Effects of age and vocabulary size on production accuracy and acoustic differentiation of young children's sibilant fricatives

Hannele Nicholson[1], Benjamin Munson[1], Patrick Reidy[2], and Jan Edwards[3]
[1]Department of Speech-Language-Hearing Sciences, University of Minnesota, Minneapolis, [2]Department of Linguistics, Ohio State University, Columbus, [3]Department of Communication Sciences and Disorders, University of Wisconsin, Madison

purplehannele@gmail.com, munso005@umn.edu, patrick.francis.reidy@gmail.com, jan.edwards@wisc.edu

ABSTRACT

This paper reports preliminary results of a study of the development of sibilant fricatives /s/ and /ʃ/ in young children. Results show that transcribed accuracy increased over the age range studied (28 to 39 months) and that children with larger vocabularies produced fricatives more accurately than ones with smaller vocabularies. Spectral characteristics were measured for productions transcribed to be sibilant. The separation between the spectra of target /s/ and target /ʃ/ also increased over the age range, and was greater in children with larger-sized vocabularies.

Keywords: Children, Fricative, Acoustic Analysis, Vocabulary Size, Sibilant

1. INTRODUCTION

Young children‘s productions of speech sounds differ greatly from those of adults. In early acquisition, toddlers’ productions deviate so far from adults that different phonetic symbols may be used to denote productions [1, 2]. In later development, differences between adults and children are revealed in other measures, such as children’s greater token-to-token durational, spectral, and kinematic variability [3].

This paper reports on transcriptions and spectral measures of the sibilant fricatives /s/ and /ʃ/ in children aged 28 to 39 months acquiring English monolingually. Productions of these sounds in the initial position of familiar real words were elicited as part of a larger project examining relationships among measures of speech production, vocabulary size, and several other aspects of phonological development and word learning. The /s/-/ʃ/ contrast was targeted for measures of speech production (and perception) because the development of fully adult-like patterns for production of this contrast begins relatively late and is very protracted [4]. There is also evidence that children’s acoustic differentiation of these sounds has a protracted time-course. [5] showed that children’s early productions of target /s/ and /ʃ/ are acoustically undifferentiated, and that over the 2- to 5-year-old age range the difference in centroid frequency (an acoustic measure of the spectral center-of-gravity along the frequency scale) between /s/ and /ʃ/ increases. [6] showed that the acoustic differentiation between target /s/ and /ʃ/ in children as old as 13 years was not yet adult-like.

The purpose of this report is threefold. The first is to describe the transcribed accuracy of /s/ and /ʃ/ productions by a sample of young children that vary in their vocabulary size, and to examine relationships among age, vocabulary size, and accuracy. The second is to describe the development of a measure to quantify the degree of contrast between /s/ and /ʃ/. The third is to describe relationships among age, vocabulary size, and acoustic differentiation between /s/ and /ʃ/.

2. METHODS

2.1. Participants

The participants were 57 children (26 boys, 31 girls) aged 28 to 39 months. The children were recruited as part of a longitudinal study of phonological and lexical development. All children passed a pure-tone hearing screening and parents reported they were native, monolingual speakers of English. As part of their participation in the project, children completed a variety of speech production and speech perception tasks, as well as standardized measures of speech and language. One of these measures was the *Expressive Vocabulary Test-2* [7]. The 57 children in this sample had scores that varied widely on this measure. Growth score values were derived from the information in the technical manual and used instead of raw scores as a measure of expressive vocabulary size, as growth score values are linear and raw scores are not.

One of the goals of the longitudinal study was to examine children who varied in socioeconomic status (SES). In this sample, we used maternal education as a proxy for SES. Children in the lowest maternal education categories (less than a high school diploma, a high school diploma, or a GED) included both children who speak African American English (AAE) and those who spoke Mainstream

American English (MAE) at home. We established a rubric (based on AAE morphological and phonological features) for determining whether families spoke AAE or MAE during the pre-test phone interview and we then confirmed the home dialect when families arrived for their first testing session. Stimuli were presented in the home dialect.

2.2. Stimuli and Elicitation Procedure

To elicit productions, a picture-prompted word repetition task was used, in which children simultaneously saw a picture of a target word and heard a recorded prompt of the target word. The target words were chosen because they were likely to be familiar to children in this age range. This was accomplished by choosing only words that were reported to be produced by at least 80% of 30-month-old children in the normative sample for the *McArthur-Bates Communicative Development Inventories* [8]. Words were selected so that all quadrants of the vowel quadrilateral would be represented equally, and so that four instances of each target sibilant adjacent to a vowel from each vowel quadrant would be elicited. This resulted in some words being presented more than once, but with a different auditory prompt. Color photographs were selected from databases of stock images to be good exemplars of the targets.

Children were tested at one of two sites. In both sites, children were in a quiet, sound-treated room. During the task, the child was seated in front of a Planar HDMI PXL2430MW 24-inch touchscreen monitor approximately 60 centimeters and was requested to repeat 95 test items that he/she heard over Klipsch BT77 speakers or Audix PH5, depending on the site. Children's productions were recorded to a steady-state recorder.

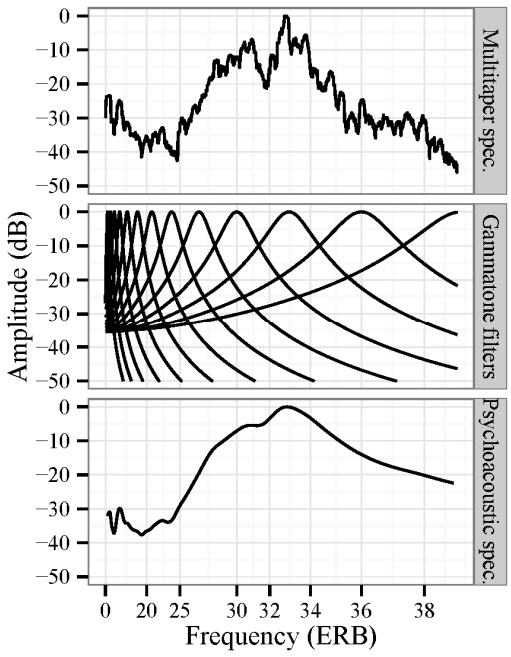
2.3. Annotation

A multi-step procedure was used to annotate the children's productions. The first step involved identifying and marking off an interval of time that included the target (i.e., first useable) production for an individual experimental trial. The second step, *turbulence tagging*, involved three stages. In the first stage, the researcher identified the production as one of a broad set of manner of articulation categories: sibilant fricative, sibilant affricate, non-sibilant fricative, non-sibilant plosive, or other. In the second stage, the researcher identified the onset and offset of the interval of turbulence for those sounds identified at the first stage as being sibilant fricatives or sibilant affricates. The third stage was *fine-grained place tagging*, where a researcher phonetically transcribed the sibilant fricatives as either [s], [ʃ], a sound intermediate between these but closer to [s], denoted [s:ʃ], or an intermediate sound closer to [ʃ], denoted [ʃ:s]. Previous research supports the use of these transcription categories, as naïve listeners' fricative-goodness judgments differ among sounds transcribed as [s], [ʃ], [s:ʃ], and [ʃ:s] [9].

2.4. Acoustic Analysis

The acoustic analysis examined the peak frequency, along the ERB-scale [10], for the productions tagged as either sibilant fricatives or sibilant affricates. To compute this measure, the middle 40 ms of frication noise of each production was extracted with a rectangular window, and then its spectrum was estimated with the multitaper spectrum [11, shown in the top panel of Figure 1], using the parameters K = 8, NW = 4 [cf. 12, 6]. This spectral estimate was then passed through a gammatone filterbank [13, 14], shown in the middle panel of Figure 1. The filterbank models the differential frequency selectivity of the auditory system. The filterbank outputs a psychoacoustic spectrum that relates the amount of excitation in a gammatone filter to its center frequency along the ERB-scale (shown in the bottom panel of Figure 1). The ERB number with the greatest level of excitation, henceforth peakERB, was used as a summary psychoacoustic measure for that token.

Figure 1: Schematic showing the calculation of the peakERB value



3. RESULTS

3.1. Transcribed Accuracy

The results of the first-stage turbulence tagging are shown in Table 1.

Table 1: Distribution of transcribed production manner categories for each target consonant (%).

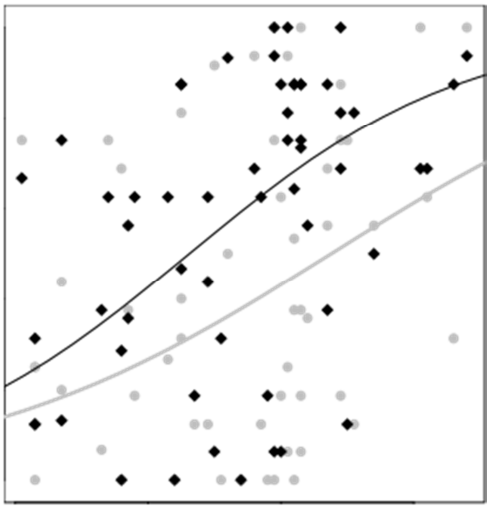
transcribed category	s	ʃ
Sibilant fricative	70	71
Sibilant affricate	8	13
Non-sibilant fricative	7	6
Non-sibilant plosive	11	6
Other	4	4

A large proportion of the tokens were transcribed as sibilant. There were some asymmetries: /s/ was more likely than /ʃ/ to be transcribed as a stop, and /ʃ/ was more likely than /s/ to be transcribed as an affricate. A logit mixed effects regression with random intercepts for talkers and a random slope for the influence of target consonant (/s/ or /ʃ/) on individual subjects examined whether the likelihood of whether a token was transcribed as sibilant (i.e., as one of the first two categories in Table 1) or non-sibilant (i.e., as one of the last three categories). A model with effects of age and consonant type but no interaction fit the data better than a simpler model with age alone ($|^2_{[df=1]}$=4.553, p=0.033). There were more sibilant productions for target /ʃ/, and targets were more likely to be sibilant with increasing age.

A second analysis examined accuracy as determined by the fine-grained place transcriptions. This analysis included only productions tagged as sibilant fricatives in the first analysis (70% of /s/ productions and 71% of /ʃ/ productions). For this analysis, /s/ targets were coded as accurate if they were transcribed as [s] or [s:ʃ] and inaccurate otherwise. /ʃ/ targets were transcribed as accurate if they were transcribed as [ʃ] or [ʃ:s]. Binary accuracy judgments were the dependent measure in a logit mixed-effects regression, with random intercepts and slopes for individual subjects. A model including age had only a marginally better fit than a model with random-effects structure only ($|^2_{[df=1]}$= 3.359, p=0.067). A model that included a factor coding consonant target had a significantly better fit than the model with age alone ($|^2_{[df=1]}$= 9.210, p=0.002). A model with an interaction did not improve model fit. Despite the lack of a statistically significant interaction, visual inspection suggested that the accuracy of /ʃ/ increases more than does the accuracy of /s/ over the age range studied, making the difference in accuracy between these two targets smaller for the older children than for the younger ones.

The next analysis examined whether there was a significant interaction between EVT-2 GSV and consonant type. In that analysis, a model with EVT-2 GSV showed a significant improvement in fit over a model with only random effects structure, and one with target consonant and EVT-2 GSV showed a further significant improvement in fit ($|^2_{[df=1]}$= 8.081, p=0.004). A model with an interaction did not improve fit beyond this; however, visual inspection of Figure 2 shows that the accuracy for both /s/ and /ʃ/ is greater for children with larger vocabularies than for children with smaller ones, and that there is a bigger difference between them for children with larger vocabularies.

Figure 2: Transcribed accuracy for /s/ and /ʃ/ targets by EVT-2 GSV.



3.2. Spectral Characteristics

The next analysis of the acoustic data examined the peakERB values for /s/ and /ʃ/ targets that had been coded as sibilant fricatives or sibilant affricates. Figure 3 shows these values as a function of age and consonant target. The black triangle and grey circles in the background are the median peakERB values for the /s/ and /ʃ/ targets. There are fewer than 57 data points for each type because one child had no sibilant productions of /s/ and two children had no sibilant productions of /ʃ/. PeakERB for /s/ was constant over the age range studied, while peakERB for /ʃ/ decreased. This is consistent with earlier studies of /s/ and /ʃ/ centroids [5].

Figure 3: Frequency of highest-amplitude peak (ERB number) for the /s/ and /ʃ/ targets by age.

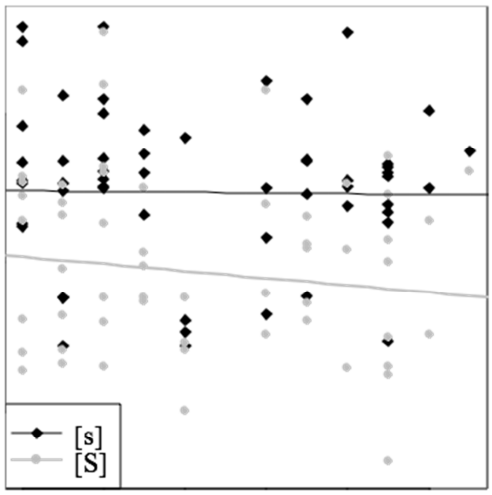
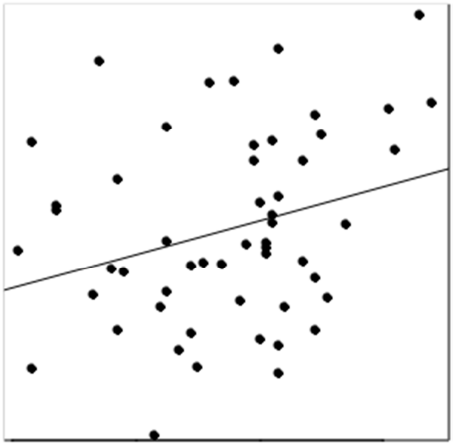


Figure 4: Individual-level slopes from the robustness of contrast analysis, plotted against the EVT-2 GS



3.3. Spectral Contrast

The next analysis of the spectral data examined the degree of acoustic contrast between the productions of /s/ and /ʃ/ target words that had been transcribed as sibilant. A logit mixed-effects regression was calculated in which consonant target (/s/ or /S/) was the dependent variable and peakERB was the independent variable. The model included random intercepts for talkers, as well as random slopes for the effect of peakERB on consonant type for individual talkers. A talker's random slope was used as a proxy measure of degree of contrast: larger slopes indicate a greater degree of separation between the peakERB values of the two targets, i.e., a greater degree of contrast.

A series of statistical analyses examined the predictors of these individual-level slopes. The first of these was a simple linear regression predicting individual-level slopes from age in months. Age was not a significant predictor. The second model was a simple regression predicting individual-level slopes from EVT-2 GSV. In this model, EVT-2 GSV was a significant predictor, albeit with a small effect size ($F[1,51] = 4.469$, $p = 0.039$, R^2=8%).

The effect of vocabulary size on the robustness of contrast is shown in Figure 4. Children with higher EVT-2 GSV values had a greater degree of contrast between /s/ and /ʃ/ targets than did children with smaller-sized vocabularies.

4. DISCUSSION

The findings in this report show that the English sibilant fricatives /s/ and /ʃ/ change even over the relatively narrow age range of 28 to 39 months. Over this span, a larger proportion of target /s/ and /ʃ/ productions are transcribed to be sibilant. More /s/ and /ʃ/ targets are transcribed to be accurate over the age range studied, though this effect only approached statistical significance. Moreover, the centroid frequency of sibilant fricatives in /ʃ/-initial words decreases significantly over this age range, resulting in an increase in the difference between /s/ and /ʃ/ centroids. Put differently, the acoustic contrast between /s/ and /ʃ/ becomes more robust over this age range.

This study found reliable effects of vocabulary size on characteristics of productions. This is true both of transcribed accuracy, as in Figure 2, and of acoustic distinctiveness, as in Figure 4. To the best of our knowledge, this is the first demonstration of an association between vocabulary size and phonetic differentiation in this age range. The finding that lexicon size is related to fricative distinctiveness invites further analyses of the specific mechanisms that drive this effect (including other measures from these same children) and how these relationships change as language develops.

7. REFERENCES

[1] Kent, R. & Forner, L. 1980. Speech segment duration in sentence recitation by children and adults. *J of Phon,*157-168.

[2] Smith, B. 1978. Temporal aspects of English speech production: A developmental perspective. *J. Phon.* 6(1). 37-67.

[3] Lee, S., Potaminos, A. Narayanan, S. 1999. Acoustic of children's speech: developmental changes of temporal and spectral parameters. *J of Acoust Soc.Am,* 103(5). 1455-1468.

[4] Smit, A.B.., Hand, L., Freilinger, J., Bernthal, J., Bird, A. 1990. The Iowa Articulation Norms Project and Its Nebraska Replication. *J. Speech, Hear. Res.*, 779-798.

[5] Li, F. 2012. Language-Specific Developmental Differences in Speech Production: A Cross-Language Acoustic Study. *Child Development, 83(4),* 1303-1315.

[6] Romeo, R., Hazan, V., Pettinato, M. 2013. Developmental and gender-related trends of intra-talker variability in consonant production. *J. Acoust. Soc. Am.,* 134(5). 3781-3792.

[7] Williams, K. 2004. *Expressive Vocabulary Test, Second edition.* Circle Pines, MN: AGS Publishing.

[8] Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. 2007. *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual* (2nd ed.). Baltimore, MD: Brookes.

[9] Stoel-Gammon, C. 2001. Transcribing the speech of young children. *Topics Lang. Dis.*21(4). 12-21.

[10] Moore B. C. J., Glasberg B. R., and Baer T. 1997. A model for the prediction of thresholds, loudness, and partial loudness. *J. Audio Eng. Soc*. 45, 224–237.

[11]Thomson, D.J. 1982. Spectrum estimation and harmonic analysis. *Proc. IEEE.* 70. 1055-1096.

[12] Koenig, L., Shadle, C., Preston, J.L., Mooshammer, C.R. 2013. Toward improved spectral measures of /s/: Results from adolescents. *Jour. Speech, Lang. Hear Res.* 56. 1175-1189.

[13] Glasberg, B., Moore, B.C.J. 1990. Derivation of auditory filter shapes from notched-noise data. *Hear. Res,* 47. 103-138.

[14] Patterson, R.D. 1976. Auditory filter shapes derived with noise stimuli. *J. Acoust. Soc. Am.*59. 640-654.

[15] Holliday, J., Reidy, P., Beckman, M., Edwards, J 2015. Quantifying the robustness of the English sibilant fricative contrast in children. *Jour. Speech, Lang. Hear Res.* doi:10.1044/2015_JSLHR-S-14-0090