ISSP 2024 - 13th International Seminar on Speech Production
13-17 May 2024, Autrans, France

Examining Speech Perception of Non-Errored Pronunciations in Children with Speech Sound Disorders

Elaine R. Hitchcock[1], Laura L. Koenig [2,3]

[1]*Montclair State University, USA*
[2]*Adelphi University, USA*
[3]*Haskins Laboratories, USA*

hitchcocke@montclair.edu, lkoenig@adelphi.edu, laura.koenig@yale.edu

Abstract

Do children with speech-sound disorders (SSDs) also differ in their speech perception? Past results suggest that perceptual difficulties are limited to sounds produced in error. Here, we assessed labeling accuracy and reaction times [RTs] in children with SSD (without voicing errors) and typical-developing [TD] peers. Stimuli were words 'boo, Pooh, doe, toe' produced by TD 2-year-olds, with VOTs that were "appropriate" (expected for the target) or inappropriate. Listener judgments were considered accurate if they matched the child's target. Results showed high listener accuracy for appropriate VOTs with no group differences. For inappropriate VOTs, children with SSD showed higher accuracy than TD, reaching significance for one comparison. RTs were faster for accurate labeling in both groups and were overall shorter children with SSD than TD peers, suggesting that children with SSD may demonstrate some differences in speech perception behavior, even for sounds not produced in error.

Keywords: speech perception, speech sound disorders, reaction time

1. Introduction

Previous studies assessing speech perception in children with speech sound disorder (SSD) suggest a) inconsistent, if any, differences from typically-developing peers (TD) and/or b) that children with SSD perceive inaccurate productions as acceptable variants of their distorted or misarticulated speech productions (Lof & Synan, 1997; Shuster, 1998). Thus, finding differences in TD and SSD perception may depend on whether or not the sounds being assessed are produced accurately or in error by the child (Locke, 1980) as well as variations in the tasks or stimuli (e.g., synthetic speech, synthetically-altered natural speech, and natural speech). Much work assessing children's speech perception has used synthetic speech, following classic studies such as Kuhl & Miller (1978); however, extending findings to natural speech is not straightforward. Perceptual judgments may also be influenced by distributional properties of the dataset (Hitchcock & Koenig, 2021; Maxwell & Weismer, 1982). The primary aim of the present work is to investigate whether TD children and those with SSD differ in their perceptual labeling of stop-initial words produced by young children. As in past work, we present data on labeling accuracy; we also add a preliminary analysis of reaction times [RTs].

2. Methods and Analysis

2.1. Participants

Listening participants included 15 monolingual English-speaking typically-developing children (TD: 9F, 6 M; age range 6;0–10;6) and 14 monolingual English-speaking children diagnosed with a speech sound disorder (SSD; 6F, 8M; age range 6;10–10;5). All children demonstrated typical language function, hearing sensitivity within normal limits, age-appropriate cognitive and motor milestones, and no significant medical or psychological history. Gender and ethnicity were not controlled. None of the children with SSD were perceived to have any voicing errors.

2.2. Listening task and stimuli

Listeners were asked to perform a forced-choice identification task in response to child-produced stimuli blocked by place of articulation (POA). All participants completed one data collection session of approximately 60–90 minutes conducted in a WhisperRoom MDL 10284 S sound booth. Stimuli were presented via Dell Latitude E6500 computers using a SB1700 soundcard and Sennheiser HD280 headphones. Stimuli consisted of a subset of single word targets from Hitchcock and Koenig (2013). Two-year old children spontaneously produced the CV target words “boo”, “pooh”, “doe”, “toe” in response to pictured stimuli. Voice onset time (VOT; Lisker & Abramson, 1964) was measured using a Pentax Computerized Speech Lab (Model-4500), referencing the acoustic waveform and wideband spectrogram. From this dataset of four words, six exemplars were chosen from each of six children with short-lag /b d/, short-lag /p t/, long-lag /b d/, and long-lag /p t/ values. For each POA and VOT category, /b d/ and /p t/ VOTs were bimodally distributed (shorter for voiced targets), separated by a 5 ms gap (see **Figure 1**). The bimodal VOT distribution consisted of four VOT ranges: Appropriate for /b d/ (0–10 ms), appropriate for /p t/ (67.5–100 ms), inappropriate for /b d/ (25–62.5 ms), and inappropriate for /p t/ (15–25 ms) (see **Figure 1**). Each listener provided 288 responses (4 target words x 6 child speakers x 6 exemplars per child speaker x 2 VOT categories, viz. appropriate and inappropriate for the target), yielding 8352 datapoints.

The design of the stimulus set (viz., toddler-produced words chosen to have bimodal VOT distributions) is a continuation from previous work (Hitchcock & Koenig, 2021). In the current context, we note the following: a) The variability inherent in young children's speech may increase the level of difficulty for listeners, i.e. provide a more sensitive test of group differences. b) Conversely, the separation between target /b d/ and /p t/

10.21437/issp.2024-63

within the short- and long-lag VOT regions may aid listeners in ascertaining the child's target.

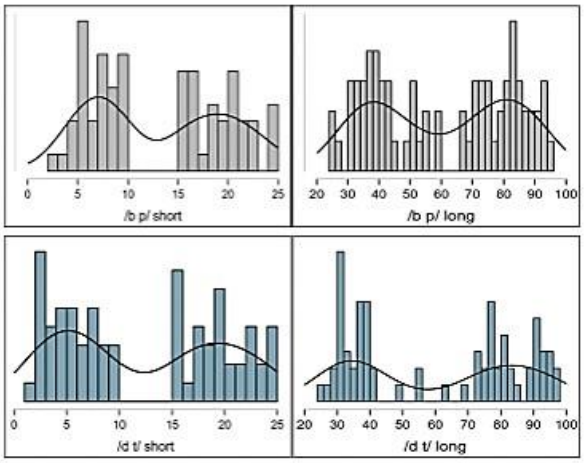

Figure 1: *Distribution of stimuli along the VOT continuum.*

2.3. Measures and processing

2.3.1 Accuracy

We classified whether listener ratings (phoneme labels) were accurate (defined as matching the speaker's intended target), and assessed RTs as described in the next paragraph. Note that children were not instructed to respond as quickly as possible.

2.3.2 Response times

Since the RTs were positively skewed, we first log-transformed the data (values that we henceforth call logRTs). We removed original RTs that were negative, which could not be log-transformed and presumably represented false starts. This removed 156 tokens from the dataset, with tokens heavily concentrated in SSD children (148/156 =95%). Three children with SSD accounted for 119 of these values. In the most extreme case (58 removed cases), we still had 83% of the child's data to analyze. We then z-transformed the data (based on the mean and SD of the full dataset), yielding logRTz. Finally, we removed logRTz values that were > |3| standard deviations from the grand mean. The final trimmed logRTz dataset contained 8084 productions.

3. Results

3.1. Accuracy

Listener responses are organized using the four categories defined above: (1) Appropriate VOTs: Productions of /p t/ with long-lag VOTs and productions of /b d/ with short-lag VOTs. (2) Inappropriate VOTs: long-lag productions of /b d/ and short-lag productions of /p t/. Results are presented in **Figures 2–3** and statistical results are summarized in Table 1.

Significant results from Shapiro-Wilks tests indicated deviation from normality for all comparisons; thus, Mann Whitney U tests were calculated to assess group differences within VOT categories. Group differences were only significant for one comparison (long-lag/inappropriate /b/). This could suggest largely comparable speech perception for TD and SSD children. Interestingly, however, for three of the four inappropriate VOT categories, accuracy was actually higher for those with SSD (albeit not always rising to the level of significance). This can be seen in **Figures 2–3**.

Table 1: *Statistics on group differences (Mann-Whitney U-values and associated p-values) for all stop consonants, with appropriate and inappropriate VOT values.*

	b	d	p	t
Appropriate VOTs				
U value	133200	135468	133650	135288
p-value	0.279	0.824	0.124	0.574
Inappropriate VOTs				
U value	128016	135792	129960	128916
p-value	*0.046	0.943	0.119	0.083

*Indicates statistical significance (p <.05).

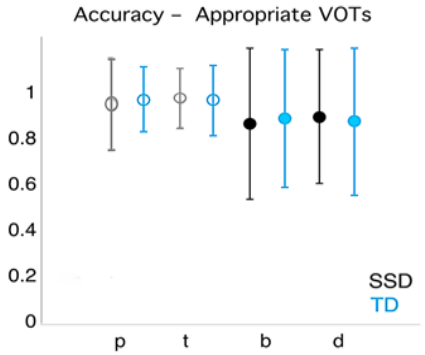

Figure 2: *Accuracy means and standard deviations for both groups – Appropriate VOT values.*

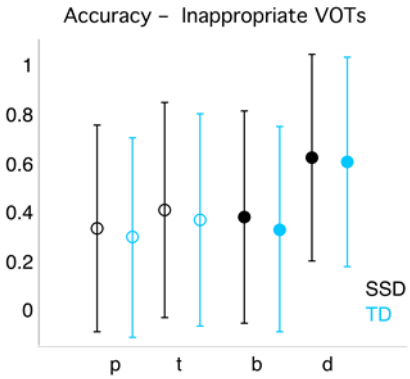

Figure 3: *Accuracy means and standard deviations for both groups – Inappropriate VOT values.*

In all cases, accuracy was much higher for appropriate VOTs (**Figure 2**) than for inappropriate VOTs (**Figure 3**) suggesting that listener judgments were mainly driven mainly by VOT. Variability is extensive in both SSD and TD groups. Unexpected high accuracy in both groups for one inappropriate VOT condition (long-lag /d/, **Figure 3**) could reflect secondary cues available in the stimuli.

3.2. Reaction times

Levene's tests of variance equality were significant across groups and accuracy measures, so we employed non-parametric statistics to test for group differences.

The data show shorter RTs for the SSD group than the TD group (SSD mean = -0.061, SD = 0.176; TD mean = 0.030, SD =

0.184). We also find shorter RTs for accurate responses than inaccurate (Accurate mean = -0.030, SD = 0.177; Inaccurate mean = 0.024, SD = 0.187). Data, split by group and accurate/inaccurate responses, are shown in **Figure 4**.

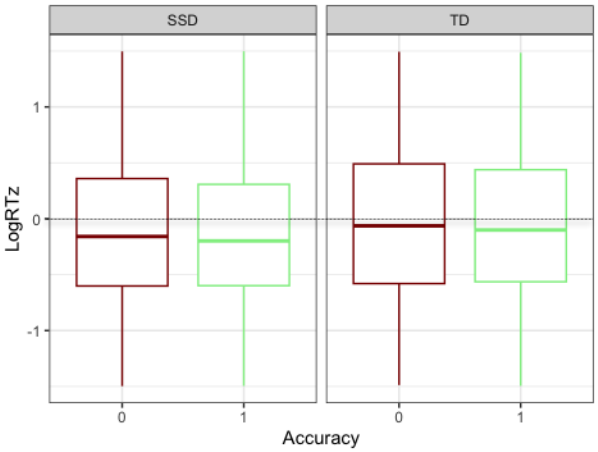

Figure 4: *Logged and z-transformed reaction times as a function of group and response accuracy (0=inaccurate, 1=accurate). The horizontal line at zero is intended to facilitate group comparisons. Outliers have been trimmed from the display.*

For both inaccurate and accurate responses, Kruskal-Wallis tests showed a highly significant group difference in logRTz (SSD < TD): Accurate responses, χ^2 = 14.400, df = 1, p<0.001; inaccurate responses, χ^2 = 10.691, df = 1, p-value = 0.001.

Evaluating whether logRTz values differed within groups as a function of accurate and inaccurate responses, we find a significant difference in the TD group: χ^2 = 4.002, df = 1, p-value = 0.046. This did not hold for the SSD group: χ^2 = 1.032, df = 1, p-value = 0.310.

Finally, we asked whether response speed differed depending on whether VOTs were appropriate or inappropriate for the target. Again, we observe a significant difference in the TD group (χ^2 = 5.359, df = 1, p-value = 0.021) but not the SSD group: χ^2 = 0.783, df = 1, p-value = 0.376. For both appropriate and inappropriate VOTs, the group difference (SSD < TD) remained significant.

As a precaution, we removed the three SSD children who contributed the greatest number of false starts and re-evaluated these conclusions (reduced dataset containing 3106 and 4261 datapoints for SSD and TD groups, respectively). Group differences remained significant in all cases. Median values are provided in Table 2. As seen before, a) all values are lower for SSD than TD; b) accurate responses are lower (faster) than inaccurate, and c) responses to appropriate VOTs are faster than to inappropriate VOTs. These values demonstrate (see also **Figures 2–3**) that group differences are quite modest.

Finally, **Figure 5** shows the logRTz values for individual listeners in both groups. There is considerable group overlap at the low end (faster RTs), but the groups diverge at the high end (slower RTs).

Table 2: Median LogRTz values divided by group (SSD, TD), response accuracy (inaccurate, accurate), and target VOT (inappropriate, appropriate).

	Response		Stimulus VOT	
	Inacc.	Acc.	Inapp.	App.
SSD	-0.115	-0.161	-0.136	-0.162
TD	0.001	-0.076	-0.005	-0.083

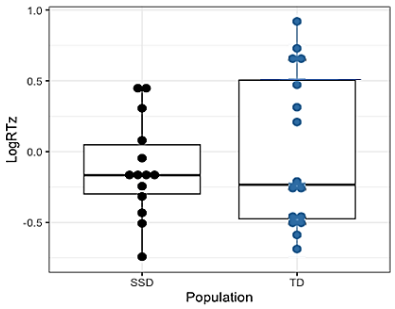

Figure 5: *Median LogRTz values for all speakers in both groups.*

4. Discussion and conclusion

4.1. Accuracy

In both child groups, labeling was highly accurate for targets with appropriate VOTs. This is consistent with previous work showing high accuracy in adults for young child productions with appropriate VOT values (Hitchcock & Koenig, 2021). The statistical results for perceptual accuracy are also generally consistent with studies suggesting that children with SSD do not show clear perceptual deficits on non-errored sounds compared to their TD peers. At the same time, slightly higher accuracy levels for *in*appropriate VOT targets in children with SSD warrants further investigation and could suggest subtle perceptual differences between groups that are not seen in other testing paradigms. Potentially, children with SSD could have less refined perceptual skills and wider boundaries in their categorical labeling functions than TD children, even for sounds that are not produced in error.

Hitchcock and Koenig (2021) explored adult labeling of toddler speech that did not incorporate the bimodal stimulus distributions used here. The adult responses to inappropriate VOT values for /p t/ were considerably lower than those observed here (11–15%). In follow-up studies, we have observed higher labeling accuracy for adults and children listening to bimodally-distributed data. This suggests that bimodal distributions of VOT within short- and long-lag ranges may lead to higher-than-expected accuracy for listener responses. To the extent that distributional characteristics of the data influenced listener responses in the current work, it appears to have had largely similar effects in both TD and SSD groups.

4.2. Reaction times

Reaction time (logRTz) data were slower for inappropriate VOTs in both groups, as one might expect. LogRTz's were also slower for inaccurate responses in both groups. Importantly, this held for both groups, and moreover, rating accuracy did not differ greatly between groups. Follow-up analyses could assess only correct responses, but this would lead to high data loss in some of the VOT categories and limit sensitivity to group differences. Perhaps the most surprising finding is that the SSD

group had faster reaction times, and this difference remained significant regardless of accurate/inaccurate responses, appropriate/ inappropriate VOTs, and removing children who had atypical (false-start) RTs. This finding, though tentative, deserves further exploration and could indicate some differences between how SSD and TD children process speech, or respond to a task like the one we presented here.

4.3. General conclusions

Listener accuracy for SSD and TD groups was largely comparable, in line with past work suggesting that children with SSD do not show clear speech perception difficulties for non-errored sounds. Interestingly, however, for inappropriate VOTs the SSD group, on average, tended to out-perform their TD peers, and this was significant in one of four comparisons. This result deserves further exploration. As part of this, we will explore individual differences among the listeners (Kong & Edwards, 2016). We also plan to assess how secondary cues in the stimuli (durational measures, f0, burst intensity) might have contributed to listener responses in TD and SSD groups.

The current RT results speak against the notion that children with SSD have a general speech perception difficulty that is manifested in slower responses, at least for non-errored sounds. Nevertheless, this modest dataset does not allow us to assert with confidence that children with speech-sound disorders are universally faster in their phonetic labeling. Along with replicating these results in larger listener groups, future work should employ more sophisticated modeling to tease apart the many possible inter-relationships among group, VOT category, response accuracy, etc.

5. Acknowledgements

The authors would like to thank the participants and their families for their ongoing cooperation throughout the study. We also express our thanks to graduate research assistants Madeline Cheyne, Amy Rosen and Dana Catalano, graduate research assistants for their support with data collection and management.

6. References

Hitchcock, E. R., & Koenig, L. L. (2013). The effects of data reduction in determining the schedule of voicing acquisition in young children. *Journal of Speech, Language, and Hearing Research, 56*(2), 441–457.

Hitchcock, E. R., & Koenig, L. L. (2021). Adult perception of stop consonant voicing in American-English-learning toddlers: Voice onset time and secondary cues. *Journal of the Acoustical Society of America, 150*(1), 460–477.

Kong, E. J., & Edwards, J. (2016). Individual differences in categorica l perception of speech: Cue weighting and executive function. *Jou rnal of Phonetics*, 59, 40-57.

Kuhl, P.K., & Miller, J.D. (1978). Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli. *Journal of the Acoustical Society of America, 63*(3), 905–917.

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

Locke, J. (1908). The inference of speech perception in the phonologically disordered child. Part I: A rationale, some criteria, the conventional tests. *Journal of Speech and Hearing Disorders, 45*(4), 431–444.

Lof, G. L., & Synan, S. T. (1997). Is there a speech discrim-ination/perception link to disordered articulation and phonology? A review of 80 years of literature. *Contemporary Issues in Communication Science and Disorders, 24(*Spring), 57–71.

Maxwell, E. M., and Weismer, G. (1982). The contribution of phonological, acoustic, and perceptual techniques to the characterization of a misarticulating child's voice contrast for stops. *Applied Psycholinguistics, 3*(1), 29–43.

Shuster, L. I. (1998). The perception of correctly and incorrectly produced /r/. *Journal of Speech, Language, and Hearing Research, 41*(4), 941–950.