JSLHR

Research Article

Quantifying the Robustness of the English Sibilant Fricative Contrast in Children

Jeffrey J. Holliday,[a,b] Patrick F. Reidy,[a] Mary E. Beckman,[a] and Jan Edwards[c]

Purpose: Four measures of children's developing robustness of phonological contrast were compared to see how they correlated with age, vocabulary size, and adult listeners' correctness ratings.
Method: Word-initial sibilant fricative productions from eighty-one 2- to 5-year-old children and 20 adults were phonetically transcribed and acoustically analyzed. Four measures of robustness of contrast were calculated for each speaker on the basis of the centroid frequency measured from each fricative token. Productions that were transcribed as correct from different children were then used as stimuli in a perception experiment in which adult listeners rated the goodness of each production.
Results: Results showed that the degree of category overlap, quantified as the percentage of a child's productions whose category could be correctly predicted from the output of a mixed-effects logistic regression model, was the measure that correlated best with listeners' goodness judgments.
Conclusions: Even when children's productions have been transcribed as correct, adult listeners are sensitive to within-category variation quantified by the child's degree of category overlap. Further research is needed to explore the relationship between the age of a child and adults' sensitivity to different types of within-category variation in children's speech.

Consonant acquisition in children can be characterized by a high degree of variability both across sounds (i.e., some consonants or features tend to be produced in an adultlike way much earlier than others) and across children (i.e., some children produce consonants in an adultlike way at a much younger age than other children). In a large majority of the studies supporting this characterization, the determination of whether or not a particular consonant or feature has been acquired is made using phonetic transcription. For example, once a certain percentage of a child's productions of a particular consonant are transcribed as correct, then the child may be said to have acquired that consonant (e.g., Prather, Hedrick, & Kern, 1975; Sander, 1972; Smit, Hand, Freilinger, Bernthal, & Bird, 1990). In the case of a feature that makes a phonological contrast between two consonants, then, the child may be said to have acquired the feature once a certain percentage of productions of each member of the contrast are transcribed as correct. By definition, phonetic transcription involves a subjective judgment of category membership. The judgment can be either at the level of a coarse-grained broad transcription that uses only one symbol for each of the consonant phonemes of the specific language that the child is learning (e.g., using /t/ for all productions of the voiceless coronal stop of English that are deemed to be correct) or at the level of a more or less fine-grained narrow transcription that symbolically represents subphonemic variation (e.g., using [t], [t̪], [ʔ], and [ɾ] to differentiate among alveolar stop, dental stop, glottal stop, and flap productions of "correct" /t/, respectively). However, even the finer-grained transcription categories are not inherently positioned on an ordinal scale, and analyses of transcriptions to determine acquisition norms typically have involved the collapsing of categories to make a binary differentiation between correct (or at least acceptable) productions and incorrect (or unacceptable) productions (see, e.g., Smit et al., 1990). More recently, researchers have used phonetic transcriptions in developing instruments, such as severity metrics, that differentiate among different types of habitual errors in children who are below age norms for consonant acquisition (see, e.g., Preston, Ramsdell, Oller, Edwards, & Tobin, 2011). However, these transcription-based metrics are still quite coarse grained relative to measures that have been used in a subset of studies that suggest differences among children whose productions have been transcribed as either correct or incorrect. These studies are of two types, both showing ways in which

[a]The Ohio State University, Columbus
[b]Indiana University, Bloomington
[c]University of Wisconsin–Madison
Correspondence to Jeffrey J. Holliday: jeffh@ling.ohio-state.edu

Editor: Jody Kreiman
Associate Editor: Karen Forrest
Received March 27, 2014
Revision received October 25, 2014
Accepted February 10, 2015
DOI: 10.1044/2015_JSLHR-S-14-0090

Disclosure: *The authors have declared that no competing interests existed at the time of publication.*

children with either typical or atypical phonological development might produce a difference between sounds that makes for a less robust contrast than what is observed in adult productions.

First, there are many studies that have shown evidence for what is known as *covert contrast*, which is when a child produces a set of contrasting segments in some way that distinguishes among them but does not lead to each of the target segments being reliably identified by adults. The child may even be able to accurately perceive the target segments in the speech of adults (Kornfeld & Goehl, 1974; Rvachew & Jamieson, 1989) despite his or her own productions being perceived by adults as incorrect. Studies have found evidence in English-acquiring children's stop productions for covert contrast in the word-initial and word-final voicing contrast (Macken & Barton, 1980; Maxwell & Weismer, 1982; Scobbie, Gibbon, Hardcastle, & Fletcher, 2000) and for the word-initial lingual place contrast (Edwards, Gibbon, & Fourakis, 1997; Forrest, Weismer, Hodge, Dinnsen, & Elbert, 1990; Gibbon, 1990; White, 2001). Covert contrast has also been documented in children's productions of the English /s/-/θ/ contrast (e.g., Baum & McNutt, 1990) and the sibilant fricative place contrasts in English and Japanese (e.g., Li, Edwards, & Beckman, 2009). In these cases, phonetic transcription was shown to be inadequate because it glosses over different types and degrees of incorrect production. More recently, covert contrast in children's productions has also been identified by asking adults to evaluate children's productions using a rating scale. For example, Munson, Edwards, Schellinger, Beckman, and Meyer (2010) found that adults rated productions that had been transcribed either as correct or as clear substitutions in ways that suggested subphonemic distinctions. They rated productions of target /s/ that were transcribed as categorical substitutions of [θ] for /s/ to be less [θ]-like than productions of target /θ/ that were transcribed as correct. Covert contrast has been shown to be clinically important in that children with phonological disorder who show evidence of a covert contrast make faster progress in therapy than children who produce no contrast at all (Tyler, Figurski, & Langsdale, 1993).

Second, other studies have suggested that phonetic transcription also risks glossing over variability within productions transcribed as correct. Even when children's productions acoustically or articulatorily deviate from adult targets, they may still be perceived as correct (Gibbon, Dent, & Hardcastle, 1993; Kewley-Port & Preston, 1974), reflecting the fact that a child's phonological development does not end once he or she produces the requisite number of correctly transcribed tokens of all segments in the language. In a large-scale study of the speech of both children ages 5 through 18 years and adults, Lee, Potamianos, and Narayanan (1999) showed that intratalker variation of segment duration and formant frequencies decreases sharply with age, reaching adultlike levels around age 12 years. However, in a more recent study of children ages 9 through 14 years, Romeo, Hazan, and Pettinato (2013) found that intratalker variation of voice onset time in /p/ and spectral mean in /s/ did not reach adultlike levels even by age 14 years and that there was no linear relationship between age and /s/-/ʃ/ or /b/-/p/ between-categories discriminability. Given that most English-acquiring children with typical development are judged to accurately produce the entire English phonological inventory by age 9 years (Smit et al., 1990), there is probably a good deal of acoustic and articulatory flexibility in terms of what gets transcribed as correct. The situation is even more complicated once the productions of children with atypical phonological development are considered. For example, Todd, Edwards, and Litovsky (2011) found that the correct productions of /s/ and /ʃ/ of children with cochlear implants showed a smaller acoustic contrast, as quantified by differences in spectral peak and means during the frication noise, relative to children in two comparison groups: chronological-age peers and hearing-age peers (i.e., children with the same duration of auditory experience). These findings suggest that it is difficult to gauge the speech development of children by relying only on transcription (cf. Hewlett & Waters, 2004).

A more fine-grained measure of speech acquisition could be one that takes variability into account, as it has also been suggested that the larger intratalker variability in children's speech may be a reflection of underdeveloped speech motor control (Smith & Goffman, 1998). Many studies have offered evidence that children's speech can be more variable than that of adults (e.g., Eguchi & Hirsh, 1969; Koenig, Lucero, & Perlman, 2008; Lee et al., 1999; Munson, 2004; Romeo et al., 2013; Sharkey & Folkins, 1985; Whiteside, Dobbin, & Henry, 2003), although not all aspects of children's speech are uniformly more variable than that of adults (Stathopoulos, 1995), and variability does not necessarily decrease monotonically with age (Smith, Kenney, & Hussain, 1996).

Nevertheless, there is greater potential for variability in children's speech, and it is known that intratalker variability can have perceptual consequences. For example, Newman, Clouse, and Burnham (2001) investigated the effect of a talker's /s/-/ʃ/ between-categories overlap and within-category dispersion on listeners' reaction time in an identification task. They found that response times were significantly slower on stimuli from a talker who exhibited between-categories overlap relative to stimuli from another talker who produced separable categories. However, response times were not significantly slower on stimuli produced by a talker who produced categories that were only barely separable (on the basis of the distance in centroid frequency between the lowest /s/ and highest /ʃ/) relative to stimuli from another talker with separable categories and a large difference between category extremes. These results suggest that it is the presence of overlap, and not variability, that slows identification. Hazan, Romeo, and Pettinato (2013) also tested the impact of intratalker variability on perception of the /s/-/ʃ/ contrast. They used stimuli produced by children ages 9 to 14 years in addition to adults and compared the effects of between-categories overlap with the effects of the distance between the category means (hereafter referred to as *between-categories distance*) while not explicitly controlling or varying within-category

dispersion. Listeners heard stimuli produced by speakers exhibiting one of three kinds of variability quantified in terms of spectral mean: categories that were very close but did not overlap, categories that were spread far apart and did not overlap, and categories that overlapped substantially. As in Newman et al. (2001), Hazan et al. (2013) found an overall effect of variability type such that reaction times were slower in response to category overlap than when the categories were close but did not overlap, suggesting that "the mere presence of overlap in a talker's categories affects the speed of perception over and above the magnitude of distance between them" (Hazan et al., 2013, p. 4). Although they found that this effect was driven mostly by the responses to stimuli produced by children, even within adult talkers the effect of category distance was smaller than that of category overlap. Last, although they did not set out to investigate the effect of within-category dispersion, they found that there was a significant overall correlation between within-category dispersion and reaction time, although it is not clear whether this effect would remain after controlling for overlap.

The measure of /s/-/ʃ/ category overlap used in Hazan et al. (2013) was the distance, in Hz, between the maximum spectral mean for /ʃ/ and the minimum spectral mean for /s/. This measure of category overlap is analogous to using the sample range as the measure of within-category variability rather than a more robust measure of statistical deviation, such as the sample standard deviation or median absolute deviation. The present study builds on this previous work by introducing an additional measure of the robustness of the English /s/-/ʃ/ contrast that is based not on the distance between any individual points in the distributions but instead on the degree to which overlap between the two distributions affects the likelihood of misidentifying the target sound for each of the sampled productions. This measure was tested by applying it, along with the three measures used in the two previous studies, to productions of English /s/ and /ʃ/ elicited from 81 children and 20 adults.

The /s/-/ʃ/ contrast was chosen in part because the voiceless sibilant fricatives are acquired relatively late despite being reasonably well attested in words in even the smaller vocabularies of preschool children. That is, looking at this contrast in preschool children provides an opportunity to compare variation across children of different ages in the robustness of contrast measures that are based on acoustic distribution with variation in accuracy measures that are based on transcription. Therefore, in this study, we first analyzed transcribed accuracy rates and the most common error patterns in productions of initial /s/ and /ʃ/ in real words elicited from the 81 children ranging in age from 2 years through 5 years (25–71 months). We then applied the four measures of robustness of contrast to the subset of the same initial /s/ and /ʃ/ productions that were transcribed as at least moderately correct by virtue of being unambiguously sibilant as well as to productions of these words elicited from 20 adults. We predicted that accuracy rates would be fairly closely related to age for both fricatives and that the robustness of contrast in the sibilant productions would be related to age to a similar extent. That is, we expected a fairly close relationship to age, although some of the younger children might have higher accuracy rates than some of the older children and, similarly, some of the younger children might have a more robust /s/-/ʃ/ contrast than some of the older children.

Last, we also report the results of a perception experiment that used goodness ratings rather than reaction times. We took a subset of 34 children's productions of /s/ and /ʃ/ in these words and in some nonwords, choosing only productions that were transcribed as correct, and used them as stimuli in a perception experiment to explore the relationship between robustness of contrast and perceived goodness. Although all of the stimuli were transcribed as correct productions, we predicted that the productions that came from children with a more robust contrast would be rated as better exemplars of the target category than productions from children with a less robust contrast, and this prediction was borne out.

Experiment 1: Production

Method

Speech Materials and Elicitation Procedure

The productions of word-initial /s/ and /ʃ/ are taken from the English part of the paidologos corpus that is described by Edwards and Beckman (2008). We used a picture-prompted auditory word repetition task to elicit children's and adults' productions of the real words and nonwords shown in Table 1. These were a subset of a larger list that included words and nonwords beginning with other lingual obstruents.[1]

For the real words, participants were presented with both a picture of a familiar object or event (e.g., a bowl of soup for *soup*, children standing under a fountain for *soak*) and the auditory stimulus (i.e., a production of the target word pronounced in a child-directed style by an adult female speaker of the target dialect) and were asked to repeat the stimulus item. The nonword repetition protocol was identical except that the pictures were of unfamiliar objects (e.g., a pile of raw turmeric, a red panda).

The real words and nonwords were elicited from each participant in one of three pseudorandom orders, which distributed trials for each target consonant in each vocalic context evenly across blocks. Elicitation was done using a tcl/tk program that, on each trial, loaded and showed the picture and then played the audio prompt once after a 300-ms delay. The audio prompt was played a second time if the first presentation of the audio prompt did not result in a clear recording of the target word. This occurred under the following circumstances: the child's first repetition was obscured by background noise, the child produced a

[1]The transcribed recordings of the children's real word productions are available to the public through the PhonBank archive at http://childes.psy.cmu.edu/media/Eng-NA/PaidoEnglish/

Table 1. Real-word and nonword stimuli used in the picture-prompted word repetition task.

	/ʃ/		/s/	
Context	**Words**	**Nonwords**	**Words**	**Nonwords**
High front vowel	shield, sheep, ship		seal, seashore, sister	sibɪθ, sigɪn, sivaɪt, sigənɑp, sibɪlaɪd, sitʃəmut, sivəblut, sitʃəkloɹ, sivɪfɹæʃ
High back vowel	shoe, chute, sugar	ʃup^{h}ɑs, ʃuvɑs, ʃumɛl, ʃunəvaɪt, ʃugɪmɪg, ʃufəkɹɑm, ʃukɪgɹaɪf, ʃubəmid, ʃunəfɹɑp	soup, super, suitcase	sugɪn, suvaɪt, subɪθ, subɪlaɪd, sutʃəmut, sugənɑp, sutʃəkloɹ, suvɪfɹæʃ, suvəblut
Midfront vowel	shape, shell, shepherd		safe, same, seven	sevaɪt, sebɪθ, segɪn, setʃəmut, segənɑp, sebɪlaɪd, sɛvɪfɹæʃ, sɛvəblut, sɛtʃəkloɹ
Mid- or low back vowel	show, shoulder, shore, shovel, shark, shop		soak, soldier, sodas, sun, soccer, sauce	sʌp^{h}on, sʌfim, sʌkɪtʃ, sʌʃəgip, sʌzɪvaɪt, sʌgənut, sɑkəpɹot, sɑpɪglok, sɑnəkɹæd

Note. The participants heard and repeated all of the real words, but there were three different audio prompts for each word, to make three lists. For the nonwords, there were not just different tokens but also different following frame portions, which were rotated among initial consonant–vowel targets across the three lists so that each participant heard only one disyllabic and two trisyllabic nonword stimuli in each vowel context.

nontarget word, the child made no response at all, or the child repeated the word very softly.

The entire elicitation session of each participant was recorded for subsequent transcription and acoustic analysis. This recording was made using a PMD660 flash card recorder (Marantz, Mahwah, NJ) and a C5900M condenser microphone with a cardioid response (AKG Acoustics, Vienna, Austria). The microphone was either mounted on a desk stand positioned about 30 cm away from the participant's mouth or held by the tester about 30 cm away from a child participant's mouth if the child was fidgety or too small to sit at the testing table at a good distance from the microphone.

Participants

A total of 81 children participated in the study. There were 20 (or 21) children from each of four age groups (2-, 3-, 4-, and 5-year-olds), with 10 girls and 10 boys per age group (except there were eleven 4-year-old boys). All children came from families of middle socioeconomic status in Columbus, Ohio, and were recorded in a quiet room at their day care centers or preschools. All children had normal speech, language, and hearing on the basis of parent report and a screening that we conducted. The screening included a hearing screening (pure-tone screening at 25 dB HL for 500, 1000, 2000, and 4000 Hz or otoacoustic emissions at 2000, 3000, 4000, and 5000 Hz) and norm-referenced measures of expressive vocabulary (Williams, 1997), receptive vocabulary (Brownell, 2000), and articulatory accuracy (Goldman & Fristoe, 2000). Any child who did not pass the hearing screening in at least one ear or who scored more than 1 *SD* below the mean on the norm-referenced measures was excluded from the current study. Any child whose parent reported that the primary language spoken in the home was not English also was excluded.

In addition to the 81 children, 20 adults completed the same two picture-prompted word or nonword repetition tasks (although they were recorded in a sound booth on the campus of The Ohio State University, Columbus). The adults also had normal speech, language, and hearing, assessed by self-report.

Transcription

All transcriptions were undertaken by native speakers and phoneticians who were not authors of the current article. A single native speaker and phonetician transcribed the initial consonant in all of the children's target productions. Productions were transcribed for both real words and nonwords. For the current study, the nonword transcriptions were used only to pick out a subset of the stimuli for a perception experiment (Experiment 2), so this section focuses on the real words.

In most cases for real words, the transcribed target production was the child's repetition in response to the first presentation of the audio prompt. However, in 94 cases, the response to the first presentation of a real word could not be transcribed because there was background noise or because the child produced the wrong word or spoke too softly; in these cases, the transcribed target production was the child's repetition in response to the second presentation of the audio prompt. In another 87 cases, the child's response to the second presentation also could not be transcribed; in these cases, the number of tokens analyzed for that target consonant for that child was reduced.

Transcription was done by both listening to each production and examining its waveform and spectrogram. Transcription was a two-step process. First, the transcriber decided if the production was correct or incorrect—a binary and categorical decision. Second, the transcriber did a fairly narrow transcription of what she heard using the consonant categories symbolized in the International Phonetic Alphabet (IPA) plus two more categories for distortion (i.e., a production not easily assigned to any consonant category symbolized in the IPA) and deletion (i.e., a production that audibly began with some other later sound, such as the following vowel target). Possible transcriptions of consonants that were not distortions or deletions included the target phoneme itself, a clear substitution of another phoneme of

English, or a non-English consonant category. Possible transcriptions also included combinations of two IPA symbols for a production that was judged to be intermediate between two sounds, such as the combination [s]:[s^{j}] for a production of some anterior sibilant sound intermediate between the English phoneme [s] and the out-of-inventory sound [s^{j}]. When using such a combination of symbols, the transcriber was required to also choose one of the two symbols as the more dominant one in the percept. This allowed a production that was coded as correct in the categorical decision at the first step of the transcription process to also be symbolized as an intermediate sound in the second step of the transcription process if it was judged to be more similar to a clearly correct production of the target consonant than to a prototypical example of the other sound. Thus, "[s]:[s^{j}]" was a possible transcription for a production of target /s/ that was judged to be correct (i.e., marginal but acceptable) as well as for a production of target /ʃ/ that was judged to be clearly incorrect (i.e., a substitution of some other more anterior sibilant fricative for the target postalveolar place). Of the 2,249 transcribable tokens, 540 were transcribed as intermediate between two categories in this way.

A second native speaker and phonetician independently transcribed 12% of the children's transcribable productions of /s/ and /ʃ/ in real words. This 12% comprised productions from two 2-year-olds, two 3-year-olds, two 4-year-olds, and two 5-year-olds. Phoneme-by-phoneme intertranscriber reliability was 84% averaged over all of the children's productions and 87% averaged over the productions of the 3- to 5-year-olds. Point-by-point agreement on whether or not a production was a sibilant fricative was 88% averaged over all of the children's productions and 90% averaged over the productions of the 3- to 5-year-olds.

Intertranscriber reliability was lower when productions of 2-year-olds were included because these productions had the lowest accuracy rate (68% for productions of 2-year-olds compared with 89% for productions of 5-year-olds). Of course, as Pye, Wilcox, and Siren (1988) pointed out, the productions that are most informative with respect to children's phonological acquisition are incorrect rather than correct productions; furthermore, productions that transcribers disagree on are particularly informative because they often are intermediate productions that don't fall clearly into a single phoneme category. This is another reason why measures such as the robustness of contrast measures examined in this article are so important for supplementing transcription-based measures.

The real-word transcriptions were used in three subsequent analyses. First, we analyzed the phonemic judgments of whether each production was correct or incorrect from the first step of transcription in order to see whether the proportion of correct tokens across the age groups mirrors the results for age of acquisition of English /s/ and /ʃ/ from earlier norming studies, such as Smit et al. (1990). Second, we analyzed the narrow phonetic encoding of each production at the second step of transcription in order to assess whether the dominant error patterns replicate findings reported in the literature on acquisition of sibilant fricatives by English-learning children. Third, we used the narrow phonetic transcriptions to also determine whether a production could be included in the spectral analysis for the quantitative measure of the /s/-/ʃ/ contrast described later.

Fricative Event Tagging

A team of five trained phoneticians (who were not the same as the transcribers and were not authors of this article) tagged fricative events in each adult's production of the target consonant in each real word and in each child's target production that was transcribed as some kind of sibilant (either a fricative or an affricate), including cases in which the consonant was transcribed as intermediate between two sounds but the primary sound was some kind of sibilant. Some additional tokens were excluded by one of the event taggers for reasons that included excessive background noise, the sibilant interval being interrupted, or the presence of overlap with the tester's voice. Note that these circumstances would not necessarily preclude phonetic transcription but could interfere with the acoustic analysis of fricative spectra. Also note that the event tagging was done prior to the design of the current study and for several other purposes. For example, the most senior member of the team (who also trained the other four) tagged half of all of the productions for an analysis of sibilant fricatives across languages (Li, 2012). These tags were also used for an independent analysis of /s/ productions related to a set of perception experiments (Munson et al., 2010) in which acoustic analysis of productions that are intermediate between [s] and [θ] was relevant. For the purposes of the current study, however, inclusion of nonsibilant fricatives could confound the analysis of the /s/-/ʃ/ contrast, so any tokens whose transcriptions contained a nonsibilant fricative element, such as [θ] or [f], were specifically excluded because the spectral measure used in the analysis described below could not serve as a reliable measure of place of articulation for fricatives with a diffuse spectrum. The final number of tokens remaining for the children's productions was 1,787.

For each token included in the spectral analysis, one of the event taggers marked the onset of frication and the fricative–vowel boundary by inspecting the spectrogram and waveform simultaneously in a Praat editor window. Each fricative's onset was marked at the earliest point at which an increase in the waveform's amplitude coincided with the presence of high-frequency energy in the spectrogram. For the fricative–vowel boundary, the onset of periodicity in the vocalic portion was first determined by inspecting the spectrogram. The fricative–vowel boundary was then marked at the zero crossing of the waveform's upswing that immediately followed the first downswing after the onset of periodicity.

The fricative events in approximately 5% of the children's tokens were independently tagged again by a second trained phonetician, who was a different member of the original team of phoneticians. This retagging was done as part of the calibration of the original tagging. Events in another 5% of the children's tokens were tagged by one of

the authors of the current article, who also was trained by the original lead tagger, so that proportionally as many tokens could be included in an evaluation of intertagger reliability as had been included in the evaluation of inter-transcriber consistency. The median absolute difference between the original tags and the two phoneticians' retags was 1.9 ms, and 85% of the tokens had an absolute difference of less than 10 ms.

Spectral Estimation and Centroid Computation

First, the waveform of each event-tagged sibilant token (i.e., the duration spanning from frication onset to the fricative–vowel boundary) was read from the source wave file into an R programming environment. The waveform was preprocessed by normalizing its amplitude so that its maximum was equal to one, but the waveform was neither pre-emphasized nor zero padded.

To estimate the spectrum of a sibilant production, the middle 50% of its amplitude-normalized waveform was extracted with a rectangular analysis window. From this, a multitaper spectrum (MTS; Thomson, 1982) was computed using parameter values K = 8 and NW = 4, where K is the number of tapers and NW is the time-bandwidth parameter. This MTS is equivalent to the pointwise average of eight statistically independent discrete Fourier transforms, computed from eight copies of the same waveform that have been shaped by eight different analysis windows. A more thorough introduction to the MTS, written for speech scientists, can be found in either Blacklock (2004) or Reidy (2013).

To compute the centroid frequency of a spectral estimate, that spectrum's amplitude values within the bandlimited frequency range 0.3 to 20.0 kHz were normalized so that they summed to one; thus, the distribution of energy across frequencies could be treated as a probability distribution over frequency. The centroid was then found by computing the expected value of this bandlimited, amplitude-normalized spectrum. In this way, the centroid represents a spectral estimate's mean frequency, or its center of gravity along the frequency scale.

The centroid values were then used to represent each participant's /s/ and /ʃ/ categories as point clouds in a one-dimensional centroid reference frame, and various structural properties of these point clouds, which indicate a participant's robustness of contrast, were calculated. The decision to compute these robustness of contrast measures from linear frequency centroid values was made out of a desire to investigate the relationship between sibilant perception and a novel robustness of contrast measure, described in detail later, such that our results would be directly comparable to previous work on the effects of a talker's robustness of contrast on a listener's perception (e.g., Hazan et al., 2013; Romeo et al., 2013). For this reason, the centroid measure and the method for computing it were jointly chosen to mirror the spectral analysis of Romeo et al. (2013), to date the most comprehensive study of such robustness of contrast measures for sibilants.

Last, because the children were tested in a room at their school rather than in a sound booth, there was a risk that background noise could distort any spectral measures computed from their productions. This background noise could be of two types: (a) transient artifacts of events such as doors closing, chairs being moved across the floor, or children screaming, or (b) persistent artifacts due to room acoustics. To ensure that transient background noises did not spectrally distort the recordings of the children, the phoneticians who tagged fricative events were instructed to exclude any and all productions that co-occurred with an audible transient background noise.

To ensure that persistent ambient background noise did not distort the spectra of the children's productions, the spectra of the background noise during the children's and adults' recordings were compared. A Welch's t test revealed no significant difference between the spectral slopes of the children's background noise ($M = -1.7013 \times 10^{-6}$) and those of the adults ($M = -1.6157 \times 10^{-6}$), $t(74.005) = -0.4833$, $p = .6303$, $d = 0.072$, which suggests that there is no reason to suppose that differences in recording environment confounded the spectral centroid measures.

Robustness of Contrast Measures

Because results of previous studies (Hazan et al., 2013; Newman et al., 2001) have suggested that the degree of a speaker's category overlap may play a more important role in consonant intelligibility than within-category dispersion or between-categories distance, our primary measure of robustness of contrast was designed to capture only the degree of overlap unconfounded by category distance. That is, two children whose /s/ and /ʃ/ categories are completely separable were treated as having equally robust contrasts even if one child's categories were closer together than those of the other child.

To estimate this degree of overlap, we calculated the percentage of a child's fricative productions whose category could be correctly predicted by the output of a mixed-effects logistic regression model built on the productions of all children in our sample. We chose to use a single mixed-effects model for all children in the corpus rather than separate regression models built for individual children because a combined model should estimate the model parameters more conservatively. It is possible that some children could have idiosyncratic production patterns that could render their fricative category distributions well separable but non-target-like. Because adult listeners already have a representation of what a good /s/ and /ʃ/ should sound like, we must estimate the robustness of a child's contrast with respect to this community-wide representation.

The overlap measure was calculated as follows. First, we built a mixed-effects logistic regression model with the following structure using the lme4 package (Bates, Maechler, & Bolker, 2013) in R:

$$\text{target} \sim 1 + \text{centroid} + (1|\text{talker}) + (0 + \text{centroid}|\text{talker})$$

The dependent variable was the target fricative, either /s/ or /ʃ/. The model contained a fixed effect of centroid frequency, with the groupwide centroid distribution centered at zero

and individual talker–level random intercepts and slopes. The model was built using only the real-word fricative productions from the 81 children described previously, of which there were 1,787. The number of tokens included in the model from each age group is shown in Table 2 (see also Figure 1).

The lme4 output returns a group-level intercept and slope for centroid and individual-level adjustments to both intercept and slope for each child. The individual-level adjustments can be added back to the group-level intercept and slope to obtain an individually fit model for each child, which lets us make a prediction for each token on the basis of its centroid, whether it is an /s/ or /ʃ/. Once a prediction has been made for each token, a percentage of tokens correctly predicted (%CP) can be calculated per child. For example, if a child has a %CP of .80, then the model was able to predict the target category of each of that child's fricative productions with 80% accuracy. We interpret %CP as an independent measure of category overlap because the distance between category means or category minimums or maximums do not figure into its calculation. We did not build a separate model for the adult productions because all 20 adult talkers' /s/ and /ʃ/ categories were linearly separable, indicating that all adult talkers had a %CP of 1.

In addition to %CP, for each talker we calculated three of the variability measures discussed in Hazan et al. (2013) and Romeo et al. (2013). Within-category dispersion was calculated as the mean of the standard deviation of the centroid frequencies of both categories—that is, $(\sigma_s + \sigma_\int)/2$. Between-categories distance was calculated as the difference between the mean centroid frequencies of both categories—that is, $\mu_s - \mu_\int$. Last, a discriminability score, $d(a)$, was calculated as the between-categories distance divided by the square root of the mean of the centroid frequency variances of the two categories—that is, $(\mu_s - \mu_\int)/\sqrt{((\mathrm{Var}_s + \mathrm{Var}_\int)/2)}$. These three measures and %CP are hereafter collectively referred to as *measures of robustness of contrast*.

Results

Transcribed Accuracy Rates and Error Patterns

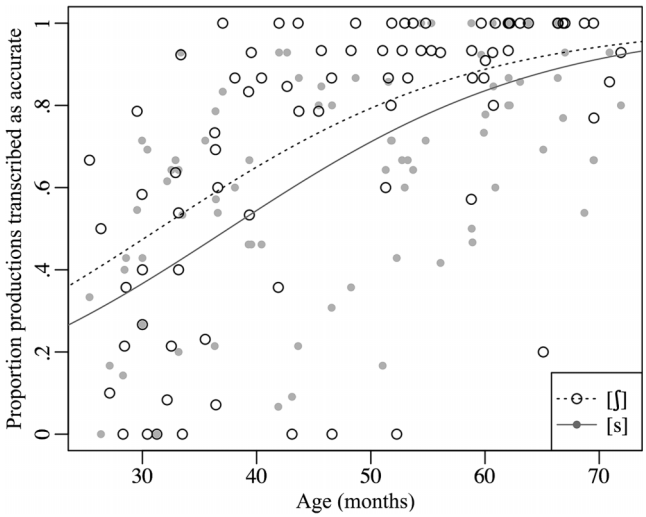
Figure 1 shows the transcribed accuracy rates of the children's productions as a function of the children's ages. The two curves in the figure are the results of a mixed-effects logistic regression with age and target consonant as fixed effects and random (individual child–level) intercepts. There was a significant effect of age ($b = .0724$, $z = 7.525$, $p < .0001$), for an estimated increase in accuracy rate of 63.0% over the age range of the children, with only five of the 2-year-olds

Table 2. Number of tokens per age group used in the logistic regression model.

Variable	Age (years)			
	2	3	4	5
Tokens (*n*)	353	413	500	521

Note. Total tokens = 1,787.

Figure 1. Transcribed accuracy rate for each child's productions as a function of age. The dashed black and solid gray lines are model curves from a mixed-effects logistic regression that predicted whether the production was transcribed as correct from age and the target consonant.



but 19 of the 5-year-olds being transcribed as correct on more than 50% of their tokens. There was also a small but significant effect of target consonant ($b = .4421$, $z = 4.203$, $p < .0001$), such that a token of /s/ was somewhat less likely to be transcribed as correct relative to a token of /ʃ/ (an estimated difference of 8.7% overall). Both of these effects are in keeping with the results of Smit et al. (1990).

Table 3 lists the three most commonly transcribed sounds for each of the two target consonants, when the production was either deemed to be correct at the first stage of transcription (left four columns) or deemed to be incorrect (right four columns). Each count in the right-hand columns adds together the number of instances of that symbol when it was transcribed alone and when it was used to transcribe the closer of the two sounds for a token that was judged to be intermediate between two sound categories. As Table 3 shows, the most frequent error transcribed for /s/ was a frontal misarticulation (i.e., substitution of the voiceless weak interdental fricative [θ]), and [θ] was the sound most commonly involved when a correct token of /s/ was transcribed as intermediate between [s] and another sound. The most frequent error transcribed for /ʃ/ also was a fronting (i.e., substitution of the other voiceless sibilant fricative [s]), and [s] was the sound most commonly involved when a correct token of /ʃ/ was transcribed as intermediate.

Both of these fronting patterns, also referred to as *dentalization* and *depalatalization*, respectively, are stereotypical errors for very young English-speaking children (Haelsig & Madison, 1986; James, 2001; Stoel-Gammon & Dunn, 1985, p. 40) and are often transcribed in children's productions of the English sibilant fricatives by speech-language pathologists when administering norm-referenced tests. However, only the second of these common error types

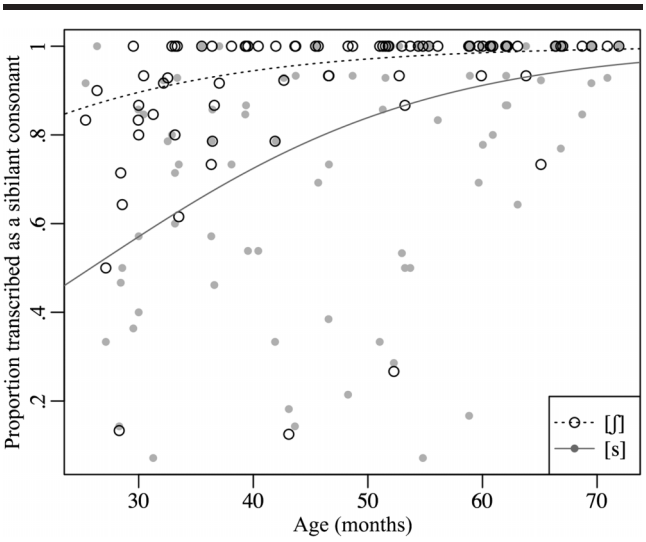
Table 3. Number (proportion) of the 710 tokens of /s/ and the 812 tokens of /ʃ/ that were judged to be correct (left four columns) and of the most commonly transcribed sounds for the 400 tokens of /s/ and the 327 tokens of /ʃ/ that were judged to be incorrect (right four columns).

/s/	Correct	/ʃ/	Correct	/s/	Incorrect	/ʃ/	Incorrect
[s] alone	578 (.81)	[ʃ] alone	722 (.89)	[θ]	85 (.21)	[s]	119 (.36)
[s]:[θ]	65 (.09)	[ʃ]:[s]	42 (.05)	[ʃ]	79 (.20)	[tʃ]	57 (.17)
[s]:[ʃ]	24 (.03)	[ʃ]:[tʃ]	16 (.02)	[ts]	63 (.16)	[θ]	23 (.07)
[s]:other	43 (.06)	[ʃ]:other	32 (.04)	Other	173 (.43)	Other	128 (.39)

is relevant for the place of articulation contrast between /s/ and /ʃ/. Thus, the 85 incorrect tokens of /s/ (and the 23 incorrect tokens of /ʃ/) that were transcribed as substitutions of [θ] and the 65 correct tokens of /s/ that were transcribed as intermediate between [s] and [θ] were not analyzable by the criterion for inclusion described in the Method section because their transcription contained a nonsibilant fricative element. In addition, the same is true of 14 tokens transcribed as substitutions of the weak palatal fricative [ç], of the affricate [kç], or of a sound that is intermediate between some sibilant fricative and [ç].

Figure 2 shows the proportion of analyzable tokens once these nonsibilant productions were excluded, child by child, again as a function of age. The dashed black and solid gray lines are model curves from a mixed-effects logistic regression that predicted whether the production was sibilant, with age and target consonant as fixed effects and random (individual child–level) intercepts. Both age (b = .0683, z = 4.959, p < .0001) and target consonant (b = 1.8698, z = 12.226, p < .0001) were significant predictors, with an estimated 46.7% increase in proportion of sibilant productions over the age range of the children and with the proportion of /ʃ/ targets produced as sibilants 14.5% greater than the proportion of /s/ targets produced as sibilants. These analyzable tokens are the productions that were included in the results described in the next sections. Eight of the 5-year-olds produced only tokens that were sibilants, whether correct or incorrect, and although only one 2-year-old produced only sibilants, there were ten 2-year-olds who produced at least 75% of their tokens as sibilant. Three notable exceptions were one 2-year-old and one 3-year-old who did not have any analyzable /s/ productions and one 4-year-old who produced only one analyzable /s/ token. These three children were excluded from all further analyses.

Figure 2. Proportion of tokens transcribed as being analyzable sibilant productions, plotted as a function of age.



Centroid Frequency

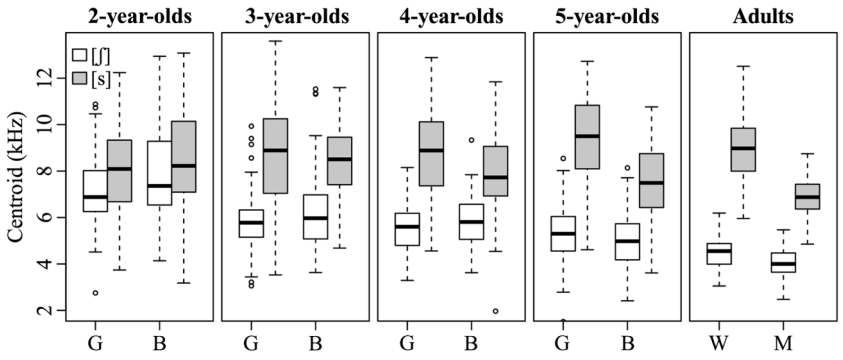
The distribution of centroid frequency across fricative target categories, age group, and gender, shown in Figure 3, presents a trend of increasing separation between the /s/ and /ʃ/ categories as age increases. To investigate the trends in mean centroid for /s/ and /ʃ/, separate two-way analyses of variance (ANOVAs) with between-subjects effects of age group (child vs. adult) and gender were run. For /s/, there was a main effect of gender, $F(1, 94) = 14.02$, $p < .001$, $\eta^2 = .122$, and a significant interaction between gender and age group, $F(1, 94) = 5.42$, $p = .022$, $\eta^2 = .047$, but no main effect of age group. Tukey's honestly significant difference (HSD) post hoc tests revealed that the only significant comparisons were those between men and both girls and women ($p < .001$ and $p < .002$, respectively), suggesting that the centroid of /s/ does not differ significantly between child and adult female speakers. For /ʃ/, there was a main effect of age group, $F(1, 94) = 35.62$, $p < .001$, $\eta^2 = .268$, but no main effect of gender or interaction between age and gender. Taken together, these results first confirm that the centroid frequency is very high for both fricatives in the youngest children, perhaps due to the effects on the "undifferentiated lingual gesture" of the generally high tongue tip in the "articulatory setting" of English (Wilson & Gick, 2014) and then suggest that in subsequent development, the centroid of /ʃ/ decreases with age for both genders, whereas the centroid of /s/ changes with age only for men. This interaction results in men having /s/ and /ʃ/ categories that are closer together than those of women.

The output of the mixed-effects logistic regression model described previously indicates that centroid frequency at the midpoint of the turbulent interval can reliably separate children's /s/ and /ʃ/ categories for productions that were transcribed as correct on at least being sibilant (b = .00147, z = 9.23, p < .001).

Robustness of Contrast

The four measures of robustness of contrast are plotted against age and raw scores for the norm-referenced measures of receptive and expressive vocabulary size in Figure 4, and summary statistics for both the children and

Figure 3. Distribution of centroid frequency across fricative target, age group, and gender. Each box plot shows the distribution of centroid for /s/ and /ʃ/ for girls (G) or women (W) and boys (B) or men (M) for each age group.



adults are reported in Table 4. The first column of panels in Figure 4 shows that %CP, between-categories distance, and $d(a)$ generally increase with age among children. On the basis of linear regression models, all three of these measures were significantly correlated with age (all comparisons $p < .00417$, the Bonferroni correction of $\alpha = .05$ for 12 comparisons), with the R^2 value for each regression printed above each individual plot. The same three measures were found to be significantly correlated with children's receptive vocabulary raw scores, shown in the second column of Figure 4. None of the measures were significantly correlated with the expressive vocabulary raw scores, shown in the third column; therefore, expressive vocabulary is excluded from further analysis. Within-category dispersion, on the other hand, decreased both with age and with increased receptive vocabulary ($p < .00417$), although the strength of these relationships ($R^2 = .156$ for age; $R^2 = .092$ for receptive vocabulary) was lower than that of the other three robustness measures ($.274 \leq R^2 \leq .346$ for age; $.178 \leq R^2 \leq .232$ for receptive vocabulary).

The relationships between the robustness measures and both age and receptive vocabulary were very similar, and a separate linear regression of receptive vocabulary raw scores against age confirmed that they were in fact highly correlated ($R^2 = .618$, $p < .001$). Because the robustness measures were more highly correlated with age than receptive vocabulary, we focus our remaining analyses on only age and gender differences. This focus also allows us to compare the children with the adults because the adults do not have vocabulary scores.

We ran separate two-way ANOVAs with between-subjects factors of age group (2-, 3-, 4-, and 5-year-olds and adults) and gender for %CP, within-category dispersion, between-categories distance, and $d(a)$. For %CP, there was a main effect of age group, $F(4, 88) = 24.22$, $p < .001$, $\eta^2 = .501$, but no main effect of gender. Tukey's HSD post hoc tests revealed significant differences ($p < .01$) between adults and all children's age groups except 5-year-olds. Among the children, %CP was significantly different between 2-year-olds and the older children, but not among the older children themselves. These results suggest a gradual increase in this measure of the robustness of contrast with age and suggest that by age 5 years the degree of overlap in children's /s/-/ʃ/ contrast may be comparable to that of adults.

For within-category dispersion, there was a main effect of age group, $F(4, 88) = 18.57$, $p < .001$, $\eta^2 = .451$, but no main effect of gender. Tukey's HSD post hoc tests revealed significant differences between adults and all children's age groups (all comparisons $p \leq .001$), indicating that even at age 5 years children still have greater levels of within-category dispersion than adults. Among the children, within-category dispersion significantly differed only between nonconsecutive age groups, suggesting that dispersion decreases quite gradually with age.

For between-categories distance, there were main effects of both age group, $F(4, 88) = 16.31$, $p < .001$, $\eta^2 = .355$, and gender, $F(1, 88) = 15.14$, $p < .001$, $\eta^2 = .089$. Tukey's HSD post hoc tests revealed no significant differences between adults and any children's age groups except for 2-year-olds. In addition, the 3-, 4-, and 5-year-olds were not significantly different from each other, indicating that changes in between-categories distance may occur between 2 and 3 years of age, after which change is gradual and may reach adult levels as early as age 3 years. This pattern of showing the most change between 2 and 3 years but less year-to-year change afterward was the same as what was found for %CP. Post hoc tests did not reveal significant gender differences within any individual age group, including adults.

For $d(a)$, there were main effects of both age group, $F(4, 88) = 38.02$, $p < .001$, $\eta^2 = .591$, and gender, $F(1, 88) = 9.93$, $p = .002$, $\eta^2 = .039$. Tukey's HSD post hoc tests revealed significant differences ($p \leq .01$) between the adults and all children's age groups and between the 2-year-olds and all other children. We found no significant differences between the 3-, 4-, or 5-year-olds, however. As with between-categories distance, the post hoc tests did not reveal significant gender differences within any individual age group, including adults.

Discussion

The output of the mixed-effects regression models indicates that centroid frequency is a reasonably good predictor of fricative category for both children and adults. It was also found that, although the mean /s/ centroid frequency did

Figure 4. Robustness of contrast measures for each child plotted against his or her age in months, receptive vocabulary raw score, and expressive vocabulary raw score. Solid lines indicate relationships that were statistically significant (α = .00417; the Bonferroni correction of α = .05 for 12 comparisons), and dashed lines indicate relationships that did not reach statistical significance. %CP = percentage of tokens correctly predicted; *d(a)* = discriminability score.

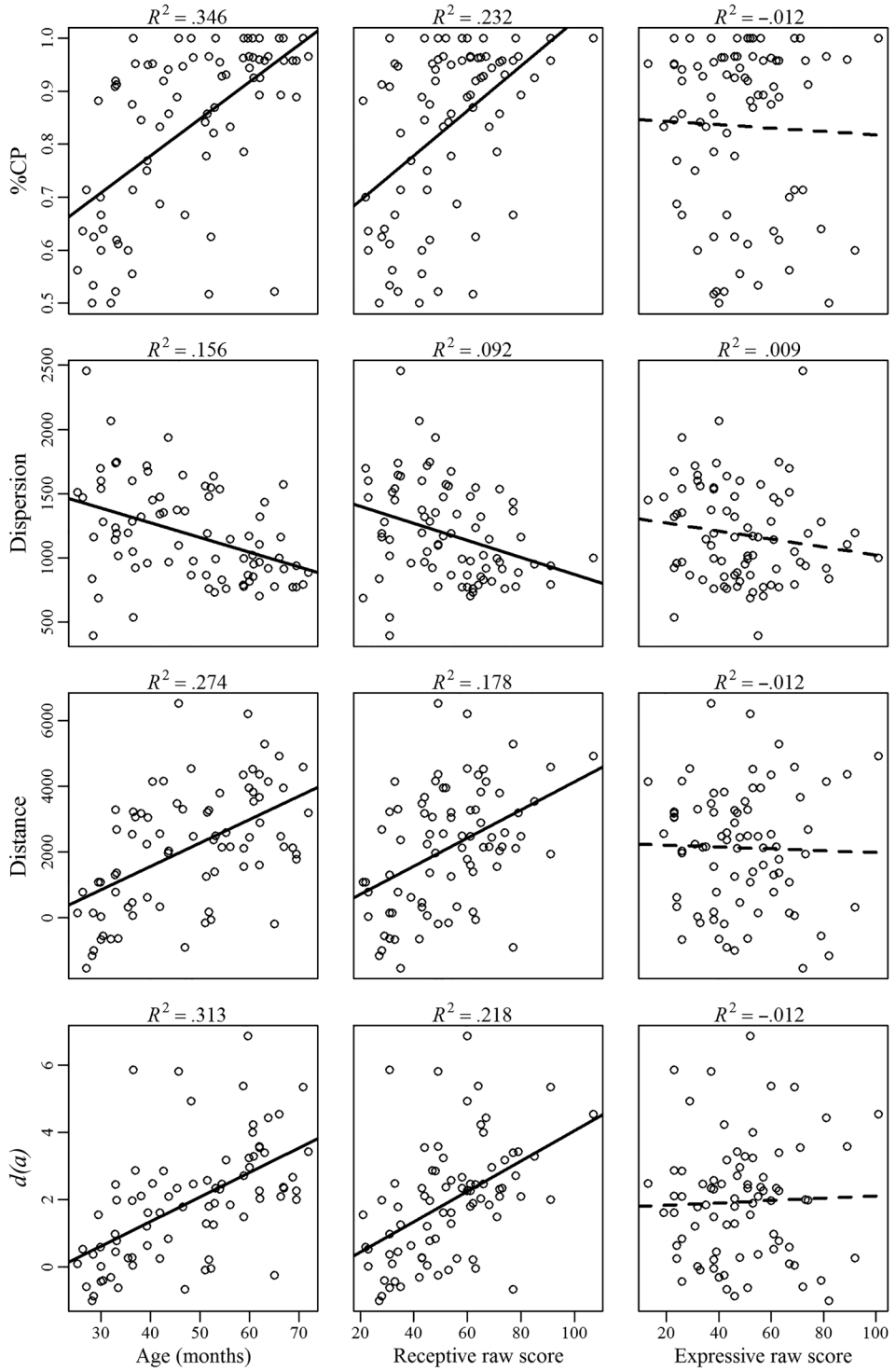


Table 4. Mean (standard deviation) of the four measures of robustness of contrast.

Variable	2-year-olds	3-year-olds	4-year-olds	5-year-olds	Adults
All participants					
%CP	.67 (.14)	.85 (.13)	.88 (.13)	.93 (.10)	1.00 (0)
Dispersion (Hz)	1367 (482)	1319 (338)	1074 (314)	984 (232)	577 (158)
Distance (Hz)	350 (1273)	2416 (1753)	2489 (1614)	3162 (1372)	3588 (1109)
d(a)	0.30 (0.95)	1.89 (1.68)	2.40 (1.75)	3.03 (1.23)	5.94 (1.86)
Female participants					
%CP	.68 (.14)	.84 (.15)	.91 (.06)	.97 (.03)	1.00 (0)
Dispersion (Hz)	1302 (346)	1316 (350)	1088 (337)	1074 (258)	627 (201)
Distance (Hz)	715 (1014)	2792 (2053)	2896 (1731)	3809 (1041)	4406 (914)
d(a)	0.57 (0.85)	2.29 (2.11)	2.82 (1.97)	3.39 (0.99)	6.83 (1.88)
Male participants					
%CP	.66 (.15)	.85 (.09)	.85 (.17)	.89 (.14)	1.00 (0)
Dispersion (Hz)	1413 (573)	1323 (345)	1059 (306)	874 (140)	528 (84)
Distance (Hz)	84 (1419)	2007 (1351)	2083 (1462)	2372 (1357)	2769 (525)
d(a)	0.11 (1.00)	1.45 (1.04)	1.99 (1.48)	2.58 (1.41)	5.06 (1.42)

Note. %CP = percentage of tokens correctly predicted; *d*(*a*) = discriminability score.

not significantly differ between children and adults, the mean /ʃ/ centroid frequency did decrease significantly with age for both genders.

The results of the robustness of contrast measures suggest that although %CP, within-category dispersion, between-categories distance, and *d*(*a*) were all significantly correlated with age, the four measures may differ in how well they capture more subtle variation. Although 5-year-olds were not significantly different from adults according to %CP or between-categories distance, all children's age groups significantly differed from adults according to within-category dispersion and *d*(*a*). Romeo et al. (2013) did not calculate %CP, but they did find that children ages 9 to 14 years had greater between-categories distance than adults, with this effect driven especially by a sudden jump in between-categories distance in 11- to 12-year-old girls. Taken together with the results of the current study, these results indicate that although between-categories distance is already at adultlike levels by age 5 years, it continues to increase for several more years until decreasing back down to adultlike levels during the teenage years.

The highly similar trends between %CP and between-categories distance suggest that the two measures may be strongly correlated, which was confirmed by linearly regressing the latter against the former ($R^2 = .680$, $p < .001$). Although %CP is a measure of category overlap that does not use between-categories distance in its calculation, it is expected that categories that are closer together are more likely to exhibit overlap, and vice versa. Therefore, although we believe it is important to not conflate the concepts of category overlap and category distance, it is also not surprising that they would pattern similarly.

Within-category dispersion and *d*(*a*) were not at adultlike levels by age 5 years. Because Romeo et al. (2013) found the same result for children ages 9 to 14 years, we can conclude that within-category dispersion decreases very gradually with age and that more development is taking place even between ages 14 and 18 years. The difference between children and adults in *d*(*a*) is likely due to the calculation of *d*(*a*) being based partly on within-category dispersion. Although between-categories distance also figures into the calculation of *d*(*a*), its effect was apparently outweighed by that of within-category dispersion.

In summary, although we found that all four robustness of contrast measures tested here were correlated with both age and receptive vocabulary size, the relationship with age among children ages 2 to 5 years was strongest for %CP and weakest for within-category dispersion. This result is seemingly at odds with the conclusions of Hazan et al. (2013) that talkers' intelligibility may be best predicted by within-category dispersion. Although the relationship between within-category dispersion and age is weaker in the current study, it remains possible that the within-category dispersion could still affect perception more than the other measures. Furthermore, because the relationship between these measures and age is not always linear (e.g., between-categories distance), it could be the case that the perception of younger children's fricatives is influenced by different factors. Our next step is to explore whether any of these four measures can capture degrees of perceived goodness more subtle than those captured through narrow phonetic transcription (Sovinski, 2011).

Experiment 2: Perception

Method

Stimuli

For the perception experiment, productions used for stimuli were chosen as follows. First, we selected productions that had been coded as correct in the first step (the phonemic judgment) of the two-step transcription process. These productions included both productions that were transcribed as the target phoneme and those that were transcribed as intermediate but closer to the target type than to the other transcribed type in the second step (the phonetic judgment) of the transcription process. Second, we included an intermediate production only if the other phoneme that

it was similar to was also a fricative. That is, a production that was coded as correct and as intermediate between [s] and [ʃ] would be included, but a production that was intermediate between [s] and [ts] or between [s] and [t] would not be included. Third, we tried to include productions from an approximately equal number of children at each age who had relatively high %CP and who had relatively low %CP. Last, as much as possible, we tried to choose an equal number of productions from children at each age, an equal number of productions from both boys and girls, and an equal number of /s/ and /ʃ/ productions. Because younger children produced relatively fewer correct real-word productions, we included nonword productions as well as real-word productions. We did this for both the younger and older children so that the distribution of stimuli made from real words and nonwords would be similar between the younger and older groups of children. The distribution of productions is shown in Table 5. Each stimulus item included the initial fricative and a 150-ms vocalic portion. The root mean square amplitude of all stimulus items was normalized to the mean RMS dB.

Participants

The participants in the perception study were 20 young adults (seven men, 13 women) enrolled in an introductory course in the Department of Communication Sciences and Disorders at the University of Wisconsin–Madison. All participants received course credit for their participation. No participants had a history of hearing loss or a speech or language disorder on the basis of self-report.

Procedure

Participants listened to six practice items and then to two blocks of the 376 stimuli, with a short break between the blocks. In one block, they rated each item in terms of its goodness as a production of /s/, and in the other block, they rated each item in terms of its goodness as a production of /ʃ/. The order of the two blocks was counterbalanced across listeners. Participants rated each production by using the mouse to click anywhere along a two-headed arrow on a computer screen. The label at the left end of the arrow was *good "s"* or *good "sh,"* and the label at the right end of the arrow was *bad "s"* or *bad "sh."* The items were presented in random order, and the experiment was self-paced. Participants were encouraged to use the entire line when rating the stimuli. (The instructions included the following: "We encourage you to use the whole line. That is, don't just click at the ends; click at the location on the line that corresponds to how good of an example you think the consonant was.") The experiment was run in E-Prime, and participants' responses were recorded automatically.

Table 5. Number of stimuli in the perception experiment.

Stimulus	2-year-olds	3-year-olds	4-year-olds	5-year-olds	Total
/s/					
Real word	34	31	35	36	136
Nonword	9	16	13	12	50
Total	43	47	48	48	186
/ʃ/					
Real word	39	47	43	43	172
Nonword	2	6	5	5	18
Total	41	53	48	48	190

Results

First, the mouse click *x*-coordinates were transformed to generalized logit values. Through this transformation the overall minimum and maximum values became negative and positive infinity, respectively, and responses with these values were discarded. The mouse click locations are hereafter referred to as the *goodness ratings* and are reported as transformed logit values. The mean and median goodness ratings were 0.098 and 0.197 for the "s" block and 0.108 and 0.204 for the "sh" block, respectively. A repeated measures ANOVA showed that goodness rating did not differ significantly across blocks, $F(1, 40) = 0.245$, $p = .623$, $\eta^2 = .0005$.

We then took the mean rating of each child's productions in each block to calculate a mean goodness rating; thus, each child was given one mean goodness rating for his or her "s" block and another mean goodness rating for his or her "sh" block. Figure 5 shows the relationship between each child's age and the mean goodness rating calculated across all of each child's productions. It shows that for both the "s" and "sh" blocks there was a general positive trend for perceived goodness to increase with age. On the basis of the R^2 value of each block, we can see that age may be a better predictor of perceived goodness for /ʃ/ ($R^2 = .251$, $p = .001$) than for /s/ ($R^2 = .107$, $p = .033$), but the relationship between age and perceived goodness does not seem particularly strong for either consonant.

We turn next to the relationship between perceived goodness and the four measures of robustness of contrast: %CP, within-category dispersion, between-categories distance, and $d(a)$. In the top two panels of Figure 6, each child's mean goodness rating is plotted instead against his or her

Figure 5. By-child mean perceived goodness plotted against age in months, separated by block.

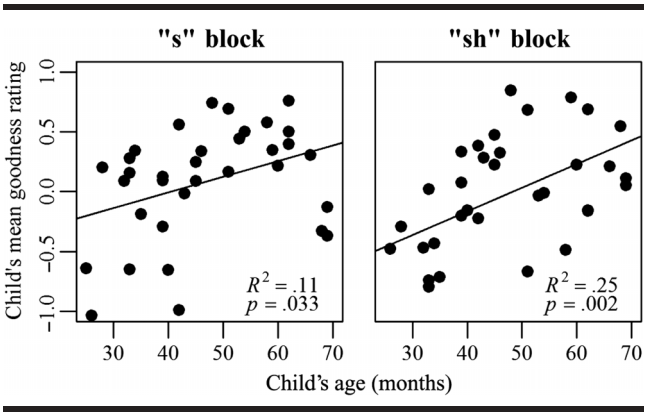
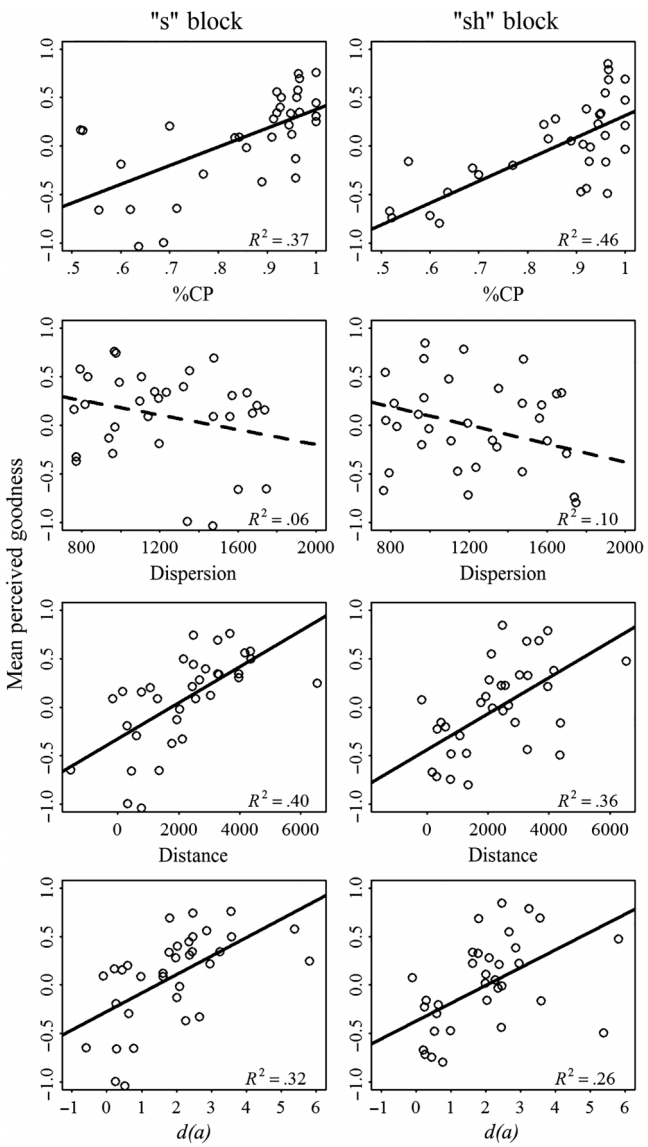


Figure 6. Mean goodness rating plotted against each measure of robustness of contrast, separated by block. Solid lines indicate relationships that were statistically significant (α = .00417; the Bonferroni correction of α = .05 for 12 comparisons), and dashed lines indicate relationships that did not reach statistical significance. %CP = percentage of tokens correctly predicted; *d*(*a*) = discriminability score.



%CP. With an R^2 value of .373 ($p < .001$) for the "s" block and .464 ($p < .001$) for the "sh" block, %CP appears to be a substantially better predictor of perceived goodness than age. The second row of panels shows the relationship between perceived goodness and within-category dispersion, which was relatively weak in both the "s" block ($R^2 = .065$, $p = .079$) and the "sh" block ($R^2 = .105$, $p = .035$).

Between-categories distance, shown in the third row of Figure 6, seems to be a good predictor of perceived goodness for both the "s" block ($R^2 = .399$, $p < .001$) and the "sh" block ($R^2 = .356$, $p < .001$), with the R^2 value for the "s" block being slightly higher than for %CP and the R^2 value for the "sh" block being a bit lower. Last, in the bottom row, *d*(*a*) appears to be a better predictor for the "s" block ($R^2 = .322$, $p < .001$) than for the "sh" block ($R^2 = .263$, $p = .001$), although both relationships are significant.

Discussion

In the perception experiment, we investigated the relationship between adult listeners' goodness judgments and each of the following independent variables: age, category overlap (%CP), within-category dispersion, between-categories distance, and discriminability, *d*(*a*). %CP and between-categories distance were similarly correlated with perceived goodness (average R^2 value across blocks was .418 for %CP and .378 for between-categories distance) and were even well correlated with each other. A remaining question is whether one measure might be more suitable than the other for quantifying robustness of contrast. Between-categories distance has the advantage of being simpler to calculate and is not bounded in the way that %CP is bounded between zero and one. As a measure of robustness of contrast, %CP predicts that all talkers with perfectly separable categories (i.e., %CP = 1) should have equally robust contrasts and be perceived as equally good. Furthermore, because %CP is a percentage, its granularity is limited by the number of tokens per talker in the available corpus. On the other hand, because between-categories distance is theoretically unbounded, it is predicted that perceived goodness or intelligibility should continuously increase with increasing distance. Previous studies have shown conflicting evidence for this claim in the perception of adult productions (Hazan & Baker, 2011; Hazan et al., 2013). In response to the stimuli produced by children ages 9 to 14 years in Hazan et al. (2013), the productions from children with greater between-categories distance levels were identified more slowly (indicating lower intelligibility) but with higher accuracy (indicating higher intelligibility). Because most of the children in that study had between-categories distance levels even higher than those of adults, the fact that the children's tokens were not overall identified more quickly or more accurately than those of adults suggests that between-categories distance is unlikely to be the primary predictor of intelligibility.

Among perception studies using stimuli produced by adults, Newman et al. (2001) concluded that distance mattered less than overlap. However, because the talker in their study who had greater overlap also had greater dispersion, the increased level of dispersion could have been driving the effect. On the other hand, Hazan and Baker (2011) found no effect of dispersion or distance on the intelligibility of adult fricative productions. These divergent findings across studies highlight the need for more studies of both child and adult talkers that look at mathematically independent measures of overlap, dispersion, and distance to better understand how intratalker variability varies both with age and across different phonological contrasts.

General Discussion

In this article, we showed that, although the transcribed accuracy of children's sibilant fricative productions generally increases with age, there is substantial variation between children within the same age group. We then quantified the robustness of children's fricative contrasts using four different measures and related these measures to not only the children's age and vocabulary size but also to adult listeners' goodness judgments of the children's fricative productions.

The findings presented here disagree with those of Hazan et al. (2013), who found that within-category dispersion was the best predictor of talker intelligibility, in that we did not find within-category dispersion to be related to perceived goodness. There are at least two possible explanations for this difference. First, Hazan et al. (2013) quantified intelligibility as listeners' response time in an identification task. Because listeners were overall very accurate at identification, it was presumed that response time would reflect the ease with which the stimuli could be identified. Although it is not clear how or whether identification response times and the goodness judgments used in the current study might pattern differently, this difference in methods should be noted. Second, the stimuli in the study of Hazan et al. (2013) were produced by older children (9- to 14-year-olds) whose level of between-categories distance was greater than even that of adults, whereas the stimuli in the current study were produced by 2- to 5-year-olds. Although we found %CP and between-categories distance to be moderately correlated with perceived goodness, it could be that once between-categories distance and %CP reach adultlike levels they affect perception less, leaving room for within-category dispersion to play a bigger role. An important difference between %CP and between-categories distance on one hand and within-category dispersion on the other is that the former are more directly related to the notion of phonological contrast. High levels of within-category dispersion may lead to categories overlapping or being close together, but dispersion in itself does not necessarily inhibit categories from being robustly differentiated. Perhaps for this reason dispersion is rightly referred to as a measure of variability in other studies (e.g., Romeo et al., 2013).

As such, within-category dispersion may not be a useful predictor of perceived goodness in very young children because variability does not necessarily reflect a lack of development. As Forrest, Elbert, and Dinnsen (2000, p. 520) pointed out,

> In some cases, low variability indicates inflexibility that limits learning, whereas increased variability is associated with periods of behavioural expansions (Tyler and Saxman, 1991; Forrest, Weismer, Dinnsen and Elbert, 1994). In other contexts, high variability restricts categorical development that may be prerequisite to the emergence of new phonemes in a child's inventory (Thelen and Smith, 1994; Forrest[, Dinnsen, and Elbert], 1997).

That is, there could be less dispersion in a younger child, with a fairly tight unimodal distribution for the two categories together, which could reflect a language-specific "undifferentiated lingual gesture," as described by Li (2012). As an alternative, children who are beginning to split a unimodal distribution of centroid frequency values into two distributions might exhibit greater within-category dispersion even as their development is reflected in greater between-categories distance and less overlap.

The relationships between the measures of robustness of contrast and the perceptual judgments in the current study were particularly interesting because only productions that were transcribed as correct were included in the perception experiment. Thus, the finding that a child's level of category overlap or between-categories distance can predict differences in perceived goodness even between correct productions suggests that adult listeners are sensitive to these within-category differences in children's productions.

What do these findings mean for speech-language pathologists who are working with children with atypical phonological development, such as children with phonological disorder or children with hearing impairment? Should clinicians continue to work on sounds even after children are perceived to produce a sound or a contrast correctly? It is unfortunate that almost no research addresses this question. In a perception study similar to the one described in this article, Bernstein, Todd, and Edwards (2013) found that tokens of /s/ produced by children with cochlear implants that were transcribed as correct were rated as less good than productions by children with normal hearing of the same age. It has been noted that speech intelligibility of children with cochlear implants is reduced relative to children with normal hearing, even for children who are implanted early and have had 7 years of experience with their cochlear implant (Peng, Spencer, & Tomblin, 2004). These findings suggest that, at least for children who have difficulty perceiving a contrast, continuing to work on consonant contrasts even after productions are perceived as correct may improve intelligibility. Furthermore, it may be useful to include additional assessments of correct production over and above the categorical transcription judgment of correct versus incorrect. These could include visual analogue rating scales with naive listeners or acoustic and psychoacoustic measures.

To conclude, this study found that the robustness of contrast between /s/ and /ʃ/, as measured by %CP and between-categories distance, gradually increased from ages 2 to 5 years. Differences were also observed between 5-year-olds and adult speakers. Differences in robustness of contrast were also reflected in adults' perceived goodness ratings, even for productions transcribed categorically as correct. These results suggest that further research is needed to evaluate whether the speech intelligibility of children with atypical phonological development would be improved if speech-language pathologists worked to ensure that children produced a robust contrast rather than just a correct contrast.

Acknowledgments

This research was supported by National Institute on Deafness and Other Communication Disorders Grant R01-02932 to

Jan Edwards and Mary E. Beckman, NSF Grant BCS-0729140 to Jan Edwards, NSF Grant BCS-0729306 to Mary E. Beckman, and National Institute of Child Health & Human Development Grant P30-HD03352 to the Waisman Center. We thank Laura Slocum for making the stimuli for the production experiment; Laura Slocum and Anne Hoffmann for recruiting and recording the children and adults for this experiment; Fangfang Li, Chanelle Mays, Oxana Skorniakova, Asimina Syrika, and Julie Johnson for tagging the edges of the target sibilant fricatives in these productions; Eunjong Kong for help with the development and analysis of Experiment 2; and Ryan Sovinski for recruiting and recording the adults for Experiment 2 (which was her master's thesis). We also thank the children who participated in Experiment 1 and their parents, as well as the participants of Experiment 2. The idea of using slopes and prediction accuracy rates from logistic regression models was presented at Interspeech 2010 in an article by the first and third authors and Chanelle Mays, and we thank the audience at Interspeech for useful comments that helped shape the further development of the idea in the current article.

References

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear Mixed-Effects Models Using S4 Classes (R Package Version 0.999999-2). Retrieved from http://CRAN.R-project.org/package=lme4

Baum, S. R., & McNutt, J. C. (1990). An acoustic analysis of frontal misarticulation of /s/ in children. *Journal of Phonetics, 18,* 51–63.

Bernstein, S., Todd, A., & Edwards, J. (2013, November). *How do adults perceive the speech of children with cochlear implants?* Poster presented at the Annual Conference of the American Speech-Language-Hearing Association, Chicago, IL.

Blacklock, O. S. (2004). *Characteristics of variation in production of normal and disordered fricatives, using reduced-variance spectral methods* (Unpublished doctoral dissertation). University of Southampton, United Kingdom.

Brownell, R. (Ed.). (2000). *Receptive One-Word Picture Vocabulary Test–Second Edition*. Novato, CA: Academic Therapy Publications.

Edwards, J., & Beckman, M. E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics & Phonetics, 22,* 939–958.

Edwards, J., Gibbon, F., & Fourakis, M. (1997). On discrete changes in the acquisition of the alveolar/velar stop consonant contrast. *Language and Speech, 40,* 203–210.

Eguchi, S., & Hirsh, I. J. (1969). Development of speech sounds in children. *Acta Otolaryngologica, 257,* 5–48.

Forrest, K., Dinnsen, D. A., & Elbert, M. (1997). Impact of substitution patterns on phonological learning by misarticulating children. *Clinical Linguistics & Phonetics, 11,* 63–76.

Forrest, K., Elbert, M., & Dinnsen, D. A. (2000). The effect of substitution pattern on phonological treatment outcomes. *Clinical Linguistics & Phonetics, 14,* 519–531.

Forrest, K., Weismer, G., Dinnsen, D. A., & Elbert, M. (1994). Spectral analysis of target-appropriate /t/ and /k/ produced by phonologically disordered and normally articulating children. *Clinical Linguistics & Phonetics, 8,* 267–282.

Forrest, K., Weismer, G., Hodge, M., Dinnsen, D., & Elbert, M. (1990). Statistical analysis of word-initial /k/ and /t/ produced by normal and phonologically disordered children. *Clinical Linguistics & Phonetics, 4,* 327–340.

Gibbon, F. (1990). Lingual activity in two speech-disordered children's attempts to produce velar and alveolar stop consonants: Evidence from electropalatographic (EPG) data. *British Journal of Disorders of Communication, 25,* 329–340.

Gibbon, F., Dent, H., & Hardcastle, W. (1993). Diagnosis and therapy of abnormal alveolar stops in a speech-disordered child using EPG. *Clinical Linguistics & Phonetics, 7,* 247–268.

Goldman, R., & Fristoe, M. (2000). *Goldman-Fristoe Test of Articulation–Second Edition*. San Antonio, TX: Pearson.

Haelsig, P. C., & Madison, C. L. (1986). A study of phonological processes exhibited by 3-, 4-, and 5-year-old children. *Language, Speech, and Hearing Services in Schools, 17,* 107–114.

Hazan, V., & Baker, R. (2011). Is consonant perception linked to within-category dispersion or across-category distance? In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 839–842). Hong Kong: City University of Hong Kong. Retrieved from http://www.icphs2011.hk/resources/OnlineProceedings/RegularSession/Hazan/Hazan.pdf

Hazan, V., Romeo, R., & Pettinato, M. (2013). The impact of variation in phoneme category structure on consonant intelligibility. *Proceedings of Meetings on Acoustics, 19,* 1–6.

Hewlett, N., & Waters, D. (2004). Gradient change in the acquisition of phonology. *Clinical Linguistics & Phonetics, 18,* 523–533.

James, D. G. H. (2001). Use of phonological processes in Australian children ages 2 to 7;11 years. *Advances in Speech-Language Pathology, 3,* 109–127.

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

Koenig, L. L., Lucero, J. C., & Perlman, E. (2008). Speech production variability in fricatives of children and adults: Results of functional data analysis. *The Journal of the Acoustical Society of America, 124,* 3158–3170.

Kornfeld, J., & Goehl, H. (1974). A new twist to an old observation: Kids know more than they say. In A. Bruck, R. Fox, & M. La Galy (Eds.), *CLS Parasession on Natural Phonology* (pp. 210–219). Chicago, IL: Chicago Linguistics Society.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105,* 1455–1468.

Li, F. (2012). Language-specific developmental differences in speech production: A cross-language acoustic study. *Child Development, 83,* 1303–1315.

Li, F., Edwards, J., & Beckman, M. E. (2009). Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers. *Journal of Phonetics, 37,* 111–124.

Macken, M. A., & Barton, D. (1980). The acquisition of the voicing contrast in English: A study of voice onset time in word initial stop consonants. *Journal of Child Language, 7,* 41–74.

Maxwell, E. M., & Weismer, G. (1982). The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics, 3,* 29–43.

Munson, B. (2004). Variability in /s/ production in children and adults: Evidence from dynamic measures of spectral mean. *Journal of Speech, Language, and Hearing Research, 47,* 58–69.

Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., & Meyer, M. K. (2010). Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of Vox Humana. *Clinical Linguistics & Phonetics, 24,* 245–260.

Newman, R. S., Clouse, S. A., & Burnham, J. L. (2001). The perceptual consequences of within-talker variability in fricative

production. *The Journal of the Acoustical Society of America, 109,* 1181–1196.

Peng, S. C., Spencer, L. J., & Tomblin, J. B. (2004). Speech intelligibility of pediatric cochlear implant recipients with 7 years of device experience. *Journal of Speech, Language, and Hearing Research, 47,* 1227–1236.

Prather, E., Hedrick, D., & Kern, C. (1975). Articulation development in children aged two to four years. *Journal of Speech and Hearing Research, 40,* 179–191.

Preston, J. L., Ramsdell, H. L., Oller, D. K., Edwards, M. L., & Tobin, S. J. (2011). Developing a weighted measure of speech sound accuracy. *Journal of Speech, Language, and Hearing Research, 54,* 1–18.

Pye, C., Wilcox, K. A., & Siren, K. A. (1988). Refining transcriptions: The significance of transcriber "errors." *Journal of Child Language, 15,* 17–37.

Reidy, P. F. (2013). An introduction to random processes for the spectral analysis of speech data. In M. E. Beckman, M. Lesho, J. Tonhauser, & T.-H. Tsui (Eds.), *Ohio State working papers in linguistics* (No. 60, pp. 67–116). Columbus, OH: The Ohio State University.

Romeo, R., Hazan, V., & Pettinato, M. (2013). Developmental and gender-related trends of intra-talker variability in consonant production. *The Journal of the Acoustical Society of America, 134,* 3781–3792.

Rvachew, S., & Jamieson, D. G. (1989). Perception of voiceless fricatives by children with a functional articulation disorder. *Journal of Speech and Hearing Disorders, 54,* 193–208.

Sander, E. (1972). When are speech sounds learned? *Journal of Speech and Hearing Disorders, 37,* 55–63.

Scobbie, J. M., Gibbon, F., Hardcastle, W. J., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. B. Broe & J. B. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon* (pp. 194–207). Cambridge, United Kingdom: Cambridge University Press.

Sharkey, S. G., & Folkins, J. W. (1985). Variability of lip and jaw movements in children and adults: Implications for the development of speech motor control. *Journal of Speech and Hearing Research, 28,* 8–15.

Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa articulation norms project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55,* 779–798.

Smith, A., & Goffman, L. (1998). Stability and patterning of speech movement sequences in children and adults. *Journal of Speech, Language, and Hearing Research, 41,* 18–30.

Smith, B. L., Kenney, M. K., & Hussain, S. (1996). A longitudinal investigation of duration and temporal variability in children's speech production. *The Journal of the Acoustical Society of America, 99,* 2344–2349.

Sovinski, R. (2011). *Perceptual validation of a robustness of contrast measure* (Unpublished master's thesis). University of Wisconsin–Madison.

Stathopoulos, E. T. (1995). Variability revisited: An acoustic, aerodynamic, and respiratory kinematic comparison of children and adults during speech. *Journal of Phonetics, 23,* 67–80.

Stoel-Gammon, C., & Dunn, C. (1985). *Normal and disordered phonology in children*. Baltimore, MD: University Park Press.

Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thomson, D. J. (1982). Spectrum estimation and harmonic analysis. *Proceedings of the IEEE, 70,* 1055–1096.

Todd, A. E., Edwards, J. R., & Litovsky, R. Y. (2011). Production of contrast between sibilant fricatives by children with cochlear implants. *The Journal of the Acoustical Society of America, 130,* 3969–3979.

Tyler, A. A., Figurski, G. R., & Langsdale, T. (1993). Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress. *Journal of Speech, Language, and Hearing Research, 36,* 746–759.

Tyler, A. A., & Saxman, J. H. (1991). Initial voicing contrast acquisition in normal and phonologically disordered children. *Applied Psycholinguistics, 12,* 453–479.

White, S. D. (2001). *Covert contrast, merger, and substitution in children's productions of /k/ and /t/* (Unpublished master's thesis). The Ohio State University, Columbus.

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

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Wilson, I., & Gick, B. (2014). Bilinguals use language-specific articulatory settings. *Journal of Speech, Language, and Hearing Research, 57,* 361–373.