JASA ARTICLE

# Adult perception of stop consonant voicing in American-English-learning toddlers: Voice onset time and secondary cues[a)]



Elaine R. Hitchcock[1,b)] and Laura L. Koenig[2,c)]
[1]*Communication Sciences and Disorders, Montclair State University, Upper Montclair, New Jersey 07043, USA*
[2]*Communication Sciences and Disorders, Adelphi University, Garden City, New York 11530, USA*

**ABSTRACT:**
Most studies of speech perception employ highly controlled stimuli. It is not always clear how such results extend to the processing of natural speech. In a series of experiments, we progressively explored the role of voice onset time (VOT) and potential secondary cues in adult labeling of stressed syllable-initial /b d p t/ produced by typically developing two-year-old learners of American English. Taken together, the results show the following: (a) Adult listeners show phoneme boundaries in labeling functions comparable to what have been established for adult speech. (b) Adult listeners can be sensitive to distributional properties of the stimulus set, even in a study that employs highly varied naturalistic productions from multiple speakers. (c) Secondary cues are available in the speech of two-year-olds, and these may influence listener judgments. Cues may differ across places of articulation and the VOT continuum. These results can lend insight into how clinicians judge child speech during assessment and also have implications for our understanding of the role of primary and secondary acoustic cues in adult perception of child speech.
 https://doi.org/10.1121/10.0005595

(Received 13 October 2020; revised 10 June 2021; accepted 21 June 2021; published online 21 July 2021)

[Editor: Jody Kreiman] Pages: 460–477

## I. INTRODUCTION

In studies of human speech perception, the most common experimental method has been to employ synthetic speech varying in one or a few acoustic parameters, using values that align with typical adult speech. How well do the results of such work extend to perception of natural speech? This is a general question of external validity; it also has considerable practical relevance. In particular, routine clinical assessment by speech-language pathologists continues to rely heavily on auditory assessment of a client's speech production. As will be described below, some past developmental work has effectively assumed that adult perception of child speech could be well-predicted from studies of adults listening to synthetic speech. The current work sought to assess the legitimacy of that assumption.

Specifically, this paper presents a series of three experiments examining adult perception of stop consonants produced by typically developing two-year-old native learners of American English. The work grew out of an earlier study, Hitchcock and Koenig (2013), which charted the emergence of the stop voicing contrast in the same set of children using a battery of measures. One of those measures, accuracy, was based on how we hypothesized adults would label children's productions, following classic acquisitional studies (see Sec. I A). The current experiments used the 2013 production data to explore directly how adults label children's stop voicing.

### A. Voice onset time (VOT) in developing speech

One of the most widely researched acoustic measures of young child speech is VOT (Lisker and Abramson, 1964). Dozens of authors have used VOT to characterize the emergence of the stop voicing contrast in multiple languages and in typically developing children as well as clinical populations.

A summary of VOT studies in English-speaking children is provided in Hitchcock and Koenig (2013). Early work (e.g., Kewley-Port and Preston, 1974; Preston *et al.*, 1968) suggested that children's first stop productions were predominantly voiceless unaspirated (short-lag VOT, about 0–20 ms). Thus, for children acquiring American English, the main developmental task is to learn to produce aspiration, or long voicing lags (about 40+ ms), as required for

[a)]Portions of this work were presented in "Secondary acoustic cues in adult perception of young children's stop productions," 177th Meeting of the Acoustical Society of America, Louisville, Kentucky, USA, May 2019; "Analyzing adult perception of stop consonant voicing in two-year-old children's speech," Annual Meeting on Phonology, New York, New York, USA, September 2017; "The role of voice onset time in the perception of English voicing contrasts in children's speech," 173rd Meeting of the Acoustical Society of America, Boston, Massachusetts, USA, June 2017; "Adult perception of stop consonants in two-year-old children's speech," American Speech-Language-Hearing Association National Convention, Denver, CO, USA, November 2015; and "Adults' perceptual voicing boundaries of 2-year-olds' citation form speech," 170th Meeting of the Acoustical Society of America, Jacksonville, Florida, USA, October 2015.

[b)]Electronic mail: hitchcocke@montclair.edu, ORCID: 0000-0003-1835-0976.
[c)]Also at: Haskins Laboratories, New Haven, CT 06511, USA, ORCID: 0000-0003-0155-7866.

https://doi.org/10.1121/10.0005595

stressed initial /p t k/. Henceforth, we will use the terms "short" and "long" VOT to refer to the expected VOT values for American English /b d/ and /p t/, respectively. (We will consider the issue of prevoicing in /b d/ later on.)

Macken and Barton (1980) proposed a three-stage sequence of VOT development for American English. In the first stage, the child shows unimodal VOT values in the short-lag range for all stops, and in the third stage, VOT values are bimodally distributed in short- and long-lag ranges for /b d g/ and /p t k/, respectively. Of particular interest is the intervening stage II, where VOT productions for contrasting categories take on a bimodal distribution within the short-lag range. That is, the children demonstrate a phonemic distinction that is not phonetically appropriate for the language. This is an example of "covert contrast," a phenomenon that has now been observed for multiple distinctions in child speech and which may in fact be widespread in development [for a recent summary, see Schellinger *et al.* (2017)].

To refer to a contrast as covert implies that adult listeners may not perceive it, and indeed stage II has been characterized as subperceptual. Specifically, Macken and Barton (1980) (pp. 49–50) stated that stage II distinctions should at least be difficult for adults to perceive and that they themselves did not believe they could differentiate between such bimodally distributed short-lag stops. Along similar lines, Kewley-Port and Preston (1974) (p. 209) supposed that adults would judge child productions of target /t/ with VOT values less than about 40 ms as phonemically ambiguous.[1] In short, these authors—while recognizing the possible influence of secondary cues to voicing—proposed that adult listeners would primarily label children's productions with reference to the VOT values that characterize the adult contrast. Hitchcock and Koenig (2013) followed in this tradition by calculating accuracy for children's stops based on their VOT values. However, several considerations subsequently led us to revisit the assumptions underlying this specific decision and adult perception of child speech more generally. We describe these considerations in Sec. I B.

### B. VOT in perception: Contextual variation, task type, secondary cues, and within-category differences

Most investigations of VOT as a perceptual cue have used synthetic speech, with a fixed vowel duration (VD) and frequencies [fundamental frequency (f0), formants] typical for an adult male speaker. In fact, stimuli developed for early investigations, such as Lisker and Abramson (1967b), were frequently borrowed by other authors (cf. Pisoni *et al.*, 1982; Zlatin and Koenigsknecht, 1975). Synthetic speech has obvious benefits: It allows for VOT values to be controlled with high precision and permits assessing the effect of a single acoustic variable without the influence of other factors that may co-vary in natural speech.[2] Early work in this area, following the model of studies such as Liberman *et al.* (1957) and Liberman *et al.* (1961), asked listeners to perform labeling and/or discrimination tasks along a continuum. The 50% crossover point between contrasting categories in labeling tasks was taken as the phoneme boundary, and higher discrimination accuracy was observed in the same region of the continuum as the phoneme boundary (see also Abramson and Lisker, 1967; Lisker and Abramson, 1967b). Elevated discrimination at the phoneme boundary and, conversely, reduced discrimination for within-category differences, represent evidence of categorical perception (Liberman *et al.*, 1967). In the strongest version of this phenomenon, listeners cannot discern differences between tokens that fall within a phonemic category. In subsequent years, authors have covaried VOT with other acoustic parameters to explore the role of secondary cues. Researchers have also expanded on the types of listening tasks that participants are asked to perform. Both of these lines of work indicate a somewhat more complex picture of the perception of voicing and aspiration, i.e., VOT.

#### 1. VOT variation

It was clear early on that VOT values in production varied as a function of many factors. In the current context, one of the most important of these is speech rate. Lisker and Abramson (1967a) observed shorter VOTs for English /p t k/ (i.e., shorter aspiration duration) in connected speech as compared to single words. This parallels the general finding that speech durations become shorter as the length of the production unit increases (e.g., from monosyllables to multisyllabic words; cf. Lehiste, 1972). The short-lag VOT category does not appear to vary much with speech rate, however (Kessinger and Blumstein, 1997; Miller *et al.*, 1986). Some work has suggested that durational modifications, or simulated changes in speech rate, can shift VOT perceptual boundaries somewhat (Miller *et al.*, 1986; Miller and Volaitis, 1989), although the manner in which durational variation is implemented may impact whether or not this result is obtained (Kessinger and Blumstein, 1998; Nakai and Scobbie, 2016).

It has been widely documented that children, on average, have slower speech rates than adults (e.g., Eguchi and Hirsh, 1969, Lee *et al.*, 1999; Smith *et al.*, 1996). Thus, overall durational differences between children and adults could affect the locations of VOT phoneme boundaries for child speech. More generally, listeners may adjust their expectations based on what they know about the speaker (e.g., child or adult, typical vs mis-articulating child; cf. Munson *et al.*, 2010). All of these considerations suggest that it is prudent to carry out a direct assessment of whether adults show comparable phoneme boundaries for child vs adult stop productions.

#### 2. Listening tasks

It has become increasingly evident that listeners can perceive not only differences between phonemes, but also within-category variation (e.g., Miller *et al.*, 1997; Wayland *et al.*, 1994). When people are asked to provide gradient judgments instead of, or in addition to, phonemic labels, they may demonstrate an awareness of subphonemic

acoustic differences. For example, Munson *et al.* (2010) selected fricative productions from a database of child speech (see Edwards and Beckman, 2008) that spanned the range from [s] to [θ], including productions transcribed as accurate (e.g., [s] for /s/), clear substitutions (e.g., [θ] for /s/), or sounds intermediate between the two (e.g., [s]:[θ], i.e., between [s] and [θ] but closer to [s]). Adult participants (students in speech-language pathology programs) used a visual analogue scale to rate each fricative on a scale with [s] at one end and [θ] at the other. The results showed graded listener responses, with significant differences among all transcription categories. Thus, these listeners demonstrated sensitivity to within-category variation.

It should be acknowledged that early authors did recognize limits on the degree to which speech perception is categorical (e.g., Liberman *et al.*, 1957; Liberman *et al.*, 1961). Pisoni and Lazarus (1974) demonstrated that within-category VOT discrimination improved under task conditions that reduced memory load. Pisoni *et al.* (1982) also observed that most untrained English-speaking listeners showed a discrimination peak between the prevoicing and voiceless unaspirated VOT ranges—smaller than the (phonemic) one between short- and long-lag VOT but greater than chance. Those authors argued that some types of experimental tasks could lead listeners to adopt a phonological strategy and concluded that adult speakers of American English retain the sensory mechanisms to detect the non-phonemic difference between prevoiced and voiceless unaspirated stops. Munson *et al.* (2010) have likewise pointed to the role of task differences in predisposing participants to adopt more versus less categorical listening strategies.

### 3. Secondary cues

In adult speech, the laryngeal actions involved in producing stop voicing distinctions can yield variations in acoustic parameters in addition to VOT (e.g., Abramson and Lisker, 1985; Lisker and Abramson, 1964), and such parameters may serve as additional cues to voicing judgments. For example, abducting the vocal folds may stretch them a bit (Sawashima *et al.*, 1981). Speakers may also learn that contraction of the cricothyroid muscle can assist in devoicing (Löfqvist *et al.*, 1989). Both of these actions could lead to an increased f0 as a concomitant of devoicing and/or aspiration, and cross-linguistic studies show such effects [see review in Kirby and Ladd (2016)]. The perceptual import of these additional or "secondary" cues to voicing has been explored extensively for adult speech. Multiple studies indicate that f0 differences contribute to voicing judgments (e.g., Haggard *et al.*, 1970; Whalen *et al.*, 1993). In natural speech, a delay in voicing onset also co-occurs with a higher onset value of first formant frequency (F1) (Lisker *et al.*, 1977), whereas F1 shows a more extensive transition to its steady-state value when voicing begins shortly after release (e.g., Cooper *et al.*, 1952). Studies separately varying VOT and the "F1 cutback" indicate that the latter also affects listener judgments (Lisker *et al.*, 1977; Soli, 1983; see also Stevens and Klatt, 1974). Finally, higher intraoral pressures in voiceless compared to voiced consonants (e.g., Arkebauer *et al.*, 1967) lead to higher burst amplitudes (i.e., intensities) in voiceless sounds. Data on whether burst intensity (BI) serves as a secondary cue are sparse and mixed; Sundara (2005) reported significant voicing effects of BI on listener judgments for Canadian French but not Canadian English.

Much less work has explored secondary cues in speech produced by children (but see Sec. I B 4 for two examples). It is important to note that one cannot assume that young children have fully mastered the constellation of gestures that adult speakers may use in achieving VOT differences. For example, whereas adult speakers may concurrently abduct the vocal folds and contract the cricothyroid muscle to produce a voiceless aspirated stop, two-year-olds are probably still learning the effects of those two articulatory actions, both separately and together. Some studies of covert contrast have also indicated that children may employ "secondary" or atypical cues to signal a phonemic contrast (Li *et al.*, 2009). Thus, the possible roles of variables such as f0, F1 cutback, and BI in adult perception of child productions remain poorly specified.

### 4. Within-category discrimination and secondary cues in child speech

Two studies of American native-English-learning children with phonological disorders have demonstrated that (a) listeners may be able to use small variations of VOT to differentiate among child targets and (b) secondary cues may play a role in adult judgments of child speech. Maxwell and Weismer (1982) described a child, age three years 11 months, who had extreme neutralization of phonological contrasts, with 20 different stops and clusters being transcribed as /d/. Acoustic measures showed that he had a three-way production distinction based on target syllable onsets: Average VOTs were 9 ms for words with initial /d/; 24 ms for words with other voiced sounds/clusters; and 34 ms for words with initial voiceless sounds/clusters. When naive listeners were asked to identify the initial consonant, their responses were usually /d/, as expected based on the transcriptions made during assessment. However, the graduate student clinicians who had worked with the child showed significantly more responses of /t/ than /d/ for the target voiceless sounds, with their slightly longer VOTs. This suggests that listeners can, under some circumstances, recognize or learn to recognize VOT differences that do not correspond to the typical adult pattern.

Forrest and Rockman (1988) collected initial stop consonant productions from three children, 3–4 years old, who had been diagnosed with phonological delay and were perceived to make voicing errors. The authors measured VOT and had listeners rate consonant voicing on a seven-point scale. For isolated words,[3] two of the three children had a significant VOT difference between voiced and voiceless targets on the order of 15 ms, and one had no VOT distinction. A comparison of listener ratings and VOT median values showed modest correlations, with $r^2$ values of 0.31–0.43

https://doi.org/10.1121/10.0005595

across the three children. To gain insight into cues other than VOT that might have influenced listener judgments, Forrest and Rockman selected tokens whose voicing judgments did not correspond well to their VOT values and assessed burst amplitude, aspiration amplitude, onset f0, low-frequency energy preceding voicing, and the F1 onset frequency and transition duration. Entering these measures along with VOT into a multiple regression yielded $r^2$ values of 0.19–0.69 across children, indicating that secondary cues could improve prediction of listener performance. For these data, the secondary cue that seemed most important was F1, although voicing judgments for some individual productions could be explained by other factors as well (e.g., BI or onset f0). The authors suggested that children with phonological disorders might be able to control acoustic features besides VOT to signal phonemic voicing contrasts, and that such differences could be recognized by some listeners.

### C. Observations and current work

The literature reviewed above indicates the following: (a) VOT category boundaries may vary as a function of speech rate and speaker characteristics; (b) listeners may demonstrate sensitivity to within-category variation in certain kinds of listening tasks; (c) secondary cues may contribute to voicing judgments; and (d) limited data from the clinical speech literature reveal some level of adult sensitivity to atypical phonemic contrasts in children. In light of these observations, for this work, we drew on our previously collected child production data to perform a series of perception experiments using adult listeners. The first two experiments were rather exploratory. The third one built on the first two to construct a carefully controlled listening experiment. For the third experiment, we also assessed possible contributions of secondary cues. Table I summarizes the questions and methods of the three experiments.

In all cases, the listeners were graduate students of speech-language pathology. Since we hoped to gain insight into the perceptual abilities that clinicians bring to assessment tasks, the fact that the students had some phonetic training was not considered to be a drawback. There is some evidence that trained listeners are more reliable in their judgments than untrained ones (Munson *et al.*, 2012), and they may be more sensitive to fine phonetic distinctions as well (Maxwell and Weismer 1982; Munson *et al.*, 2012). As stated earlier, perceptual judgments remain a cornerstone of clinical assessment, and implementing an effective intervention program relies on an accurate assessment. Our aim, therefore, was to ascertain how the labeling behavior of (somewhat) experienced listeners relates to standard acoustic measures that have been found to correlate with the stop voicing distinction. Establishing such correspondences is particularly important in light of the growing awareness of covert contrast as a developmental phase and greater availability of acoustic analysis methods in clinical settings. Cases of covert contrast provide insight into a child's developing system, have prognostic value (Forrest *et al.*, 1990; Tyler *et al.*, 1990), and lie at the heart of the phonetics–phonology interface insofar as they represent a (possibly nascent) phonemic contrast that is not appropriately instantiated phonetically (cf. Eckman *et al.*, 2015). Although we use more controlled listening tasks than encountered in standard clinical settings, the results can help establish the limits of what clinicians might be expected to perceive regarding a child's voicing errors or emerging voicing contrast.

TABLE I. An outline of unique and overlapping criteria across experiments I, II, and III.

| | Experiment I | Experiment II | Experiment III |
|---|---|---|---|
| Question | Are adult phoneme boundaries for child speech similar to those obtained for synthetic speech based on adult values? | Is adult labeling affected by the child's intended target and within-category stimulus distribution? | Is adult labeling affected by the child's intended target and within-category stimulus distribution? |
| *N* child speakers | 10 | 6 | 6 |
| VOT continua | 0–100 ms | −169–488 ms, i.e., including prevoicing and exaggerated aspiration | 0–100 ms |
| | <30 ms: /b d/ targets; >30 ms: /p t/ targets | Equal numbers of vcd/vcls targets overall, but /b d/ targets were mostly prevoiced and most /p t/ targets had very long aspiration | Equal numbers of vcd/vcls targets within each 10-ms bin |
| No. of categories | 10 categories: 10 ms bins from 0–100 ms | 3 categories: prevoiced and short, ambiguous, long | 3 categories: short, ambiguous, long |
| Category definitions | | Appropriate for /b d/: <20 ms; Ambiguous: 20–30 ms; Appropriate for /p t/: >30 ms | Appropriate for /b d/: 0–20 ms; Ambiguous: 20–30 ms; Appropriate for /p t/: 30–100 ms |
| Tokens per category/per speaker | 2 | 4 | 4 |
| *N* tokens | 200 (100 per POA) | 288 (144 per POA) | 288 (144 per POA) |
| *N* presentations per token | 2 | 2 | 2 |
| *N* listeners | 20 | 20 | 20 |
| Total ratings | 8000 | 11 520 | 11 520 |

### D. Production data

Stimuli were selected from the database collected for Hitchcock and Koenig (2013). In that work, ten typically developing two-year-old children were recorded every other week for a total of eight sessions, i.e., over a four-month period. Child ages at the beginning of the experiment were 27–29 months. Parents were given pictures of four target words (*doe, toe, boo, Pooh*) ahead of time to familiarize the children with the stimuli. Recording sessions targeted a minimum of 15 analyzable tokens of each of the four stops and used the pictures paired with a verbal prompt (e.g., *The ghost says…*) to encourage single-word production of the targets. For analysis, tokens were discarded based on interference (noise, yelling), if syllables did not begin with a stop at the appropriate place of articulation (POA), and/or if they were produced in direct imitation.

The production analysis included 6339 tokens, an average of ca. 20 productions per target phoneme (/b/, /p/, /d/, /t/) for each child and session. VOT was measured for all productions using an acoustic waveform and wideband spectrogram in the PENTAX Medical (Montvale, NJ) Computerized Speech Lab (model 4500) program. Tokens for the current work were chosen from this dataset according to criteria that varied by experiment, as described below (see also Table I).

## II. EXPERIMENT I

### A. Purpose

In the first experiment, we asked whether adults listening to child speech would show phoneme boundaries (50% crossover values in the labeling functions) comparable to those previously obtained for synthetic speech based on adult values (see the Appendix). Alternatively, phoneme boundaries for /b p/ and /d t/ could be shifted to longer VOT values for children's speech given their typically longer speech durations (Miller and Volaitis, 1989). Exploratory experiment I was intended to be a simple test of the assumptions behind the classic studies of Kewley-Port and Preston (1974) and Macken and Barton (1980) and, by extension, the work of Hitchcock and Koenig (2013), wherein the authors supposed that adult labeling of naturally produced child speech would follow the English phoneme boundaries previously established in the literature regardless of the child's intended target. All of those papers acknowledged the possibility of secondary cues but nevertheless assumed that VOT values would primarily drive adult phoneme judgments.

### B. Methods

#### 1. Listeners

Participants were 20 monolingual English speakers, ages 22–42 years [mean = 25 years; standard deviation (SD) = 5 years]. All were graduate students in a speech-language pathology program; had pure-tone thresholds of 20 dB at 500, 1000, 2000, and 4000 Hz based on bilateral audiometric screening; and were blind to the purpose and hypothesis of the experiment.

#### 2. Stimuli, presentation, and scoring

For all ten children in the original database, two productions were chosen for each 10 ms VOT bin between 0 and 100 ms, leading to a total of 200 productions at each POA, fully balanced across children. All tokens less than 30 ms VOT corresponded to target /b d/; all tokens greater than 30 ms VOT corresponded to target /p t/.[4] (To improve readability, throughout the text, we will round VOT values; the actual cutoff for /b d/ tokens here was 29.99 ms, whereas that for /p t/ was 30 ms. Analogous boundary definitions hold for subsequent experiments.) Stimuli were presented on an IBM Lenovo Thinkpad L420 using a Praat presentation script (Boersma and Weenink, 2010). Presentation was blocked by POA, so that listeners completed two separate labeling tasks. Within blocks, each token was presented twice, and stimulus items were fully randomized. (Presenting more than one repetition per stimulus increases the number of tokens for statistical analysis and has considerable precedent in the literature; see, e.g., Lisker and Abramson, 1970; Nittrouer and Studdert-Kennedy, 1987; Pisoni and Lazarus, 1974; Whalen *et al.*, 1993). Thus, 8000 listener judgments were obtained for each POA (200 tokens × 2 repetitions × 20 listeners). During the listening task, participants wore Sennheiser (Wedemark, Germany) HD280 Pro closed-ear, circumaural headphones and were seated in a sound-treated booth. Tokens could be replayed up to two times, and listeners chose their response by clicking on one of two phoneme options: /b/ /p/ or /d/ /t/.[5] Time to complete each listening block (bilabial or alveolar) was about 15–20 min. Responses were scored as accurate if the listener judgment matched the word elicited from the child. [Note that this differs from the definition of accuracy in Hitchcock and Koenig (2013), where we scored child productions as accurate based on what we thought the adult label would be.]

### C. Results and discussion

Labeling functions, combining data across all listeners, are presented in Fig. 1.[6] The figure shows 50% crossovers, i.e., the *x* axis intercept where uncertainty is maximal (the "phoneme boundary"). We also carried out a probit analysis for both POAs to estimate that boundary more precisely (Coady *et al.*, 2005). That analysis yielded 50% crossovers at 28.13 ms for bilabials and 30.18 ms for alveolars. These values are comparable to what previous authors have observed (cf. Appendix).

The data also allowed us to evaluate how accurately the children's productions were judged. Results, averaged across each 10-ms bin, are given in Table II. In all cases but three, listeners were essentially at ceiling in labeling the child's productions in accordance with the target. The three exceptions fall in the range of 20–30 ms for bilabials and 20–30 and 30–40 ms for alveolars, where inaccuracy rates

https://doi.org/10.1121/10.0005595

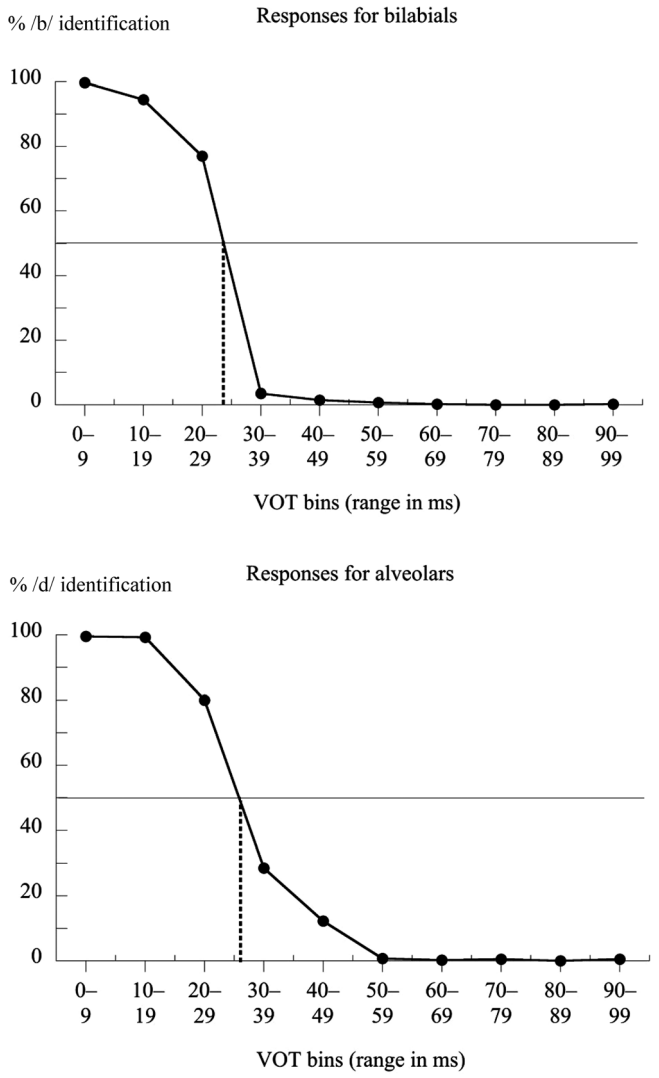


FIG. 1. Labeling functions for bilabials (top) and alveolars (bottom). Combined data for 20 listeners. The dotted vertical line marks the location of the 50% crossover. To improve readability, the *x* axes suppress decimal places at the upper (right) edge of each bin, e.g., the bin 0–9.99 ms is represented as 0–9 ms in the axis label.

TABLE II. Percent accuracy (i.e., adult labels that matched the child's target) for experiment I showing combined data for 20 listeners. As in Fig. 1, we suppress decimal places; the first VOT bin was actually 0–9.99 ms, and likewise henceforth. Note that all productions in the range 0–29 ms correspond to targets /b d/; 30–100 ms correspond to targets /p t/.

| Target | VOT bin (ms) | Bilabials (%) | Alveolars (%) |
|---|---|---|---|
| Voiced | 0–9 | 99.8 | 99.6 |
| Voiced | 10–19 | 94.5 | 99.3 |
| Voiced | 20–29 | 77.0 | 80.0 |
| Voiceless | 30–39 | 96.5 | 81.7 |
| Voiceless | 40–49 | 98.5 | 97.0 |
| Voiceless | 50–59 | 99.3 | 99.3 |
| Voiceless | 60–69 | 99.8 | 99.8 |
| Voiceless | 70–79 | 100.0 | 99.5 |
| Voiceless | 80–89 | 100.0 | 100.0 |
| Voiceless | 90–99 | 99.8 | 99.6 |

were about 20%. The 20 listeners were fairly consistent in their overall accuracy ratings (range across listeners of 94%–98% accuracy). Further, listeners were quite reliable in labeling the two presentations of each word with the same phoneme (96.5%; 3856/4000).

Taken together, these crossover functions and accuracy results lend general support to the premise that adult perceptual judgments of child speech are largely similar to what have been reported in studies using synthetic speech based on adult values. More specifically, the crossover was not shifted to longer VOTs, as might occur if longer speech durations affected listener judgments of child speech. Cases where adults misjudged the children's targets were, as one would expect, mostly in the regions of maximum uncertainty based on the VOT crossover functions.

Of course, these data were drawn from naturally produced speech, where a variety of secondary cues may have contributed to listener judgments. VOT values also covaried with target, in that shorter VOT values corresponded to target /b d/ and longer values corresponded to target /p t/. Secondary cues will be addressed in experiment III. The main goal for experiment II was to include targets that were and were not appropriate to the target along the VOT continuum.

## III. EXPERIMENT II

### A. Purpose

In contrast to experiment I, experiment II included equal numbers of cases where the VOT value did and did not match the target, i.e., tokens of target /p t/ that had negative or short-lag VOTs and tokens of /b d/ that had long-lag VOTs. We expected that productions with VOTs less than about 25–30 ms would tend to be labeled as voiced, and those greater than about 30 ms would tend to be labeled as voiceless, regardless of the intended voicing category. On the other hand, secondary cues could mitigate the effects of ambiguous or inappropriate VOTs, so that listeners might show higher-than-expected accuracy in judging the child's target in these cases.

For exploratory experiment II, we also employed a larger range of VOT values than in experiment I. This was of interest for two reasons. First, Macken and Barton (1980) observed very long-lag values of VOT ("overshoot") for some children, and Hitchcock and Koenig (2013) found that all of the ten children evaluated showed at least a brief period of overshoot. Second, although the Hitchcock and Koenig (2013) data generally supported the conventional claim that early stop productions tend to be in the short-lag VOT range, the dataset did contain examples of prevoicing for all children, and some children used it extensively in several sessions. Thus, tokens in experiment II were chosen to expose listeners to some of the extreme values found in child speech.

### B. Methods

#### *1. Listeners*

Participants were 20 monolingual English speakers, ages 22–33 years (mean = 26 years; SD = 4 years). As before, all

were graduate students in a speech-language pathology program; had pure-tone hearing thresholds of 20 dB at 500, 1000, 2000, and 4000 Hz based on bilateral screening; and were blind to the purpose and hypothesis of the experiment.[7] None of the listeners in experiment II participated in experiment I.

### 2. Stimuli, presentation, and scoring

Our initial intent was to select stimuli from all ten children in Hitchcock and Koenig (2013), as in experiment I. However, some children had insufficient productions with inappropriate VOT values (e.g., target /p/ with short-lag VOT). Recognizing that individual child speakers might have unique patterns, we sought to ensure that as many children as possible were represented in a balanced fashion across the VOT range. As a result, tokens selected for this experiment came from six of the ten children. For each of these six, we were able to obtain four tokens of each target phoneme in three VOT ranges: Appropriate for /b d/ (<20 ms); appropriate for /p t/ (>30 ms), and ambiguous (20–30 ms). In eight cases (of 96) we allowed up to 3 ms error for "ambiguous" tokens, so that, e.g., a production of /t/ with VOT = 32 ms was included in that category. This provided us with 24 tokens per VOT category, a fairly wide range of productions for adults to label.

This selection procedure yielded 288 productions (144 per POA) for the perceptual testing (6 children × 3 VOT ranges × 4 words × 4 productions). Some very long VOTs at both the long (aspirated) and negative (prevoiced) ends of the continuum were included (see Fig. 2) to explore whether such within-phoneme differences are perceptually salient to adult listeners.

The listening setup was the same as in experiment I. Tokens were randomized within blocks of bilabial and alveolar productions, with two presentations per token, yielding 11 520 ratings (288 bilabial and alveolar tokens × 2 presentations × 20 listeners). The time to complete the task for each POA was about 20 min. Scoring was as in experiment I as well, i.e., accuracy scores reflect whether adult labels matched the child's intended production, regardless of whether or not VOT was appropriate for the target.

## C. Results and discussion

Across all data, individual listener accuracy ranged from 62% to 70%. Listeners were again quite reliable in labeling repeated presentations identically (90.5%, or 5215/5760 of all pairs).

To organize the results, we present listener responses for three situations:

(1) Appropriate VOTs: In this experiment, these are productions of /p t/ with long-lag VOTs and productions of /b d/ with short-lag VOTs or prevoicing.
(2) Inappropriate VOTs: Productions of /p t/ with prevoicing or short-lag VOTs and productions of /b d/ with long-lag VOTs.
(3) Ambiguous VOTs: Values of VOT that are around the phoneme boundary, in the region of 20–30 ms VOT.

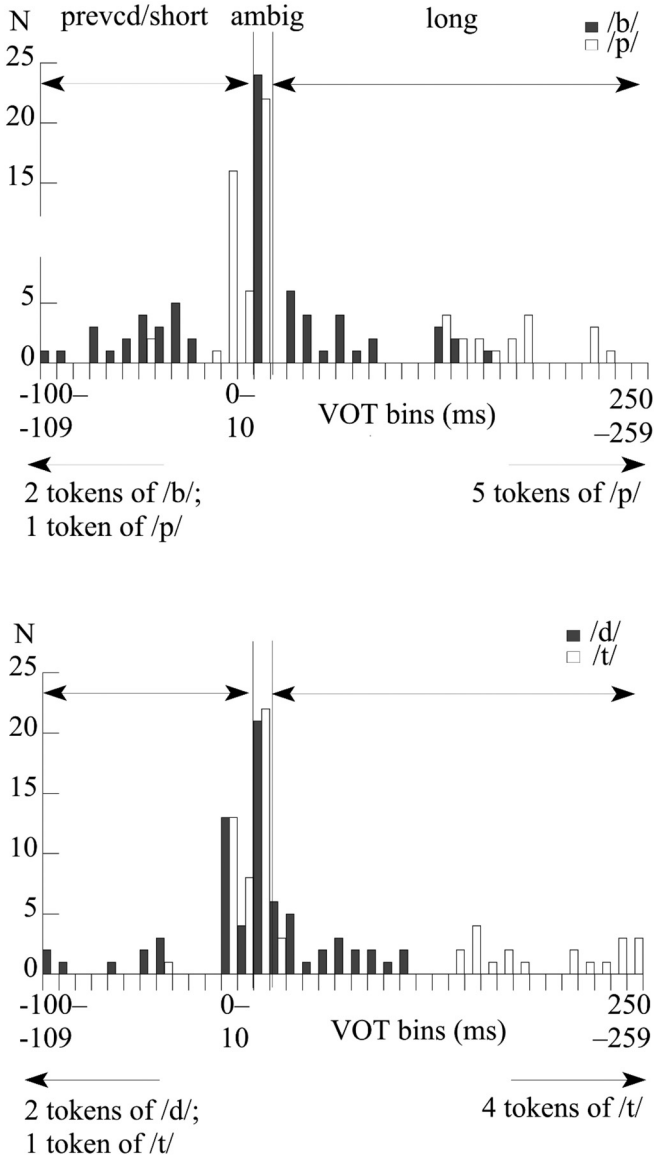


FIG. 2. VOT distributions of /b p/ and /d t/ productions used in experiment II. As indicated, a few extreme outliers had VOTs beyond the scale shown.

### 1. Labeling for appropriate VOTs

Accuracy data are provided in Fig. 3. As in experiment I, adults labeled children's productions at near-ceiling levels of accuracy when the VOT values were appropriate for the target: 99.4% for /b/ and 95.1% for /d/ with VOTs < 20 ms, and 99.7% for /p/ and 92.0% for /t/ with VOTs > 30 ms. Chi-square tests showed all of these to differ from chance at $p < 0.001$.

### 2. Labeling for inappropriate VOTs

If VOT values are a strong determiner of listener judgments, one would expect low accuracy rates for target /b d/ with long VOTs and target /p t/ with prevoicing or short VOTs. This was the case for target /p t/ with prevoicing or

https://doi.org/10.1121/10.0005595

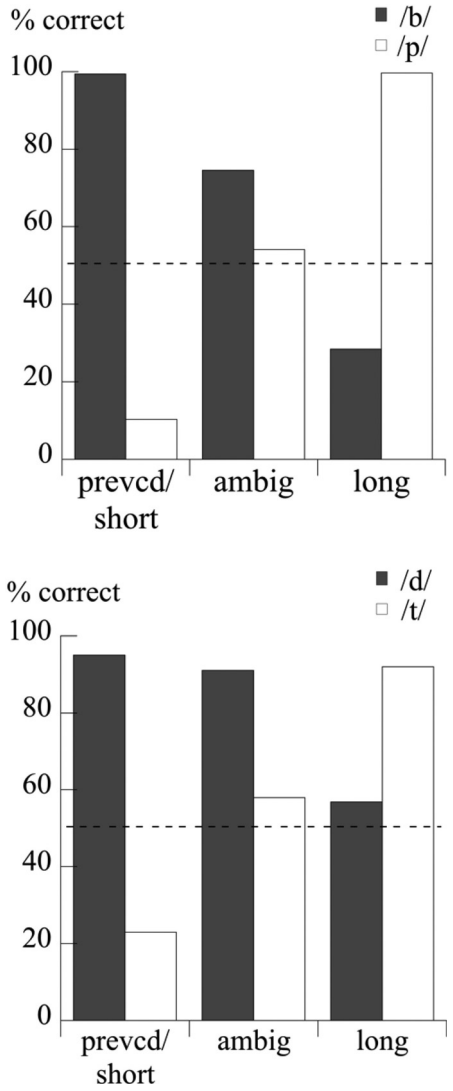


FIG. 3. Accuracy data for bilabials (left) and alveolars (right) for experiment II. Response data were combined for the 20 listeners. The *x* axes show the VOT category of the child's production. The prevcd/short-lag categories include both short-lag (0–20 ms) and prevoiced tokens (<0 ms). The long-lag categories were aspirated, i.e., had VOT values greater than 30 ms. The dotted line at 50% indicates chance responding.

short VOTs (10.3% and 22.9%, respectively; $p < 0.001$) as well as target /b/ with long VOTs (28.4%, $p < 0.001$). Interestingly, cases of target /d/ with long VOTs were labeled as /d/ at greater than chance levels (56.8%, $\chi^2 = 17.604$, $p < 0.001$).

### 3. Labeling for ambiguous VOTs

For VOTs in the ambiguous range, we expected that listeners might respond essentially at chance. On the contrary, the data show that listeners were rather accurate for target /b d/ in this range, at 74.6% and 91.0%, respectively ($p < 0.001$ in both cases). Accuracy was lower for voiceless targets, but listeners still performed better than chance: For /p/, 54.1% ($\chi^2 = 6.3375$, $p < 0.012$), and for /t/, 57.9% ($\chi^2 = 24.067$, $p < 0.001$).

### 4. Possible reasons for high accuracy with ambiguous and inappropriate VOTs

One sensible explanation for the unexpectedly high accuracy in cases of ambiguous VOTs is that listeners drew on secondary cues. For target /d/ with long-lag values, it could be that listeners were actually relying on secondary cues more than VOT. Secondary cues will be evaluated in experiment III. That said, an alternative (or additional) interpretation is also possible for this dataset.

As indicated earlier, one aim of experiment II was to include some of the extreme VOT values that are produced by typical two-year-olds. In the original production dataset, when children used prevoicing, it was usually for target /b d/, i.e., in productions of *boo* and *doe*. Likewise, productions with exaggerated aspiration were largely seen for productions of *Pooh* and *toe*. Thus, the decision to include productions at the extreme ends of the VOT range led to a situation where a number of /b d/ productions had prevoicing, whereas all but a few productions of /p t/ with inappropriate VOTs were short lag. Conversely, in the long lag range, intended /p t/ productions tended to have longer VOTs than intended /b d/. In fact, inspection of Fig. 2 shows that voiced and voiceless targets were essentially bimodally distributed within the ranges appropriate for /b d/, viz., prevoiced or short-lag VOT, and those appropriate for /p t/, viz., long lag.

These distributional characteristics of the stimuli could have influenced listeners' labeling criteria, albeit in a complex way. Conceivably, the presence of exaggerated aspiration for target /p t/ could have led listeners to accept productions of /b d/ with longer-than-expected VOTs as members of the voiced category—for the alveolars, even for tokens with long-lag VOTs. A possible explanation for why this did not extend to /b/ could lie in the specifics of the stimulus distributions. Figure 2 shows that long-lag alveolars were fully bimodal with nearly a 30 ms gap between the /d/ and /t/ long lag productions. For the bilabials, however, there was some overlap in the VOT region of 130–160 ms. Given that the stimuli were blocked by POA, listeners might have developed somewhat different criteria for the subcomponents of the listening task. In the ambiguous range, /p t/ also showed better than chance accuracy (if lower than what was observed for /b d/). This could reflect a boundary shift based on the presence of prevoicing. These possibilities will be revisited in Sec. V.

If, indeed, listeners' labeling behavior was affected by the distributional characteristics of the stimuli, it implies that they were sensitive to subphonemic differences in VOTs. Experiment III controlled for within-category distributions of phoneme targets to remove potential confounds on listener behavior arising from statistical characteristics of the stimuli.

## IV. EXPERIMENT III

### A. Purpose

Experiment III was undertaken to explore listeners' labeling when voiced and voiceless targets were more

balanced along the VOT continuum. Since the distributional imbalances of experiment II were mainly at the VOT extremes, experiment III excluded prevoicing and exaggerated aspiration and considered only VOT values of 0–100 ms. Stimuli were selected within three VOT ranges: Appropriate for /b d/ (short lag), appropriate for /p t/ (long lag), and ambiguous; see Table I and Sec. IV B 2 for further details. We also obtained acoustic measures to explore the possible role of secondary cues. Following the work reviewed in Sec. I B 3, we chose the following for analysis: f0, F1, BI, and VD.

## B. Methods

### 1. Listeners

Participants were 20 monolingual English speakers, ages 23–34 years (mean = 25 years; SD = 3 years). As in experiments I and II, all were graduate students in a speech-language pathology program; had pure-tone thresholds of 20 dB at 500, 1000, 2000, and 4000 Hz based on bilateral screening; and were blind to the purpose and hypothesis of the experiment. None of the listeners in experiment III participated in either of the first two experiments.

### 2. Stimuli, presentation, and scoring for listening task

Experiment III stimuli were selected from the same six children as in experiment II. The three VOT ranges were as follows: Appropriate for /b d/ (0–20 ms), appropriate for /p t/ (30–100 ms), and ambiguous (20–30 ms). The exclusion of prevoicing and very long aspiration (>100 ms) meant that a large number of productions from experiment II in appropriate VOT ranges were replaced for this stimulus set: 100% of /p t/ productions in experiment II had exaggerated aspiration and were replaced with values between 30 and 100 ms. For /b d/, 79% had prevoicing (14/24 /d/, 24/24 /b/) and were replaced with short-lag VOT tokens. Replacements in other VOT ranges were less common and arose primarily in an attempt to refine the VOT category ranges and balance tokens across the 100 ms continuum. Last, two child participants lacked a single short-lag token to meet the inclusion criteria of 12 productions per child per target phoneme across the three VOT ranges as defined for this experiment. Rather than eliminate those participants, we opted to include one additional token from another child in the relevant bin to maintain the voiced-voiceless balance. With these adjustments, we obtained a stimulus set that had a largely equivalent representation of target /b d/ and /p t/ in each 10 ms bin over the continuum (see Fig. 4) yet still included six of the original ten speakers. Table VIII (second column) summarizes these changes between experiments II and III.

As in experiment II, the selection process resulted in 144 productions per POA, creating a total of 288 productions for perceptual testing (6 children × 3 VOT ranges × 4 words × 4 productions). The listening setup was consistent with experiment I and experiment II: Tokens were presented in fully randomized fashion, with two presentations per token, yielding 5760 ratings per POA category (144 tokens × 2 presentations × 20 listeners; 11 520 total ratings). The time to complete each block (bilabial, alveolar) was about 20 min. As before, accuracy scores signify a match between adult labels and the child's intended production, regardless of whether VOT was appropriate for the target.

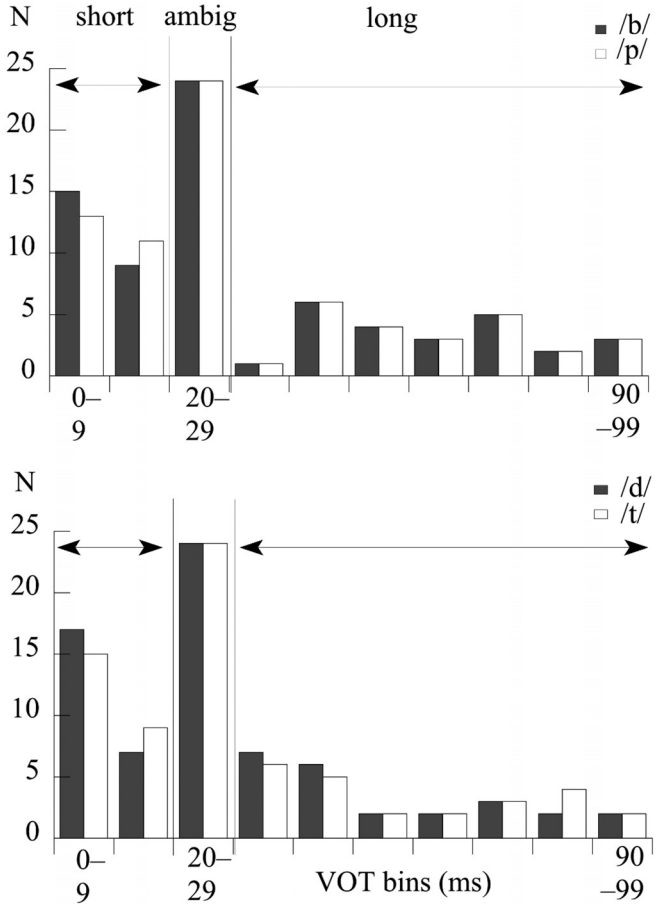


FIG. 4. Distributions of /b p/ and /d t/ productions across the VOT range used in experiment III.

### 3. Acoustic analysis

For acoustic analysis, speech samples were digitized at 11 025 Hz with a cutoff frequency of 5500 Hz.[8] Measures were interactively made by the first author in Praat (Boersma and Weenink, 2010) using a waveform and narrowband spectrogram (window length = 0.05 s). For each production, measures were made of f0 at voicing onset, F1 at voicing onset, BI, and VD. Onsets and offsets of glottal vibration were tagged using the acoustic waveform. The f0 tracking used the autocorrelation method with pitch range settings adjusted as needed to ensure that the pitch trace aligned with the harmonics as seen in the spectrogram. Formant tracking used a maximum number of five formants and maximum formant frequency of 5500 Hz. For BI, the time of the stop release was identified by tagging the onset of acoustic energy following the silent stop gap. Burst intensity was taken as an average over 2 ms from the release. VD was taken as the duration between voicing onset and offset.

https://doi.org/10.1121/10.0005595

Interrater reliability was assessed by asking acoustically trained graduate students to remeasure all four acoustic variables for a minimum of 15 productions per POA randomly selected across all recording sessions for each child. On several occasions, the student(s) did more than the recommended 15/48 tokens per variable (ca. 30%). As a result, 281/384 (73%) of the data were remeasured. The final number of tokens remeasured per variable and the corresponding reliability data are shown in Table III. All measures for original and remeasured tokens were highly correlated, and mean differences between raters were small.

#### 4. Discriminant analyses

To explore possible effects of secondary cues on listener judgments, the four acoustic measures, f0, F1, BI, and VD, were entered into discriminant analyses to assess how well they predicted the listeners' categorical labeling judgments. VOT was not included as a predictor variable because our analysis was divided by VOT range, i.e., it is inherent to the voiced/voiceless categorization. As will be seen in Sec. IV C 4, most listener accuracy ratings were quite accurate or quite inaccurate, with few productions of intermediate accuracy. Thus, for the discriminant analyses the data were divided into sets of *high accuracy* and *low accuracy*. Stimuli were assigned a rating of high accuracy if >66% of listener ratings aligned with the child's intended production; a low accuracy rating was assigned if <33% of listener ratings matched the target.

### C. Results and discussion

Across all data, individual listener accuracy ratings ranged from 61% to 66%. Consistent with experiments I and II, listeners were also quite reliable in assigning the same label to repeated presentations (92.45%, or 5325/5760 pairs). Average listener accuracies for experiment III are shown in Fig. 5. It is convenient again to discuss these in sets of appropriate, inappropriate, and ambiguous VOTs.

#### 1. Labeling for appropriate VOTs

Similar to experiments I and II, labeling accuracy was quite high when VOTs were appropriate for the target: All percentages were 93% or higher, and all $\chi^2$ analyses yielded $p < 0.001$.

TABLE III. Interrater reliability for acoustic measures: f0, F1, BI, and VD per POA category.

| POA category | Measure | *r* | Mean difference | SD of difference | *N* |
|---|---|---|---|---|---|
| bp | F0 (Hz) | 1.000[a] | 0.13 Hz | 1.22 Hz | 42 |
| dt | F0 (Hz) | 0.996[a] | 0.55 Hz | 5.58 Hz | 42 |
| bp | F1 (Hz) | 1.000[a] | 1.80 Hz | 0.80 Hz | 42 |
| dt | F1 (Hz) | 0.930[a] | 14.14 Hz | 40.56 Hz | 42 |
| bp | BI (dB) | 0.937[a] | 1.01 dB | 2.82 dB | 32 |
| dt | BI (dB) | 0.896[a] | 2.10 dB | 3.35 dB | 25 |
| bp | VD (s) | 0.989[a] | 0.00 s | 0.03 s | 28 |
| dt | VD (s) | 0.786[a] | 0.04 s | 0.10 s | 28 |

[a]Correlation is significant at the 0.01 level (two-tailed).

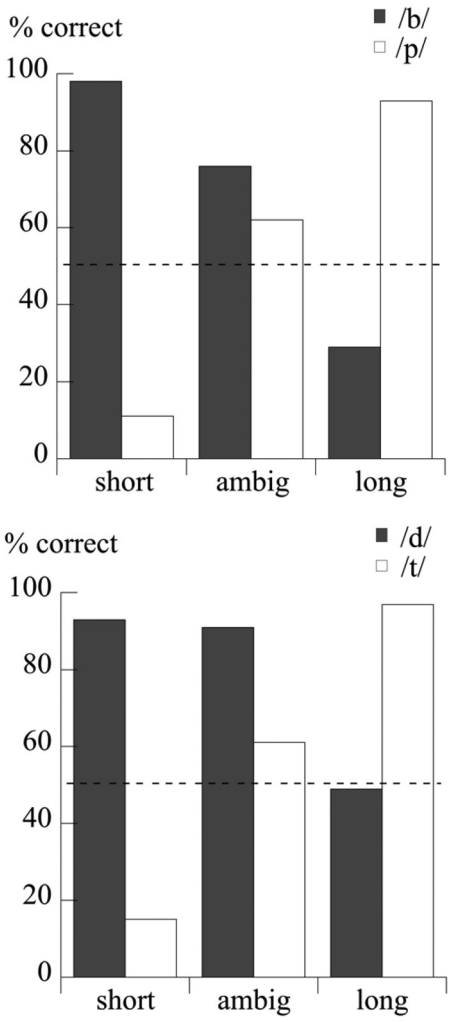


FIG. 5. Accuracy data for bilabials (left) and alveolars (right) for experiment III. Response data were combined for the 20 listeners. The *x* axes show the VOT category of the child's production. The short-lag categories had VOT values of 0–20 ms. The long lag categories had VOT values of 30–100 ms. The dotted line at 50% indicates chance responding.

#### 2. Labeling for inappropriate VOTs

As in experiment II, listeners had low accuracy rates for three of the four cases: long-lag /b/ = 29%; short-lag /p/ = 11%; short-lag /t/ = 15%; $p < 0.001$ for all three $\chi^2$ analyses. Unlike experiment II, for /d/ targets with long VOTs, listener accuracy was 49%, statistically at chance ($\chi^2 = 0.204\,17$, $p = 0.6514$). Listener judgments therefore appeared to be driven mainly by VOT.

#### 3. Labeling for ambiguous VOTs

As in experiment II, labeling for productions with ambiguous VOTs was at better than chance accuracy, particularly for target /b d/: 76% for /b/, 91% for /d/, 62% for /p/, and 61% for /t/. All $\chi^2$ analyses again yielded $p < 0.001$. (Recall that this category was minimally altered from

experiment II). Targets /b d/ were again labeled correctly more often than /p t/ in the ambiguous range. Average listener accuracies in experiments II and III will be compared in greater detail in Sec. V.

### 4. Secondary cues

A series of discriminant analyses was run to determine how well listener accuracy could be predicted based on the acoustic variables of f0, F1, and BI (potential secondary cues) and VD (a proxy for speech rate). Since the data were divided according to the three VOT regions (see next paragraph), the actual VOT value in ms was not entered into the analysis because it could not be treated as an independent contributor.

Of the full set of 144 productions per POA, only 15 (<7% of /b p/, <4% of /d t/) were labeled with "intermediate" accuracy, i.e., 33%–66%. Seven of these were bilabials in the ambiguous range, and the rest were scattered among VOT categories. Since it was not possible to carry out meaningful analyses for these "confusing" tokens, they were excluded from the subsequent analyses, and the remaining data were bifurcated into low (<33%) and high (>66%) accuracy tokens. Analyses were run separately in the three VOT ranges (short, ambiguous, and long), since we expected that the utility of secondary cues might vary depending on the phonetic nature of the production (e.g., a cue relevant for long-lag VOTs might not be relevant for short-lag VOTs). We further expected that in the ambiguous range, it may be necessary to access secondary cues to accomplish categorization. Additionally, studies of secondary cues have typically reported rather small shifts in crossover boundaries (e.g., Lisker, 1975; Lisker *et al.*, 1977; Miller and Volaitis, 1989; Whalen *et al.*, 1993) rather than large-scale changes extending to the extremes of the continua.

Before performing the discriminant analyses, the data were reviewed for skew. Productions[9] with f0 values >600 Hz (5.5% of bilabial tokens; 2% of alveolar tokens) and/or VDs >800 ms (7% and 1% for bilabials and alveolars, respectively) were excluded from all subsequent analyses. We also assessed correlations among variables to ensure that the data met the assumptions for discriminant analyses. These are provided in Table IV. By and large, the correlations were weak ($|r| < 0.2$) or moderate ($|r| = 0.2$–$0.4$). Only 4 of the 36, or 11%, were strong ($|r|$ values 0.401–0.573). On the whole, therefore, the data did not show extensive collinearity, suggesting that each acoustic measure was contributing unique information to the analyses. Descriptive statistics for the four acoustic variables are given in Table V. Despite high within-category variability, these measures differed noticeably across VOT categories. Data for individual child speakers are provided in the supplemental materials.[8]

Standardized canonical discriminant function coefficients varied across POA and VOT categories (see Table VI). Correct classification, based on the existing group size per variable, reached approximately 80% for ambiguous VOTs at both POAs and long lag alveolars. No relationships reached significance using the standard threshold of $p < 0.05$, presumably as a result of high within- and cross-speaker variability. The ANOVA did show $F$ values of 2.85 or greater and $p$ values of 0.10 or less for the following: Burst intensities for short-lag and ambiguous bilabials, f0 for long-lag bilabials, and F1 for long-lag alveolars. For long-lag bilabials, higher accuracy was associated with higher f0s. Lower F1s corresponded to lower listener accuracy for the long-lag alveolars. Unexpectedly, lower burst intensities were associated with lower accuracy for short-lag and ambiguous bilabials.

In sum, the discriminant analyses support Forrest and Rockman's (1988) claims that secondary cues may influence adult listener perception of child speech, albeit in ways that can vary with VOT values and POA. The standardized coefficients did not indicate that secondary cues had larger effects for ambiguous VOTs (see Table VI). It was the case, however, that the discriminant analyses classified ambiguous VOTs rather well (>80%). One issue that must be

TABLE IV. Correlations among the acoustic variables entered into the discriminant analyses: f0, F1, BI, and VD.

| /bp/ | | | | | /dt/ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Measure | f0 | F1 | BI | VD | Measure | f0 | F1 | BI | VD |
| Short f0 | — | | | | Short f0 | — | | | |
| Short F1 | 0.099 | — | | | Short F1 | 0.150 | — | | |
| Short BI | 0.046 | −0.165 | — | | Short BI | −0.231 | −0.229 | — | |
| Short VD | 0.189 | 0.268 | −0.573 | — | Short VD | 0.529 | −0.105 | −0.296 | — |
| Amb[a] f0 | — | | | | Amb f0 | — | | | |
| Amb F1 | −0.155 | — | | | Amb F1 | 0.235 | — | | |
| Amb BI | −0.028 | −0.013 | — | | Amb BI | −0.408 | −0.137 | — | |
| Amb VD | 0.105 | −0.370 | 0.134 | — | Amb VD | 0.103 | 0.139 | −0.352 | — |
| Long f0 | — | | | | Long f0 | — | | | |
| Long F1 | 0.105 | — | | | Long F1 | 0.401 | — | | |
| Long BI | −0.005 | −0.358 | — | | Long BI | 0.040 | −0.101 | — | |
| Long VD | 0.196 | −0.082 | −0.216 | — | Long VD | 0.265 | 0.063 | −0.311 | — |

[a]Ambiguous (Amb).

https://doi.org/10.1121/10.0005595

TABLE V. Descriptive statistics by POA and VOT category for each acoustic measure: f0, F1, BI, and VD per POA and VOT category.

| Descriptive Statistics | | | | | | |
|---|---|---|---|---|---|---|
| POA | Variable | VOT category | *N* | Range | Mean | SD |
| /bp/ | f0 (Hz) | Short | 41 | 303.33 | 360.90 | 67.61 |
| | | Amb | 36 | 353.84 | 357.03 | 77.71 |
| | | Long | 36 | 344.88 | 352.42 | 79.54 |
| /dt/ | f0 (Hz) | Short | 45 | 273.60 | 339.17 | 72.13 |
| | | Amb | 44 | 332.56 | 327.16 | 76.79 |
| | | Long | 38 | 305.75 | 336.05 | 80.24 |
| /bp/ | F1 (Hz) | Short | 41 | 473.83 | 578.76 | 103.16 |
| | | Amb | 36 | 380.34 | 595.05 | 92.48 |
| | | Long | 36 | 442.17 | 559.78 | 97.33 |
| /dt/ | F1 (Hz) | Short | 45 | 781.82 | 848.51 | 152.69 |
| | | Amb | 44 | 914.18 | 781.91 | 166.32 |
| | | Long | 38 | 846.48 | 816.89 | 195.93 |
| /bp/ | BI (dB) | Short | 41 | 31.96 | 67.63 | 6.86 |
| | | Amb | 36 | 22.79 | 58.88 | 5.51 |
| | | Long | 36 | 38.14 | 63.54 | 6.96 |
| /dt/ | BI (dB) | Short | 45 | 26.90 | 66.17 | 5.48 |
| | | Amb | 44 | 24.15 | 61.46 | 5.95 |
| | | Long | 38 | 28.53 | 64.93 | 6.85 |
| /bp/ | VD (s) | Short | 41 | 0.51 | 0.31 | 0.12 |
| | | Amb | 36 | 0.46 | 0.33 | 0.10 |
| | | Long | 36 | 0.36 | 0.34 | 0.08 |
| /dt/ | VD (s) | Short | 45 | 0.56 | 0.32 | 0.11 |
| | | Amb | 44 | 0.62 | 0.36 | 0.15 |
| | | Long | 38 | 0.55 | 0.41 | 0.12 |

considered is that some of these results may reflect token- or child-specific patterns. This will be addressed in Sec. V.

As an additional exploration of the role of secondary cues, we compared the results from experiments I and III. The stimulus sets for both of these spanned the same VOT range, 0–100 ms. In experiment I, all tokens at the short-lag end of the range, <20 ms, corresponded to target /b d/, and all tokens at the long-lag end of the range, >30 ms, corresponded to target /p t/. In experiment III, all 10 ms VOT bins contained both voiced and voiceless targets, with numbers balanced insofar as possible. We focus here on the alveolars, since for the bilabials, we only had one token each of /b p/ in the 30-ms VOT bin (cf. Fig. 4), produced by a single child. (For the record, the bilabials do show the same pattern as obtained for the alveolars.) We further focus on the region from 0 to 50 ms, which both includes the 50% crossover region and avoids the longer VOT bins that had limited data and some imbalances for the 80–90 ms bin. Figure 6 re-plots the labeling data from experiments I and III, overlaid, for that range. In experiment III, for /d/ targets, the VOT crossover shifted rightward. This shift suggests that secondary cues might have played a mitigating role, leading listeners to label stops as voiced even though the VOT exceeded 30 ms. Conversely, when the child's target was /t/, the crossovers shifted to the left, suggesting again that other cues can override a short VOT to some extent. As a comparison, Whalen *et al.* (1993) found that the effect of altering a single secondary cue (f0) shifted phonemic boundaries by approximately 10 ms. Our Fig. 6 shows that, when the stimuli were changed from single consistent child targets to a combination of voiced and voiceless targets, the phoneme boundary shifted about 10 ms to the left and about 15 ms to the right. The greater magnitude of the boundary shifts seen in our data can be explained by the fact that our stimuli had natural variation not only in f0, but also in BI and F1, i.e., there were potentially multiple secondary cues available.

TABLE VI. Standardized canonical discriminant function coefficients for f0, F1, BI, and VD per POA and VOT category.

| | /bp/ | | | /dt/ | | |
|---|---|---|---|---|---|---|
| | Short | Amb | Long | Short | Amb | Long |
| *N* | 41 | 36 | 36 | 45 | 44 | 38 |
| f0 | −0.339 | 0.320 | **0.893**[a] | 0.628 | −0.090 | 0.490 |
| F1 | −0.193 | −0.550 | 0.388 | 0.010 | **0.765** | **−0.908** |
| BI | **1.094** | **0.810** | 0.098 | 0.496 | 0.402 | 0.243 |
| VD | 0.351 | −0.258 | −0.237 | **−0.919** | −0.533 | 0.433 |
| % within category[b] | 59 | 83 | 69 | 58 | 80 | 79 |

[a]Bold indicates the standardized coefficient greatest in magnitude per POA and VOT category.
[b]Denotes the percentage of original grouped cases correctly classified.

## V. SUMMARY AND GENERAL DISCUSSION

This work explored adult perception of stop consonant voicing in productions from typically developing two-year-old speakers of American English. We sought to assess (1) whether adult listeners' phoneme boundaries for child speech were similar to those obtained previously for synthetic speech based on adult values; (2) how the distribution of /b p/ and /t d/ targets across the VOT continuum, with and without extreme VOT values, influenced listener

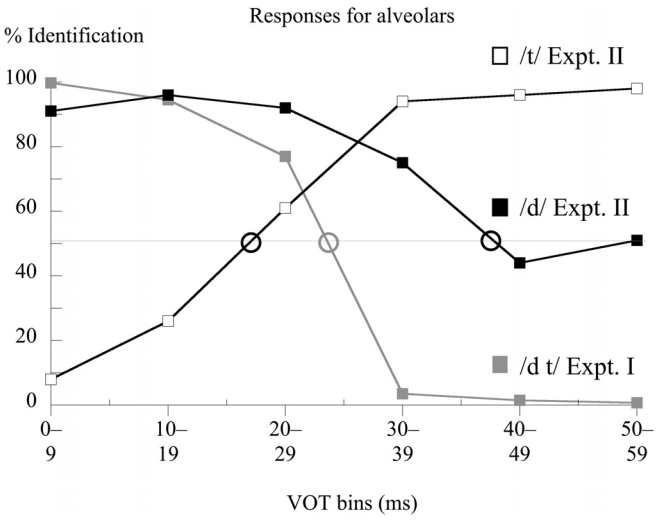


FIG. 6. Labeling functions for /d t/ in experiment I (gray) and experiment III (black, /d/; unfilled, /t/). Combined data for 20 listeners. Circles show the 50% crossovers. To improve readability, the *x* axis suppresses decimal places at the upper (right) edge of each bin, e.g., the bin 0–9.99 ms is represented as 0–9 ms in the axis label.

judgments; and (3) whether listener judgments of voicing might reflect the influence of identifiable secondary acoustic cues.

The /b p/ and /d t/ phoneme boundaries established in experiment I were similar to those found in past research, indicating that the criteria by which mature listeners judge voicing in the speech of young children are roughly comparable to those used for synthetic and adult speech. This finding provides some support for the traditional practice of using adult-based VOT values to infer how adults will label children's productions (Hitchcock and Koenig, 2013; Kewley-Port and Preston, 1974; Macken and Barton, 1980). Despite a lack of directly comparable adult data for the specific words used in this experiment, it is likely that our child speakers showed longer speech durations, overall, than what would be observed in adult speech (cf. Lee *et al.*, 1999). Whereas some authors have found that VOT category boundaries may shift as a function of speech rate (e.g., Miller and Volaitis, 1989), our listeners did not demonstrate an overall shift in their phonemic boundaries toward longer VOT values. It is possible, however, that slower speech rates in children make listeners more tolerant of exaggerated VOTs (Miller *et al.*, 1986; cf. also Munson *et al.*, 2010). At the same time, the accuracy data from experiments II and III suggest that adult voicing judgments may be affected by the nature of the stimulus set. Interpreting these results in the context of previous work requires consideration of both the stimuli and the listening task.

### A. On stimulus characteristics

The listener accuracy data for experiments II and III showed several statistically significant differences (see Table VII). Sampling error is not likely to account for all of these, particularly given that our listener characteristics were quite similar across the two studies, and in both cases, we collected a large number of listener responses. As summarized in Table VIII, some accuracy differences between the experiments (viz., for long-lag /p t/) can be ascribed to characteristics of the specific tokens chosen. That is, in these two cases, listeners heard entirely different productions in experiments II and III. In other cases, however, the differences may reflect statistical properties of the listening test as a whole. In particular, the presence of prevoicing, exaggerated aspiration, or both in experiment II could have had the effect of shifting listener boundaries for particular categories, either because of the presence of extreme values or the resulting largely bimodal distributions for /b/ vs /p/ and /d/ vs /t/ along the VOT continuum.

TABLE VII. Results ($p$ values) of the test of two proportions per POA and VOT category, comparing accuracy results across experiments II and III. Listener accuracy proportions are included for those comparisons that were statistically significant.

| | | Test of two proportions | | Listener accuracy | |
|---|---|---|---|---|---|
| POA | VOT category | Value | Asymptotic significance (two-sided) | Experiment II | Experiment III |
| /b/ | Short | 6.076 | **0.014**[a] | 0.994 | 0.980 |
| | Amb | 0.719 | 0.396 | | |
| | Long | 0.162 | 0.687 | | |
| /p/ | Short | 0.138 | 0.711 | | |
| | Amb | 11.702 | **0.001** | 0.541 | 0.620 |
| | Long | 60.728 | **0.000** | 0.997 | 0.930 |
| /d/ | Short | 4.441 | **0.035** | 0.951 | 0.930 |
| | Amb | 0.104 | 0.747 | | |
| | Long | 10.840 | **0.001** | 0.568 | 0.490 |
| /t/ | Short | 21.291 | **0.000** | 0.229 | 0.150 |
| | Amb | 2.078 | 0.149 | | |
| | Long | 26.684 | **0.000** | 0.920 | 0.970 |

[a]Bold indicates significant $p$ values using a standard criterion of 0.05.

In stressed syllable-initial position, American English has traditionally been said to contrast short-lag and long-lag VOT, with the presence of prevoicing being optional and, possibly, speaker-specific (e.g., Lisker and Abramson, 1964). Abramson and Lisker (1967) did not observe a discrimination peak between prevoiced and voiceless unaspirated tokens for English, and indeed some perception studies using American-English listeners have only employed positive values of VOT (e.g., Brady and Darwin, 1978; Lisker *et al.*, 1977; Miller and Volaitis, 1989). This could be taken to imply that prevoicing is irrelevant in English, and listeners may simply ignore it. On the other hand, Pisoni *et al.* (1982) observed sensitivity to prevoicing in certain listening tasks. Some changes in listener accuracy from experiments II to III are consistent with this possibility (Table VIII).

Exaggerated aspiration may also have affected listener judgments in experiment II (Table VIII). Although variations in the degree of aspiration are not known to function phonemically in the world's languages (Cho and Ladefoged, 1999; Lisker and Abramson, 1964), there is at least one precedent in the literature suggesting listener sensitivity to exaggerated aspiration. Miller and Volaitis (1989) allowed listeners to choose */p/ for synthetic productions with very long VOTs. For shorter (but not longer) syllable durations, the /p–*p/ labeling function was similar in steepness to that for /b–p/, suggesting that listeners imposed a category boundary between regular and exaggerated aspiration.

More generally, previous reports indicate that the nature of the dataset may shift listener judgments of voicing. Diehl *et al.* (1978) reported that identifications of ambiguous productions could be shifted toward judgments of /p/ when preceded by a prevoiced token and conversely toward /b/ when preceded by a VOT of 100 ms, i.e., rather long aspiration. Since our data were fully randomized within blocks, simple ordering effects can be ruled out, but listeners still could have been sensitive to statistical properties of the stimulus sets. Such statistical learning has been discussed extensively in regard to infant speech perception [see review in Romberg and Saffran (2010)] but has also been documented in adult listeners to some degree; for studies of voicing perception, see Baese-Berk (2019), Brady and Darwin (1978), Maye and Gerken (1999), and Pegg and Werker (1997).

https://doi.org/10.1121/10.0005595

TABLE VIII. Possible explanations for significant accuracy differences between experiments II and III. For experiment II, given the presence of exaggerated aspiration, only for /p t/, and prevoicing, mainly for /b d/, we cannot determine to what extent the results reflect the actual VOT values vs the fact that the two cognates at each POA were bimodally distributed.

| Target and VOT category | Token replacements, experiment II to experiment III | Listener accuracy change, experiment II to experiment III | Possible source | Comments |
|---|---|---|---|---|
| /p/ Long | 24/24 | ↓ 6.7% | Experiments had different tokens (removing very long aspiration for experiment III) | |
| /t/ Long | 24/24 | ↑ 5.0% | | |
| /t/ Short | 4/24 | ↓ 7.9% | Prevoicing in experiment II | Shorter-than-expected VOTs for /t/ more acceptable in the presence of prevoicing for /d/ |
| /p/ Amb | 2/24 | ↑ 7.9% | Long aspiration in experiment II | Ambiguous VOTs for /p/ less acceptable in the presence of exaggerated aspiration of /p/ |
| /d/ Long | 3/24 | ↓ 7.8% | Long aspiration in experiment II | Longer-than-expected VOTs for /d/ more acceptable in the presence of exaggerated aspiration of /t/ |
| /b/ Short | 24/24 | ↓ 1.4% | Prevoicing or extreme aspiration in experiment II | Listener accuracy higher when the phonemic difference is exaggerated |
| /d/ Short | 14/24 | ↓ 2.1% | | |

Brady and Darwin (1978) observed that ambiguous VOT values of 25 or 35 ms were more likely to be labeled as voiceless when those values fell at the long end of the 20-ms VOT range in that listening condition. In Maye and Gerken (1999), listeners were trained on end points of the continuum and achieved discrimination accuracy rates of about 60% for tokens with bimodally distributed VOTs in the narrow range of ca. 0–10 ms. Interestingly, even with the data largely balanced with 10-ms bins for experiment III, our listeners were nevertheless more accurate than chance for all four phonemes in the ambiguous range. A *post hoc* assessment of the data showed that, in fact, VOT values for /b d/ skewed short within that 10 ms range, and those for /p t/ skewed long. Traditionally, such small VOT differences were thought to be subperceptual (Hirsh, 1959), but in light of more recent work, it is clear that this assumption needs to be revisited. We plan a follow-up study to explore this.

It is likely that the degree to which adults can demonstrate statistical learning depends on both the presentation methods and the stimulus set among other factors. Of the explanations in Table VIII, the most compelling evidence for statistical learning comes from the cases where the child productions were largely overlapping for experiments II and III. This was true for target /p/ with ambiguous VOTs (22/24 identical) and target /d/ with long VOTs (21/24 identical). Here, changes in listener behavior can most clearly be ascribed to the nature of the dataset as a whole.

### B. On task characteristics

The data reported here come from labeling tasks. In everyday communicative situations, listeners are mainly engaged in trying to determine what the speaker said. Routine clinical practices, including assessment and measures of treatment efficacy, are also based on a (trained) listener making categorical judgments of phonemic accuracy (cf. Li, 2008). In contrast, previous studies assessing the perception of subphonemic differences and/or the extent to which speech perception is categorical have asked listeners to provide goodness judgments (Munson *et al.*, 2010) or manipulated the nature of discrimination tasks (Gerrits and Schouten, 2004; Pisoni *et al.*, 1982; cf. also Pisoni and Tash, 1974).

Although we chose a labeling task in an attempt to better capture real-world, clinically relevant conditions, it is still the case that our listeners took part in rather lengthy tasks where they heard numerous productions of two words per listening block. It is not entirely clear how these results would carry over to more naturalistic situations. As noted in the Introduction, Macken and Barton (1980) and Kewley-Port and Preston (1974) expected that adult listeners would not easily recognize a voicing distinction in children in stage II or with ambiguous VOTs, a supposition that was confirmed in early stages of the Hitchcock and Koenig (2013) project. Listening to spontaneous child productions in real time is quite different from a situation wherein participants are asked to focus on a specific phonetic distinction. The results of Maxwell and Weismer (1982) are instructive in this context: Their child speaker was initially identified as making extensive voicing errors, but the student clinicians were able to differentiate target productions of voiceless sounds in a controlled listening situation. The fact that cues to a distinction are available in the child's speech does not mean that listeners can necessarily use them in all situations, but extensive exposure to one child's speech may allow listeners to detect atypical distinctions (cf. Munson *et al.*, 2012). It is an empirical question to what extent our results from speech-language pathology students can be extended to listeners without phonetic training.

### C. Secondary cues and individual differences

In experiment I, VOT values covaried with the target (/b d/ or /p t/), making it impossible to know to what degree listeners may have used secondary cues along with VOT.

Experiments II and III remedied this (see Table I for details). In both of these experiments, listener accuracy was greater than chance for all appropriate and ambiguous VOT conditions. High listener accuracy for appropriate VOTs was expected; however, greater than chance accuracy for ambiguous VOTs was surprising and suggests that secondary cues were present in the stimuli. High classification accuracies in the discriminant analyses in the ambiguous range at both POAs further indicate that the children were successfully producing secondary cues to voicing. This finding, based on a large number of listener judgments, corroborates past reports (e.g., Forrest and Rockman, 1988), using more limited datasets, that appropriate secondary cues are available in young child speech. The fact that the coefficients in the discriminant analyses were not significantly higher in the ambiguous range could have resulted from the data trimming that we did to remove outliers of f0 and VD. That is, some extreme values of f0 or VD might have contributed to listener accuracy but were not reflected in the discriminant analysis results. The limited number of tokens in each VOT bin, combined with stimulus variability, could also have led to a type II error in this analysis.

In both experiments II and III, listener judgments in the ambiguous category were more accurate for /b d/ targets than /p t/ targets. This could relate to the fact that the short-lag category, being developmentally earlier, is more practiced. It may also be that the added articulatory action of vocal-fold abduction required for aspiration inherently makes /p t/ more difficult, leading to greater variability in both primary and secondary cues.

Speaker-specific variation is extensive in child speech, and such variability is evident in our measures of f0, F1, VD, and BI (see supplemental materials).[8] Our perceptual results, combining listener data across a small number of tokens from six children, cannot reveal whether adults made use of different secondary cues for different speakers, but a brief *post hoc* exploration suggests that this may have been so in some cases. For example, child C5 showed rather extreme variation in f0 across voicing targets, whereas C4 showed very little. This could potentially explain why listeners' judgments were more accurate for C5 than C4 in multiple categories (see supplemental Table I).[8] The developmental trajectory of secondary cues to voicing and other contrasts clearly requires further examination (see also Karlsson *et al.*, 2003).

### D. Adult perception and child speech: Other considerations

In an attempt to assess how adult listeners process child speech, we drew on a large dataset of elicited productions. We included tokens from multiple children in each experiment, to gain insight into how adults judge child speech generally, rather than obtaining results specific to a particular speaker [see Forrest and Rockman (1988) and Holliday *et al.* (2015) for similar considerations]. This approach has the advantage of increasing ecological validity, and it is comparable to what clinicians encounter in a typical work setting. It also has clear disadvantages. Despite having more than 6000 productions to choose from, it was not possible to exercise complete control over what was included in listening tasks, as can be done using synthetic speech. Our listeners could draw on secondary cues rather than being forced to rely on a single manipulated cue such as VOT. At the same time, our listeners had to contend with the variability of six or more speakers and the token-to-token variability (e.g., in duration, intonation, voice quality) that arises generally in natural speech and all the more so when one needs to engage a two-year-old long enough to get 15+ productions of the same four words in multiple sessions. To the extent that aspects of our results mirror those of more highly controlled perception studies, it suggests that some characteristics of speech perception are rather resilient in the face of wide-ranging stimulus variation.

Studies assessing adult perception of natural child speech are not highly common [one exception being Munson *et al.* (2010)]. Nevertheless, such research can contribute not only to a fuller picture of adult speech perception, but also to our understanding of speech development: Children's production development is likely driven, in part, by the success (or not) of their communicative attempts. Work of this nature is also relevant in clinical assessment and measures of treatment efficacy, where children are primarily judged on the basis of adult perception of their speech. Clinicians may use a phonetic inventory to assess the age appropriateness of a child's speech (Bleile, 2004), determine treatment targets (Gierut, 2005), or establish discharge criteria. Perceived errors in voicing or, alternatively, undetected covert voicing contrasts may contribute to misdiagnosing or overestimating or underestimating the severity of a child's speech sound disorder. Use of acoustic analysis in addition to transcription could mitigate these issues.

## VI. CONCLUSIONS

Through a series of experiments, we sought to better understand how adult listeners perceive stop consonant voicing in the speech of two-year-olds. The results do not indicate clearly different phoneme boundaries for children compared to adults. They do, however, suggest that secondary cues are available in young child speech and that adults can make use of them. High classification accuracy along with listener accuracy above chance for child productions with ambiguous VOTs provide support this conclusion. Listener behavior also seems to show effects of distributional properties of the dataset, even for tasks including multiple speakers and the extensive token-to-token variability characteristic of child speech.

## ACKNOWLEDGMENTS

We thank Karen Forrest, who, in an assessment of the Hitchcock and Koenig (2013) submission, raised questions regarding adult perception of children's voicing that ultimately led us to undertake this set of experiments. We also express our thanks to the child speakers and their

parents; to the listening participants; and to Kristen Shubeck, Colleen Russell, Francesca Ilardi, Jennifer Caceres, Carly Truglio, and Lea Marucci, graduate research assistants at Montclair State University.

## APPENDIX

See Table IX for selected 50% VOT bilabial and alveolar crossover values from the literature.

TABLE IX. Selected 50% VOT bilabial and alveolar crossover values from the literature, in (whole) ms. Values with "ca." were estimated from graphs in the original sources. Results for Kuhl and Miller (1978) are from human participants only (i.e., excluding the chinchillas).

| Study | Words/syllables | 50% crossover (ms) |
|---|---|---|
| Kuhl and Miller (1978) | /bɑ/ /pɑ/ | 27 |
| | /dɑ/ /tɑ/ | 35 |
| Lisker and Abramson (1967a) | /bɑ/ /pɑ/ | ca. 22 |
| | /dɑ/ /tɑ/ | ca. 35 |
| Pisoni and Lazarus (1974) | /bɑ/ /pɑ/ | 25–30 |
| Zlatin (1974) | bees, peas | 30 |
| | bear, pear | 34 |
| | dime, time | 27 |

[1]Additional evidence for these impressions is the following: An inclusion criterion for the toddlers in Hitchcock and Koenig (2013) was no perceptible voicing contrast as judged by the first author, a speech-language pathologist with considerable expertise in child speech. Subsequent analyses demonstrated that four of those children did produce significant VOT differences between voicing categories, i.e., they were in subperceptual stage II.

[2]It should be noted that the synthetic VOT continua from the Haskins group also incorporated a changing F1 onset frequency (F1 "cutback"). This is discussed more in Sec. I B 3.

[3]Forrest and Rockman (1988) also reported VOT values in connected speech; for simplicity, we only discuss the isolated word productions here. The connected speech results were slightly different, but not in a way that changes our general points.

[4]Hitchcock and Koenig (2013) originally used accuracy cutoffs of 20 ms for bilabials and 30 ms for alveolars, given findings of somewhat different crossover values across POAs (see Appendix). They subsequently did a follow-up analysis for the bilabials using the same 30 ms criterion for both places of articulation. For simplicity, here we used the single common threshold.

[5]We allowed listeners the option to replay productions because we thought the task might be rather difficult, and the decision has precedent in studies of "difficult" tasks (e.g., Ingvalson *et al.*, 2017; Xie and Fowler, 2013). This decision did have the drawback of not allowing us to assess reaction times in a simple fashion, and studies of speech perception in typical adults have not typically allowed this. It is possible that our results therefore reflect some improved performance over what would be obtained if listeners had to supply a response after a single production. An assessment of individual variation in the response times for the task does suggest that only a few listeners (1–3 in each experiment) made use of the replay option.

[6]A reviewer questioned our use of a line graph here, observing that our VOT continuum is not, strictly speaking, a continuous scale. The same is true of other traditional VOT studies employing synthetic continua. (Our bins did have some internal variation, i.e., a range of VOT values, but that range varied across bins.) We follow the graphical tradition established in early work (e.g., Liberman *et al.*, 1957) and adopted subsequently (e.g., Coady *et al.*, 2005) whereby the "crossover" function is presented as a line graph. Early work simply found the $x$ axis intercept of the 50% $y$ axis value as the phoneme boundary, i.e., the boundary was estimated from the graph. To improve precision, here we followed Coady *et al.* (2005) by using a probit analysis to fit the data and estimate the 50% crossovers for the two places of articulation.

[7]Because of an experimenter error, one participant did not pass the 20 dB threshold at 4000 Hz in one ear but still completed the experiment. The participant detected 4000 Hz at 25 dB in that ear. The data from that participant were not qualitatively different from the rest, so were retained here.

[8]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0005595 for data analyzing whether down-sampling might have affected our burst intensity measures, particularly in the case of smaller child vocal tracts. In response to a reviewer query, we explored this possibility in a set of data from adults and one child speaker from the original database. Briefly: Original and re-sampled data for children and adults were very highly correlated ($r > 0.999$), and analyses of variance (ANOVAs) with sampling rate as the independent variable yielded no significant differences ($p$ values $> 0.9$). Original and down-sampled data did show a significant difference across consonants, with the largest effects on [t]. Average differences between original and down-sampled data were quite small: <0.02 dB for adults and <0.2 dB for the child (lower values for the down-sampled data). Given the very small magnitude of the effect, we do not believe that this is of concern. Also, see the supplementary material for a summary of accuracy judgments split by speaker and VOT category, acoustic measures (f0, F1, BI, and VD) for all children, and figures showing distributions of the four acoustic measures for each child speaker, phoneme, and VOT category.

[9]These cutoff values were chosen to reduce skew in the data and also based on past work. For f0, Keating and Buhr (1978) found that 93% of productions fell in the range of 0–600 Hz for their two-year-olds. For duration, based on Smith (1978) and Kubaska and Keating (1981), 800 ms represents a reasonable high-end cutoff for monosyllables similar to ours.

Abramson, A. S., and Lisker, L. (**1967**). "Discriminability along the voicing continuum: Cross-language tests," in *Proceedings of the 6th International Congress of Phonetic Sciences*, Prague, Czech Republic, September 7–13, pp. 569–573.

Abramson, A. S., and Lisker, L. (**1985**). "Relative power of cues: F0 shift versus voice timing," in *Linguistic Phonetics: Essays in Honor of Peter Ladefoged*, edited by V. Fromkin (Academic, New York), pp. 25–33.

Arkebauer, H. J., Hixon, T. J., and Hardy, J. C. (**1967**). "Peak intraoral air pressures during speech," J. Speech Lang. Hear. Res. **10**(2), 196–208.

Baese-Berk, M. M. (**2019**). "Interactions between speech perception and production during learning of novel phonemic categories," Attn. Percept. Psychophys. **81**(4), 981–1005.

Bleile, K. (**2004**). *Manual of Articulation and Phonological Disorders: Infancy through Adulthood*, 2nd ed. (Delmar, Clifton Park, NY).

Boersma, P., and Weenink, D. (**2010**). "Praat (version 5.1.35) [computer program]," http://www.praat.org (Last viewed 06/09/2021).

Brady, S. A., and Darwin, C. J. (**1978**). "Range effect in the perception of voicing," J. Acoust. Soc. Am. **63**(5), 1556–1558.

Cho, T., and Ladefoged, P. (**1999**). "Variation and universals in VOT: Evidence from 18 languages," J. Phon. **27**(2), 207–229.

Coady, J. A., Kluender, K. R., and Evans, J. L. (**2005**). "Categorical perception of speech by children with specific language impairments," J. Speech Lang. Hear. Res. **48**(4), 944–959.

Cooper, F. S., Delattre, P. C., Liberman, A. M., Borst, J. M., and Gerstman, L. J. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**(6), 597–606.

Diehl, R. L., Elman, J. L., and McCusker, S. B. (**1978**). "Contrast effects on stop consonant identification," J. Exp. Psychol. Hum. Percept. Perform. **4**(4), 599–609.

Eckman, F., Iverson, G., and Song, J. (**2015**). "Overt and covert contrast in L2 phonology," J. Second Lang. Pronunc. **1**(2), 254–278.

Edwards, J., and Beckman, M. E. (**2008**). "Methodological questions in studying consonant acquisition," Clin. Linguist. Phon. **22**(12), 937–956.

Eguchi, S., and Hirsh, I. J. (**1969**). "Development of speech sounds in children," Acta Otolaryngol. Suppl. **257**, 1–51.

Forrest, K., and Rockman, B. K. (**1988**). "Acoustic and perceptual analysis of word-initial stop consonants in phonologically disordered children," J. Speech Hear. Res. **31**(3), 449–459.
Forrest, K., Weismer, G., Hodge, M., Dinnsen, D. A., and Elbert, M. (**1990**). "Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children," Clin. Linguist. Phon. **4**(4), 327–340.
Gerrits, E., and Schouten, M. E. H. (**2004**). "Categorical perception depends on the discrimination task," Percept. Psychophys. **66**(3), 363–376.
Gierut, J. A. (**2005**). "Phonological intervention: The how or the what?," in *Phonological Disorders in Children: Clinical Decision Making in Assessment and Intervention*, edited by A. G. Kamhi and K. E. Pollock (Brookes, Baltimore, MD), pp. 201–210.
Haggard, M., Ambler, S., and Callow, M. (**1970**). "Pitch as a voicing cue," J. Acoust. Soc. Am. **47**(2B), 613–617.
Hirsh, I. J. (**1959**). "Auditory perception of temporal order," J. Acoust. Soc. Am. **31**(6), 759–767.
Hitchcock, E. R., and Koenig, L. L. (**2013**). "The effects of data reduction in determining the schedule of voicing acquisition in young children," J. Speech Lang. Hear. Res. **56**(1), 441–457.
Holliday, J. J., Reidy, P. F., Beckman, M. E., and Edwards, J. (**2015**). "Quantifying the robustness of the English sibilant fricative contrast in children," J. Speech Lang. Hear. Res. **58**(3), 622–637.
Ingvalson, E. M., Lansford, K. L., Fedorova, V., and Fernandez, G. (**2017**). "Cognitive factors as predictors of accented speech perception for younger and older adults," J. Acoust. Soc. Am. **141**(6), 4652–4659.
Karlsson, F., Sullivan, K. P. H., van Doorn, J., and Czigler, P. E. (**2003**). "Då or Tå, På or Bå — Seeing is believing!," in *Proceedings of the 15th International Conference on Phonetic Sciences*, August 3–9, Barcelona, Spain, Vol. 2, pp. 1967–1970.
Keating, P., and Buhr, R. (**1978**). "Fundamental frequency in the speech of infants and children," J. Acoust. Soc. Am. **63**(2), 567–571.
Kessinger, R. H., and Blumstein, S. E. (**1997**). "Effects of speaking rate on voice-onset time in Thai, French, and English," J. Phon. **25**(2), 143–168.
Kessinger, R. H., and Blumstein, S. E. (**1998**). "Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies," J. Phon. **26**(2), 117–128.
Kewley-Port, D., and Preston, M. (**1974**). "Early apical stop production: A voice onset time analysis," J. Phon. **2**(3), 195–210.
Kirby, J. P., and Ladd, D. R. (**2016**). "Effects of obstruent voicing on vowel F0: Evidence from 'true voicing' languages," J. Acoust. Soc. Am. **140**(4), 2400–2411.
Kubaska, C. A., and Keating, P. A. (**1981**). "Word duration in early child speech," J. Speech Hear. Res. **24**(4), 615–621.
Kuhl, P. K., and Miller, J. D. (**1978**). "Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli," J. Acoust. Soc. Am. **63**(3), 905–917.
Lee, S., Potamianos, A., and Narayanan, S. (**1999**). "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," J. Acoust. Soc. Am. **105**, 1455–1468.
Lehiste, I. (**1972**). "The timing of utterances and linguistic boundaries," J. Acoust. Soc. Am. **51**(6B), 2018–2024.
Li, F. (**2008**). "The phonetic development of voiceless sibilant fricatives in English, Japanese and Mandarin Chinese," Ph.D. dissertation, Ohio State University, Columbus, OH.
Li, F., Edwards, J., and Beckman, M. E. (**2009**). "Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers," J. Phon. **37**(1), 111–124.
Liberman, A. M., Cooper, F. S., Shankweiler, D. P., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psych. Rev. **74**(6), 431–461.
Liberman, A. M., Harris, K. S., Eimas, P. D., Lisker, L., and Bastian, J. (**1961**). "An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance," Lang. Speech **4**(4), 175–195.
Liberman, A. M., Harris, K. S., Hoffman, H. S., and Griffith, B. C. (**1957**). "The discrimination of speech sounds within and across phoneme boundaries," J. Exp. Psychol. **54**(5), 358–368.
Lisker, L. (**1975**). "Is it VOT or a first-formant transition detector?," J. Acoust. Soc. Am. **57**(6), 1547–1551.
Lisker, L., and Abramson, A. S. (**1964**). "A cross-language study of voicing in initial stops: Acoustical measurements," Word **20**(3), 384–422.
Lisker, L., and Abramson, A. S. (**1967a**). "Some effects of context on voice onset time in English stops," Lang. Speech **10**(1), 1–28.
Lisker, L., and Abramson, A. S. (**1967b**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th International Congress of Phonetic Sciences*, Prague, Czech Republic, September 7–13, pp. 563–567.
Lisker, L., and Abramson, A. S. (**1970**). "The voicing dimension: Some experiments in comparative phonetics," in *Proceedings of the 6th International Congress of Phonetic Sciences, Prague 1967*, edited by B. Hála, M. Romportl, and P. Janota (Academia Publishing House, Czechoslovak Academy of Sciences, Prague), pp. 563–567.
Lisker, L., Liberman, A. M., Erickson, D. M., Dechovitz, D., and Mandler, R. (**1977**). "On pushing the voice-onset-time (VOT) boundary about," Lang. Speech **20**(3), 209–216.
Löfqvist, A., Baer, T., McGarr, N. S., and Story, R. S. (**1989**). "The cricothyroid muscle in voicing control," J. Acoust. Soc. Am. **85**, 1314–1321.
Macken, M. A., and Barton, D. (**1980**). "The acquisition of the voicing contrast in English: A study of voice onset time in word initial stop consonants," J. Child Lang. **7**, 41–74.
Maxwell, E. M., and Weismer, G. (**1982**). "The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops," Appl. Psycholing. **3**(1), 29–43.
Maye, J., and Gerken, L. (**1999**). "Learning phonemes without minimal pairs," in *Proceedings of the 24th Annual Boston University Conference on Language Development*, November, Boston, MA, Vol. 2, pp. 522–533.
Miller, J. L., Green, K. P., and 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), 106–115.
Miller, J. L., O'Rourke, T. B., and Volaitis, L. E. (**1997**). "The internal structure of phonetic categories: Effects of speaking rate," Phonetica **54**(3), 121–137.
Miller, J. L., and Volaitis, L. E. (**1989**). "Effect of speaking rate on the perceptual structure of a phonetic category," Percept. Psychophys. **46**(6), 505–512.
Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K. (**2010**). "Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana," Clin. Linguist. Phon. **24**(4), 245–260.
Munson, B., Johnson, J. M., and Edwards, J. (**2012**). "The role of clinical experience in speech-language pathologists' perception of subphonemic detail in children's speech," Am. J. Speech Lang. Pathol. **21**(2), 124–139.
Nakai, S., and Scobbie, J. M. (**2016**). "The VOT category boundary in word-initial stops: Counter-evidence against rate normalization in English spontaneous speech," Lab. Phon. **7**(1), 13.
Nittrouer, S., and Studdert-Kennedy, M. (**1987**). "The role of coarticulatory effects in the perception of fricatives by children and adults," J. Speech Hear. Res. **30**, 319–329.
Pegg, J. E., and Werker, J. F. (**1997**). "Adult and infant perception of two English phones," J. Acoust. Soc. Am. **102**(6), 3742–3753.
Pisoni, D. B., Aslin, R. N., Perey, A. J., and Hennessy, B. L. (**1982**). "Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants," J. Exp. Psychol. Hum. Percept. Perform. **8**(2), 297–314.
Pisoni, D. B., and Lazarus, J. H. (**1974**). "Categorical and noncategorical modes of speech perception along the voicing continuum," J. Acoust. Soc. Am. **55**(2), 328–333.
Pisoni, D. B., and Tash, J. (**1974**). "Reaction times to comparisons within and across phonetic categories," Percept. Psychophys. **15**(2), 285–290.
Preston, M. S., Yeni-Komshian, G., Stark, R. E., and Port, D. K. (**1968**). "Developmental studies of voicing in stops," Haskins Lab. Status Rep. **13/14**, 181–184.
Romberg, A. R., and Saffran, J. R. (**2010**). "Statistical learning and language acquisition," Wiley Interdisciplinary Rev. Cogn. Sci. **1**(6), 906–914.
Sawashima, M., Hirose, H., Hibi, S., Yoshioka, H., Kawase, N., and Yamada, M. (**1981**). "Measurements of the vocal fold length by use of stereoendoscope—A preliminary study," Annu. Bull. Res. Inst. Logoped. Phoniatr. **15**, 9–16.

Schellinger, S. K., Munson, B., and Edwards, J. (**2017**). "Gradient perception of children's productions of /s/ and /θ/: A comparative study of rating methods," Clin. Linguist. Phon. **31**(1), 80–103.
Smith, B. L. (**1978**). "Temporal aspects of English speech production: A developmental perspective," J. Phon. **6**(1), 37–67.
Smith, B. L., Kenney, M. K., and Hussain, S. (**1996**). "A longitudinal investigation of duration and temporal variability in children's speech production," J. Acoust. Soc. Am. **99**(4), 2344–2349.
Stevens, K. N., and Klatt, D. H. (**1974**). "The role of formant transitions in the voiced-voiceless distinction for stops," J. Acoust. Soc. Am. **55**(3), 653–659.
Soli, S. D. (**1983**). "The role of spectral cues in discrimination of voice onset time differences," J. Acoust. Soc. Am. **73**(6), 2150–2165.
Sundara, M. (**2005**). "Acoustic-phonetics of coronal stops: A cross-language study of Canadian English and Canadian French," J. Acoust. Soc. Am. **118**(2), 1026–1037.
Tyler, A. A., Edwards, M. L., and Saxman, J. H. (**1990**). "Acoustic validation of phonological knowledge and its relationship to treatment," J. Speech Lang. Hear. Res. **55**(2), 251–261.
Wayland, S. C., Miller, J. L., and Volaitis, L. E. (**1994**). "The influence of sentential speaking rate on the internal structure of phonetic categories," J. Acoust. Soc. Am. **95**(5), 2694–2701.
Whalen, D. H., Abramson, A. S., Lisker, L., and Mody, M. (**1993**). "*F*0 gives voicing information even with unambiguous voice onset times," J. Acoust. Soc. Am. **93**(4), 2152–2159.
Xie, X., and Fowler, C. A. (**2013**). "Listening with a foreign-accent: The interlanguage speech intelligibility benefit in Mandarin speakers of English," J. Phon. **41**(5), 369–378.
Zlatin, M. A. (**1974**). "Voicing contrast: Perceptual and productive voice onset time characteristics of adults," J. Acoust. Soc. Am. **56**(3), 981–994.
Zlatin, M. A., and Koenigsknecht, R. A. (**1975**). "Development of the voicing contrast: Perception of stop consonants," J. Speech Hear. Res. **18**, 541–553.