# Phonetic Second Language Perception and Production

Noa Attali[1]

[1] Rutgers, The State University of New Jersey, New Brunswick

Primary Advisor: Karin Stromswold, Professor, Psychology and Cognitive Science

Author Note

Correspondence concerning this article should be addressed to Noa Attali, . E-mail: noa.attali@gmail.com

# Abstract

This study focuses on tying measures of successful communication to acoustic features of second language (L2) speech, as a step towards explaining associations between successes of perception and production in a second language. We examine how L2 English speakers produce and perceive voiced and voiceless English plosives. Three inter-related questions are addressed. 1) How - on an acoustic-phonetic level - do L2 speakers produce phonemes that differ between their L1 and L2? 2) What acoustic-phonetic features do L1 speakers use to understand phonemes said by L2 speakers? 3) What is the relationship between L2 phoneme perception and production? We fit linear mixed models to test how voicing, speaker age, and acoustic features predict accuracy and reaction time of L2 production and L2 perception. We find that the L2 speech sound system reflects the development of hybrid phonotactic constraints on phonetic categories that are similar between a speaker's L1 and L2. We further find that phonetic perception precedes phonetic production.

*Keywords:* second language speech, phonology, perception, production, speaker age, Speech Learning Model

# Phonetic Second Language Perception and Production

# Contents

## List of Figures

## List of Tables

## Introduction

How do second language (L2) learners produce and perceive phonemes in their second language, and in what way are production and perception capabilities related? Perception and production may be considered processes with differing neurophysiological properties, expressions, constraints, and motivations, but they may also be considered two sides of the same functional coin, communication, within a speaker's general spoken language system. Here we focus on one aspect of communication, ease or success of information transfer as measured by accuracy and reaction time of understanding. Biases towards successful information transfer are assumed to at least partly affect both production and perception through generalization across speakers' respective previous experiences with successes and failures of understanding (Jaeger, 2013). Furthermore, we analyze L2 perception and production modalities, and their interconnection for a speaker, in accordance with Flege's Speech Learning Model (SLM) for the acquisition of L2 speech sounds (Flege, 1995). SLM suggests that a central task in L2 learning is to maintain contrasts between first language (L1) and L2 phonetic categories in a common phonological space; moreover, that the ability to successfully produce L2 phonetic contrasts is constrained by the ability to successfully perceive them. Overall, in seeking associations between the functional success of speakers' productions and perceptions and the acoustic properties of their speech, we hope to build a fine-grained picture of the properties and functions of a speaker's phonological system.

We determine functional success as *understandability* on the basis of listener perceptions. Presumably, for a single production of a speaker in a particular context, we can find patterns and consistencies across the responses of multiple listeners, which reflect on characteristics of that production. Previous studies focusing on the perception of L2 speech according to L1 listener judgments have analyzed listener understanding along several dimensions. Following the terminology of Munro and Derwing (1995a), the first is accuracy of understanding, referred to as *intelligibility* and measured by whether the listeners' indication of what they heard matches with what the speaker intended to say. The second is ease of understanding or *comprehensibility*, which could also be understood as *processing difficulty* since comprehensibility scores tend to correlate with reaction time data (Munro & Derwing, 1995b). A third measure, *accentedness*, corresponds to the degree to which listeners judge utterances to conform to expected native-like patterns. The three dimensions are largely independent (Munro & Derwing, 1995a), though comprehensibility and accentedness may be highly correlated for some listeners (Munro & Derwing, 1995b). The current study, in order to avoid falsely indicating to listeners that their task was to judge the subjective correctness of the speech they heard, did not ask listeners for overt judgments regarding the quality of stimuli. Instead we measured accuracy and reaction time of understanding in a forced-choice task as intelligibility and comprehensibility, respectively. For clarity, in describing our study, we will refer directly to the measures themselves and not to the dimension of understanding to which they theoretically refer. Please see Figure 1 for an outline of terminology.

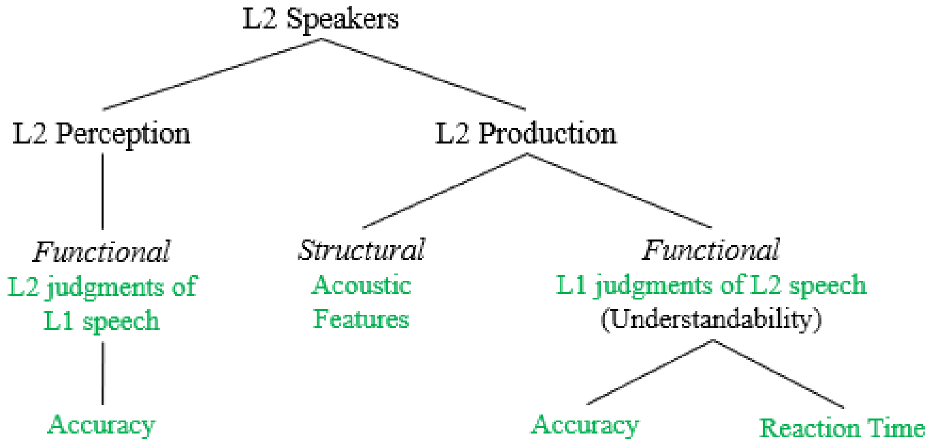


*Figure 1*. Terminology

In the following sections, we review the literature on L2 production. We then review SLM as it relates to the current study. Study 1 considers connections between the structural and functional properties of L2 production (i.e., when an L2 speaker's speech is highly understandable, what is the L2 speaker doing acoustically?). Study 2 considers connections between the structural properties of L2 production and the functional properties of L2 perception (i.e., when an L2 speaker possesses high understanding of his/her L2, what does the speaker do acoustically?). Finally in Study 3 we summarize the relationship between speech production and perception in an L2. When is L2 perception more successful than L2 production? We expect that the better an L2 speaker can understand an L1 speaker's speech, the better her speech is understandable to an L1 speaker, but that an L2 speaker's perception can be more but not less successful than his/her production.

## Functional L2 Speech Production

Previous studies on the accuracy of L1 speaker perceptions of L2 speech have focused on the effects of impressionistically measured contextual and listener characteristics. Prejudice by L1 speakers against foreign accents, sometimes formulated as irritation due to speech errors or evaluational bias against dissimilar others (e.g. Lambert, 1967), may affect accuracy of understanding (Albrechtsen, Henriksen, & Faerch, 1980; Anderson-Hsieh & Koehler, 1988; Fayer & Krasinski, 1987), rendering accents "costly" in communication (Munro & Derwing, 1995b). Conversely, greater native listener exposure to, or experience or familiarity with, the speakers' first language improves listener accuracy though not ease of understanding (Gass & Varonis, 1984; Kennedy & Trofimovich, 2008). Very fast or very slow speaking rate adversely affects accuracy, especially when accentedness is also judged to be high (Anderson-Hsieh & Koehler, 1988; Munro & Derwing, 2001). Furthermore, an interlanguage speech intelligibility benefit has been documented (Bent & Bradlow, 2003, Hayes-Harb, Smith, Bent, and Bradlow (2008)), such that for native English listeners, native English speakers are more accurately understood than nonnative speakers, while for nonnative English listeners, nonnative English speech is at least as well understood.

However, there tends to be a high degree of similarity in responses regarding not only accuracy but also ease and degree of accentedness from listeners of diverse backgrounds to the same speech stimuli (Munro, Derwing, & Morton, 2006). This suggests that features specific either to the speaker or to the speech overwhelmingly affect information transfer. In particular, ease of understanding and degree of accentedness seem, in comparison with accuracy, to be closely related to phonology and phonological articulation (Flege, 1988), although it is important to note, again, that predictor measures tend to be impressionistic. Deviance from nativelike phonemes and syllable structure, but most robustly from nativelike prosody, affect judgments of phonological error (Anderson-Hsieh, Johnson, & Koehler, 1992). Additionally, the effect of speaker on the ease of a listener's speech processing may be heightened when speakers have a foreign accent; McLennan and González (2012) found that talker-specific effects on speech perception by a given listener are greater for L2 speakers compared to L1 speakers.

**Effects of Speaker Age on L2 Production.** Speaker age is one of the most important factors in the understandability of L2 speech, as a sensitive period for the acquisition of a nonnative phonology has been documented to limit the degree to which older learners of a second language master its phonological system (Oyama, 1976). As reviewed by J. S. Johnson and Newport (1989), the literature suggests that in studies of immigrants, age of arrival to a host country is the only predictor of L2 proficiency; and that late learners have an initial and short-lived advantage over early learners which is reversed when measured by ultimate attainment. Age affects degree of accentedness, as adults who learned their L2 in childhood receive better accentedness scores than adults who learned their L2 in adulthood (Flege, 1988). However, there is debate about the extent to which the amount of experience with L2, beyond an initial rapid learning phase, affects pronunciation.

One view holds that adult L2 pronunciation fossilizes early (see Selinker, 1972). Within the debate over the role of maturational and nonmaturational factors affecting L2 acquisition, the strong formulation of the Sensitive Period Hypothesis is the Critical Period

Hypothesis (CPH; for a review, see Pallier, 2007). In the strict tradition, Lenneberg (1967) suggests there is a limited time period between birth and puberty in which language can be acquired and beyond which language acquisition and competency declines. Lenneberg (1967) utilized indirect behavioral evidence to propose that, beyond puberty, the brain irreversibly loses neural plasticity necessary for language acquisition, or that a speaker's L1 fixes the functional neural connections in the cortex (Penfield, 1965).

Alternatively, amount and goodness of input matters, and it is exactly age-related differences in the amount of L2 input that determine age as a factor in proficiency (Flege et al., 2006), especially when the population of L2 speakers of interest are immigrants to the country where that L2 is a native language. Generally standing against neurological maturation accounts are accounts of L2 articulatory errors as due to failures in L2 perception (for a review, see Flege, 1995). An L2 speaker's years of learning their L2 involves learning what sounds to pay attention to and in what way, which creates a perceptual "grid", "sieve", or "filter" through which the L2 speaker also perceives L2 sounds (see Flege, 2003 for a review). Specifically, Flege's Speech Learning Model (SLM, discussed in more detail below) suggests that perceptual targets are necessary to guide sensorimotor learning of L2 sounds (Flege, 1995). Accurate perceptual targets in turn depend on sufficient phonetic input of a native language for an L2 learner of that language (Flege, 1991). Accounts that focus on processing rather than neurology as a factor in "ultimate attainment" may be considered input hypotheses.

SML, which locates perception module as the locus of success and failure of understanding in an L2, is supported by evidence that amount of experience with or training in L2 affects the ability to perceive L2 phones more than the ability to produce those phones (see Sakai & Moorman, 2018 for a review). Perception "outstrips" production, perhaps because internal auditory-perceptual representations of phones are more flexible, or subjective to a longer sensitive period, than motor speech abilities (Flege, 1988). However, the Motor Theory of speech perception (MT) is a third account for constraints on the acquisition and development of L2 production and perception which directly contradicts SML (for a review, please see Galantucci, Fowler, & Turvey, 2006; Liberman & Mattingly, 1985; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). MT locates the motor system as the determinant of both successful L2 perception and production. Phonetic categories are stored as abstract articulatory programs, and these articulatory programs form the basis of the perception of phonetic categories. Thus, given the terms of the argument between SML and MT, greater attention should be paid to the speaker's phonetics system underlying both perception and production in L2.

**Effects of Acoustic Features on L2 Production.** In this study, we focus on acoustic features, rather than suprasegmental features like prosody, as possible phonological features that affect information transfer between speaker and listener. Cross-linguistically, specific acoustic features can crucially affect what phoneme native speakers of a language perceive a phone to be. But it is not clear how features of L1 and L2 enter into L2 productions. L1 phonological features could transfer, interfere with, or influence L2 pronunciation in the L2 learning process (Chan & Li, 2000; see Flege, 1988). Nor is it clear what features native speakers of one language rely on to parse nonnative speech. It is possible that native speakers rely on their experience with the features of their language when listening to a speaker identified as having a foreign accent, and therefore better understand nonnative speakers who use of nativelike acoustic features. Alternatively, native speakers may ignore or update their knowledge of nativelike phonological features in contexts where they are unlikely to consistently encounter them. There is evidence that listeners rapidly adapt to accents in terms of both their processing accuracy and speed (Bradlow & Bent, 2003; Clarke & Garrett, 2004), as part of a more general ability to adapt to talker-specific and across-talker variability during speech processing (Kleinschmidt & Jaeger, 2015; see also Pajak, Fine, Kleinschmidt, & Jaeger, 2016 on the idea of continual revision in the context of both a learner's L1 and L2). However, most studies on adaptation focus on adaptation to sentences, in which semantic and syntactic context may provide powerful guidance for comprehension which would not be present when listeners only hear words out of context.

## Speech Learning Model

In this project, we operationalize hypotheses about the effects of acoustic features on successful L2 perception and production through the Speech Learning Model (SLM), which characterizes L2 learning in terms of position-sensitive phonetic categories (Flege, 1995). Typically the term phoneme is used to refer to an abstract category that can be used to distinguish between words in a language. For example, /p/ and /b/ are phonemes in English because they are the distinguishing categorical difference in the phonological minimal pair *pill/bill*. However, "SLM focuses on the learning of *position-sensitive* phonetic categories, not phonemes, because the learning of L2 speech sounds may vary according to position" (Flege, 2018, p. 3). Phonetic categories according to Flege refer to long-term memory representations that specify the position-sensitive aspects or features of the phonemes relevant to the sound system of a language. We will use the term *phone* to refer to a speech sound without specifying whether it is phonemic or phonetic in the language. L2 learners' perception of the features of an L2 phoneme create mental representations that, in turn, guide production. There is thus a unidirectional relation between perception and production, such that L2 speakers are not expected to be able to produce a phone they cannot perceive. Moreover, segmental perceptual ability is expected to be constant over a speaker's lifetime. Amount and goodness of input in the L2, rather than neurophysical or biological constraints on perceptual and sensorimotor abilities, determine the proficiency attained by L2 learners.

SLM posits that the development of phonetic categories in L2 learning involves the application of the same mechanisms and processes necessary in L1 learning. Crucially, in learning an L2, a speaker maintains L1 and L2 phonetic categories in a common phonological space: phonetic categories depend on the properties of all (both L1 and L2) phones that are determined to be the realization of that category. Therefore L2 speakers are proficient to the extent that they can establish and maintain contrast between L1 and L2 categories. That is to say, L2 learning involves the decision to assimilate L2 sounds to L1 phonetic categories, to amend phonetic categories according to L2 input, or to form novel categories for L2 phones. The decision depends on perceived differences between L1 and L2 phones. A new category can be established for an L2 phone if L2 learners perceive enough difference between that L2 phone and its closest L1 phone. Thus Flege (1988) has pointed out the issue of "equivalent classification limits" in a phonetic system: phones in L2 that only differ physically from a counterpart in L1 may be regarded by the learner as being the realization of a phonetic category that has already been established, and may therefore not be learned as a new category. These perceptually linked L1 and L2 phones (diaphones) will come to be produced similarly to each other. On the other hand, phones in L2 that do not have a direct L1 counterpart could be produced in a native-like accent by any learner. But the novel category established for these L2 phones may still not resemble a native speaker's due either to the L2 learner's need to maintain contrast between L1 and L2 (e.g., by exaggerating features of the L2 category) or to the L2 learner using different features to represent the phone.

The current project investigates production and perception of English by people whose L1 is Cantonese and L2 is English. English and Cantonese are typologically distant; the former is a Germanic language and the latter belongs to the Sino-Tibetan language family (Chan & Li, 2000). The phonemes under consideration are the six plosive consonants: /p, b, d, t, k, g/. The plosives constitute a class of phones that, at least as a subset, exist in every natural spoken language. The bilabial plosives are commonly /p, b/. The alveolar plosives are /t, d/. The velar plosives are /g/. In both Cantonese and English, the plosives are phonemes in that they distinguish between words. For example, in order to successfully communicate, English speakers must be able to perceive and produce a categorical difference between the first phone in each of the following words - *pill*/*bill*, *dill*/*till*, *kill*/*gill*. However, the acoustic-phonetic features of phonemes differ within and between the languages. We will use slashes to refer to abstract phonemic categories (e.g., the category /p/) and brackets to refer to the acoustic-phonetic characteristics of the realization of phonemes (e.g., the unaspirated voiceless bilabial [p]).

**Cantonese and English Plosives.** In English, the set of plosive phonemes often have a voicing contrast. Voicing is determined physiologically by vocal cord vibration during the closure of the stop and results acoustically in a periodic waveform. The voiced plosives are /b, d, g/ (e.g., *bill*, *dill*, *gill*) and the voiceless plosives are /p, t, k/ (e.g., *pill*, *till*, *kill*). /p/ is realized as unaspirated voiceless [p], /b/ is realized as unaspirated voiced [b], and so on. English orthography seems to regularize the voicing distinction.

However, the realization of plosive contrasts is irregular and, moreover, the contrast between a native English speakers' perception of a voiced or unvoiced plosive may depend on factors other than physiological voicing during the closure of the stop (Lisker & Abramson, 1964). In utterance-initial and word-initial positions the distinction may be realized by aspiration rather than voicing. Aspiration is determined as a burst of air emitted between the plosive onset release and the vowel onset. In such a case, all the plosives are voiceless, and the /b, d, g/ set are realized as unaspirated voiceless [p, t, k] while the /p, t, k/ set are realized as aspirated voiceless [$p^h$, $t^h$, $k^h$]. That is to say, a native speaker may produce the initial consonant in *bill* as [p] rather than [b] and the initial consonant in *pill* as [$p^h$] rather than [p]. The result is that unaspirated voiceless [p, t, k] may be the phonetic realization of either of the plosive sets, depending on the context. Please see Figures 2 and 3 for example waveforms of voiced and aspirated voiceless bilabial, alveolar, and velar English plosives. The figures shows that voiced plosives, like vowels, have a periodic waveform, while aspiration is indicated by pronounced aperiodicity between the plosive release burst and the onset of the vowel.

The position of the plosive in the word has other implications for its realization. Probabilistic durational patterns condition the perception of voicing (Raphael, 1972). For

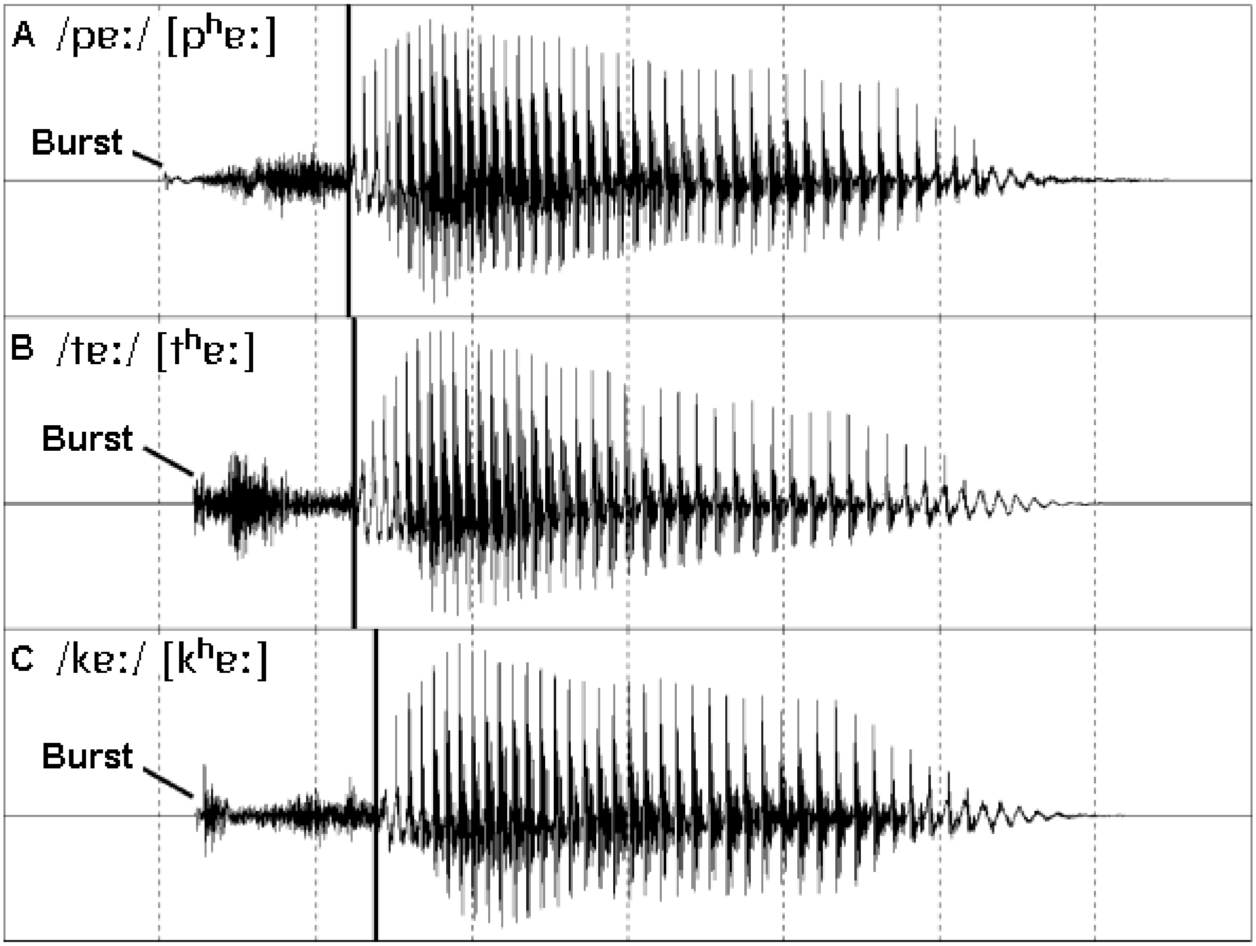


*Figure 2*. Example unvoiced plosive-vowel waveforms, adapted from Mannell (2008)

words with plosives in the syllable-final position (e.g. *cup* or *cub*, hereafter referred to as coda words), the vowel anticipating the voiced plosive (e.g. *cub*) tends to be greater in duration than the vowel anticipating the unvoiced plosive (e.g. *cup*). For words with plosives in the syllable-initial position (e.g. *back* or *pack*, hereafter referred to as onset words), the VOT of the unvoiced plosive (e.g. *pack*) tends to be longer than the VOT of the voiced plosive (e.g. *back*). Varying only the durations of these cues (e.g. of the vowel in *cub*) is sufficient to vary listeners' perception of the word (e.g. of perceiving either *cub* or *cup*). Furthermore, speakers tend to demonstrate phonetic adaptation to the immediate future, such that they are more likely to de-voice voiced coda words (e.g., pronouncing *cub* as *cup*) since codas are followed by an unvoiced period of time, while the opposite occurs

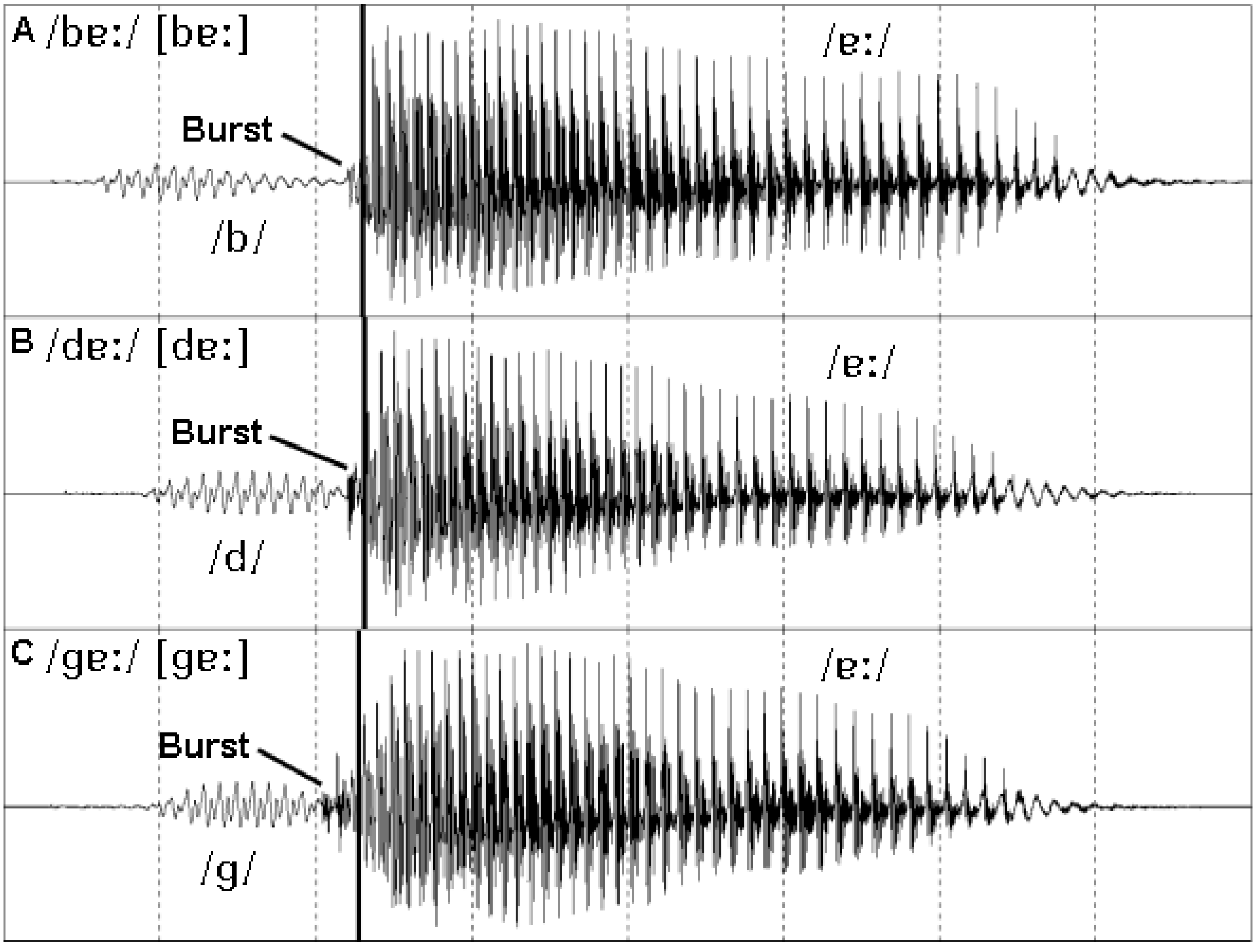


*Figure 3*. Example voiced plosive-vowel waveforms, adapted from Mannell (2008)

for unvoiced onsets that are followed by a voiced period of time. Please see Figure 4 for examples of VOT and vowel duration, the two most important durational cues to voicing in English.

Unlike English, Cantonese consistently uses aspiration to distinguish voiced from unvoiced plosives in onset position (Chan & Li, 2000). In other words, all Cantonese plosives are voiceless and the /b, d, g/ set are realized as unaspirated voiceless plosives [p, t, k] while the /p, t, k/ set are realized as aspirated voiceless plosives [$p^h$, $t^h$, $k^h$]. Orthography regularizes these distinctions such that [p, t, k] are represented as “b”, “d”, and “g”, respectively, and [$p^h$, $t^h$, $k^h$] are represented as “p”, “t”, and “k”, respectively. Cantonese speakers can therefore be reasonably expected to always produce unaspirated

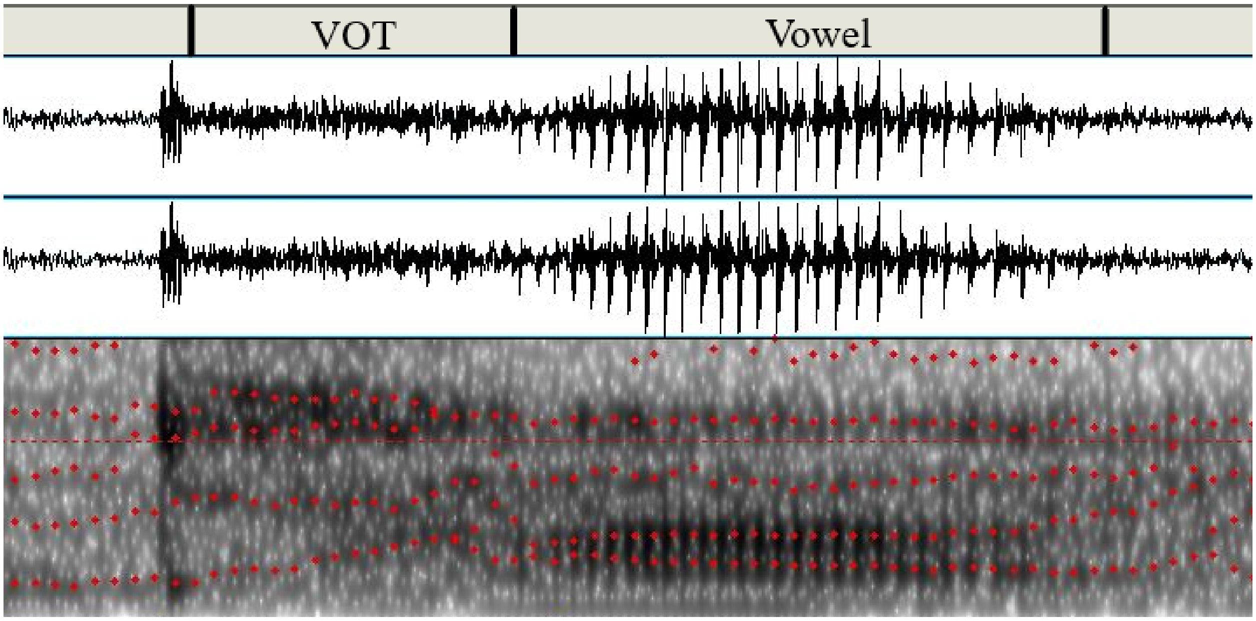
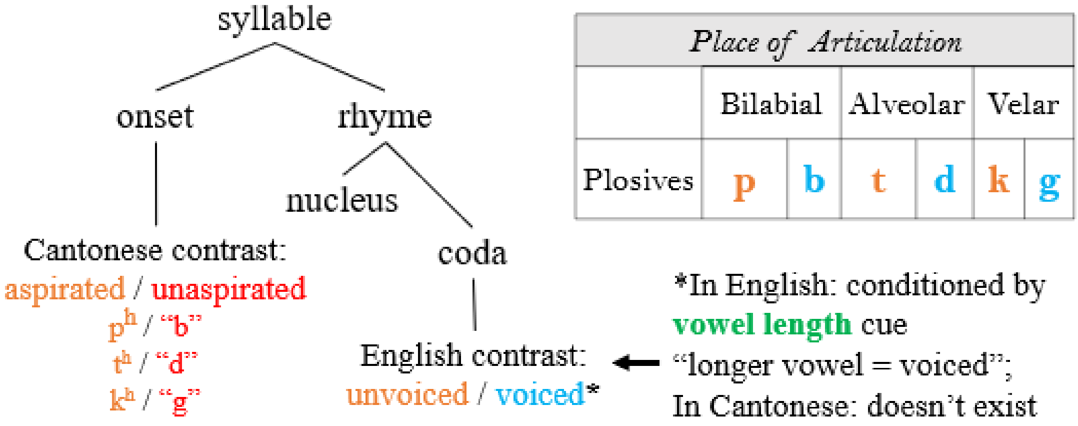


*Figure 4*. Example waveform marked for VOT and vowel duration

voiceless stops like [p] for an English /b/, though phonetically that /b/ may need to be realized as [b] especially in coda position. The evidence is that ESL Cantonese speakers do not generally use the voicing contrast. Additionally, ESL education may intensify preferment for the aspiration contrast because it emphasizes teaching each letter of the alphabet in its word-initial position (e.g., *boy* for /b/), in which native speakers of English can also use the aspiration contrast.

Numerous other differences regarding plosives in Cantonese and English may affect L1 understanding of L2 speech. According to Chan and Li (2000), while all English plosives may occur in the onset, nucleus, or coda of a word, in Cantonese coda position only permits /p, t, k/. Also unlike in English, in coda position plosives are more often unreleased and therefore unaspirated, which Chan and Li (2000) suggest neutralizes the contrast between aspirated and unaspirated plosives and may give the impression that Cantonese speakers swallow coda plosives in English. Please see Figure 5 for a visualization of a comparison between English and Cantonese phonetic categories for plosives.



| | *Place of Articulation* | | | | | |
|---|---|---|---|---|---|---|
| | Bilabial | | Alveolar | | Velar | |
| Plosives | p | b | t | d | k | g |

*Figure 5*. Position-dependent phonetic features of English and Cantonese plosives

**Hypotheses.** For the purposes of this project, we operationalize SLM hypotheses for L2 production and perception of English plosives as follows. Successful L2 perception and production will reflect the extent to which L2 speakers maintain a contrast between L1 and L2 phones. The following hypotheses are an attempt to build a full picture of the possibilities of L2 comprehension and production. However, the studies described in this thesis focus on a subset of these hypotheses, in particular on the role of speaker age and durational acoustic-phonetic features for L2 production and perception. Although the full set of hypotheses for L2 production includes consideration of the role of aspiration, the extent to which L2 speakers successfully maintain a contrast between L1 and L2 phonetic categories may reflect a *tradeoff* between consistent use of aspiration and consistent use of durational features. Therefore, the following studies will focus on the extent to which durational acoustic-phonetic features condition the perception of voicing for L2 speakers' perceptions and productions, in order to understand the extent to which both L1 and L2 features are used in L2 speech.

*Question 1: How do L1 Acoustic-Phonetic Features Influence the Perception and*

## *Production of L2 Phonemes?*

As discussed above, Flege has argued that acoustic-phonetic features of the phonemes in one's native language can influence the perception and production of L2 phonemes in a variety of ways. At one extreme, L2 speakers may use the acoustic-phonetic features of their L1 when they produce and perceive a second language (Hypothesis 1 below) and at the other extreme, L2 speakers may use acoustic-phonetic features that are extremely similar to those used by native speakers (Hypothesis 4 below). In this section, we outline 4 logically possible alternatives, spelling out specific empirical predictions of each alternative.

### *Hypothesis 1*: Equivalent Classification

According to the Equivalent Classification Hypothesis, L2 speakers use their L1 phonological systems to produce and perceive phonemes in an L2.

*L2 Comprehension*: When they listen to English, L1 Cantonese speakers will perceive aspirated English plosive onsets as voiceless and unaspirated English onsets as voiced, and they will perceive all English plosive codas as voiceless. E.g., for bilabials:

- In onset position:
  - L2 speakers will perceive [$\text{p}^{\text{h}}$] as /p/.
  - L2 speakers will perceive [p] and [b] as /b/.
- In coda position:
  - L2 speakers will perceive [$\text{p}^{\text{h}}$], [p], and [b] as /p/.

*L2 Production*: L1 Cantonese speakers will produce voiceless English plosive onsets as aspirated voiceless plosives and voiced English plosive onsets as unaspirated voiceless plosives. They will always produce English plosive codas as aspirated voiceless plosives. E.g.,

- In onset position:
  - L2 speakers will produce /p/ as [$p^h$].
  - L2 speakers will produce /b/ as [p].
- In coda position:
  - L2 speakers will produce /p/ and /b/ as [$p^h$].

### *Hypothesis 2*: Feature Adjustment

According to the Feature Adjustment Hypothesis, L2 speakers use their L1 phonological systems to produce and perceive phonemes in an L2, but they may change their use of their L1 systems, for example by generalizing its position-dependent phonetic features to apply to phones in other positions in a word.

*L2 Comprehension*: When they listen to English, L1 Cantonese speakers will perceive aspirated English plosive onsets as voiceless and unaspirated plosive onsets as voiced. They will generalize this rule to codas and perceive English aspirated codas as voiceless and English unaspirated codas as voiced. E.g.,

- In onset and coda position:
  - L2 speakers will perceive [$p^h$] as /p/.
  - L2 speakers will perceive [p] and [b] as /b/.

*L2 Production*: L2 speakers will produce voiceless onsets as aspirated voiceless plosives and voiced onsets as unaspirated voiceless plosives. L2 speakers will generalize the rule to codas and produce voiceless codas as aspirated voiceless plosives and voiced codas as unaspirated voiceless plosives. E.g.,

- In onset and coda position:

  - L2 speakers will produce /p/ as [$\text{p}^{\text{h}}$].
  - L2 speakers will produce /b/ as [p].

### *Hypothesis 3*: Hybrid Phonotactic Constraints

According to the Hybrid Phonotactic Constraints Hypothesis, L2 speakers use both their L1 and L2 phonological systems to produce and perceive phonemes in an L2. Phonetic features native to both their L1 and L2 systems inform the categorization of phones in L2 speakers' L2 production and perception. However, L1 features override L2 features as phonetic category features.

*L2 Comprehension*: When they listen to English, L1 Cantonese speakers will perceive aspirated plosive onsets as voiceless and unaspirated onsets as voiced. They will generalize this rule to English codas and will perceive aspirated English codas as voiceless and unaspirated codas as voiced. VOT and vowel duration will also influence their perception of phonemes: they will tend to perceive English plosive onsets that have long VOTs as being voiceless and those with short VOTs as being voiced. They will also tend to perceive English plosive codas with long vowel durations as being voiced and those with short vowel durations as being unvoiced. However, aspiration will take precedence over VOT or vowel duration as a category feature. E.g.,

- In onset position:
  - L2 speakers will always perceive [$\text{p}^{\text{h}}$] as /p/.
  - L2 speakers will perceive unaspirated plosives with short VOTs as /b/.
  - L2 speakers will perceive unaspirated plosives with long VOTs as /p/.
- In coda position:
  - L2 speakers will always perceive [$\text{p}^{\text{h}}$] as /p/.
  - L2 speakers will perceive unaspirated plosives with long vowel durations as /b/.

  - L2 speakers will perceive unaspirated plosives with short vowel durations as /p/.

*L2 Production*: When they say English words, L1 Cantonese speakers will produce voiceless English plosive onsets as aspirated voiceless plosives and voiced English plosive onsets as unaspirated voiceless plosives. They will generalize this rule to English codas, and will produce voiceless English codas as aspirated voiceless plosives and voiced English codas as unaspirated voiceless plosives. They will also tend to produce voiced English plosive onsets with shorter VOTs than unvoiced English plosive onsets, and they will tend to produce voiced English plosive codas with longer vowel durations than unvoiced English plosive codas. E.g.,

- In onset position:
  - L2 speakers will produce /p/ as [p$^h$] with longer VOT than for /b/.
  - L2 speakers will produce /b/ as [p] with shorter VOT than for /p/.
- In coda position:
  - L2 speakers will produce /p/ as [p$^h$] with shorter vowels than for /b/.
  - L2 speakers will produce /b/ as [p] with longer vowels than for /p/.

### *Hypothesis 4*: Category Formation

According to the Category Formation Hypothesis, L2 speakers do not use their L1 phonological systems to produce and perceive phonemes in an L2. Instead, L2 speakers develop mental representations of phonetic categories that closely resemble those of native speakers of the L2.

*L2 Comprehension*: When they listen to English, L1 Cantonese speakers will use the same acoustic features to perceive English plosives as native English speakers. They will perceive English plosive onsets that have long VOTs as being voiceless and those with short

VOTs as being voiced. They will perceive English plosive codas with long vowel durations as being voiced, and those with short vowel durations as being unvoiced. Aspiration will play a secondary role to VOT and vowel duration: aspirated English onsets will always be perceived as being voiceless, and perception of unaspirated onsets will depend on the phone's VOT. The rare instances of aspirated codas will always be perceived as voiceless. Unaspirated codas will be perceived as voiced or voiceless depending on the duration of the preceding vowel. E.g.,

- In onset position:
  - L2 speakers will always perceive aspirated plosives as /p/.
  - Unaspirated plosives with long VOT will be perceived as /p/.
  - Unaspirated plosives with short VOT will be perceived as /b/.
- In coda position:
  - L2 speakers will always perceive aspirated plosives as /p/.
  - Unaspirated plosives with long vowel duration will be perceived as /b/.
  - Unaspirated plosives with short vowel duration will be perceived as /p/.

*L2 Production*: When they say English words, L1 Cantonese speakers will do so in a manner similar to that of native English speakers. Voiceless plosive onsets will have longer VOTs than voiced plosive onsets, though the category boundary may not be the same or as sharp as for native English speakers. Voiceless onsets will sometimes be aspirated and sometimes not, but voiced onsets will never be aspirated. Voiced plosive codas will have longer vowel durations than unvoiced codas, and neither voiced or unvoiced codas will be aspirated. E.g.,

- In onset position:
  - L2 speakers will always produce /p/ with longer VOT than for /b/.

  - L2 speakers will always produce /b/ with shorter VOT than for /p/.
  - L2 speakers will usually aspirate /p/.
  - L2 speakers will never aspirate /b/.

- In coda position:
  - L2 speakers will always produce /p/ with shorter vowel duration than for /b/.
  - L2 speakers will always produce /b/ with longer vowel duration than for /p/.
  - L2 speakers will never aspirate /b/ or /p/.

### *Question 2: What is the Role of Speaker Age on the Production and Perception of L2 Phonemes?*

As discussed above, one hotly debated issue in the second language acquisition literature is the role of input and age on the production and perception of L2 phonology.

### *Critical Period Hypothesis*

The Critical Period Hypothesis (CPH) posits that L2 phonologies fossilize at an early age. If CPH is correct, we would predict that adult L2 speakers will not perceive or produce L2 phonemes in a more "native-like" way than adolescent L2 speakers who learned their L2 at the same age.

### *Speech Learning Model: Input Hypothesis*

Input hypotheses, in particular the Speech Learning Model (SLM), posit that L2 phonologies are adaptive over an L2 learner's lifespan. If the SLM hypothesis is correct, we would predict that adult L2 speakers will perceive and produce L2 phonemes in a more native-like way than adolescent L2 speakers because adults would have had more exposure to L2 phonology than adolescents.

### *Question 3: What is the Relationship between Production and Perception of L2 Phonemes?*

As discussed above, another debated issue in the second language acquisition literature is the extent to which L2 phonetic production and perception determine each other, and why.

#### *SLM Hypothesis: Perception Underlies Production*

According to SLM, perception creates mental representations that guide production. It is not possible to produce a phonetic contrast that one cannot perceive. Therefore SLM predicts that perceptual accuracy will generally outstrip production accuracy, and production accuracy will never outstrip perceptual accuracy.

#### *Motor Theory (MT) Hypothesis: Production Underlies Perception*

Acccording to MT, phonetic categories are stored as abstract articulatory programs which guide perception of phonetic categories. In essence, perception of phonetic categories is parasitic on the production of phonetic categories. Therefore, MT predicts that perceptual accuracy will never outstrip production accuracy.

## **Study 1: L2 Production**

Study 1 focuses on functional and structural measures of L2 production. We look for the acoustic-phonetic and speaker features tied to the accuracy and reaction time of L2 listener perceptions of L1 productions differing in the voicing of the plosive in onset or coda position. Speakers were adults and adolescents whose L1 was Hong Kong Cantonese and L2 was English, recorded in a study by Terry Au (ms). Listeners were L1 English speakers.

## Methods

**Participants.** Speakers were seventy-eight speakers whose L1 was Hong Kong Cantonese and L2 was English (henceforth referred to as L2 participants). Forty-two of the L2 participants were adolescents (ages 11-13 years, 36% male) and thirty-six were adults (ages 18-22 years, 31% male). L2 participants were recruited at Hong Kong University as part of a training program to help participants perceive and produce English words that include plosives (Au, ms). All participants had been learning English from L2 English speakers in a classroom since ages 5 or 6. Listeners were ninety-six native English-speaking college students (ages 18-22 years, 36% male) recruited at Rutgers University (henceforth referred to as L1 participants). L1 participants were compensated $10 or received course credit.

**Stimuli.** The stimuli words were plosive onset and coda minimal pair words originally recorded by the L2 speakers as part of a pretest for the training study (see Appendix). L2 speakers read aloud the words embedded in "I say . . . " phrases such as "I say cab". Each of the 78 L2 participants said 24 words (18 coda and 6 onset), for a total of 1,872 stimuli words. Please see the Appendix for the words.

**Experimental Procedure.** Words said by L2 participants that did not have at least 2 of 3 target phonemes were removed from further analysis. For example, if an L2 participant said *crat* or *ca* instead of *cat*, the word was kept in the pool, but if an L2 participant said *cram* the word was removed from further analysis. The remaining set of 1,672 words said by L2 participants was divided into 8 groups with each of the 8 groups having an equal number of voiced and unvoiced onset and coda words. L1 listeners listened to and judged a total of 411 onset words (206 said by L2 adolescents, and 205 said by L2 adults) and 1251 coda words (610 said by L2 adolescents, 641 said by L2 adults). L1 participants listened to L2 productions presented over Sennheiser HD 202 headphones and chose which of two written words on a computer screen matched the word they heard.

Each L1 participant judged one-eighth of the L2 productions. E-prime version 2.0 software presented the stimuli trials and recorded L1 participants' choices and reaction times. Order of presentation of trials was randomized for each L1 participant and each word was judged by 7 or 8 different L1 participants.

## Analysis

**Acoustic Analysis.** For each word in the production data, trained coders used the acoustic analysis software Praat (Boersma & Weenink, 2018) to measure aspects of low-level acoustic features relevant to the contrast between plosives. Coders were one native English speaker and one bilingual English speaker. Coders were blind to the L2 speaker's age and the identity of the words spoken. The inter-rater concordance rate for duration of acoustic features was over 99%.

For onset words, voice onset time (VOT) duration was marked as the length of time between the release of the onset stop consonant and the onset of periodicity marking the vowel. If there was aspiration, duration and mean aspiration intensity were measured. Vowel duration was measured using the .wav method to mark the start of the vowel and the F2 method to mark the end of the vowel. For coda words, vowel duration was also measured using the .wav method to mark the start of the vowel and F2 method to mark the end of the vowel. The duration of closure was measured as the length of time between the end of the vowel and the onset of voicing for the consonant, which meant that it excluded release bursts for the final phoneme. The duration of the final phoneme was measured as the length of time between the end of the duration of closure and the end of any voicing or aspiration associated with the phoneme.

**Statistical Analysis.** Analyses were conducted in R (R Core Team, 2018). Linear mixed effects models (Bates, Mächler, Bolker, & Walker, 2015) were fit for the effects of speaker age and the relevant acoustic properties of the stimuli words on the two measures of understanding of those words, accuracy and reaction time. To test the durational acoustic and speaker features underlying L1 understanding of L2 onsets, generalized and general linear mixed-effects models were fit for the effects of VOT (centered at its mean), voicing (unvoiced/voiced), and speaker age (adolescent/adult) in the onset words on accuracy and reaction time, respectively. To test the features affecting L1 understanding of L2 codas, generalized and general linear mixed-effects models were fit for the effects of vowel duration (centered at its mean), voicing (unvoiced/voiced), and speaker age (adolescent/adult) in the coda words on accuracy and reaction time, respectively. One L1 participant's data was removed from analysis for extremely fast reaction times and very low recorded accuracy of understanding.

Models included interactions of fixed effects in order to account for the predictions that understanding varies by features that themselves vary by voicing and age. Speaker, listener, and item were random effects in all models. Stimuli words that differ only in whether the target plosive was voiced or unvoiced were considered an item. For example, for the purposes of the analyses, *bay/pay* were treated as a single onset word item, and *cab/cap* were treated as a single coda word item. P-values for fixed effects in the logit mixed models were obtained via Wald *chi*-square tests. P-values for fixed effects in the general models were obtained via Wald *t*-tests (Kuznetsova, Brockhoff, & Christensen, 2017). For the general linear models, random intercepts and slopes were fit for speaker, listener, and item; for the logit mixed models, only random intercepts were included before the models failed to converge.

## Results

Table 1 summarizes descriptive statistics for the onset and coda productions of the L2 speakers and corresponding perceptions of the L1 speakers, which was the measure of L2 production accuracy. These results are tentatively in line with the hypothesis that L2 speaker age is associated with better L2 production accuracy, because onset productions by adults were better and more quickly understood by L1 participants than were onset productions by adolescents. Moreover, as with onsets, codas spoken by adults were better and more quickly understood than codas spoken by adolescents. Results are also tentatively in line with the hypothesis that L2 speakers develop hybrid phonotactic constraints on L2 phonetic categories, because it seems that L2 speakers use both aspiration and durational cues to voicing. Use of durational features is indicated by the longer VOT for unvoiced than for voiced onsets, with the difference between average voiced and unvoiced VOT greater in the adolescent group than in the adult group. Adolescent productions also showed a greater standard deviation. For codas, vowel length was also longer on average for voiced than for unvoiced codas. Use of aspiration as a cue to voicing is indicated by differences in accuracy for voiced and unvoiced plosives, perhaps due to the L2 speakers' consistently aspirated unvoiced plosives. Unvoiced onsets were better understood by L1 speakers than were voiced onsets, with voiced onsets spoken by adolescents the least well understood. Unvoiced codas were also better understood than voiced codas, with accuracy of L1 participants' judgments below 50% for voiced codas. The following sections present the output of the mixed models tying these acoustic and speaker features to L1 listener understanding.

**Features Associated with Accuracy.** Table 2 summarizes the results of the L2 production accuracy models. Please see Figures 6 and 7 for visualizations of L2 production accuracy as measured by L1 accuracy of judgments for L2 onsets and codas. The graphs show generalized linear smooths on the accuracy of each L1 judgment for onsets and codas. Accuracy varies according to the durational cue to voicing in onsets and codas, for both adult and adolescent speakers. However, even in the range of the durational feature at which the duration of that feature is not used to differentiate between voiced and unvoiced plosives - the range at which the linear smooths for voiced and unvoiced plosives overlap - L1 listener accuracy exceeds 50%. By implication, there is another cue to voicing, aspiration of unvoiced onset and coda plosives, in the L2 productions.

As depicted in Figure 6, for onset words, a main effect was found for voicing ($\beta =$

Table 1

*Features of L2 productions and L1 Accuracy and RT for the Same Words*

| | | Durational feature (ms) | | L1 Accuracy | | L1 RT (s, all trials) | |
|---|---|---|---|---|---|---|---|
| Voicing | Age | *M* | *SD* | Pct. Correct | *SD* | *M* | *SD* |
| **Coda words** | | | | | | | |
| Unvoiced | Adolescents | 143.467 | 47.449 | 76.7% | 0.423 | 2.531 | 2.448 |
| Unvoiced | Adults | 137.957 | 39.287 | 79.3% | 0.405 | 2.062 | 1.308 |
| Voiced | Adolescents | 156.930 | 58.955 | 38.5% | 0.487 | 2.560 | 1.390 |
| Voiced | Adults | 160.355 | 52.676 | 47.2% | 0.499 | 2.214 | 1.499 |
| **Onset words** | | | | | | | |
| Unvoiced | Adolescents | 102.050 | 35.362 | 93.6% | 0.245 | 2.207 | 1.270 |
| Unvoiced | Adults | 93.041 | 20.135 | 95.8% | 0.202 | 1.826 | 0.845 |
| Voiced | Adolescents | 25.877 | 21.989 | 79.3% | 0.405 | 2.490 | 1.971 |
| Voiced | Adults | 32.914 | 15.250 | 85.3% | 0.354 | 1.940 | 0.940 |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

-2.13, $SE$ = 0.30) with higher accuracy for unvoiced words (95% correct) than voiced words (82% correct). No other main effects were found. However, there was an interaction between voicing and VOT ($\beta$ = -16.40, $SE$ = 6.16), with accuracy decreasing with longer VOTs for voiced words but not for unvoiced words. Finally, a three-way interaction was found ($\beta$ = 27.31, $SE$ = 13.66). This interaction according to the predictions of the mixed model is visualized in Figure 16.

For voiced words said by adult speakers, accuracy increased with longer VOT, whereas for the same voiced words said by adolescent speakers, accuracy decreased with longer VOT. That is to say, when voiced VOT was longer, adult speakers were more likely to be accurately understood than were adolescent speakers. This difference between

Table 2

*GLMM estimates for L1 speakers' accuracy for L2 productions*

| | Coda | | Onset | |
|---|---|---|---|---|
| **Fixed Effects** | | | | |
| Intercept | 1.138 | *** | 3.304 | *** |
| Durational feature | −6.085 | *** | −2.492 | |
| Voiced | −1.628 | *** | −2.125 | *** |
| Adults | 0.082 | | 0.516 | |
| Durational feature X Voiced | 13.342 | *** | −16.396 | *** |
| Durational feature X Adults | −0.242 | | −1.501 | |
| Voiced X Adults | 0.315 | *** | 0.643 | |
| Durational feature X Voiced X Adults | 6.583 | *** | 27.313 | * |
| **Random Variances** | | | | |
| Listener | 0.038 | | 0.870 | |
| Subject | 0.084 | | 0.198 | |
| Item | 0.037 | | 0.136 | |
| Deviance | 12059.390 | | 2004.550 | |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

*$p$ < .05. **$p$ < .01. ***$p$ < .001.

speakers did not exist for the unvoiced words, nor for the voiced words with shorter VOT.

As depicted in Figure 7, for coda words, there was a main effect of voicing ($\beta$ = -1.63, $SE$ = 0.07) with higher accuracy for unvoiced words (78% correct) than voiced words (43% correct). There was also a main effect of vowel duration ($\beta$ = -6.09, $SE$ = 1.14), with lower accuracy for words with longer vowels. Crucially, there was an interaction between voicing and vowel duration ($\beta$ = 13.34, $SE$ = 1.33). As can be seen in Figure 17 which visualizes

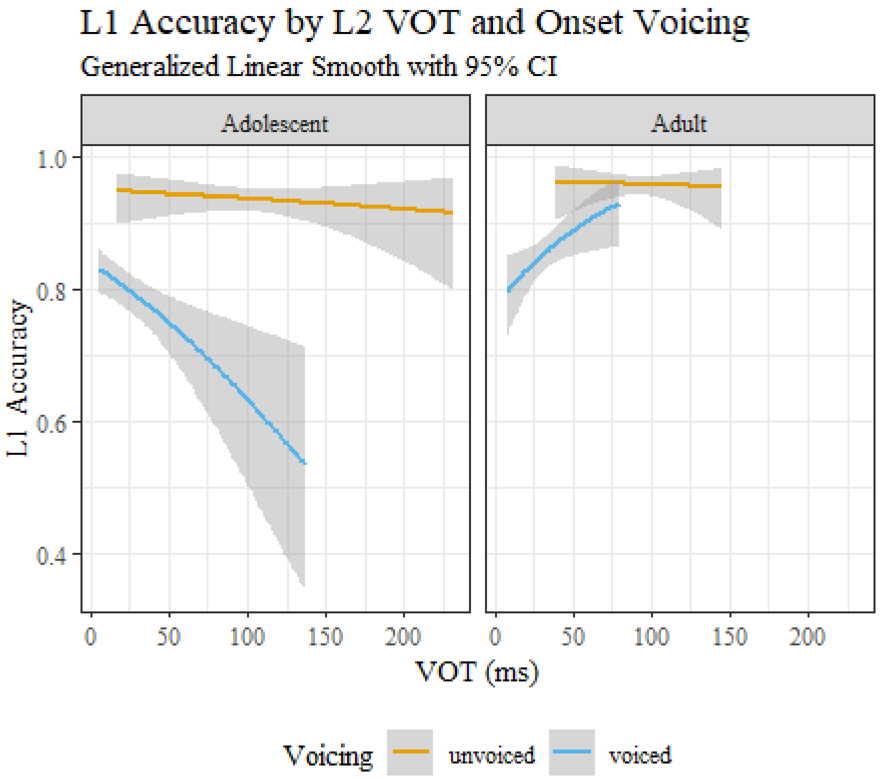


*Figure 6*. L1 Listeners' Accuracy Identifying L2 Speakers' Onsets

mixed model predictions for L2 coda production accuracy, for voiced words, accuracy increased with longer vowels, whereas for unvoiced words, accuracy decreased with longer vowels. There was no main effect of age, but there was an interaction between age and voicing ($\beta = 0.32$, $SE = 0.10$), with the difference in accuracy between voiced and unvoiced plosives higher for words said by adolescents than adults. The important three-way interaction of voicing, vowel duration, and age ($\beta = 6.58$, $SE = 2.06$) reflected that the opposing patterns of accuracy for voicing and vowel duration were greater for words said by adults than those said by adolescents. This interaction according to the predictions of the mixed model is visualized in Figure 17.

The three-way interaction is mediated by the opposing patterns of accuracy per voiced and unvoiced vowel duration. Overall, for voiced words, accuracy increased with vowel duration, while for unvoiced words, accuracy decreased with vowel duration.

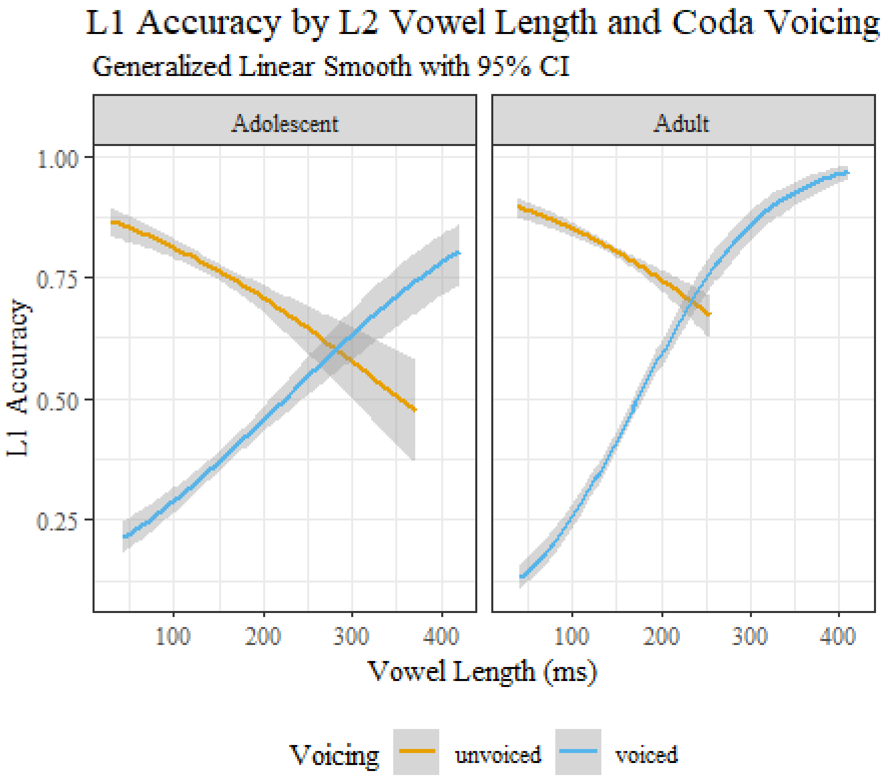


*Figure 7*. L1 Listeners' Accuracy Identifying L2 Speakers' Codas

Furthermore, for voiced words said by adult speakers as compared with adolescent speakers, accuracy increased more with longer vowel duration, whereas the same did not occur for the unvoiced words. That is to say, the greatest difference between accuracies for adult and adolescent speakers occurred at long vowel durations for voiced codas.

**Features Associated with Reaction Time.** Table 3 summarizes the results of the reaction time models. Please see Figures 8 and 9 for visualizations of L2 production RT as measured by L1 RT of judgments for L2 onsets and codas. RT does not seem to vary according to the durational cue to voicing in onsets and codas. Instead, the difference seems to be a function of speaker age.

As depicted in Figure 8, for onset words, there was a main effect of voicing ($\beta =$ 310.07, $SE = 108.49$) with faster reaction times for unvoiced words (201829) than for voiced words (221169). No other main effects or interactions were found. The mixed model

Table 3

*LMM estimates for L1 speakers' RTs (ms) for L2 productions (all trials)*

| | Coda | | Onset | |
|---|---|---|---|---|
| **Fixed Effects** | | | | |
| Intercept | 2551.351 | *** | 2139.514 | *** |
| Durational feature | 671.803 | | 1009.946 | |
| Voiced | 24.322 | | 310.067 | *** |
| Adults | −460.460 | *** | −302.567 | |
| Durational feature X Voiced | −191.352 | | −3195.852 | |
| Durational feature X Adults | 658.717 | | −1379.465 | |
| Voiced X Adults | 89.686 | | −161.589 | |
| Durational feature X Voiced X Adults | −1078.008 | | 6256.992 | |
| **Random Variances** | | | | |
| Listener | 206506.903 | | 126442.112 | |
| Subject | 49617.719 | | 68845.068 | |
| Item | 1285.471 | | 0.000 | |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

*$p$ < .05. **$p$ < .01. ***$p$ < .001.

predictions are visualized in Figure 18.

As depicted in Figure 9, for coda words, there was a main effect of age ($\beta$ = -460.46, $SE$ = 72.04) with faster reaction times for words spoken by adults (212951) than than adolescents (254526). No other main effects or interactions were found. Mixed model predictions are visualized in Figure 19. In light of these results, we ran the same models predicting RT only for correct trials.

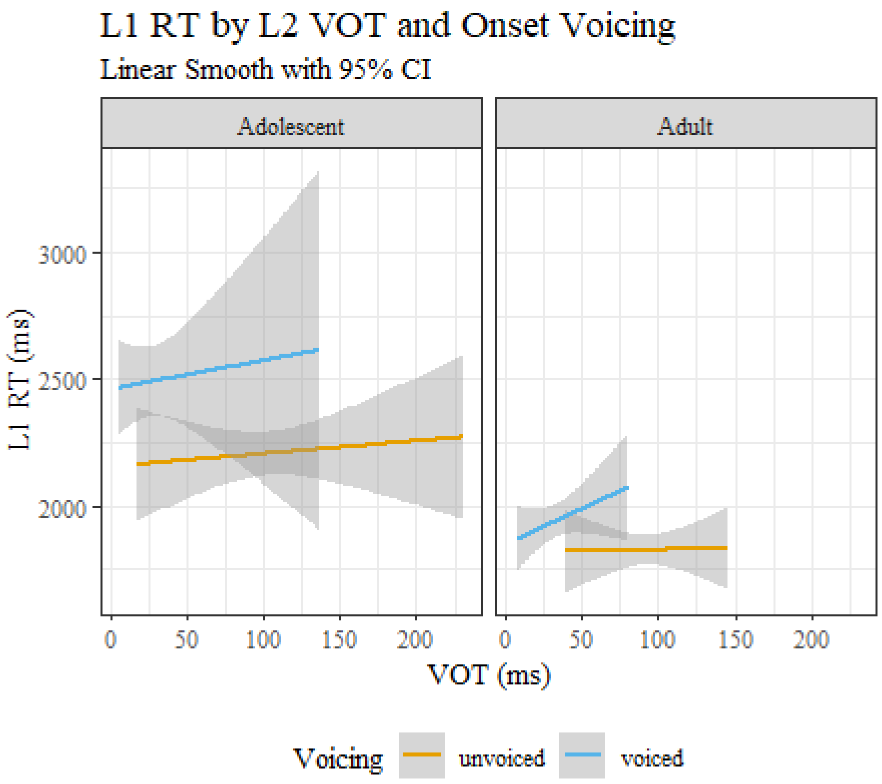


*Figure 8*. L1 Listeners' RTs for Identifying L2 Speakers' Onsets (all trials)

**Features Associated with Reaction Time for Correct Trials.** Table 4 summarizes the results of the reaction time models for correct trials (henceforth referred to as accurate L1 RT). Please see Figures 10 and 11 for visualizations of L2 production RT as measured by accurate L1 RT of judgments for L2 onsets and codas. Main effects of age and voicing on accurate L1 listener responses are in evidence which were not observed for all L1 responses in the previous models.

As depicted in Figure 10, for onset words, there was a main effect of voicing ($\beta$ = 278.64, $SE$ = 86.40) with faster reaction times for unvoiced words (200295) than for voiced words (212618). There was also a main effect of age ($\beta$ = -277.41, $SE$ = 117.72) with faster reaction times for adults (188183) than for adolescents (224406). No other main effects or interactions were found. The model predictions are visualized in Figure 20.

As depicted in Figure 11, for coda words, there was a main effect of age ($\beta$ = -404.68,

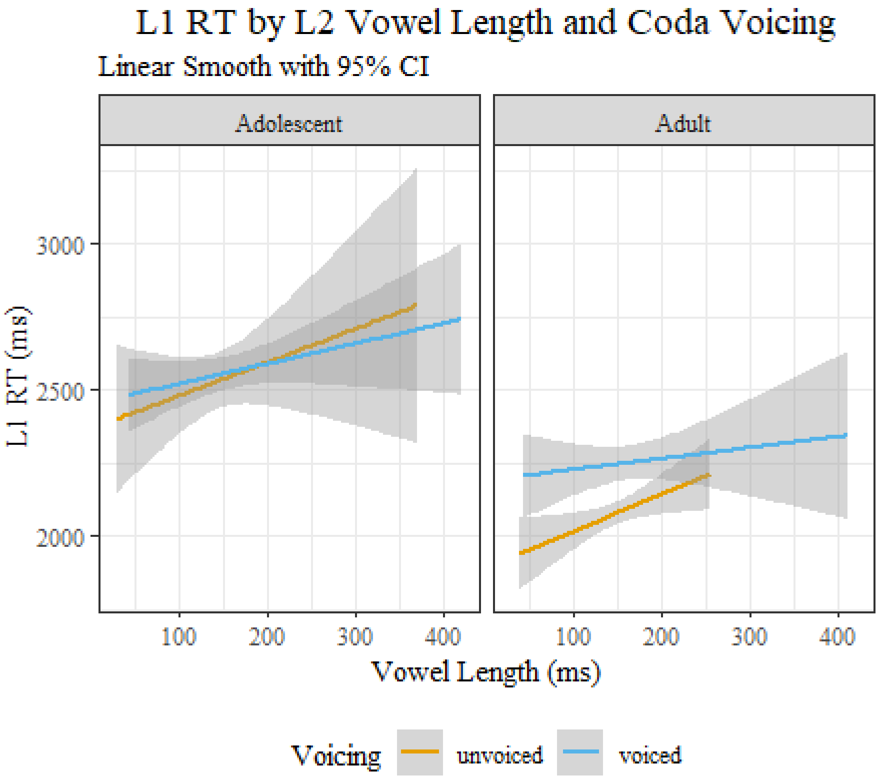


*Figure 9*. L1 Listeners' RTs for Identifying L2 Speakers' Codas (all trials)

$SE$ = 64.52) with faster reaction times for words spoken by adults (210039) than than adolescents (248414). There was a main effect of voicing ($\beta$ = 177.05, $SE$ = 50.20) with faster reaction times for unvoiced words (219391) than for voiced words (241016). There was also a main effect of vowel duration ($\beta$ = 1,289.65, $SE$ = 655.06) with faster reaction times for shorter vowel durations. No interactions were found. Model predictions are visualized in Figure 21.

## Discussion

Since listener participants were native English speakers, we expected higher accuracy and lower reaction time of L1 listener responses to L2 speaker productions that showed more native English-like phonological features. The expectation was confirmed for accuracy. In native English the perception of a plosive in coda position is conditioned by

Table 4

*LMM estimates for L1 speakers' RTs (ms) for L2 productions (correct trials)*

| | Coda | | Onset | |
|---|---:|---|---:|---|
| **Fixed Effects** | | | | |
| Intercept | 2456.028 | *** | 2114.379 | *** |
| Durational feature | 1289.646 | * | 936.937 | |
| Voiced | 177.049 | *** | 278.635 | *** |
| Adults | −404.677 | *** | −277.408 | * |
| Durational feature X Voiced | −1039.920 | | 135.067 | |
| Durational feature X Adults | −94.146 | | −1035.625 | |
| Voiced X Adults | 25.921 | | −131.087 | |
| Durational feature X Voiced X Adults | −1217.067 | | 3987.691 | |
| **Random Variances** | | | | |
| Listener | 188953.164 | | 122528.493 | |
| Subject | 43996.509 | | 57929.900 | |
| Item | 1679.137 | | 0.000 | |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

*$p$ < .05. **$p$ < .01. ***$p$ < .001.

the preceding vowel duration, and indeed for voiced words, accuracy increased with longer vowels, whereas for unvoiced words, accuracy decreased with longer vowels. Additionally the perception of a plosive in onset position is conditioned by the plosive VOT, and indeed voiced onsets were better understood the shorter their VOT, although unvoiced onsets were not better understood the longer their VOT. But unvoiced onset plosives had a high accuracy of understanding in general relative to the voiced plosive. There may have been a ceiling effect for unvoiced onset plosives, rendering it difficult to observe a change in accuracy varying by VOT duration.

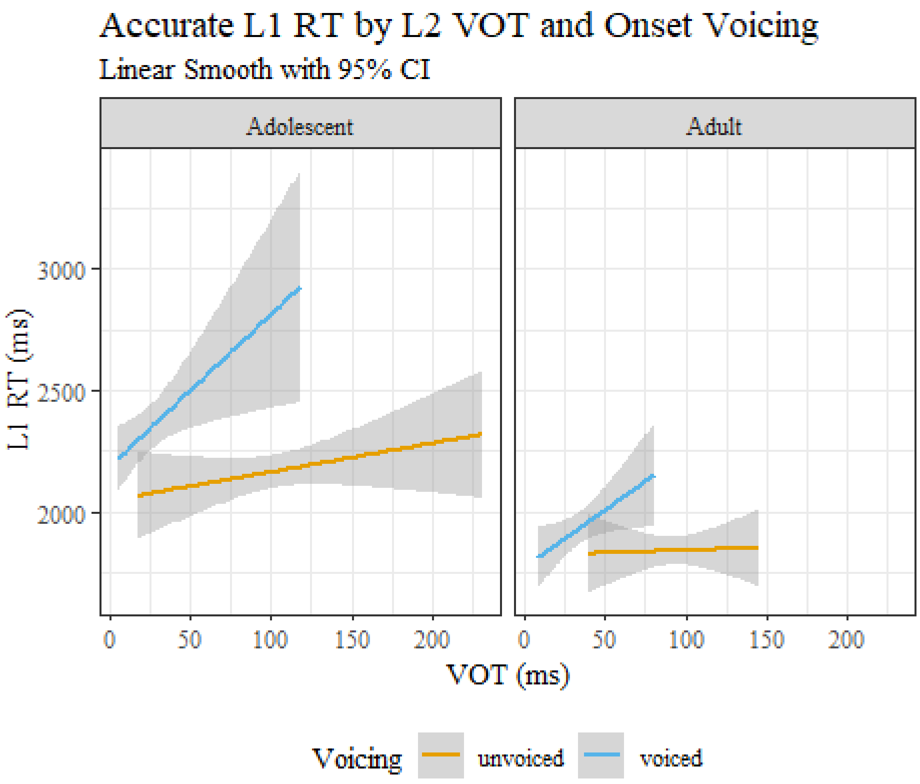


*Figure 10*. L1 Listeners' RTs for L2 Speakers' Onsets (correct trials only)

More importantly for consideration of L2 production, the results reflect that L2 speakers developed hybrid phonotactic constraints on L2 phonetic categories, because L2 speakers seem to use both English and Cantonese cues to voicing. First, L2 speakers used durational cues to voicing which are present in English but not in Cantonese. Unvoiced onsets were produced with longer VOT than voiced onsets, and voiced codas were produced with longer vowel length than unvoiced codas. These results rule out the hypotheses that L2 speakers developed equivalent classification of L1 and L2 phones or that they merely adjusted L2 phonetic features according to their L1, because neither of these hypotheses support the possibility of L2 speakers using a feature not in their L1.

Second, the results imply that L2 speakers consistently used the native Cantonese aspiration feature to differentiate betweeen voiced and unvoiced plosives in onset position, and that they moreover generalized the rule to plosives in coda position. L2 adults were

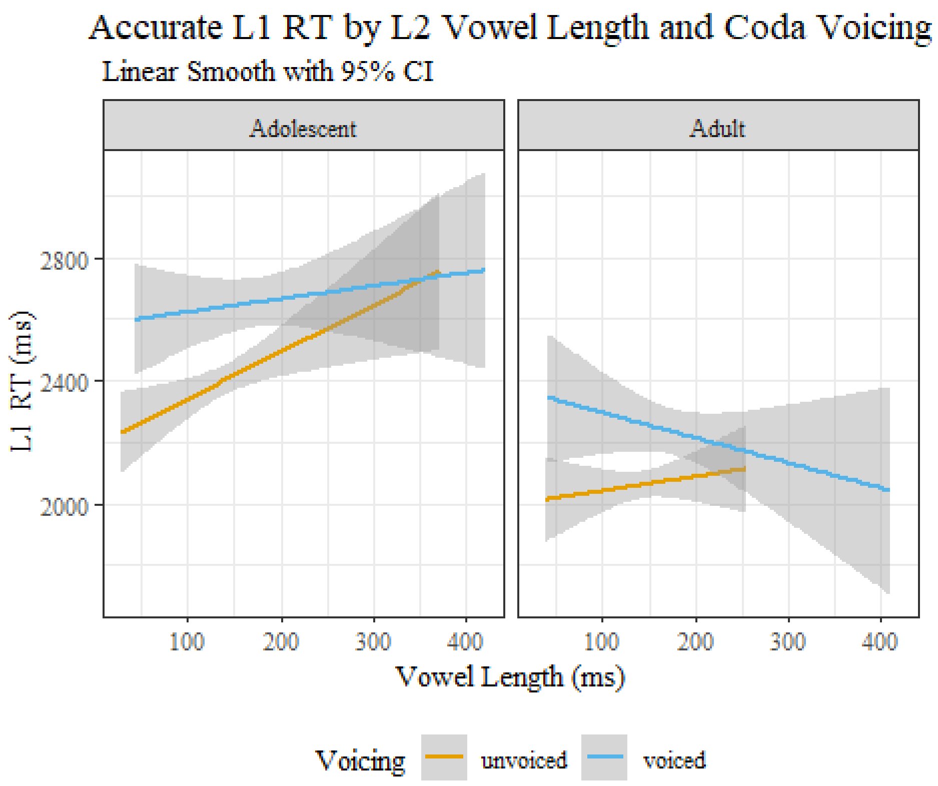


*Figure 11*. L1 Listeners' RTs for L2 Speakers' Codas (correct trials only)

understood above chance on voiced onsets even with very long VOTs. In particular, as depicted in Figure 7, L2 adults were understood with approximately 75% accuracy for both voiced and unvoiced codas that had the same vowel duration of approximately 225 ms. L2 adolescents were also understood at above chance in a similar range of vowel duration. In other words, in cases when vowel duration was not used to differentiate voicing, L2 speakers were accurately understood more often than not. This suggests that they provided listeners with another signal that a plosive was voiced or unvoiced, namely that they consistently aspirated unvoiced plosives.

Additionally, mean accuracy of native English-speakers' comprehension of L2 speech was lower for voiced onsets compared with unvoiced onsets and lower for voiced codas compared with unvoiced codas. Moreover, the longer the vowel, the lower the accuracy of L1 understanding. The L1 listeners were familiar with the aspiration contrast in onset position, but they differed from the L2 speakers in that the aspiration contrast in English is implemented with less regularity, and the English unaspirated unvoiced onset plosive is an ambiguous realization of either the voiced or unvoiced corresponding plosive. Therefore L1 listeners heard the aspirated unvoiced onset plosive unambiguously as the unvoiced realization, but had greater difficulty understanding that the L2 speakers were using the ambiguous unaspirated plosive as the voiced plosive. L1 listeners would then perceive the aspirated unvoiced coda as a hyperarticulated unvoiced plosive, but would perceive the unaspirated unvoiced coda as a truly ambiguous signal. Alternatively, words may have been pronounced with especially long vowels when L2 speakers were especially uncertain of what they were saying, and that speaker uncertainty may have led to a more ambiguous realization of the target word.

The results suggest that voiced plosives with especially long durational features were the most problematic for L1 listener understanding. An interaction between the durational feature, voicing, and L2 speaker age was seen in both the onset and coda mixed models. Underlying the interaction in both cases was the relatively large difference between the L2 production accuracy of adult speakers compared with adolescent speakers when the durational feature for the voiced plosive was long. In both cases, use of the aspiration contrast by the L2 speakers could not have constituted an unambiguous voicing signal to the L1 listeners. A particularly ambiguous signal could be considered a situation with a heavier cognitive load for L1 listeners, which would intensify differential reactions to other aspects of the signal normally masked under easier conditions. Thus, it may have been the case that the adult L2 speakers made more consistent use of the native English-like voicing feature compared with the adolescent speakers. A higher variability in the productions of the L2 adolescents supports this possibility.

Altogether, the associations between L2 production accuracy and the L2 speakers' use of hybrid acoustic-phonetic features for the categorization of voiced and unvoiced plosives suggests that L2 production success was tied to the L2 speakers' use of these acoustic-phonetic features. L1 and L2 speakers thus met each other half-way for successful communication.

The greater accuracy of L2 adults compared to L2 adolescent is consistent with the predictions of the SLM and not CPH. In particular, according to SML, more years of experience with an L2 involves more L2 input, which is beneficial for L2 phonological proficiency because it provides more opportunity for the L2 speaker perception of the differences between very similar phonetic categories.

Finally, unlike accuracy, reaction time was not well predicted by our models, although reaction time in correct trials was better predicted by speaker and acoustic features than was reaction time in general. Reaction time was understood to indicate processing difficulty. Reflecting the greater difficulty of voiced plosives in comparison with unvoiced plosives, listeners in correct trials took a longer time understanding voiced words compared with unvoiced words. Reflecting that adult speakers were better understood than adolescent speakers, listeners in correct trials took a longer time understanding words said by adolescent as compared with adult speakers. Furthermore, codas with longer vowels were more slowly understood than codas with shorter vowels. In line with the greater insensitivity of onset accuracy compared with coda accuracy to acoustic features, due perhaps to a ceiling effect for onset understandability, there was no main effect of VOT on reaction time for onsets. Altogether, reaction time was not as sensitive to phonological features as is accuracy.

Considering our results on accuracy and reaction time together, we suggest that reaction time corresponds to processes of understanding that are so automatized as to be largely insensitive to the relative difficulty of processing a particular utterance, except in unusually uncertain situations. Accuracy thus does not predict reaction time, even though accuracy is associated with the immediate acoustic features available in a speech signal and could have been assumed to depend on cognitive mechanisms that preface or underlie those on which reaction time depends. But if an utterance is inaccurately understood, it does not seem to be necessarily more slowly understood. Listeners must decide on what they heard - mapping phonetics to phonology - as quickly as possible. Listeners do not necessarily slow down due to processing error, only due to processing ambiguity.

## Study 2: L2 Perception

Study 2 focuses on tying L2 understanding of L1 speech to the acoustic features of an L2 speaker's productions. We maintain previous hypotheses on the role of speaker age and expand on the hypothesis of the relation between L2 perception and production. Following the general hypothesis of SLM that perception creates the mental representations that guide production, we look for associations between successes of L2 perception of L1 words and corresponding use of L1 durational cues to voicing in L2 production of the same words.

### Methods

**Participants.** The participants were the same L2 speakers who participated in Study 1 (42 adolescents, 36 adults).

**Stimuli.** The stimuli words were the same as those used in Study 1 (see Appendix).

**Experimental Procedure.** Using the elicitation procedure described in Study 1, L2 participants read each stimuli word aloud once. Using the identification procedure described in Study 1, L2 participants listened to each stimuli word presented over headphones and chose which of two words matched the word they heard (e.g., they heard *cab* and had to choose between *cab* and *cap*). The words that the L2 participants listened to were said by a native English speaking adult.

### Analysis

**Acoustic Analysis.** Analyses of production data was the same as described in Study 1.

**Statistical Analysis.** As in Study 1, generalized linear mixed effects models (Bates et al., 2015) were fit for the effects of speaker age and the relevant acoustic properties of the stimuli words on accuracy of understanding. However, instead of predicting L1 listener accuracy (the measure of L2 production accuracy), we predicted L2 listener accuracy of the same words said by L1 speakers (the measure of L2 perceptual accuracy). Random intercepts were fit for L2 speaker as any greater random effects structure led the models to fail to converge.

A generalized linear mixed-effects model was fit for the effects of VOT (centered at its mean), voicing (unvoiced/voiced), and speaker age (adolescent/adult) in the onset words on accuracy of L2 speaker judgments. A mixed model was fit for the effects of vowel duration (centered at its mean), voicing (unvoiced/voiced), and speaker age (adolescent/adult) in the coda words on L2 accuracy.

## Results

Table 5 summarizes descriptive statistics for the productions of the L2 speakers and their corresponding perceptions of native English speech. Results are tentatively in line with the hypothesis that L2 speakers rely on aspiration of unvoiced plosives to determine voicing in onsets, because for L2 adults and adolescents perceiving both onset words, unvoiced words were better understood than their voiced counterpart. In line with the possibility that L2 speakers default to the expectation that English codas, like Cantonese codas, are unvoiced, adults and adolescents also understood unvoiced codas better than voiced codas. Onset words were on the whole better understood than coda words. In line with the hypothesis that speaker age benefits perception, adolescents were much worse than adults at perceiving voiced and unvoiced codas, as well as voiced onsets. As seen in Study 1, results generally suggest that L2 speakers develop hybrid phonotactic constraints on L2 phonetic categories, because it seems that L2 speakers use both aspiration and durational features to perceive and produce voicing. Use of durational features in production is indicated by the longer VOT for unvoiced than for voiced onsets, and the longer vowel length for voiced than for unvoiced codas. Use of durational features in perception is indicated by greater than 50% L2 perceptual accuracy for voiced words. The following sections present the output of the mixed models tying these L2 acoustic and speaker features to L2 understanding of L1 speech.

Table 5

*Features of L2 productions and L2 Perceptual Accuracy of the Same Words*

| | | Durational feature (ms) | | L2 Accuracy | |
|---|---|---|---|---|---|
| Voicing | Age | *M* | *SD* | Pct. Correct | *SD* |
| **Coda words** | | | | | |
| Unvoiced | Adolescents | 143.467 | 47.449 | 64.1% | 0.480 |
| Unvoiced | Adults | 137.957 | 39.287 | 79.8% | 0.402 |
| Voiced | Adolescents | 156.930 | 58.955 | 60.4% | 0.489 |
| Voiced | Adults | 160.355 | 52.676 | 81.9% | 0.385 |
| **Onset words** | | | | | |
| Unvoiced | Adolescents | 102.050 | 35.362 | 94.4% | 0.230 |
| Unvoiced | Adults | 93.041 | 20.135 | 97.9% | 0.142 |
| Voiced | Adolescents | 25.877 | 21.989 | 75.6% | 0.430 |
| Voiced | Adults | 32.914 | 15.250 | 93.0% | 0.255 |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

**Features Associated with Accuracy.** Table 6 summarizes the results of the durational feature models. Please see Figures 12 and 13 for visualizations of L2 perceptual accuracy of L1 speech and how it correlates with L2 speakers' use of acoustic-phonetic cues to voicing. The graphs show generalized linear smooths on the accuracy of each L2 judgment for onsets and codas. L2 speakers' accuracy seems correlated with the L2 speakers' use of the durational cue to voicing in onsets and codas, for both adult and adolescent speakers. As in Study 1, even in the range of the durational feature at which the duration of that feature is not used to differentiate between voiced and unvoiced codas - the range at which the linear smooths for voiced and unvoiced codas overlap - L2 perceptual accuracy exceeds 50%. Accuracy for onsets is at ceiling and seems less sensitive to the durational feature. By implication, there is another cue to voicing especially for onsets through aspiration of unvoiced plosives.

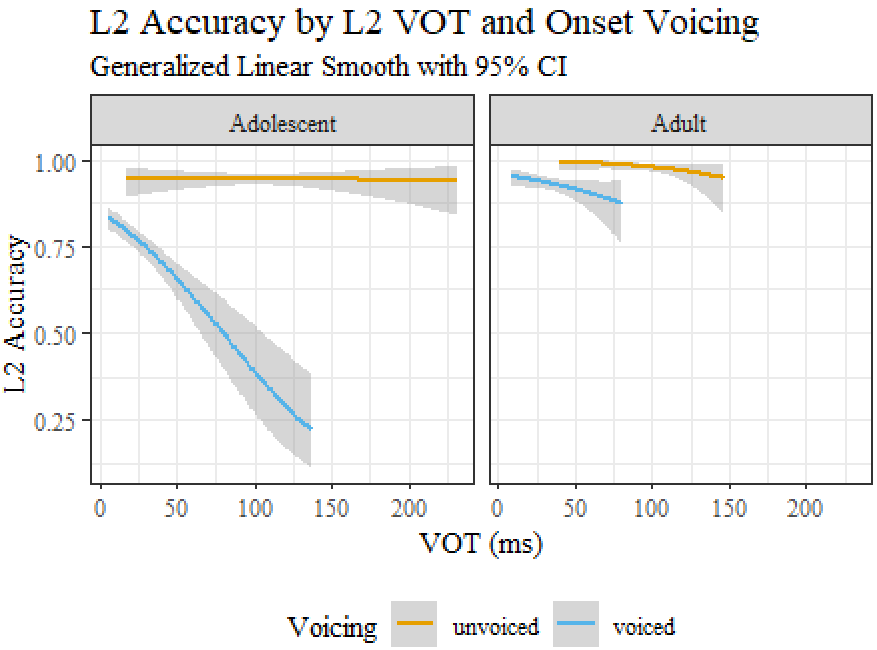


*Figure 12*. L2 Perceptual Accuracy and L2 Speakers' Onsets

As depicted in Figure 12 and in line with results from Study 1, for onset words, a main effect was found for voicing ($\beta$ = -2.61, $SE$ = 0.39) with higher accuracy for unvoiced words (96% correct) than voiced words (84% correct). There was also a main effect of age ($\beta$ = 10.25, $SE$ = 1.77) with adults showing higher accuracy (96% correct) than adolescents (85% correct). There was no interaction between voicing and VOT, but there

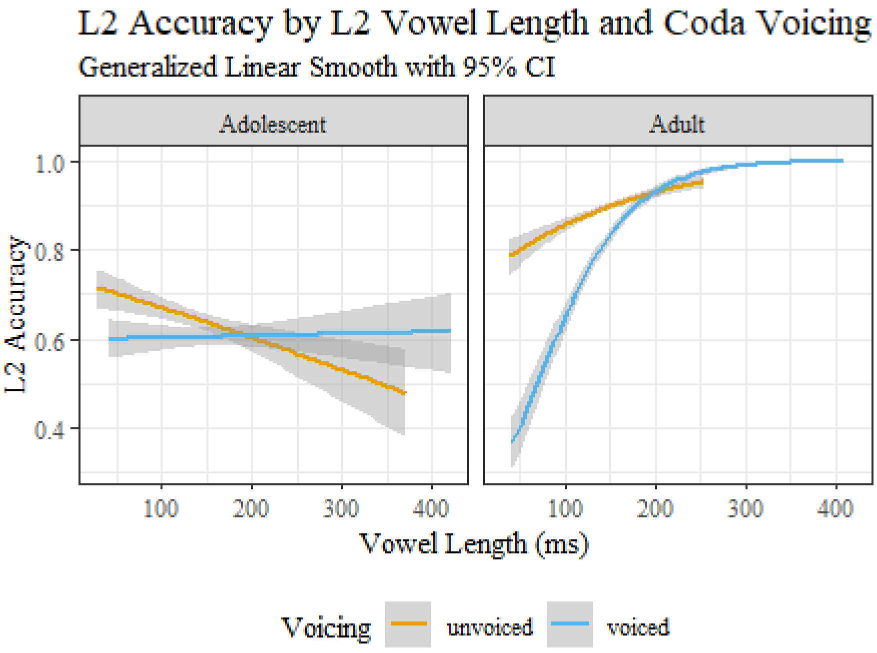


*Figure 13.* L2 Perceptual Accuracy and L2 Speakers' Codas

was an interaction between age and VOT ($\beta$ = -118.99, $SE$ = 21.44), with accuracy increasing with shorter VOTs for adults but not for adolescents. There was also an interaction between age and voicing ($\beta$ = -4.88, $SE$ = 1.19), with a greater difference between unvoiced and voiced words for adolescents compared to adults. A three-way interaction was found ($\beta$ = 130.87, $SE$ = 23.93). This interaction is visualized in Figure 22 which shows the mixed model predictions for onset accuracy.

Adolescents were always more accurate on unvoiced than voiced words regardless of VOT length, whereas adults were equally accurate on voiced and unvoiced onsets when they produced long VOT but not when they produced short VOT. Thus age mediated the extent to which there was an association between how well L2 speakers understood L1 speech and how well L2 speakers produced differences in the durational feature.

As depicted in Figure 13, for coda words, there was a main effect of voicing ($\beta$ = -0.17, $SE$ = 0.07) with higher accuracy for unvoiced words (73% correct) than voiced words (72% correct). There was also a main effect of age ($\beta$ = 1.24, $SE$ = 0.18) with

Table 6

*GLMM estimates for L2 speakers' accuracy for L1 productions*

| | Coda | | Onset | |
|---|---:|---|---:|---|
| **Fixed Effects** | | | | |
| Intercept | 0.643 | * | 6.056 | *** |
| Durational feature | −2.284 | * | −0.157 | |
| Voiced | −0.167 | * | −2.611 | *** |
| Adults | 1.239 | *** | 10.254 | *** |
| Durational feature X Voiced | 4.076 | *** | 4.145 | |
| Durational feature X Adults | 6.770 | *** | −118.994 | *** |
| Voiced X Adults | 0.439 | *** | −4.881 | *** |
| Durational feature X Voiced X Adults | 11.921 | *** | 130.872 | *** |
| **Random Variances** | | | | |
| Subject | 0.499 | | 17.308 | |
| Item | 0.489 | | 1.643 | |
| Deviance | 11329.780 | | 1045.150 | |

*Note.* Durational feature is vowel duration for coda and VOT for onset.

*$p < .05$. **$p < .01$. ***$p < .001$.

adults showing higher accuracy (81% correct) than adolescents (62% correct). There was also a main effect of vowel duration ($\beta$ = -2.28, $SE$ = 1.16), with lower accuracy for words with longer vowels. There was an interaction between voicing and vowel duration ($\beta$ = 4.08, $SE$ = 1.31). As can be seen in Figure 23, accuracy increased with longer vowels for voiced but not for unvoiced words. There was also an interaction between age and voicing ($\beta$ = 0.44, $SE$ = 0.11), with the difference in accuracy between voiced and unvoiced plosives higher for adults than adolescents. The important three-way interaction of voicing, vowel duration, and age ($\beta$ = 11.92, $SE$ = 2.38) reflected that the opposing patterns of accuracy for voicing and vowel duration were greater for words said by adults than those said by adolescents. This interaction is visualized in Figure 23, which shows the mixed model predictions for L2 perceptual accuracy of L1 coda words.

The three-way interaction is mediated by the opposing patterns of accuracy per voiced and unvoiced vowel duration. Overall, for voiced words, accuracy increased with vowel duration. Furthermore, for voiced words and adult speakers as compared with adolescent speakers, accuracy increased more with longer vowel duration, whereas the same did not occur for the unvoiced words. That is to say, the greatest difference between accuracies for adult and adolescent speakers occurred at long vowel durations for voiced codas, as in Study 1. Adolescents' perceptual accuracy was reflected less in production that was sensitive to durational features, whereas adults were more accurate on voiced codas when they produced long vowel duration. As with onsets, age mediated the extent to which there was an association between how well L2 speakers understood L1 speech and how well L2 speakers produced differences in the durational feature.

## Discussion

The perceptual accuracy of the L2 participants was sensitive to native English-like durational cues to voicing, as reflected by the associations between L2 perceptual accuracy and the acoustic-phonetic characteristics of L2 productions of the same words. In native English the perception of a plosive in coda position is conditioned by the preceding vowel duration, and indeed L2 speakers who showed higher accuracy of perception also produced longer vowels for voiced words, whereas L2 speakers who showed low accuracy of perception produced longer vowels for unvoiced words. This effect was strong for L2 adults compared with L2 adolescents. L2 speakers' perceptions of L1 onsets and codas suggest that L2 speakers were able to use both VOT and vowel duration as well as aspiration to contrast voiced and unvoiced plosives, because L2 accuracy on voiced onsets and codas was higher than complete misunderstanding or 50% misunderstanding, as would be predicted by hypotheses that L2 speakers rely only on aspiration.

However, features of production showed less sensitivity to L2 perception of a plosive in onset position by the plosive VOT. Unvoiced onset plosives had a high accuracy of L2 understanding in general relative to the voiced plosive. There may have been a ceiling effect for unvoiced onset plosives, rendering it difficult to observe a change in accuracy varying by VOT duration. But in particular, as no main or interaction effect was observed for VOT length on adolescent accuracy, it seems that L2 adolescent productions were not closely tied to their perception of VOT as a cue to voicing. Adults were more accurate on unvoiced onsets with short as compared with long VOT. Altogether, L2 adults and adolescents likely relied more on aspiration to perceive and produce the contrast in voiced and unvoiced onsets.

Thus the results, as in Study 1, reflect that L2 speakers developed hybrid phonotactic constraints on L2 phonetic categories, because L2 speakers used both durational and aspiration cues to voicing. More importantly for consideration of how L2 speakers' phonetic perception is tied to phonetic production, L2 speakers' perceptual accuracy in judging the same words spoken by an L1 speaker was associated with their use of hybrid phonotactic constraints. In line with the hypothesis that perception does not lag behind production, L2 perception did not show that L2 speakers developed equivalent classification of L1 and L2 phones or that they merely adjusted L2 phonetic features according to their L1, because neither of these hypotheses support the possibility of L2 speakers understanding voiced codas well and therefore using the duration feature not in their L1.

The difference between results for adult and adolescent speakers conformed to the SLM hypothesis that speaker age benefits phonetic perception or the connection between phonetic perception and production. Adults had higher average perceptual accuracy than adolescents in general. Adolescents were always more accurate on unvoiced than voiced words regardless of VOT length, whereas adults were equally accurate on voiced and unvoiced onsets when they produced long VOT but not when they produced short VOT. Thus age mediated the extent to which there was an association between how well L2 speakers understood L1 speech and how well L2 speakers produced differences in the durational feature. Perhaps this is due to L2 adolescents' production abilities lagging farther behind their perceptual abilities as compared with L2 adults. The SML hypothesis that a speaker's age or years of experience learning their L2, rather than their initial age of learning, determines their L2 phonological proficiency.

## Relationship between L2 Perception and Production

Study 3 focuses on tying L2 judgments of L1 speech to L1 judgments of L2 speech. We are interested in understanding if L2 speakers' perceptions are better than their productions and if L2 speakers' productions are better than their perceptions. The former possibility is in line with the notion that perception precedes production because perception creates the mental representations of phonetic categories that subsequently guides production. The Speech Learning Model in particular predicts that perception can never be worse than production. The latter possibility in in line with the notion that production precedes perception because phonetic categories are stored as abstract articulatory programs which subsequently guide the perception of phonetic categories. The Motor Theory in particular predicts that production will never be worse than perception. Hence we look for how speaker age and voicing are tied to the difference between L2 perceptual and production accuracy for each word.

### Methods

**Participants.** Participants are the same L2 and L1 participants as in Study 1.

**Stimuli.** The stimuli words are the same as in Study 1 (please see Appendix).

## Analysis

In order to investigate the relationship between phoneme perception and production, we calculated the difference between the mean perceptual accuracy and mean production accuracy score for each stimuli word for each participant. If the “Perception - Production” score is greater than 0, this means that, consistent with SML, perception outstrips production. If the Perception - Production score is less than 0, this means that, consistent with MT, production outstrips perception.

As in the previous studies, general linear mixed effects models (Bates et al., 2015) were fit for effects on understanding of onset and coda words. To test if L2 perception leads L2 production, a general linear mixed-effects model was fit for the effects of voicing (unvoiced/voiced) and speaker age (adolescent/adult) in the onset words on the Perception - Production score of each word. A second model was fit for the effects of voicing (unvoiced/voiced) and speaker age (adolescent/adult) in the coda words on the Perception - Production score of each word. Models included interactions of fixed effects in order to account for the predictions that understanding varies by voicing and age. Speaker, listener, and item were random effects in all models. P-values for fixed effects were obtained via Wald $t$-square tests. Random intercepts were fit for speaker, listener, and item as any greater random effects structure led the models to fail to converge.

## Results

Table 7 summarizes the results of the models.

For onset words a main effect was found for voicing ($\beta$ = -0.04, $SE$ = 0.01) with a more positive difference between perception and production for voiced words (0.02) than for unvoiced words (0.01). In particular, adults were better at perceiving than producing voiced onsets. There was no main effect of age. But there was an interaction between voicing and age ($\beta$ = 0.09, $SE$ = 0.02). The difference in the perception-production

Table 7

*LMM estimates for difference between mean L2 production and perception accuracy per each word*

| | Coda | | Onset | |
|---|---|---|---|---|
| **Fixed Effects** | | | | |
| Intercept | −0.127 | *** | 0.005 | |
| Voiced | 0.345 | *** | −0.038 | *** |
| Adults | 0.136 | *** | 0.013 | |
| Voiced x Adults | −0.008 | | 0.093 | *** |
| **Random Variances** | | | | |
| Listener | 0.001 | | 0.003 | |
| Subject | 0.014 | | 0.013 | |
| Item | 0.010 | | 0.015 | |

*$p < .05$. **$p < .01$. ***$p < .001$.

difference for adults and adolescents was greater for voiced than for unvoiced words. Figure 14 visualizes the model output. The y axis shows mean difference per word between production and perception accuracy. If the value of the difference is greater than 0, perception outstrips production, and if the value is less than 0, production outstrips perception. If the value straddles 0, then neither perception nor production outstrips the other. As the figure show, the greatest difference in the perception-production difference for adults and adolescents was for voiced onsets.

However, as depicted in Figure 14, the difference between perception and production overlapped with zero for adolescents perceiving and producing voiced onsets and for both adolescents and adults perceiving and producing unvoiced onsets. Results thus indicate that L2 production and L2 perception were not significantly different from each other, on

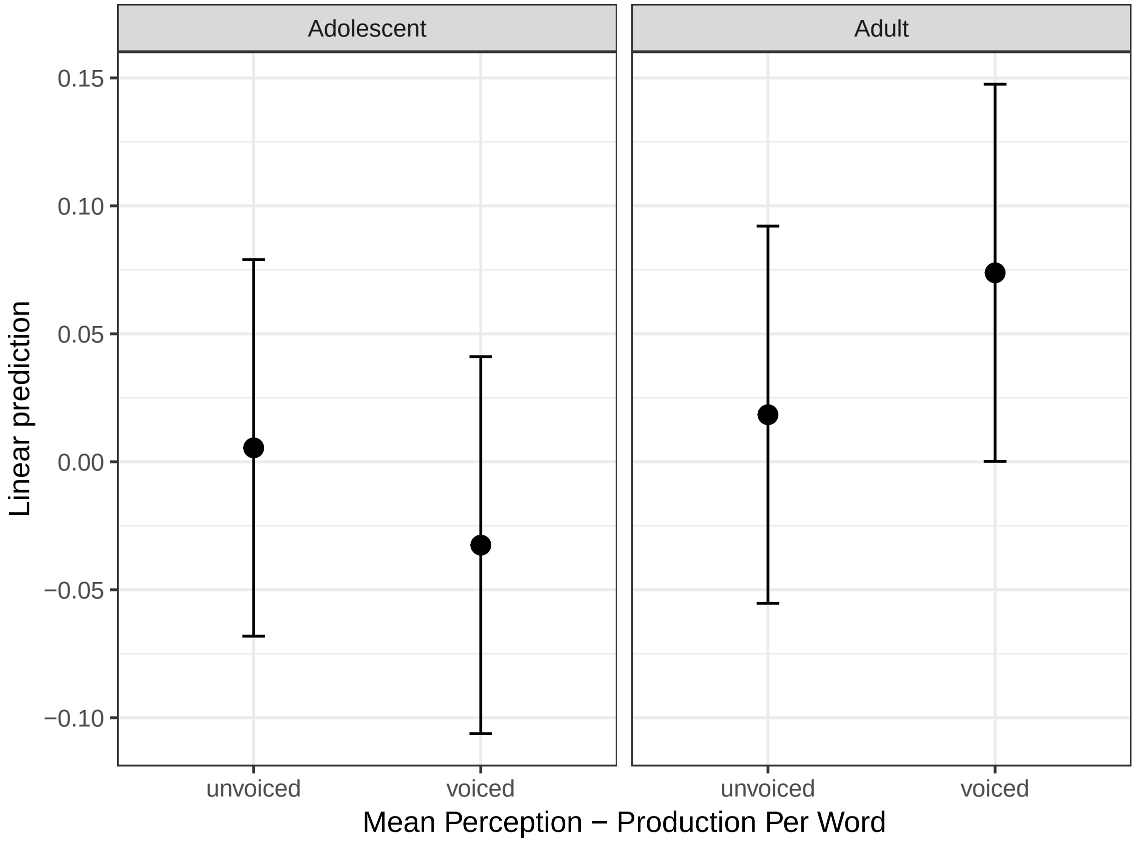


*Figure 14.* Model Predictions for Mean Difference Between L2 Production and Perception Accuracy Per Onset Words

average, per each onset word, except in the case of adults and voiced onsets.

For coda words, there was a main effect of voicing ($\beta = 0.34$, $SE = 0.01$) with a more positive difference between perception and production for voiced words (0.30) than for unvoiced words (0.02). There was also a main effect of age ($\beta = 0.14$, $SE = 0.03$) with adults showing a more positive difference between perception and production (0.81) than adolescents (0.62). Model results are visualized in Figure 15.

Overall, as depicted in Figure 15, the difference between perception and production only overlapped with zero for adults perceiving and producing unvoiced codas. Adults and adolescents perceiving and producing voiced codas were better at perception than production. However, adolescents were worse at perceiving unvoiced codas than they were

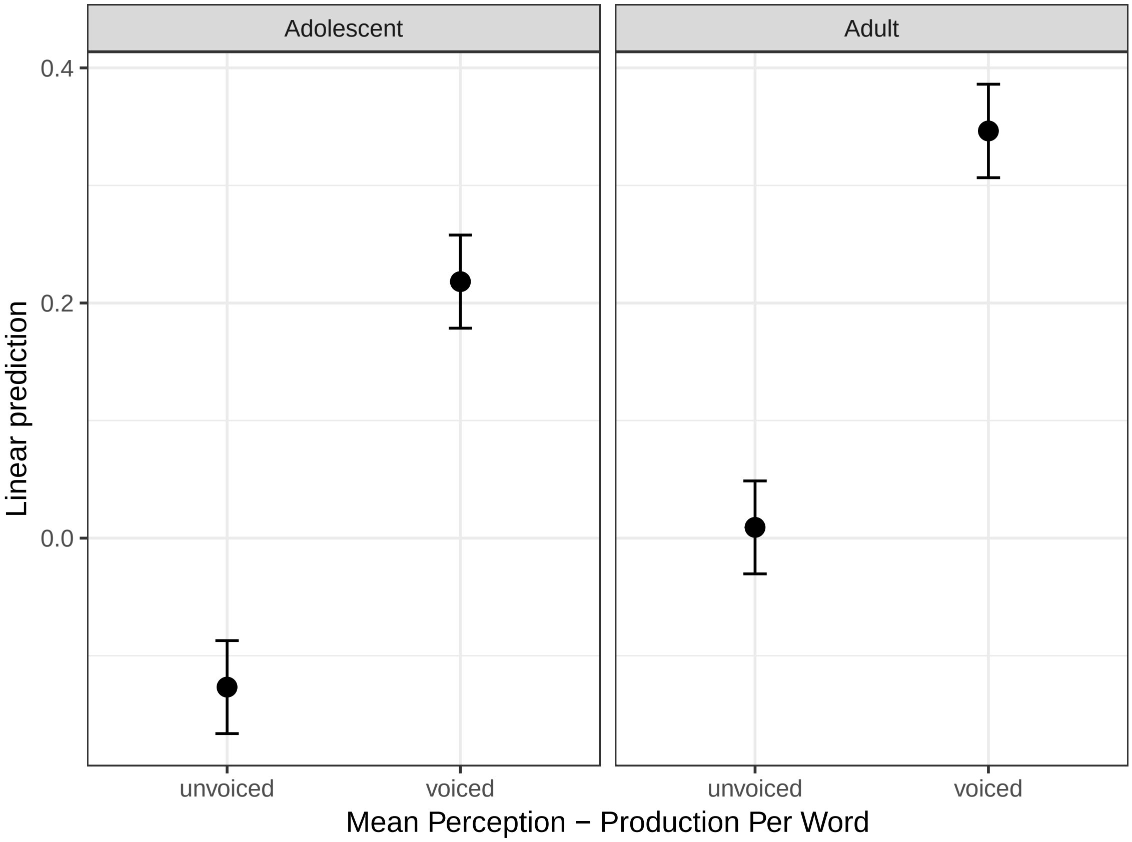


*Figure 15.* Model Predictions for Mean Difference Between L2 Production and Perception Accuracy Per Coda Words

at producing unvoiced codas.

### Discussion

Adult and adolescent L2 speakers were more accurate at perceiving than at producing voiced codas. L1 speakers' productions of codas would not have relied on aspiration to differentiate voiced from unvoiced plosives, which suggests that L2 speakers were able to use primarily the vowel duration contrast to perceive coda voicing. Furthermore, results from Study 1, which implied that L2 speakers developed hybrid phonotactic constraints on phonetic categories, suggest that L2 speakers' ability to produce voiced codas was associated with use of both aspiration and vowel duration as features differentiating between voiced and unvoiced codas. But because L1 English listeners would not have relied on lack of aspiration to determine that a coda was voiced, L2 speakers' productions of voiced codas were thus less accurately understood than their perceptions of the same. Results thus seem in line with the SLM hypothesis that phonetic perception can outstrip phonetic production. These results directly contradict the Motor Theory prediction that perception cannot be more accurate than production.

However, adult and adolescent L2 speakers were on average equally accurate at perceiving and producing voiced and unvoiced onset plosives, although adult perception of voiced onsets was a little better than their production. These results are not able to adjudicate between SLM and Motor Theory predictions. However, we suggest that since L1 and L2 speakers could agree on the use of aspiration as a cue to voicing in onsets, L2 speakers did not need to further develop the acoustic-phonetic features defining phonemic categories in onsets to the same extent as in codas. L2 production and L2 perception can be equally accurate on the basis of the use of aspiration, which brings successes of perception and production close to ceiling. Only in the case of voiced onsets, where the VOT feature plays a larger role in defining voicing, adults' perceptions were slightly better than their productions, which is in line with the previous results that adults were better than adolescents at perceiving the VOT feature.

Finally, adolescent L2 speakers were more accurate at producing than perceiving unvoiced onsets. This may have been the case because adolescents as compared with adults, as shown in Studies 1 and 2, relied less on VOT as a cue to voicing in onsets. While L1 speakers would have been able to accurately perceive unvoiced onsets as aspirated unvoiced plosives, adolescent L2 speakers may have found it difficult to perceive unaspirated unvoiced plosives as realizations of unvoiced onsets, thereby leading to a discrepancy between L2 adolescent production and perception accuracy.

## General Discussion

Studies 1, 2, and 3 jointly address the three broad questions asked in the hypotheses about L2 production and perception. The first question was, how do L1 acoustic-phonetic features influence the perception and production of L2 phonemes? We assume that a speaker's L1 phonological system influences or interferes with their L2 system to an extent, but this interference can take multiple forms. SLM, which centrally posits that L2 speakers must maintain contrasts between L1 and L2 phonetic categories in order to successfully perceive and produce in their L2, provides a useful framework for operationalizing hypotheses between two extremes.

At one extreme, which we called the Equivalent Classification Hypothesis, L2 speakers may completely assimilate an L2 phone to an L1 phone such they only rely on the acoustic-phonetic features of their L1 when they produce and perceive their L2. We did not find evidence for equivalent classification. For example, the hypothesis predicts that L1 Cantonese speakers, when they listen to English, will perceive all English plosive codas as voiceless, because Cantonese only permits voiceless plosives in coda position. In such a case, Cantonese speakers would never perceive English as voiced. However, L2 adolescents accurately perceived English voiced codas about 60% of the time and L2 adults accurately percieved English voiced codas about 80% of the time. Moreover, in the case of equivalent classification L1 Cantonese speakers are expected to always produce English plosive codas as aspirated voiceless plosives, which, in the absence of other features that L1 English listeners could use identify that a coda is voiced, would lead all L2 codas to be perceived as unvoiced. However, when L2 adolescents and adults intended to produce a voiced coda, they were accurately understood nearly 80% of the time. It does not seem to be the case that L2 speakers rely only on their L1 to produce and perceive in their L2.

The Feature Adjustment Hypothesis articulates a second, less extreme possibility. L2 speakers use their L1 phonological systems to produce and perceive phone in an L2, but they may amend the use of position-dependent phonetic features to apply to phones in other positions in a word. For example, L1 Cantonese speakers perceiving and producting English would recognize that English does have voiced plosives in coda position. L2 speakers would therefore be able to identify and articulate voiced codas, but they would do so according to the acoustic-phonetic features that define voicing in Cantonese phonemes, namely by always aspirating unvoiced plosives and never aspirating voiced plosives. The current studies did not report on the presence and intensity of aspiration in L2 participants' productions and perceptions. However, the case of feature adjustment predicts that L2 speakers do not use acoustic-phonetic features native to their L2 and nonnative to their L1. Since we found evidence that L2 participants produced and perceived durational acoustic-phonetic cues to voicing in English, we rule out the possibility that L2 speakers only rely on their L1 phonological system to define their L2 system.

L2 speakers used durational cues to voicing which are present in English but not in Cantonese. In productions by both L2 adults and adolescents, unvoiced onsets were produced with longer VOT than voiced onsets, and voiced codas were produced with longer vowel length than unvoiced codas. These features of production were tied to functional success. When L2 speakers used the dominant durational feature for voicing in English, they were more accurately understood by English listeners. Findings were especially strong for coda words, as L1 listeners more accurately understood voiced plosives with longer vowels and unvoiced plosives with shorter vowels. We found moreover that when L2 speakers accurately understood L1 speech they were more likely to use the dominant feature for voicing in their own speech. These findings were especially strong for coda words, as L2 speakers who better understand English plosives produced voiced plosives with longer vowels and unvoiced plosives with shorter vowels.

Use of English durational features is in line with the other two hypotheses for how L1 acoustic-phonetic features influence the perception and production of L2 phonemes. The more extreme hypothesis, which we called Category Formation, would suggest that L2 speakers develop phonetic categories that closely resemble native speakers' categories. Category Formation predicts that L2 speakers do not use their L1 system to produce and perceive their L2. Thus, in tandem with the use of English durational features, L2 speakers would be expected to aspirate in a native English-like way. In particular, L1 Cantonese speakers would never aspirate plosives in coda position, although, like L1 English speakers, they may usually aspirate unvoiced plosives in onset position. In contrast, the Hybrid Phonotactic Constraints Hypothesis suggests that L2 speakers combine use of L1 and L2 phonological systems to produce and perceive phonemes in an L2. Phonetic features native to both their L1 and L2 systems would inform the categorization of phones in L2 speakers' L2 production and perception, with priority given to L1 features. This latter hypothesis predicts that while durational cues would condition the perception of voicing in unaspirated plosives, L2 speakers would always perceive aspirated plosives in both onset and coda position as unvoiced phones. Likewise, L2 productions would contain durational cues to voicing but would also consistently aspirated unvoiced and not aspirate voiced plosives. Results were in line with the hypothesis that L2 speakers develop hybrid phonotactic constraints, because in addition to the findings that L2 speakers use durational features, multiple results imply the regular use of aspiration.

Onset words were more accurately perceived and produced than coda words, on average. For onset words, there was high accuracy of both L2 perception and production in cases of even very short VOT in unvoiced words, reflecting that L2 speakers may have relied on perceiving and producing aspiration to characterize the contrast between voiced and unvoiced plosive onsets. L2 productions of unvoiced onset plosives were tied to a high accuracy of L2 perception in general relative to the voiced plosive and regardless of the plosive VOT. There may have been a ceiling effect for unvoiced onset plosives, rendering it difficult to observe a change in accuracy varying by VOT duration. But in particular, L2 adolescent productions did not seem closely tied to their perception of VOT as a cue to voicing, because no main or interaction effect was observed for VOT length on L2 adolescent accuracy of perception.

Results imply that L2 speakers moreover generalized the rule to plosives in coda position. L2 adults were understood more than 50% of the time on voiced onsets even with very long VOTs. In particular, L2 adults were understood with approximately 75% accuracy for both voiced and unvoiced codas that had the same vowel duration of approximately 225 ms. L2 adolescents were also understood at above chance in a similar range of vowel duration. In other words, in cases when vowel duration was not used to differentiate voicing, L2 speakers were accurately understood more often than not. This suggests that they provided listeners with another signal that a plosive was voiced or unvoiced, namely that they aspirated unvoiced plosives. That voiced plosives with especially long durational features were the most problematic for L1 listener understanding may have been the result of the ambiguity created for L1 listeners when L2 speakers used aspiration to contrast voicing - that is to say, when L2 productions defined voiced plosives as unaspirated plosives, all other features being equal. An interaction between the durational feature, voicing, and L2 speaker age was seen in both the onset and coda mixed models. Underlying the interaction in both cases was the relatively large difference between the L2 production accuracy of adult speakers compared with adolescent speakers when the durational feature for the voiced plosive was long. In both cases, use of the aspiration contrast by the L2 speakers could not have constituted an unambiguous voicing signal to the L1 listeners.

The second broad question is the role of speaker age on the production and perception of L2 phonemes. In the current study L2 participants all began learning L2 at age 5, but one group was aged about 12 years old and the other group was aged about 20 years old. The Critical Period Hypothesis would predict that the older group would not outperform the younger group, because L2 phonologies fossilize early. On the other hand, input hypotheses, in particular SLM, predict that the older group would perceive and produce L2 phonemes more accurately than the younger group, because L2 phonologies are adaptive and rely on amount and goodness of L2 input, which adults would have had more of than adolescents. Results are in line with SLM. Adult L2 productions and perceptions were more accurate than adolescent productions and perceptions for voiced codas, unvoiced coda, voiced onsets, and unvoiced onsets. Adult L2 speakers also seemed to implement the vowel duration coda contrast and the aspiration onset contrast with greater consistency than adolescent L2 speakers in both production and perception. An interesting finding is that adults produced on average a smaller difference between voiced and unvoiced vowel durations, and voiced and unvoiced VOT, as compared with adolescents. A smaller, more consistent difference in the use of durational features to constrast phones may reflect greater efficiency in production.

The last question that these studies intended to address was the relationship between production and perception of L2 phonemes. SLM argues that perception creates mental representations that guide production, such that L2 production accuracy cannot be better than L2 perceptual accuracy. MT argues that phonetic categories are stored as abstract articulatory programs which guide perception of phonetic categories, such that L2 perceptual accuracy cannot be better than production accuracy. We found that L2 adults and adolescents were more accurate on average at perceiving than at producing voiced codas, which were exactly the cases when hybrid phonotactic constraints on phonetic categories would have allowed L2 speakers to accurately understand L2 English speakers but not to produce voiced plosives in a manner unambiguously signalling voicing to L1 listeners.

## Conclusion

L2 speakers' use of both durational features found in English but not Cantonese, and aspiration features used more regularly in Cantonese than in English, suggests that L2 speakers developed hybrid phonotactic constraints on their mental representations of L2 phones. These phonetic categories guided L2 production such that L2 perception and L2 production were found to be associated, with production either lagging behind or equal to perception for voiced and unvoiced onset and coda plosives. Altogether, the associations between L2 perception accuracy and the L2 speakers' use of hybrid acoustic-phonetic features for the categorization of voiced and unvoiced plosives suggests that L2 perception and production success was tied to the L2 speakers' use of these acoustic-phonetic features. L1 and L2 speakers thus met each other half-way for successful communication.

Accented speech is not an impediment to communication, as multiple factors of the listener as well as of the speaker and listener relative to each other affect listener understanding. Furthermore, the expectation may be incorrect that older learners of a second language face a disadvantage as compared with younger learners. Our study shows that older L2 speakers were more accurately and quickly understood. Finally, our study removed semantic and syntactic context from the speech stimuli, possibly rendering the task of understanding between speaker and listener as difficult as possible. Under everyday circumstances, context is likely to have an ameliorative effect on communication, particularly by aiding in adaptation to talker-specific variability.

### Future Studies

The current studies focused on the role of durational acoustic-phonetic features in L2 production and perception. We assumed, since these durational features together with aspiration are the primary cues to voicing according to the literature, that instances of successful L2 production and perception not attributable to use of the durational feature were attributable to use of the aspiration feature. However, future study should measure the presence and intensity of plosive aspiration in both onset and coda position and consider its role in successful speech. Future study can also ask how multiple features work together. If speakers and listeners are efficient, there are still multiple ways in which they may pursue minimal effort to achieve successful communication. It may be possible to observe a tradeoff in the use of acoustic-phonetic features such that, for example, the presence of a certain cue to the identity of a word precludes the presence of a different cue to word identification. If such a tradeoff is observed, it would indicate that the goal of communicative success powerfully determines the implementation of a phonological system. On the other hand, mappings between acoustic-phonetic features and phonemic categories may be largely insensitive to communicative context, such that it would be easily possible for speakers to "overdetermine" the identity of a word through the use of multiple signals to its identity.

It would also be fruitful to consider the role of L2 speaker experience in more specific terms than speaker age. The current study assumes that increased L2 speaker age involves increased experience with native L1 input so that older L2 speakers have more opportunity than younger L2 speakers to perceive differences between L2 and L1 phones. This may have occurred in the present study through older Cantonese speakers' increased experience with English TV, songs, or media as compared with younger Cantonese speakers. However, future study should more carefully measure and differentiate between experience with native L1 input and experience with nonnative L1 input, which was not controlled for here.

To further investigate whether L2 listeners use their L1 acoustic-phonetic features when they listen to a second language, in the future, we will conduct acoustic analyses of the native English speech used in the present studies to determine the extent to which L2 listeners' perceptual accuracy reflect the perception and use of aspiration versus durational cues. For example, if L2 listeners always perceive aspirated English onsets as unvoiced and unaspirated onsets as voiced regardless of the VOT, this would be consistent with the Feature Adjustment hypothesis, whereas if both VOT and aspiration play a role, this would be consistent with the Hybrid Phonotactic Constraints hypothesis.

Additionally, the L1 Cantonese participants' data analyzed in this thesis were the pre-training data for adolescents and adults in a training study by Terry Au (ms). If the Speech Learning Model view of the role of input is correct, then we would expect that both adolescents and adults would benefit from training and would have better post-training perception of voiced codas. If, on the other hand, the Critical or Sensitive Period Hypothesis is correct, we would expect that training would be futile for the adults who were well beyond the sensitive period. Adolescents, on the other hand, might show modest benefits from training. In a future study, we will address the role of age and input by analyzing the degree of coda improvement in adolescents and adults.

The original L1 Cantonese data set also contained data from 9 and 10 year old children. In a future study, we will use the data from these younger children to further investigate the Input and Sensitive Period hypotheses. Some researchers (e.g., Lenneberg, 1969) have argued that native-like second language learning strictly ends at puberty. If this is correct, we would expect that only pre-pubescent children would benefit from training. If, on the other hand, puberty is a soft border, then the younger children would benefit the most from the training procedure, followed by the older children, with the adults showing little or no improvement. If, on the other hand, the SML input hypothesis is correct, we would expect all 3 groups to benefit equally from the training.

Finally, in the L1 Cantonese training study, participants only received perceptual training and no training on producing voiced plosives. We could further investigate the relationship between speech production and perception by looking to see whether only perceptual training on plosives leads to improvements in plosive production. Affirmative results would be inconsistent with MT but consistent with SLM.

## Acknowledgements

For the portion of the study conducted in Hong Kong, I would like to thank the participants, parents, schools, and the Research Grants Council of Hong Kong (HKU740511H) for their support of this research. For the portion conducted in the US, thanks to the participants, members of the Language Acquisition and Processing Lab, the

Aresty Foundation, and the Honors College of Rutgers, New Brunswick for their support.

In particular, I would like to express my sincerest gratitude to my advisor, Karin Stromswold, for teaching and guiding me through all stages of this research.

I would also like to thank Ana Rinzler for advising and training me in acoustic analysis using Praat. My sincere thanks go also to Melchi Michel for teaching me to use R, and to David Kleinschmidt for advice and comments on analysis.

All my love and thanks to my family. I couldn't have written this thesis without you.

# Figures

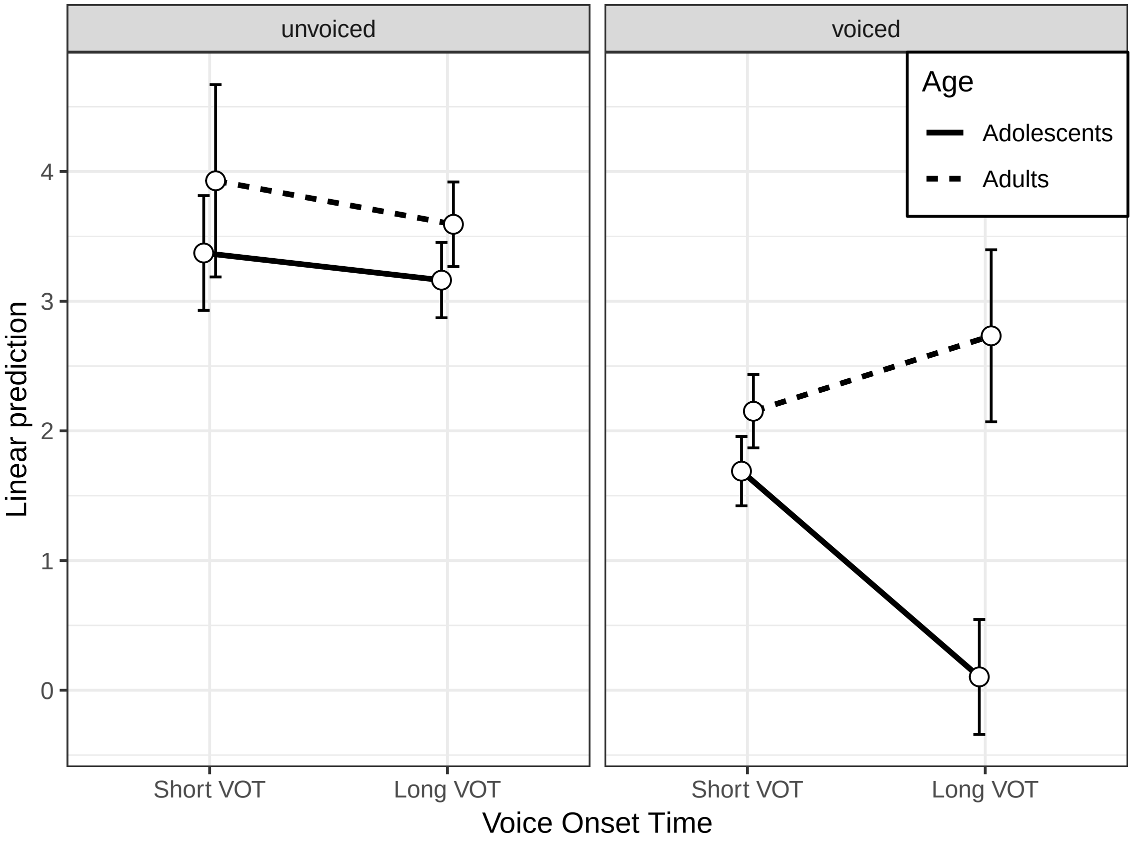


*Figure 16*. Model Predictions for L1 Listeners' Accuracy for L2 Speakers' Onset Words (all trials). Each point on the graph indicates mean predicted accuracy in log-odds for each combination of the levels of voicing and speaker age. Bars on each point indicate standard errors on mean accuracy. Since durational feature is a continuous predictor without levels, predictions are reported for points that are a standard deviation above and below the mean durational feature.

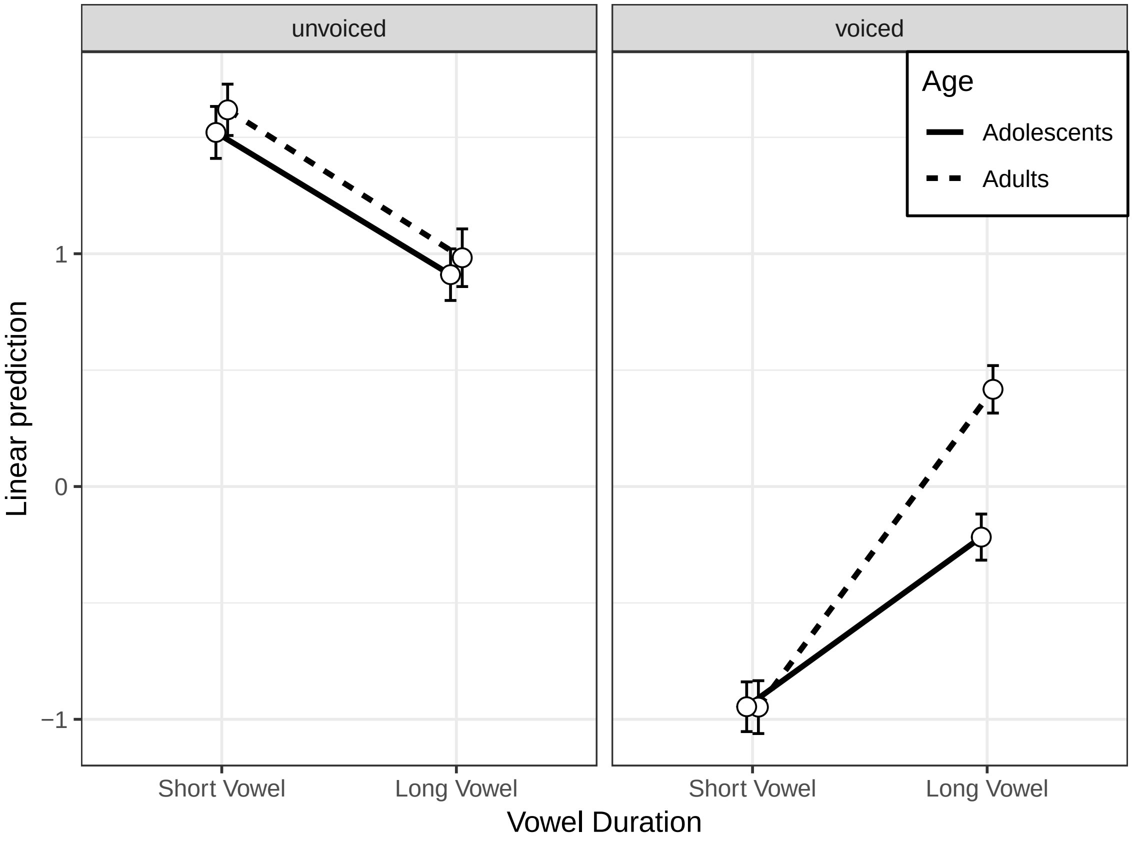


*Figure 17.* Model Predictions for L1 listeners' Accuracy for L2 Speakers' Coda Words (all trials).

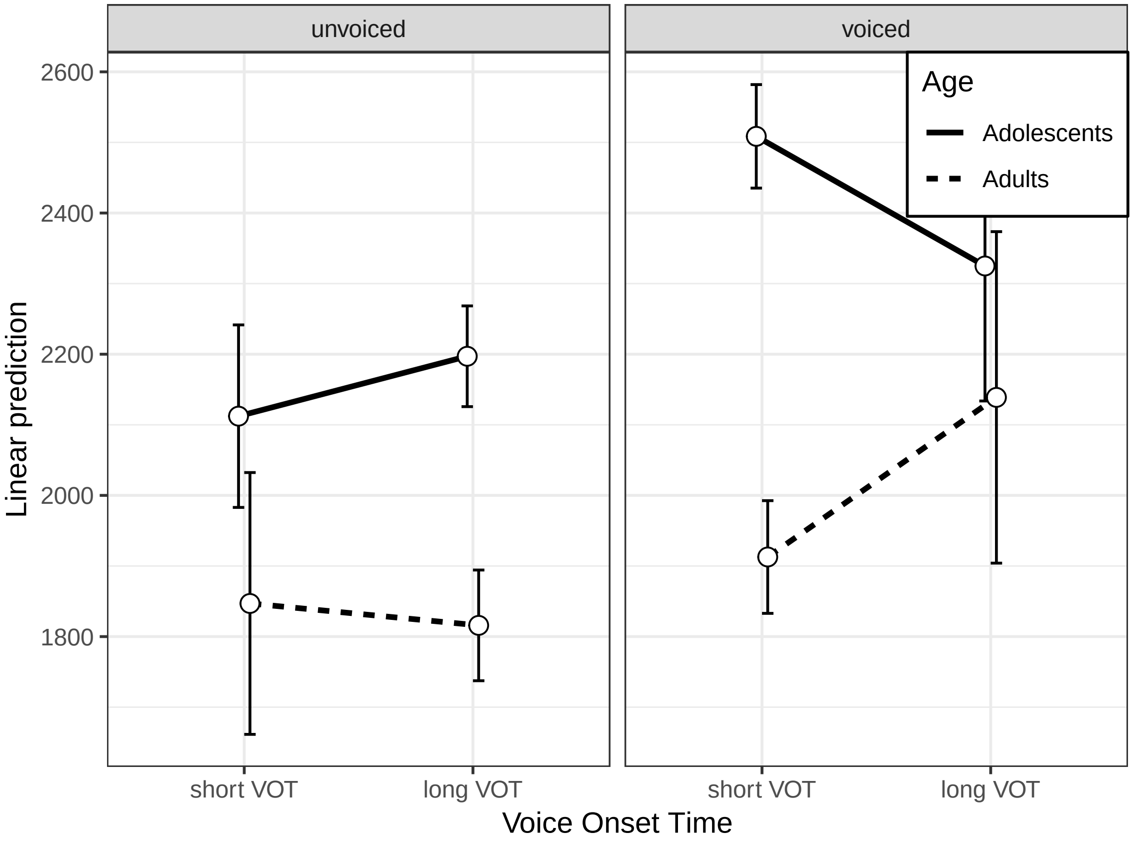


*Figure 18*. Model Predictions for L1 listeners' RTs for L2 Speakers' Onset Words (all trials)

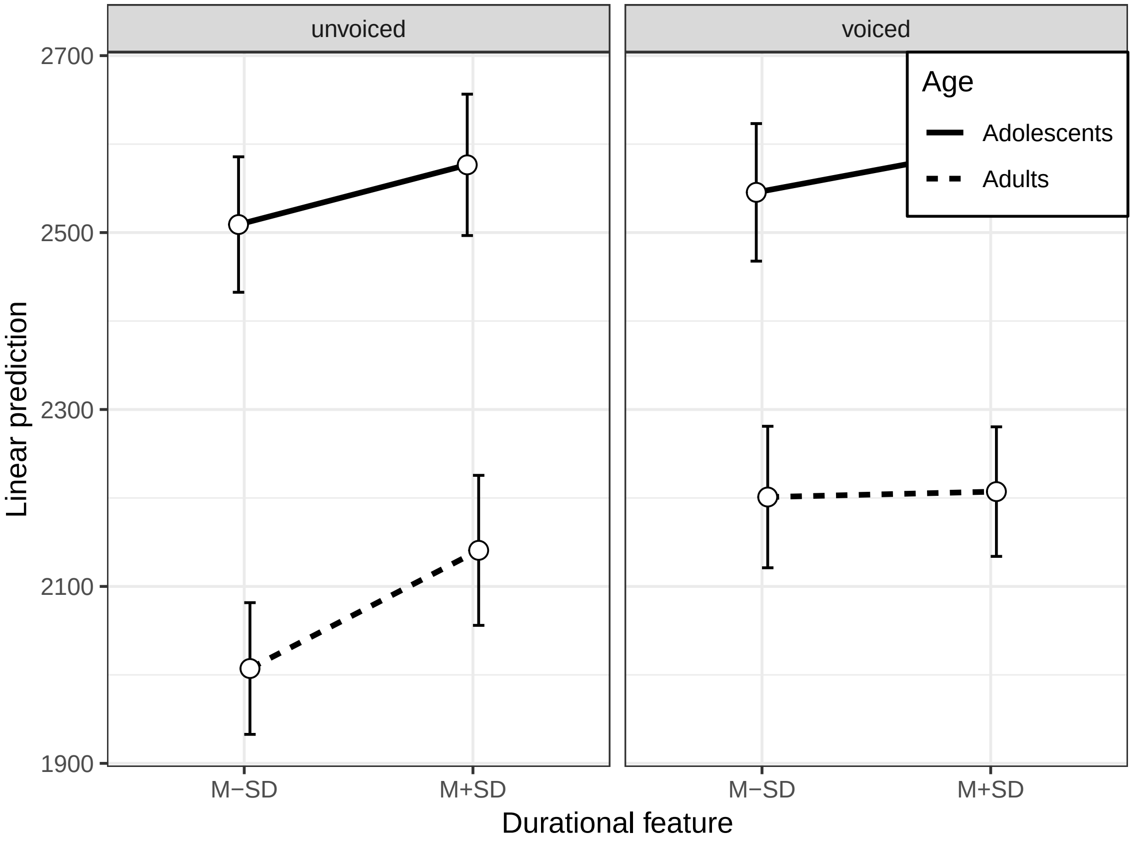


*Figure 19*. Model Predictions for L1 Listeners' RTs for L2 Speakers' Coda Words (all trials)

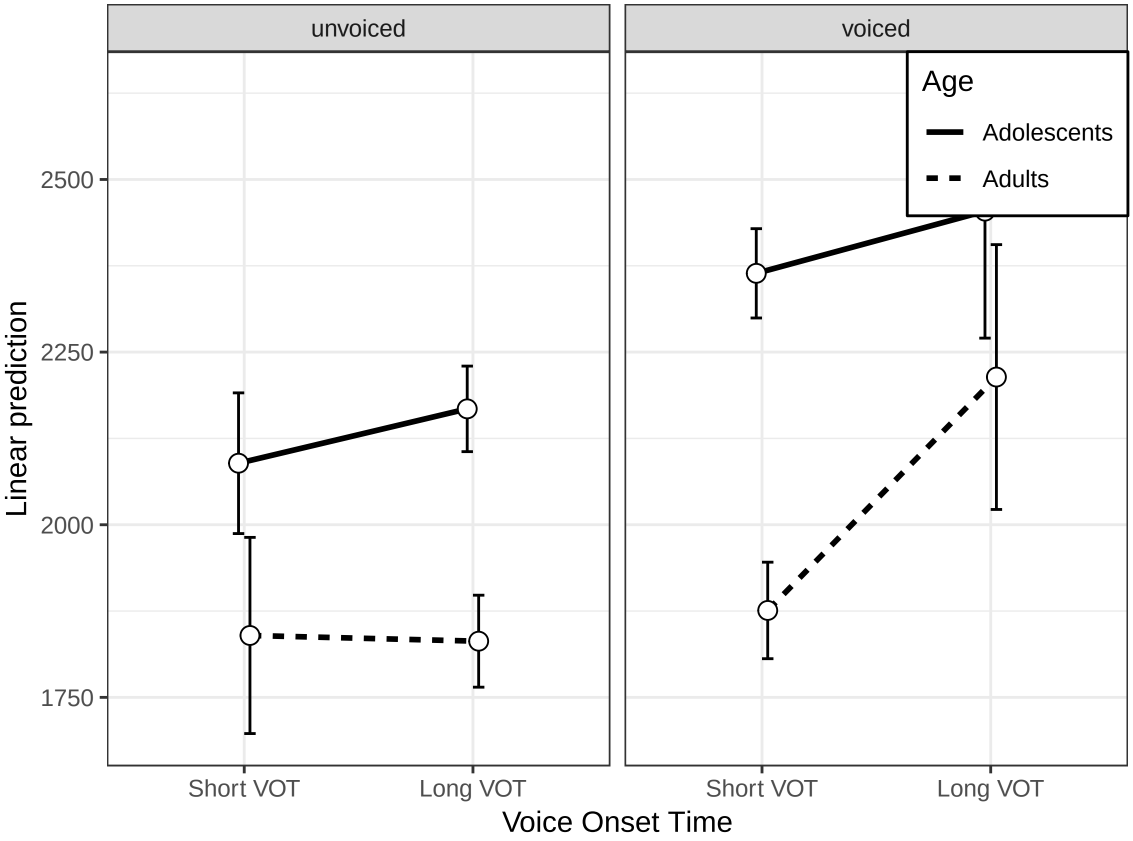


*Figure 20.* Model Predictions for L1 Listeners' RTs for L2 Speakers' Onset Words (correct trials only)

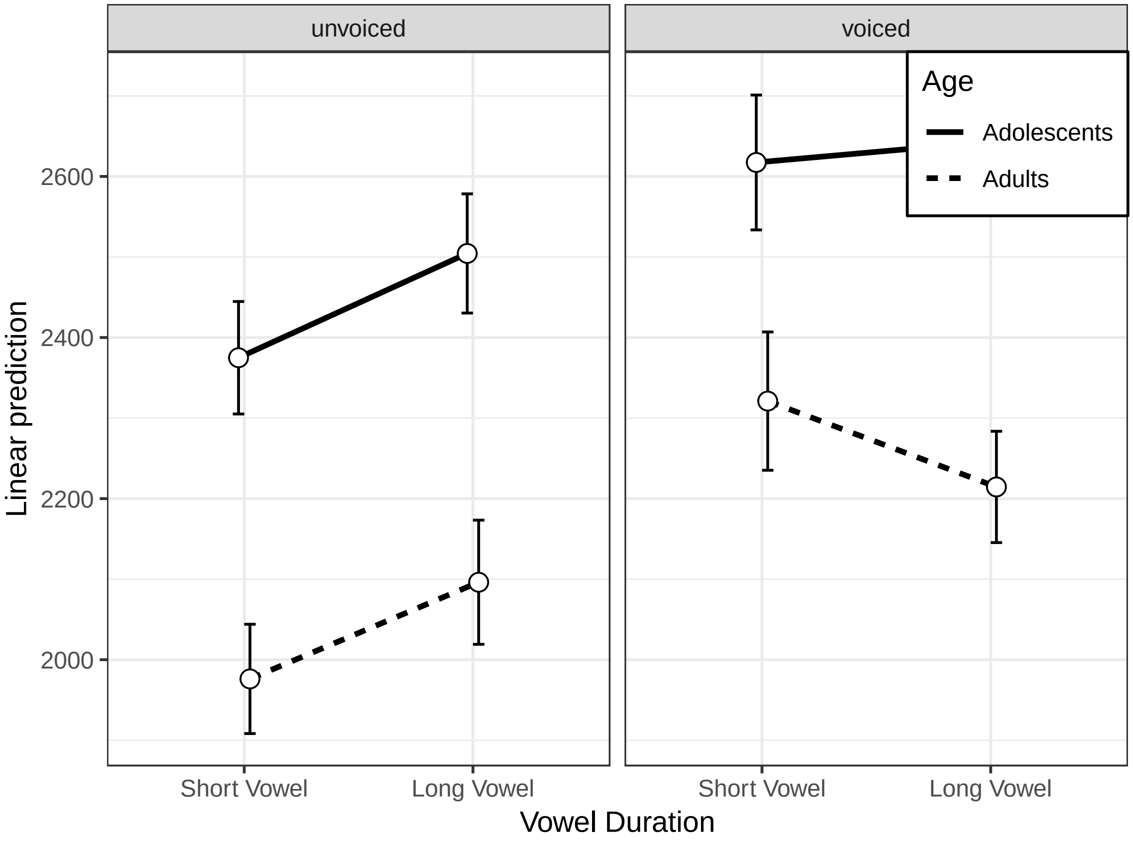


*Figure 21*. Model Predictions for L1 Listeners' RTs for L2 Speakers' Coda Words (correct trials only)

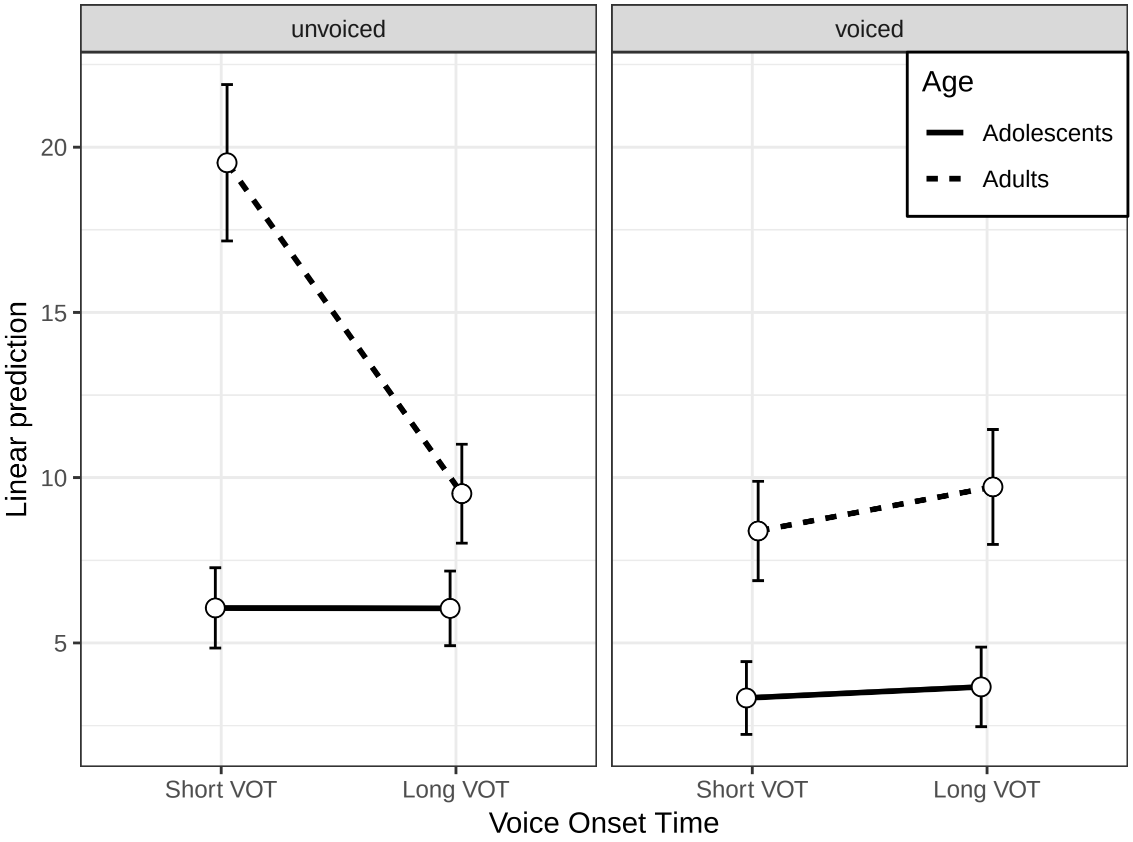


*Figure 22.* Model Predictions for L2 Perceptual Accuracy of L1 Onset Words

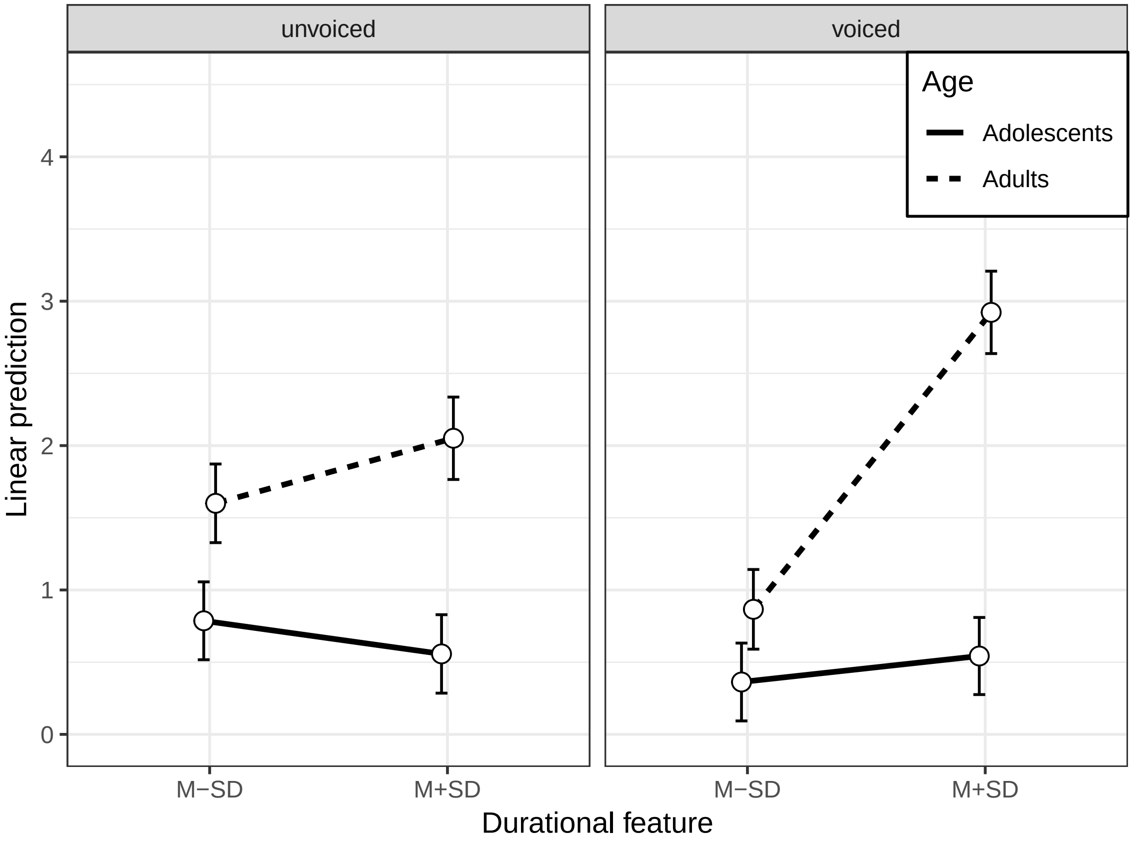


*Figure 23.* Model Predictions for L2 Perceptual Accuracy of L1 Coda Words

## References

Albrechtsen, D., Henriksen, B., & Faerch, C. (1980). NATIVE speaker reactions to learners'SPOKEN interlanguage 1. *Language Learning*, *30*(2), 365–396.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*(4), 561–613.

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentais, prosody, and syllable structure. *Language Learning*, *42*(4), 529–555.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi:10.18637/jss.v067.i01

Bent, T., & Bradlow, A. R. (2003). The interlanguage speech intelligibility benefit. *The Journal of the Acoustical Society of America*, *114*(3), 1600–1610.

Boersma, P., & Weenink, D. (2018). Praat: Doing phonetics by computer. version 6.0. 37.

Bradlow, A. R., & Bent, T. (2003). Listener adaptation to foreign-accented english. In *Proceedings of the 15th international congress of phonetic sciences* (pp. 2881–2884). Universitat Autònoma de Barcelona Barcelona.

Chan, A. Y., & Li, D. C. (2000). English and cantonese phonology in contrast: Explaining cantonese esl learners' english pronunciation problems. *Language Culture and Curriculum*, *13*(1), 67–85.

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented english. *The Journal of the Acoustical Society of America*, *116*(6), 3647–3658.

Fayer, J. M., & Krasinski, E. (1987). Native and nonnative judgments of intelligibility and irritation. *Language Learning*, *37*(3), 313–326.

Flege, J. E. (1988). Factors affecting degree of perceived foreign accent in english sentences. *The Journal of the Acoustical Society of America*, *84*(1), 70–79.

Flege, J. E. (1991). Perception and production: The relevance of phonetic input to l2 phonological learning. *Crosscurrents in Second Language Acquisition and Linguistic Theories*, *2*, 249–289.

Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. *Speech Perception and Linguistic Experience.*

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. *Phonetics and Phonology in Language Comprehension and Production: Differences and Similarities*, *6*, 319–355.

Flege, J. E. (2018, July). The speech learning model (slm) account of how japanese speakers learn english /r/ and /l/. Sophia University, Tokyo; http://jimflege.com/files/Sophia_rl_talk5.pdf.

Flege, J. E., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in english sentences produced by korean children and adults. *Journal of Phonetics*, *34*(2), 153–175.

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, *13*(3), 361–377.

Gass, S., & Varonis, E. M. (1984). The effect of familiarity on the comprehensibility of nonnative speech. *Language Learning*, *34*(1), 65–87.

Hayes-Harb, R., Smith, B. L., Bent, T., & Bradlow, A. R. (2008). The interlanguage speech intelligibility benefit for native speakers of mandarin: Production and perception of english word-final voicing contrasts. *Journal of Phonetics*, *36*(4), 664–679.

Jaeger, T. F. (2013). Production preferences cannot be understood without reference to communication. *Frontiers in Psychology*, *4*, 230.

Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of english as a second language. *Cognitive Psychology*, *21*(1), 60–99.

Kennedy, S., & Trofimovich, P. (2008). Intelligibility, comprehensibility, and accentedness of l2 speech: The role of listener experience and semantic context. *Canadian Modern Language Review*, *64*(3), 459–489.

Kleinschmidt, D. F., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. doi:10.18637/jss.v082.i13

Lambert, W. E. (1967). A social psychology of bilingualism. *Journal of Social Issues*, *23*(2), 91–109.

Lenneberg, E. H. (1967). The biological foundations of language. *Hospital Practice*, *2*(12), 59–67.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1–36.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*(6), 431.

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word*, *20*(3), 384–422.

McLennan, C. T., & González, J. (2012). Examining talker effects in the perception of native-and foreign-accented speech. *Attention, Perception, & Psychophysics*, *74*(5), 824–830.

Munro, M. J., & Derwing, T. M. (1995a). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97.

Munro, M. J., & Derwing, T. M. (1995b). Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech*, *38*(3), 289–306.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of l2 speech the role of speaking rate. *Studies in Second Language Acquisition*, *23*(4), 451–468.

Munro, M. J., Derwing, T. M., & Morton, S. L. (2006). The mutual intelligibility of l2 speech. *Studies in Second Language Acquisition*, *28*(1), 111–131.

Oyama, S. (1976). A sensitive period for the acquisition of a nonnative phonological system. *Journal of Psycholinguistic Research*, *5*(3), 261–283.

Pajak, B., Fine, A. B., Kleinschmidt, D. F., & Jaeger, T. F. (2016). Learning additional languages as hierarchical probabilistic inference: Insights from first language processing. *Language Learning*, *66*(4), 900–944.

Pallier, C. (2007). Critical periods in language acquisition and language attrition. *Language Attrition: Theoretical Perspectives*, 155–168.

Penfield, W. (1965). Conditioning the uncommitted cortex for language learning. *Brain*, *88*(4), 787–798.

R Core Team. (2018). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raphael, L. J. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in american english. *The Journal of the*

*Acoustical Society of America*, *51*(4B), 1296–1303.

Sakai, M., & Moorman, C. (2018). Can perception training improve the production of second language phonemes? A meta-analytic review of 25 years of perception training research. *Applied Psycholinguistics*, *39*(1), 187–224.

Selinker, L. (1972). Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching*, *10*(1-4), 209–232.

# Appendix

## Stimuli (adapted from Terry Au, ms)

Each stimuli word was recorded by speaker participants embedded within an “I say . . . ” phrase such as “I say pack”. Listener participants heard the full phrase.

### Phonological pair: /p/ - /b/

- with the plosives in onset position
  - pack; back
  - pay; bay
  - peak; beak
  - pet; bet
  - pill; bill
- with the plosives in coda position
  - cap; cab
  - lap; lab
  - nip; nib
  - rope; robe
  - cop; cob
  - cup; cub
  - mop; mob
  - nap; nab
  - rip; rib
  - tap; rab

### Phonological pair: /t/ - /d/

- with the plosives in onset position

  - teal; deal
  - tie; die
  - time; dime
  - tone; done
  - tuck; duck

- with the plosives in coda position
  - bet; bed
  - bit; bid
  - fat; fad
  - fate; fade
  - bat; bad
  - coat; coda
  - feet; feed
  - got; god
  - mat; mad
  - not; nod

**Phonological pair: /k/ - /g/**

- with the plosives in onset position
  - cane; gain
  - cap; gap
  - coat; goat
  - con; gone
  - cot; got
- with the plosives in coda position
  - dock; dog

- duck; dug
- jock; jog
- peck; peg
- back; bag
- lock; log
- muck; mug
- pick; pig
- rack; rag
- tack; tag