Correlates to intelligibility in deviant child speech – comparing clinical evaluations to audience response system-based evaluations by untrained listeners

Sofia Strömbergsson [1], *Christina Tånnander*[2]
[1] Department of Speech Music and Hearing, KTH, Stockholm, Sweden
[2] Swedish Agency for Accessible Media, Johanneshov, Sweden
sostr@kth.se, christina.tannander@mtm.se

10.21437/Interspeech.2013-585

Abstract

The severity of speech impairments can be measured in different ways; whereas some metrics focus on quantifying the specific speech deviations, other focus on the functional effects of the speech impairment, e.g. by rating intelligibility. This report describes the application of a previously untested method to the domain of deviant child speech; an audience response system-based method where listeners' responses are continuously registered during playback of speech stimuli. 20 adult listeners were given the task of clicking a button whenever they perceived something unintelligible or deviant during playback of child speech stimuli. The untrained listeners' responses were compared to clinical evaluations of the same speech samples, revealing a strong correlation between the two types of measures. Furthermore, patterns of how listeners' different experiences influence their clicking responses were explored. Qualitative analysis linking listener clicks to triggering events in the speech samples demonstrates the potential of the click method as an instrument for identification of features in children's speech that are most detrimental to intelligibility – insights that may have important implications for the selection of speech targets in clinical intervention.

Index Terms: child speech, intelligibility, listening tests

1. Introduction

Deviant speech production is communicatively impairing; impaired intelligibility often leads to misunderstandings, which in turn may cause communicative breakdowns. For children, early in typical speech acquisition as well as later in impaired speech development, these difficulties are especially pronounced when interacting with people they do not already know [1] [2].

Although speech intelligibility is an important consideration in many clinical decisions, e.g. when determining whether a child should be enrolled in therapy, and when evaluating progress in intervention, it is not trivially assessed. In clinical practice, speech-language therapists (SLTs) often estimate intelligibility by means of impressionistic statements [3] [4], despite the fact that these methods are susceptible to subjectivity and inconsistency. An alternative, quantitative, approach is to rate intelligibility using a scale, e.g. on a Likert scale from 1 to 7, where 1 indicates that the sample is essentially unintelligible, and 7 that the sample is essentially intelligible, e.g. [3]. However, despite their ease of use [3] [5], and their evidenced reliability within and across listeners [6], rating-scale measures of intelligibility stand the risk of being too coarse to be of real clinical value [6]. A third approach to assess intelligibility is by using word identification tests, where the examiner writes down the words (s)he understands and notes what (s)he does not understand. The percentage of intelligible words across all words can then be used as a measure of intelligibility, see e.g. [5] [4].

Although it might be intuitive to assume a strong correlation between a child's speech production skills and the perceived intelligibility of his or her speech, the relation between the two factors is just weak [1] [7]. This reflects the fact that intelligibility is dependent also on factors outside of the speaker, e.g. listener characteristics, the communicative context, and the message content. To address this complexity, a measure of *functional intelligibility* has recently been suggested in [8], designed to assess the severity of speech sound disorders (SSDs) with respect to the functional consequences of impaired speech production, e.g. the degree to which a particular child is understood by unfamiliar people. However, the most widely used metric of severity of SSDs is still the Percentage of Consonants Correct (PCC) [7]. This measure is calculated as the proportion of correctly produced consonants (as judged by a trained clinician) across all target consonants in a speech sample. Despite its established reliability and validity as a quantitative measure of severity of involvement [7], there are also some limitations associated with the PCC metric. First, it is not applicable to very unintelligible speech; the number of target consonants can only be estimated if the target words are known, and that is not always the case in unintelligible speech. Second, calculating PCC is a time-consuming process, and is therefore more often used in experimental research than by clinicians [3].

Of the contextual factors influencing intelligibility, the listener's familiarity with the speaker has been shown to play an important role; family members are, for instance, better at glossing a child's intended words than unfamiliar people [1]. Regarding the listener's experience with hearing deviant speech in general, its influence on intelligibility rating remains more obscure; although it has been found that in intelligibility experiments with speech of hearing-impaired speakers, listeners with experience of this type of speech give higher intelligibility scores than inexperienced listeners [9], this effect interacts with the linguistic content and speaker characteristics. Smaller or no effects on listener's intelligibility ratings have been reported in the context of dysarthric speech [10]. When rating the severity of phonologically impaired speech, high congruence has been found between experienced and inexperienced listeners [7].

Audience Response Systems (ARS) have long been used in concurrent evaluations of e.g. movies and screenplays, where many subjects are asked to click a button when they like (or dislike) what they see. The method has also been used for time-efficient evaluation of speech synthesis by many subjects, where the listeners are asked to click whenever they hear something they do not like or do not understand [11]. Applying the ARS-based method to recordings of deviant

speech presents itself as an interesting opportunity. First, compared to standard methods of intelligibility or severity rating, the method allows for fast collection of ratings from many listeners, thus strengthening the reliability of the ratings. Second, the use of naïve listeners as raters gives an indication of the extent of the everyday problems these children experience when communicating with unfamiliar people, thus relating to the concept of functional intelligibility [8]. Third, the real-time ratings can serve as pointers to salient speech problems, indicating what speech phenomena are most disturbing for the listeners. If coupled with qualitative speech analysis, this information can go far beyond standard measures of intelligibility/severity.

1.1. Research questions

The research questions addressed in the present study are:

1) How do untrained listeners' online ratings of intelligibility relate to standard measures of severity performed by SLTs?
2) Are the listeners' ratings dependent on their experience of speech analysis and/or their experience of communicating with pre-school aged children?
3) What features in the children's speech are the most salient correlates to unintelligibility?

The answers to the first and second questions will extend our current knowledge of the ecological validity of the standard methods for assessing severity of speech production problems. The third question will be explored by detailed qualitative analysis of one speech sample; from the preliminary results presented here, the potential value of the ARS-based method as an instrument of locating sources of unintelligibility in the speech of children will be demonstrated.

2. Method

2.1. Speech data

Conversational speech was recorded from 7 Swedish children, 5 of them diagnosed with phonological impairment (PI), and 2 children with typical speech. All children with PI exhibited velar fronting in their speech, i.e. a pattern of substituting [t], [d], [n] for /k/, /g/ /ŋ/, respectively. All of them also exhibited additional phonological processes, e.g. stopping of fricatives and/or weakening of /r/ or /l/. Of the two children with typical development (TD), only the younger one presented with deviant (albeit age-adequate) speech, with /r/-weakening and substitution of [ʃ] for /s/. The collected speech characteristics of all recorded children are presented in Table 1. At recruitment, consent forms, which complied with Swedish ethical guidelines for subject participation were used.

In the recording situation, the children and an adult (a certified SLT or an SLT student) talked about toys or pictures, visible to both of them. The children were recorded with a Zoom H2 recorder with a 44 kHz sampling frequency. Samples of continuous child speech were extracted manually from the child-adult conversations. Samples containing distracting noises or longer sequences of overlapping speech were not included. For each recording, samples were then sequentially concatenated, with intervening periods of silence (1 sec), to form approximately 1 minute long speech samples. In all, 17 such speech samples were used in the listening test.

Table 1. *Characteristics of the recorded children. Children 1-5 are diagnosed with PI, whereas children 6-7 are children with TD.*

Child	Age	Speech deviations
1	4;7	Velar fronting, stopping, /r/-weakening, neutralizing voicing contrast, cons. cluster reductions
2	4;11	Velar fronting, /r/-weakening, /l/-weakening, cons. cluster reductions
3	5;4	Velar fronting, stopping, /r/-weakening, cons. cluster reductions
4	5;8	Velar fronting, stopping, /r/-weakening, cons. cluster reductions
5	5;3	Velar fronting, cons. cluster reductions
6	4;2	-
7	3;3	/r/-weakening, /s/ → [ʃ]

2.2. Clinical evaluations

Percentage of Consonants Correct (PCC) was calculated for all speech samples, along the procedures described in [7]. Additionally, Percentage of Velar Consonants Correct (PCC-V) was calculated analogously. PCC and PCC-V were first calculated by one rater and then by another (both of which were either a certified SLT or an SLT student). Coding reliability was determined for all speech samples; the mean difference between pairs of PCC scores was 0.5%, and for PCC-V, the mean difference was 3%. For each speech sample, the average between the two raters' PCC scores was designated as the PCC value for that sample. Similarly, an averaged PCC-V value was designated as the PCC-V value for each sample.

Two speech samples were subject to qualitative analysis, in which the first author (a certified SLT) used Wavesurfer [12] to mark and label all speech deviations, with regards to the phonological processes described in Table 1. Segments perceived as unintelligible were assigned the label "unintelligible", and typically ranged across several words. From the resulting timestamps, the midpoint and the duration of the event were passed on to distributional analysis.

2.3. ARS-based listening test

20 adults participated as listeners in the ARS-based listening test; their age varied between 23 and 70 (M = 41, SD = 11). The gender distribution was 45% female and 55% male. The listeners reported varying experience of interacting with young children (40% of the listeners reported that they were used to talking to children under 6 years of age). Moreover, the listeners indicated whether they were "phoneticians, speech scientists or similar", or not; 50% of the listeners reported a speech science background. None of the listeners worked with pre-school-aged children professionally.

One of the 17 speech samples (from a child with PI) was selected as an introductory sample and was not included in the analysis. The remaining 16 samples were randomized for each subject, and implemented in a web-based ARS listening test. Before starting the test, the listeners were informed that the speech samples had been extracted from conversations between children and an adult discussing pictures or objects visible to both of them. The listeners were instructed to listen to the speech samples and to strike any keyboard key (or

mouse key) whenever they perceived something unintelligible or deviant during playback of the child speech stimuli. The average number of clicks over all listeners and all speech samples was used in the weighting of each individual listener's clicks, so that clicks from listeners who do not click very often is worth more than clicks from listeners who click more frequently.

2.4. Data analysis

The distribution of the weighted clicks and the distribution of manually annotated speech events were analyzed by means of Kernel Density Estimation (KDE) estimations. The analysis resembles a histogram, but the produced curve is continuous and smooth.

3. Results

3.1. PCC scores vs. click-test scores

Table 2 displays the PCC scores and the scores from the ARS-based test for all speech samples. A Pearson's product-moment correlation determined the relationship between PCC scores, PCC-V scores and click scores for the speech samples. This analysis showed a strong, negative correlation between PCC and the number of clicks: $r = -.81$, $n = 16$, $p < .001$, and a weaker, negative correlation between PCC-V score and the number of clicks: $r = -.55$, $n = 16$, $p = .03$.

Visual inspection of the results in Table 2 also reveals that the scores of the children with typical speech – samples 15 and 16 – differ from the scores of the other samples; here the PCC scores are higher, and particularly for sample 15, the number of clicks is smaller compared to the other samples.

3.2. Variation across listeners

The clicking frequency varied extensively across listeners, with a listener's total number of clicks ranging from 60 to 478 clicks ($M = 210$, $SD = 109$). A one-way ANOVA showed that the variation between listeners could neither be attributed to the listeners' varying experience with child speech, $F(1, 16) = .15$, $p = .70$, nor to their varying experience with speech analysis, $F(1,16) = .22$, $p = .64$. Neither was there an interaction between these factors, $F(1, 16) = .06$, $p = .80$. There was also no correlation between the listeners' age and their total number of clicks, $r = .38$, $n = 20$, $p = .10$. The variation in the number of clicks per speech sample could not be linked to the order of presentation in test, $r^2 < .001$, $p = .84$.

In order to explore whether listeners clicked more or less often depending on the position within the speech sample, another linear regression analysis was conducted. This showed a significant but weak negative correlation ($r^2 = .001$, $p < .001$).

Table 2. *Percentage of Consonants Correct (PCC), Percentage of Velar Consonants Correct (PCC-V) and click scores for all speech samples.*

Sample	Child	PCC	PCC-V	N of clicks (weighted)
1	1	58%	23%	15
2	1	51%	12%	21
3	1	56%	0%	21
4	2	77%	39%	13
5	2	73%	52%	13
6	3	66%	11%	13
7	4	74%	12%	9
8	4	77%	23%	15
9	4	62%	20%	23
10	4	73%	33%	14
11	5	86%	13%	8
12	5	79%	10%	9
13	5	85%	17%	13
14	5	73%	4%	14
15	6	95%	100%	2
16	7	87%	100%	9

3.3. Speech correlates to clicks

Figure 1 illustrates how listener clicks and annotated speech events (i.e. segmental deviations or stretches of unintelligible speech) are distributed in speech sample 15 (see Table 2). As this sample represents a child with adult-like speech, few or no peaks are expected. And indeed, as illustrated in the figure, only three speech production problems were annotated (represented by the peaks e_1, e_2 and e_3). The click peaks c_1 and c_2 both occur about one second after the event peaks e_1 and e_2 respectively, and considering the absence of other event peaks in the vicinity, it is reasonable to assume that these particular clicks were evoked by these particular events. However, the fact that the third click peak (c_3), occurs *before* the event peak, warrants an alternative explanation. Inspection of the preceding vicinity of the peak in the speech sample reveals that the child produces the phrase "Det är t-rexar" (*It is t-*

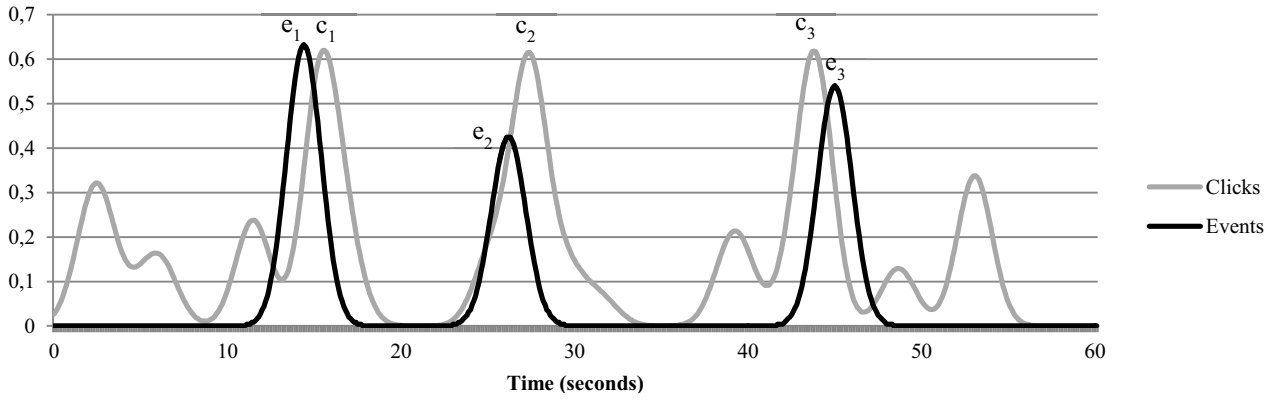

Figure 1. *Kernel density estimate curves for listener clicks and speech events in speech sample 15.*

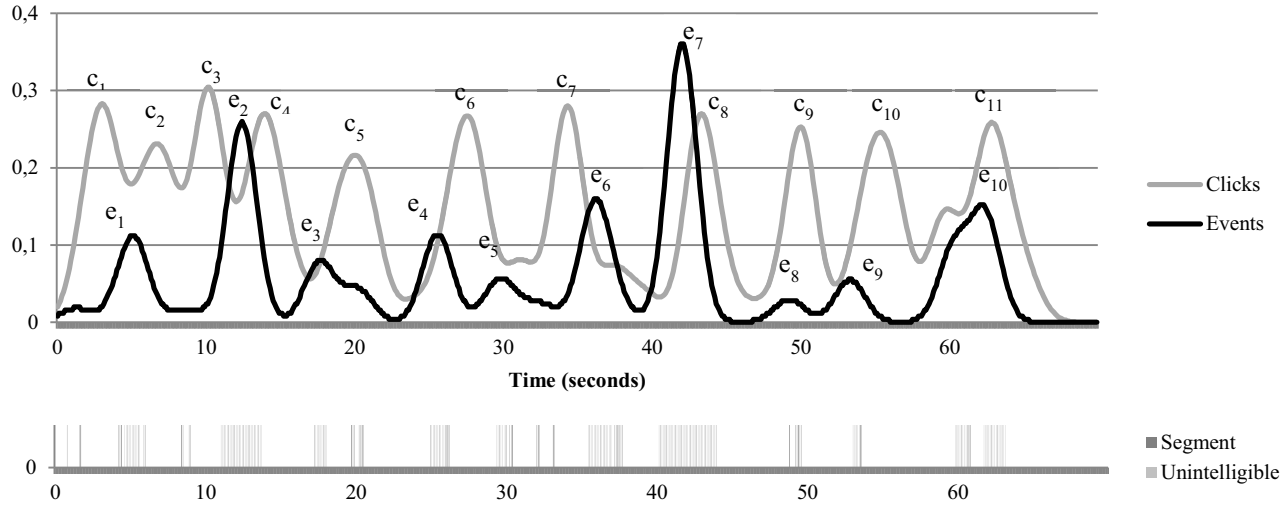

Figure 2. *Top panel: Kernel density estimate curves for listener clicks and speech events in speech sample 4. Bottom panel: Raw distribution of annotated speech events, split into isolated segments and stretches of unintelligible speech.*

-rexes) at about one second before the click peak. Considering that the word "t-rexar" (t-rexes) probably is self-invented, and that it is not preceded by contextual cues revealing that dinosaurs is the topic of the discussion, it is likely that many listeners perceived the word as unintelligible, which would explain the peak c_3. The fact that the annotated event e_3 (a case of /r/-weakening) is not followed by a peak with corresponding amplitude in the click curve, indicates that this deviation was less disturbing to the listeners.

A different pattern can be observed in Figure 2, which illustrates speech events and listener clicks for sample 4 (see Table 2). As expected, there are more peaks in both curves, and their interconnectedness is therefore less obvious. However, despite of their difference in amplitude, nearly all event peaks tightly precede a peak in the click curve. The peaks e_5 and e_6 deviate from this pattern, however, as neither of them is clearly associated with c_7. As indicated in the bottom panel, there are segmental problems at about 1 second before c_7, and a possible interpretation is that these events might have evoked the clicks represented by c_7, whereas the stretch of speech annotated as unintelligible was less disturbing to the listeners.

4. Discussion

We have presented the application of an ARS-based method of evaluating intelligibility to the domain of deviant and typical child speech. The results have been validated against a standard clinical measure of severity, revealing a strong correlation between the two. We have demonstrated the potential in the ARS-based method to highlight places of interest in the speech signal, in this case the speech deviance patterns that are most deteriorating for intelligibility.

As the PCC measure can only be estimated for words that are understood, unintelligible words and passages are not included in the PCC score. Hence, the PCC scores of the most unintelligible children may be artificially high. At the same time, unintelligible passages are likely to evoke a large number of clicks. The strong correlation between the PCC score and number of clicks might therefore be underestimated.

Several listeners commented that they found the task difficult, and that they felt as if they had clicked entirely at random. However, the patterns observed in the visual inspection of the distribution curves of clicks and speech events indicate a more structured pattern. Although these findings are preliminary, considering the small amount of data analyzed, and the uncertainties surrounding the length of the lapse between speech event and listener response, there are some conclusions to be drawn from the qualitative analysis of the correlation between speech events and click responses. First, when the speech signal contains few events, they can quite easily be linked to peaks in the click distribution. However, when speech events are more frequent, or more distributed in time, there will be more variation in the click distribution, and therefore it will be more difficult to identify what event evoked which peak. Second, when the instructions to the listeners are vague, e.g. "Click whenever you hear something you do not understand or something you perceive as deviant", it will be difficult to interpret what the distribution of clicks signifies. This shortcoming will be addressed in future work.

Examination of the relative contribution to decreased intelligibility of different speech production problems would address the paucity of established norms in this area. By including untrained listeners in the evaluation of intelligibility, rapid collection of data would be enabled – data representing a simulation of the situation where children talk to unfamiliar people. This prospect could constitute a valuable contribution to the increasing interest in functional intelligibility.

5. Acknowledgements

This work was funded partly by The Swedish Graduate School of Language Technology, and partly by the Promobilia foundation. SLT students Ida Andersson and Catarina Lång performed recordings and speech assessments. The web-based platform for the ARS-test was provided by Södermalms Talteknologiservice (STTS). Jens Edlund produced the KDE curves.

6. References

[1] J. Kwiatkowski and L. D. Shriberg, "Intelligibility Assessment in Developmental Phonological Disorders: Accuracy of Caregiver Gloss," *Journal of Speech and Hearing Research,* vol. 35, pp. 1095-1104, 1992.

[2] J. Coplan and J. R. Gleason, "Unclear Speech: Recognition and Significance of Unintelligible Speech in Preschool Children," *Pediatrics,* vol. 82, pp. 447-452, 1988.

[3] M. Gordon-Brannan and B. Williams Hodson, "Intelligibility/Severity Measurements of Prekindergarten Children's Speech," *American Journal of Speech-Language Pathology,* vol. 9, pp. 141-150, 2000.

[4] R. Kent, G. Miolo and S. Bloedel, "The Intelligibility of Children's Speech: A Review of Evaluation Procedures," *American Journal of Speech-Language Pathology,* vol. 3, pp. 81-95, 1994.

[5] P. Flipsen Jr, "Measuring the intelligibility of conversational speech in children," *Clinical Linguistics and Phonetics,* vol. 20, no. 4, pp. 303-12, 2006.

[6] N. Schiavetti, "Scaling procedures for the measurement of speech intelligibility," in *Intelligibility in speech disorders*, Philadelphia, John Benjamins, 1992, pp. 11-34.

[7] L. D. Shriberg and J. Kwiatkowski, "Phonological Disorders III: A Procedure for Assessing Severity of Involvement," *Journal of Speech and Hearing Disorders,* vol. 47, pp. 256-270, 1982.

[8] S. McLeod, L. J. Harrison and J. McCormack, "The Intelligibility in Context Scale: Validity and Reliability of a Subjective Rating Measure," *Journal of Speech Language and Hearing Research,* vol. 55, pp. 648-656, 2012.

[9] R. B. Monsen, "The oral speech intelligibility of hearing-impaired talkers," *Journal of Speech and Hearing Disorders,* vol. 48, pp. 286-296, 1983.

[10] M. Walshe, N. Miller, M. Leahy and A. Murray, "Intelligibility of dysarthric speech: perceptions of speakers and listeners," *International Journal of Language & Communication Disorders,* vol. 43, no. 6, pp. 633-348, 2010.

[11] J. Edlund, A. Hjalmarsson and C. Tånnander, "Unconventional methods in perception experiments," in *Proceedings of Nordic Prosody XI*, Tartu, Estonia, 2012.

[12] K. Sjölander and J. Beskow, "WaveSurfer - an open source speech tool," in *Proceedings of ICSLP 2000, 6th Intl Conf on Spoken Language Processing*, Beijing, China, 2000.