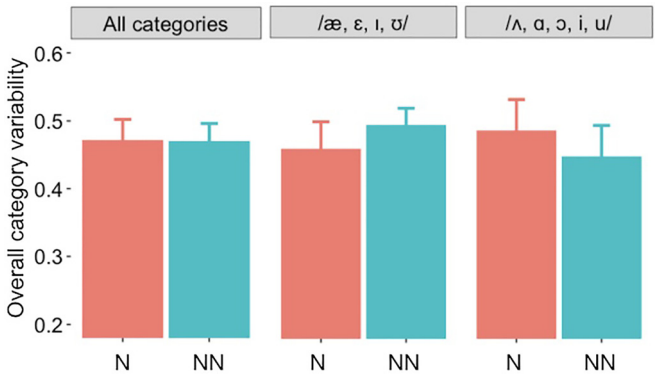

FIG. 6. (Color online) Category dispersion in study 1 shown for all vowels (left), as well as separately for vowels predicted to show increased dispersion in NN speech than in N speech under the closest-neighbor hypothesis (middle), and those predicted to not show increased dispersion under this hypothesis (right).

simple effect matches our predictions based on theories of L2 speech perception and production, but the fact that this simple effect does not reach significance relativizes the strength of the support this result provides for our predictions. For the latter simple effect, we expected a null effect. In terms of significance, that is what we observe, although numerically category dispersion for these vowels was *smaller* for non-native speech compared to native speech.

The effect on overall variability seems to be driven primarily by changes in the dispersion along F2. For the SD of F1, there was neither significant main effect of accent nor an accent-by-expectation interaction ($ps > 0.24$). For the SD of F2, there was a significant interaction between accent and expectation ($\hat{\beta} = 0.072$, SE $= 0.032$, $t = 3.275$, $p = 0.001$). Simple effects analysis revealed the same pattern for overall dispersion, although again both simple effects did not reach significance. Specifically, there was a trend for greater F2 variability for the four vowels expected to be so ($\hat{\beta} = 0.105$, SE $= 0.060$, $t = 1.725$, $p = 0.08$); there was no difference between native and non-native speech for the remaining categories ($\hat{\beta} = -0.040$, SE $= 0.059$, $t = -0.683$, $p = 0.49$).

c. Category-by-category comparison of native and non-native speech. Our third and final analysis of category variability compares non-native to native speech separately for each of the nine vowels. The model specification was identical to that employed in our analyses of category means and separability. Simple effects for the SDs of F1, F2, and overall variability are reported in Table V. L2 speakers exhibited greater variability for /ɛ/ (F1 and F2), /i/ (F1 only), and /ɪ/ (F1 only) compared to L1 speakers. There were no differences between the two speaker groups for /ɑ/, /ɔ/, /ʌ/, and /ʊ/. Non-native speakers actually had smaller variability for /æ/ (F1 only) and /u/ (F2 only) than native speakers. Non-native speakers showed greater overall variability for /ɛ/ and smaller variability for /u/ compared to native speakers.

In sum, we find little evidence that Mandarin-accented English speech is *universally* more variable in its production of English vowels. Non-native speech was only more variable for a single category, namely /ɛ/. Non-native speakers were actually *less* variable (more precise) than native speakers in their production of /u/. If one considers variability along just F1, there would appear to also be evidence that non-native speech exhibits greater variability for /i/ and /ɪ/ and less variability for /æ/.

TABLE V. Comparison of vowel category variability between N and NN speakers [$\Delta\sigma$ (NN-N)] based on mixed-effects regression with a gamma-distributed outcome (log-link). Each row shows the simple effect of accent (NN-N). Results are bold if there is a significant difference between N and NN speech in terms of overall variability.

Mixed-effects models: Lobanov-normalized F1 × F2 category variability						
Vowel	Measure	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\sigma$ (NN-N)
ʌ	F1	−0.030	0.069	−0.440	0.660	
	F2	0.062	0.077	0.796	0.426	
	Overall	0.002	0.058	0.041	0.967	
ɑ	F1	0.079	0.069	1.143	0.253	
	F2	0.089	0.077	1.146	0.252	
	Overall	0.062	0.059	1.056	0.291	
ɔ	F1	−0.094	0.069	−1.363	0.173	
	F2	−0.074	0.077	−0.958	0.338	
	Overall	−0.099	0.058	−1.686	0.092	
æ	F1	−0.207	0.069	−3.003	0.003**	N > NN
	F2	0.057	0.077	0.735	0.462	
	Overall	−0.112	0.058	−1.917	0.055†	
ɛ	**F1**	**0.178**	**0.069**	**2.591**	**0.010****	**NN > N**
	F2	**0.364**	**0.077**	**4.707**	**0.000*****	**NN > N**
	Overall	**0.257**	**0.058**	**4.394**	**0.000*****	**NN > N**
i	F1	0.202	0.069	2.929	0.003**	NN > N
	F2	−0.080	0.077	−1.036	0.300	
	Overall	0.041	0.059	0.702	0.483	
ɪ	F1	0.159	0.069	2.304	0.021*	NN > N
	F2	0.038	0.077	0.494	0.622	
	Overall	0.091	0.058	1.562	0.118	
u	**F1**	**−0.013**	**0.069**	**−0.189**	**0.850**	
	F2	**−0.220**	**0.077**	**−2.843**	**0.004****	**N > NN**
	Overall	**−0.151**	**0.059**	**−2.575**	**0.010***	**N > NN**
ʊ	F1	0.051	0.069	0.736	0.462	
	F2	−0.049	0.077	−0.640	0.522	
	Overall	−0.010	0.058	−0.173	0.862	

d. Effects of phonetic context on category variability. *Post hoc* analyses presented in the SI assessed the possibility that some of the accent differences in category variability might be caused by the failure to consider phonetic context. A comparison of the definitions of context-dependent and -independent category variability illustrates the potential problem:

$$\text{Context} - \text{dependent variability of } /\text{æ}/ \text{ along F1} = \sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n_j}(\text{F1}_{\text{token } i \text{ of } /\text{æ}/} - \hat{\mu}_{\text{F1}_{/\text{æ}/ \text{ in context } j}})^2}{n}},$$

$$\text{Context} - \text{independent variability of } /\text{æ}/ \text{ along F1} = \sqrt{\frac{\sum_{i=1}^{n}(\text{F1}_{\text{token } i \text{ of } /\text{æ}/} - \hat{\mu}_{\text{F1 of } /\text{æ}/})^2}{n}}.$$

https://doi.org/10.1121/10.0001141

When context-specific category means are farther apart (i.e., when the context effect is large), tokens belonging to a particular context level (e.g., tokens followed by voiced consonants vs tokens followed by voiceless consonants) might be close to the corresponding context-specific category mean $\hat{\mu}_{\mathrm{F1}_{/æ/\text{ in context } j}}$ but at the same time are far from the overall category mean $\hat{\mu}_{\mathrm{F1\,of\,}/æ/}$. The greater the phonetic context effect is, the greater inflation there would be in the calculation of context-independent variability. This would not be of concern for the comparison of native and non-native speech if phonetic context effects were the same in size in both accents (since native and non-native speakers in our sample produced vowels in the same contexts). However, as we described in our analysis of category means, phonetic context effects seem to impact vowel means differently in native and non-native speech. Crucially, the effects are sometimes smaller in the non-native speech. Failing to control for context in the calculation of category dispersion thus risks artificially inflating the within-talker variability of native speech. In particular, it is possible that an inflated estimation of variability in the native speech contributed to the two surprising results (/æ/ and /u/) where we observed reduced variability in the non-native speech.

The *post hoc* analyses presented in the SI do not change the conclusions with regard to the across-the-board or closest-neighbor hypotheses. The comparison of context-dependent category variability across native and non-native speech largely returns the same results as the main analysis of context-independent category variability. At least in the present dataset, there is no convincing evidence that phonetic context systematically confounds the overall lack of accent differences in within-talker variability. We note, however, that the smaller variability of /æ/ (along F1 and overall) in non-native compared to native speech seems to be no longer significant (although still in the same direction) when phonetic context is considered. Similarly, the other surprising results—smaller F2 variability of /u/ in non-native speech—is completely removed after controlling for phonetic context. We return to these points in Sec. IV.

4. Comparing the orientation of dispersion (covariation between F1 and F2)

The covariation of F1 and F2 determines the orientation of vowel categories in the formant space. It is possible that non-native vowels are systematically different from native speech in terms of the direction in which they are dispersed instead of being overall more dispersed.

For each unique combination of vowel category and speaker, we calculated correlations between F1 and F2. These by-speaker by-category F1-F2 correlations were analyzed in a mixed-effects linear regression using the same predictors and coding as in Sec. II D 1. Table VI summarizes the results. The degree of F1-F2 correlations was significantly different between native and non-native speech for /æ/ ($\hat{\beta} = -0.169$, SE $= 0.064$, $t = -2.613$, $p = 0.010$) and /ʊ/ ($\hat{\beta} = -0.162$, SE $= 0.064$, $t = -2.505$, $p = 0.013$). No significant differences were observed for other categories.

Figure 7 visualizes the accent-specific (N vs NN) distribution of vowel categories. Each ellipse was determined by taking the averaged mean of F1 and F1 values as well as the averaged covariance matrix across all talkers. In Fig. 7, it is apparent that the seemingly decreased variability for /æ/ along F1 in non-native speech (see Table V) is primarily driven by the non-native *co*variance between F1 and F2 in non-native speech. And, although the *magnitude* of dispersion for /ʊ/ does not differ between native and non-native speech, its *orientation* does. This highlights the importance of analyzing cue covariation when trying to understand differences between native and non-native speech. In Sec. IV, we link our findings to L2 learning theories and discuss in greater detail why these two categories are particularly affected in its orientation.

E. Discussion

In line with previous work on L2 productions, we find that non-native speech differs significantly from native speech in terms of many of its category *means*. Also in agreement with expectations and previous work, we find that neighboring vowel categories are significantly less separable in non-native speech compared to native speech. The present database thus replicates

TABLE VI. Comparison of F1-F2 cue correlation between N and NN speakers $\Delta\rho$ (NN-N) based on mixed-effects regression. Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Cue correlation between Lobanov-normalized F1 and F2					
Vowel	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\rho$ (NN-N)
ʌ	0.005	0.064	0.078	0.938	
ɑ	0.037	0.064	0.575	0.566	
ɔ	0.018	0.064	0.279	0.780	
æ	0.169	0.064	2.613	0.010**	Weaker negative correlation in NN
ɛ	−0.103	0.064	−1.603	0.111	
i	−0.039	0.064	−0.601	0.549	
ɪ	−0.098	0.064	−1.512	0.133	
u	0.047	0.064	0.725	0.469	
ʊ	0.162	0.064	2.505	0.013*	Positive correlation in NN; negative correlation in N

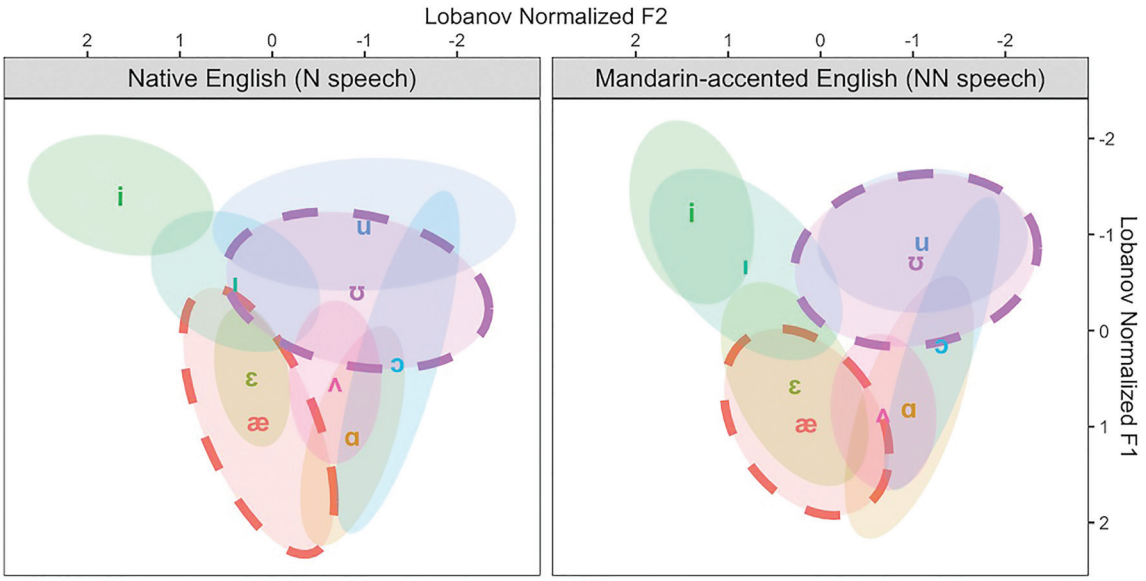

FIG. 7. Vowel category distributions in F1-F2 space by accent averaged across all talkers within an accent. Ellipses show 95% CIs based on the variance-covariance matrix averaged across the talker-specific variance-covariance matrices and centered around the mean of the talker-specific means. Bold lines highlight the two categories for which F1-F2 correlations differed between native and non-native speech.

well-documented properties of non-native speech (e.g., Flege *et al.*, 1997), including findings of previous work on Mandarin-accented English (Flege, 2003; Wang and van Heuven, 2006).

One somewhat surprising result is the lack of a significant L1-L2 difference in the separability of /ɛ/-/æ/ in study 1. This English contrast is presumably difficult for native-Mandarin speakers, whose realization has been reported to be perceptually confusable for native-English listeners (e.g., Evanini and Huang, 2012; Jia *et al.*, 2006). One possibility is that the non-native speakers tested in the present study were more proficient in English compared to speakers tested in past work. Another possibility is that dialectal differences in Mandarin, notably present in our native-Mandarin speakers, may add to differences in phonetic realization of this contrast (Jia *et al.*, 2006; Mou *et al.*, 2018). A third possibility is that the often reported /ɛ/-/æ/ confusion in Mandarin-accented English is not mediated only by the spectral distance between the two categories but is also affected by the orientation of category dispersion. We return to this point in Sec. IV.

These clear differences in category means and separability stand in contrast to our findings for category dispersion. We find no evidence in support of the across-the-board hypothesis that L2 vowel productions are inherently less precise. Rather, we find that the effect of L2 speech on category dispersion differs between vowels. This finding is consistent with evidence from recent studies that the amount of category variability in L2 speech seems to be category and cue specific (see the discussion in Vaughn *et al.*, 2019).

Indeed, most theories on L2 speech production predict that production of L2 categories are affected by the categories' place within both the speaker's L1 and L2 phonologies. Based on these theories, we developed the closest-neighbor hypothesis and categorized vowels into two groups—those expected to exhibit increased variability in non-native speech (/æ, ɛ, ɪ, ʊ/) and those not predicted to exhibit increased variability (/ʌ, ɑ, ɔ, i, u/). The two groups of vowels were indeed affected differently in non-native speech compared to native speech. However, not all of the specific predictions were borne out. We therefore discuss vowel-specific patterns.

We expected *no* increase in variability both for vowels that have phonologically equivalent (or at least very similar) counterparts in Mandarin (/ɑ, ɔ, i, u/) and for vowels that have no nearby neighbor in Mandarin (/ʌ/). These predictions were met for four of the five vowels: for /ɑ/, /ɔ/, /i/, and /ʌ/, we found no difference in category dispersion between the two speaker groups—neither in the amount of dispersion nor in the orientation of dispersion (for /ɔ/, overall dispersion was marginally *smaller* in non-native speech, $p = 0.092$). For /u/, we expected no increase in variability and found *de*creased variability.

We expected increased variability in non-native speech for vowels that are poor exemplars of a nearby closest-neighbor category in the L1 (/æ, ɛ, ɪ, ʊ/). This prediction was met most clearly only for one category: there was greater overall variability for /ɛ/ in non-native speech. For /ɪ/, we found no significant difference between native and non-native speech in overall dispersion, although the difference was approaching marginal significance in the predicted direction ($p = 0.118$; see Table V). The results for the remaining two vowel categories for which we expected increased variability did not match our predictions. For both /æ/ and /ʊ/, we found no significant increase in overall dispersion in non-native speech. In fact, the overall dispersion of /æ/ was marginally *smaller* in non-native speech

https://doi.org/10.1121/10.0001141

($p = 0.055$). In Sec. IV, we return to these unexpected results and offer explanations, including phonetic context effects on these vowels.

To sum up study 1, we found no support for the across-the-board hypothesis. At least the experienced L2 speakers in our database do not show a general increase in variability in vowel production. The overall pattern was compatible with the predictions of the closest-neighbor hypothesis, but /ɪ, æ, ʊ, u/ exhibited less category variability than predicted. These results suggest that the closest-neighbor hypothesis fails to account for additional factors influencing non-native speakers' production. We will return to this point in Sec. IV, together with evidence from study 2.

III. STUDY 2

One possibility for the comparatively small differences in variability found in study 1 is that study 1 (like previous work) focused on realizations of contrasts that employ cues that are also used for the same class of contrasts in the native language of the L2 speakers. As discussed in Vaughn *et al.* (2019, pp. 24–25), it is possible that non-native speakers may show increased variability primarily for cues that they are less familiar with from their L1. Study 2 presents the first test of this possibility. To this end, we consider a case where L1 and L2 differ in the set of acoustic cues in signaling particular phonological contrasts.

Specifically, study 2 compares syllable-final (coda) stops produced by native-English speakers and native-Mandarin speakers (L2 speakers). We focus on the duration of preceding vowels, the closure interval, and the burst release. Voicing in coda stop variants in English is associated with longer vowels, shorter closure intervals, and shorter burst releases. The primary cues in English tend to be vowel and closure duration (e.g., Flege and Hillenbrand, 1987). Bursts are often not audibly released in L1 English, especially for voiceless stops (e.g., Deelman and Connine, 2001). Unlike English, Mandarin has no word-final stops. While Mandarin *does* have stop distinctions in word *onsets* [e.g., *ba* vs *pa* (*dad* vs *afraid*); *da* vs *ta* (*big* vs *stamp*); *gu* vs *ku* (*aunt* vs *cry*)], phonetic features that constitute the primary cues to stop voicing in English differ between the word-initial and -final positions (e.g., Klatt, 1975; Raphael, 1972) and tend to do so across languages (e.g., Abramson and Tingsabadh, 1999; Flege and Eefting, 1987; Flege and Wang, 1989; Lisker and Abramson, 1964). Whereas English contrasts onset homorganic stop pairs (/b/-/p/, /d/-/t/, and /g/-/k/) in terms of voicing (cued primarily through voice onset time), Mandarin contrasts onset stops in terms of aspiration (cued by burst, among other cues). At a more abstract level, these contrasts in English and Mandarin onset stops are encoded in similar ways: short-lag stops (English voiced, Mandarin voiceless unaspirated) contrast with long-lag stops (English voiceless, Mandarin voiceless aspirated). Perhaps as a result of such similarity, native-Mandarin L2 speakers of English tend to use the burst cue—which distinguishes homorganic onset stop pairs—to distinguish voicing in coda stops (e.g., Flege, 1989; Xie *et al.*, 2017).

Word-final stop voicing thus poses a different type of challenge to L1 Mandarin learners than the acquisition of the English vowel system. Whereas the latter requires the acquisition of novel categories, the phonetic features that constitute the primary cues to vowels in English (formants) are also used to distinguish between Mandarin vowels. In contrast, word-final stop voicing requires L1 Mandarin learners of English to perceive and produce features that are not used—or at least not in the same phonological, articulatory, and perceptual contexts—in their native language Mandarin. Previous work has found that acoustic differences between voiced and voiceless coda stops in Mandarin-accented English tend to be smaller than those in native English (Bent *et al.*, 2008; Flege and Wang, 1990; Flege *et al.*, 1992; Hayes-Harb *et al.*, 2008). That is, L1 Mandarin productions of English coda stop voicing exhibit non-nativeness in terms of their central tendencies.

Here, we ask whether Mandarin-accented English exhibits increased variability compared to native English. We again test the across-the-board hypothesis and contrast it with a more specific hypothesis about *which* aspects of L2 production exhibit increased variability. Specifically, we expect the lack of articulatory practice with cue manipulations that do not occur in the L1 to result in increased variability. This prediction is derived from the "feature hypothesis" of SLM (Flege, 1995), which states that L2 phonetic features that are not phonologically contrastive in the L1 are more difficult for L2 speakers to grasp, leading to non-native pronunciations (for practice-induced reduction in category variability, see Kartushina *et al.*, 2016; Kartushina and Frauenfelder, 2014).

We predict that native-Mandarin L2 speakers of English exhibit greater within-category dispersion for vowel and closure duration—two cues that are not contrastively used in Mandarin—and possibly deviation from native covariation involving either of these cues. We refer to this as the *cue-specific* hypothesis, which we contrast with the across-the-board hypothesis that non-native speakers are inherently less precise in their production regardless of categories or cues.

A. Materials

Recordings from ten English speakers and ten Mandarin speakers (same as those in study 1) are analyzed. The stimuli consisted of 76 words (32 voiced, 44 voiceless) each repeated 3 times by each speaker. This resulted in 96 voiced (15 /b/, 63 /d/, and /18 /g/) and 132 voiceless (33 /p/, 36 /k/, and 63 /t/) tokens to be used in the following analyses. A total of 12 tokens (2.6%) were not included due to mispronunciations, leaving a total of 4548 tokens to analyze.

We measured three durational cues that signal voicing in word-final stops: vowel length, closure length, and burst length. Replicating past work (e.g., Flege *et al.*, 1992), we observed a higher ratio of unreleased bursts among English speakers (mean = 14.0%, SD = 13.2%) than among Mandarin speakers (mean = 2.3%, SD = 1.8%). To accurately assess overall category variability and cue covariation structure, we only included tokens for which all three cues were available (excluding another 8%). As a result, 4176 tokens were included in the following analyses (Table VII). More tokens were excluded from native speech than non-native speech due to the fact that some of the native-English speakers in our database habitually omitted burst releases (replicating previous work, e.g., Connine *et al.*, 1994). Critically, the unbalanced exclusion rates, if anything, are biases *against* the result we find below: if we did not exclude tokens with omitted bursts from the analysis, this would *in*crease the estimates of category variability for native speakers. In short, the exclusion criterion we apply here is both justified on *a priori* grounds and should make it *easier* to detect increased variability in non-native speakers.[11] After exclusions, there were 15 or more tokens per stop category per speaker even for speakers who had a high rate of unreleased bursts (with a single exception of a native speaker with only 8 tokens for category /t/).

To control for individual variability in speaking rates, we calculated the ratio of cue duration to total word duration as a proportional measure to control for variations in speaking rates.[12] In what follows, we refer to the cue/word duration ratio as *normalized durations* for simplicity. Additional analyses of the raw cue values are included in the SI and confirm the results presented here. Voiced categories (/b, d, g/) and voiceless categories (/p, t, k/) were analyzed separately.

B. Results

We proceed in the same order as in study 1. We first analyze differences in category means between native and non-native speech, followed by a comparison of category separability. Then, we analyze category dispersion, including both the magnitude and orientation of dispersion.

TABLE VII. Number of tokens for each stop category elicited for analysis for N and NN speakers. The last two rows indicate the number of tokens included in the analysis after removing mispronunciations and tokens with no detectible bursts.

Stop	b	d	g	p	t	k	Voiced	Voiceless	Total
Tokens elicited per speaker	15	63	18	33	63	36	96	132	228
Total tokens (N speech)	130	577	179	287	423	358	886	1068	1954
Total tokens (NN speech)	139	602	178	328	617	358	919	1303	2222

1. Comparing native and non-native category means

As in study 1, we employed mixed-effects linear regressions over the combined data from all six stop categories by both native and non-native speakers. The analysis was thus based on 120 data points (=10 speaker * 2 accents * 6 stop categories), where each data point was a speaker's category mean. The analysis contained voicing (sum-coded, voiced = 1, voiceless = −1), accent (sum-coded, NN speakers = 1, N speakers = −1), their interaction, and place of articulation (treatment-coded with alveolar as the reference level) as fixed-effect predictors, as well as the maximal random effect structure (by-talker intercepts). The effect of place of articulation was significant in all models. Since it served only as a control predictor and did not pertain to our questions on differences between native and non-native speakers, our reports here focus on the effects of accent and voicing. The full models are reported in the SI.

Following study 1, separate models were fitted for the three acoustic cues: vowel, closure, and burst. Table VIII reports the simple effects of accent for both voiced and voiceless categories. For voiced categories, non-native speakers had marginally significantly shorter burst durations than native speakers. For voiceless categories, non-native speakers had significantly longer vowel durations and shorter closure than native speakers. This replicates past findings for coda stop voicing in Mandarin-accented English (Bent *et al.*, 2008; Flege *et al.*, 1992; Hayes-Harb *et al.*, 2008) but with a sample size about one order of magnitude larger than in previous work. The specific pattern we find suggests that native-Mandarin speakers of English have a greater distinction in bursts but attenuated distinction in vowel and closure compared to native speakers of English. The next analysis assesses this more directly.

2. Comparing the separability of neighboring categories: The cases of /b/-/p/, /d/-/t/, and /g/-/k/

Following the procedure described in study 1, we determined the separability of voiced and voiceless categories. For each category, its separability from the neighboring category was operationalized as the average distance of individual tokens to the midpoint position of the neighboring category (e.g., from each /b/ token to the center of the

TABLE VIII. Comparison of coda stop category means between N and NN speakers [$\Delta\mu$ (NN-N)] based on mixed-effects linear regression. Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Vowel × closure × burst category means						
Stop	Measure	Coef $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\mu$ (NN-N)
Voiced	Vowel	0.004	0.008	0.504	0.619	
	Closure	0.007	0.006	1.073	0.295	
	Burst	−0.016	0.008	−2.023	0.055†	Shorter burst
Voiceless	Vowel	0.024	0.008	3.112	0.005**	Longer vowel
	Closure	−0.038	0.006	−6.000	0.000***	Shorter closure
	Burst	0.005	0.008	0.582	0.566	

https://doi.org/10.1121/10.0001141

/p/ category and from each /k/ token to the center of the /g/ category). The results, shown in Table IX, provide clear evidence that coda stop voicing is less separable in Mandarin-accented English than native English. This replicated previous work and provides an explanation for the well-documented difficulty of native listeners to recognize word-final stop voicing in Mandarin-accented speech (e.g., Bent *et al.*, 2008; Hayes-Harb *et al.*, 2008; Xie and Fowler, 2013).

3. Comparing the magnitude of category dispersion

Next, we compare non-native against native speech in terms of within-talker within-category variability for coda stops. In parallel with study 1 and the omnibus tests presented in past work, we start by testing whether non-native speech exhibits increased category variability in the data pooled across all six stop categories. Following study 1, we first present an analysis of categories' overall variability in the three-dimensional cue space defined by vowel, closure, and burst.

Also as in study 1, we additionally analyze variability along the three separate cue dimensions. Unlike in study 1, these latter analyses are of particular interest to us: these analyses allow us to investigate whether cues that L2 speakers have little previous practice with from their L1 are affected differently than cues that are employed in L2 speakers' native language. If articulatory precision is indeed affected by the long-term practice of cue manipulation, we expect to see increased variability in non-native speech for unfamiliar cues (vowel and closure durations); we do not expect a difference between native and non-native speech for familiar cues (burst duration).

Second, we conduct separate analyses of the dispersion of voiced and voiceless stop categories, comparing native and non-native speech. While category variability (as measured by voice onset times) tends to be larger for voiceless stops than voiced stops in word-initial position (e.g., Allen and Miller, 1999), we know of no previous work that assesses category variability for coda stops. We do not have predictions as to whether any potential difference in variability between native and non-native speech are more pronounced for voiced or voiceless stops.

We analyze variability in terms of the *coefficient of variation*. This measure—defined as SD divided by mean—corrects for dependencies between category means and their variability. A common concern in comparing measures of variability (e.g., SDs) for inherently bounded variables (such as durations, which cannot be smaller than zero) is that variability tends to increase with increasing means. The coefficient of variation is the commonly used approach to correct—or at least reduce—this problem, and it has been used in other research on category variability (e.g., Smith and Kenney, 1994; Whiteside *et al.*, 2003).[13]

For each measure, a generalized mixed-effects model with a gamma distributed outcome (log-link) was fitted with the same predictors and coding as described in Sec. II. For the overall variability in the three-dimensional cue space, there was no effect of accent for either voiced or voiceless categories (simple effects are presented in Table X; full model results are reported in the SI). This suggests that non-native speech does not exhibit overall increased variability. This result replicates the finding for vowels in study 1.

The separate analyses of the three cues revealed that variability in burst duration—the cue native-Mandarin speakers are familiar with—was affected differently in non-native speech, compared to both vowel and closure duration—the cues native-Mandarin speakers are less familiar with (see Table X). For burst durations, there was greater variability in non-native speech than native speech for voiced categories, but there was no difference for voiceless categories.

4. Comparing the orientation of dispersion

We followed the same procedure as in study 1. The results are shown in Table XI and visualized in Fig. 8. For both voiced and voiceless categories, there was a weaker vowel-burst correlation (which was negative) in non-native speech than in native speech. For voiced categories, the closure-burst correlation also differed between native and non-native speech: the correlation changed from positive in native speech to negative in non-native speech. No other differences were statistically significant.

TABLE IX. Comparison of category separability of coda stop contrasts in N and NN speech [Δseparability (NN-N)] based on mixed-effects linear regression. Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Vowel × closure × burst distance to contrastive category						
Category	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	Comparison	Δseparability (NN-N)
b	−0.022	0.008	−2.593	0.013*	b → p center	N > NN
p	−0.026	0.007	−3.526	0.002**	p → b center	N > NN
d	−0.033	0.007	−4.699	0.000***	d → t center	N > NN
t	−0.037	0.007	−5.222	0.000***	t → d center	N > NN
g	−0.024	0.008	−3.001	0.005**	g → k center	N > NN
k	−0.025	0.007	−3.362	0.003**	k → g center	N > NN

TABLE X. Comparison of coda stop category variability between N and NN speakers [$\Delta\sigma$ (NN-N)] based on mixed-effects regression with a gamma-distributed outcome (log-link). Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Vowel × closure × burst category variability						
Stop	Measure	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\sigma$ (NN-N)
Voiced	Vowel	−0.042	0.046	−0.914	0.361	
	Closure	0.026	0.040	0.648	0.517	
	Burst	0.115	0.052	2.223	0.026*	NN > N
	Overall	0.006	0.029	0.198	0.843	
Voiceless	Vowel	−0.049	0.046	−1.053	0.292	
	Closure	0.052	0.040	1.320	0.187	
	Burst	−0.049	0.052	−0.952	0.341	
	Overall	−0.013	0.029	−0.453	0.651	

C. Discussion

Our analyses highlight three findings. First, replicating past work, we find that Mandarin-accented English maintains a greater distinction in burst lengths for coda stop voicing compared to native-English speech but has diminished distinction in vowel and closure durations (e.g., Hayes-Harb *et al.*, 2008; Xie and Fowler, 2013). Moreover, the separability of voiced and voiceless stops was consistently smaller in Mandarin-accented English at all places of articulation. These results replicate previous work based on smaller databases (e.g., Flege *et al.*, 1992), as well as our own results for vowels in study 1. Second, there was little evidence that non-native speakers had increased variability in the realization of coda stop voicing when multiple cues were considered, either separately or jointly, with the exception that voiced categories had more variable bursts in Mandarin-accented English. This finding is in line with the cue-specific hypothesis.

Third, also in agreement with the cue-specific hypothesis, the degree of cue covariation appeared to differ between native and non-native speech. Specifically, non-native speakers showed a comparable amount of cue variation for the two primary cues used by native-English speakers, vowel and closure. On the other hand, vowel-burst covariation (and to some extent, closure-burst covariation) was smaller in non-native speech than in native speech. It is possible that as the non-native speakers attempt to maintain a voicing contrast by varying burst duration, they distort the covariation structure with other cues (vowel and closure) compared to native speech.

Taken together, our results suggest that there was comparatively little difference between native and non-native speech in terms of category variability. In terms of significance patterns, we found striking differences in category means as well as covariation between the relevant cues and relatively small to no differences in variability. This mirrors the results of study 1 on vowel production.

IV. GENERAL DISCUSSION

We set out to investigate whether non-native speech exhibits greater within-category within-talker variability than native speech. Given the conflicting findings in the literature on this issue, our immediate goal was to test the hypothesis in a comparatively high-powered dataset against a broader range of phonological categories. To this end, we compared productions of nine English vowels and six coda stops by native speakers of American English and Mandarin speakers who learned English as an L2. Across two studies, we found little evidence in support of the assumption that non-native speakers are *generally* more variable or less precise in their realization of L2 sounds. This result replicates some previous work (Smith *et al.*, 2019; Vaughn *et al.*, 2019). The present study further reveals two aspects of non-native speech that shed light on the seemingly conflicting results of previous work. First, although we do not observe an across-the-board increase in category variability in non-native speech, we do see evidence compatible with a more nuanced view of L1-to-L2 influence in terms of category dispersion. Second, a stark difference between native and non-native speech lies in the orientation of category dispersion—i.e., how cues *co*vary during production. The structure of cue covariation in non-native speech suggests that category variability is best understood in the joint phonetic space defined by multiple relevant phonetic cues (e.g., F1-F2 space for vowels). Before we elaborate on these two points, we briefly summarize our findings for category means and category separability in order to facilitate comparison with past work.

A. Category-specific L1-to-L2 influence in category means and separability

Studies 1 and 2 conceptually replicate past findings that L2 production often differs from native speech of the L2 as

TABLE XI. Comparison of cue correlation for coda stops between N and NN speakers [$\Delta\rho$ (NN-N)] based on mixed-effects linear regression. Each row shows the simple effect of accent (NN-N).

Mixed-effects models: Pairwise cue correlation						
Stop	Measure	Coefficient $\hat{\beta}$	SE ($\hat{\beta}$)	t	p	$\Delta\rho$ (NN-N)
Voiced	Vowel-closure	0.069	0.037	1.879	0.068	
	Vowel-burst	0.136	0.037	3.686	0.001***	Weaker negative correlation in NN
	Closure-burst	−0.114	0.043	−2.633	0.012*	Negative correlation in NN; positive correlation in N
Voiceless	Vowel-closure	−0.017	0.037	−0.470	0.641	
	Vowel-burst	0.098	0.037	2.665	0.011*	Weaker negative correlation in NN
	Closure-burst	0.014	0.043	0.335	0.740	

https://doi.org/10.1121/10.0001141

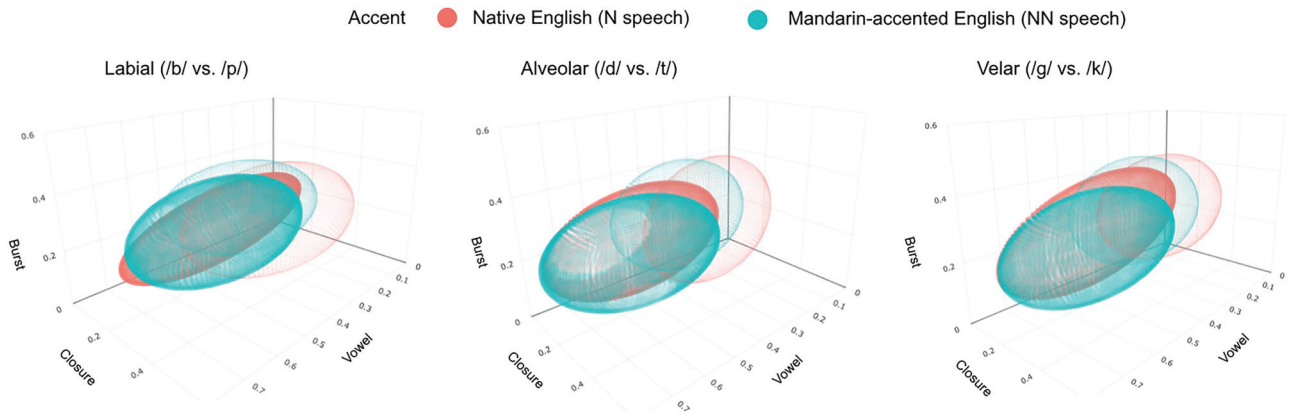

FIG. 8. Coda stop categories in the three-dimensional cue space defined by vowel, closure, and burst duration. Colors show N and NN speech averaged across all talkers within that accent. Ellipses in darker colors represent voiced categories, and ellipses in lighter colors represent voiceless categories. Ellipses show 95% CIs based on the variance-covariance matrix averaged across the talker-specific variance-covariance matrices and centered around the mean of the talker-specific means.

characterized by the central tendencies of categories (for review, see Flege, 2007). In study 1, we found that the vowel category means in Mandarin-accented English differed from native-English variants for almost all categories in at least one phonetic dimension (F1 or F2). In study 2, we again replicated previous work that Mandarin-accented speakers maintained a phonological distinction between voiced and voiceless stops in coda position. Their production clearly differed from native-English counterparts (Flege *et al.*, 1992; Flege and Wang, 1989). Mandarin-accented speakers produced greater separability in burst length, a non-primary cue for English coda stop voicing. This corroborates earlier findings (e.g., Flege *et al.*, 1992; Hayes-Harb *et al.*, 2008; Xie *et al.*, 2018). Across both studies, non-native speakers showed less separability for pairs of neighboring categories in a multidimensional acoustic space compared to native speakers. These results are consistent with predictions by PAM and SLM: L2 sounds that are assimilated into a single L1 category—which causes poor discrimination in non-native speakers (e.g., Flege *et al.*, 1997; Jia *et al.*, 2005; Wang and Munro, 1999)—are likely to be less distinguished in production.

Following previous work (see Smith *et al.*, 2019; Vaughn *et al.*, 2019), our primary analysis did not consider the effect of phonetic contexts. A number of studies have shown that L2 listeners are impacted differently than L1 listeners by context effects differently in perception (e.g., Levy, 2009). Relatively little is known, however, about how phonetic context affects L2 *production*. We addressed this question in *post hoc* analyses summarized in the SI. We found that non-native speakers exhibit similar context effects as native speakers, although the effects are often reduced.

One question for future research is *why* non-native speakers might be affected differently by certain phonetic contexts. One possibility is that the reduced effects of phonetic context reflect the fact that these effects are at least partially phonologized and thus need to be *learned*. Another possibility is that the reduced context effects reflect a stylistic difference between careful speech and casual speech—non-native speakers might choose more careful registers when being recorded. Regardless of the specific explanation, this promises to be an interesting venue for future work—in particular, it would seem to raise questions about the extent to which theories of L2 learning can account for the differences in context effects (e.g., van Leussen and Escudero, 2015).

B. Category-specific and cue-specific L1-to-L2 influence in category dispersion

A large body of work has documented the influence of L1 sound systems on L2 phonetic production in terms of category means. Comparatively little is known about how differences and similarities between L1 and L2 phonology impact category variability. The current work compared the across-the-board hypothesis—that non-native speech is generally more variable than native speech—against a more nuanced hypothesis that the influence of L1 phonology on variability in L2 speech production is category and even cue specific. Neither study found support for the across-the-board hypothesis. Combined with results from recent work (e.g., Smith *et al.*, 2019; Vaughn *et al.*, 2019), we conclude that non-native speakers, at least those who are in a late stage of L2 acquisition, are not generally more variable in their production. Therefore, whatever cross-language interaction there is during L2 speech production, it does not seem to pervasively impact the realization of sounds at the phonetic level. Is there, then, any evidence of L1-to-L2 influence on the production variability of specific categories or cues?

In study 1, we examined vowel productions in native American English and Mandarin-accented English. Study 1 thus investigated a case where the phonetic cues (F1 and F2) that define L2 contrasts are present in the L2 speakers' mother tongue. The nine English vowel categories we investigated differ, however, in terms of their similarity to the closest Mandarin vowel categories. We tested the closest-neighbor hypothesis that category variability in L2 is affected by how particular L2 categories are assimilated into

a closest-neighbor L1 category. Specifically, under this hypothesis, we expected no difference between native and non-native speech in category variability for five out of the nine vowels we tested: /ʌ, ɑ, ɔ, i, u/. Tokens of these categories are either good exemplars of the L1 counterparts or they are uncategorized (i.e., they have no parallel L1 sounds). We expected increased category variability for /æ, ɛ, ɪ, ʊ/, which constitute poor exemplars of the closest L1 category. Study 1 found the closest-neighbor hypothesis supported in that the two classes of categories indeed showed different degrees of category variability in non-native speech relative to native speech. However, the results of study 1 also suggested additional differences within each of the two classes of vowels. In particular, we did not observe any difference for /ɪ/ and /ʊ/ in category variability, and /u/ and /æ/ were actually less variable in non-native speech than in native speech.

In the SI,[14] we explored the possibility that native and non-native speakers are affected by the phonetic context differently, and this difference contributes, in part, to the observed difference in vowel category variability. These analyses do not change the support for the across-the-board hypothesis: even after controlling for phonetic context, non-native speakers did not exhibit across-the-board increased category variability. With regard to the two surprising findings, however, we did find that the category variability for /æ/ and /u/ was no longer significantly smaller in non-native speech once context is taking into account. It is therefore possible that phonetic contexts explain these two otherwise surprising results. In the remainder of this section, we discuss alternative (mutually compatible) explanations for these and other findings that further highlight the potential role of L1 phonology in L2 production. Our intent in doing so is to develop specific hypotheses to be tested in future work.

The closest-neighbor hypothesis makes the simplifying assumption that it is sufficient to consider only the closest ("competing") category. This simplifying assumption is not uncommon in research on L2 speech perception (e.g., Flege, 2003; Flege *et al.*, 1997) and production (e.g., Bosch & Ramon-Casas, 2011; or, for that matter, native production, cf. Wedel *et al.*, 2018). While this simplifying assumption can serve as a productive starting point, it is known that the perception and production of an L2 category can be affected by L1 categories beyond the closest counterparts of the L2 sound (Flege, 1995; see also Tyler, 2019). For instance, SLM postulates that L1 and L2 production share a common phonetic space (Flege, 2007). Two consequences follow from this hypothesis: first, additional Mandarin vowels, particularly those which do not exist in English but are in the proximity of English vowel categories, bear influence on L1 Mandarin speakers' production of English vowels; second, as L2 speakers strive to maintain a distinction between both L1 and L2 categories, an L2 category may "dissimilate" from a similar L1 category to the extent that their phonetic space is more crowded than either L1 or L2 alone (Flege *et al.*, 2003). We consider how these two consequences might explain three otherwise unexpected results of study 1: (a) the significantly smaller variability of /u/ in Mandarin-accented English, (b) the lack of increased variability for /ɪ/ and /ʊ/ in Mandarin-accented English, and (c) the significantly smaller variability of /æ/ in Mandarin-accented English.

Regarding (a), Mandarin contrasts three high vowels, /i/-/y/-/u/, whereas English contrasts only two high vowels, /i/-/u/. The presence of a neighbor category /y/ along the F2 dimension potentially exerts pressure on the expansion of /u/ and /i/ in Mandarin. This pressure might impact English /i/ and /u/ as well [see Fig. 9(a)]: either because L2 speakers approximated their production of English /i/ and /u/ to Mandarin /i/ and /u/ or because they attempt to "preserve phonetic contrast among the elements of the L1 and L2 subsystems" (Flege, 2003, p. 487). Either way, we would expect reduced dispersion of these L2 categories in non-native speech compared to native speech (there is no vowel competitor along F2 in English). This is indeed what we observed in study 1 (see Fig. 7): for /u/, F2 variability was significantly smaller in L2 speech; for /i/, the effect went in the same direction but was not significant (see Table V).

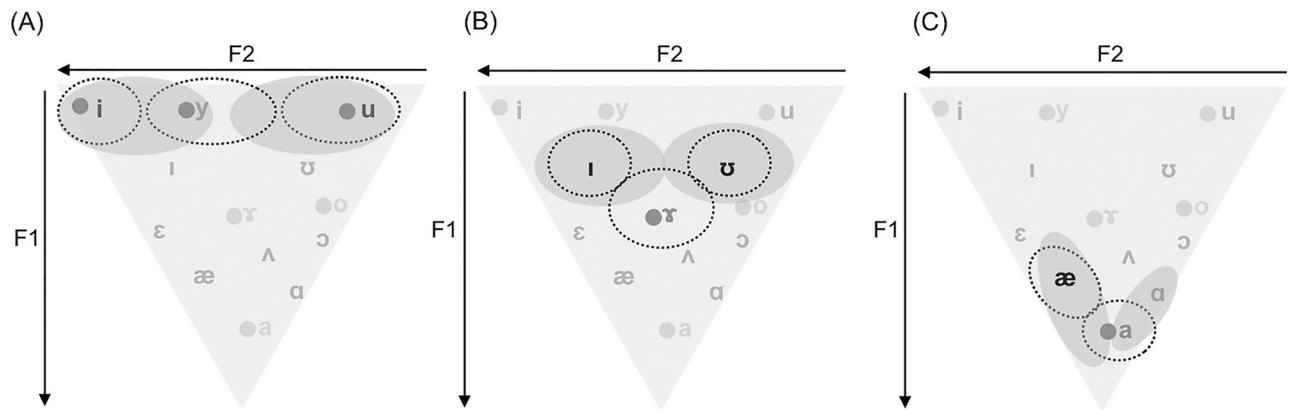

FIG. 9. Hypothesized influences that L1 Mandarin vowel categories might exert on the formation and production of English vowels by native-Mandarin learners of L2 English. (a) The potential expansion of /i/ and /u/ with (shaded ellipses) or without (dotted ellipses) the presence of /y/. (b) The potential expansion of /ɪ/ and /ʊ/ with (shaded ellipses) or without (dotted ellipses) the presence of /ɝ/. (c) The potential expansion of /æ/ with (shaded ellipses) or without (dotted ellipses) the presence of /a/.

https://doi.org/10.1121/10.0001141

Of note, although /y/ is a front vowel, Mandarin /y/ actually has acoustic properties quite similar to English /u/ and, perceptually, it may be less front than indicated by the phonological classification (Chang *et al.*, 2011; similarly, French /y/ is acoustically closer to English /i/ but it is perceptually assimilated to English /u/; see Strange *et al.*, 2004). This would explain why there was a greater decrease in category variability for /u/ than for /i/.

Regarding (b), English has two vowels /ɪ/ and /ʊ/ (closely neighboring /i/ and /u/, respectively), whereas Mandarin does not have similar equivalents. The effects of closest neighbors alone would predict increased variability in non-native speech for these two categories if they are assimilated into /i/ and /u/. On the other hand, Mandarin vowels /y/ and /ɤ/ (neither is present in English) may exert influences on the non-native production of /ɪ/ and /ʊ/. Specifically, although /y/ is unlikely to directly affect the expansion of /ɪ/ and /ʊ/ given their differences in vowel height, the presence of /y/ in Mandarin might impact /ɪ/ and /ʊ/ in the same way it impacts /i/ and /u/ via assimilation. Then, as with /i/ and /u/, we expect to see decreased variability along F2 in Mandarin-accented English for /ɪ/ and /ʊ/. In addition, the Mandarin mid vowel /ɤ/ is argued to be highly variable in its phonetic forms (e.g., Mou *et al.*, 2018). Given a lack of suitable quantitative data on the phonetic distribution of Mandarin /ɤ/, we can only speculate that it probably occupies a rather expanded region near the upper-middle vowel space. Hence, it is possible that it expands into a higher F1 region where English /ɪ/ and /ʊ/ are located [see Fig. 9(b)]. This would put Mandarin /ɤ/ phonetically close to English /ɪ/ and /ʊ/ (despite their clear phonological distinctions), constraining the expansion of /ɪ/ and /ʊ/ in non-native production. Combined with the effect of closest neighbors, the presence of /y/ and /ɤ/ in Mandarin thus offers a potential explanation as to why we found no (significant) differences between the native and non-native category variability for /ɪ/ and /ʊ/.

Last, with regard to (c), /æ/ and /ʊ/ had smaller category variability than expected for the non-native speech. At the same time, they both showed a difference between native and non-native speech in terms of cue covariation structure. As shown in Fig. 7, these changes implied that the shape of the two categories and hence its position relative to other categories in the phonetic space differs between native and non-native speech. It is possible that the shape of Mandarin /u/ bears some influence on F1-F2 covariation in Mandarin-accented English /ʊ/. The fact that both /u/ and /ʊ/ showed a positive F1-F2 correlation in Mandarin-accented speech in study 1 appears to be consistent with past work on Mandarin vowels (Mou *et al.*, 2018). Although Mandarin /a/ and English /ɑ/ are considered phonologically equivalent—both being the point vowel in the same corner of the vowel space, Mandarin /a/ is phonetically closer to /æ/ than English /ɑ/. It is possible that its position, combined with inherent articulatory constraints, has affected the shape and orientation of Mandarin-accented English /æ/ [see Fig. 9(c)]. This change in both the shape and orientation of the /æ/ category would explain the relatively low intelligibility of Mandarin-accented /æ/ in general (e.g., Jia *et al.*, 2006).

In sum, several results that are not predicted by the close-neighbor hypothesis might receive an explanation once we take into account the broader L1 phonological inventory: if some of the neighboring L1 categories (Mandarin) occupy parts of the phonetic space that are not (equally) occupied by L2 categories (English), this could constrain the variability of the L2 category in L2 speakers' productions compared to native speakers of the L2. Depending on the placement of L1 neighbors in F1-F2 space, this can lead to *reduced* variability in non-native compared to native speech or to *increased* variability. At this point, these are hypotheses based on *post hoc* analyses of our data. Adequate future tests of these hypotheses will require large-scale production data from speakers' L1 and L2, ideally while holding phonotactic contexts as comparable as possible across the L1 and L2 productions. This points to a challenging yet important venue for future work.

In study 2, we examined the production of word-final stops in native American English and Mandarin-accented English. We tested the hypothesis that L1 Mandarin speakers' lack of experience of certain acoustic cue manipulation (vowel and closure duration) leads to increased variability in these cues for L2 English contrasts. The present results suggest that reduced experience with contrastive cue use does *not* necessarily make the production more variable as we found no significant difference between native and non-native speech in the variability of coda stop voicing cues. This was the case in terms of both overall category variability and variability along vowel or closure duration alone. This absence of differences in category variability stands in contrast with the clear difference in category means between native and non-native speech.

As we discuss next, the view that the phonetic forms of L2 production originate from L1 phonological constraints might also hold the key to understanding differences in category-specific cue covariation structure between native and non-native speech.

C. L1- L2 differences in cue covariation structure

Simultaneous manipulation of multiple cues during articulation is one type of difficulty that seems to persist even among experienced L2 speakers. Phonetic cues to the same phonological contrast often covary. For instance, shorter voice onset times tend to co-occur with lower F0s. In English, both are more likely in voiced than in voiceless onset stops (e.g., Chodroff and Wilson, 2018; Kingston and Diehl, 1994; Kirby and Ladd, 2015; but see Clayards, 2018). Such covariation can result from general articulatory constraints—for example, certain F1-F2 combinations do not naturally result from typical human anatomy and jaw-cycles (e.g., Schwartz *et al.*, 2012, and references therein)—or reflect language-specific phonology. Research on L2 speech production has investigated to what extent certain cues

covary with category identity and how this affects native perception of non-native speech or non-native perception of native speech (e.g., Holt and Lott, 2006; Schertz *et al.*, 2015). Comparatively little attention has been given to cue covariation structure *within* phonological categories—the present focus.

We examined patterns of cue covariation for individual categories in non-native speech and compared it to native speech. In study 1, differences in cue covariation emerged for two categories that were hypothesized to be difficult based on their perceptual assimilation status: /æ/ and /ʊ/. These differences indicate that /æ/ and /ʊ/ were oriented differently in the native and non-native phonetic space. In study 2, we observed reduced strength of cue correlations in the durational cues for coda stops in non-native speech (weaker vowel-burst correlations and weaker closure-burst correlations). These findings corroborate the concern we raised in the Introduction (see Fig. 1): investigations into the *magnitude* of category dispersion should take into account the *orientation* of dispersion.

Deviations from the expected (native) within-category covariation structure—like those observed in studies 1 and 2—are predicted to contribute to native listeners' perception difficulty by any model that links categories' shape in the phonetic space to categorization (e.g., Kleinschmidt and Jaeger, 2015; Pierrehumbert, 2002; Walsh *et al.*, 2010). For example, non-native orientation of the category dispersion causes the category to interact with and be affected by different neighboring categories than is the case in native speech. As a result, the exact nature of competition among L2 categories may differ between native and non-native speech. This difference may be particularly pronounced when the phonetic space is already crowded—for instance, during vowel production in L2s with many vowel categories. Consider /ɪ/, for example. Although we only examine its separability from /i/ in the current study, it may also overlap with /e/ or /ɛ/ and cause competition with these other categories in perception. A question for future research is whether L1-L2 differences in cue correlations are indeed particularly likely to occur for categories that are absent in L1 and/or for phonetic cues that are not phonologically relevant in L1.

Another important issue to be addressed in future work is the relative contributions of these three potential sources of non-nativeness. Specifically, future work should assess the effect of non-native category means, variability, and cue covariation on both *separability* and *intelligibility*. We use the former term—separability—to refer to the in-principle distinguishability of categories from the perspective of a listener who "knows" the true cue distributions of those categories in non-native speech (such as a listener with sufficient exposure to that type of accent). We use the latter term—intelligibility—to refer to the intelligibility of non-native speech from the perspective of listeners that assume native cue distributions. We anticipate that more fully specified computational models (e.g., Kleinschmidt and Jaeger, 2015; Pajak *et al.*, 2013; Todd *et al.*, 2019) will play a critical role in addressing both of these questions.

D. Limitations and future directions

The present study analyzed non-native speech from relatively proficient L2 speakers. It remains an open question whether production precision changes with L2 experience or production proficiency. It is well established that more experienced L2 learners have more native-like segmental productions in terms of category means (e.g., Bohn and Flege, 1992; Fabra and Romero, 2012; Jia *et al.*, 2006). Focusing on the very beginning stage of L2 acquisition, recent work shows that explicit training significantly reduces category variability in L2 learners' production (Kartushina *et al.*, 2016). Combined with our findings, it is reasonable to believe that as L2 learners become more experienced, they are not only aiming to produce more native-like targets (shift in category means) but also learning to improve precision around the targets (reduction in category variability). Interestingly, there appears to be a link between L1 production and early-stage L2 production in individual talkers' stability of production. Kartushina and colleagues (Kartushina *et al.*, 2016; Kartushina and Fraunfelder, 2014) found that if an L2 category has an acoustically close counterpart in an L2 learner's L1, then the within-talker variability of the L2 category is influenced by how variable the talker is when producing the L1 counterpart category. Such a relationship between L1 and L2 production within a talker may persist even among experienced L2 speakers (Bradlow *et al.*, 2018; Bradlow *et al.*, 2017). It is therefore possible that at any stage of L2 learning, a talker's stability of production is the joint result of individual traits (Bradlow *et al.*, 2018) and the establishment of L2 categorical representations. If one considers novice L2 learners and experienced L2 speakers (as tested in our study) to be at two end points of L2 phonetic learning, it begs the question of what the learning trajectory looks like. In particular, what factors, if any, help to reduce within-talker variability in naturalistic L2 production settings when no explicit feedback is given? It is also an open question where the starting point of articulatory precision would be for L2 contrasts that use cues absent in L1 such as vowel lengths in coda stops.

In the present study, we collected production data using a single task conducted at a single time point. More work is needed to find out whether our findings generalize across tasks and over time. There is evidence, for example, that native speakers are more variable in spontaneous speech compared to reading or elicited speech as we used here (DiCanio *et al.*, 2015). Considering that articulatory control requires effort and L2 speech production is already cognitively more taxing than L1 speech production, it is possible that spontaneous L2 speech would demonstrate even greater within-category variability.

Last, it is important to note that we sampled a relatively homogenous group of L2 speakers who had similar L2 learning experiences via formal classroom instructions and had attained a good level of proficiency. Our data do not speak about whether category variability in L2 speech is affected by L2 proficiency or length of experience on a broader scale.

https://doi.org/10.1121/10.0001141

Research on L2 perception has revealed a tremendous amount of individual variation (e.g., Mayr and Escudero, 2010). So far, experience-related heterogeneity among L2 speakers has almost exclusively focused on talker means. We suggest that characterizing speakers' production variability will likely provide a more comprehensive picture of individual speakers' developmental paths in L2 phonetic production.

V. CONCLUSION

We have investigated differences in category-specific cue distributions between native and non-native speech with a focus on within-talker category variability. Our results suggest that non-native speakers of a second language do not show greater variability across the board. Although there is evidence that interlanguage competition affects L2 production (e.g., Amengual, 2018; Costa *et al.*, 2003), we do not find strong across-the-board effects on category dispersion. Together with other recent work (e.g., Vaughn *et al.*, 2019), this casts doubt on the assumption that non-native speakers are inherently more variable in speech production.

Instead, we observe category- and cue-specific effects from L1 to L2 transfer, manifested in category means, category variability, as well as cue covariation within a category. The nature of these effects is broadly compatible with the principles of existing theories on L2 speech perception and production. At the same time, we also see that a simplifying assumption employed in much of the empirical tests of these theories—the focus on the closest L1 neighbor—is limiting. Here, we have discussed that one additional source of L1-to-L2 influence, namely the presence and location of additional L1 categories—categories beyond similar L1 categories to which L2 categories are assimilated—also constrains the dispersion of L2 categories in the phonetic space.

Research in L2 speech phonetics has traditionally focused on the central tendency in talkers' production. Recently, more work on native language learning has looked beyond category means and started to quantify the influence of the kind of token-to-token variability on perception and phonetic representations (e.g., Clayards *et al.*, 2008; Nixon *et al.*, 2016; Kronrod *et al.*, 2016). We therefore consider it particularly important for future work to develop databases that contain distributional cue information of native and non-native speech of the target L2, as well as non-native speakers' L1 language, preferably using speech instances from similar phonological contexts.

ACKNOWLEDGMENTS

Analyses presented here depend on a speech database designed by X.X. and Ruolan Li. Annotations were provided by Ruolan Li and Nicole Viyeto. We are grateful for feedback from Zachary Burchill, Wednesday Bushong, Chigusa Kurumada, Shaorong Yan, and, in particular, Ruolan Li. This work was funded in part by the National Institutes of Health (NIH) Grant No. R01 HD075797 to T.F.J. The views expressed here do not necessarily reflect those of the funding agency.

TABLE XII. Stimuli used in study 1.

Word	Vowel	Word	Vowel	Word	Vowel	Word	Vowel
Drug	ʌ	Draw	ɔ	Pet	ɛ	Lib	I
Tuck	ʌ	Cords	ɔ	Said	ɛ	Live	I
Bud	ʌ	Cores	ɔ	Ten	ɛ	Sin	I
Duck	ʌ	Falls	ɔ	Vet	ɛ	Sing	I
Gum	ʌ	Gong	ɔ	Wren	ɛ	Thin	I
Rung	ʌ	Moss	ɔ	Beach	i	Whip	I
Some	ʌ	Gone	ɔ	Beak	i	Win	I
Sun	ʌ	Batch	æ	Clean	i	Wing	I
But	ʌ	Crack	æ	Glean	i	Killed	I
Cut	ʌ	Patch	æ	Green	i	Lick	I
Dug	ʌ	Lab	æ	Peach	i	Lip	I
Fun	ʌ	Lad	æ	Peak	i	Silt	I
Suck	ʌ	Lap	æ	Please	i	Sit	I
Rum	ʌ	Path	æ	Breathe	i	Gild	I
Run	ʌ	Ram	æ	Feed	i	Gloom	u
Gun	ʌ	Ran	æ	Feet	i	Groom	u
Clock	ɑ	Rang	æ	Fields	i	Doom	u
Cards	ɑ	Sack	æ	Neat	i	Lose	u
Carve	ɑ	Sag	æ	Need	i	Fool	u
God	ɑ	Tab	æ	Peas	i	Pool	u
Got	ɑ	Tap	æ	Seem	i	Room	u
Mob	ɑ	Bad	æ	Seen	i	Soon	u
Mop	ɑ	Bag	æ	Team	i	Cooed	u
Cop	ɑ	Pan	æ	Teen	i	Good	ʊ
Fall	ɑ	Pat	æ	Beads	i	Bull	ʊ
Fond	ɑ	Sad	æ	Beep	i	Look	ʊ
Not	ɑ	Tan	æ	Clear	I	Pull	ʊ
Pond	ɑ	Sends	ɛ	Crick	I	Put	ʊ
Pot	ɑ	Bed	ɛ	Dig	I	Soot	ʊ
Rod	ɑ	Beg	ɛ	Thick	I		
Shot	ɑ	Mess	ɛ	Thing	I		
Hop	ɑ	Pen	ɛ	Trip	I		

TABLE XIII. Stimuli used in Study 2.

Word	Stop	Word	Stop	Word	Stop	Word	Stop
Lab	b	Cooed	d	Cut	t	Sit	t
Lib	b	Feed	d	Feet	t	Bag	g
Mob	b	Find	d	Got	t	Beg	g
Robe	b	Fond	d	Late	t	Dig	g
Tab	b	Gild	d	Neat	t	Drug	g
Beep	p	God	d	Night	t	Dug	g
Cop	p	Good	d	Not	t	Sag	g
Hop	p	Killed	d	Pat	t	Beak	k
Lap	p	Lad	d	Pet	t	Clock	k
Lip	p	Need	d	Pot	t	Crack	k
Mop	p	Pond	d	Pout	t	Crick	k
Rope	p	Pound	d	Put	t	Duck	k
Tap	p	Ride	d	Shout	t	Lick	k
Trip	p	Rode	d	Silt	t	Look	k
Whip	p	Sad	d	Soot	t	Peak	k
Wipe	p	Said	d	Vet	t	Sack	k
Bad	d	Slide	d	Wait	t	Suck	k
Bed	d	Rod	d	White	t	Thick	k
Bud	d	But	t	Shot	t	Tuck	k

TABLE XIV. Language background information of NN speakers. NA, nonapplicable.

Speaker	Regional dialect of Mandarin	Age of L2 acquisition (yr)	Age of arrival in U.S. (yr)	Length of residence (months)
1	Taiwan	5	16	NA
2	Beijing	6	19	6
3	Taiwan	10	15	NA
4	Beijing	12	NA	NA
5	Shanghai	7	18	5
6	Chengdu	11	NA	24
7	Jiangsu	7	15	36
8	Shandong	12	26	42
9	Tianjin	7	24	60
10	Beijing	12	NA	NA

APPENDIX A: STIMULI MATERIALS

This appendix provides all stimuli analyzed in studies 1 and 2 (see Tables XII and XIII, respectively). Each word was repeated three times by each speaker.

APPENDIX B: LANGUAGE BACKGROUND INFORMATION OF NON-NATIVE SPEAKERS

This appendix provides information about the language background of the non-native speakers (L1 Mandarin, L2 English) used in studies 1 and 2 (see Table XIV). All speakers were male and enrolled at a U.S. university as undergraduate or graduate students at the time of recording. All speakers learned English as an L2 in classroom settings in Mandarin-speaking regions. The column showing "Regional dialect" in Table XIV indicates the dialectal region of the speakers.

[1]Wade *et al.* (2007) presented but did not analyze average within-category correlations for native- and Spanish-accented English talkers. For differences between native and non-native speech in the correlations of category means *across* categories and talkers, see Chodroff and Baese-Berk (2019).

[2]The English mid vowels /e, o/ are often realized as diphthongs and are not considered here. We consider /ɑ/-/ɔ/ as phonologically distinct in General American English (Hillenbrand *et al.*, 1995). As the results show, our speakers do maintain a /ɑ/-/ɔ/ distinction.

[3]In the terminology of PAM, /ʊ/ and /ɪ/ involve *category-goodness assimilation*, where another nearby L2 category is the better exemplar of the closest L1 neighbor (Best, 1995). Both /æ/ and /ɛ/ involve *single-category assimilation*—they are both poor exemplars of Mandarin /ɤ/, whose allophonic variants occupy a large F1-F2 space, including *some* of the space occupied by English /æ/ and /ɛ/ (Mok, 2012; Chen *et al.*, 2001). In line with this classification, English /æ/ and /ɛ/ are mutually confusable for Mandarin speakers (e.g., Chen *et al.*, 2001; Jia *et al.*, 2006; Thomson *et al.*, 2009).

[4]Mandarin has no unrounded mid-low back vowels like /ʌ/. Through reference to descriptions in past work (Chen *et al.*, 2001), we determine that /ʌ/ is likely to be between Mandarin /a/ and /ɤ/ without being particularly close to either of these two L1 categories.

[5]Smith *et al.* (2019) avoided this potential confound by only investigating *one* phonetic context (*h*V*d* words). This does, however, leave open the notion of whether their findings generalize to other phonetic contexts.

[6]We note that Lobanov-normalization has the *potential* to affect the comparison of category means and dispersion in unintended ways. Specifically, Lobanov-normalization involves division by the SD of formant values across all categories. This SD is a function of both the *within-category* variability of the formant summed across all categories and the *across-category* variability in the category means of that formant. Similar potential concerns would apply to any normalization that corrects for overall F1/F2 variability across vowels. Critically, additional analyses over raw non-normalized formant values reported in the SI avoid this issue and replicate all findings. No other analyses were conducted unless explicitly mentioned.

[7]The *R* formula for these models is category statistic (e.g., here, category mean) $\sim 1 +$ vowel/accent $+ (1|$speaker). Since we are comparing by-speaker statistics (instead of token-level statistics), it was neither required nor possible to include by-word random effects.

[8]The only exceptions were the F1 of /i/ and /ɑ/ and the F2 of /u/, which no longer differ significantly between native and non-native speech once phonetic context is controlled for. We note, however, that this might simply reflect the loss in power resulting from the small amount of data available for each context (see SI).

[9]The SI presents additional analyses using an alternative measure of between-category separability. The results are qualitatively identical to those presented here. The alternative measure is based on the average distance of vowel tokens from one category to all tokens of the competing category. This measure thus takes into account the token dispersion of the competing category instead of just the competing category's center. Similar measures have been employed in studies on phonetic competition (see McCloy *et al.*, 2015; Xie and Myers, 2018).

[10]The SI reports additional analyses for an alternative measure of the overall F1-F2 variability: the product of the eigenvalues of the category's F1-F2 covariance matrix. This product is proportional to the size of the ellipse area (squared) covered by a category (e.g., Wade *et al.*, 2007; Mou *et al.*, 2018). The results of this additional analysis were consistent with the results reported here but were more conservative.

[11]Specifically, burst omission is *not* an extreme realization of the gradient burst cue: whereas released bursts are, on average, shorter (i.e., closer to zero duration) for voiced stops, burst omission is vastly more common for voiceless stops (72% of omissions) than voiced stops (28% of omissions). Recoding omitted bursts as zero duration bursts thus would make no sense (see SI for further detail). The SI also reports auxiliary analyses (requested by an anonymous reviewer) of all vowel durations, not excluding those from tokens without released burst. These analyses replicate all results reported in the main text.

[12]Although non-native speakers have been found to speak more slowly in general (e.g., Baese-Berk and Morrill, 2015), there was no difference ($p = 0.53$) in speaking rates between native and non-native speaker groups used in the present study in the present tasks as measured by average word duration (514 ms in native speech vs 496 ms in non-native speech). This is consistent with Sadat *et al.* (2012), who found duration differences only in phrases and not in isolated nouns.

[13]This correction was unnecessary in Study 1 because Lobanov-normalization successfully removed correlations between SDs and the mean. The SI presents analyses of SDs for Study 2 to facilitate further comparison to study 1.

[14]See supplementary material at https://doi.org/10.1121/10.0001141 for detailed analyses on the effects of phonetic context on vowel production.

Abramson, A. S., and Tingsabadh, K. (**1999**). "Thai final stops: Cross-language perception," Phonetica **56**(3–4), 111–122.

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (**2009**). "Comprehension of familiar and unfamiliar native accents under adverse listening conditions," J. Exp. Psychol. Hum. Percept. Perform. **35**(2), 520–529.

Allen, J. S., and Miller, J. L. (**1999**). "Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words," J. Acoust. Soc. Am. **106**(4), 2031–2039.

Amengual, M. (**2018**). "Asymmetrical interlingual influence in the production of Spanish and English laterals as a result of competing activation in bilingual language processing," J. Phonetics **69**, 12–28.

Antoniou, M., Tyler, M. D., and Best, C. T. (**2012**). "Two ways to listen: Do L2-dominant bilinguals perceive stop voicing according to language mode?," J. Phonetics **40**(4), 582–594.

Baayen, R. H., Davidson, D. J., and Bates, D. M. (**2008**). "Mixed-effects modeling with crossed random effects for subjects and items," J. Mem. Lang. **59**(4), 390–412.

Baese-Berk, M. M., and Morrill, T. H. (**2015**). "Speaking rate consistency in native and non-native speakers of English," J. Acoust. Soc. Am. **138**(3), EL223–EL228.
Bent, T., Bradlow, A. R., and Smith, B. L. (**2008**). "Production and perception of temporal patterns in native and non-native speech," Phonetica **65**(3), 131–147.
Berthele, R. (**2019**). "Policy recommendations for language learning: Linguists' contributions between scholarly debates and pseudoscience," J. Euro. Sec. Lang. Assoc. **3**(1), 1–11.
Best, C. T. (**1994**). "The emergence of native-language phonological influences in infants: A perceptual assimilation model," Dev. Speech Percept. **167**(224), 233–277.
Best, C. T. (**1995**). "A direct realist view of cross-language speech perception," in *Speech Perception and Linguistic Experience* (York Press, Timonium, MD), pp. 171–206.
Best, C. T., and Tyler, M. D. (**2007**). "Nonnative and second-language speech perception: Commonalities and complementarities," in *Language Experience in Second Language Speech Learning: In Honor of James Emil Flege* (John Benjamins, Amsterdam), Vol. 1334, pp. 1–47.
Boersma, P., and Weenink, D. (**2018**). "Praat: Doing phonetics by computer (version 6.0.40) [computer program]," http://www.praat.org (Last viewed 10/25/2019).
Bohn, O. S., and Flege, J. E. (**1992**). "The production of new and similar vowels by adult German learners of English," Stud. Second Lang. Acquis. **14**(2), 131–158.
Bosch, L., and Ramon-Casas, M. (**2011**). "Variability in vowel production by bilingual speakers: Can input properties hinder the early stabilization of contrastive categories?," J. Phonetics **39**(4), 514–526.
Bradlow, A. R. (**1995**). "A comparative acoustic study of English and Spanish vowels," J. Acoust. Soc. Am. **97**(3), 1916–1924.
Bradlow, A. R., Kim, M., and Blasingame, M. (**2017**). "Language-independent talker-specificity in first-language and second-language speech production by bilingual talkers: L1 speaking rate predicts L2 speaking rate," J. Acoust. Soc. Am. **141**(2), 886–899.
Bradlow, A. R., Blasingame, M., and Lee, K. (**2018**). "Language-independent talker-specificity in bilingual speech intelligibility: Individual traits persist across first-language and second-language speech," Journal of the Association for Laboratory Phonology **9**(1).
Chang, C. B., Yao, Y., Haynes, E. F., and Rhodes, R. (**2011**). "Production of phonetic and phonological contrast by heritage speakers of Mandarin," J. Acoust. Soc. Am. **129**(6), 3964–3980.
Chen, Y., Robb, M., Gilbert, H., and Lerman, J. (**2001**). "Vowel production by Mandarin speakers of English," Clin. Linguist. Phon. **15**(6), 427–440.
Chodroff, E., and Baese-Berk, M. (**2019**). "Constraints on variability in the voice onset time of L2 English stop consonants," in *Proceedings of the 19th International Congress of Phonetic Sciences. International Congress of Phonetic Sciences*, August 4-10, Australia.
Chodroff, E., and Wilson, C. (**2018**). "Predictability of stop consonant phonetics across talkers: Between-category and within-category dependencies among cues for place and voice," Linguist. Vanguard **4**(s2), 20170047.
Clayards, M. (**2018**). "Individual talker and token covariation in the production of multiple cues to stop voicing," Phonetica **75**(1), 1–23.
Clayards, M., Tanenhaus, M. K., Aslin, R. N., and Jacobs, R. A. (**2008**). "Perception of speech reflects optimal use of probabilistic speech cues," Cognition **108**(3), 804–809.
Connine, C. M., Blasko, D. G., and Wang, J. (**1994**). "Vertical similarity in spoken word recognition: Multiple lexical activation, individual differences, and the role of sentence context," Percept. Psychophys. **56**(6), 624–636.
Costa, A., Colomé, À., Gómez, O., and Sebastián-Gallés, N. (**2003**). "Another look at cross-language competition in bilingual speech production: Lexical and phonological factors," Bilingualism: Lang. Cognit. **6**(3), 167–179.
Cutler, A., Weber, A., and Otake, T. (**2006**). "Asymmetric mapping from phonetic to lexical representations in second-language listening," J. Phonetics **34**(2), 269–284.
Darcy, I., and Krüger, F. (**2012**). "Vowel perception and production in Turkish children acquiring L2 German," J. Phonetics **40**(4), 568–581.
Deelman, T., and Connine, C. M. (**2001**). "Missing information in spoken word recognition: Nonreleased stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **27**(3), 656.
DiCanio, C., Nam, H., Amith, J. D., García, R. C., and Whalen, D. H. (**2015**). "Vowel variability in elicited versus spontaneous speech: Evidence from Mixtec," J. Phonetics **48**, 45–59.
Escudero, P. (**2005**). "Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization," Ph.D. thesis, LOT Dissertation Series 113, Utrecht University.
Escudero, P., and Bion, R. A. H. (**2007**). "Modeling vowel normalization and sound perception as sequential processes," in *Proceedings of the 16th International Congress of Phonetic Sciences*, pp. 1413–1416.
Evanini, K., and Huang, B. (**2012**). "Production of English vowels by speakers of Mandarin Chinese with prolonged exposure to English," Proc. Meet. Acoust. **18**(1), 060004.
Fabra, L. R., and Romero, J. (**2012**). "Native Catalan learners' perception and production of English vowels," J. Phonetics **40**(3), 491–508.
Feldman, N. H., Griffiths, T. L., and Morgan, J. L. (**2009**). "The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference," Psych.l Rev. **116**(4), 752.
Flege, J. E. (**1987**). "The production of 'new' and 'similar' phones in a foreign language: Evidence for the effect of equivalence classification," J. Phonetics **15**(1), 47–65.
Flege, J. E. (**1989**). "Differences in inventory size affect the location but not the precision of tongue positioning in vowel production," Lang. Speech **32**(2), 123–147.
Flege, J. E. (**1995**). "Second language speech learning: Theory, findings, and problems," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Research*, edited by W. Strange (York Press, Timonium, MD), Vol. 92, pp. 233–277.
Flege, J. E. (**2007**). "Language contact in bilingualism: Phonetic system interactions," Lab. Phonol. **9**, 353-382.
Flege, J. E., Bohn, O. S., and Jang, S. (**1997**). "Effects of experience on non-native speakers' production and perception of English vowels," J. Phonetics **25**(4), 437-470.
Flege, J. E., and Eefting, W. (**1987**). "Cross-language switching in stop consonant perception and production by Dutch speakers of English," Speech Commun. **6**(3), 185–202.
Flege, J. E., and Hillenbrand, J. (**1987**). "A differential effect of release bursts on the stop voicing judgments of native French and English listeners," J. Phonetics **15**(2), 203–208.
Flege, J. E., and Liu, S. (**2001**). "The effect of experience on adults' acquisition of a second language," Stud. Second Lang. Acquis. **23**(4), 527–552.
Flege, J. E., and MacKay, I. R. (**2004**). "Perceiving vowels in a second language," Stud. Second Lang. Acquis. **26**(1), 1–34.
Flege, J. E., Munro, M. J., and Skelton, L. (**1992**). "Production of the word-final English /t/–/d/contrast by native speakers of English, Mandarin, and Spanish," J. Acoust. Soc. Am. **92**(1), 128–143.
Flege, J. E., Schirru, C., and MacKay, I. R. (**2003**). "Interaction between the native and second language phonetic subsystems," Speech Commun. **40**(4), 467–491.
Flege, J. E., and Wang, C. (**1989**). "Native-language phonotactic constraints affect how well Chinese subjects perceive the word-final English /t/-/d/ contrast," J. Phonetics **17**(4), 299–315.
Gahl, S., Yao, Y., and Johnson, K. (**2012**). "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech," J. Mem. Lang. **66**(4), 789–806.
Goldrick, M., Runnqvist, E., and Costa, A. (**2014**). "Language switching makes pronunciation less nativelike," Psychol. Sci. **25**(4), 1031–1036.
Hayes-Harb, R., Smith, B. L., Bent, T., and Bradlow, A. R. (**2008**). "The interlanguage speech intelligibility benefit for native speakers of Mandarin: Production and perception of English word-final voicing contrasts," J. Phonetics **36**(4), 664–679.
Hazan, V., Romeo, R., and Pettinato, M. (**2013**). "The impact of variation in phoneme category structure on consonant intelligibility," Proc. Meet. Acoust. **19**(1), 060103.
Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**(5), 3099–3111.
Holt, L. L., and Lotto, A. J. (**2006**). "Cue weighting in auditory categorization: Implications for first and second language acquisition," J. Acoust. Soc. Am. **119**(5), 3059–3071.
James, A. R. (**1984**). "Syntagmatic segment errors in non-native speech," Linguistics **22**(4), 481–506.

Jia, G., Strange, W., Wu, Y., Collado, J., and Guan, Q. (**2006**). "Perception and production of English vowels by Mandarin speakers: Age-related differences vary with amount of L2 exposure," J. Acoust. Soc. Am. **119**(2), 1118–1130.
Jongman, A., and Wade, T. (**2007**). "Acoustic variability and perceptual learning," in *Language Experience in Second Language Speech Learning* (John Benjamins, Amsterdam), pp. 135–150.
Kartushina, N., and Frauenfelder, U. H. (**2014**). "On the effects of L2 perception and of individual differences in L1 production on L2 pronunciation," Front. Psychol. **5**, 1246.
Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., and Golestani, N. "Mutual influences between native and non-native vowels in production: Evidence from short-term visual articulatory feedback training," J. Phonetics **57**, 21–39 (**2016**).
Kingston, J., and Diehl, R. L. (**1994**). "Phonetic knowledge," Language **70**(3), 419–454.
Kirby, J. P., and Ladd, D. R. (**2015**). "Stop voicing and F0 perturbations: Evidence from French and Italian," in *International Congress of Phonetic Sciences*.
Klatt, D. H. (**1975**). "Voice onset time, frication, and aspiration in word-initial consonant clusters," J. Speech Hear. Res. **18**(4), 686–706.
Kleinschmidt, D. F., and Jaeger, T. F. (**2015**). "Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel," Psychol. Rev. **122**(2), 148–203.
Kronrod, Y., Coppess, E., and Feldman, N. H. (**2016**). "A unified account of categorical effects in phonetic perception," Psychon. Bull. Rev. **23**(6), 1681–1712.
Lahiri, A., and Marslen-Wilson, W. (**1991**). "The mental representation of lexical form: A phonological approach to the recognition lexicon," Cognition **38**(3), 245-294.
Labov, W. (**2006**). "A sociolinguistic perspective on sociophonetic research," J. Phonetics **34**(4), 500-515.
Levy, E. S. (**2009**). "Language experience and consonantal context effects on perceptual assimilation of French vowels by American-English learners of French," J. Acoust. Soc. Am. **125**(2), 1138-1152.
Liljencrants, J., and Lindblom, B. (**1972**). "Numerical simulation of vowel quality systems: The role of perceptual contrast," Language 839-862.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**(3), 384–422.
Luce, P. A., and Pisoni, D. B. (**1998**). "Recognizing spoken words: The neighborhood activation model," Ear and Hearing **19**(1), 1.
Mayr, R., and Escudero, P. (**2010**). "Explaining individual variation in L2 perception: Rounded vowels in English learners of German," Bilingual.: Lang. Cognit. **13**(3), 279–297.
McClelland, J. L., and Elman, J. L. (**1986**). "The TRACE model of speech perception," Cognitive psychology **18**(1), 1-86.
McCloy, D. R., Wright, R. A., and Souza, P. E. (**2015**). "Talker versus dialect effects on speech intelligibility: A symmetrical study," Lang. Speech **58**(3), 371–386.
Mok, P. P. (**2013**). "Does vowel inventory density affect vowel-to-vowel coarticulation?," Language and Speech **56**(2), 191-209.
Moon, S. J., and Lindblom, B. (**1994**). "Interaction between duration, context, and speaking style in English stressed vowels," J. Acoust. Soc. Am. **96**(1), 40-55.
Moreton, E. (**2004**). "Realization of the English postvocalic [voice] contrast in F1 and F2," J. Phonetics **32**(1), 1–33.
Mou, Z., Chen, Z., Yang, J., and Xu, L. (**2018**). "Acoustic properties of vowel production in Mandarin-speaking patients with post-stroke dysarthria," Scientific Rep. **8**(1), 14188.
Munro, M. J., and Derwing, T. M. (**1995**). "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," Lang. Learn. **45**(1), 73–97.
Newman, R. S., Clouse, S. A., and Burnham, J. L. (**2001**). "The perceptual consequences of within-talker variability in fricative production," J. Acoust. Soc. Am. **109**(3), 1181–1196.
Nixon, J. S., van Rij, J., Mok, P., Baayen, R. H., and Chen, Y. (**2016**). "The temporal dynamics of perceptual uncertainty: Eye movement evidence from Cantonese segment and tone perception," J. Mem. Lang. **90**, 103–125.
Oh, Y. R., Kim, M., and Kim, H. K. (**2008**). "Acoustic and pronunciation model adaptation for context-independent and context-dependent pronunciation variability of non-native speech," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, March, pp. 4281–4284.
Olson, D. J. (**2013**). "Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production," J. Phonetics **41**(6), 407–420.
Owren, M. J. (**2008**). "GSU Praat Tools: Scripts for modifying and analyzing sounds using Praat acoustics software," Behavior Research Methods **40**, 822–829.
Pajak, B., Bicknell, K., and Levy, R. (**2013**). "A model of generalization in distributional learning of phonetic categories," in *Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, August, pp. 11–20.
Pierrehumbert, J. B. (**2002**). "Word-specific phonetics," in *Laboratory Phonology VII*, edited by C. Gussenhoven and N. Warner (de Gruyter, Berlin), pp. 101–139.
Raphael, L. J. (**1972**). "Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English," J. Acoust. Soc. Am. **51**(4B), 1296–1303.
Romeo, R., Hazan, V., and Pettinato, M. (**2013**). "Developmental and gender-related trends of intra-talker variability in consonant production," J. Acoust. Soc. Am. **134**(5), 3781–3792.
Romero-Rivas, C., Martin, C. D., and Costa, A. (**2015**). "Processing changes when listening to foreign-accented speech," Front. Hum. Neurosci. **9**, 167.
Sadat, J., Martin, C. D., Alario, F. X., and Costa, A. (**2012**). "Characterizing the bilingual disadvantage in noun phrase production," J. Psycholinguist. Res. **41**(3), 159–179.
Schertz, J., Cho, T., Lotto, A., and Warner, N. (**2015**). "Individual differences in phonetic cue use in production and perception of a non-native sound contrast," J. Phonetics **52**, 183–204.
Schmale, R., Hollich, G., and Seidl, A. (**2011**). "Contending with foreign accent in early word learning," J. Child Lang. **38**(5), 1096–1108.
Schwartz, J. L., Boë, L. J., Badin, P., and Sawallis, T. R. (**2012**). "Grounding stop place systems in the perceptuo-motor substance of speech: On the universality of the labial–coronal–velar stop series," J. Phonetics **40**(1), 20–36.
Smith, B. L., Johnson, E., and Hayes-Harb, R. (**2019**). "ESL learners' intra-speaker variability in producing American English tense and lax vowels," J. Second Lang. Pronunc. **5**(1), 139–164.
Smith, B. L., and Kenney, M. K. (**1994**). "Variability control in speech production tasks performed by adults and children," J. Acoust. Soc. Am. **96**(2), 699–705.
Stibbard, R. (**2004**). "The spoken English of Hong Kong: A study of co-occurring segmental errors," Language, Culture and Curriculum **17**(2), 127–142.
Strange, W., Levy, E., and Lehnholf, R., Jr. (**2004**). "Perceptual assimilation of French and German vowels by American English monolinguals: Acoustic similarity does not predict perceptual similarity," J. Acoust. Soc. Am. **115**(5), 2606.
Thomson, R. I., and Derwing, T. M. (**2014**). "The effectiveness of L2 pronunciation instruction: A narrative review," Appl. Linguist. **36**(3), 326–344.
Thomson, R. I., Nearey, T. M., and Derwing, T. M. (**2009**). "A modified statistical pattern recognition approach to measuring the crosslinguistic similarity of Mandarin and English vowels," J. Acoust. Soc. Am. **126**(3), 1447–1460.
Todd, S., Pierrehumbert, J. B., and Hay, J. (**2019**). "Word frequency effects in sound change as a consequence of perceptual asymmetries: An exemplar-based model," Cognition **185**, 1–20.
Tyler, M. D. (**2019**). *PAM-L2 and Phonological Category Acquisition in the Foreign Language Classroom* (Aarhus University, Denmark).
Van Leussen, J.-W., and Escudero, P. (**2015**). "Learning to perceive and recognize a second language: The L2LP model revised," Front. Psychol. **6**, 1000.
Vaughn, C., Baese-Berk, M., and Idemaru, K. (**2019**). "Re-examining phonetic variability in native and non-native speech," Phonetica **76**(5), 327–358.
Wade, T., Jongman, A., and Sereno, J. (**2007**). "Effects of acoustic variability in the perceptual learning of non-native-accented speech sounds," Phonetica **64**(2-3), 122–144.
Walsh, M., Möbius, B., Wade, T., and Schütze, H. (**2010**). "Multilevel exemplar theory," Cognit. Sci. **34**(4), 537–582.

Wang, X., and Munro, M. J. (**1999**). "The perception of English tense-lax vowel pairs by native Mandarin speakers: The effect of training on attention to temporal and spectral cues," in *Proceedings of the 14th International Congress of Phonetic Sciences*, University of California, Berkeley, CA, Vol. 3, pp. 125–128.

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

Wedel, A., Nelson, N., and Sharp, R. (**2018**). "The phonetic specificity of contrastive hyperarticulation in natural speech," J. Mem. Lang. **100**, 61–88.

Weil, S. A. (**2003**). "The impact of perceptual dissimilarity on the perception of foreign accented speech," Doctoral dissertation, The Ohio State University.

Whiteside, S. P., Dobbin, R., and Henry, L. (**2003**). "Patterns of variability in voice onset time: A developmental study of motor speech skills in humans," Neurosci. Lett. **347**(1), 29–32.

Wiese, R. (**1997**). "Underspecification and the description of Chinese vowels," in *Studies in Chinese Phonology*, edited by W. Jialing and N. Smith (Mouton de Gruyter, Berlin), pp. 219–249.

Witteman, M. J., Weber, A., and McQueen, J. M. (**2014**). "Tolerance for inconsistency in foreign-accented speech," Psychon. Bull. Rev. **21**(2), 512–519.

Wright, R. (**2004**). "Factors of lexical competition in vowel articulation," in *Laboratory Phonology VI*, edited by J. Local, R. Ogden, and R. Temple (Cambridge University Press, New York), pp. 75–87.

Xie, X., and Fowler, C. A. (**2013**). "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English," J. Phonetics **41**(5), 369–378.

Xie, X., and Myers, E. (**2018**). "Left inferior frontal gyrus sensitivity to phonetic competition in receptive language processing: A comparison of clear and conversational speech," J. Cognit. Neurosci. **30**(3), 267–280.

Xie, X., Theodore, R. M., and Myers, E. B. (**2017**). "More than a boundary shift: Perceptual adaptation to foreign-accented speech reshapes the internal structure of phonetic categories," J. Exp. Psychol. Hum. Percept. Perform. **43**(1), 206.

Xie, X., Weatherholtz, K., Bainton, L., Rowe, E., Burchill, Z., Liu, L., and Jaeger, T. F. (**2018**). "Rapid adaptation to foreign-accented speech and its transfer to an unfamiliar talker," J. Acoust. Soc. Am. **43**(4), 2013–2031.

Yuan, J., and Liberman, M. (**2008**). "Speaker identification on the SCOTUS corpus," J. Acoust. Soc. Am. **123**(5), 3878.