Quantifying the robustness of the English sibilant fricative contrast in children

Journal: Journal of Speech, Language, and Hearing Research

Manuscript ID: JSLHR-S-14-0090

Manuscript Type: Research Article

Date Submitted by the Author: 27-Mar-2014

Complete List of Authors: Holliday, Jeffrey; Indiana University, Second Language Studies; The Ohio State University, Linguistics
Reidy, Patrick; The Ohio State University, Linguistics
Beckman, Mary; The Ohio State University, Linguistics
Edwards, Jan; UW-Madison, Communicative Disorders

Keywords: Phonological development, Acoustics, Children, Speech production
Quantifying the robustness of the English sibilant fricative contrast in children

Jeffrey J. Holliday\textsuperscript{a,b,*}, Patrick F. Reidy\textsuperscript{a}, Mary E. Beckman\textsuperscript{a}, and Jan Edwards\textsuperscript{c}

a Department of Linguistics, Ohio State University
b Department of Second Language Studies, Indiana University Bloomington
c Department of Communication Sciences and Disorders, University of Wisconsin-Madison

Emails: {jeffh, reidy, mbeckman}@ling.ohio-state.edu, jedwards2@wisc.edu

* Corresponding author

Jeffrey J. Holliday
Department of Second Language Studies
Indiana University
Memorial Hall M03
1021 E. Third St.
Bloomington, IN 47405 USA
+1 812 855 0032

Abstract
**Purpose:** Four measures of children's developing robustness of phonological contrast were compared to see how they correlated with age, with vocabulary size, and with adult listeners' "correctness" ratings.

**Method:** Word-initial sibilant fricative productions from 81 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 based on the centroid frequency measured from each fricative token. Productions from different children that were transcribed as correct 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 how the relationship between age of the child and adults’ sensitivity to different types of within-category variation.

**Key Words:** Phonological development, acoustics, children, speech production
Introduction

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 adult-like way much earlier than others) and across children (i.e. some children produce consonants in an adult-like 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., Sander, 1972; Prather, Hedrick, & Kern, 1975; Smit et al. 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/). 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 et al. 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 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 et al. 2000) and also for the word-initial lingual place contrast (Forrest et al., 1990; Gibbon, 1990; Edwards, Gibbon, & Fourakis, 1997; 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 (2009) found that adults rated productions that had been transcribed either as “correct” or as “clear substitutions” in ways that suggested sub-phonemic distinctions. Specifically, 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 (Kewley-Port & Preston,
1974; Gibbon, Dent, & Hardcastle, 1993), 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 aged 5-18 and adults, Lee, Potamianos and
Narayanan (1999) showed that intra-talker variation of segment duration and formant frequencies
decreases sharply with age, reaching adult-like levels around age 12. However, in a more recent study of
children aged 9-14, Romeo, Hazan, and Pettinato (2013) found that intra-talker variation of VOT in /p/
and spectral mean in /s/ did not reach adult-like levels even by age 14, and that there was no linear
relationship between age and /s/-/ʃ/ or /b/-/p/ between-category discriminability. Given that most
typically-developing English-acquiring children are judged to accurately produce the entire English
phonological inventory by age 9 (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 (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 (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 presence of increased intra-talker variability in children’s
speech may be a reflection of underdeveloped speech motor control (Smith & Goffman, 1998). Many
studies have offered some evidence that children’s speech can be more variable than that of adults (e.g.
Eguchi & Hirsh, 1969; Koenig, Lucero & Perlman, 2008; Lee, Potamianos & Narayanan, 1999; Munson,
2004; Romeo, Hazan & Pettinato, 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 also known that intra-talker variability can have perceptual consequences. For example, Newman, Clouse, and Burnham (2001) investigated the effect of a talker’s /s/-/ʃ/ between-category overlap and within-category dispersion on listeners’ reaction time in an identification task. They found that while listeners responded more slowly to stimuli that came from a talker who exhibited between-category overlap, relative to a talker who produced separable categories, there was less of an effect on response time when the stimuli came from two talkers who each produced separable /s/ and /ʃ/ categories, but who differed in the distance between the extremes of those categories; thus, it seemed to be the presence of between-category overlap that slowed identification. Hazan, Romeo, and Pettinato (2013) also tested the impact of intra-talker variability on perception of the /s/-/ʃ/ contrast. They used stimuli produced by children aged 9-14 in addition to adults and they compared the effects of between-category overlap to the effects of the distance between the category means (hereafter, “between-category 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. Lastly,
while they did not set out to investigate the effect of within-category dispersion, they found 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 along the spectral mean Hz scale between the maximum value for /ʃ/ and the minimum value for /s/. This method of calculating overlap does not treat linearly separable sets of categories uniformly, and instead measures different degrees of non-overlap. 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 is tested by applying it, along with three previously used measures, 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. We will therefore first describe transcribed accuracy rates and the most common error patterns in productions of initial /s/ and /ʃ/ in real words elicited from the 81 children with an age of 2;1 (years;months) to 5;9. We will then apply 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 in productions of these words elicited from 20 adults. To anticipate the results, the data patterns show that while accuracy rates are closely related to age for both fricatives, the robustness of contrast in the sibilant productions does not strictly depend on age: some of the younger children have a more robust /s/-/ʃ/ contrast than some of the older children. Lastly, we take 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 use 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 who have a less robust contrast, and this prediction was borne out.

**Experiment 1: Production**

**Methods**

**Speech materials and elicitation procedure**

The productions of word-initial /s/ and /ʃ/ are 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 (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).

--------- Insert Table 1 here ---------

The real words and nonwords were elicited from each participant in one of three pseudo-random orders, which distributed trials for each target consonant in each vocalic context evenly across blocks,

¹The transcribed recordings of the children’s real word productions are available to the public through the PhonBank archive at http://childes.talkbank.org/media/Eng-NA/paidologos/0wav/.
using a customized program that, on each trial, loaded and showed the picture and then played the audio prompt once after a 300 ms delay. There was also a replay button that the tester was instructed to use to present the audio prompt a second time when eliciting a child’s productions in case the first presentation of the audio prompt did not result in a clear recording of the target word. The most common cases were trials where the child’s repetition after the first presentation of the audio prompt was obscured by background noise, but the tester was instructed to replay the prompt also on trials where the child perseverated and said the preceding target word again as the audio prompt was played, on trials where the child said a different word, or when the child made no response or repeated the word very softly.

The entire elicitation session was recorded for subsequent transcription and acoustic analysis. This recording was made using a Marantz PMD660 flash card recorder and an AKG C5900M condenser microphone with a cardioid response. The mic 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 mic.

**Participants**

81 children participated in the recording. There were 20 children from each of four different age groups (2-, 3-, 4-, and 5-year-olds), with 10 females and 10 males per age group. The children were from middle socioeconomic status (SES) families in Columbus, OH. They were recorded at their daycare centers or preschools in a quiet room. All children had normal speech, language, and hearing, based on 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 one standard deviation 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-
nonword repetition tasks (although they were recording in a sound booth on the Ohio State University campus). The adults also had normal speech, language, and hearing, assessed by self-report.

**Transcription**

A single native speaker/phonetician transcribed the initial consonant in all of the children’s target productions. 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 91 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, etc., so that the transcribed target production was the child’s repetition in response to the second presentation of the audio prompt. In another 136 cases, the child’s response to the second presentation also could not be transcribed, so that the number of tokens analyzed for that target consonant for that child was reduced. Counts were reduced in this way for 19 of the 20 two-year-olds, 19 of the 20 three-year-olds, 17 of the 21 four-year-olds, and 14 of the 20 five-year-olds. In 75% of these cases, the reduction was by just one or two of the 15 tokens for one or both consonant targets, and in all but a handful of the other cases, the reduction was by no more than three tokens for either consonant type. The notable exceptions were one of the youngest 2-year-olds (who produced only four transcribable tokens of /s/), one of the older 3-year-olds (who produced only eight transcribable tokens of /ʃ/), and the youngest 5-year-old (who produced only nine transcribable tokens of /s/).

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 plus two more categories for “distortion” (a production not easily assigned to any consonant category symbolized in the IPA) and “deletion” (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 IPA symbol for the target phoneme itself (e.g., [s] for target /s/), the IPA symbol for a clear substitution of another phoneme of English (e.g., the weak fricative [θ] for target /s/ or the sibilant affricate [ʃ] for target /ʃ/), or the IPA symbol for a non-English consonant category (e.g. sibilant affricate [ts] for target /s/ or sibilant fricative [sʲ] or weak fricative [ç] for target /ʃ/). 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]:[sʲ]/” for a production of some anterior sibilant sound intermediate between the English phoneme [s] and the out-of-inventory sound [sʲ]. 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, so that a production that was coded as correct in the categorical decision at the first step of the transcription process could also be symbolized as such an intermediate sound in the second step of the transcription process if it was judged to more similar to a clearly correct production of the target consonant than to a prototypical example of the other sound. Thus, “[s]:[sʲ]” 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 2249 transcribable tokens, about a quarter (539) were transcribed as intermediate between two categories in this way.

We used these transcriptions in three ways in 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 use the narrow
phonetic transcriptions also to determine whether a production could be included in the spectral analyses for the quantitative measure of the “place of articulation” contrast. A token was deemed to be “analyzable” for the quantitative measure of place if the consonant was transcribed as some kind of sibilant (either a fricative or an affricate), or if the consonant was transcribed as intermediate between two sounds, both of which were sibilants of some kind.

**Fricative event tagging**

A trained phonetician marked the onset of frication and the fricative-vowel boundary for each analyzable token 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.

**Spectral estimation and centroid computation**

In the interest of comparability to previous work on the structural properties of the English voiceless sibilant contrast, our spectral estimation methods followed, as closely as possible, those of Romeo et al. (2013). 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 (.wav) file into an R programming environment using functionality from the tuneR package (Ligges, 2014). Each waveform was pre-processed by normalizing its amplitude so that it was equal to one. No pre-emphasis filter was applied to the waveform. The spectrum of either sibilant was estimated from the middle 50% of each amplitude-normalized frication waveform; thus, the number of samples from which a spectral estimate was computed varied according to the duration of the waveform.

To estimate either spectrum, the middle 50% of a waveform was extracted with a rectangular analysis window, and from this an eighth-order multitaper spectrum (MTS) (Thomson, 1982), with
parameters $K = 8$ and $NW = 4$, was computed using functionality from the `multitaper` package (Rahim and Burr, 2013). Briefly, an eighth-order MTS can be thought of as 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. The parameter $K$ determines the number of statistically independent DFTs that are averaged to yield the MTS; thus, an eighth-order MTS will always have $K = 8$. The parameter $NW$ determines the shapes of the analysis windows applied to the $K$ copies of the waveform; it is conventional to use a parameter value of $NW = K/2$ (cf. Percival & Walden, 1993). 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 0.3–20 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.

**Robustness of contrast measures**

Because results of previous studies (Newman et al., 2001; Hazan et al., 2013) 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-category 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 will be treated as having equally robust contrasts even if one child’s categories are 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 measure was calculated as follows. First, we built a mixed effects logistic regression model with the following structure using the \texttt{lme4} package (Bates et al., 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 group-wide centroid distribution centered at zero, and individual talker-level random intercepts and slopes. The model was built using fricative productions from the 81 children described above. We used all of each child’s productions that were transcribed as any kind of sibilant, rather than just those transcribed as a correct production. The number of tokens included in the model is shown in Table 2 (see also Figure 1).

\begin{table}
\centering
\caption{Number of tokens included in the model.}
\begin{tabular}{|c|c|}
\hline
Child & Tokens \\
\hline
Child 1 & 500 \\
Child 2 & 450 \\
Child 3 & 400 \\
\hline
\end{tabular}
\end{table}

The \texttt{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, based on 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 0.80, 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 also 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 (i.e. (σₚ + σᵢ)/2). Between-category distance was calculated as the difference between the mean centroid frequencies of both categories (i.e. μₚ – μᵢ). Lastly, a discriminability score, d(a), was calculated as the between-category distance divided by the square root of the mean of the centroid frequency variances of each category (i.e. (μₚ – μᵢ)/√((Varₚ + Varᵢ)/2). These three measures and %CP will hereafter be 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 is a significant and substantial effect of age, with all but one of the youngest children being transcribed as correct on fewer than 50% of their tokens and many of the oldest children being transcribed as correct on 90% or more of their tokens. There is also a small but significant effect of target consonant, such that a token of /s/ is somewhat less likely to be transcribed as correct relative to a token of /ʃ/. 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, either when the production was deemed to be incorrect at the first stage of transcription (left four columns) or deemed to be correct (right four columns). Each count in the left 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 the table shows, the most frequent error transcribed for /s/ is a “frontal misarticulation” (i.e., substitution of the voiceless weak interdental fricative [θ]) and [θ] is also the sound most commonly involved when a token of /s/ is transcribed as “borderline” correct. The most frequent error transcribed for /ʃ/ also is a “fronting” (i.e., substitution of the other voiceless sibilant fricative [s]) and [s] is also the sound most commonly involved when a token of /ʃ/ is transcribed as “borderline” correct.

Both of these “fronting” patterns are stereotypical errors for very young English-speaking children and they are errors often transcribed in children’s productions of the English sibilant fricatives by speech language pathologists when administering tests such as the Goldman-Fristoe Test of Articulation. However, only the second of these common error types 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 also 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 our methods. Moreover, the same is true of the 12 tokens transcribed as substitutions of the weak palatal fricative [ç] or affricate [kç] or as intermediate between some sibilant fricative and [ç].
Figure 2 shows the proportion of “analyzable” tokens once such productions are 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 with age and target consonant as fixed effects and random (individual child-level) intercepts. These “analyzable” tokens are then the productions that were included in the results described in the next sections. Many of the older children produced only tokens that were sibilants, whether correct or incorrect, and even the youngest children tended to produce many sibilants, particularly for /ʃ/. The exceptions were the two-year-old and one three-year-old who did not produce any analyzable /s/ tokens, and the one four-year-old who produced only one analyzable /s/ token.

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 ANOVAs with between-subjects effects of age group (child vs. adult) and gender were run. For /s/, there was a main effect of gender ($F(1,94) = 13.98, p < .001$) and a significant interaction between gender and age group ($F(1,94) = 5.44, p = .022$), but no main effect of age group ($p = .194$). Tukey HSD post-hoc tests revealed that the only significant comparisons were those between adult males and both child and adult females ($p < .001$ and $p < .002$, respectively), suggesting that the centroid of /s/ does not change significantly between child and adult female speakers. The difference between child and adult males was marginally significant ($p = .055$), which may reflect the substantial increase in lip thickness when males go through puberty (Vorperian et al., 2011) and the effect that this change has of providing a cavity anterior to the teeth even when there is no labial constriction. For /ʃ/, there was a main effect of age group ($F(1,94) =$
35.62, \( p < .001 \), 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 and Gick, in press), and then suggest that in subsequent development, the centroid of /ʃ/ decreases with age for both genders, but the centroid of /s/ changes with age only for men. This interaction results in adult men having /s/ and /ʃ/ categories that are closer together than those of adult women.

The output of the mixed effects logistic regression model described above indicates that centroid frequency at the midpoint of the turbulent interval can reliably separate children’s /s/ and /ʃ/ categories (\( \beta = 0.00147, z = 10.29, p < .001 \)).

--- Insert Figure 3 here ---

**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 adults are reported in Table 4. Because the calculation of the robustness measures requires there be a sufficient number of tokens from each category, the three children who produced zero or only one analyzable /s/ token were excluded from further analysis.

The first column of panels in Figure 4 shows that %CP, between-category distance, and \( d(a) \) generally increase with age among children. All three of these measures were significantly correlated with age (all comparisons \( p < .001 \)). The same three measures were found to be significantly correlated with children’s receptive vocabulary raw scores (all comparisons \( p < .002 \)), shown in the second column of Figure 4, but none of the measures were correlated with the expressive vocabulary raw scores shown in
the third column. Within-category dispersion, on the other hand, decreased with age ($p < .001$), although the strength of its relationship with age among the children ($R^2 = .16$) was lower than that of %CP ($R^2 = .35$), between-category distance ($R^2 = .28$), or $d(a)$ ($R^2 = .32$). There was no significant relationship between within-category dispersion and either receptive or expressive vocabulary raw scores.

Because age and receptive vocabulary were so highly correlated with each other (Pearson’s $r = 0.80$), and the robustness measures were more highly correlated with age than receptive vocabulary anyway, we focus our remaining analyses on only age and gender differences. This focus also allows us to compare the children with the adults, as the adults do not have vocabulary scores. Expressive vocabulary will also be excluded from further analyses, as it was not shown to be predictive of any of the robustness measures.

We ran separate two-way ANOVAs with between-subjects factors of age group (2-, 3-, 4-, 5-year-olds, and adults) and gender for %CP, within-category dispersion, between-category distance, and $d(a)$. For %CP, there was a main effect of age group ($F(4,88) = 24.22$, $p < .001$) but no main effect of gender ($p = .183$). Tukey 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 amongst the older children themselves. These results suggest a gradual increase with age, and that by age 5 the degree of overlap in children’s /s/-/ʃ/ contrast may be comparable to that of adults.

---------- Insert Figure 4 here ----------

---------- Insert Table 4 here ----------

For within-category dispersion, there was a main effect of age group ($F(4,88) = 18.58$, $p < .001$)
but no main effect of gender ($p = .499$). Tukey 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 children still have greater levels of within-category dispersion than adults. Among the children, within-category dispersion significantly differed between only non-consecutive age groups, suggesting that dispersion decreases quite gradually with age.

For between-category distance, there were main effects of both age group ($F(4,88) = 16.31, p < .001$) and gender ($F(1,88) = 15.10, p < .001$). Tukey HSD post-hoc tests revealed significant differences between adults and all children’s age groups except for 5-year-olds. Additionally, the 3-, 4-, and 5-year-olds were not significantly different from each other, indicating that changes in between-category distance may occur between 2 and 3 years of age, after which change is gradual and may reach adult levels by age 5. This pattern of showing the most change between 2 and 3 years but less year-to-year change afterwards was the same as what was found with %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 ($F(4,88) = 38.27, p < .001$) and gender ($F(1,88) = 9.92, p = .002$). Tukey HSD post-hoc tests revealed significant differences ($p < .001$) 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-category 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 good predictor of fricative category for both children and adults. It was also found that the difference in mean /s/ centroid frequency does not significantly differ between children and adults, although there was a marginally significant gender interaction reflecting the slight difference between child and adult males. The mean /ʃ/ centroid frequency did decrease significantly with age for both genders, however. In their
study of children aged 9-14 and adults, Romeo et al. (2013) found the opposite trend, in which the
centroid frequency for /s/ but not /ʃ/ differed significantly between children and adults.

The results of the robustness of contrast measures suggest that while %CP, within-category
dispersion, between-category distance, and \( d(a) \) are 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-category distance, all children’s age
groups significantly differed from adults according to within-category dispersion and \( d(a) \). Although
Romeo et al. (2013) did not calculate %CP, they did find that children aged 9-14 had greater between-
category distance than adults, with this effect driven especially by a sudden jump in between-category
distance in 11- to -12-year-old girls. Taken together with the current study, these results indicate that
although between-category distance is already at adult-like levels by age 5, it continues to increase for
several more years until decreasing back down to adult-like levels during the teenage years. This age
trend in between-category distance may also explain the seemingly disparate trends in centroid frequency
discussed above: although 5-year-olds’ mean /s/ centroid frequency did not differ significantly from
adults, the fact that Romeo et al. (2013) found that the children in their study had a higher mean /s/
centroid frequency than adults agrees with those children also having greater between-category distance
than adults.

The highly similar trends between %CP and between-category distance suggest that the two
measures may be strongly correlated. Pearson’s \( r \) indicates that this is the case \((r = 0.78, df = 76, p
< .001)\). Although %CP is a measure of category overlap that does not use between-category 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 adult-like levels by age 5. Because Romeo et al.
(2013) found the same result for children aged 9-14, 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. 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-category 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 vocabulary size, the relationship with age among children aged 2-5 was strongest for %CP and weakest for within-category dispersion. This result is seemingly at odds with the conclusions of Hazan et al. (2013), who concluded 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 impact perception more than the other measures. Furthermore, because the relationship between these measures and age is not always linear (e.g. between-category distance), it could be the case that the perception of younger children’s fricatives are influenced by different factors. Accordingly, 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

Methods

Stimuli

Stimuli for the perception experiment were chosen as follows. First, we selected stimuli 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 intermediate stimuli only if the other phoneme that the production 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. Finally, insofar as possible, we tried to choose an equal number of productions from children at each age, equal numbers of productions from both boys and girls, and equal numbers of /s/ and /ʃ/ productions. Because there were relatively fewer correct productions from the younger children, we included nonword productions as well as real word production. We did this for both the younger and older children so that the distribution of stimuli from real words and nonwords would be similar between the younger and older groups of children. Each stimulus item included the initial fricative and a 150 ms vocalic portion. All stimulus items were normalized for amplitude. The stimuli for the perception experiment are shown in Table 5.

------------ Insert Table 5 here ------------

Participants

The participants in the perception study were 20 (7 male, 13 female) young adults 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, based on self report.

Procedure

Participants listened to 6 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
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.

**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 were discarded. The mouse click locations will hereafter be referred to as the “goodness ratings” and will be reported as transformed logit values. The mean and median goodness ratings were 0.14 and 0.25 for the “s”-block, and 0.13 and 0.23 for the “sh”-block, respectively, indicating that neither block was perceived as overall better than the other.

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. Based on the $R^2$ value of each block we can see that age was a better predictor of perceived goodness for /ʃ/ ($R^2 = .18, p = .009$) than for /s/ ($R^2 = .11, p = .029$).

We turn next to the relationship between perceived goodness and the four measures of robustness of contrast under investigation: %CP, within-category dispersion, between-category 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 %CP. With an $R^2$ of .38 ($p < .001$) for the “s”-block and .43 ($p < .001$) for the “sh”-block, %CP appears to be a substantially better predictor of perceived goodness than age-in-months. The second row of panels shows the relationship between perceived goodness and within-category dispersion, which was not statistically significant for either the “s”-block ($R^2 = .03, p = .15$), or the “sh”-block ($R^2 = .01, p = .42$).

Between-category distance, shown in the third row of Figure 6, seems to be a good predictor of perceived goodness for both the “s”-block ($R^2 = 0.35, p < .001$) and the “sh”-block ($R^2 = 0.20, p = .005$), although the $R^2$ values for this measure were substantially lower in the “sh”-block than those for %CP. Lastly, in the bottom row of panels, $d(a)$ appears to be a better predictor for the “s”-block ($R^2 = .29, p < .001$) than for the “sh”-block ($R^2 = .13, p = .022$), although both relationships are significant. The less predictive power of $d(a)$ in the “sh”-block is probably due to the influence of within-category dispersion, upon which the calculation of $d(a)$ is partially based.

**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-category distance, and discriminability ($d(a)$). In summary, we found that age was modestly correlated with perceived goodness, and that this relationship was stronger in the “sh”-block than in the “s”-block, perhaps reflecting the relationship between centroid frequency and age for /ʃ/ but not for /ʃ/. Among the four measures of robustness of contrast, %CP was the most strongly correlated with perceived goodness in both the “s”-block and “sh”-block. Between-category distance was also moderately correlated with perceived goodness, but this relationship was less strong in the “sh”-block. Within-category dispersion was not correlated with perceived goodness at all, and the discriminability
score \( d(a) \) was less predictive of perceived goodness than \%CP or between-category distance, likely due to its calculation being partially based on within-category dispersion.

Although perceived goodness was better predicted by \%CP than by between-category distance, because these two measures performed similarly – and are 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-category distance has the advantage of being simpler to calculate, and is also not bounded in the way that \%CP is bounded between 0 and 1. 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-category 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 aged 9 to 14 in Hazan et al. (2013), the productions from children with greater between-category 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-category 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-category 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, but because the talker in their study who had greater overlap also had greater dispersion, it could have been the increased level of dispersion that was 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 intra-talker variability varies both with age and across different
phonological contrasts.

**General Discussion**

In this paper, we showed that while 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 adult listeners’ goodness judgments of the children’s fricative productions.

The findings presented here disagree with 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. While 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 Hazan et al. (2013)’s study were produced by older children (9- to 14-year-olds) whose level of between-category distance was greater than even that of adults, while the stimuli in the current study were produced by 2- to 5-year-olds. Although we found %CP and between-category distance to be well correlated with perceived goodness, it could be that once between-category distance and %CP reach adult-like levels they impact perception less, leaving room for within-category dispersion to play a bigger role. An important difference between %CP and between-category 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, and perhaps for this reason 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) point 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 et al., 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). Alternatively, 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-category distance and less overlap.

The relationship 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-category 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? Unfortunately, there is almost no research that addresses this question. In a perception study similar to the one described in this paper, Bernstein, Todd, & Edwards (2013) found that tokens of /s/ produced by children with CIs that were transcribed as being correct were rated as less good than productions by children with NH 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, 2006). These findings suggest that, at least for children who have difficulty perceiving a contrast, it may improve their intelligibility to continue to work on consonant contrasts even after productions are perceived as correct. Furthermore, it may be useful to include additional assessments of correct production over and above the categorical transcription judgment of correct vs. incorrect. These could include VAS rating scales with naive listeners or acoustic/psychoacoustic measures.

To conclude, this study found that the robustness of contrast between /s/ and /ʃ/, as measured by %CP and between-category distance gradually increased from 2 to 5 years of age. Differences were also observed between 5-year-olds and adult speakers. Differences in robustness of contrast were also reflected in adults’ 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.

Acknowledgements

This research was supported by NIDCD Grant R01-02932 to Jan Edwards, Mary E. Beckman, and Benjamin Munson, NSF Grant No. BCS-0729140 to Jan Edwards, NSF Grant No. BCS-0729306 to Mary E. Beckman, and NICHD Grant P30-HD03352 to the Waisman Center. We thank Laura Slocum for making the stimuli for the production experiment and both Laura Slocum and Anne Hoffmann for recruiting and recording the children and adults for this experiment; Fangfang Li, Channelle Mays, Oxana Skorniakova, Asimina Syrika, Julie Johnson, and Jeffrey Kallay 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 MS thesis). We also thank the children who participated in study 1 and their parents, as well as the participants of study 2. The idea of using slopes and prediction accuracy rates from logistic regression models was presented earlier at INTERSPEECH 2010 in a paper by the first and third authors and Channelle Mays, and we thank the audience at INTERSPEECH for useful comments that helped to shape
the further development of the idea in the current paper.

References

Bates, D., Maechler, M., & Bolker, B. (2013). lme4: Linear mixed-effects models using S4 classes. R package version 0.999999-2. http://CRAN.R-project.org/package=lme4


Table 1. Real word and nonword stimuli used in the picture-prompted word repetition task. 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 different following “frame” portions, which were rotated among initial CV targets across the three lists, so that each participant only heard one disyllabic and two trisyllabic nonword stimuli in each vowel context.

<table>
<thead>
<tr>
<th>context</th>
<th>/ʃ/ words</th>
<th>/ʃ/ nonwords</th>
<th>/s/ words</th>
<th>/s/ nonwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>high front vowel</td>
<td>shield, sheep, ship</td>
<td>seal, seashore, sister</td>
<td>sibθ, sign, siven, sivenop, sibilaθ, sifamut, sivałut, sivałkloθ, sivałfæf</td>
<td></td>
</tr>
<tr>
<td>high back vowel</td>
<td>shoe, chute, sugar</td>
<td>soup, super, suitcase</td>
<td>sugm, suvaŋ, sibθ, sibilaθ, sifamut, sugenop, sivałkloθ, suvałfæf, suvałblut</td>
<td></td>
</tr>
<tr>
<td>mid front vowel</td>
<td>shape, shell, shepherd</td>
<td>safe, same, seven</td>
<td>sevam, sebθ, segm, sifamut, segenop, sibilaθ, sevałfæf, sevałblut, sevałkloθ</td>
<td></td>
</tr>
<tr>
<td>mid or low back vowel</td>
<td>show, shoulder, shore, shovel, shark, shop</td>
<td>soak, soldier, sodas, sun, soccer, sauce</td>
<td>sap'ôn, sapim, saktif, saktigip, sažtvam, saganut, sakəmət, saptigək, sanałkæd</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Number of tokens per age group used in the logistic regression model.

<table>
<thead>
<tr>
<th>Age in years</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of tokens</td>
<td>353</td>
<td>414</td>
<td>500</td>
<td>521</td>
<td>1788</td>
</tr>
</tbody>
</table>
Table 3. Counts of most commonly transcribed sounds for correct and incorrect /s/ and /ʃ/. The counts of the 710 tokens of /s/ and the 812 tokens of /ʃ/ that were judged to be correct (left panels) and the most commonly transcribed sounds for the 400 tokens of /s/ and the 327 tokens of /ʃ/ that were judged to be incorrect (right panels). The numbers in parentheses in rows 2-4 in the left two columns are proportions relative to the number of intermediate transcriptions, whereas those in the right two columns are proportions relative to the number of errors.

<table>
<thead>
<tr>
<th></th>
<th>/s/ correct</th>
<th>/ʃ/ correct</th>
<th>/s/ incorrect</th>
<th>/ʃ/ incorrect</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s] alone</td>
<td>578</td>
<td>722</td>
<td>119 (0.36)</td>
<td>119 (0.36)</td>
</tr>
<tr>
<td>[s]:[θ]</td>
<td>65 (0.49)</td>
<td>42 (0.47)</td>
<td>79 (0.20)</td>
<td>57 (0.17)</td>
</tr>
<tr>
<td>[s]:[ʃ]</td>
<td>24 (0.18)</td>
<td>16 (0.18)</td>
<td>63 (0.16)</td>
<td>23 (0.07)</td>
</tr>
<tr>
<td>[s]:other</td>
<td>91 (0.69)</td>
<td>32 (0.36)</td>
<td>173 (0.43)</td>
<td>128 (0.39)</td>
</tr>
</tbody>
</table>
Table 4. Mean and standard deviation of the four robustness of contrast measures.

<table>
<thead>
<tr>
<th></th>
<th>2-year olds</th>
<th>3-year olds</th>
<th>4-year olds</th>
<th>5-year olds</th>
<th>Adults</th>
</tr>
</thead>
<tbody>
<tr>
<td>%CP</td>
<td>0.67 (0.14)</td>
<td>0.85 (0.13)</td>
<td>0.88 (0.13)</td>
<td>0.93 (0.10)</td>
<td>1.00 (0)</td>
</tr>
<tr>
<td>Dispersion</td>
<td>1367 (482)</td>
<td>1319 (338)</td>
<td>1074 (314)</td>
<td>984 (232)</td>
<td>577 (158)</td>
</tr>
<tr>
<td>Distance</td>
<td>350 (1273)</td>
<td>2416 (1753)</td>
<td>2489 (1614)</td>
<td>3162 (1372)</td>
<td>3588 (1109)</td>
</tr>
<tr>
<td>(d(a))</td>
<td>0.30 (0.95)</td>
<td>1.89 (1.68)</td>
<td>2.40 (1.75)</td>
<td>3.03 (1.23)</td>
<td>5.94 (1.86)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2-yo females</th>
<th>3-yo females</th>
<th>4-yo females</th>
<th>5-yo females</th>
<th>Adult females</th>
</tr>
</thead>
<tbody>
<tr>
<td>%CP</td>
<td>0.68 (0.14)</td>
<td>0.84 (0.15)</td>
<td>0.91 (0.06)</td>
<td>0.97 (0.03)</td>
<td>1.00 (0)</td>
</tr>
<tr>
<td>Dispersion</td>
<td>1302 (346)</td>
<td>1317 (350)</td>
<td>1088 (337)</td>
<td>1074 (258)</td>
<td>627 (201)</td>
</tr>
<tr>
<td>Distance</td>
<td>715 (1014)</td>
<td>2784 (2051)</td>
<td>2896 (1731)</td>
<td>3809 (1041)</td>
<td>4406 (914)</td>
</tr>
<tr>
<td>(d(a))</td>
<td>0.57 (0.85)</td>
<td>2.28 (2.08)</td>
<td>2.82 (1.97)</td>
<td>3.39 (0.99)</td>
<td>6.83 (1.88)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>2-yo males</th>
<th>3-yo males</th>
<th>4-yo males</th>
<th>5-yo males</th>
<th>Adult males</th>
</tr>
</thead>
<tbody>
<tr>
<td>%CP</td>
<td>0.66 (0.15)</td>
<td>0.85 (0.09)</td>
<td>0.85 (0.17)</td>
<td>0.89 (0.14)</td>
<td>1.00 (0)</td>
</tr>
<tr>
<td>Dispersion</td>
<td>1413 (573)</td>
<td>1323 (345)</td>
<td>1059 (306)</td>
<td>874 (140)</td>
<td>528 (84)</td>
</tr>
<tr>
<td>Distance</td>
<td>84 (1419)</td>
<td>2007 (1351)</td>
<td>2083 (1462)</td>
<td>2372 (1357)</td>
<td>2768 (525)</td>
</tr>
<tr>
<td>(d(a))</td>
<td>0.11 (1.00)</td>
<td>1.45 (1.04)</td>
<td>1.99 (1.48)</td>
<td>2.58 (1.41)</td>
<td>5.06 (1.42)</td>
</tr>
</tbody>
</table>
Table 5. Number of stimuli in the perception experiment.

<table>
<thead>
<tr>
<th></th>
<th>2-year-olds</th>
<th>3-year-olds</th>
<th>4-year-olds</th>
<th>5-year-olds</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>/s/ stimuli</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real word</td>
<td>34</td>
<td>31</td>
<td>35</td>
<td>36</td>
<td>136</td>
</tr>
<tr>
<td>Non-word</td>
<td>9</td>
<td>16</td>
<td>13</td>
<td>12</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>43</td>
<td>47</td>
<td>48</td>
<td>48</td>
<td>186</td>
</tr>
<tr>
<td><strong>/ʃ/ stimuli</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Real word</td>
<td>39</td>
<td>47</td>
<td>43</td>
<td>43</td>
<td>172</td>
</tr>
<tr>
<td>Non-word</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>5</td>
<td>18</td>
</tr>
<tr>
<td>Total</td>
<td>41</td>
<td>53</td>
<td>48</td>
<td>48</td>
<td>190</td>
</tr>
</tbody>
</table>
Figure captions

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.

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

Figure 3. Distribution of centroid frequency across fricative target, age group, and gender. Each boxplot shows the distribution of centroid for /s/ and /ʃ/ for females (left) and males (right) for each age group.

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.

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.
Figure 5
169x101mm (72 x 72 DPI)