Assessing the Utility of Judgments of Children's Speech Production Made by Untrained Listeners in Uncontrolled Listening Environments

Benjamin Munson[1]

[1] Department of Speech-Language-Hearing Science,
University of Minnesota, Minneapolis, MN USA

Munso005@umn.edu

Abstract

Speech perception experiments conducted on the internet have numerous benefits over traditional laboratory studies. For example, they have the potential benefit of including many more participants than can be recruited in traditional laboratory-based studies, including members of special hard-to-recruit groups. In the study of phonological development, they have the potential for generating judgments of children's speech production accuracy that reflect the consensus of the child's speech community, rather than the judgments of a small group of phonetically trained individuals. Internet-based studies have the potential disadvantage of being conducted in less controlled listening environments with more variable equipment than laboratory studies. This study compared the performance of two groups of listeners: ones participating in a sound booth with high-quality headphones and ones participating over the internet with whatever equipment was available. Listeners participated in a series of tasks rating children's productions of the sounds /s/, /ʃ/, /θ/, /d/, /g/, /t/, and /k/. Using a variety of dependent measures, no group differences were found between internet and laboratory listeners. The potential utility of these judgments for studies of phonological development are discussed.

Index Terms: speech perception, speech-sound development

1. Introduction

Recent research suggests that at least some aspects of speech-sound acquisition can be understood as the gradual differentiation of similar sounds. For example, a recent cross-sectional study by [1] showed that Japanese and English acquiring preschool children produce gradually larger differences in spectral centroids /s/ and /ʃ/ (in English) or /ɕ/ (in Japanese) across the 2 to 5 year age range. The continuous nature of speech-sound acquisition is also evident in studies of 'covert contrasts', i.e., reliable acoustic differences between pairs of sounds transcribed to the same. These have been shown in typically developing children for a variety of contrasts, including voicing contrasts [2], obstruent place contrasts [3],[4], and manner of articulation contrasts [5]. Covert contrasts have also been shown in the speech of children with speech sound disorder [6]. [7] showed that these are of potential clinical significance of covert contrasts: children with SSD with evidence of covert contrast progressed through speech therapy more quickly and generalized correct production more readily than children without covert contrasts.

Phonetic transcriptions are often used in large scale normative studies of speech-sound development, such as [8], and diary studies, such as [9]. They are used almost exclusively in clinical assessments of the speech of children with SSD. Transcription is ill-suited to assess the gradual acquisition of sounds as described in the previous paragraph. Acoustic and articulatory analyses are possible solutions to this problem, though both have challenges to being implemented in real-world settings. Acoustic analysis in particular requires a signal that is relatively free from background noise. Direct articulatory measures can be relatively easy to collect (i.e., through ultrasound) though the equipment needed to make them and process them can be expensive.

One other solution is to gather perceptual judgments from multiple listeners. The average of these judgments could be used as a continuous measure of production. Figures 1 and 2 illustrate the viability of this, by examining the average or visual-analog scale judgments (Figure 1, based on data from [11]), or binary judgments (Figure 2, based on the data from [10]) of children's productions of /s/ and /ʃ/. The correlation between both types of judgment and the centroid frequency of the fricative is high. More importantly, the ratings for sounds that are intermediate between the endpoints are themselves intermediate. This property makes suggests that these ratings might be a useful gradient perceptual measure of children's speech-production accuracy.

Judgments like those in Figures 1 and 2 could be useful proxies for the careful acoustic measures in observational studies of speech-sound development. For example, consider the diary study of /t/ and /k/ development presented by [9]. Perceptual ratings of that child's productions might have revealed a different and more gradual pattern of acquisition than was suggested by the categorical transcriptions from that study. Judgments like these might also be useful in documenting gradual progress during speech therapy by children who are learning new sounds.

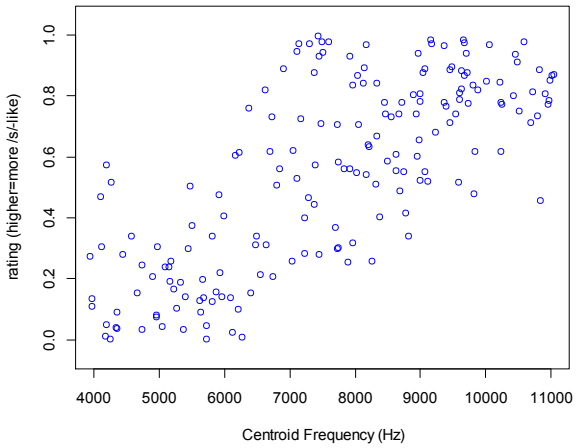

Figure 1: *The relation between visual-analog scale judgments of children's productions of target /s/ and /ʃ/ and the centroid frequencies of the fricatives being rated.*

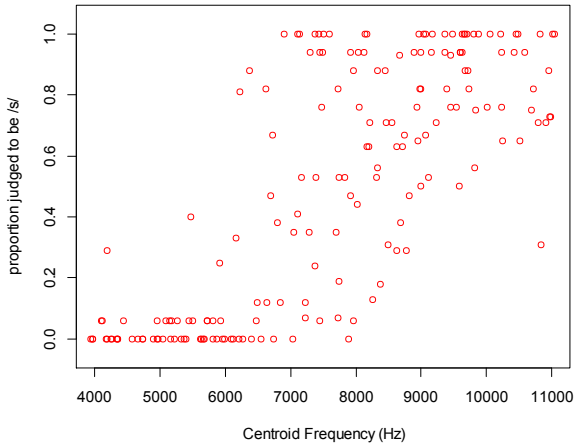

Figure 2: *The relation between average forced-choice categorizations of children's productions of target /s/ and /ʃ/ and the centroid frequencies of the fricatives being rated.*

The advent of tools for crowd-sourcing work from large groups of listeners who perform work over the internet (i.e., Amazon Mechanical Turk, [12], [13) makes this type of data collection feasible. Large number of listeners on internet-based data-systems like could provide ratings of large number of children's productions, the averages of which could be used as dependent measures in clinical and observation studies of children's speech-sound development. As argued by [14], these data would have the advantage of being more gradient than phonetic transcriptions, as well as reflecting the consensus judgments of a speech community, rather than the potentially idiosyncratic judgments of a single transcriber.

Prior to conducting this type of research, we must first determine whether the relatively uncontrolled listening conditions associated with listening in the field systematically affect the ratings that naïve listeners give to children's speech stimuli. A small number of studies have successfully used internet-based listeners in speech perception experiments. For example, [15] found that these listeners adjust their perception of stop voicing based on different types of exposure, as is predicted by Bayesian models of speech perception. However, stop voicing is robustly encoded even in noisy signals. Stop and fricative place contrasts are often dependent on relatively high-frequency spectral cues. These cues might be more susceptible to masking from ambient noise in real-world listening environments that internet listeners might be in when making perception judgments. This might limit the extent to which listeners can perceive subtle differences in high-frequency spectral information, particularly in /s/ and /t/. Hence, a study is needed to determine whether stop and fricative place contrasts can be robustly and reliably perceived by internet-based listeners.

This study examined listeners' ratings of speech that was elicited from typically developing children. We compared the performance of listeners participating in a sound booth using high quality headphones and ones participating over the internet using whatever equipment they had available. We compared the average ratings (i.e., whether the groups had different biases to particular endpoint sounds), the extent to which they used the full range of ratings available to them (i.e., the degree of categoricity of perception), and, for some sound contrasts, the correlation between ratings and the sounds' acoustic characteristics.

2. Methods

2.1 Stimuli

The stimuli were 60 productions by children, 15 productions of target /s/ and /ʃ/, 15 of target /s/ and /θ/, 15 of target /d/ and /g/, and 15 of target /t/ and /k/. These were a subset of productions that had been used as stimuli in two previous studies [16] and [17]. They were selected from the παιδολογος [paidologos] corpus of children's productions [18]. The tokens were chosen based on ratings that had been provided in previous studies, and by their acoustic characteristics, so that they formed a kind of naturally occurring continuum from one endpoint sound to the other.

Acoustic measures of the 15 /s/-/ʃ/ stimuli are presented in Tables 1 below. As this table show, the 15 /s/-/ʃ/ stimuli varied widely in the acoustic characteristics that differentiate between adults' productions of these sounds. Acoustic measures of the other three sets of stimuli are not presented here because preliminary analyses showed them not to be strongly correlated with the listeners' judgments in this experiment. This is likely because of the small number of tokens used in this experiment, as previous studies with a larger number of tokens have found a relationship between listeners' ratings and the sounds' acoustic characteristics (i.e., [16], [17]). It is noteworthy to point out that the 15 stimuli chosen for each of these blocks had been found previously by [16] and [17] to vary in the ratings that they elicited from both phonetically trained and untrained listeners, so we can assume that they represent a perceptual continuum between the endpoint.

Table 1. *Selected acoustic characteristics of the /s/-/ʃ/ stimuli.*

M1 (Hz)	Onset F2 (Hz)
4961	2981
5084	2416
5167	2651
5810	3289
5872	2240
6376	2473
6695	2782
7369	2046
7472	2126
8317	2848
9584	2853
9630	2225
9881	1734
10449	2125
10946	2257

The productions were elicited using a picture-prompted repetition task, in which children viewed a picture while hearing an audio prompt of the word. They repeated the word. All of the tokens that were selected to be stimuli in this study were transcribed as either an endpoint stimulus, a substitution

of one endpoint for the other (i.e., a [t] for /k/ error), or, following the suggestions of [19], something intermediate between the endpoint sounds. The stimuli were truncated to contain only the initial consonant-vowel sequence. This ensured that listeners would not use lexical knowledge when responding.

2.2 Listeners

2.2.1 Internet

Internet listeners (n=21) were elicited through advertisements made on social media and through word of mouth. Recruitment materials indicated that listeners should be native speakers of English without a current or former speech, language, or hearing impairment.

2.2.2 Laboratory

Laboratory listeners (n=20) were recruited through fliers on the University of Minnesota campus and through word of mouth. The criteria were the same as those for the internet.

2.3 Procedures

Participants in both groups responded through a custom interface that was designed by the Survey Development Service of the University of Minnesota College of Liberal Arts. The laboratory listeners all used the same Internet browser to respond; the internet listeners used whatever browser they chose.

On each trial, listeners used the audio player embedded in their Web browser to play a clip of a single production. Internet listeners were encouraged to listen through the highest-quality headphones that they had available, and to listen over a speaker only if headphones were not available. Listeners made a judgment along a 9-point scale anchored by "the 's' sound" and "the 'sh' sound", "the 's' sound" and "the 'th' sound", "the 't' sound" and "the 'k'' sound", or "the 'd' sound" and "the 'gh' sound." Figure 3 shows a display that was used to elicit a rating for a sound in the /s/-/ʃ/ block. Listeners could play the sound more than once, but were asked to listen to each token no more than twice before they made their rating.

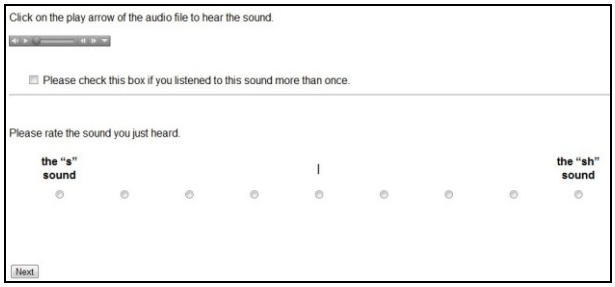

Figure 3. *Example response screen for the /s/-/ʃ/ rating task.*

Ratings were blocked by contrast type. Every listener received the same quasi-random order of stimuli and of blocks. In addition to the speech-rating trials, listeners were asked a variety of questions about their attitudes and beliefs about speech-sound development and variation in speech perception. This was done to accommodate the design of a larger study on the influence of clinical training on speech-language pathologists' perception of children's speech [20].

3. Results

3.1 Categoricity of response

The first dependent measure examined was the extent to which the responses used the entire 9-point scale, as opposed to just the end-point values. The proportion of responses from each listener that were the endpoints (i.e., ratings of 1 or 9) were calculated. Box plots of these data are shown in Figure 4. These data were subjected to separate non-parametric Mann-Whitney U tests to examine whether the proportion of 1 and 9 responses differed between the internet listeners and the laboratory listeners for the four stimulus sets. None of the differences were significant (Mann-Whitney U=207.000, Wilcoxon W=438.000, z=-0.079, p=0.937 for /s/-/ʃ/ Mann-Whitney U=192.500, Wilcoxon W=402.500, z=-0.458, p=0.647 for /s/-/θ/, Mann-Whitney U=193.000, Wilcoxon W=424.000, z=-0.447, p=0.655 for /d/-/g/, Mann-Whitney U=172.000, Wilcoxon W=403.000, z=-0.997, p=0.319 for /t/-/k/).

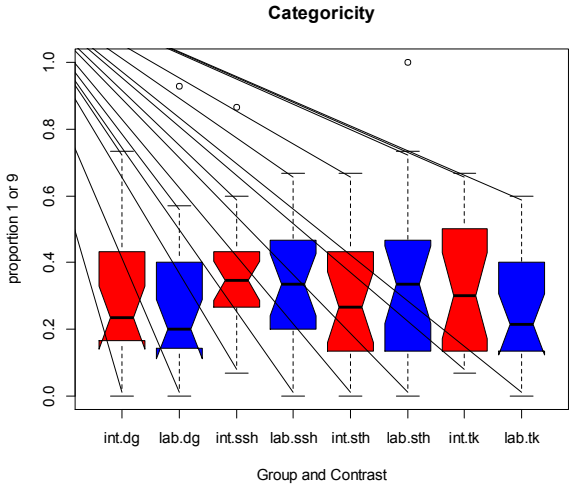

Figure 4: *Proportion of Endpoint responses by Internet listeners (red) and laboratory listeners (blue) for the four contrast types.*

3.2 Average Ratings

The next set of analyses examined whether the two groups differed in the average ratings that they gave to the stimuli, to determine whether one set of listening conditions biased people to respond to a particular endpoint. Each listener's average rating for the four sets of stimuli were calculated. These are shown in Figure 5. These were used as dependent measures in a series of Mann-Whitney U Tests comparing the groups. None of the differences were significant (Mann-Whitney U=157.000, Wilcoxon W=367.000, z=-1.385, p=0.166 for /s/-/ʃ/, Mann-Whitney U=171.000, Wilcoxon W=402.000, z=-1.018, p=0.309 for /s/-/θ/, Mann-Whitney U=204.500, Wilcoxon W=414.500, z=-0.144, p=0.886 for /d/-/g/, Mann-Whitney U=196.500, Wilcoxon W=427.500, z=-0.352., p=0.725 for /t/-/k/).

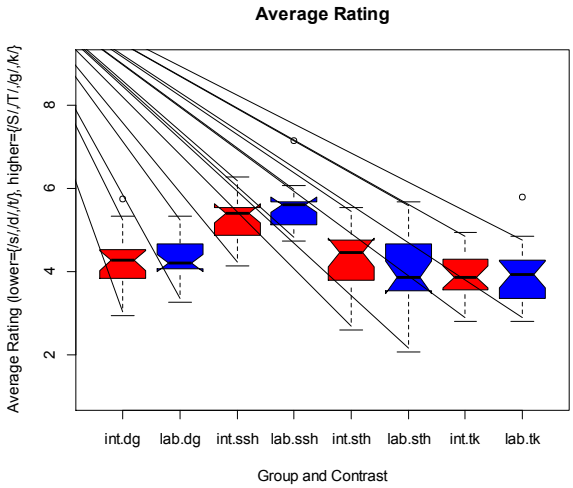

Figure 5: *Average Rating for the Internet Listeners (red) and the Laboratory Listeners (blue) for the four contrast types.*

3.4 Correlation between ratings and sounds' acoustic characteristics

The last set of analyses examined the extent to which the listeners' judgments could be predicted by the acoustic characteristics of the stimuli. A series of regressions were conducted in which individual listeners' ratings were predicted by the acoustic characteristics of the stimuli. These were conducted only for the /s/-/ʃ/ stimuli, as described in section 2. For the /s/-/ʃ/ regressions, the average R^2 for the listeners in the laboratory was 53.7%, while the average for the internet listeners was 59.3%. This difference was not significant, Mann-Whitney U=159.000, Wilcoxon W=390.000, z=-1.330, p=0.183. A plot of the average ratings for the /s/-/ʃ/ stimuli by the fricative centroid is shown in Figure 6.

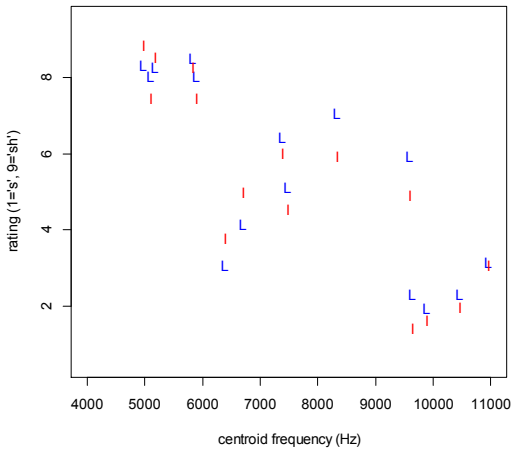

Figure 6: *Average ratings for the 15 /s/-/ʃ/ tokens by the internet listeners (I, in red) and the laboratory listeners (L), plotted as a function of the centroid frequencies of the fricatives being rated.*

4. Discussion and Conclusions

The results of this study suggest that judgments of children's speech made by listeners over the internet are functionally equivalent to those made in a research laboratory, at least for the four sets of contrasts examined in this paper. The presumably less favorable listening environments of the internet listeners did not bias their responses toward one endpoint sound. Most importantly, both sets of listeners provided intermediate judgments for sounds that were intermediate between the endpoint stimuli. This finding means that judgments of children's speech-production accuracy that are crowd-sourced from the internet can potentially be useful to assess intermediate productions in both observational studies and in measuring speech sound learning by children undergoing speech therapy.

5. Acknowledgements

This research was supported by NIH grant DC 02932 to Jan Edwards, by NSF grant BCS0729277 to Benjamin Munson, and by funds from the University of Minnesota College of Liberal Arts. Sam Kortes, Marie Meyer, and Tom Lindsey provided useful support for this research.

6. References

[1] Li, F. (2012). Language-specific developmental differences in speech production: a cross-linguistic acoustic study. *Child Development, 83*, 1303-1315 .

[2] Macken, M., & Barton, D. (1980). A longitudinal study of the acquisition of the voicing contrast in American English word-initial stops, as measured by voice onset time. *Journal of Child Language, 7*, 41-74.

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

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

[5] Scobbie, J., Gibbon, F., Hardcastle, W., & Fletcher, P. (2000). Covert contrast as a stage in the acquisition of phonetics and phonology. In M. Broe & J. Pierrehumbert (Eds), *Papers in Laboratory Phonology V* (p. 194-206). Cambridge: Cambridge University Press

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

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

[8] Smit, A. B., Hand, L., Freilinger, J. J., Bernthal, J. E., & Bird, A. (1990). The Iowa Articulation Norms Project and its Nebraska replication. *Journal of Speech and Hearing Disorders, 55*, 779-798.

[9] Inkelas, S., & Rose, I. (2007). Positional neutralization: a case study from child language. *Language, 84*, 707-736.

[10] Urberg Carlson, K., & Munson, B. (submitted). *Research Note: A Comparative Study of Perceptual Ratings of Children's Production of Sibilant Fricatives.* Manuscript under consideration.

[11] Li, F., Munson, B., Edwards, J., Yoneyama, K., & Hall, K.C. (2011). Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development. *Journal of the Acoustical Society of America, 129*, 999-1011.

[12] Buhrmester, M., Kwang, T., & Gosling, S. (2011). Amazon's Mechanical Turk: A new source for inexpensive, yet high-quality, data? *Psychological Science, 6*, 3-5.

[13] Germine L, Nakayama K, Duchaine BC, Chabris CF, Chatterjee G, Wilmer JB. (2012). Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic Bulletin and Review, 19*, 847-57.

[14] Edwards, J., Beckman, M.E., & Munson, B. (in press). Cross-language differences in speech sound acquisition. In M. Redford (Ed.), *Handbook of Speech Production*. Hoboken, NJ: John Wiley and Sons.

[15] Kleinschmidt, D., & Jaeger, T.F. (2012). A continuum of phonetic adaptation: Evaluating an incremental belief-updating model of recalibration and selective adaptation. In Miyake, N., Peebles, D. & Cooper, R.P. (Eds.), *Proceedings of the 34th Annual Conference of the Cognitive Science Society.* Sapporo, Japan. Cognitive Science Society.

[16] Julien, H and Munson, B. (2012). Modifying Speech to Children based on their Perceived Phonetic Accuracy. *Journal of Speech, Language, and Hearing Research,* 55, 1836-1849.

[17] Munson, B., Johnson, J., & Edwards, J. (2012). The Role of Experience in the Perception of Phonetic Detail in Children's Speech: A Comparison of Speech-Language Pathologists with Clinically Untrained Listeners. *American Journal of Speech-Language Pathology*, 24, 124-139.

[18] Edwards, J., & Beckman, M.E. (2008). Methodological questions in studying consonant acquisition. *Clinical Linguistics and Phonetics, 22*, 937-956.

[19] Stoel-Gammon, C. (2001). Transcribing the speech of young children. *Topics in Language Disorders, 21*, 12-21.

[20] Meyer, M.K. (2012). *Do attitudes and practice patterns predict the perception of children's speech? Evidence from a web-based audio survey of Speech-Language Pathologists.* M.A. Thesis, University of Minnesota.