GLOSSA psycholinguistics

Schertz, J., Adil, F., & Kravchuk, A. (2023). Underpinnings of explicit phonetic imitation: perception, production, and variability. *Glossa Psycholinguistics*, 2(1): 4, pp. 1–51. DOI: https://doi.org/10.5070/G601123

eScholarship University of California

Underpinnings of explicit phonetic imitation: perception, production, and variability

Jessamyn Schertz, Department of Language Studies, University of Toronto Mississauga, Canada; Department of Linguistics, University of Toronto, Canada, jessamyn.schertz@utoronto.ca

Fatima Adil, Department of Language Studies, University of Toronto Mississauga, Canada

A. Kravchuk, Department of Language Studies, University of Toronto Mississauga, Canada

This work tests the relative role of perception- and production-based predictors, and the relationship between them, in imitation of artificial accents varying in voice onset time (VOT), using a paradigm designed to target distinct sub-processes of imitation. We examined how explicit imitation of sentences differing systematically in voice onset time (VOT) was influenced by the type of VOT manipulation (lengthened vs. shortened) and by the presence vs. absence of voice-related variability in exposure. In contrast to previous work, participants imitated shortened as well as lengthened VOT, albeit with both qualitative and quantitative differences across the two manipulation types. The presence of voice-related variability inhibited imitation, but this inhibition was mitigated by a preceding session with no voice-related variability (i.e., sentences were acoustically identical except for VOT). We then tested the extent to which individual performance on the accent imitation task was related to performance on three other tasks: 1) discrimination of the target accents, 2) imitation of words in isolation drawn from a VOT continuum, and 3) discrimination of these same words. Performance on the accent discrimination task and the word-level imitation task, but not the word-level discrimination task, were independently predictive of accent imitation. Results are consistent with a conceptualization of explicit imitation as the sum of automatic phonetic convergence processes overlaid with distinct, controlled perceptual and articulatory factors that pattern differently across individuals. Phonetic imitation should not be considered as a monolithic skill, and models predicting variation in imitative ability must consider not only the potential sources of individual variability, but also at what level these sources of variability exert their influence.

1. Introduction

Phonetic imitation is a complex behavior, requiring accurate perception, identification, and (re-) production of the relevant characteristics of the target of imitation. Previous work has shown that listeners are sensitive to, and can to some extent spontaneously reproduce, phonetic properties of familiar accents when asked to mimic them from memory (e.g., Flege and Hammond, 1982; Mora et al., 2014), and that they can imitate properties of novel (D'Imperio et al., 2014; Spinu et al., 2018) or artificial (Spinu et al., 2020) accents, or phonetic variation in words or syllables (Olmstead et al., 2013; Dufour & Nguyen, 2013). The variability found within these studies reflects the wide range of imitative ability in the real world.

Compared to the substantial body of work on implicit phonetic convergence, there has been less examination of speakers' performance when explicitly instructed to imitate the phonetic properties of another talker. This work is an initial step in a broader research program working toward an accurate model of the processes underlying explicit phonetic imitation and the details of its conditioning factors. Using a novel paradigm designed to test imitation and its sub-processes, we test the relative roles of perceptual and articulatory sub-processes in predicting individual variability in imitation of artificial accents characterized by shortened and lengthened voice onset time (VOT), and we explore how talker variability might constrain this sort of imitation.

1.1 The cognitive architecture of explicit imitation

The term *phonetic imitation* encompasses a wide range of phenomena, and it is used in studies differing in the target of imitation (e.g., words in isolation vs. natural stretches of speech by an interlocutor), instructions (e.g., with vs. without an explicit directive to imitate), and task (e.g., speech shadowing vs. natural conversation). The current work focuses on explicit imitation of systematic phonetic variation, using a task in which participants are instructed to imitate English sentences characterized by different "accents" that vary minimally from canonical speech in a single phonetic feature: the VOT of voiceless stops.

Faithful imitation requires successful realization of several processes that are each on their own quite complex. As an example, we consider imitation of speech in which the VOTs of voiceless stops are consistently shortened, one of the artificial accents used in the current work. First, the imitator must be able to *discriminate* the feature(s) characterizing the accent as distinct from canonical speech (i.e., be capable of perceiving the difference between the shortened vs. canonical VOT). They must also be able to *articulate* the variant (i.e., be capable of producing a voiceless stop with shortened VOT). The imitator must *identify* which acoustic information is relevant to their imitative goal; if the goal is to imitate an accent, this involves separating the linguistically-relevant feature(s) of the target accent (i.e., the shortened VOT) as distinct from non-linguistically-relevant properties such as those inherent to a speaker's voice (e.g., a speaker's

raw f0 values). The feature also must be recognized as a systematic property of the accent, rather than specific to a single episode (*generalization*), and the link between this general feature and the accent must be encoded in the perceptual representation. Finally, the imitator must *select* the feature in production when imitating the accent.

While this work focuses on explicit imitation of artificial accents, it is also relevant to consider the extent to which the above processes are shared by other types of imitation. Several direct comparisons have found that explicit instructions elicit more imitation than implicit methods. For example, Dufour & Nguyen (2013) showed that participants who were explicitly asked to imitate showed closer approximation to the formant values of a model talker than participants who were simply asked to repeat the words they heard (see also Clopper & Dossey, 2020; Pardo et al., 2010; Sato et al., 2013). To account for these differences, Dufour & Nguyen (2013) proposed that a general automatic alignment mechanism is a shared component of both explicit and implicit imitation, but that explicit instructions to imitate invoke additional "attentional" processes, directing perceptual attention to the specific indexical features of the model talker's speech. In other words, successful explicit imitation differs from implicit imitation in the extent of attention allocated to the input. We further propose that the explicit imitation may also invoke qualitatively different perception- and production-related processes that are not necessarily required for implicit imitation, as detailed in the following paragraph. We conceptualize explicit imitation as encompassing all processes involved in implicit imitation, but overlaid by additional, "controlled," perception and production processes.

A prerequisite to any type of imitation is the ability to perceive and produce the relevant variability (*discrimination* and *articulation*). Given these capabilities and a model of speech processing including a direct perception-production link (e.g., Goldinger, 1998; Pickering & Garrod, 2013), implicit phonetic convergence can, in principle, occur as an automatic consequence of the interaction between incoming speech and the listeners' representations (though there is strong evidence that mediating factors play a role, even in implicit convergence; see Coles-Harris 2017 for discussion).[1] On the other hand, explicit imitation also requires additional, controlled

[1] While the details of the cognitive architecture of phonetic convergence are far from established, there is good evidence that it involves automatic processes resulting from the interaction of the perception and production systems, but that it is also mediated by other factors (Coles-Harris, 2017). In this work, for simplicity, we use *automatic* to refer to the set of processes resulting in implicit phonetic convergence, in contrast to the *controlled* factors we lay out as necessary to explicit imitation. We do not mean to imply that implicit phonetic convergence results from an unmediated perception-production link, or that there is no level of control involved in implicit phonetic convergence. Rather, the contrast is that (successful) explicit imitation requires these controlled factors, while automatic phonetic convergence does not. Whether the differing performance in explicit vs. implicit imitation tasks found in previous work (e.g., Dufour & Nguyen, 2013; Clopper & Dossey, 2020) is best described as a quantitative difference in the influence of similar mediating factors, or a qualitative difference in which mediating factors play a role, is an area for future work.

processes, whose trajectory cannot be determined automatically based on the combination of input and representation, but which must be guided by other factors. Successful explicit imitation requires not only greater attention to the indexical properties of a talker's speech (Dufour & Nguyen, 2013), but also qualitative decisions about what to pay attention to, because the choice of which features are relevant (*identification*) will differ based on the imitative goal. For example, consider an adult speaker of General American English asked to imitate a child who produces systematically shortened VOTs of voiceless stops. The child's speech will differ in (at least) two ways from the adult's speech: shorter VOT, and an overall higher pitch due to smaller vocal folds. Which of these differences is imitated depends on the imitative goal: if they aim to imitate the *accent* of the speaker, the adult might modify their VOT, but if they aim to imitate the *age* of the speaker, they might instead modify the pitch of their voice. If their goal is to produce the word in a way that was indistinguishable from what they heard, they might modify both. Another decision that must be made by the imitator, in cases where the material to be imitated is not identical to the input, is the scope of generalization. For example, if the imitator in the example above has only heard voiceless stops starting with /p/ (but not /t/ or /k/) in the child's speech, they must decide whether the shortened VOT is indeed a systematic property of the speaker's "accent," and if so, whether this property generalizes to other segments (like /t/ or /k/), or whether it is specific to /p/. Finally, the *selection* process in production is also controlled by the imitator. For example, the imitator might be reluctant to produce a shortened VOT, even if they are able to articulate it and have recognized it as a property of the accent, perhaps because of social connotations associated with doing so, or perhaps because they are simply uncomfortable diverging from their usual speech norms.

In sum, we conceptualize successful explicit imitation as the combination of automatic processes of phonetic alignment, overlaid with additional controlled processes. Since there is overlap in the processes involved in implicit and explicit imitation (minimally, the ability to discriminate and articulate the relevant variability), findings from the larger body of research on implicit imitation may be relevant to explicit imitation as well, and we therefore consider them in formulating our research questions and hypotheses, and when contextualizing our results within the broader body of work. At the same time, the differences between the two types of imitation, outlined above and supported by the different empirical findings for implicit vs. explicit imitation in past work, highlight the importance of considering the different types of imitation as distinct entities.

1.2 Predictors of variability in imitation

Understanding the sources of variability underlying imitative performance is of both theoretical and practical interest, but it is also a challenge. Because multiple subcomponents are essential to successful imitation, a breakdown at any level will inhibit imitation, so it is often impossible to

determine the underlying reason for differences in imitation across groups or individuals simply by looking at the results of an imitation task and how these results correlate with behavioural or neuroanatomical measures. Some previous work has acknowledged this issue: for example, both Nielsen (2011) and Zellou & Brotherton (2021) bring up both perceptual and production-based mechanisms that could underlie the patterns of results found in their work on imitation. Furthermore, Reiterer et al. (2011) found that foreign speech pronunciation aptitude was associated with differences in neural activation in both speech-motor and auditory-perceptual areas.

Perceptual difficulties are often assumed to be the primary cause of lack of imitation (e.g., Olmstead et al., 2013), and theoretical accounts of phonetic convergence include critical roles for perception (e.g., Pickering & Garrod, 2013; Sancier & Fowler, 1997). However, most empirical work on native-language imitation does not directly test the role of perception. One notable exception, Kim and Clayards (2019), examined whether perceptual weighting of spectral and durational cues to the English /ɛ/ – /æ/ contrast was related to the imitation of these acoustic dimensions in an explicit imitation task. Results showed that individuals who relied more on spectral cues in perception showed greater imitation of durational, but not spectral differences, while individual reliance on durational cues was not predictive of imitation of either dimension. The authors concluded that general perceptual acuity, rather than attention to specific phonetic dimensions, modulates phonetic imitation. On a group level, Nielsen & Scarborough (2015) found evidence that linguistic selectivity in imitation has a perceptual basis: English listeners were better at discriminating artificially-lengthened than artificially-shortened VOT, providing a perception-based explanation for an earlier finding of asymmetrical imitation (Nielsen, 2011; discussed in detail below). Overall, however, there is no clear or straightforward evidence for the role of individual perceptual patterns in predicting native-language imitation.

If there has been little work directly testing the role of perception in imitation, there has been even less exploring the role of articulation. The one exception we are aware of is work by Reiterer et al. (2013), who, in the context of a neuroimaging study, also took a measure of spectro-temporal acoustic variability, which they termed "articulation space," in German speakers' productions of L1 and L2 speech. The participants also completed a separate foreign language imitation task, where they were asked to directly repeat sentences of Hindi, an unfamiliar language, with performance assessed perceptually by native speakers of Hindi. Participants who received higher ratings on the imitation task also had a larger articulation space, and the authors concluded that articulatory flexibility facilitates imitative ability.

Compared to the small number of studies exploring perceptual or articulatory predictors of imitation, there has been considerable interest in examining which characteristics of individual speakers, and their attitudes toward the interlocutor or the target speech, might predict imitation. A multitude of factors, including bilingualism (Spinu et al., 2020), musical experience (Coumel et al., 2019), and general cognitive processes (working memory: Reiterer et al., 2011;

neurocognitive flexibility: Reiterer et al., 2013) have been shown to correlate with individuals' extent of imitation. The list of social or personality-related factors proposed to condition imitation expands even further when considering the better-studied domain of phonetic convergence (see Wade et al., 2020 for a review). While intriguing, findings can be inconsistent and often fail to replicate across studies, making it difficult to draw firm conclusions about the individual characteristics governing imitation (Cohen Priva & Sanker, 2020; Wade, 2022).

One reason that the search for straightforward predictors of imitative ability may be elusive could be the multiple perception- and production-based processes underlying imitation laid out above, each of which is likely in and of itself affected by different factors. For example, the ability to detect small magnitudes of phonetic difference, the propensity to generalize, and the willingness to diverge from habitual speech norms, which might be expected to most strongly affect discrimination, generalization, and selection respectively, will not necessarily pattern together across individuals. These different patterns of variability could therefore obscure any relationship between overall imitation and each of the processes individually.

1.3 Imitation of VOT in English voiceless stops

Imitation of VOT in English voiceless stops has been well-studied, and previous work has consistently found imitation of lengthened VOT (e.g., Nielsen, 2011; Shockley et al., 2004; Wade et al., 2020). However, results on imitation of shortened VOT are mixed. Flege & Hammond (1982) found reduced VOTs in English speakers asked to spontaneously mimic their idea of Spanish-accented speech, indicating that it is possible for English speakers to imitate reduced VOTs. However, imitation of reduced VOT has not been shown to be elicited in any lab-based shadowing/exposure studies. Notably, in a direct test of lengthened vs. shortened VOT imitation, Nielsen (2011) showed that participants produced longer VOTs after exposure to voiceless-stop-initial words where VOTs had been lengthened by 40 ms, but no such effect was found after exposure to analogous stimuli with VOTs shortened.

One possible explanation for the asymmetry, proposed by Nielsen (2011), is that shortened VOT might be less perceptually salient than lengthened VOT. This was supported by a subsequent perception study by Nielsen & Scarborough (2015), who found that listeners were more accurate in discriminating similar lengthened VOT than shortened VOT stimuli from those with natural VOT values. This could also explain the discrepancy between the results of Nielsen (2011) and those of Flege & Hammond (1982), who did find shortened VOT imitation: the participants in Flege & Hammond (1982) may have been basing their imitations on their familiarity with (or stereotypes of) English spoken by L1 Spanish speakers, which may be characterized by unaspirated stops, consistent with the phonetic realization of phonologically voiceless stops in Spanish. These unaspirated stops would have lower VOT values than the

shortened VOT stimuli used in Nielsen (2011) (which were on average 30 ms for /p/-initial words, as compared to a range of 7–18 ms for /p/-initial Spanish words reported in Flege & Eefting 1987). They are also likely to be perceived by English listeners as phonologically voiced stops, further increasing their perceptual salience.

In our study, we tested imitation of voiceless stops with shortened, as well as lengthened, VOT using more extreme values than those used in Nielsen (2011): our shortened VOT values were within the normal range of voiced stops (i.e., shortened /t/ was 15 ms, which is characteristic of, and therefore likely to be perceived as, /d/). Therefore, if anything, we might expect *increased* salience relative to the lengthened VOT condition, which does not ever result in a change of phonological category. In addition, we used an explicit imitation paradigm, in contrast to the implicit pre/post-exposure paradigm in Nielsen's study, which we expected would direct more attention to phonetic characteristics of target stimuli (Dufour & Nguyen, 2013). If the lack of shortened VOT imitation was driven by limited perceptual salience, these differences in our stimuli and paradigm should result in imitation of shortened VOT.

To get a full view of the nature of imitation, it is important to consider distributional patterns as well as average values, as the shape of the distributions of imitated VOT values provides important information that can be obscured when only considering differences in mean values across conditions. The VOT distributions pre- and post-exposure to lengthened VOT in Nielsen (2011) were strikingly similar in shape, just shifted such that the post-exposure distribution was characterized by slightly higher values overall (mean 7 ms). This suggests that participants' imitations were best characterized as small but consistent increases in VOT. By contrast, when participants were asked to mimic Spanish-accented English, Flege & Hammond (1982) found that the distribution of VOT values was characterized by two clear clusters around 30 ms and 90 ms. This bimodal pattern suggested that there are distinct types of productions: some tokens were characterized by broad, categorical changes in VOT, while others remained consistent with a canonical English pronunciation. This distribution suggests that unaspirated stops that often characterize Spanish-accented English may have been perceived as a separate phonological category (voiced stops) in some participants and/or utterances, and were produced as such in imitation. Although Flege & Hammond (1982) and Nielsen (2011) differ substantially from one another in their methodologies, the contrast is consistent with a potential qualitative difference between imitation of shortened VOT, which can potentially be perceived as a different phonological category, and lengthened VOT, which cannot. Our study provides a direct comparison of imitation of lengthened vs. shortened VOT; comparison of the distributions across the two conditions allows us to determine if differences observed across previous studies might be ascribed to qualitative, rather than or in addition to quantitative, differences.

1.4 Artificial accents and talker variability

In order to test explicit imitation of different accents (i.e., systematic, linguistically-relevant phonetic characteristics) as well as perception of these differences, we required a paradigm where the target accents were presented in direct juxtaposition. To preview the procedure, which is described in detail in later sections, participants were asked to 1) listen to two talkers with different "accents" saying the same sentence, 2) to repeat after each talker, imitating the accent of each, and 3) to decide whether a third talker best matched the first or second talker they had heard. This direct juxtaposition of contrasting accents, along with the need for phonetically controlled stimuli that would allow us to isolate imitation of a single feature, led us to use artificial accents, and also led us to directly test the effect of voice-related variability on perception and imitation of systematic phonetic differences, as we describe in this subsection.

Different accents in the real world are generally produced by different talkers, so ideally, to make the task plausible, we wanted the two accents in a given trial to be heard as having two different voices. On the other hand, using two different talkers undermines the experimental control that is critical for isolating imitation of a single phonetic feature: speech naturally produced by two different talkers will always have many acoustic differences, particularly when using sentence-level stimuli, so participants might perceive and imitate these differences, instead of or in addition to our target feature of VOT.

To maintain full control over the acoustics of the material presented to participants, while still allowing for the voice-related variability that is present in naturally-occurring accents, we created artificial accents using sentence-level stimuli that differ minimally in specific phonetic features. Artificial accents can be constructed in a variety of ways: by using a speech synthesizer (e.g., Maye et al., 2008), by manipulating natural speech (Yu et al., 2013), or by recording natural productions produced with systematic phonetic/phonological differences (Adank & Janse, 2010; Spinu et al., 2020), and they have been shown to elicit phonetic imitation in past work. For example, using an explicit imitation task, Spinu et al. (2020) found that bilinguals showed more imitation than monolinguals of an artificial accent of English which varied from a canonical accent in four features (/ɛ/ -> /jɛ/, intervocalic /l/ -> /ɾ/, /ə/-epenthesis in s + stop clusters, and a novel intonation pattern in tag questions), and Yu et al. (2013) showed that participants produced longer VOTs after exposure to a narrative with systematically lengthened VOT of voiceless stops.

In order to directly compare imitation of multiple artificial accents, we manipulated the VOT of voiceless stops from a single set of baseline recordings to create three "accents" varying only in VOT: one set of stimuli had canonical VOT, approximating the talker's natural production values, while the other two sets were manipulated to have either systematically shortened or lengthened VOT. Apart from the VOT, the three "accents" were acoustically identical. We then created multiple "voices" for each accent by scaling the f0 and formants. Crucially, because only overall

f0 and formant scaling were modified, there were no linguistically-relevant differences between the different voices. In sum, we created a bank of stimuli representing three different accents, with multiple voices per accent, while maintaining complete control over all other acoustic characteristics of the sentences.

The inclusion of multiple voices, in contrast to previous work looking at imitation of artificial accents, brought up the question of how the presence of voice-related variability might affect attention to – and therefore imitation and discrimination of – the target VOT differences. In order to determine what effect this might have, we compared performance in a condition characterized by multiple voices, as described in the previous paragraph, to another condition which included no voice-related variability at all: participants imitated and discriminated sentences that differed *only* in VOT: in other words, with different artificial accents (e.g., shortened vs. canonical VOT), but with the same "voice."

Findings from several strands of work make different predictions about how performance might differ across conditions where voice-related variability is present vs. absent. First, there is evidence that acoustic variability introduced by the inclusion of multiple talkers invokes a general phonetic processing cost, perhaps attributable to the additional processing resources needed to separate talker-specific from phonetically-relevant information (e.g., Mullennix & Pisoni, 1990; Mullennix et al., 1989). Support for this idea comes from findings that talker variability has been shown to hinder performance on some perceptual learning tasks, particularly those not requiring generalization (see Baese-Berk, 2018, for discussion and examples). Under this view, which assumes that phonetic detail is retained during processing (Goldinger, 1998), we would expect that decreased sensitivity or attention to the target VOT differences in discrimination and imitation would lead to worse performance with multiple voices, as compared to the condition with no voice-related variability.

In the studies discussed above that found decreased performance in the presence of multiple talkers, the stimuli were indeed produced by different human talkers, and therefore exhibited natural talker-related variability. In our study, on the other hand, the different voices were fully controlled except for scaling of f0 and formants, characteristics not relevant to phonetic perception in English. Based on findings that natural talker variability hindered performance on spoken word identification, but minimal differences in amplitude or overall f0 did not (Bradlow et al., 1999; Sommers et al., 1994; Sommers and Barcroft, 2006), Sommers and Barcroft (2006) proposed the Phonetic Relevance Hypothesis: variability that alters acoustic properties that are important for phonetic identification imposes processing costs on perception, but non-phonetically-relevant variability does not. Under this view, the presence of our highly controlled different "voices", specifically constructed not to vary in any phonetically-relevant dimension, would not be expected to affect imitation or perception of VOT differences. In this case, we would expect to see identical performance in the conditions with and without voice-related variability.

Finally, there is reason to expect that voice-related variability might actually improve performance on tasks where generalization is required, as is the case in some portions of the current study. Talker variability has been shown to facilitate generalization in both dialect classification (Clopper & Pisoni, 2004) and accent adaptation (Bradlow & Bent, 2008; Schmale et al., 2012). These studies are not directly analogous to the tasks in the current work, since these previous studies compare performance when there are multiple talkers per accent vs. a single talker per accent, whereas the current work compares multiple talkers per accent to a complete lack of talker variability (i.e., the same voice is used in both accents). However, the fundamental prediction from these previous studies still applies: variability inherent in a multi-talker context may provide information about which dimensions are linguistically relevant (and therefore helpful in informing perceptual judgments in different contexts), and which can be attributed to idiosyncratic or indexical properties of a single talker (and therefore only relevant to that talker). We included a combination of exposed and novel sentences in the discrimination phase of our experiment, allowing us to test whether the effect of voice-related variability would facilitate generalization.

1.5 Current study

This work examines explicit imitation of artificial accents varying in a single feature (VOT), using a preregistered design and analysis.[2] Our paradigm consists of four tasks: artificial accent imitation, artificial accent discrimination, word-level VOT imitation, and word-level VOT discrimination. All tasks involved imitation and discrimination of VOT differences; however, the accent tasks consisted of imitation of sentences, as opposed to individual words. The sentence-level accent tasks differed from the word-level tasks in that they were designed to invoke the goal of imitating an "accent" rather than idiosyncratic properties of a specific token of a word. We did this by including variability in sentences and talkers (such that the tasks could not be done without some level of generalization), and by explicitly invoking the concept of "accent" in instructions to the participants.

The two accent-level tasks, artificial accent imitation and discrimination, were combined into a single task presented to participants. In each of a series of *trial sets*, participants heard and imitated a pair of sentences with different artificial accents, then completed an ABX task, deciding whether a third sentence matched the first or second accent they had heard. Stimuli for the accent-level tasks were manipulated in three fully crossed conditions (**Table 1**). First, the AccentType manipulation varied in the nature of the artificial accent: one condition tested imitation and discrimination of lengthened vs. canonical VOT, while the other tested shortened vs. canonical VOT. Second, to test the role of voice-related variability, the TalkerMatch manipulation varied

[2] Stimuli, results, analysis code, preregistration information, and a document detailing changes made to the preregistered analysis during the review process are available on OSF at the following link: https://osf.io/zve4c/.

in whether the sentences within each trial were spoken with the same voice or different voices. Third, the ABX discrimination task included two types of X sentences: half were the same as the A and B sentences in terms of linguistic content, while the other half were different sentences (SentenceMatch). This factor was included to test generalization to new words and segments, and to reinforce the instructions that the task should be done by considering differences as a general feature of an "accent" (i.e., a systematic property of the way a person talks), as opposed to an idiosyncratic property of individual words or utterances.

Table 1: Conditions used for accent imitation and discrimination tasks. These conditions were fully crossed, with each participant hearing all 8 possible combinations (2 AccentType * 2 TalkerMatch * 2 SentenceMatch).

Variable	Levels	Description
AccentType	Shortened VOT	Canonical vs. shortened VOT
	Lengthened VOT	Canonical vs. lengthened VOT
TalkerMatch	SameVoice	Talkers in a trial have the same "voice" (f0/formants)
	DifferentVoice	Talkers in a trial differ in "voice" (f0/formants)
SentenceMatch (discrimination task only)	SameSentence	New sentence is the same as exposure sentences
	DifferentSentence	New sentence differs from exposure sentences

The other two tasks tested word-level VOT imitation and discrimination. Stimuli for these tasks were tokens of a single word, spoken in a single voice, manipulated to differ only in VOT. In contrast to the accent-level tasks, word-level imitation and discrimination were presented as two separate tasks: for imitation, participants heard and were asked to listen to and imitate tokens drawn from a VOT continuum, and for discrimination, participants completed an ABX task with tokens from the same continuum.

The four tasks included in the study were chosen to tap into different sub-processes involved in explicit imitation of artificial accents. **Table 2** shows the mapping between the tasks and the five processes laid out above in Section 1.1. The ability to discriminate VOT differences is necessary for successful completion of all tasks. This is in fact the only prerequisite to completion of the word-level ABX discrimination task, which can be done by shallow acoustic matching of tokens of a single word. The accent discrimination task, on the other hand, additionally requires identification of VOT as a relevant feature, while at the same time identifying other variability, such as f0 of the talker's voice, as *not* a relevant feature, when assessing similarity. It also requires generalization of the VOT difference to different segments and words in order to accurately discriminate a sentence with different linguistic content (in the DifferentSentence condition).

Table 2: Descriptions of the sub-processes of imitation given in Section 1.1 (first column), and their assumed mapping to the four tasks in the current work, with 'x' indicating that a given task requires a given process, and '(x)' indicating that the process is not strictly required for imitation to occur, but is likely to be invoked in the task.

	Word-level discrimination	**Accent discrimination**	**Word-level imitation**	**Accent imitation**
	Discrimination of words differing in VOT	*Discrimination of sentence-level artificial accents differing in VOT*	*Imitation of words differing in VOT*	*Imitation of sentence-level artificial accents differing in VOT*
Discrimination: Ability to perceive Feature X	x	x	x	x
Identification of Feature X as relevant to the contrast	-	x	(x)	(x)
Generalization of Feature X to broader domain	-	x	-	-
Articulation: Ability to produce Feature X	-	-	x	x
Selection: Choice to articulate Feature X	-	-	x	x

Turning to imitation, both tasks require the ability to discriminate the differences, and the ability and decision to produce them (articulation and selection). The question of whether identification is involved in the imitation tasks is less straightforward. To the extent that fully automatic alignment processes operate in all imitation tasks, VOT imitation would be expected to occur even without any active identification of VOT as a relevant dimension. However, following the view of Dufour and Nguyen (2013), we expect that explicit instructions to imitate direct additional attention to the properties of the target stimuli and encourage identification of the dimension that is relevant to the target contrast. Furthermore, although the relevant dimension (VOT) is the same in both the word-level and accent imitation tasks, it is important to note that the two tasks differ in imitative goal (imitate differences in the word vs. differences in two accents), in the nature of the stimuli (sentence- vs word-level), and in the fact that the accent task includes additional voice-related variability (in one of the two TalkerMatch conditions); any of these differences might affect the ease of identification of this dimension. Neither imitation task tests generalization to new material.

The aim of the current study is to test the role of several factors expected to influence explicit imitation of artificial accents varying in VOT. First, we explore the extent to which individual performance in artificial accent imitation is correlated with three tasks (accent discrimination, word-level imitation, and word-level discrimination) designed to target different sub-processes of imitation. For those tasks shown to be significant predictors, we also test to what extent they

themselves are related. These comparisons provide information about which factors might be the most important predictors of variability in artificial accent imitation. For example, if overall perceptual acuity is the primary predictor of imitative ability, we would expect variability in the accent imitation task to be correlated with all three other tasks (and that all tasks would in fact be correlated), since all require discrimination of VOT differences. On the other hand, if production-based factors (articulation and/or selection) are primary, we would expect to find that the word-level imitation task is predictive of the accent imitation task, and that this is independent of performance on the other tasks. If the ability to recognize and/or generalize VOT as a property of an artificial accent, even in the face of talker variability, is primary, then we would expect to find that the accent discrimination task is predictive, independent of other tasks. Second, we test the effect of voice-related variability on imitation and discrimination of systematic VOT differences by comparing performance in the SameVoice vs. DifferentVoice conditions of the accent-level tasks. We expect it to be more difficult to perceive – and therefore imitate – differences in VOT in the DifferentVoice condition, due to the additional acoustic variability introduced by the different voices. However, since the acoustic variability was controlled to only differ in non-phonetically-relevant characteristics (overall f0 and formant scaling), this voice-related variability may be automatically filtered out (Sommers & Barcroft, 2006), in which case we would expect no difference in the two conditions. It is also possible that the variability of voices in the DifferentVoice condition facilitates generalization in classifying novel items, and we therefore expect that any inhibitory effect of the DifferentVoice condition on the perception task might be smaller – or reversed – when test items are novel sentences. Finally, we compare explicit imitation of shortened vs. lengthened VOT, using more extreme values than used in previous lab-based studies. We also compare distributional patterns across imitations of the two VOT manipulations.

2. Method

2.1 Participants

Forty-two native speakers of English completed both sessions of the experiment (22 female/20 male, mean age 32, age range 19—64, based on self-reports). All learned English as a first language (10 had exposure to other languages in the home as well: 1 Cantonese, 1 French, 2 Mandarin, 1 Portuguese, 3 Spanish, 2 Tagalog).[3] Our recruitment target, based on an a priori power

[3] As suggested in previous work (Spinu et al., 2018), bilinguals may have better imitative ability, and we might in particular expect that participants with short-lag voiceless stops in their native inventory (e.g., Spanish, Portuguese, Tagalog) might be better at imitating shortened VOTs than English monolingual speakers. We conducted exploratory analyses to see whether the 6 participants who fit this profile showed different patterns in either perception or production: specifically, we coded participants as *multilingual* (i.e., these 6 participants) or *monolingual* and tested whether the effect of this factor was significant in any of the statistical models. No significant effects were found; however, given the small number in the group, we do not interpret this as a strong indication of lack of differences, and this is an important question to explore in future work.

analysis,[4] was 40 participants, but given the online experiment setting and the fact that participants were recorded using their own equipment, we anticipated having to recruit substantially more than we would be able to use, and we ended up with 42 usable participants. We recruited participants through Prolific, an online platform for behavioral experiments, with the initial session open to users who had reported all of the following in their registered profiles: 1) born and currently residing in the United States or Canada; 2) learned English as a first language; and 3) no language- or hearing-related difficulties. We then invited participants back for the second session if 1) their productions and their responses to our language background survey indicated that they met the native English language requirement; 2) their recordings were of sufficiently good quality for analysis; and 3) they performed at above 75% on the catch trials, which were included to ensure people were paying attention to the task and completing it as instructed. Sixty-nine additional participants completed Session 1 but were not invited back because of bad or overly noisy recording quality (n = 28), failure to meet the native English language requirement (n = 23), recording completely missing (n = 12), technical issues with the experiment platform (n = 4), or failure to meet the threshold of accuracy on catch trials (n = 2). An additional 14 participants completed Session 1 and were invited to Session 2 but did not participate (n = 6) or had low-quality or missing recordings in Session 2 (n = 8), so their data is not included here.

2.2 Artificial accent tasks: Materials

Stimuli for the accent tasks (imitation and discrimination) consisted of eight English sentences, recorded by a female native speaker of English. Each sentence began with a list containing two target words that began with voiceless stops and that had no real-word voiced minimal pairs (e.g., ***Parrots****, ferrets,* ***cats*** *and fish live with me in my home*). The list format was expected to elicit a pause between words, minimizing coarticulation with the preceding word. There were no other onset-initial stops in the sentences, and target stops were roughly balanced for place of articulation (5 /p/, 5 /t/, 6 /k/). Four of the sentences, *exposure sentences*, were used in all phases of the experiment (exposure, imitation, and discrimination; discussed below), while the other four sentences, *generalization sentences*, were only used to test generalization in the discrimination phase. The full list of sentences is included in Appendix A.

[4] For our power analysis, we performed simulations of the accent discrimination experiment using effect sizes and standard deviations based on pilot work using a similar task. Since the power analysis and sample size is based on the accent discrimination task, if null effects are found in other tasks, we will need to take into account the possibility of lower power when interpreting those results. To arrive at our planned sample size, we performed 1000-iteration simulations of our planned statistical analysis for the accent discrimination task for increasing n in multiples of 4 (because we had 4 conditions) until we surpassed 80% power in the simulation to capture all relevant effects (Intercept, i.e., accuracy above chance, TalkerMatch, and TalkerMatch*AccentType).

The VOT of each target stop was annotated, beginning just before the stop burst and ending at the beginning of periodicity in the following vowel, then was manipulated using the PSOLA algorithm in Praat (Moulines & Charpentier, 1990) to create three versions differing in VOT duration: shortened, canonical, or lengthened.[5] Values for manipulation were chosen based on global consideration of several factors. First, canonical values needed to roughly match production values in the natural recordings (94, 110, and 99 ms for /p t k/ respectively), and the shortened and lengthened versions needed to differ equally from the canonical version. We also wanted the shortened set to be contrast-threatening, with VOT values that would be plausible for voiced stops. The final versions of the shortened versions varied in how "voiced" they sounded, with some sounding voiceless, some voiced, and some ambiguous; the implications of this will be discussed in Section 3. Finally, the VOT values differed by place of articulation, reflecting natural production differences (Cho & Ladefoged, 1999). The final values for shortened/canonical/lengthened, respectively, were as follows: /p/: 5/80/155 ms; /t/: 15/95/175 ms; /k/: 25/110/195 ms.

From each of these sentences, we created versions with 14 different "voices" by scaling the f0 and formants in multiple steps (using the Praat *Change Gender* function), with formant ratios ranging from 0.775 for the "lowest" voice to 1.1 for the "highest" (1.0 representing the natural production values), and f0 median ranging from 130 to 260 Hz. Endpoints were chosen based on pre-testing in order to maximize the range of voices while still maintaining relative naturalness. This resulted in a continuum ranging from a low-pitched voice with spectral characteristics of a large vocal tract, to a high-pitched voice with characteristics of a small vocal tract. In total, 42 versions (14 voices * 3 VOT) of each of the 8 sentences were created for use in the accent imitation/discrimination tasks.

An additional set of sentences was used for practice trials and catch trials in the accent tasks. These sentences, recorded by a different female speaker, included two words with coda rhotics (e.g., *The bike was much slower than the car*). Two versions of each sentence were recorded, one with the target words including the rhotics, and the other without, as would be produced in a non-rhotic dialect. This contrast was chosen because it was expected to be relatively salient and familiar to participants. Rhotic and non-rhotic versions of the target word were spliced into the same frame sentence to create two versions which only varied in the target word. Multiple versions with different "voices" were created as described above.

2.3 Artificial accent tasks: Procedure

Artificial accent imitation and discrimination were assessed in a single task consisting of 16 *trial sets*. Each trial set tested imitation and discrimination of a sentence spoken by two *talkers* who were said to have two different accents. To minimize participant fatigue, the task was split into

[5] This was a different manipulation technique than the "copy-and-splice" method used in Nielsen (2011). While we think it unlikely, this methodological discrepancy could contribute to differences in results across the two studies.

two sessions of 8 trial sets each, completed on two separate days. The average time between sessions was 7 days (minimum 4, maximum 12).

Each trial set consisted of three phases: *exposure*, *imitation*, and *discrimination*. **Figure 1** shows a visualization of the procedure for a single trial set, including the exact instructions presented to participants. First, in the *exposure* phase, participants heard an utterance spoken first in one accent, followed by the same utterance spoken in another accent. This phase was repeated two times. The two accents were represented visually as silhouettes of individual talkers (Talker F and Talker J in **Figure 1**), with each relevant silhouette appearing during playback. Second, in the *imitation* phase, the two utterances would be played again, with a pause after each one and a prompt for participants to imitate, and this phase was also repeated two times. Finally, in the *discrimination* phase, participants would complete six trials in which they would hear the two utterances once again, followed by a third utterance (represented by a third silhouette) and be asked to classify it as one of the two accents. Participants were told whether their response was correct after each trial, and they were shown their current mean accuracy across the entire experiment.

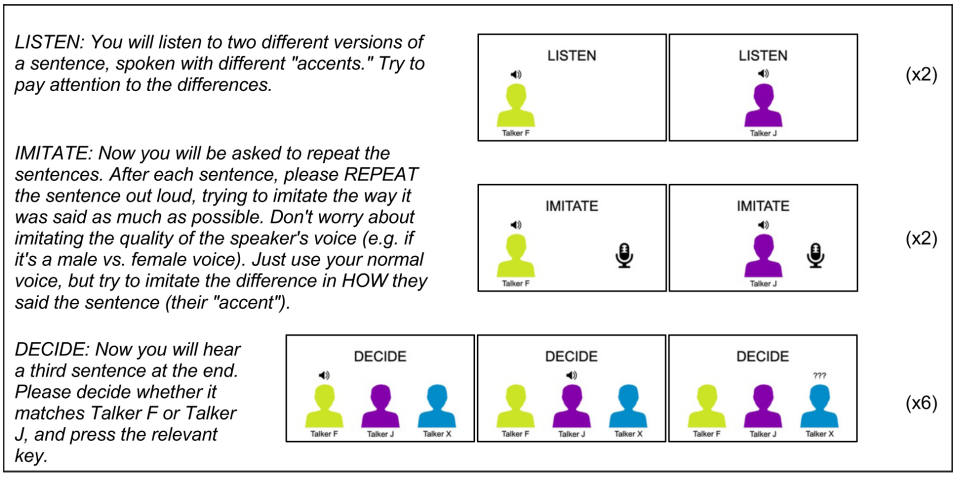

Figure 1: Summary of the procedure for a single trial set in the accent imitation/discrimination task, including the exact instructions given to the participants for each phase (exposure, imitation, discrimination). The LISTEN and IMITATE phases were each repeated twice, and there were six discrimination questions.

2.4 Artificial accent tasks: Design

The stimuli in each trial set differed as a function of three factors, summarized in **Table 1**. AccentType (Lengthened vs. Shortened VOT) refers to whether the canonical-VOT stimulus (which was spoken by Talker F in all trials)[6] was compared to artificially lengthened or shortened

[6] We chose to keep the order constant because our design was already fairly complex, and we didn't have specific predictions about how this would affect performance. We speculate that if anything, additional variability due to different order would have made the task overall a bit more difficult (in all conditions).

VOT (Talker J). TalkerMatch (SameVoice vs. DifferentVoice) refers to whether the utterances in each trial were presented in the "same voice" (i.e., identical f0/formant parameters) or "different voices," in order to test how variability in voices affects performance. Finally, SentenceMatch (SameSentence vs. DifferentSentence) tests how well participants could generalize their discrimination to new lexical content. In the SameSentence discrimination trials, the third utterance to be classified was the same as the exposure sentences; in the DifferentSentence trials, the third utterance had different lexical content. The two sentences used in the exposure and imitation phases were always identical.

All participants participated in all combinations of conditions, across two separate sessions (approximately 30 minutes each). Each session included 8 trial sets from a single TalkerMatch condition: half of the participants heard the SameVoice condition in Session 1 and the DifferentVoice condition in Session 2, and the other half of the participants heard the opposite. Both AccentTypes were present in each session, blocked so that half of the participants heard four lengthened-VOT trials followed by four shortened-VOT trials, and the other half heard the shortened-VOT trials first. Between the two blocks was a set of catch trials. The full experimental sequence for a sample participant is given in Appendix B.

Recall that 14 different "voices," varying in f0 and formants, were created for each sentence. In the SameVoice session, all stimuli within a block were the same voice (however, a different voice was used for the first vs. second half of the session, corresponding to the two AccentType conditions; see Appendix B). In the DifferentVoice session, the two exposure/imitation sentences in each trial were presented in perceptibly different voices, and the new sentence to be discriminated was different than either of the two exposure voices. In other words, in the DifferentVoice condition, three distinct voices were present in each trial set, and different sets of voices were heard across different trial sets, whereas in the SameVoice condition, a single voice was heard throughout an entire block. Information about acoustics and distribution of the voices is given in Appendix C.

Unlike the other two factors, the SentenceMatch factor was not blocked; instead, in each trial set, there were 6 discrimination trials: three SameSentence trials followed by three DifferentSentence trials, with the order of trials within each SentenceMatch type randomized by participant. The voice of the sentence to be discriminated was always the same as the two exposure voices in the SameVoice condition, and different than either of the two exposure voices in the DifferentVoice condition. The accent (canonical vs. noncanonical) of Talker X was balanced equally across trial sets and conditions, such that half of the discrimination trials matched Talker F and half matched Talker J.

To summarize, across the two accent task sessions, each participant imitated 128 stops, excluding practice/catch trials (4 sentences * 2 stops/sentence * 2 accents (canonical vs. modified) * 2 repetitions * 2 AccentType * 2 TalkerMatch), and responded to 96 discrimination trials (4 sentences * (3 SameSentence + 3 DifferentSentence trials) * 2 AccentType * 2 TalkerMatch).

2.5 Word-level tasks: Materials

Stimuli for the word-level tasks were created from a single natural production of the English word “toast” from a female native speaker of English (different than the speaker in the accent tasks). A five-step series was created with VOT varying from 15 to 175 ms in increments of 40 ms, using the same method for VOT manipulation as for the accent-level task materials, with endpoints chosen to match the shortened and lengthened values for /t/ in the accent tasks. No voice manipulations were done.

2.6 Word-level tasks: Procedure and design

In the second experimental session, after completing the accent tasks, participants completed the word-level ABX discrimination task on productions of the word *toast* drawn from the VOT series described above (instructions: “You will do a short version of the task you did earlier, but with single words. In each trial, you will hear three words. Please decide if the third word matches the first or second, pressing ‘f’ or ‘j’ to indicate your choice.”). Listeners completed 28 critical trials, consisting of all possible combinations of one- or two-step (40 or 80 ms) differences (7 trials), repeated four times. An additional 3 trials with a four-step difference (160 ms) were included in the trials to ensure that there would be trials that were obviously different; these trials are not included in the analysis. Order of trials was pseudo-randomized and held constant for all participants. This task took less than five minutes.

Finally, participants completed the word-level imitation task, in which they heard and imitated three repetitions of the stimulus set, for a total of 15 productions (instructions: “You will hear the word ‘toast’ multiple times. After hearing each word, please repeat it out loud, trying to imitate exactly how it was said.”). The order of the trials was pseudo-randomized and held constant for all participants. There were no practice trials for this task, but since we expected that the first production might be anomalous, we included an extra trial at the beginning, and did not include this in the analysis. This task took approximately five minutes.

2.7 Production data: Acoustic analysis

The following acoustic landmarks were manually annotated for the initial voiceless stops in all target words in both the accent and word-level imitation tasks: 1) beginning of the stop burst, as visible in the waveform; 2) onset of periodicity of the following vowel, as visible in the waveform; 3) end of the following vowel, as indicated by the end of stable formants for F2 and above (given the difficulty in isolating the following vowel in the words *parrot* and *poem*, the “vowel” interval for these words included the diphthong for *poem* and half of the /VrV/ sequence in *parrot*. VOT was calculated as the duration from (1) to (2), and vowel duration from (2) to (3). In some cases, the target sound and/or entire word was missing due to technical or recording errors ($n = 85$), or the signal was too noisy to annotate VOT ($n = 16$). We also excluded tokens from analysis if

the participant produced a different place or manner of articulation than the target sound (69 /t/ > /p/; 7 /k/ > /p/; 11 other), or the consonant was omitted (n = 8, all /p/-initial words). In total, 5219 (out of 5376) tokens were included in the analysis for the accent task and 591 (out of 630) for the analysis of the word-level task. One participant was omitted from analysis of the word-level production task (and comparisons between this task and others) because they had no data corresponding to two of the five continuum steps.

2.8 Statistical analysis

All statistical analyses were performed in R (R Core Team, 2020). For analysis of group results for all tasks, we used mixed-effects logistic (for perception) and linear (for production) regression models, using the package lme4 (Bates et al., 2015). P-values for the linear models were computed using the lmerTest package (Kuznetsova et al., 2017), and an alpha-level of 0.05 was used as the threshold for significance. In the case of significant interactions, we performed follow-up tests using the *phia* package (De Rosario-Martinez, 2015) to test whether the effect of interest held at each level of the other factor(s). Along with our primary predictor variables, we also performed exploratory analyses to examine whether participants' age affected performance, given the relatively large age range of our participants and previous findings of changes in speech perception across the lifespan (e.g., Incera & McLennan, 2018; McLennan, 2006). To do this, for each task, we ran an additional model including the continuous variable of age (centered), and performed likelihood ratio tests, using the anova() function in R, comparing models with and without age, to determine whether inclusion of age significantly improved the model. Details of the model structure vary by task and will be introduced in the relevant subsection of the results. To give a concrete idea of effect sizes, we also provide percentages in terms of mean accuracy rates (for discrimination) and VOT values (for imitation) across levels of each condition.

3. Results

3.1 Accent tasks: Catch trials

Practice/catch trials (presented at the beginning of the experiment and between blocks of the accent tasks) were included to ensure that participants understood the task and were imitating the accent, as opposed to the voice of the talker. These trials consisted of a sentence containing two coda rhotics, with one accent spoken with the rhotics present and the others with the rhotics absent. Participants completed 4 imitation and 16 discrimination trials (11 of the 168 total imitation trials were missing due to recording errors). Imitation trials were coded for presence/absence of a coda rhotic for each target word, based on the perception of a phonetically trained research assistant. We report the accuracy (for discrimination) and percentage of target imitations in which the presence/absence of rhotic matched the stimulus, both in total across the full experiment, and broken down by Session/TalkerMatch condition.

Participants had very high discrimination accuracy for these catch trials across all Session/TalkerMatch conditions (Session 1 SameVoice: 98%; Session 1 DifferentVoice: 98%; Session 2 SameVoice: 100%; Session 2 DifferentVoice: 98%). Imitation of target sounds in the catch trials was also highly accurate. Target words with rhotics present (i.e., the canonical North American English pronunciation) were imitated with rhotics 97% of the time (Session 1 SameVoice: 97%; Session 1 DifferentVoice: 97%; Session 2 SameVoice: 95%; Session 2 DifferentVoice: 98%). Target words with rhotics absent (i.e., a noncanonical pronunciation for North American English speakers) were imitated without rhotics 85% of the time (Session 1 SameVoice: 79%; Session 1 DifferentVoice: 79%; Session 2 SameVoice: 92%; Session 2 DifferentVoice: 93%).

3.2 Artificial accent imitation results

VOT values for imitations of the target words in the artificial accent imitation task, as well as the VOT values of the stimuli themselves, are shown in **Figures 2** and **3**, and statistical results are shown in **Table 3**. For purposes of analysis, we calculated a by-trial measure of *VOT-imitation* that captures the difference in VOT for each target segment across the two accents in each imitation trial, transformed such that positive values indicate a change in the expected direction of imitation. For trials in the lengthened VOT condition, this was [*VOT during imitation of lengthened accent*] – [*VOT during imitation of canonical accent*], and vice versa for the shortened VOT condition. We used a mixed-effects linear regression model to test whether there was imitation, as well as whether the extent of VOT-imitation differed based on AccentType (*lengthened* vs. shortened VOT), TalkerMatch (*same* or different voices), and Session (*1* vs. 2). The three predictor variables (reference levels in italics above) were included, as well as all interactions, mirroring the structure of the perception model (excluding the factor SentenceMatch, which was not relevant for the production data because the imitated sentences in a given trial were always the same). Random intercepts were included for Participant and Sentence ID, as well as random by-participants slopes for TalkerMatch and AccentType, and by-item slopes for TalkerMatch.[7] All factors were centered (–0.5, 0.5). In interpreting the model, the estimate of the intercept represents the extent of VOT imitation (in ms) across all conditions, and the estimate corresponding to each fixed factor represents the difference in VOT imitation between the two levels of the factor. Likelihood ratio tests showed that including age as an additional predictor did not significantly improve model fit from the model without age, either when it was included as both a main effect and in interaction with Session and TalkerMatch ($\chi^2 = 10.26$, $p = 0.248$), or when it was included as only a main effect with no interactions ($\chi^2 = 0.23$, $p = 0.635$).

[7] The statistical model reported here differed from that of the preregistered analysis in two ways: the addition of Session, and the addition of by-Sentence random effects, both of which resulted in a more conservative model than the planned model. Full details are available in the supplementary materials.

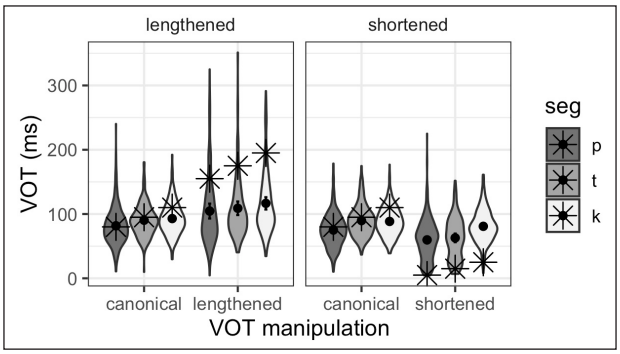

Figure 2: Raw VOT values produced in imitation of sentences with canonical vs. lengthened VOT (left panel) or canonical vs. shortened VOT (right panel), broken down by place of articulation. Large asterisks indicate the VOT values in the stimuli. Error bars show 95% confidence intervals of by-participant means.

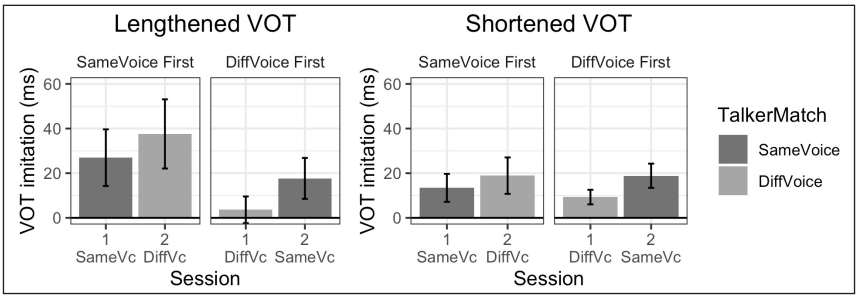

Figure 3: Accent imitation task: VOT imitation (difference in VOT between two accents on a given trial, with positive values indicating a difference in the expected direction) for lengthened (left) and shortened (right) VOT conditions. Participants who heard the SameVoice condition in Session 1 and DifferentVoice in Session 2 are shown in the left panel of each graph, and the other half of the participants are shown in the right panel. Error bars show 95% confidence intervals of by-participant means.

Table 3: Statistical results from a linear mixed-effects model predicting VOT imitation (in ms) in the accent imitation task. Reference levels are in italics, and significant effects are shaded.

Factor	**Estimate**	**Std. Error**	***t* value**	***p* value**
(Intercept)	18.469	2.932	6.299	< 0.001
Session (*1* vs. 2)	10.141	1.865	5.438	< 0.001
TalkerMatch (*same* vs. different)	–1.881	1.865	–1.009	0.319
AccentType (*lengthened* vs. short)	–5.905	5.375	–1.099	0.285
Session * TalkerMatch	23.583	10.071	2.342	0.024
Session * AccentType	–4.240	2.667	–1.584	0.113
TalkerMatch * AccentType	–0.481	2.676	–0.180	0.857
Session * TalkerMatch * AccentType	–39.172	15.191	–2.579	0.014

The significant (positive) intercept indicates that participants imitated the VOT differences between the two accents, showing on average an 18 ms VOT difference, in the expected direction, between the two productions imitating the two different accents in a given trial. Main effects indicated more imitation in Session 2 than in Session 1, but no overall difference between the two TalkerMatch conditions or between the two AccentType conditions. However, there was a two-way interaction between Session and TalkerMatch, with follow-up tests indicating that there was a significant effect of TalkerMatch during Session 1, with more imitation in the SameVoice than the DifferentVoice condition, and no effect during Session 2 (rather, there was a trend in the opposite direction, with DifferentVoice being numerically higher than SameVoice) (Session 1: $\chi^2 = 6.48$, $p = 0.011$; *Session 2:* $\chi^2 = 3.41$, $p = 0.065$). There was also a significant three-way interaction of Session, TalkerMatch, and AccentType: follow-up tests of simple effects of AccentType at each combination of Session and TalkerMatch indicate that there was an effect of AccentType, with greater imitation for lengthened than shortened VOT in the Session 1 SameVoice condition ($\chi^2 = 3.90$, $p = 0.048$), and the Session2 DifferentVoice condition ($\chi^2 = 7.19$, $p = 0.007$), but no significant effects in the other two Session * TalkerMatch conditions (both $p > 0.1$). Finally, we performed follow-ups to test whether participants showed significant imitation in all Session * TalkerMatch conditions. The imitation effect was not significant in the least imitative, Session1-DifferentVoice condition, but was significant in the other three conditions (Session1-DifferentVoice: $\chi^2 = 2.49$, $p = 0.114$, all others $p < .001$).

We also examined the shape of the distributions of imitated VOTs. Density plots of the distribution of the full set of participants' VOT values are shown in **Figure 4**, broken down by Session and TalkerMatch condition. Overall, distributions of shortened VOT are shifted lower, and lengthened VOT are shifted higher, than the canonical distributions, although this is much less apparent in the Session 1 DifferentVoice condition than the other three Session/Voice combinations, reiterating the result shown above that imitation was inhibited in the DifferentVoice condition for those participants who completed it during the first session. However, observation of the distributions of the conditions where imitation was found (i.e., all except for the Session 1 DifferentVoice condition) reveals differences that are masked by the group means shown above. Specifically, the distributional patterns are different in the shortened condition, where there is clear evidence of bimodality, than in the lengthened condition, where there is not. This is reminiscent of differences found in previous work: recall that the distributions of post-exposure VOT lengthening in Nielsen (2011) indicated a small-but-consistent modification (i.e., a shifted distribution of the same shape), whereas distributions of shortened VOT in spontaneous mimicry of Spanish-accented English in Flege and Hammond (1982) indicated more categorical behavior (i.e., a bimodal distribution).

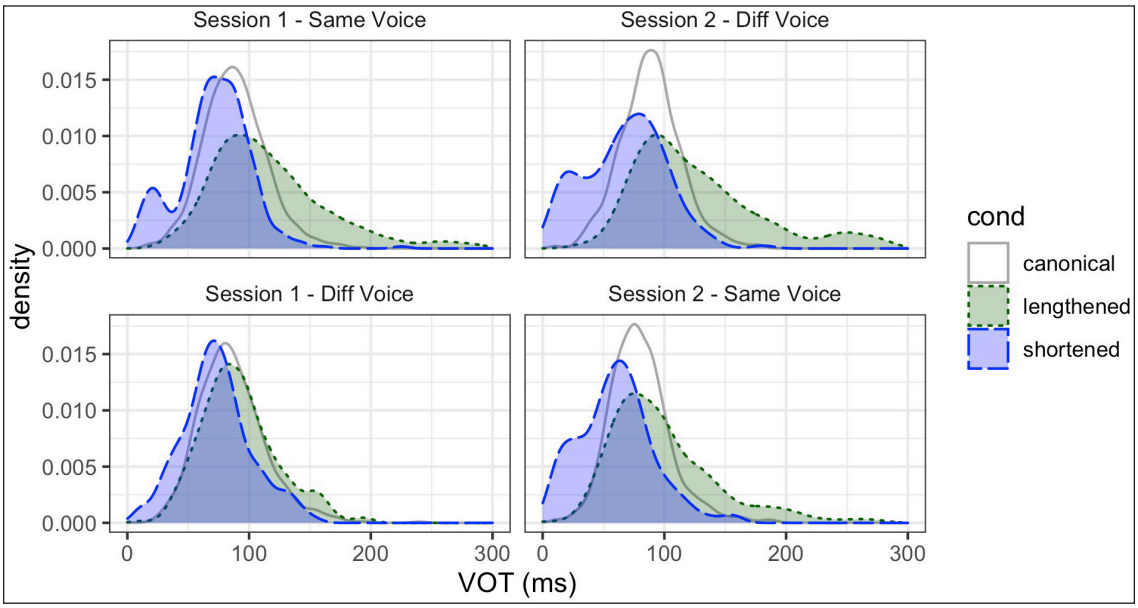

Figure 4: Density plots showing the distribution of raw VOTs in imitations of the canonical (grey) and manipulated accents (lengthened: green fill and dotted line; shortened: blue fill and dashed line), broken down by Session and TalkerMatch condition.

In order to look at the distributional patterns of imitation in more detail, we examine the results broken down by item, shown in **Figure 5**. For these item-based plots, we only include data from Session 2: we exclude Session 1 data because we are interested here in examining distributional patterns when there *is* imitation, and imitation appeared to be categorically inhibited in the Session 1 DifferentVoice condition (there *was* imitation in Session 1 for half of the participants, those who heard the SameVoice condition first; however, including this would mean that this half of the participants would be over-represented in the distributional analysis). In **Figure 5**, imitations of the manipulated (lengthened or shortened) VOT are overlaid on imitations of the canonical VOT, broken down by token. First, we consider the distributions of imitations of shortened VOTs, shown in the top panels. There is clear evidence of bimodality in most of the /p/ and /t/ tokens, with the lower peak indicating that some participants (by-participant distributions are shown in Appendix D) imitated these as unaspirated stops, presumably because they perceived them as voiced rather than voiceless stops. In contrast, there is no evidence of bimodality in imitations of shortened VOT for tokens beginning with /k/, suggesting that these tokens were consistently heard as voiceless stops, despite their low VOT.

Turning to the lengthened imitations, recall that we expected a reflection of the findings of Nielsen (2011): a rightward-shifted version of the canonical imitations. Our data instead show distributions with slightly offset peaks, but notably, with a substantial rightward skew. In contrast to the shortened condition, the shape of the distribution for lengthened imitations is fairly

consistent across all words (**Figure 5**). Observation of by-participant results indicates that this shape arises from substantial heterogeneity both within and across participants: some participants showed small-but-consistent modifications, some showed larger, consistent modifications, and some showed more variable, flatter distributions. Furthermore, as can be seen in **Figure 5**, some productions were hyper-lengthened, i.e., longer than the value of the model talker.

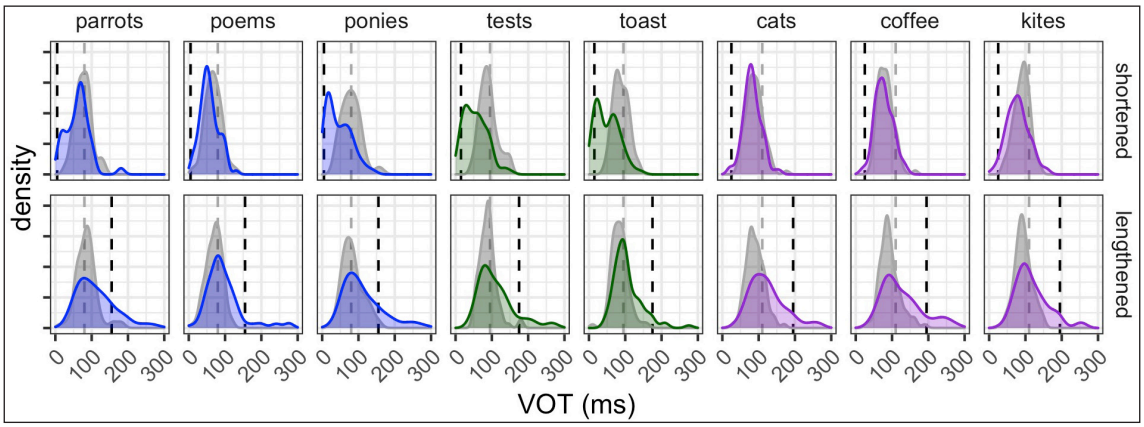

Figure 5: Density plots showing the distribution of VOTs in imitations of the canonical (grey) and manipulated (blue for /p/, green for /t/, purple for /k/) accent in each trial, broken down by token, in the shortened (top) and lengthened (bottom) conditions (Session 2 data only). Dashed lines indicate the model's VOT values for canonical (grey) and manipulated (black) conditions.

Observation of the distributions of shortened and lengthened VOT imitation calls into question the validity of the comparison of imitation across the two accent types: while the statistical analysis above suggested a similar *degree* of imitation, the *nature* of this imitation is quite different. The bimodality in the shortened condition also brings up the question of whether shortened VOT imitation effect might be solely attributable to a categorically different phonological target. To examine this possibility, and to provide a comparison of shortened vs. lengthened VOT imitation that eliminates this difference, we performed a post-hoc analysis on the subset of words beginning with /k/ (*cats*, *coffee*, and *kites*) in Session 2 (504 trials): these words showed no evidence of bimodality, so any imitation found in this subset is unlikely to be driven by perception of a categorically different segment.

Our post-hoc analysis was designed to test two questions: first, whether we still find imitation of shortened VOT after excluding tokens that were sometimes perceived as voiced, and second, whether the extent of imitation differs between shortened and lengthened VOT in this subset of stimuli. We used a linear mixed-effects model with the same structure as above (excluding the factor of Session) to test how the extent of imitation of /k/ words was predicted by AccentType, TalkerMatch, and their interaction, including random intercepts for Participant and Item, as well as random by-participants slopes for TalkerMatch and AccentType. A significant effect of

AccentType showed that there was more imitation of lengthened than shortened VOTs (30 ms for lengthened vs. 8 ms for shortened; $\beta = -22.19$, $SE = 5.23$, $t = -4.244$, $p = 0.003$), and a follow-up test indicated that there was indeed significant imitation for the shortened VOT stops ($\chi^2 = 7.84$, $p = 0.003$). Effects of TalkerMatch and the AccentType * TalkerMatch interactions were not significant ($p > 0.1$). We further note that the shape of the distributions for these shortened VOT imitations (**Figure 5**) shows less variability, and does not show the same type of skew, as the distributions of their lengthened VOT counterparts; rather, they are symmetrical and reflective of the shifted distributions that were found for *lengthened* VOT in Nielsen (2011). Implications of this will be discussed below.

To summarize the VOT results from the accent imitation task, participants showed imitation of accents differing in lengthened and shortened VOT, but the extent of imitation was not significant when participants heard different voices in the first session. Taking into account data from all words, the mean imitation values between the normal and modified VOT were 6 ms for the least accurate, Session1-DifferentVoice condition, compared with 21 ms in Session1-SameVoice, 28 ms in Session2-DifferentVoice, and 18 ms in Session2-SameVoice. There were no overall effects of AccentType, but in one condition (Session2-DifferentVoice), there was greater imitation for lengthened VOT than for shortened VOT. Although the overall extent of imitation was similar across the two AccentTypes, the shape of the distributions differed. A clear bimodal distribution for shortened VOT imitations was attributable to categorical perception and production of a different phonological category (voiced stops) for some tokens. However, the overall shortened VOT imitation effect cannot be fully attributable to this, as a post-hoc test including only tokens with no evidence of voiced perception (i.e., /k/-initial words) found significant imitation of shortened VOT (despite the lower power due to a smaller sample size), although the extent of imitation was of smaller magnitude than was found for lengthened VOT.

3.2.1 Vowel duration differences

Based on previous work (Nielsen, 2011), we expected that speakers might interpret – and imitate – VOT differences as overall duration/rate differences, and therefore wanted to test whether speakers also produced differences in vowel duration. For each production, we calculated the difference in vowel duration between the production imitating the modified-VOT (lengthened/shortened) sentence and the canonical-VOT sentence. We then tested whether this difference was significantly different than zero (i.e., if speakers showed modification of vowel duration), and whether the extent of difference varied across conditions. We tested this in a parallel way to the VOT difference analysis above, using a mixed-effects linear regression model with the same predictor variables (Session, TalkerMatch, and AccentType) and random effects structure, with estimates representing the average vowel duration difference corresponding to that factor. **Figure 6** shows the average vowel duration differences across conditions, and statistical results

are given in **Table 4**. The effect of AccentType is clear in the graph, with speakers showing consistent differences in the lengthened, but not the shortened condition. When imitating lengthened VOT, speakers also increased their vowel duration, while when imitating shortened VOT, there was not a clear pattern of change in vowel duration.

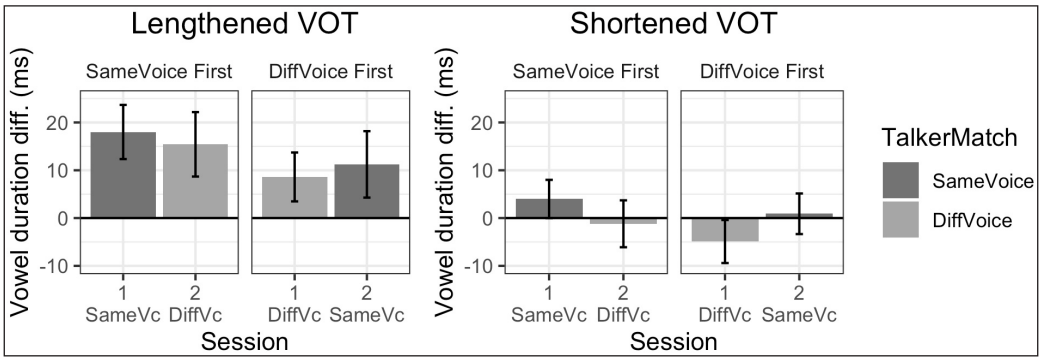

Figure 6: Accent imitation task: By-trial differences in vowel duration between productions imitating modified (lengthened/shortened VOT) vs. canonical values, across all conditions. Positive values indicate that the vowel duration following the modified sentence was longer than that following the canonical VOT sentence. Participants who heard the SameVoice condition in Session 1 and the DifferentVoice condition in Session 2 are shown in the left panel of each graph, and the other half of the participants are shown in the right panel. Error bars show 95% confidence intervals of by-participant means.

Table 4: Statistical results from a linear mixed-effects model predicting vowel duration differences (in ms) in the accent imitation task. Reference levels are in italics, and significant effects are shaded.

Factor	**Estimate**	**Std. Error**	***t* value**	***p***
(Intercept)	6.53	1.599	4.085	< 0.001
Session (*1* vs. 2)	0.216	1.275	0.170	0.866
TalkerMatch (*same* vs. different)	–3.853	1.275	–3.022	0.004
AccentType (*lengthened* vs. short)	–13.524	2.317	–5.838	< 0.001
Session * TalkerMatch	10.345	5.618	1.842	0.073
Session * AccentType	0.316	2.342	0.135	0.892
TalkerMatch * AccentType	–2.746	2.342	–1.173	0.241
Session * TalkerMatch * AccentType	–5.834	8.087	–0.721	0.475

In the statistical results, the significant (positive) intercept indicates that participants produced overall longer vowel durations for the modified compared to the canonical values, but there were main effects for TalkerMatch (with greater differences for Same vs. DifferentVoice)

and AccentType (with greater differences for lengthened vs. shortened AccentType), indicating that this difference may not hold across all conditions. Confirming the patterns seen in the graph, follow-up tests looking at the effect in the two AccentTypes separately show that the vowel duration difference is significantly greater from zero for the lengthened, but not for the shortened, condition (lengthened: $\chi^2 = 45.67$, $p < 0.001$; shortened: $\chi^2 = 0.01$, $p = 0.908$). Following up on the effect of TalkerMatch, we found that the vowel duration difference was significantly different from zero for both. There was also a trending two-way interaction between Session and TalkerMatch, which, when followed up, mirrored patterns from the VOT analysis above: the vowel duration difference was not significant in the Session1-DifferentVoice condition, but it was in the other three (Session1-DifferentVoice: $\chi^2 = 0.66$, $p = 0.416$, all others $p < .05$).

To summarize, although the only durational difference in the exposure stimuli was VOT, our participants also produced longer vowel durations when imitating lengthened VOT (by 14 ms on average). They did not modify their vowel durations when imitating the difference between shortened and canonical VOT (0 ms difference on average). The vowel duration modification was smaller in the DifferentVoice condition, particularly in Session 1. Increased vowel duration for the lengthened-VOT condition was expected: it has been found in other work (Wade et al., 2020), and it can be plausibly attributed to the idea that participants interpret the VOT differences at least in part as an overall durational difference, and therefore increase global duration in their imitation. However, speech rate differences alone cannot explain the overall findings of VOT imitation: as in Nielsen (2011), there was proportionally more VOT lengthening than vowel duration lengthening. In addition, there was no decrease in vowel duration for the shortened VOT condition.

3.3 Artificial accent discrimination results

Listeners' discrimination accuracies in the accent discrimination task are shown in **Figures 7** and **8**, and statistical results are in **Table 5**. We used a mixed-effects logistic regression model to evaluate the effect of our test variables on listeners' responses. The binary response variable was listeners' accuracy in each trial, and the fixed predictors, with reference levels in italics, were Session (*1* vs. 2), TalkerMatch (*same talker* vs. different talker), SentenceMatch (*same sentence* vs. different sentence), and AccentType (*lengthened* vs. shortened), as well as the three-way interactions between Session * TalkerMatch * AccentType and between Session * TalkerMatch * SentenceMatch (we did not include interactions involving AccentType and SentenceMatch because we did not have reason to expect that the extent of generalization would differ based on the AccentType). Random intercepts were included for Participant and Sentence ID (i.e., Item), as well as random by-participant slopes for TalkerMatch and AccentType, and by-item slopes for TalkerMatch. All factors were centered (—0.5, 0.5). In interpreting the model, the estimate of the intercept represents the log odds of an accurate response across all conditions, and the estimate corresponding to each fixed factor represents the difference in log odds between the two levels

of the factor. Likelihood ratio tests showed that including age as an additional predictor did not significantly improve model fit from the model without age, either when it was included as a main effect and in interaction with Session and TalkerMatch ($\chi^2 = 2.03$, $p = 0.731$), or when it was included as only a main effect involved in no interactions ($\chi^2 = 0.12$, $p = 0.728$).

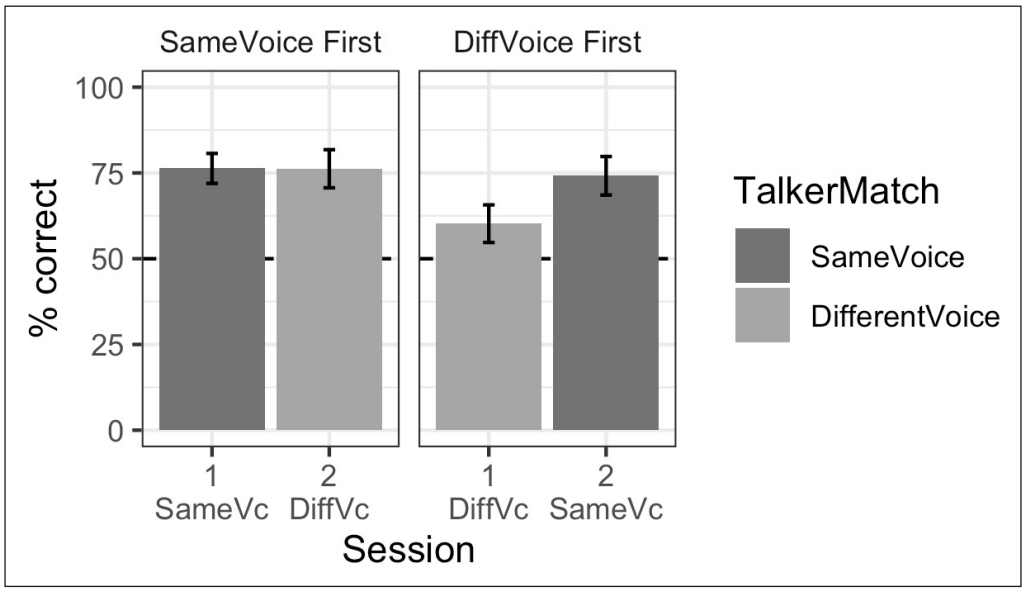

Figure 7: Accent discrimination task: Percentage accuracy by Session and TalkerMatch. Error bars show 95% confidence intervals of by-participant means.

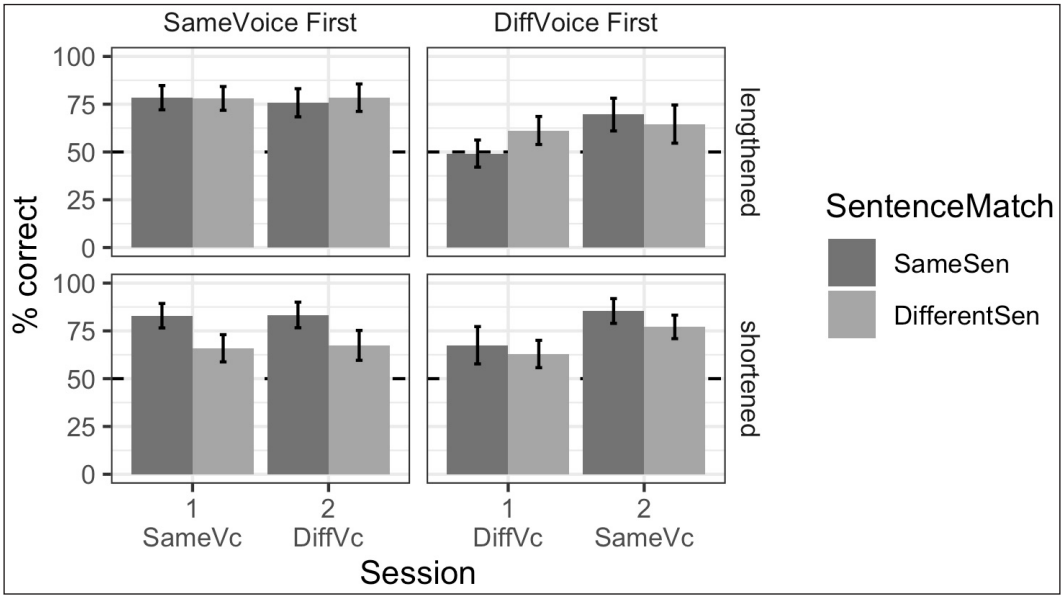

Figure 8: Accent discrimination task: Percentage accuracy broken down by all variables. Error bars show 95% confidence intervals of by-participant means.

Table 5: Statistical results from a logistic mixed-effects model predicting accuracy in responses on the accent discrimination task. Reference levels are in italics, and significant effects are shaded.

Factor	Estimate	Std. Error	*z* value	*p* value
Intercept	1.126	0.150	7.500	< 0.001
Session (*1* vs. 2)	0.392	0.109	3.583	< 0.001
TalkerMatch (*same* vs. different)	–0.391	0.127	–3.086	0.002
AccentType (*lengthened* vs. short)	0.302	0.264	1.142	0.254
SentenceMatch (*same* vs. different)	–0.330	0.238	–1.385	0.166
Session * TalkerMatch	0.992	0.393	2.527	0.011
Session * AccentType	0.293	0.154	1.905	0.057
Session * SentenceMatch	–0.256	0.152	–1.681	0.093
TalkerMatch * AccentType	–0.123	0.201	–0.609	0.542
TalkerMatch * SentenceMatch	0.375	0.198	1.893	0.058
Session * TalkerMatch * AccentType	–1.533	0.548	–2.800	0.005
Session * TalkerMatch * SentenceMatch	–0.645	0.305	–2.117	0.034

The significant (positive) intercept indicates that listeners were above chance overall, and that performance differed overall by Session and TalkerMatch, as expected, with listeners performing better in Session 2 than Session 1, and better in the SameVoice condition than in the DifferentVoice condition. However, there was a significant interaction between these two variables, and follow-up tests show that the effect of TalkerMatch was only significant in Session 1 (Session 1: $\chi^2 = 14.66$, $p < .001$; Session 2: $\chi^2 = 0.20$, $p = 0.654$). AccentType and SentenceMatch did not show significant main effects, but there were significant three-way interactions between each of these two variables and Session * TalkerMatch. We performed follow-up tests to determine whether there were significant effects of AccentType and/or SentenceMatch in different Session/TalkerMatch conditions. For AccentType, there was a significant effect in the SameVoice condition in Session 2 ($\chi^2 = 7.33$, $p = 0.007$), with the shortened accent condition eliciting higher accuracy than the lengthened condition, but no significant effects in the other three Session * TalkerMatch combinations (all $p > 0.1$). For SentenceMatch, there were no significant simple effects in any of the four Session * TalkerMatch conditions, although there was a trending effect for higher performance for previously exposed sentences (SameSentence) than novel sentences in the SameVoice condition, in both sessions (Session 1 SameVoice: $\chi^2 = 3.85$, $p = 0.050$; Session 2 SameVoice: $\chi^2 = 2.90$, $p = 0.088$; Session 1 and Session 2 DifferentVoice conditions, both $p > 0.1$). We also confirmed via follow-up tests that the effect of TalkerMatch was significant in Session 1, but not Session 2, when broken down by both levels of AccentType and TalkerMatch,

respectively. Finally, we performed follow-ups to test whether listeners were above chance in all Session * TalkerMatch conditions (i.e., even in the least accurate Session1-DifferentVoice condition), and they were (Session 1 Different talker: $\chi^2 = 6.10$, $p = 0.014$, all others $p < .001$).

To summarize, listeners were able to classify accents differing in lengthened and shortened VOT at above-chance levels, with performance substantially hindered (albeit still above chance) when presented with different voices in the first session. Mean accuracy rates were 60% in that condition, compared with 76% in Session1-SameVoice, 76% in Session2-DifferentVoice, and 74% in Session2-SameVoice. There was no consistent pattern found for AccentType, but in one condition (Session2-SameVoice), listeners were more accurate on the shortened than the lengthened condition. Listeners generalized to new sentences, with no significant difference between new and familiar sentences (although there was a trending effect for worse performance on new sentences in the SameVoice condition). This ability to generalize to new sentences supports the idea that participants were not simply performing shallow acoustic matching when doing this task, but were basing their responses on their awareness of systematic differences in the phonetic realization of voiceless stops as a category.

3.4 Word-level discrimination

The word-level discrimination task was an ABX discrimination task on pairs of tokens of the word *toast* that differed by either one or two steps along a five-step VOT continuum (15, 55, 95, 135, and 175 ms). We expected that ease of discrimination would differ based on the acoustic distance between the two (1 or 2 steps), as well as based on whether or not the pair straddles a phonological category boundary (with higher accuracy expected in between-category pairs). As we expected the category boundary to fall between 15 and 55 ms (consistent with previous work on English alveolar stop perception, e.g., 35 ms in Kuhl and Miller, 1978, and approximately 25 ms in Benkí, 2001), we created a factor Shared Category indicating whether the target pair of stimuli was *between-category* (all pairs including a token with 15 ms VOT), or *within-category* (all other pairs, since all other tokens had 55 ms VOT or greater). We analyzed the results with a mixed-effects logistic regression model predicting accuracy from the two predictor variables Step Distance and Shared Category, both simple-coded as (–0.5, –.5), as well as their interaction. A random by-subjects intercept and random by-subjects slopes for Step Difference and Shared Category (uncorrelated with the random intercepts) were included. Discrimination accuracy is shown in **Figure 9**, and statistical results are shown in **Table 6**.

The significant (positive) intercept indicates that listeners were above chance overall. The effect of Step Difference was not significant, although it trended in the expected direction, with listeners showing numerically higher accuracy for the two-step than for the one-step difference. As expected, listeners showed higher accuracy for pairs that likely straddled a category boundary (i.e., 15 vs. 55 ms, or 15 vs. 95 ms) than those that were likely perceived as the same category

(73% vs. 59% accuracy on average). A follow-up test showed that despite having lower accuracy, the within-category pairs were still discriminated significantly above chance ($\chi^2 = 15.37$, $p < 0.001$). There was no interaction between Step Difference and Shared Category. Likelihood ratio tests showed that including age as an additional predictor did not significantly improve model fit ($\chi^2 = 0.02$, $p = 0.876$).

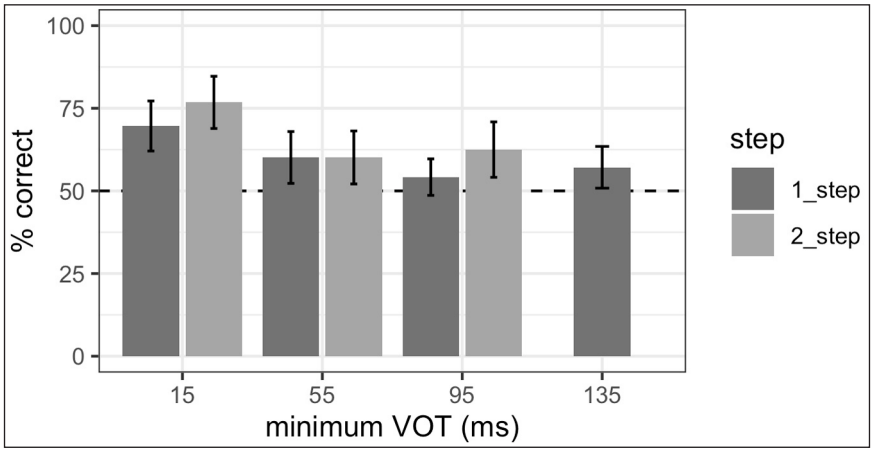

Figure 9: Word-level discrimination task: Percentage accuracy on an ABX discrimination task, broken down by VOT step distance (1 step = 40 ms; 2 steps = 80 ms) and minimum VOT of the pair. Error bars show 95% confidence intervals of by-participant means.

Table 6: Statistical results from a logistic mixed-effects model predicting accuracy in responses on the word-level discrimination task. Reference levels are in italics, and significant effects are shaded.

Factor	Estimate	Std. Error	*z* value	*p* value
(Intercept)	0.740	0.101	7.331	< 0.001
StepDifference (*1* vs. 2)	0.284	0.149	1.913	0.056
SharedCategory (*between* vs. within)	–0.700	0.173	–4.056	< 0.001
StepDifference * SharedCategory	–0.209	0.299	–0.700	0.484

3.5 Word-level imitation

We tested whether participants modified their VOT and vowel duration when asked to imitate tokens drawn from the same five-step VOT continuum described above. Mean values for VOT and vowel duration as a function of continuum step are shown in **Figure 10**, and statistical results are shown in **Table 7**. Statistical analysis was done using two mixed-effects linear regression models, with the response variables being VOT in one model and vowel duration in the other, with a predictor variable of VOT step (scaled to z-scores and treated as a continuous variable) in both

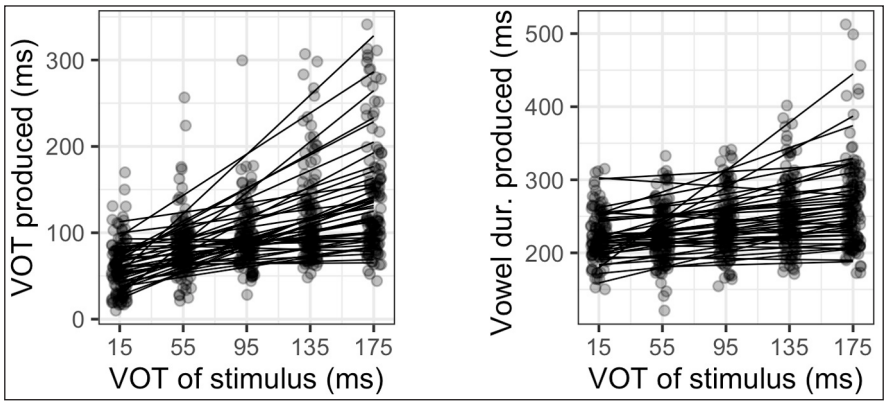

Figure 10: Word-level imitation task: VOT (left) and vowel duration (right) of productions following stimuli from a VOT continuum. Dots show raw values; lines show best-fit regression lines by participant.

Table 7: Statistical results from two linear mixed-effects models predicting VOT and vowel duration from VOT step in the word-level imitation task. Significant effects are shaded.

	VOT (ms)				**Vowel duration (ms)**			
Factor	**Estimate**	**Std. Error**	***t* value**	***p* value**	**Estimate**	**Std. Error**	***t* value**	***p* value**
(Intercept)	101.87	5.14	19.82	< 0.001	242.78	5.53	43.88	< 0.001
VOT step	26.89	3.51	7.66	< 0.001	15.85	2.95	5.36	< 0.001

cases. Random by-participant intercepts and slopes for VOT step were included in both models. In both models, the effect of VOT step was significant, indicating that participants increased both their VOT and their vowel duration as the VOT of the stimulus increased. To give a concrete example of the effect size, the mean VOT values for imitations at the endpoints of the continuum were 62 ms (following the shortest, 15 ms step) to 142 ms (following the longest, 175 ms step), while mean vowel durations at these endpoints were 224 and 266 ms. As can be seen in the figures, while there were some participants who showed much greater effects of VOT step than others, most participants showed an overall increase. Likelihood ratio tests showed that including age as an additional predictor did not significantly improve model fit ($\chi^2 = 1.91$, $p = 0.385$).

3.6 Relationship between tasks

To compare performance across tasks on an individual level, we calculated by-participant indices for each of the four tasks: the accent imitation and discrimination tasks (completed over two sessions) and the word-level imitation and discrimination tasks (completed at the end of Session 2 by all participants). Recall that in the accent imitation and discrimination tasks, there were large discrepancies in performance between participants in the different TalkerMatch conditions

during Session 1: both discrimination accuracy and extent of imitation were much lower in the DifferentVoice condition than the SameVoice condition. On the other hand, during Session 2, performance was largely similar for all participants, regardless of whether they were listening to the same or different voices. This discrepancy presented a challenge in our task of calculating individual indices of imitative ability. The most straightforward approach, calculating an index based on the full dataset from each participant, would result in systematically lower values for participants who had heard the DifferentVoice condition first, but this would not reflect an actual difference in imitative ability. Therefore, we used only Session 2 data to calculate individual indices for the accent tasks, which we take to be a better indicator of imitative ability. It should therefore be kept in mind that half of the participants' data is from the SameVoice and the other half is from the DifferentVoice condition; however, as we did not see systematic differences between these conditions in Session 2, we think that this is a reasonable indicator of their overall imitative and discriminative ability.

Indices were calculated as follows. For the perception tasks (both accent and word-level), indices were taken to be each participant's overall accuracy across target trials. For the accent imitation task, the index was calculated as the mean VOT imitation using the same metric as for the group-level analysis (by-trial VOT difference between the two accents, transformed such that positive values correspond to changes in the expected direction). For the word-level imitation task, the index was the slope of the best-fit regression line predicting an individual's VOT from the stimulus VOT step (i.e., the slope of one of the lines shown above in **Figure 10**). Our primary analysis tested how artificial accent imitation was predicted by each of the other three tasks, and we further tested the relationship between 1) word-level discrimination and word-level imitation and 2) accent discrimination and word-level imitation, as discussed below.

Scatterplots of the comparisons of interest are shown in **Figure 11**, and statistical results are shown in **Table 8**. Pearson's product-moment correlations were computed for each of these using the cor.test function in R. All correlations between accent imitation and each of the three other tasks were positive; those with the accent discrimination and word-level imitation tasks were significant, while that with the word-level discrimination task was not. The two significant correlations provide evidence that success in artificial accent imitation was predicted both by subjects' ability to perceive the difference in the different accents (as evidenced by the significant correlation with the accent discrimination task) and their low-level VOT imitation ability (as evidenced by the significant correlation with the word-level imitation task). It is possible that these are not independent predictors: perhaps talented imitators are high performers in both of these tasks. In this case, we should also see a significant correlation between individual performance on the word-level imitation and the accent discrimination tasks. However, this correlation was not significant. As with any null result, it is possible that the effect was simply not strong enough to be detected with an experiment of our sample size. Based on a post-hoc power analysis, our

design had sufficient power to detect correlations of $r = 0.45$ or higher with the tests we used (i.e, 'moderate' or stronger, according to Evans 1996), and therefore may not have been sensitive enough to pick up weaker effects.[8] Still, if there is an effect, it is likely very weak, indicating a small amount of shared variance, and we therefore take this as evidence of independence between these two predictors of accent imitation. Finally, we tested the relationship between word-level imitation and discrimination, and found no significant correlation.

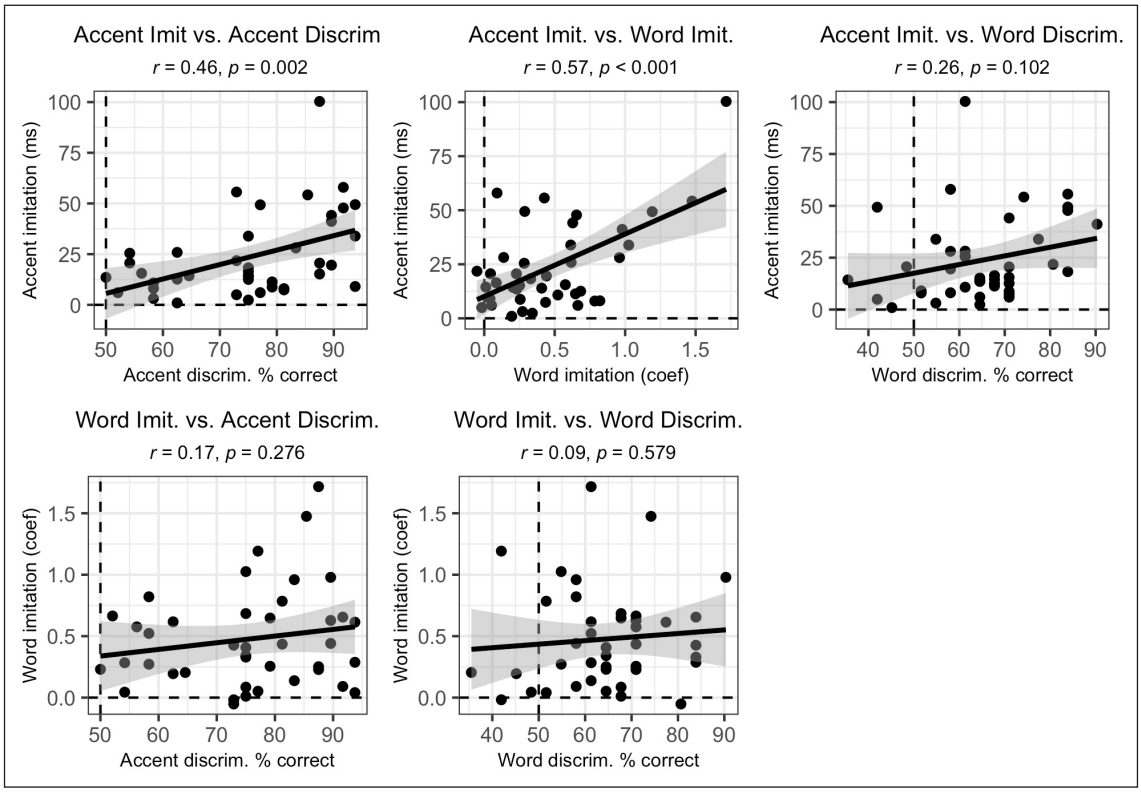

Figure 11: Relationship between individual participants' performance on the tasks: Accent imitation vs. accent discrimination, word imitation, and word discrimination (top), word-level imitation vs. accent discrimination and word-level discrimination (bottom right). Error bars show 95% confidence intervals surrounding the best-fit regression line. Dotted lines show chance performance (for discrimination tasks) or no imitation (for imitation tasks). Correlation coefficients and p-values are shown; full statistical results are provided in **Table 8**.

[8] Our preregistered design did not include a power analysis for the correlational analyses, but we performed a post-hoc power analysis with the goal of determining the minimum correlation coefficient that would be detectable given our sample size and variance present in our data (we would like to thank an anonymous reviewer for this suggestion). The simulation-based analysis consisted of calculating the minimum threshold for the correlation coefficient that would result in 80% significance across 5000 iterations, and the threshold ended up being between .4 and .45 for all pairwise comparisons. Details are available in the supplementary material.

Table 8: Results of correlation analyses testing relationships between performance across tasks. For correlations involving accent imitation, results are shown for three different calculations of the accent imitation index: an overall index based on all words (left), and indices based on the subset of /k/-initial words, with either lengthened (middle) or shortened VOT (right).

	Imitation of all words			**Imitation of /k/ words with lengthened VOT**			**Imitation of /k/ words with shortened VOT**		
Comparison	***r***	***t***	***p***	***r***	***t***	***p***	***r***	***t***	***p***
Accent imitation vs. Accent discrimination	0.46	3.26	0.002	0.46	3.27	0.002	0.22	1.41	0.166
Accent imitation vs. Word-level imitation	0.57	4.37	<0.001	0.38	2.60	0.013	0.12	0.74	0.462
Accent imitation vs. Word-level discrimination	0.26	1.67	0.102	0.27	1.75	0.088	0.34	2.32	0.025
Word-level imitation vs. Accent discrimination	0.17	1.11	0.276						
Word-level imitation vs. Word-level discrimination	0.09	0.56	0.579						

The analyses done above were based on imitation data from all words. However, recall that shortened-VOT imitations appeared to be characterized by categorical segmental substitutions in many words. When considering only the subset of data where this issue did not arise (i.e., /k/-initial words used in the post-hoc analysis for the accent imitation task), we found different imitation behavior for shortened vs. lengthened VOT stimuli, calling into question the validity of an aggregate individual imitation index. We therefore calculated a by-participant imitation index in the same way as above for shortened and lengthened VOT separately, considering the /k/-word subset. Using these indices, we first ran a correlation test to determine participants' consistency in their extent of imitation across the two accent types; in other words, whether participants with the greatest lengthened VOT imitation also showed the greatest shortened VOT imitation. The correlation was not significant ($r = 0.23$, $t = 1.46$, $p = 0.151$).

We then revisited the same correlations done above: those between the accent imitation task and the three other tasks. However, instead of using the aggregate imitation index based on all words, we ran two sets of post-hoc correlation analyses: one set using the lengthened index and one set using the shortened index, both calculated from the subset of words beginning with /k/. We tested whether individual performance on each of the three tasks was predictive of lengthened and shortened accent imitation separately, within this subset of data, with results shown in **Table 8**. Participants' imitation of lengthened VOT patterned the same way as in the original correlational analysis: performance on the accent discrimination and word-level imitation tasks was predictive of performance on the accent imitation task. On the other hand,

imitation of shortened VOT was not significantly correlated with either of these sub-tasks, but *was* significantly correlated with the word-level discrimination task, the one task that had *not* been correlated with the aggregate value of imitation across all words.

4. Discussion

4.1 Summary of results

This work tested the effects of talker variability on explicit imitation of artificial accents with lengthened and shortened VOT, and it explored how individual variability in tasks targeting different perception- and production-based sub-processes predicted imitative performance. Participants imitated artificial accents characterized by shortened, as well as lengthened, VOT, in contrast to previous work showing only lengthened-VOT imitation (Nielsen 2011). Some of the shortened VOT imitation was attributable to the fact that certain tokens were perceived and imitated as a different phonological category (i.e., /p/ perceived and imitated as voiced /b/). Interestingly, even when these tokens were omitted, imitation of shortened VOT was found, albeit to a lesser degree than imitation of lengthened VOT; however, results of this post-hoc analysis need to be interpreted with caution, since it is only based on a subset of data and since our assumptions of perception of the stimuli as voiced vs. voiceless were based on indirect evidence.

Both imitation and discrimination of VOT differences were inhibited by voice-related variability: participants initially presented with a multi-voice condition (DifferentVoice) showed no imitation and had less accurate discrimination performance than participants in a condition with no voice-related variability (SameVoice). However, participants who first completed the condition with no voice-related variability showed equally good performance on the subsequent multi-voice condition, indicating that the inhibitive effect of variability was mitigated by prior experience with the no-variability condition.

Individual imitative performance was independently predicted by two subtasks: accent discrimination and word-level imitation. Performance on the word-level discrimination task was not correlated with overall imitation, but was found to be correlated with the extent of imitation of shortened VOT in a post-hoc analysis. Exploratory analyses did not provide any differences for age-based differences in explicit imitation; however, given that our study was not designed to test this factor, we cannot draw strong conclusions about this and leave more systematic investigation of the effect of age on imitation for future work.

4.2 Sub-processes of imitation

Our first research question focused on the roles of different sub-components of imitation in predicting imitative performance by examining correlations between artificial accent imitation and each of three tasks designed to tap into various sub-processes of variation: accent discrimination,

word-level imitation, and word-level discrimination. We found that imitative performance was significantly correlated with the first two of these tasks: individuals who imitated artificial accents more faithfully were more accurate in discrimination of the same artificial accents, and these individuals also showed more faithful imitation of word-level VOT differences. The predictiveness of the accent imitation task indicates that the ability to identify and/or generalize the relevant feature (VOT) as a property of an artificial accent is an important predictor of imitative performance. The predictiveness of the word-level imitation task shows that variability in production-based processes (articulatory precision, flexibility, and/or willingness to diverge from production norms) is similarly important. Furthermore, the fact that these two tasks were independently predictive, uncorrelated with one another, is critical for two reasons. First, it shows that the relevant variability cannot be ascribed to general factors external to the specific tasks (e.g., motivation to complete experimental tasks) or to factors shared by the two tasks (e.g., low-level VOT discrimination acuity). Second, it suggests that performance on the sub-processes targeted by each task patterns differently across individuals, such that an individual might excel at certain sub-components of imitation but not others.

On the other hand, performance on the word-level discrimination task was not predictive of overall performance on the artificial accent imitation task, suggesting that variability in low-level perceptual ability is not a primary driver of differences in accent imitation. This does not mean that low-level perception is irrelevant to accent imitation; in fact, in a post-hoc analysis involving a targeted subset of data, we found that word-level discrimination was correlated with imitation of certain shortened VOT in some cases, as is discussed in more detail below. However, the overall lack of correspondence between the word-level discrimination task and the accent imitation task, as well as the lack of correspondence between the word-level discrimination and word-level imitation task, suggests that sources of variability targeting other sub-processes override differences in low-level perception when accounting for differences in explicit imitation.

The independence of articulatory vs. perceptual predictors of artificial accent imitation highlights the idea that there are multiple, distinct reasons why individuals may vary in imitative ability. While our results are only directly applicable to explicit imitation, we propose that the idea of distinct perception- and production-based predictors may in part account for the difficulty in identifying robust predictors of stable individual differences in imitation more generally (e.g., Cohen Priva & Sanker, 2020; Wade, 2022). While there has been extensive interest in phonetic imitation and its predictors, there has been little research into the level of influence of these predictors. Conceptualizing imitation as a set of sub-processes could provide a clearer framework for testing predictions about characteristics expected to influence imitative ability. It is plausible that there are indeed systematic cognitive traits governing individual variation in imitation, but that simply testing for correlations between the proposed traits and an imitation task may not be the most fruitful way of identifying them. Instead, it is first important to identify which

sub-process the trait is expected to influence: some may be predicted to influence perceptual processes (e.g., musicality, Coumel et al., 2019), some to influence low-level production (e.g., "articulation space"; Reiterer et al., 2013), while others may plausibly be expected to influence both (e.g., focus, or lack of attention-shifting; Yu et al., 2013). Making explicit predictions about the level at which a given trait is expected to have an effect, and testing individual sub-processes separately, would result in tighter predictions and higher-powered analyses that would be more likely to result in robust, replicable effects – and facilitate the search of isolating individual predictors amidst the noise of a cognitively complex process.

4.3 The role of talker variability in accent imitation and perception

Based on effects of variability found in past work in accent classification tasks (e.g., Clopper & Pisoni, 2004), phonetic processing (Mullennix et al., 1989; Mullennix & Pisoni, 1990), and accent adaptation/perceptual learning (review in Baese-Berk, 2018), we expected that voice-related variability might hinder performance in both the imitation and discrimination tasks. We did indeed see heavily degraded performance in both tasks for those participants who completed the multiple-voice condition (DifferentVoice) before the no-variability condition (SameVoice). However, we found no disruption at all for those participants who had prior experience with the no-variability condition.

What is the source of the inhibitive effect of voice-related variability? It does not appear to be attributable to a general increased processing cost in the context of greater acoustic variability (e.g., Mullennix et al., 1989; Mullennix & Pisoni, 1990), since a low-level effect of this nature would be expected to persist even after experience with a no-variability condition. Instead, we posit that this occurred at the level of identification: the presence of additional voice-related acoustic variability may have made it more difficult to identify the target feature (VOT) as a relevant feature of the target of imitation. For those participants who first completed the no-variability condition, the acoustic homogeneity may have directed attention to VOT as an appropriate dimension to use as a grouping variable, and this knowledge was retained and used in the subsequent variable-voice session.

The same participants who showed low performance in the variable-voice condition successfully imitated and discriminated catch trials characterized by a rhotic vs. nonrhotic distinction. This indicates that the decreased performance in the presence of voice-related variability is not attributable to general confusion with the task, and also shows that the inhibitory effect is not categorical. Some features appear to be more salient and easier to pick up on as a group-level property of an accent; an interesting topic for future work would be to examine which features are more easily identified as properties of an accent (as with the rhotic/nonrhotic distinction in the current work) and which may need more directing of attention (as with VOT in the current work).

An additional hypothesis was that talker-related variability might facilitate performance in conditions requiring generalization (e.g., Clopper & Pisoni, 2004; Bradlow & Bent, 2008; Schmale et al., 2012). We found little evidence in support of this: while a trending effect of lower accuracy for the generalization sentences in the single-talker condition hints that this may be something to explore further, we cannot make strong claims based on the current findings.

Previous work has posited that, while natural talker-related variability will incur phonetic processing costs, acoustic variability that is purely attributable to properties of a talker's voice, and not relevant to phonetic identification, is automatically filtered out and therefore does not incur these same processing costs (e.g., overall f0: Sommers & Barcroft, 2006; see also Bradlow et al., 1999; Sommers et al., 1994). The substantial inhibitory effect found in the current work goes against this idea; the voice-related variability clearly inhibited performance on the tasks in this work, despite the fact that the variability was carefully controlled to vary only in properties not relevant for English phonetic identification (overall f0 and formant scaling). One difference between this work and work where no inhibitory effect was found is the nature of the manipulations: while f0 was manipulated in Sommers and Barcroft (2006), none of the previous studies manipulated formant scaling. While it is possible that this additional acoustic variability is the source of the different results, we think it more likely that the discrepancy is due to the different tasks used in the different studies: spoken word identification in the previous studies, compared with imitation and discrimination of group-level features in the current work, which requires more active engagement in deciding which phonetic information is relevant to the task. A direct test of the influence of minimal spectral scaling differences on word identification would be necessary to confirm this.

Finally, our results may point to some practical implications for training paradigms for second language sound training and/or perceptual learning. While high-variability training paradigms have been shown to enhance these kinds of learning, their efficacy is also fairly inconsistent both within and across studies (e.g., Baese-Berk, 2018; Perrachione et al., 2011). As we have seen in the current work, increased variability may make it more difficult to pay attention to, and therefore identify and encode, relevant phonetic detail, presenting a barrier to learning. Ensuring that learners are aware of the relevant dimensions prior to exposure or training, either through explicit instructions or by drawing their attention to the relevant dimensions by including *less* variability in an early stage of training, as in the no-variability condition in the current work, may provide a relatively efficient way to maximize the effects of high-variability training.

4.4 Explicit vs. implicit imitation of VOT differences

While we did not directly test the difference between explicit vs. implicit imitation in the current study, our findings have implications for the similarities and the differences between them. Recall that previous work has found a greater degree of imitation in explicit compared to implicit tasks (Dufour & Nguyen, 2013; Pardo et al., 2010; Sato et al., 2013). Dufour and Nguyen (2013)

posited that the two processes share an automatic general mechanism; however, Sato et al. (2013) found no correlation between the same participants' performance on explicit and implicit tasks, calling this relationship into question.

These insights from previous work fit in with our conceptualization of the processes underlying explicit imitation and with our empirical results. Under our view, explicit imitation subsumes any automatic processes involved in implicit phonetic convergence, but also requires additional, distinct "controlled" processes that are governed by additional, distinct conditioning factors, including things like willingness to diverge from speech norms, metalinguistic awareness, or experience with imitation. We speculate that there may be greater variability in these traits than there is in those processes governing implicit imitation. If this is the case, then we would expect to see not only more imitation in explicit than implicit tasks (as shown in earlier work), but also more variability, and this is exactly what we found in our lengthened VOT condition. Recall that participants' imitations of lengthened VOT followed a right-skewed distribution, suggesting that some participants in our experiment were willing and able to make relatively large VOT adjustments, and that there was substantial between-participant variability in the extent of imitation. This skewed distribution stands in contrast to results of Nielsen (2011), in which distributional patterns suggested small but consistent changes in participants' VOTs after exposure to lengthened VOT. Therefore, in comparison to Nielsen's (2011) findings of lengthened-VOT imitation, we found both a larger imitative effect (mean 22 ms overall in the current study, vs. 7 ms in Nielsen's) and a more variable distribution. While the larger magnitude of effect is likely partially attributable to the more extreme VOT values in our lengthened stimuli, we think that differences in the nature of the task augment the effect, and crucially, that the task difference accounts for the greater individual variability found in the distributions.

On the other hand, the distribution for shortened VOT, when considering the subset of stimuli which did not show categorically different perception, showed a shifted distribution, suggesting small-but-consistent changes with little across-participant variability. While any interpretations of this data must remain speculative given the post-hoc nature of the analysis, we propose that a production-based constraint for contrast preservation may be partially responsible for differences in the nature of imitation of shortened, as compared to lengthened, VOT. Given that there *is* imitation of shortened VOT, this is not a strong, categorical constraint completely inhibiting shortening of VOT. Instead, we propose that the controlled sources of variability that are responsible for individual variation in lengthened VOT imitation are inhibited in the shortened condition, leaving less room for individual variability in imitation.

Our correlation analyses provide some tentative support for this idea. We might expect to find that performance in tasks where low-level perceptual acuity plays a central role is more predictive of more "automatic" types of imitation than it is predictive of imitation where "controlled" processes play a larger role. Consistent with this, we found that performance on a

word-level discrimination task (where we assume low-level perceptual acuity is more central) was predictive of variability in shortened, but not lengthened, VOT imitation, while the tasks involving controlled factors (accent discrimination and word-level imitation) were predictive of lengthened, but not shortened, VOT imitation.

4.5 Limitations, future directions, and conclusion

The methodology used in this work diverged from previous work in several ways, such that caution is warranted in the interpretation of our results and in comparison with previous work. First, we designed our tasks to try to tap into different sub-processes of explicit imitation of systematic phonetic variation; however, it is difficult to confirm whether we successfully targeted the intended processes. For example, we designed the accent vs. word-level tasks to try to tap into different targets of imitation (general properties of an accent vs. specific properties of a token) by including talker variability and, in the discrimination stage, sentence variability in the accent discrimination task, and this difference was reinforced in instructions given to participants. Performance on the generalization questions in discrimination suggests that participants were indeed interpreting the variation as we intended, but a direct test of generalization in production would be necessary to confirm this. More broadly, we think that the question of what the intended target is in imitation tasks, and how this might differ based on the instructions and/or presentation context, is one worthy of investigation in and of itself.

Performance on our word-level discrimination task was lower than expected, with 73% accuracy for between-category trials, when we would have expected close to ceiling for these stimuli. While we do not have any clear explanation for this relatively low performance, or the lack of significant effect of Step Difference, it may be due to the specific baseline token we used. Perceptual sensitivity to VOT differences with the same raw values may differ across words (as evidenced by the different imitation strategies shown for the various words in the accent imitation task, **Figure 5**), or even based on the properties of a particular token of a word. It is possible that the stimuli in the continuum used for our word-level task were particularly difficult to discriminate. This could occur if, for example, secondary acoustic cues in the baseline token contributed to a bias toward /t/ perception, even in the low-VOT range, essentially creating a completely within-category series of stimuli. This possibility points to the importance of using multiple words in this sort of perception task; in future work using the current paradigm, it would be preferable to use the same tokens in both the word- and accent-level tasks, such that only the methodology differs between the two tasks.

While we have drawn tentative conclusions about some aspects of the nature of, and factors conditioning, explicit imitation, as distinct from the better-studied phenomenon of implicit imitation/phonetic convergence, this study did not include a direct comparison between the two types of imitation. Doing so in the future would allow for a test of our predictions about the

differences between the two types of imitation outlined above. Another future step, which can be done straightforwardly with this paradigm, would be to expand the range and number of features to further test the role of linguistic selectivity and perceptual salience in imitation. The paradigm also can easily incorporate naturalistic accents, allowing for tests of how factors like familiarity with a specific accent and its social connotations affect phonetic imitation.

Overall, this work explored the nature of explicit imitation of systematic phonetic variation in artificially-constructed accents. Our results provided evidence for the independent roles of perception- and production-based processes in predicting individual imitative ability; this highlights the fact that future work examining how individual differences in social, cognitive, and/or linguistic traits influence imitation should consider the level at which the target trait is expected to exert its influence. Imitation was substantially hindered by voice-related variability during exposure, indicating that even "phonetically irrelevant" variability (f0 and formant scaling) affects the ability to identify features of an accent, and that exposure in the context of less variability can direct attention to the relevant contrast. Finally, we found imitation of shortened as well as lengthened VOT, in contrast to previous work, but saw both qualitative and quantitative differences in imitation of the two manipulations. We hope that the framework provided here can be used in the future to build on and test the generalizability of our results with different linguistic features, with the broader aim of arriving at a fuller understanding of the linguistic, social, and cognitive factors governing explicit imitation.

Appendix A

Table A1: The following sentences were used for the accent-level imitation and discrimination tasks. The shaded sentences at the bottom were used as practice/catch trials.

Exposure (SameSentence)	**Generalization (DifferentSentence)**
Coffee, toast, eggs, and cereal are what I ate this morning.	Love, caring, patience, and fairness are all virtues they have.
Parrots, ferrets, cats, and fish live with me in my home.	Tigers, pythons, lions, and lizards are animals that scare him.
Tests, poems, essays, and journals will be used for evaluation of marks.	Teacher, chemist, journalist, or singer are jobs I'd like in the future.
Ponies, kites, novels, and art were things I liked when I was young.	Curry, tofu, rice, and chives are the foods she wants for lunch.
Practice/catch trials:	
The bike is much slower than the car.	I used sugar to make the coffee sweeter.

Appendix B

Table B1: Summary of the course of the accent tasks for a sample participant. Colors indicate different "voices" (different f0/formants).

SESSION 1 (SameVoice condition)					
Accent-Type	**Trial Set**	**Phase**	**Talker F (canonical)**	**Talker J (noncanonical)**	**Talker X**
Length-ened VOT	1	Exposure	"Parrots, ferrets…"	"Parrots, ferrets…"	--
			"Parrots, ferrets…"	"Parrots, ferrets…"	--
		Imitation	"Parrots, ferrets…"	"Parrots, ferrets…"	--
			"Parrots, ferrets…"	"Parrots, ferrets…"	--
		Discrimination: SameSen (order randomized)	"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (F)
			"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (J)
			"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (F)
		Discrimination: DifferentSen (order randomized)	"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (J)
			"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (F)
			"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (J)

(Contd.)

SESSION 1 (SameVoice condition)					
Accent-Type	**Trial Set**	**Phase**	**Talker F (canonical)**	**Talker J (noncanonical)**	**Talker X**
	2 3 4	*Same as Trial Set 1, with different sentences for each trial set. The same voice was used for all trial sets in this block.*			
Short-ened VOT	5 6 7 8	*Same as Trial Sets 1-4, but with shortened instead of lengthened VOT as the noncanonical Talker J. All stimuli in this block had the same voice, and this voice was different from the one used in the first block.*			
SESSION 2 (DifferentVoice condition)					
Length-ened VOT	**Trial Set**	**Phase**	**Talker F (canonical)**	**Talker J (noncanonical)**	**Talker X**
	1	Exposure	"Parrots, ferrets…"	"Parrots, ferrets…"	--
			"Parrots, ferrets…"	"Parrots, ferrets…"	--
		Imitation	"Parrots, ferrets…"	"Parrots, ferrets…"	--
			"Parrots, ferrets…"	"Parrots, ferrets…"	--
		Discrimination: SameSen (order randomized)	"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (J)
			"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (F)
			"Parrots, ferrets…"	"Parrots, ferrets…"	"Parrots, ferrets…" (J)
		Discrimination: DifferentSen (order randomized)	"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (F)
			"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (J)
			"Parrots, ferrets…"	"Parrots, ferrets…"	***"Tigers, pythons…"*** (F)
	2 3 4	*Same as Trial Set 1, with different sentences for each trial set. The same set of voices was used for all trial sets in this block.*			
Short-ened VOT	5 6 7 8	*Same as Trial Sets 1-4, except:* - *Noncanonical Talker J had shortened instead of lengthened VOT.* - *The four voices used for Talkers F and J were different than that of the first block.*			

Appendix C

Table C1: Acoustic properties used to create the different voices used in the tasks. T13 is based on the natural production of a female native speaker of English.

Talker	Formant shift ratio	median f0	Talker	Formant shift ratio	median f0
T4	0.775	130	T11	0.95	200
T5	0.8	140	T12	0.975	210
T6	0.825	150	T13	1	220
T7	0.85	160	T14	1.025	230
T8	0.875	170	T15	1.05	240
T9	0.9	180	T16	1.075	250
T10	0.925	190	T17	1.1	260

Table C2: Talkers T8 and T13 were used for trials in the SameVoice conditions. For the DifferentVoice conditions, the following combinations were used. Full information about which voices were used in each trial, as well as the stimuli themselves, are included in the supplementary material.

Talker F	Talker J	Talker X	Talker F	Talker J	Talker X
T4	T13	T6	T12	T5	T6
T4	T13	T7	T12	T5	T7
T4	T13	T10	T12	T5	T10
T4	T13	T11	T12	T5	T11
T4	T13	T14	T12	T5	T14
T4	T13	T15	T12	T5	T15
T8	T17	T6	T16	T9	T6
T8	T17	T7	T16	T9	T7
T8	T17	T10	T16	T9	T10
T8	T17	T11	T16	T9	T11
T8	T17	T14	T16	T9	T14
T8	T17	T15	T16	T9	T15

Appendix D

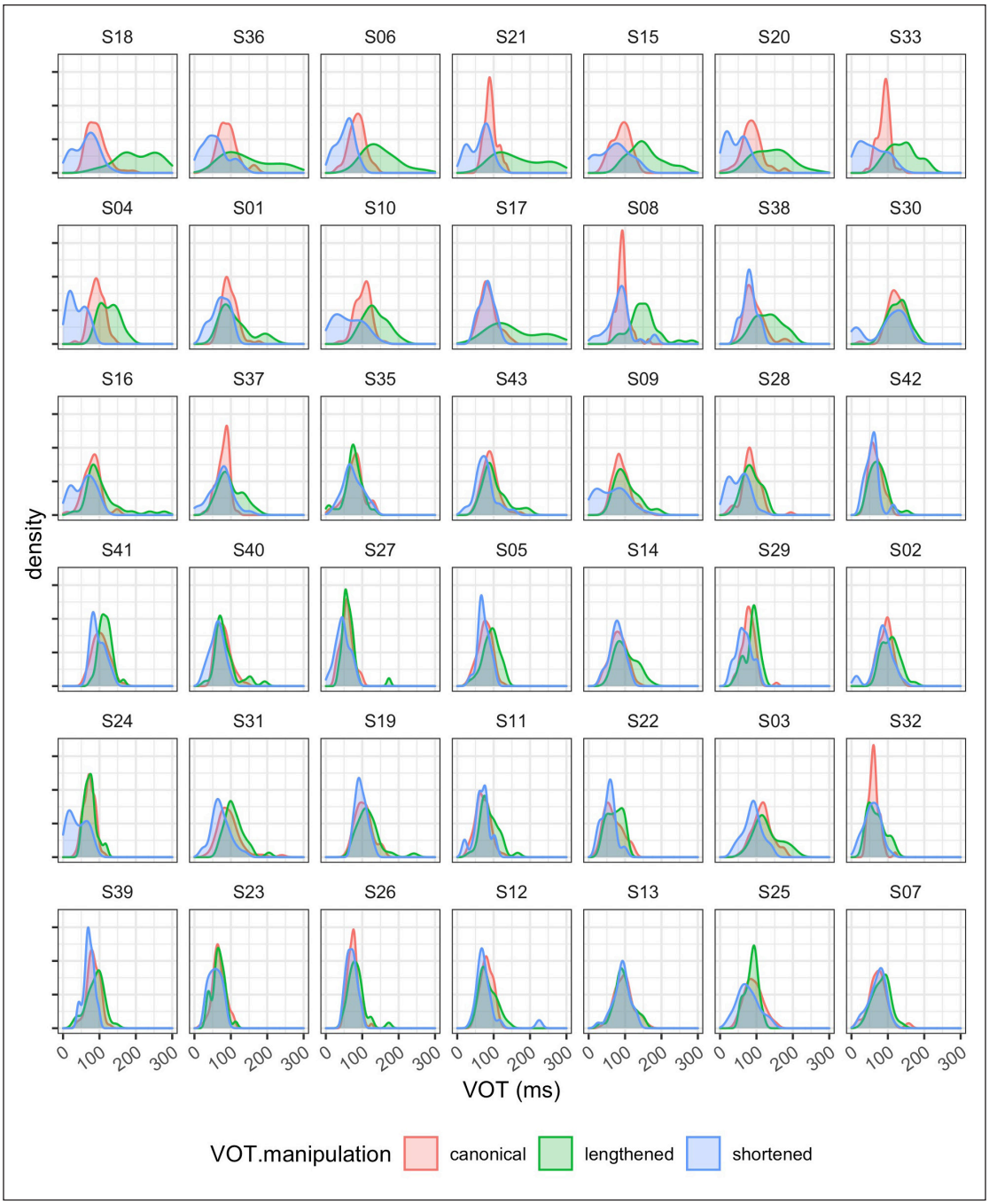

Figure D1: By-participant distributions of VOT values in Session 2 of the accent-level imitation task. Participants are sorted by extent of imitation (greatest to least). Canonical productions from both lengthened and shortened conditions are grouped together.

Data availability

This study was preregistered on OSF (Open Science Framework): https://doi.org/10.17605/OSF.IO/SZW6N.

Stimuli, results, analysis code, preregistration information, and a document detailing changes made to the preregistered analysis during the review process are available on OSF at the following link: https://osf.io/zve4c/.

Ethics and consent

This research was approved by the Research Ethics Board of the University of Toronto (Protocol Number 36939).

Funding information

This work was supported by a grant to J.S. by the Natural Sciences and Engineering Research Council of Canada.

Acknowledgements

Thanks to Jiaying Li, Mariam Galytskyy, Grace Meany, Monika Krizic, Lisa Sullivan, and Michelle Sun for help with experiment implementation and analysis, and to Hanna Zhang for feedback on an earlier draft.

Competing interests

The authors have no competing interests to declare.

References

Adank, P., & **Janse, E.** (2010). Comprehension of a novel accent by young and older listeners. *Psychology and Aging*, *25*(3), 736–740. DOI: https://doi.org/10.1037/a0020054

Baese-Berk, M. (2018). Perceptual learning for native and non-native speech. *Psychology of Learning and Motivation*, *68*, 1–29. DOI: https://doi.org/10.1016/bs.plm.2018.08.001

Bates, D., Maechler, M., Bolker, B. M., & **Walker, S. C.** (2015). Lme4: Linear mixed-effects models using eigen and S4. https://cran.r-project.org/web/packages/lme4/index.html

Benkí, J. R. (2001). Place of articulation and first formant transition pattern both affect perception of voicing in English. *Journal of Phonetics*, *29*(1), 1–22. DOI: https://doi.org/10.1006/jpho.2000.0128

Bradlow, A. R., & **Bent, T.** (2008). Perceptual adaptation to non-native speech. *Cognition*, *106*(2), 707–729. DOI: https://doi.org/10.1016/j.cognition.2007.04.005

Bradlow, A. R., Nygaard, L. C., & **Pisoni, D. B.** (1999). Effects of talker, rate, and amplitude variation on recognition memory for spoken words. *Perception & Psychophysics*, *61*(2), 206–219. DOI: https://doi.org/10.3758/bf03206883

Cho, T., & **Ladefoged, P.** (1999). Variation and universals in VOT: Evidence from 18 languages. *Journal of Phonetics*, *27*(2), 207–229. DOI: https://doi.org/10.1006/jpho.1999.0094

Clopper, C. G., & **Dossey, E.** (2020). Phonetic convergence to Southern American English: Acoustics and perception. *The Journal of the Acoustical Society of America*, *147*(1), 671. DOI: https://doi.org/10.1121/10.0000555

Clopper, C. G., & **Pisoni, D. B.** (2004). Effects of talker variability on perceptual learning of dialects. *Language and Speech*, *47*(3), 207–239. DOI: https://doi.org/10.1177/00238309040470030101

Cohen Priva, U., & **Sanker, C.** (2020). Natural leaders: Some interlocutors elicit greater convergence across conversations and across characteristics. *Cognitive Science*, *44*(10), e12897. DOI: https://doi.org/10.1111/cogs.12897

Coles-Harris, E. H. (2017). Perspectives on the motivations for phonetic convergence. *Language and Linguistics Compass*, *11*(12), e12268. DOI: https://doi.org/10.1111/lnc3.12268

Coumel, M., Christiner, M., & **Reiterer, S. M.** (2019). Second language accent faking ability depends on musical abilities, not on working memory. *Frontiers in Psychology*, *10*, 257. DOI: https://doi.org/10.3389/fpsyg.2019.00257

D'Imperio, M., Cavone, R., & **Petrone, C.** (2014). Phonetic and phonological imitation of intonation in two varieties of Italian. *Frontiers in Psychology*, *5*. DOI: https://doi.org/10.3389/fpsyg.2014.01226

De Rosario-Martinez, H. R. (2015). Package 'Phia,' https://CRAN.R-project.org/package=phia

Dufour, S., & **Nguyen, N.** (2013). How much imitation is there in a shadowing task? *Frontiers in Psychology*, *4*, 346. DOI: https://doi.org/10.3389/fpsyg.2013.00346

Evans, J. W. (1996). Straightforward statistics for the behavioral sciences. Pacific Grove, CA: Brooks/Cole Publishing.

Flege, J. E., & **Eefting, W.** (1987). Production and perception of English stops by native Spanish speakers. *Journal of Phonetics*, *15*, 67–83. DOI: https://doi.org/10.1016/S0095-4470(19)30538-8

Flege, J. E., & **Hammond, R. M.** (1982). Mimicry of non-distinctive phonetic differences between language varieties. *Studies in Second Language Acquisition*, *5*(1), 1–17. DOI: https://doi.org/10.1017/s0272263100004563

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–279. DOI: https://doi.org/10.1037/0033-295x.105.2.251

Incera, S., & **McLennan, C. T.** (2018). Bilingualism and age are continuous variables that influence executive function. *Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition*, 25(3), 443–463. DOI: https://doi.org/10.1080/13825585.2017.1319902

Kim, D., & **Clayards, M.** (2019). Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*, *34*(6), 769–786. DOI: https://doi.org/10.1080/23273798.2019.1582787

Kuhl, P. K., & **Miller, J. D.** (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *The Journal of the Acoustical Society of America*, *63*(3), 905–917. DOI: https://doi.org/10.1121/1.381770

Kuznetsova, A., Brockhoff, P. B., & **Christensen, R. H. B.** (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). DOI: https://doi.org/10.18637/jss.v082.i13

Maye, J., Aslin, R. N., & **Tanenhaus, M. K.** (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. In *Cognitive Science* (Vol. 32, Issue 3, pp. 543–562). DOI: https://doi.org/10.1080/03640210802035357

McLennan, C. T. (2006). The time course of variability effects in the perception of spoken language: Changes across the lifespan. *Language and Speech*, 49, 113–125. DOI: https://doi.org/10.1177/00238309060490010701

Mora, J. C., Rochdi, Y., & **Kivistö-de Souza, H.** (2014). Mimicking accented speech as L2 phonological awareness. *Language Awareness*, *23*(1–2), 57–75. DOI: https://doi.org/10.1080/09658416.2013.863898

Moulines, E., & **Charpentier, F.** (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453–467. DOI: https://doi.org/10.1016/0167-6393(90)90021-Z

Mullennix, J. W., & **Pisoni, D. B.** (1990). Stimulus variability and processing dependencies in speech perception. *Perception & Psychophysics*, *47*(4), 379–390. DOI: https://doi.org/10.3758/bf03210878

Mullennix, J. W., Pisoni, D. B., & **Martin, C. S.** (1989). Some effects of talker variability on spoken word recognition. *The Journal of the Acoustical Society of America*, *85*(1), 365–378. DOI: https://doi.org/10.1121/1.397688

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, *39*(2), 132–142. DOI: https://doi.org/10.1016/j.wocn.2010.12.007

Nielsen, K., & **Scarborough, R.** (2015). Perceptual asymmetries between greater and lesser vowel nasality and VOT. *Proceedings of ICPhS*. Glasgow.

Olmstead, A. J., Viswanathan, N., Aivar, M. P., & **Manuel, S.** (2013). Comparison of native and non-native phone imitation by English and Spanish speakers. *Frontiers in Psychology*, *4*, 475. DOI: https://doi.org/10.3389/fpsyg.2013.00475

Pardo, J. S., Jay, I. C., & **Krauss, R. M.** (2010). Conversational role influences speech imitation. *Attention, Perception & Psychophysics*, *72*(8), 2254–2264. DOI: https://doi.org/10.3758/BF03196699

Perrachione, T. K., Lee, J., Ha, L. Y. Y., & **Wong, P. C. M.** (2011). Learning a novel phonological contrast depends on interactions between individual differences and training paradigm design. *The Journal of the Acoustical Society of America*, *130*(1), 461–472. DOI: https://doi.org/10.1121/1.3593366

Pickering, M. J., & **Garrod, S.** (2013). An integrated theory of language production and comprehension. *The Behavioral and Brain Sciences*, *36*(4), 329–347. DOI: https://doi.org/10.1017/S0140525X12001495

R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org

Reiterer, S. M., Hu, X., Erb, M., Rota, G., Nardo, D., Grodd, W., Winkler, S., & **Ackermann, H.** (2011). Individual differences in audio-vocal speech imitation aptitude in late bilinguals: Functional neuro-imaging and brain morphology. *Frontiers in Psychology*, *2*(271). DOI: https://doi.org/10.3389/fpsyg.2011.00271

Reiterer, S. M., Hu, X., Sumathi, T. A., & **Singh, N. C.** (2013). Are you a good mimic? Neuro-acoustic signatures for speech imitation ability. *Frontiers in Psychology*, *4*(782). DOI: https://doi.org/10.3389/fpsyg.2013.00782

Sancier, M. L., & **Fowler, C. A.** (1997). Gestural drift in a bilingual speaker of Brazilian Portuguese and English. *Journal of Phonetics*, *25*(4), 421–436. DOI: https://doi.org/10.1006/jpho.1997.0051

Sato, M., Grabski, K., Garnier, M., Granjon, L., Schwartz, J. L., & **Nguyen, N.** (2013). Converging toward a common speech code: Imitative and perceptuo-motor recalibration processes in speech production. *Frontiers in Psychology*, *4*, 422. DOI: https://doi.org/10.3389/fpsyg.2013.00422

Schmale, R., Cristia, A., & **Seidl, A.** (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science*, *15*(6), 732–738. DOI: https://doi.org/10.1111/j.1467-7687.2012.01175.x

Shockley, K., Sabadini, L., & **Fowler, C. A.** (2004). Imitation in shadowing words. *Perception & Psychophysics*, *66*(3), 422–429. DOI: https://doi.org/10.3758/BF03194890

Sommers, M. S., & **Barcroft, J.** (2006). Stimulus variability and the phonetic relevance hypothesis: Effects of variability in speaking style, fundamental frequency, and speaking rate on spoken word identification. *The Journal of the Acoustical Society of America*, *119*(4), 2406–2416. DOI: https://doi.org/10.1121/1.2171836

Sommers, M. S., Nygaard, L. C., & **Pisoni, D. B.** (1994). Stimulus variability and spoken word recognition. I. Effects of variability in speaking rate and overall amplitude. *The Journal of the Acoustical Society of America, 96*(3), 1314–1324. DOI: https://doi.org/10.1121/1.411453

Spinu, L. E., Hwang, J., & **Lohmann, R.** (2018). Is there a bilingual advantage in phonetic and phonological acquisition? The initial learning of word-final coronal stop realization in a novel accent of English. *International Journal of Bilingualism*, *22*(3), 350–370. DOI: https://doi.org/10.1177/1367006916681080

Spinu, L., Hwang, J., Pincus, N., & **Vasilita, M.** (2020). Exploring the use of an artificial accent of English to assess phonetic learning in monolingual and bilingual speakers. In *Proceedings of Interspeech 2020*. DOI: https://doi.org/10.21437/interspeech.2020-2783

Wade, L. (2022). Experimental evidence for expectation-driven linguistic convergence. *Language, 98*(1). DOI: https://doi.org/10.1353/lan.0.0257

Wade, L., Lai, W., & **Tamminga, M.** (2020). The reliability of individual differences in VOT imitation. *Language and Speech*, 23830920947769. DOI: https://doi.org/10.1177/0023830920947769

Yu, A. C. L., Abrego-Collier, C., & **Sonderegger, M.** (2013). Phonetic imitation from an individual-difference perspective: Subjective attitude, personality and "Autistic" Traits. *PloS One*, *8*(9), e74746. DOI: https://doi.org/10.1371/journal.pone.0074746

Zellou, G., & **Brotherton, C.** (2021). Phonetic imitation of multidimensional acoustic variation of the nasal split short-a system. *Speech Communication*, *135*, 54–65. DOI: https://doi.org/10.1016/j.specom.2021.10.005