Subsegmental Representation in Child Speech Production:

Structured Variability of Stop Consonant Voice Onset Time in American English and Cantonese

Eleanor CHODROFF[1*], Leah BRADSHAW[2], and Vivian LIVESAY[3]

[1]*University of York, Department of Language and Linguistic Science, Heslington, York, UK YO10 5DD*

[2]*University of Zurich, Institute of Computational Linguistics, Andreasstrasse 15, 8050 Zurich, Switzerland*

[3]*Mount Holyoke College, Department of Psychology and Education, 50 College Street, South Hadley, Massachusetts, USA 01075*

*Author to whom correspondence should be addressed. Electronic mail:

eleanor.chodroff@york.ac.uk

Acknowledgments

We gratefully acknowledge Colin Wilson for early discussion and contributions related to this work. Many thanks to Alessandra Golden for help in processing the English VOT data, as well as Justin Lo and Adas Li for help with Cantonese transliteration and translation. We are also grateful to Emily Atkinson and Marilyn Vihman for helpful discussion and feedback, and to Mara Breen for coordinating remote research opportunities for Mt. Holyoke College undergraduate students. Finally, we thank the Mt. Holyoke's Lynk Summer Funding Program for financially supporting the third author for work on this project.

Abstract

Voice onset time (VOT) of aspirated stop consonants is marked by variability and systematicity in adult speech production. The present study investigated variability and systematicity of voiceless aspirated stop VOT from 161 two- to five-year-old talkers of American English and Cantonese. Child VOT was more variable than adult VOT, but VOT means were comparable between children and adults. Additional aspects of child VOT structure parallel adult patterns and inform our understanding of early speech production. Across children in both languages, child-specific VOT means were strongly correlated between [t^h] and [k^h]. This correlation has previously been observed in adult VOT and may reflect a constraint of "target uniformity" that minimizes variation in the phonetic realization of a shared distinctive feature. The findings suggest that target uniformity is not merely a product of a mature grammar, but may instead shape speech production representations in children as young as two years of age.

Introduction

Child speech is marked by considerable variability at the phonetic and phonological levels, and in many ways, deviates from adult speech patterns. At the phonetic level, differences in anatomy and motor control between children and adults have sizable influences on the phonetic output (e.g., Eguchi & Hirsch, 1969; Koenig, 2000; Macken & Barton, 1980; Smith & Kenney, 1998). At the phonological level, child word forms differ from those of adults with segmental and featural modifications. Processes including segment deletion, insertion, and metathesis are quite commonplace; subsegmental changes such as denasalization, vocalization, fronting, velarization, and stopping, among others, have also been attested (e.g., Barlow & Gierut, 1999; Haelsig & Maddison, 1986; Smith, 1973). Nevertheless, aspects of these levels and the interface between them may still resemble the mature grammar. In the present study, we investigate the extent to which child speech patterns parallel adult speech patterns with a focus on aspirated stop consonant voice onset time (VOT) in two- to five-year-old talkers of American English and Cantonese, and discuss the implications of these findings for the development of speech production representations.

VOT Production in English and Cantonese

Across languages, VOT is the primary correlate to the stop voicing contrast (Lisker & Abramson, 1964). In word-initial position, American English and Cantonese have voiceless aspirated stops with long-lag VOT at three primary places of articulation. Phonologically, English aspirated stops are argued to be phonological surface segments of underlying /p t k/ which contrast with /b d g/.[1] Cantonese aspirated stops are argued to be phonological surface

[1] Voiceless aspirated stops have ‘long-lag’ VOT, in which the interval from the stop release to the onset of voicing is delayed. Voiceless unaspirated stops have ‘short-lag’ VOT, in which the interval is short, and phonetically voiced stops have ‘lead’ or negative VOT, in which phonetic voicing precedes the stop release. American English /b d g/ are most frequently produced with short-lag VOT, but can also be produced with lead VOT, especially in certain dialects (Hunnicutt & Morris, 2016).

segments of underlying /p^h t^h k^h/ which contrast with /p t k/.[2] In English, the phonetic places of articulation are bilabial, alveolar and velar, and in Cantonese, the places of articulation are bilabial, dental/alveolar, and velar. For simplicity, we refer to the three places as labial, coronal, and dorsal, and use the square-bracketed [p^h t^h k^h] to refer to the surface segments present in both languages. Previous studies have reported highly comparable VOT values between American English and Cantonese aspirated stops in adult speech, despite the fact that the two languages are otherwise typologically distinct (see Table 1).

[2] The dental/alveolar place of articulation has an additional contrast between an aspirated and plain affricate (/$\widehat{ts}^h$/ vs /$\widehat{ts}$/), and the velar place of articulation has additional contrast between aspirated and plain labialized voiceless stops (/k^{wh}/ vs /k^w/). As these latter two contrasts involve secondary articulations and are not found in English, they are not included in the present study, but should be investigated in future research.

Table 1. Previously reported VOT values in milliseconds from adult talkers of English and Cantonese. M refers to the mean, SD to the standard deviation, R_M to the range of talker means, R_{SD} to the range of talker standard deviations, and R_I to the range of individual VOT values.

	Study	[p^h]	[t^h]	[k^h]	Notes
American English: Adult	Lisker & Abramson (1964)	M = 58 R_I = 20:120	M = 70 R_I = 30:105	M = 80 R_I = 50:135	4 talkers, isolated speech
	Zlatin & Koenigsknecht (1976)	M = 84 SD = 25	M = 87 SD = 26	M = 91 SD = 24	20 talkers, isolated speech
	Koenig (2000)	M = 45 SD = 12 R_M = 34:80 R_{SD} = 7:16	M = 56 SD = 11 R_M = 36:68 R_{SD} = 7:13	–	14 talkers, each grand mean and SD calculated from reported talker means, isolated speech (with short carrier phrase)
	Yao (2007)	M = 41 SD = 25	M = 51 SD = 28	M = 58 SD = 26	19 talkers, spontaneous speech
	Chodroff & Wilson (2017)	M = 89 SD = 27 R_M = 56:139 R_{SD} = 12:27	M = 98 SD = 28 R_M = 57:156 R_{SD} = 10:26	M = 99 SD = 24 R_M = 67:137 R_{SD} = 11:20	24 talkers, isolated speech (with short carrier phrase)
	Chodroff & Wilson (2017)	M = 51 SD = 9 R_M = 28:78 R_{SD} = 11:35	M = 61 SD = 9 R_M = 40:96 R_{SD} = 9:34	M = 56 SD = 8 R_M = 36:79 R_{SD} = 11:30	180 talkers, connected speech
Cantonese: Adult	Lisker & Abramson (1964)	M = 77 R_I = 35:110	M = 75 R_I = 45:95	M = 87 R_I = 70:115	1 talker, isolated speech
	Clumeck, Barton, Macken, & Huntington (1981)	M = 74 SD = 20 R_I = 30:144	M = 83 SD = 21 R_I = 43:160	M = 91 SD = 22 R_I = 52:162	8 talkers, isolated speech
	Lee Oi Yee (1997)	M = 68	M = 73	M = 74	10 talkers, means estimated from figure, isolated speech

Empirical evidence suggests that most English and Cantonese-speaking children acquire aspirated stops around two to three years of age (ENGLISH: Barton, 1976; Gilbert, 1977; Macken & Barton, 1980; Major, 1976; Kewley-Port & Preston, 1974; Zlatin & Koenigsknecht, 1976; CANTONESE: Stokes & To, 2002). Similar findings have also been reported for children of other languages with a voiceless aspirated stop series (e.g., THAI: Gandour, Petty, Dardarananda, Dechongkit, & Mukngeon, 1986; MANDARIN: Yang, 2018). Child production of VOT in voiceless aspirated stops is, however, not immediately adult-like.

Some studies report longer VOT values for children than adults, suggesting an 'overshoot' phase (Barton & Macken, 1980; Gilbert, 1977; Lee Oi Yee, 1997; Millasseau, Bruggeman, Yuen, & Demuth, 2021), whereas others report shorter VOT values for children than adults (e.g., Kewley-Port & Preston, 1974; Macken & Barton, 1980; Zlatin & Koenigsknecht, 1976). Regardless of the relative direction, child VOT is consistently characterized by greater variability than adult productions (Barton & Macken, 1980; Clumeck et al., 1981; Eguchi & Hirsch, 1969; Koenig, 2000; Yang, 2018; Zlatin & Koenigsknecht, 1976).

Increased variability in child VOT likely arises from underdeveloped laryngeal control, which refers to the management of glottal abduction degree, vocal fold tension, as well as transglottal pressure and flow (Koenig, 2000). Moreover, these skills may not fully mature until mid-adolescence (Koenig, 2000; Redford, 2019; Smith & Zelaznik, 2004). Indeed, talker-specific VOT standard deviations of English aspirated stops are almost twice as large for five-year-old talkers than for adult talkers, and similar findings have been observed for Cantonese-learning children (see Table 2).[3]

[3] The laryngeal-control explanation stands in contrast to a potential "complex timing" argument, in which children may struggle to time the articulations of the necessary oral and laryngeal gestures. Koenig (2000) identified that variability in duration increased not only for stop consonants, but also [h], which does not involve an oral articulation, thereby contradicting the complex timing argument.

Table 2. Previously reported VOT values in milliseconds from child talkers of English and Cantonese. M refers to the mean, SD to the standard deviation, R_M to the range of talker means, R_{SD} to the range of talker standard deviations, and R_I to the range of individual VOT values.

	Study	[p^h]	[t^h]	[k^h]	Notes
American English: Children	Zlatin & Koenigsknecht (1976)	M = 65 SD = 35	M = 69 SD = 34	M = 80 SD = 34	10 two-year-olds, isolated speech
	Barton & Macken (1980)	M = 66 SD = 31 R_M = 32:93 R_{SD} = 13:15	M = 90 SD = 25 R_M = 61:109 R_{SD} = 13:15	M = 114 SD = 19 R_M = 92:126 R_{SD} = 13:15	3 two-year-olds, grand means and SDs calculated from reported talker means multiple speech styles
	Barton & Macken (1980)	M = 84 SD = 13 R_M = 72:98 R_{SD} = 2:11	M = 131 SD = 40 R_M = 99:176 R_{SD} = 2:6	M = 136 SD = 13 R_M = 127:145 R_{SD} = 7:7	3 two-year-olds, grand means and SDs calculated from reported talker means isolated speech
	Barton & Macken (1980)	M = 75 SD = 36 R_I = 3:174	M = 79 SD = 44 R_I = 4:236	M = 86 SD = 39 R_I = 27:224	4 four-year-olds, multiple speech styles
	Barton & Macken (1980)	M = 83 SD = 38	M = 81 SD = 25	M = 91 SD = 33	4 four-year-olds, isolated speech
	Koenig (2000)	M = 53 SD = 21 R_M = 27:104 R_{SD} = 12:29	M = 58 SD = 22 R_M = 39:108 R_{SD} = 13:32	—	7 five-year-olds, isolated speech (with short carrier phrase)
Cantonese: Children	Clumeck et al. (1981)	M = 27 SD = 22 R_I = 5:49	M = 52 SD = 41 R_I = 10:113	M = 54 SD = 17 R_I = 38:80	1 two-and-a-half-year-old (first), isolated speech
	Clumeck et al. (1981)	M = 50 SD = 27 R_I = 18:84	M = 85 SD = 54 R_I = 20:252	M = 51 SD = 28 R_I = 4:102	1 two-and-a-half-year-old (second), isolated speech
	Clumeck et al. (1981)	M = 132 SD = 24 R_I = 102:160	M = 87 SD = 34 R_I = 48:108	M = 170 SD = 75 R_I = 98:264	1 three-year-old, isolated speech
	Clumeck et al. (1981)	M = 60 SD = 35 R_I = 21:155	M = 76 SD = 24 R_I = 45:134	M = 77 SD = 32 R_I = 47:148	1 four-year-old, isolated speech
	Lee Oi Yee (1997)	M = 59	M = 43	M = 68	10 two-year-olds, means estimated from figure, isolated speech
	Lee Oi Yee (1997)	M = 77	M = 79	M = 92	10 three-year-olds, means estimated from figure, isolated speech
	Lee Oi Yee (1997)	M = 77	M = 73	M = 88	10 four-year-olds, means estimated from figure, isolated speech

Structured VOT Variability

While VOT varies extensively across talkers, it is also highly structured among stop categories, at least in adult speech production. We discuss three primary types of structure observed in adult VOT—between-category VOT structure, within-category VOT structure, and "contextual" VOT structure—and present our proposed analysis and predictions for each type of structure with respect to child speech production. In particular, we investigate this structure in a large spoken corpus with 161 two- to five-year-olds in American English (81 talkers) and Cantonese (80 talkers). In children, VOT structure may need to develop over time as both the motor and linguistic systems mature. In this case, we would expect a lower degree of VOT structure in children than adults. Alternatively, children may already possess some *mature* phonetic representations despite limitations in motor control. In this case, we would expect a comparable degree of VOT structure in children as in adults.

Between-category VOT Structure

In many, if not all languages, the mean VOT of /p/ is lower than that of /k/, reflecting a consistent ordinal relationship (e.g., Cho & Ladefoged, 1999; Chodroff, Golden, & Wilson, 2019; Lisker & Abramson, 1964). The mean VOT of /t/ is frequently intermediate to that of /p/ and /k/, though some variability in this ranking is not uncommon (e.g., Chodroff & Wilson, 2017; Docherty, 1992; Gandour et al., 1986; Millasseau, Bruggeman, Yuen, & Demuth, 2019; Stuart-Smith, Sonderegger, Rathcke, & Macdonald, 2015; Suomi, 1980). Moreover, language- and talker-specific mean VOTs are strongly correlated among stop categories with a shared laryngeal feature, particularly among aspirated stops. Across 68 languages, the correlation of language-specific mean VOTs between [p^h] and [t^h] was 0.85, between [t^h] and [k^h] 0.81, and between [t^h] and [k^h] 0.82 (Chodroff et al., 2019). Among

American English aspirated stops, correlations of talker-specific mean VOT exceeded 0.90 in isolated speech, 0.75 in connected speech, and 0.81 in spontaneous speech (Chodroff & Wilson, 2017). These findings indicate that in general, a talker with a long VOT for [p^h] will also have a long VOT for both [t^h] and [k^h], reflecting a consistent and strong *linear* relationship of VOT between differing places of articulation (e.g., Chodroff & Wilson, 2017; Koenig, 2000; Newman, 2003; Theodore et al., 2009; Zlatin, 1974).

Why might adults be so consistent in their production of VOT across place of articulation? One explanation is that talkers have highly similar, or even identical, phonetic targets for the laryngeal realization of [p^h], [t^h], and [k^h]. This constraint on phonetic realization is referred to as *target uniformity* (Chodroff & Wilson, 2022), and it maximizes identity of phonetic targets (i.e., the motor or auditory goals for a speech sound) that correspond to a given distinctive feature value, such as [+spread glottis] (Chodroff & Wilson, 2017; see also Schwartz, Boë, & Abry, 2007). For aspirated stop consonants, the phonetic targets corresponding to [+spread glottis] likely involve the glottal spreading gesture and its timing relative to the oral release. VOT would then be the measurable acoustic output of the phonetic target, and not the target itself. Theoretically, an individual could have a unique laryngeal phonetic target for each of [p^h], [t^h], and [k^h]. It is physically possible for a talker to specify [k^h] with a phonetic target that has a shorter mean VOT than that for [p^h], but across adult talkers, this rarely, if ever, happens. The posited uniformity constraint restricts such arbitrary variation in the phonetic target specification, and instead requires near identity across the laryngeal phonetic targets for [p^h], [t^h], and [k^h].

The extent to which such between-category VOT structure applies to child speech production has been scarcely investigated. Infants' developing motor systems do indeed result in greater VOT variability, but despite this immature laryngeal control, between-category VOT structure might also be present in child speech. Whalen, Levitt, & Goldstein

(2007) found the expected VOT ranking in unaspirated stops produced by English- and French-babbling infants (9 and 12 months of age), but the infants had not yet acquired aspiration, which adds a layer of complexity to articulation. In addition, Koenig (2000) found significant correlations of talker-specific VOT medians and maxima between [p^h] and [t^h] across five-year-old and adult talkers (median VOT $r = 0.78$; maximal VOT $r = 0.79$, each $p < 0.05$), but the correlation across five-year-olds alone was not reported. To our knowledge, no study has yet investigated the linear relationship between aspirated stops across children younger than five.

To investigate between-category structure in children, we assess the linear and ordinal relationships *between* talker-specific mean VOTs of [t^h] and [k^h]. A strong linear relationship suggests that the phonetic targets for the laryngeal realization of [t^h] and [k^h] are not independent of each other, but rather yoked together—even with a developing motor system. The ordinal relationship is also reported for comparison with previous literature.

Within-category VOT Structure

In addition to between-category VOT structure, within-category VOT structure has been observed in the relationship between talker-specific means and standard deviations: adult talkers with longer VOTs typically also have a higher variance (Chodroff &Wilson, 2017). This relationship likely reflects a general motor principle which relates increased movement time to increased movement error (Schmidt, Zelaznik, Hawkins, Frank, & Quinn, 1979; Turk & Shattuck-Hufnagel, 2014). It could be that children are so variable in speech production that no clear relationship emerges between the VOT mean and standard deviation. To investigate within-category structure, we examine the extent to which a talker's mean VOT is correlated with its standard deviation *within* a stop category: a phonetic target that corresponds to a long VOT should also have a high standard deviation. A correlation here

would suggest that the motor movement principle is present early in child speech production; a child aiming for phonetic targets with a more delayed voice onset will also have more variable realizations.

Contextual VOT Structure

VOT is subject to substantial variability both within and across talkers. VOT variability is structured by prosodic, environmental, and social factors (e.g., Cole, Choi, Kim, & Hasegawa, 2003; Mielke & Nielsen, 2018; Yao, 2009). VOT varies systematically by place of articulation (Fischer-Jørgensen, 1954), following vowel quality (Klatt, 1975; Nearey & Rochet, 1994), speaking rate (Miller, Green, & Reeves, 1986; Theodore et al., 2009), prosodic domain (Cho & Keating, 2009), lexical properties (e.g., Goldinger & Van Summers, 1989; Yao, 2009), age (e.g., Koenig, 2000; Stuart-Smith et al. 2015), gender (e.g., Koenig, 2000; Swartz, 1992), and across talkers, particularly for aspirated stops. Some talkers have characteristically long VOTs, whereas others have characteristically short VOTs (Chodroff & Wilson, 2017; Docherty, 1992; Theodore, Miller, & DeSteno, 2009). Among children, the influence of place of articulation is generally similar in direction to adult talkers (Zlatin & Koenigsknecht, 1976; Whalen et al., 1997), but as far as we are aware, many other factors have not yet been explored.

To investigate contextual VOT structure, we assess the factors governing variability in VOT for word-initial, prevocalic [t^h] and [k^h], including place of articulation, following vowel context, speaking rate, number of syllables, talker age, and talker gender. We expect these factors to have similar effects on VOT realization in children as in adults.

Present study

The present study investigates the degree to which two- to five-year-old children demonstrate adult-like VOT structure in the voiceless aspirated stop series of American English and Cantonese. Considerable evidence now supports the presence of VOT structure in adult speech within and between stop categories (Chodroff & Wilson, 2017, 2022; Chodroff et al., 2019; Hullebus, Tobin, & Gafos, 2018; Johnson, 2021; Puggaard & Goldshtein, 2020; Tanner, Sonderegger, & Stuart-Smith, 2020).[4] Identifying the extent of this structure in child speech informs our understanding of the acquisition and cognitive representation of phonetic targets in language production: structure may reveal maturity in the representation despite relatively immature motor control.

The study repurposes American English and Cantonese speech data from the Paidologos Project on Cantonese, English, Greek, and Japanese child speech production (Edwards & Beckman, *n.d.*). American English and Cantonese were selected primarily based on data availability and because they each have a voiceless aspirated stop series; they are otherwise typologically distinct. For both American English and Cantonese, we investigate the extent of variability and structure in [t^h] and [k^h] VOT as produced by two- to five-year-old talkers. Despite the presence of [p^h] in both languages, we focus on [t^h] and [k^h] as the Paidologos corpora included lingual consonants only. We examine the range of VOT means and standard deviations, followed by the structure of VOT along the three primary dimensions discussed above. For each analysis, we compare the child-specific values to corresponding adult values when available. For American English adults, we use the Chodroff & Wilson (2017) adult VOT data from isolated speech for all analyses. For Cantonese, we use the Clumeck et al.

[4] Johnson (2021) examined the degree of VOT structure in Cantonese and English aspirated stops produced by adult Cantonese-English bilingual talkers in spontaneous speech. The magnitudes of the reported correlations were somewhat lower than those found for comparable monolingual English talkers, even in spontaneous speech (cf., Chodroff & Wilson, 2017). This could reflect an aspect of the bilingual status of the talkers, or statistical fluctuation in the measurement of this structure.

(1981) adult VOT data from isolated speech for comparison of VOT means and standard deviations. The observed patterns of variability shed light on the specification and structure of phonetic targets across speech sounds in child speech production.

American English

American English: Methods

The analysis employed the English portion of the Paidologos Corpus, which contains isolated speech produced by 81 children (40 female) aged 2;0 to 5;11, and time-aligned transcripts (Edwards & Beckman, *n.d.*). There were approximately 10 male and 10 female talkers per age (two-year-olds: 9 F, 11 M; three-year-olds: 10 F, 10 M; four-year-olds: 10 F, 11 M; five-year-olds: 11 F, 9 M). Within each age group, approximately half were considered "young" (Y: less than 6 months to the year) and half "old" (O: more than 6 months to the year; two-year-olds: 8 Y, 12 O; three-year-olds: 10 Y, 10 O; four-year-olds: 13 Y, 8 O; five-year-olds: 11 Y, 9 O). All participants came from middle socioeconomic status families, were typically developing, and had normal hearing (Edwards & Beckman, *n.d.*). Each child completed a picture-naming task with stimuli consisting of an image and a corresponding sound file. Infants were instructed to repeat the word; repeat responses were elicited if the response was different from the prompted word or if the tester thought the target sequence would be impossible to transcribe for any reason. The decision to employ the imitation paradigm was employed to ensure all participants had the same stimulus presentation. Audible productions, including repetitions, were retained in the recording and transcribed by native-speaker trained phoneticians. All recordings were made in quiet rooms at one or more preschools in central Ohio (Edwards & Beckman, 2008a, 2008b).

All stimuli began with /t k s/ and preceded the broad vowel categories of /i e a o u/; these consonants and vowels are present in the four languages investigated in the original corpus

(Edwards & Beckman, 2008a). For our current analysis, we examined only those tokens beginning with [t^h] or [k^h]. There were 29 unique words: 14 beginning with [t^h] and 15 beginning with [k^h]. These consonants preceded the vowels [i ɪ eɪ ɛ ʌ ɑ ɔ oʊ ʊ u], and were loosely balanced across the five broad vowel categories (Edwards & Beckman, 2008b). The stimulus list of words beginning with aspirated stops can be found in the Appendix.

The transcript for each recording was aligned to the audio with the Penn Forced Aligner (Yuan & Liberman, 2008). Word-initial stop boundaries were extended by 40 ms in each direction to create an interval of analysis for AutoVOT, an automatic VOT aligner that aims to identify the locations of the stop release and onset of voicing (Keshet, Sonderegger, & Knowles, 2014). All stop boundaries were manually corrected to align with the stop release, marked by the transient in the waveform, and the onset of voicing in the following vowel, marked by the start of periodicity in the waveform or the presence of the voice bar in the spectrogram. We retained all intended stop consonant productions for which the child produced an identifiable obstruent in initial position; this included a few repetitions for certain words. An identifiable obstruent was defined by the presence of aperiodicity, indicative of frication, in the waveform. Speaking rate was measured as the duration of the vowel immediately following the stop consonant (see also Theodore et al., 2009). Vowel offsets were also manually corrected for this estimation.

In total, there were 2,226 word-initial, prevocalic stops for analysis: the two-year-olds produced 580 tokens (per-child range: 24 to 39), the three-year-olds 526 tokens (per-child range: 14 to 31), the four-year-olds 580 tokens (per-child range: 17 to 30), and the five-year-olds 540 tokens (per-child range: 12 to 33). The median number of tokens per child was 13 for [t^h] (range: 5 to 16) and 15 for [k^h] (range: 7 to 22).

American English: Results

The VOT means and standard deviations for [t^h] and [k^h] for each age group are reported in Tables 3 and 4, along with the range of talker-specific means and standard deviations. The range of VOT means for [t^h] and [k^h] overlapped considerably with corresponding adult means reported in Chodroff & Wilson (2017). A handful of tokens were produced with short-lag VOT; however, all talker-specific mean VOTs were well within the long-lag VOT range (> 35 ms). Moreover, only about 4% of VOT tokens were less than 35 ms, and these were produced at an approximately equal rate across age groups (two: 21 tokens, three: 22, four: 27, five: 24).

While the mean VOTs were similar between children and adults, children had considerably larger standard deviations relative to adults. The maximum talker-specific child standard deviation was 64 ms for [t^h] and 93 ms for [k^h]; both stops had sizable ranges of child-specific standard deviations. In contrast, adult stop-specific standard deviations are typically between 10 and 30 ms for word-initial aspirated stops in isolated speech (Chodroff & Wilson, 2017).

Table 3. American English talker-specific means and standard deviations of [t^h] VOT in milliseconds presented for each age group and overall. The grand mean, SD, and range of talker means are taken over the distribution of talker-specific VOT means. The mean talker SD and range of talker SDs are taken over the distribution of talker-specific VOT standard deviations. Adult values calculated from the Chodroff & Wilson (2017) isolated speech data are reproduced here for comparison.

	[t^h]			
Age	Grand Mean (SD)	Range of Talker Means	Mean Talker SD	Range of Talker SDs
2	103 (33)	55 – 185	42	16 – 105
3	92 (19)	63 – 130	39	17 – 90
4	81 (15)	56 – 128	28	17 – 46
5	87 (24)	53 – 159	32	14 – 63
Combined	91 (25)	53 – 185	35	14 – 105
Adult	98 (23)	57 – 156	16	10 – 26

Table 4. American English talker-specific means and standard deviations of [k^h] VOT in milliseconds presented for each age group and overall. The grand mean, SD, and range of talker means are taken over the distribution of talker-specific VOT means. The mean talker SD and range of talker SDs are taken over the distribution of talker-specific VOT standard deviations. Adult values calculated from the Chodroff & Wilson (2017) isolated speech data are reported here for comparison.

	[k^h]			
Age	Grand Mean (SD)	Range of Talker Means	Mean Talker SD	Range of Talker SDs
2	99 (26)	63 – 162	43	24 – 100
3	87 (15)	63 – 134	32	22 – 49
4	80 (19)	49 – 139	29	8 – 83
5	85 (22)	44 – 152	27	9 – 46
Combined	88 (22)	44 – 162	34	8 – 100
Adult	99 (18)	67 – 137	16	11 – 20

With respect to between-category structure, child-specific mean VOTs were highly correlated between [t^h] and [k^h] across two- to five-year-olds (see Table 5 and Figure 1). The correlation was slightly lower than the adult VOT correlation for [t^h] and [k^h] in isolated speech reported in Chodroff & Wilson (2017; $r = 0.98$), but still strong at $r = 0.80$. A moderate to strong and significant correlation was also observed within each age group. With respect to the ordinal relationship, 58% of children had a [t^h] VOT mean slightly longer than their respective [k^h] VOT mean. A simple linear regression model predicting a child-specific VOT mean for [k^h] from the corresponding VOT mean for [t^h] revealed that at overall lower VOT values for [t^h], [k^h] was more likely to be greater than [t^h], but as the VOT for [t^h] increased, the distance between [t^h] and [k^h] decreased until the ranking ultimately reversed ($\beta_0 = 23.70$, $\beta_1 = 0.71$, each $p < 0.001$).

Additional within-category structure was also observed: child-specific VOT means and standard deviations were moderately correlated within each stop category, though not all correlations reached significance (Table 6). Overall, the within-category correlations across children parallel adult VOT correlations which were also moderate in magnitude ([t^h]: $r = 0.53$, [k^h]: $r = 0.43$; Chodroff & Wilson, 2017).

Table 5. Correlations between American English talker-specific means for [t^h] and [k^h] along with 95% bootstrapped confidence intervals. Adult data calculated from the Chodroff & Wilson (2017) isolated speech study is reported here for comparison. ** reflects $p < 0.001$

	[t^h] – [k^h]		
Age	*r*	95% CI	[t^h] < [k^h]
2	0.82**	[0.63, 0.93]	45%
3	0.73**	[0.44, 0.88]	50%
4	0.75**	[0.22, 0.95]	48%
5	0.81**	[0.38, 0.96]	40%
Combined	0.80**	[0.70, 0.89]	46%
Adult	0.95**	[0.86, 0.98]	54%

Table 6. Correlations of American English talker-specific means and corresponding standard deviations along with 95% bootstrapped confidence intervals. Adult values from the Chodroff & Wilson (2017) isolated speech data are reported here for comparison. * reflects $p < 0.01$. ** reflects $p < 0.001$

	[t^h] mean – SD		**[k^h] mean – SD**	
Age	*r*	95% CI	r	95% CI
2	0.82**	[0.55, 0.94]	0.71**	[0.38, 0.90]
3	0.60*	[0.37, 0.81]	0.53*	[-0.01, 0.87]
4	0.38	[-0.22, 0.78]	0.58*	[-0.43, 0.80]
5	0.49	[0.18, 0.68]	0.68*	[0.14, 0.89]
Combined	0.68**	[0.52, 0.82]	0.66**	[0.52, 0.81]
Adult	0.53*	[0.22, 0.74]	0.43	[-0.02, 0.70]

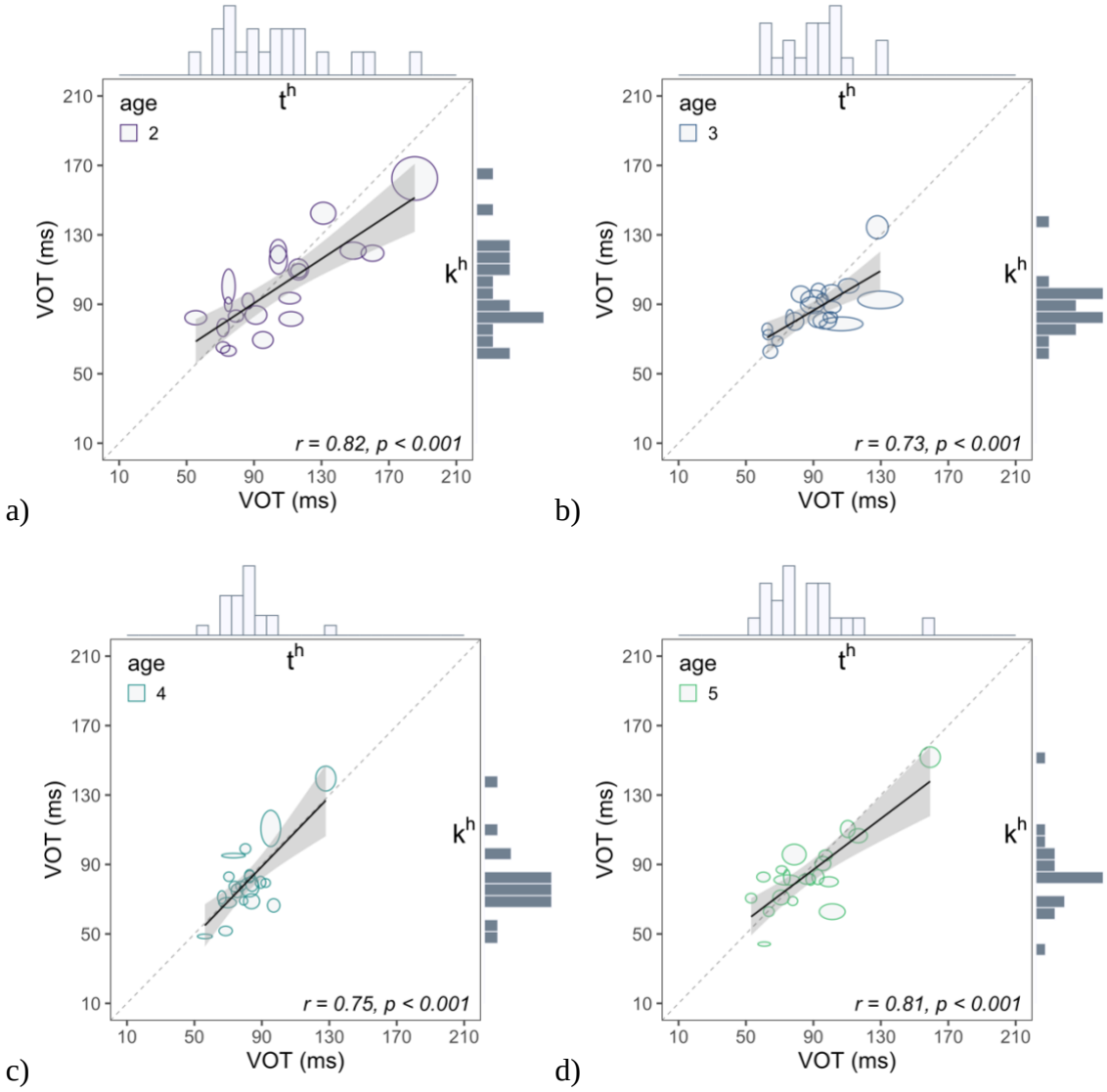

Figure 1. Variation and covariation of VOT means (ms) across American English child talkers at a) age two, b) age three, c) age four, and d) age five. Each ellipsis is centered on the paired talker-specific means for [t^h] and [k^h], and the diameter reflects the stop-specific standard deviation, scaled by 0.5. Marginal histograms show variation in talker means. The dashed line reflects the line of equality, the solid line the best-fit linear regression line, and the gray shading the local confidence interval around the best-fit linear regression line.

Variation in individual VOT values was analyzed with a linear mixed-effects model with fixed effects of place of articulation, speaking rate, number of syllables, following vowel height, following vowel tenseness, age, gender, the interaction between height and tenseness, and the full interactions between speaking rate, number of syllables, and age. The model also included a random by-child intercept and slope for place of articulation. All fixed effects were weighted effect coded in the following manner: place of articulation (coronal = 1, dorsal = -0.85), number of syllables (one = 1, two = -1.23), vowel height (high [i ɪ u ʊ] = 1, non-high [eɪ ɛ ʌ oʊ ɔ ɑ] = -0.54), vowel tenseness (tense [i eɪ ʌ u oʊ ɔ ɑ] = 1, lax [ɪ ɛ ʊ] = -2.95), and age (*age2*: two = 1, five = -1.07; *age3*: three = 1, five = -0.97; *age4*: four = 1, five = -1.07), and gender (female = 1, male = -1). Weighted effect coding is similar to sum coding but corrects for unbalanced data in the estimation of effects (Nieuwenhuis, te Grotenhuis, & Pelzer, 2017; te Grotenhuis, Pelzer, Eisinga, Nieuwenhuis, Schmidt-Catran, & Konig, 2017). As described above, speaking rate was measured as the following vowel duration, and was centered on the mean (174 ms) to facilitate model interpretation.

The model revealed significant effects of number of syllables, vowel height, vowel tenseness, age, as well as significant interactions between vowel height and tenseness, and speaking rate and age. VOTs in monosyllabic words were significantly longer than those in disyllabic words (*syllables*: β = 7.25, *t* = 4.72). Significantly longer VOTs were also observed before high vowels (*height*: β = 8.20, *t* = 4.27), before tense vowels (*tenseness*: β = 2.75, *t* = 2.98), and particularly before high, tense vowels: VOTs preceding high tense vowels ([i u]) were slightly longer than those preceding non-high tense vowels ([eɪ ʌ oʊ ɔ a]) (*height x tenseness*: β = 3.36, *t* = 3.20). The VOT of two-year-olds was significantly longer than average, by almost 13 ms (*age2:* β = 13.28, *t* = 3.24); this was further modulated by a small, but significant interaction with speaking rate, in which VOT increased by about 5 ms for each

100 ms increase in vowel duration (*rate x age2:* β = 0.05, *t* = 2.79). The VOT of four-year-olds was significantly shorter than average by approximately 10 ms (*age4*: β = -9.75, *t* = -2.36); the VOT of three-year-olds did not significantly differ from the mean (*age3*: β = -1.81, *t* = -0.43).

The main effects of place of articulation, speaking rate, and gender did not significantly influence VOT (*place*: β = -2.44, *t* = -1.79; *rate*: β = 0.02, *t* = 1.79; *gender*: β = 0.42, *t* = 0.18). The interactions between speaking rate and the three- or four-year-old age groups, number of syllables and each age group, and the three-way interactions between speaking rate, number of syllables, and each age group all failed to reach significance (*rate x age3*: β = -0.02, *t* = -0.99; *rate x age4*: β = 09.91, *t* = -0.46; *syllables x age2:* β = 0.51, *t* = 0.32; *syllables x age3:* β = -1.63, *t* = -0.96; *syllables x age4*: β = 1.22, *t* = 0.67; *rate x syllables x age2*: β = -0.02, *t* = -1.40; *rate x syllables x age3*: β = 0.02, *t* = 1.02; rate x *syllables x age4*: β = 0.03, *t* = 1.41).

American English: Discussion

Among American English children, considerable variation and structure emerged in the VOT of [t^h] and [k^h]. The range of child VOT means for [t^h] and [k^h] did not differ substantially from the corresponding ranges of adult VOT means in isolated speech (Chodroff & Wilson, 2017). Consistent with previous studies, child VOT standard deviations were much higher than adult VOT standard deviations for both [t^h] and [k^h] (see also Koenig, 2000). Notably, the child VOT SDs reported here were on average two times larger than the corresponding adult VOT SDs reported in Chodroff & Wilson (2017). Moderate correlations were observed between the VOT mean and SD, indicating increased variability with longer VOT targets.

With regards to between-category structure, a strong correlation of talker-specific child mean VOT between [t^h] and [k^h] was observed for each age group (r = 0.73 to r = 0.82). Though these correlations are not as high as the corresponding correlations in adult isolated speech (r = 0.95), they are comparable to figures in adult connected speech (r = 0.77). The highly predictable relationship between [t^h] and [k^h] VOT indicates that American English-learning children clearly represent the structured relationship between the laryngeal targets of [t^h] and [k^h].

Beyond the linear relationship, between-category structure can also be assessed with the ordinal (rank) relationship of VOT. For just over half of the children, the VOT mean for [t^h] exceeded that of [k^h], though statistically, the difference in VOT between [t^h] and [k^h] was negligible and not significant. This corresponds to previous findings of adult American English speech where 46% of adult talkers also produced a marginally longer VOT mean for [t^h] than [k^h] (Chodroff & Wilson, 2017; see also Docherty, 1992; Stuart-Smith et al., 2015; Yao, 2009). This inconsistency in rank somewhat contrasts with the broader generalization that VOT increases with more posterior places of articulation (Cho & Ladefoged, 1999); however, the linear relationship was still highly predictable: at lower VOT values, [t^h] was more likely to be shorter than [k^h], and as the overall VOT increased, the VOT mean for [t^h] was likely to surpass that of [k^h].

Within a given stop category, the phonetic targets for overall VOT might also account for the corresponding variance, thus reflecting VOT structure within a stop category. Indeed, moderate to strong correlations were observed between child-specific VOT means and SDs for both [t^h] and [k^h].

Contextual VOT variation partially paralleled adult speech patterns. Speaking rate (as an interaction with age), number of syllables, and the following vowel (high, tense vowels in particular) significantly accounted for variation in child VOT. These factors also influenced

VOT in adult speech, but at times only in the connected speech style as opposed to the isolated speech style. Specifically, a following [i] or [u] corresponded to a significantly longer VOT in adult connected speech, mirroring the significant height by tenseness interaction in child VOT, but this influence was not observed in isolated speech (Chodroff & Wilson, 2017). Nevertheless, speaking rate and number of syllables (only tested in connected speech) had significant influences on adult VOT production.

Overall age differences were also observed: relative to average, two-year-olds had significantly longer VOTs, whereas four-year-olds had shorter VOTs. The longer VOTs observed for the two-year-olds partially reflected a stronger influence of speaking rate and number of syllables on VOT for their age group, but these enhanced effects could not entirely account for the duration. This could reflect a general overshoot phase of VOT production (Macken & Barton, 1980), or simply a bias in the sample. It is worth noting that in the isolated speech style, some adult mean VOTs were just as long as the mean VOTs from two-year-olds. The long VOT values found here may not be so atypical for any age, at least in this isolated speech style.

Cantonese

In the following section, we investigate variability and systematicity in the VOT of [t^h] and [k^h] across 80 two- to five-year old Cantonese talkers. Cantonese parallels American English in having voiceless aspirated stops in word-initial position. While the underlying featural representation of these stops could differ between languages, the same laryngeal feature is argued to be present for each Cantonese aspirated stop (i.e., [+spread glottis], [-voice], or even ["X"], where X can refer to any shared feature). Based on the shared featural representation, we expect Cantonese-learning children to parallel their American English-learning peers in VOT production.

Cantonese nevertheless does differ from English along many phonological dimensions, most notably by having an expanded inventory of aspirated consonants. In addition to the simple aspiration contrasts (/p^h t^h k^h/ vs /p t k/), Cantonese has an aspiration contrast on the alveolar affricate (/$\widehat{ts}^h$/ vs $\widehat{ts}$/) and labialized velar voiceless stop (/k^{wh}/ vs /k^w/). As these do not occur in English and given their increased articulatory complexity (e.g., affrication and labialization), we leave the realization of these stops to future research; nonetheless, the larger inventory could impact the acquisition trajectory or structural relationships between places of articulation.

Cantonese: Methods

This analysis employed the Cantonese portion of the Paidologos Corpus (Edwards & Beckman, *n.d.*). The corpus contents and collection procedure mirrored that of the English portion, but was based at the Chinese University of Hong Kong. 80 talkers, aged two to five, were recorded in quiet rooms in one or more nurseries in Hong Kong. 10 male and 10 female talkers were recruited for each age year. For children aged two to four, half of the children were young for the age group (e.g., 2;0 to 2;5) and half were old for the age group (e.g., 2;6 to 2;11). For children in the five-year-old group, 15 were "young" for the group, and 5 were "old". All participants came from middle socioeconomic status families, were typically developing, and had normal hearing (Edwards & Beckman, *n.d.*). Participants were asked to complete a picture naming task in which an image and prerecorded audio stimulus were presented to the child. The child's production and any subsequent repetitions were recorded. As in the English methods, all stimuli began with /t k s/ and preceded the broad vowel categories of /i e a o u/. For Cantonese, the initial consonants were [t t^h $\widehat{ts}$ $\widehat{ts}^h$ s k k^h k^w k^{wh}]. For comparison with English, we analyzed words beginning with [t^h] and [k^h]. There were 27 unique words: 12 beginning with [t^h] and 15 beginning with [k^h]. These consonants were followed by the vowels [iː iːu ɪ ɛː ei a aːi ɐu ɔː ɔːi ʊ uːy]; as for English, the stimuli were

loosely balanced across the five broad vowel categories (Edwards & Beckman, 2008a, 2008b). A full list of the words and initial syllable transcriptions can be found in the Appendix.

To obtain VOT measurements, boundaries of word-initial stop consonants were submitted to AutoVOT, which automatically identified the stop release and onset of voicing (Keshet et al., 2014). All stop boundaries were manually corrected to align with the burst release and the onset of voicing, using the same acoustic landmarks noted in section 2.1. Any instance in which the child did not produce the intended stop-initial prompt was excluded from analysis.

In addition to manual refinement of the VOT boundaries, syllable boundaries were also marked. Since it was difficult to identify a clear boundary before a nasal, we used the syllable rhyme duration (vowel and following consonant, if present) as a proxy for speaking rate instead of the following vowel duration.

In total, there were 2,017 word-initial, prevocalic stops for analysis: the two-year-olds produced 495 tokens (per-child range: 21 to 27), the three-year-olds 516 tokens (per-child range: 24 to 29), the four-year-olds 512 tokens (per-child range: 16 to 30), and the five-year-olds 494 tokens (per-child range: 11 to 29). The median number of tokens per child was 12 for [t^h] (range: 3 to 14) and 15 for [k^h] (range: 7 to 17).

Cantonese: Results

As shown in Tables 7 and 8, stop-specific means and standard deviations ranged considerably across talkers. The means were comparable to previously reported adult VOT means from isolated speech which range from approximately 70 to 90 ms for [t^h] and [k^h] (Clumeck et al., 1981; Lee Oi Yee, 1997; Lisker & Abramson, 1964). Nevertheless, several children under the age of 5 still produced short-lag VOTs for the intended long-lag stops. Specifically, seven VOT means from five unique child talkers were well within the short-lag

range (< 35 ms; three two-year-olds, one three-year-old, and one four-year-old), and approximately 10% of VOT tokens were less than 35 ms. The rate of short-lag VOT production decreased with age (two: 89 tokens, three: 73, four: 35, five: 7).

Table 7. Cantonese means and standard deviations of [t^h] VOT in milliseconds presented for each age group and overall. The grand mean, SD, and range of talker means are taken over the distribution of talker-specific VOT means. The mean talker SD and range of talker SDs are taken over the distribution of talker-specific VOT standard deviations. Adult values come from Clumeck et al. (1981) and are reported here for comparison.

	[t^h]			
Age	Grand Mean (SD)	Range of Talker Means	Mean Talker SD	Range of Talker SDs
2	76 (26)	25 – 116	36	12 – 75
3	76 (23)	36 – 117	36	17 – 79
4	83 (24)	19 – 123	29	11 – 55
5	93 (19)	64 – 129	26	12 – 48
Combined	82 (24)	19 – 129	32	11 – 79
Adult	83 (21)	—	—	—

Table 8. Cantonese talker-specific means and standard deviations of [k^h] VOT in milliseconds presented for each age group and overall. The grand mean, SD, and range of talker means are taken over the distribution of talker-specific VOT means. The mean talker SD and range of talker SDs are taken over the distribution of talker-specific VOT standard deviations. Adult values come from Clumeck et al. (1981) and are reported here for comparison.

	[k^h]			
Age	Grand Mean (SD)	Range of Talker Means	Mean Talker SD	Range of Talker SDs
2	82 (24)	31 – 117	38	18 – 51
3	81 (21)	34 – 132	34	17 – 52
4	90 (24)	26 – 128	30	11 – 52
5	102 (21)	66 – 145	35	18 – 74
Combined	89 (24)	26 – 145	34	11 – 74
Adult	91 (22)	—	—	—

As in English, VOT variation was highly structured between [t^h] and [k^h]: talker-specific means VOTs were strongly correlated between [t^h] and [k^h] across two- to five-year old talkers (see Table 9 and Figure 2). A simple linear regression revealed that the difference between [t^h] and [k^h] VOT means decreased with overall higher VOT values ($\beta_0 = 15.45$, $\beta_1 = 0.89$, each $p < 0.01$). The ordinal relationship between [t^h] and [k^h] VOT means conformed to

the expected generalization in which VOT increases with more posterior places of articulation for 73% of Cantonese-learning children. As shown in Table 10, talker-specific child means and standard deviations were moderately correlated within each of the stop categories, though as before, not all correlations reached significance.

Table 9. Correlations between Cantonese talker-specific means for [t^h] and [k^h] along with 95% bootstrapped confidence intervals. * reflects $p < 0.01$; ** reflects $p < 0.001$.

	[t^h] – [k^h]		
Age	*r*	95% CI	[t^h] < [k^h]
2	0.85**	[0.68, 0.94]	70%
3	0.76**	[0.49, 0.89]	70%
4	0.87**	[0.59, 0.96]	80%
5	0.84**	[0.62, 0.94]	70%
Combined	0.85**	[0.76, 0.91]	73%

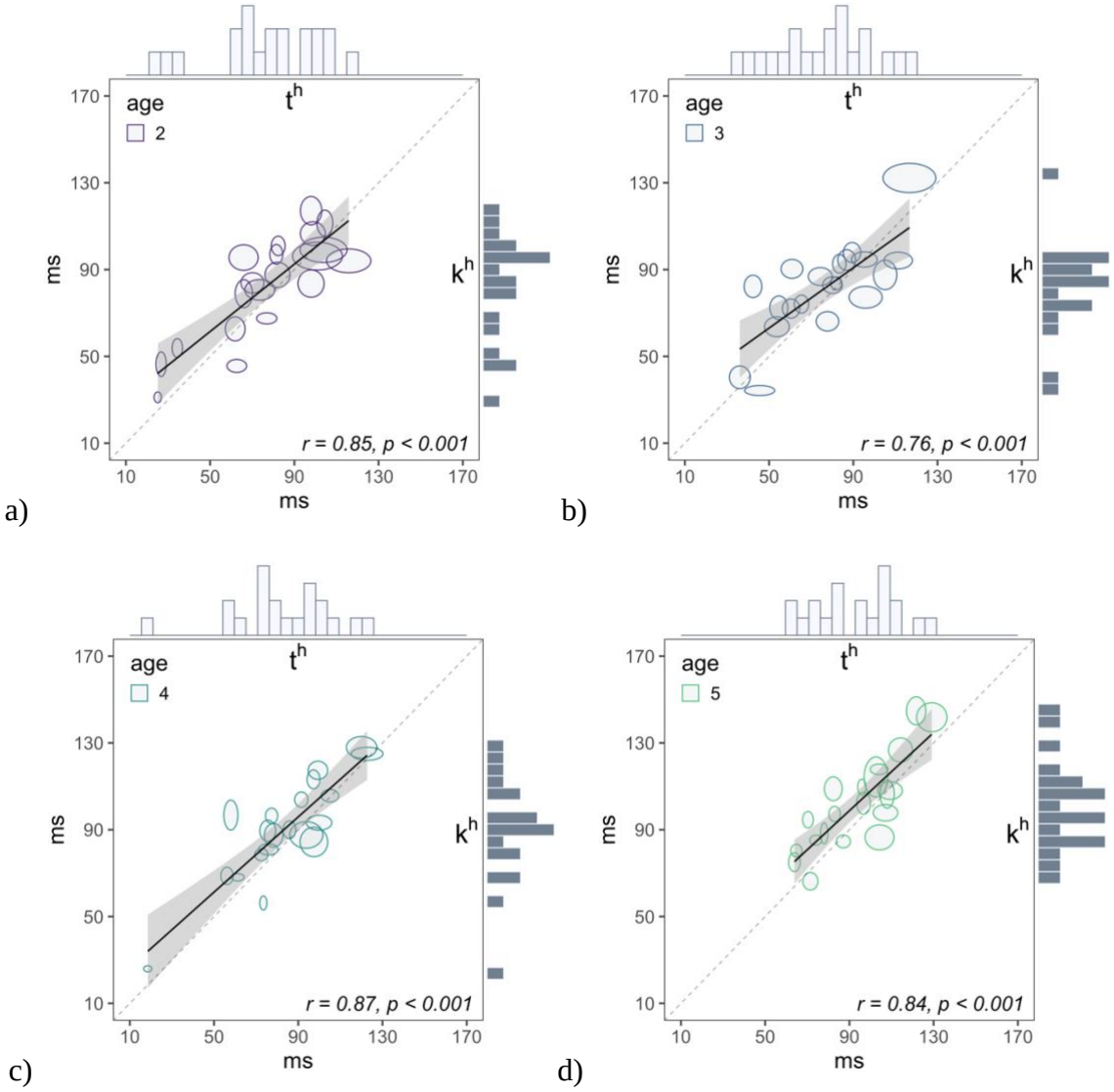

Figure 2. Variation and covariation of VOT means (ms) across Cantonese child talkers at a) age two, b) age three, c) age four, and d) age five. Each ellipsis is centered on the paired talker-specific means for [t^h] and [k^h], and the diameter reflects the stop-specific standard deviation, scaled by 0.5. Marginal histograms show variation in talker means. The dashed line reflects the line of equality, the solid line the best-fit linear regression line, and the gray shading the local confidence interval around the best-fit linear regression line.

Table 10. Correlations of Cantonese talker-specific means and corresponding standard deviations along with 95% bootstrapped confidence intervals. * reflects $p < 0.01$; ** reflects $p < 0.001$.

	[t^h] mean – SD		**[k^h] mean – SD**	
Age	*r*	95% CI	r	95% CI
2	0.69**	[0.39, 0.86]	0.67*	[0.19, 0.58]
3	0.43	[-0.16, 0.73]	0.47	[-0.19, 0.88]
4	0.69**	[0.47, 0.84]	0.41	[-0.05, 0.74]
5	0.75**	[0.54, 0.87]	0.49	[0.01, 0.72]
Combined	0.48**	[0.32, 0.61]	0.42**	[0.22, 0.59]

A linear mixed-effects regression was used to assess factors governing variability in VOT. Categorical predictors were weighted effect coded (te Grotenhuis et al., 2017) and continuous variables were centered on their respective means. The model included fixed effects of place of articulation (coronal = 1, dorsal = -0.80), speaking rate (mean: 233 ms), number of syllables (two = 1, three = -2.01), following vowel height (high [iː iːu ɪ u uːy ʊ] = 1, non-high [ɛː ei a aːi ɐu ɔː ɔːy] = -1.07), following vowel tenseness (tense [iː iːu ɛː ei a aːi ɐu ɔː ɔːy u uːy] = 1, lax [ɪ ʊ] = -4.42), age (*age2*: two = 1, five = -1.00; *age3*: three = 1, five = -1.04; *age4*: four = 1, five = -1.04), and gender (female = 1, male = -0.97). All interactions between speaking rate, number of syllables, and age were also included.[5] The random effect structure included an intercept and slope for place of articulation by participant and an intercept for item. Speaking rate was approximated using the syllable rhyme duration (duration of the following vowel and coda nasal consonant if present).

The model revealed significant effects of place of articulation, number of syllables, vowel height, and vowel tenseness. Coronal stops were shorter than dorsal stops by approximately 6 ms (*place*: $\beta = -3.08$, $t = -3.49$). The number of syllables in the word significantly modulated the initial VOT: VOT in disyllabic words was approximately 4 ms longer than VOT in trisyllabic words (*syllables*: $\beta = 2.02$, $t = 3.41$). VOT preceding high

[5] Like English, Cantonese also has a tenseness contrast between vowels that can appear in open or closed syllables, and an additional length distinction among tense vowels (Tse, 2019; Zee, 2003). The present dataset included only a few high lax vowels ([ɪ ʊ]) so we could not include the interaction between height and tenseness as had been done for the English analysis.

vowels was also significantly longer than VOT preceding non-high vowels, by about 10 ms (*height:* β = 5.17, *t* = 5.88). VOT preceding tense vowels was also significantly longer than preceding lax vowels (*tenseness:* β = 2.81, *t* = 6.50).

Main effects of speaking rate, age and gender failed to reach significance. Though trending in the expected direction, speaking rate did not significantly influence VOT (*rate*: β = 0.02, *t* = 1.92). Much of the VOT variation that could be related to speaking rate was accounted for by the number of syllables in the word. Comparable VOT values were also found across two- to five-year-olds (*age2*: β = -6.93, *t* = -1.60; *age3*: β = -6.70, *t* = -1.57; *age4*: β = 0.98, *t* = 0.23) and between male and female talkers (*gender*: β = -0.73, *t* = -0.29). All two- and three-way interactions between speaking rate, number of syllables, and age were also not significant (*rate x syllables*: β = 0.00, *t* = -0.61; *rate x age2*: β = 0.01, *t* = 0.51; *rate x age3*: β = -0.01, *t* = -0.36; *rate x age4*: β = -0.01, *t* = -0.44; *syllables x age2*: β = -0.33, *t* = -0.32; *syllables x age3*: β = -0.33, *t* = -0.33; *syllables x age4*: β = 0.29, *t* = 0.27; *rate x syllables x age2*: β = 0.02, *t* = 1.64; *rate x syllables x age3*: β = 0.00, *t* = -0.36; *rate x syllables x age4*: β = 0.00, *t* = -0.27).

Cantonese: Discussion

The average child VOT means did not differ considerably from previously reported adult values that were around 70 to 90 ms for aspirated stop categories (Clumeck et al., 1981; Lee Oi Yee, 1997; Lisker & Abramson, 1964); however, considerable individual variability was observed. Several two- and three-year-old talkers had not yet acquired consistent long-lag VOT; stop-specific means for eight of the talkers were below 50 ms (four two-year-olds, three three-year-olds, one four-year-old). All but one talker above the age of four had average VOTs well within the long-lag VOT range (e.g., ≥ 50 ms). Consistent with previous child VOT studies, the standard deviations of Cantonese children were slightly higher than the corresponding standard deviations of adult talkers (e.g., Koenig, 2000). Clumeck et al. (1981)

observed adult standard deviations of approximately 21 and 22 ms for [t^h] and [k^h], whereas the average child standard deviation in the present study was 27 and 28 ms for these same categories. In addition, the talker-specific child means and SDs were moderately and significantly correlated with one another.

With respect to between-category VOT structure, strong correlations of talker-specific mean VOT were observed between [t^h] and [k^h] for each age group. Cantonese children clearly represent the structured relationship between stop VOTs and reliably produce this relationship between stop categories. Moreover, the expected ordinal relationship in VOT between [t^h] and [k^h] was also observed for most Cantonese-learning children, and the difference between them was significant. Overall shorter VOT values also corresponded to a slightly greater VOT difference between the two places of articulation. Within-category VOT structure was observed in the moderate to strong correlations between VOT means and SDs for both [t^h] and [k^h]. Finally, contextual influences of place of articulation, number of syllables and following vowel context also significantly influenced VOT in Cantonese child speech: [k^h] had a longer VOT than [t^h], and longer VOTs were observed in words with fewer syllables. In addition, longer VOTs were observed before high or tense vowels. In contrast to American English, no effect of age or speaking rate (as approximated by rhyme duration) was observed.

General Discussion

The present study identified considerable VOT variability in aspirated stops produced by American English and Cantonese two- to five-year-old talkers. Consistent with previous studies, the within-category standard deviations of child talkers were larger than those of adult talkers, but in contrast, child VOT means were very similar to corresponding adult VOT means in each language. Although a repetition approach was used to collect these materials, the increased variability in child speech relative to adult speech importantly indicates that

children were not merely reproducing the exact form of the adult prompt. Moreover, child VOT production was still highly structured: Between stop categories, strong correlations of VOT means were identified across talkers of each age group and in each language. Within each stop category, child-specific means and corresponding SDs were mostly positively correlated with one another. Child VOT variability was also structured by many of the same contextual factors implicated in adult speech production. These findings indicate that American English and Cantonese children represent VOT structure among aspirated stops to a reasonably high degree and in a manner comparable to adult speech. These findings have implications for the representation of subsegmental detail, its relation to the constraint of uniformity on phonetic realization, as well as active debates in language acquisition regarding innate or emergent representations.

Representation of Subsegmental Detail and Uniformity

The present study highlights clear similarities in the phonological and phonetic representation of aspirated stop consonants between children and adults. Across talkers, the correlation between [t^h] and [k^h] VOT means was strong, and within a talker, the difference between [t^h] and [k^h] means was generally small. Moreover, most child VOT means were well within the adult VOT range, suggesting comparable phonetic targets — at least in the average. Critical to note is the fact that children observe this structure, despite having underdeveloped motor and specifically laryngeal control (Green, Moore, Higashikawa, & Steeve, 2000; Goffman & Smith, 1999; Koenig, 2000). This developing laryngeal control is naturally reflected in the substantially larger standard deviations of VOT relative to adults. That is, the physical instantiation of stop VOT is more variable in children than adults.

These findings are also consistent with a constraint of target uniformity on phonetic realization: strong VOT covariation among aspirated stops should arise due to underlying

uniformity in the mapping from the laryngeal feature (e.g., [+spread glottis]) to the corresponding phonetic target across segments, regardless of the place of articulation (Chodroff & Wilson, 2017, 2022; Keating, 2003). Overall, the phonological representation of aspirated stops appears to be comparable to that of adult talkers with a common feature yoking [t^h] and [k^h] together. In child speech production, as in adult speech production, the representation of a given segment is not as an independent entity, but rather as a member of a natural class, connected by a shared subsegmental representation. Moreover, the phonetic representation of the laryngeal feature, where the abstract phonetic target is approximated by the average VOT, is also very adult-like for most two- to five-year-olds.

Variation in Uniformity

As a constraint on phonetic realization, target uniformity minimizes deviations in the phonetic targets that correspond to a distinctive feature value across relevant segments. Though the correlations between VOT means were overall strong, the strength of the relationship did differ by language. The VOT correlations in American English were slightly weaker than those in Cantonese, and the ranking between [t^h] and [k^h] VOTs was less consistent in English than in Cantonese. The Cantonese ranking followed the expected ordinal relationship based on universal patterns ([t^h] < [k^h]). Interestingly, while the American English ranking was more variable across children, the same pattern was observed across American English-speaking adults. Specifically, the VOT of [t^h] was frequently longer than that of [k^h], especially at overall long VOT values.

One possible explanation for this difference in ranking could be a weaker [k^h] occlusion in English than in Cantonese. The duration necessary for sufficient intraoral pressure would end up being longer, giving rise to a slightly shorter VOT for [k^h] than might otherwise be expected. Evidence for a weak stop occlusion could be the presence of a multiple burst

release or frication in the release, which is indeed relatively common in dorsal stop productions (Lavoie, 2001; Olive, Greenwood, & Coleman, 1993).

In a post-hoc analysis, the stop productions from the five-year old talkers were labeled for a single burst release or a multiple burst release. A logistic mixed-effects regression revealed that, contrary to expectation, Cantonese stops were significantly more likely to have a multiple burst release than English stops (β = 0.37, $p < 0.01$), and conforming to expectation, [k^h] was more likely to be produced with a multiple burst release in both languages (β = 0.41, $p < 0.001$). No interaction was observed between language and stop category on the release type (single or multiple; β = -0.02, p = 0.80). Based on impression of the audio and waveform, aspirated stops in Cantonese child speech were overall more likely to be fricated than those in English. This line of investigation was therefore inconclusive with respect to the VOT ranking.

Alternatively, it may be that target uniformity has a *weaker* influence in American English than in Cantonese. This is not to say that uniformity has no influence: the expected deviation from the VOT values for [t^h] and [k^h] given uniformity of articulatorily-specified targets is very small, and the correlations are still strong. The unexpected but relatively consistent VOT ranking in American English could instead be accounted for by a broader constraint of "pattern" uniformity, in which talkers maintain consistent differences between phonetic targets of related speech sounds (Chodroff & Wilson, 2022).

Development of Subsegmental Representation

The current findings strongly suggest a subsegmental representation of a laryngeal feature in speech production in talkers as young as two years of age: some feature must tie the productions of e.g., [t^h] and [k^h] together. Barring the evident variability in laryngeal control, the observed between-category VOT structure between stop categories is already

robust at an early age. Nevertheless, for the given isolated speech style, the adult correlational structure between [t^h] and [k^h] is still tighter (r = 0.95). This difference between children and adults could be due to one of the following scenarios: 1) children have a detailed cognitive representation of the relationship, but struggle in the articulation, or 2) the cognitive representation of the relationship could still be under development, and the partial relationship is reflected in articulation, or 3) children are still developing the cognitive representation and have difficulty in the articulation. Regardless, the subsegmental representation between [t^h] and [k^h] is already apparent. In light of this, however, is this representation innate or does it emerge from the perception of ambient speech data?

It is well-established that perceptual sensitivity to phonetic detail, especially VOT, emerges at a very young age. By one month of age, infants demonstrate the capacity to discriminate between voiced and voiceless stop consonants along the VOT continuum (Aslin, Jusczyk, & Pisoni, 1998; Eimas, Siqueland, Jusczyk, & Vigorito, 1971; McMurray & Aslin, 2005). At eight months, infants show sensitivity to within-category VOT variation (McMurray & Aslin, 2005), and at one year, sensitivity to mispronunciations of familiar words involving the laryngeal feature (Swingley & Aslin, 2000). Finally, five- and seven-year-old children display adult-like identification and discrimination of stop consonants (Wolf, 1973). These findings show that children are highly attuned to VOT variation and its relevance to the voicing feature at a young age.

Given sensitivity to VOT variation for stop voicing discrimination, children may also be attuned to VOT systematicity among stops with a shared laryngeal feature. A child could reasonably acquire a rich representation of VOT and its laryngeal representation for speech production directly from ambient data and very early on (e.g., PRIMIR: Werker & Curtin, 2005). How exactly such perceptual input relates to production representations, however, is complicated. If information is indeed “coupled” between the perceptual and production

representations, children could mimic these perceptual representations in speech production (McMurray & Farris-Trimble, 2012; Redford, 2019; Schwartz, Basirat, Ménard, & Sato, 2012).

However, one important caveat needs to be raised for an emergent representation: a child would need to recognize that VOT must be calibrated by talker. If the child tracked the [t^h] VOTs from talker A and the [k^h] VOTs from talker B, consistency in the between-category structure would be unlikely to arise given the considerable cross-talker variability in overall VOT values. Social sensitivity to the *source* of the content has indeed been proposed for handling multi-talker input for child speech perception, so such talker-specific tracking of VOT is not improbable (Tripp, Feldman, & Idsardi, 2021). Indeed, how multi-talker input might affect any emergent representations of this feature is open for debate. In perception, infant word learning frequently improves with multi-talker input or highly variable input from a single talker along multiple acoustic dimensions (e.g., Bulgarelli & Bergelson, 2022; Galle, Apfelbaum, & McMurray, 2010; Rost & McMurray, 2009, 2010; cf., Quam, Knight, & Gerken, 2017). In production, multi-talker input also improves production accuracy and speed in four-year-olds, whereas single talker input does not (Richtsmaier, Gerken, Goffman, & Hogan, 2009).

Alternatively, the representations necessary for systematic laryngeal realization could be innate or arise from an innate general pressure for representational economy in cognition (e.g., Maddieson, 1995; Schwartz et al., 2007). Re-use of the laryngeal gesture and laryngeal-oral timing relationship across places of articulation would be economical in representation (e.g., Smith & Zelaznik, 2004). However, while target uniformity is related to a general principle of economy, it does not require perfect re-use of phonetic targets across speech sounds (Chodroff & Wilson, 2022). In fact, the VOT patterns suggest the laryngeal phonetic targets might differ slightly between places of articulation. Regardless, the constraint

enforces a high degree of similarity across speech sounds in the phonetic targets corresponding to the shared distinctive feature.[6] Such a pressure for economical representations could be argued to arise from a general, but innate bias in cognition.

Indeed, English- and French-learning infants at 6, 9, and 12 months of age demonstrate a reasonably consistent ordinal VOT relationship in babbled unaspirated stops (Whalen et al., 2007). Such an ordinal relationship is expected if children are indeed attempting similar laryngeal targets across stop place of articulation. Whereas development of a uniform set of laryngeal phonetic targets might suggest that the between-category relationships strengthen over time, we instead observe a stable relationship across age groups.

Regardless of the representational origins, these findings indicate an early and strong presence of the subsegmental relationship between [t^h] and [k^h] in speech production, pointing towards an early or even innate constraint on phonetic realization. By the age of two, children have developed not only a fine-grained perceptual representation of the laryngeal contrast in stop consonants, but also, a rich production representation of the laryngeal feature for voiceless aspirated stops.

Conclusion

Overall, child VOT production was more variable than adult VOT production, but the means and overall VOT structure were highly comparable. All three forms of structure indicate a rather mature representation of phonetic detail in child speech production. Of note is the strong between-category VOT correlations that are consistent with universal patterns of VOT (Chodroff et al., 2019). The highly predictable linear relationship of VOT may reflect a constraint of target uniformity on phonetic realization, in which the phonetic targets

[6] Note that representational economy in this sense does not necessarily mean ease of articulation (cf. Lindblom, 1986 for a type of economy that is related to ease of articulation); the phonetic target could be articulatorily complex, but its re-use across multiple speech sounds would still be economical (e.g., Maddieson, 1995).

corresponding to the laryngeal feature of both [t^{h}] and [k^{h}] are yoked together and highly similar. This structure reveals that children represent phonetic detail that is below the level of a word (e.g., Vihman, 2017). Overall these findings shed light on our understanding of phonetic representations in child speech.

Future studies should extend this analysis of phonetic systematicity and variation to additional segments, phonetic dimensions, and languages in child speech production. For example, structured phonetic variation has been observed across adult talkers in the F1 of vowels with a shared height, such as mid vowels, [e] and [o] (e.g., Schwartz & Ménard, 2019; Oushiro, 2019; Salesky, Chodroff, Pimentel, Wiesner, Cotterell, Black, & Eisner, 2020; Watt, 2000), and in spectral properties of sibilant fricatives with a shared place of articulation (e.g., Salesky et al., 2020; Chodroff & Wilson, 2022). It would be worthwhile to examine whether such structure is also observed in other segments and phonetic dimensions across child talkers.

Supplemental Material

The data and analysis scripts for this study are available on OSF at https://osf.io/hmz27/?view_only=99fb3db9e267440791248c4b63b3e47a.

References

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1998). Speech and auditory processing during infancy: constraints on and precursors to language. In W. Damon (Ed.), *Handbook of Child Psychology: Vol. 2. Cognition, Perception, and Language* (p. 147–198). John Wiley & Sons Inc.

Barlow, J. A., & Gierut, J. A. (1999). Optimality theory in phonological acquisition. *Journal of Speech, Language, and Hearing Research*, *42*(6), 1482–1498.

Barton, D. (1976). *The role of perception in the acquisition of phonology* (Doctoral dissertation). University College London (University of London).

Barton, D., & Macken, M. A. (1980). An instrumental analysis of the voicing contrast in word-initial stops in the speech of four-year-old English-speaking children. *Language and Speech*, *23*(2), 159–169.

Bulgarelli, F., & Bergelson, E. (2022). Talker variability shapes early word representations in English-learning 8-month-olds. To appear in *Infancy*. doi: 10.31234/osf.io/rxyjc

Cho, T., & Keating, P. A. (2009). Effects of initial position versus prominence in English. *Journal of Phonetics*, *37*(4), 466–485. https://doi.org/10.1016/j.wocn.2009.08.001

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

Cole, J. S., Choi, H., Kim, H., & Hasegawa-Johnson, M. (2003). The effect of accent on the acoustic cues to stop voicing in Radio News speech. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 15–18). Barcelona, Spain.

Chodroff, E., Golden, A., & Wilson, C. (2019). Covariation of stop voice onset time across languages: evidence for a universal constraint on phonetic realization. *The Journal of the Acoustical Society of America*, *145*(1), EL109–EL115. doi: 10.1121/1.5088035

Chodroff, E., & Wilson, C. (2017). Structure in talker-specific phonetic realization: covariation of stop consonant VOT in American English. *Journal of Phonetics, 61*, 30–47. doi: 10.1016/j.wocn.2017.01.001

Chodroff, E., & Wilson, C. (2022). Uniformity in phonetic realization: Evidence from sibilant place of articulation in American English. *Language, 98*(2). doi: 10.1353/lan.0.0259

Clumeck, H., Barton, D., Macken, M. A., & Huntington, D. A. (1981). The aspiration contrast in Cantonese word-initial stops: data from children and adults. *Journal of Chinese Linguistics*, *9*(2), 210–225.

Docherty, G. (1992). *The timing of voicing in British English obstruents*. Berlin: Walter de Gruyter.

Edwards, J. R., & Beckman, M. E. (2008a). Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development. *Language Learning and Development*, *4*(2), 122–156. doi: 10.1080/15475440801922115

Edwards, J. R., & Beckman, M. E. (2008b). Methodological questions in studying consonant acquisition. *Clinical Linguistics & Phonetics*, *22*(12), 937–956. https://doi.org/10.1080/02699200802330223

Edwards, J. R., & Beckman, M. E. (n.d.). Paidologos Project. *PhonBank.* Retrieved July 5, 2021, from https://phonbank.talkbank.org/access/Paidologos.html.

Eguchi, S., & Hirsch, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngol, 257*, 1-51.

Eimas, P. D., Siqueland, E. R., Jusczyk, P. W., & Vigorito, J. (1971). Speech perception in infants. *Science*, *171*(3968), 303–306.

Galle, M. E., Apfelbaum, K. S., & McMurray, B. (2015). The Role of Single Talker Acoustic

Variation in Early Word Learning. *Language Learning and Development*, *11*(1), 66–79. https://doi.org/10.1080/15475441.2014.895249

Gandour, J., Petty, S. H., Dardarananda, R., Dechongkit, S., & Mukngoen, S. (1986). The acquisition of the voicing contrast in Thai: a study of voice onset time in word-initial stop consonants. *Journal of Child Language*, *13*(3), 561–572. doi: 10.1017/S0305000900006887

Gilbert, J. V. (1977). A voice onset time analysis of apical stop production in 3-year-olds. *Journal of Child Language*, *4*(1), 103–110. doi: 10.1017/S0305000900000507

Goffman, L., & Smith, A. (1999). Development and phonetic differentiation of speech movement patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *25*(3), 649–660. doi: 10.1037/0096-1523.25.3.649

Goldinger, S. D., & Van Summers, W. (1989). Lexical neighborhoods in speech production: A first report. *The Journal of the Acoustical Society of America*, *85*(S1), S97-S97.

Green, J. R., Moore, C. A., & Steeve, R. W. (2000). The physiologic development of speech motor control: Lip and jaw coordination. *Journal of Speech, Language, and Hearing Research*, *43*(1), 239–255.

Haelsig, P. C., & Madison, C. L. (1986). A study of phonological processes exhibited. *Language, Speech, and Hearing Services in Schools*, *17*(2), 107–114.

Hullebus, M. A., Tobin, S. J., & Gafos, A. I. (2018). Speaker-specific structure in German voiceless stop voice onset times. *Proceedings of Interspeech 2018*, 1403–1407. doi: 10.21437/Interspeech.2018-2288

Hunnicutt, L., & Morris, P. (2016). Pre-voicing and aspiration in Southern American English. *University of Pennsylvania Working Papers in Linguistics*, *22*(1), 215–224. https://doi.org/10.1017/CBO9781107415324.004

Johnson, K. A. (2021). Leveraging the uniformity framework to examine crosslinguistic

similarity for long-lag stops in spontaneous Cantonese-English bilingual speech. *Proceedings of Interspeech*. Brno, Czech Republic.

Keating, P. A. (2003). Phonetic and other influences on voicing contrasts. In M. Solé, D. Recasens, & J. Romero (Eds.), *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 20–23). Barcelona, Spain.

Keshet, J., Sonderegger, M., Knowles, T. (2014). AutoVOT: a tool for automatic measurement of voice onset time using discriminative structured prediction [Computer program]. Version 0.91, retrieved August 2014 from https://github.com/mlml/autovot/.

Kewley-Port, D., & Preston, M. S. (1974). Early apical stop production: a voice onset time analysis. *Journal of Phonetics*, *2*(3), 195–210.

Klatt, D. (1975). Voice onset time, frication, and aspiration in word-initial consonant clusters. *Journal of Speech and Hearing Research*, *18*(4), 686–706.

Koenig, L. L. (2000). Laryngeal factors in voiceless consonant production in men, women, and 5-year-olds. *Journal of Speech, Language and Hearing Research*, *43*(5), 1211–1228.

Lavoie, L. M. (2001). *Consonant strength: phonological patterns and phonetic manifestations*. New York: Garland Publishing, Inc.

Lee Oi Yee, C. (1997). *A study of voice onset time in word-initial stop consonants by Cantonese-speaking children* (Undergraduate dissertation). The University of Hong Kong. Retrieved from https://hub.hku.hk/bitstream/10722/56255/1/ft.pdf?accept=1.

Lindblom, B. (1986). Phonetic universals in vowel systems. *Experimental Phonology*, 13-44.

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

Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: a study of voice-onset time in word-initial stop consonants. *Journal of Child Language*,

7(1), 41–74.

Maddieson, I. (1995). Gestural economy. *Proceedings of the 13th International Congress of Phonetic Sciences*, 574–577. Stockholm, Sweden.

Major, R. (1976). *Phonological differentatiation of a bilingual child* (Doctoral dissertation). The Ohio State University. Retrieved from https://eric.ed.gov/?id=ED149644.

McMurray, B., & Aslin, R. N. (2005). Infants are sensitive to within-category variation in speech perception. *Cognition*, *95*(2), B15–B26. doi: 10.1016/j.cognition.2004.07.005

McMurray, B., & Farris-Trimble, A. (2012). Emergent information-level coupling between perception and production. *The Oxford Handbook of Laboratory Phonology*, (October 2020), 1–26. doi: 10.1093/oxfordhb/9780199575039.013.0015

Mielke, J., & Nielsen, K. Y. (2018). Voice Onset Time in English voiceless stops is affected by following postvocalic liquids and voiceless onsets. *The Journal of the Acoustical Society of America*, *144*(4), 2166–2177. https://doi.org/10.1121/1.5059493

Millasseau, J., Bruggeman, L., Yuen, I., & Demuth, K. (2019). Durational cues to voicing and place contrasts in Australian English. *Proceedings of the 19th International Congress of Phonetic Sciences*, 3759–3762.

Millasseau, J., Bruggeman, L., Yuen, I., & Demuth, K. (2021). Temporal cues to onset voicing contrasts in Australian English-speaking children. *The Journal of the Acoustical Society of America*, *149*(1), 348–356. doi: 10.1121/10.0003060

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica*, *43*(1–3), 106–115.

Nearey, T. M., & Rochet, B. L. (1994). Effects of place of articulation and vowel context on VOT production and perception for French and English stops. *Journal of the International Phonetic Association*, *24*(1), 1–18.

https://doi.org/10.1017/S0025100300004965

Newman, R. S. (2003). Using links between speech perception and speech production to evaluate different acoustic metrics: a preliminary report. *The Journal of the Acoustical Society of America, 113*(5), 2850–2860. doi: 10.1121/1.1567280

Nieuwenhuis, R., te Grotenhuis, H. F., & Pelzer, B. J. (2017). Weighted effect coding for observational data with wec. *The R Journal, 9*(1), 477–485. doi: 10.32614/RJ-2017-017

Olive, J. P., Greenwood, A., & Coleman, J. (1993). *Acoustics of American English speech: a dynamic approach*. Springer Science & Business Media.

Oushiro, L. (2019). Linguistic uniformity in the speech of Brazilian internal migrants in a dialect contact situation. In S. Calhoun, P. Escudero, M. Tabain, & P. Warren (Eds.), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019* (pp. 686–690). Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Puggaard, R., & Goldshtein, Y. (2020). Realization and representation of plosives in Jutlandic varieties of Danish: variation in phonetics predicts variation in phonology. Paper presented at the 17th Annual Conference of Laboratory Phonology (LabPhon17), Virtual.

Quam, C., Knight, S., & Gerken, L. (2017). The distribution of talker variability impacts infants' word learning. *Laboratory Phonology, 8*(1).

Redford, M. A. (2019). Speech production from a developmental perspective. *Journal of Speech, Language, and Hearing Research, 62*(8S), 2946–2962. doi: 10.1044/2019_JSLHR-S-CSMC7-18-0130

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science, 12*(2), 339–349. https://doi.org/10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive variability in early word learning. *Infancy*, *15*(6). https://doi.org/10.1016/j.micinf.2011.07.011.Innate

Salesky, E., Chodroff, E., Pimentel, T., Wiesner, M., Cotterell, R., Black, A. W., & Eisner, J. (2019). A corpus for large-scale phonetic typology. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4526–4546. Association for Computational Linguistics.

Schmidt, R. A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn Jr, J. T. (1979). Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review*, *86*(5), 415.

Schwartz, J.-L., Basirat, A., Ménard, L., & Sato, M. (2012). The perception-for-action-control theory (PACT): a perceptuo-motor theory of speech perception. *Journal of Neurolinguistics*, *25*(5), 336–354. doi: 10.1016/j.jneuroling.2009.12.004

Schwartz, J.-L., Boë, L., & Abry, C. (2007). Linking the dispersion-focalization theory (DFT) and the maximum utilization of the available distinctive features (MUAF) principle in a perception-for-action-control theory (PACT). In M. J. Solé, P. S. Beddor, & M. Ohala (Eds.), *Experimental Approaches to Phonology* (pp. 104–124). Oxford University Press.

Schwartz, J.-L., & Ménard, L. (2019). Structured idiosyncrasies in vowel systems. *Unpublished Manuscript*. doi: 10.31219/osf.io/b6rdv

Smith, N. V. (1973). The acquisition of phonology: A case study. Cambridge University Press.

Smith, B. L., & Kenney, M. K. (1998). An assessment of several acoustic parameters in children's speech production development: longitudinal data. *Journal of Phonetics*, *26*(1), 95–108. doi: 10.1006/jpho.1997.0061

Smith, A., & Zelaznik, H. N. (2004). Development of functional synergies for speech motor coordination in childhood and adolescence. *Developmental Psychobiology*, *45*(1), 22–33. doi: 10.1002/dev.20009

Stokes, S. F., & To, C. K. S. (2002). Feature development in Cantonese. *Clinical Linguistics and Phonetics*, *16*(6), 443–459. doi: 10.1080/02699200210148385

Stuart-Smith, J., Sonderegger, M., Rathcke, T., & Macdonald, R. (2015). The private life of stops: VOT in a real-time corpus of spontaneous Glaswegian. *Laboratory Phonology*, *6*(3–4), 505–549. https://doi.org/10.1515/lp-2015-0015

Suomi, K. (1980). *Voicing in English and Finnish stops: a typological comparison with an interlanguage study of the two languages in contact*. Helsinki: Turun yliopiston suomalaisen.

Swartz, B. L. (1992). Gender difference in voice onset time. *Perceptual and Motor Skills*, *75*(3), 983–992.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. doi: 10.1016/S0010-0277(00)00081-0

Tanner, J., Sonderegger, M., & Stuart-Smith, J. (2020). Structured speaker variability in Japanese stops: relationships within versus across cues to stop voicing. *The Journal of the Acoustical Society of America*, *148*(2), 793–804. doi: 10.1121/10.0001734

te Grotenhuis, M., Pelzer, B., Eisinga, R., Nieuwenhuis, R., Schmidt-Catran, A., & Konig, R. (2017). When size matters: advantages of weighted effect coding in observational studies. *International Journal of Public Health*, *62*(1), 163-167.

Theodore, R. M., Miller, J. L., & DeSteno, D. (2009). Individual talker differences in voice-onset-time: contextual influences. *The Journal of the Acoustical Society of America*, *125*(6), 3974–3982. doi: 10.1121/1.3106131

Tse, H. (2019). Vowel shifts in Cantonese?: Toronto vs. Hong Kong. *Asia-Pacific*

Language Variation, 5(1), 67-83. doi: 10.1075/aplv.19001.tse

Tripp, A., Feldman, N. H., & Idsardi, W. J. (2021). Social Inference May Guide Early Lexical Learning. *Frontiers in Psychology*, *12*(May), 1–19. https://doi.org/10.3389/fpsyg.2021.645247

Turk, A., & Shattuck-Hufnagel, S. (2014). Timing in talking: what is it used for, and how is it controlled? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658), 20130395.

Vihman, M. M. (2017). Learning words and learning sounds: Advances in language development. *British Journal of Psychology*, *108*(1), 1-27.

Watt, D. J. L. (2000). Phonetic parallels between the close-mid vowels of Tyneside English: are they internally or externally motivated? *Language Variation and Change*, *12*(1), 69–101. doi: 10.1017/S0954394500121040

Werker, J., & Curtin, S. (2005). PRIMIR: a developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234. doi: 10.1207/s15473341lld0102_4

Whalen, D. H., Levitt, A. G., & Goldstein, L. M. (2007). VOT in the babbling of French- and English-learning infants. *Journal of Phonetics*, *35*, 341–352. doi: 10.1016/j.wocn.2006.10.001

Wolf, C. G. (1973). The perception of stop consonants by children. *Journal of Experimental Child Psychology*, *16*(2), 318–331. doi: 10.1016/0022-0965(73)90170-7

Yang, J. (2018). Development of stop consonants in three-to six-year-old Mandarin-speaking children. *Journal of Child Language*, *45*(5), 1091–1115. doi: 10.1017/S0305000918000090

Yao, Y. (2007). Closure duration and VOT of word-initial voiceless plosives in English in spontaneous connected speech. *UC Berkeley Phonology Lab Annual Report*, 183–225.

Yao, Y. (2009). Understanding VOT variation in spontaneous speech. In M. Pak (Ed.), *Current Numbers in Unity and Diversity of Languages* (pp. 1122–1137). Seoul: Linguistic Society of Korea.

Yuan, J., & Liberman, M. Y. (2008). Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics '08*, 5687–5790. doi: 10.1121/1.2935783

Zlatin, M. A. (1974). Voicing contrast: perceptual and productive voice onset time characteristics of adults. *The Journal of the Acoustical Society of America*, *56*(3), 981–994. doi: 10.1121/1.1903359

Zlatin, M. A., & Koenigsknecht, R. A. (1976). Development of the voicing contrast: a comparison of voice onset time in stop perception and production. *Journal of Speech and Hearing Research*, *19*(1), 93–111.

Zee, E. (2003). Frequency analysis of the vowels in Cantonese from 50 male and 50 female speakers. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 1117-1120). Universitat Autònoma de Barcelona. Barcelona.

Appendix

Table 11. Stimulus list of English words beginning with aspirated stops.

Word	IPA Transcription
teacher	[tʰiʧəɹ]
tickle	[tʰɪkəɬ]
tail	[tʰeɪɬ]
taste	[tʰeɪst]
tent	[tʰɛnt]
tongue	[tʰʌŋ]
taco	[tʰɑkoʊ]
tall	[tʰɔɬ]
torn	[tʰɔɹn]
toad	[tʰoʊd]
toast	[tʰoʊst]
tooth	[tʰuθ]
tube	[tʰub]
tuna	[tʰunə]
key	[kʰi]
kicking	[kʰɪkʰɪŋ]
kitchen	[kʰɪʧən]
ketchup	[kʰɛʧəp]
cake	[kʰeɪk]
cave	[kʰeɪv]
color	[kʰʌləɹ]
cutting	[kʰʌɾɪŋ]
car	[kʰɑɹ]
coat	[kʰoʊt]
cocoa	[kʰoʊkʰoʊ]
cone	[kʰoʊn]
cookie	[kʰʊkʰi]
cooking	[kʰʊkʰɪŋ]
cougar	[kʰugəɹ]

Table 12. Stimulus list of Cantonese words beginning with simple aspirated stops.

Word	IPA Transcription	English Translation
天空	[tʰiːn55 hʊŋ55]	‘sky’
甜品屋	[tʰiːm21 pɐn35 ʊk5]	‘dessert bar/patisserie’
貼紙	[tʰiːp3 ʧiː35]	‘sticker’
停車	[tʰɪŋ21 ʧʰɛː55]	‘stop the car’
聽音樂	[tʰɛːŋ55 jɐm55 ŋɔːk2]	‘listen to music’
踢波	[tʰɛːk3 pɔː55]	‘play football’
頭髮	[tʰɐu21 fat3]	‘hair’
探熱針	[tʰam33 jiːt2 ʧɐm55]	‘thermometer’
太陽	[tʰaːi33 jœːŋ21]	‘sun’
拖鞋	[tʰɔː55 haːi35]	‘slippers’
糖果	[tʰɔːŋ21 kʷɔː35]	‘sweets’
通心粉	[tʰʊŋ55 sɐm55 fɐn35]	‘macaroni’
虔誠	[kʰiːn21 sɪŋ21]	‘pious’
揭開	[kʰiːt3 hɔːi55]	‘flip open’
翹住手	[kʰiːu23 ʧy22 sɐu35]	‘folding arms’
傾偈	[kʰɪŋ55 kɐi35]	‘chat’
企鵝	[kʰei23 ŋɔː35]	‘penguin’
騎膊馬	[kʰɛː21 pɔːk3 ma23]	‘piggyback/sit on shoulders’
球場	[kʰɐu21 ʧʰœːŋ21]	‘sportsground’
扣鈕	[kʰɐu33 lɐu35]	‘button up’
卡通片	[kʰa55 tʊŋ55 pʰiːn35]	‘cartoon’
鈣片	[kʰɔːi33 pʰiːn35]	‘calcium tablet’
窮人	[kʰʊŋ21 jɐn21]	‘poor person’
曲奇餅	[kʰʊk5 kʰei21 pɛːŋ35]	‘cookie’
潰爛	[kʰuːy35 lan22]	‘fester’
繪畫樂	[kʰuːy35 wa35 lɔːk2]	‘the joy of drawing’
箍牙	[kʰu55 ŋa21]	‘wear braces’