Real-time registration of listener reactions to unintelligibility in misarticulated child speech

Ivonne Contardo[1], *Anita McAllister*[1], *Sofia Strömbergsson*[2]
[1] *Division of Speech-Language Pathology, CLINTEC, Karolinska Institutet, Sweden*
[2] *KTH Speech, Music and Hearing, Stockholm, Sweden*
ivonne.contardo@student.ki.se, anita.mcallister@ki.se, sostr@kth.se

Abstract

This study explores the relation between misarticulations and their impact on intelligibility. 30 listeners (17 clinicians and 13 untrained listeners) were given the task of clicking a button whenever they perceived something unintelligible during playback of misarticulated child speech samples. No differences were found between the clinicians and the untrained listeners regarding clicking frequency. The distribution of listener clicks correlated strongly with the clinical evaluations of the same samples. The distribution of clicks was also related to manually annotated speech errors, allowing examination of links between events in the speech signal and reactions evoked in listeners. Hereby, we demonstrate a viable approach to ranking speech error types with regards to their impact on intelligibility in conversational speech.

Introduction

Children with speech disorders often present with systematic speech error patterns. As a communicative consequence, intelligibility is often reduced. For these children, as well as for younger children following a typical course of speech acquisition, communication is especially affected when interacting with people they do not already know (Coplan & Gleason, 1988; Kwiatkowski & Shriberg, 1992).

Speech intelligibility is an important consideration in many clinical decisions; however, it is not trivially assessed. Standard measures of intelligibility may be questioned with respect to their functional relevance (McLeod, Harrison, & McCormack, 2012). For one, the child's ability to produce isolated words in a clinical setting may not be very representative of how well he or she is understood when communicating with unfamiliar people. Second, no currently available clinical measure of intelligibility exposes what causes reductions of intelligibility. In order to link levels of (un)intelligibility to features in speech production, intelligibility assessments need to be complemented with assessments of the type and degree of speech impairment.

The most widely used metric of severity of speech disorders is the Percentage of Consonants Correct (PCC; Shriberg & Kwiatkowski, 1982). 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 (*ibid*.), the PCC metric is associated with some limitations, e.g. relating to its application to highly unintelligible speech. And although it may seem 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 is weak (Kwiatkowski & Shriberg, 1992). Hence, linking speech production problems to levels of intelligibility requires alternative approaches.

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 unfamil-

iar people (Kwiatkowski & Shriberg, 1992), and clinicians have been found to evaluate misarticulated speech as more intelligible compared to untrained listeners (Lundeborg & McAllister, 2007; McGarr, 1983). In order to understand how clinical evaluations reflect the child's everyday communicative challenges, there is value in calibrating clinical evaluations against evaluations of other listeners.

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 (Edlund, Hjalmarsson, & Tånnander, 2012). Applying the ARS-based method to recordings of misarticulated speech presents itself as an interesting opportunity. First, the method allows for fast collection of ratings from many listeners, thus strengthening the reliability of the ratings. Second, recruiting untrained listeners as raters gives an indication of the extent of the everyday problems that 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 to listeners. If coupled with qualitative speech analysis, this information can go far beyond standard measures of intelligibility/severity.

Strömbergsson & Tånnander (2013) describe a first exploration of applying the ARS method to the domain of misarticulated speech. However, despite demonstrating the potential of the ARS method as an instrument for identification of features in children's speech that are most detrimental to intelligibility, the study is limited in several respects. First, the instructions provided to the listeners were unclear, thus restricting the interpretation of the listeners' responses. Second, the qualitative analysis was very limited, which precluded any conclusions regarding what specific speech error types evoke the most reactions in listeners. The present study aims to address these issues.

Research questions

In the present study, the following research questions are addressed:

1. Are there any differences between SLPs and untrained listeners in their real-time reactions to unintelligibility in samples of misarticulated child speech?
2. To what extent do real-time reactions to unintelligibility reflect results of standard clinical measures?
3. To what extent do specific speech errors contribute to decreased intelligibility?

Method

Conversational speech was recorded from 7 preschool-aged Swedish children exhibiting speech production deficits. Speech production was assessed by means of LINUS (Blumenthal & Lundeborg Hammarström, 2013); speech production characteristics are summarized in Table 1. Parental evaluation of intelligibility was assessed by means of the Intelligibility in Context Scale (ICS) (McLeod et al., 2012).

In the recording situation, the children and an adult (an SLP 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. Sequences of continuous child speech were extracted manually from the child-adult conversations, and sequentially concatenated to form one-minute long speech samples. In all, 11 such speech samples were combined into a listening script, with one sample serving as an introductory item, and excluded from analysis.

For the 10 conversational speech samples, the Percentage of Consonants

Table 1. Descriptions of the recorded children. Age is given as years;months.

ID	Age	Speech errors
A	3;8	Added voicing, velar fronting, stopping, /r/-weakening, cluster reductions, omission of /ç/ and /h/, assimilations and metatheses, /e/ not established.
B	5;0	/r/-weakening, stopping, omission of /s/, cluster reductions, assimilations.
C	6;1	/r/-weakening, stopping, /m/ → [b], /n/ → [d], /ɧ/ → [s], /ŋ/ → [g].
D	6;0	/r/-weakening.
E	5;6	Velar fronting, cluster reductions.
F	5;8	/ʈ/ → [t], cluster reductions.
G	5;6	/r/-weakening, /l/-weakening, stopping, labialization, cluster reductions, assimilations, epentheses, /b/ → [v], /e/ not established.

Correct (PCC) was calculated along the procedures described in Shriberg & Kwiatkowski (1982), by two independent experimenters. Inter-judge reliability between the experimenters was .96 (Crohnbach's alpha). For each sample, the average of the two judges' PCC measures served as the final PCC measure for that particular sample.

All 10 conversational speech samples were also subject to qualitative analysis. Here, the first author used Wavesurfer (Sjölander & Beskow, 2000) to mark and label all speech errors occurring in the samples. Stretches of unintelligible speech were assigned the label "unintelligible", and typically ranged across several words. From the resulting timestamps, the midpoints of the events were passed on to further analysis.

30 adults participated in the ARS-based listening test; their age varied between 25 and 61 (M = 35.80, SD = 9.56). The gender distribution was 10:20 (male:female). 13 of the listeners were SLPs, all with experience of working with children. There was no difference between the SLPs and the 17 untrained listeners with regards to age: $t(28) = .36$, $p = .72$). The SLPs' working experience ranged between 8 months and 23 years (M = 9.03, SD = 8.10).

The 10 conversational speech samples were randomized for each listener, and implemented in a web-based ARS listening test. The listeners were instructed to listen to the speech samples and to click any keyboard key (or mouse key) whenever they perceived something unintelligible during playback. All listener clicks were registered during playback. The average number of clicks over all listeners and all speech samples was used in the weighting of the listeners' clicks, so that clicks from listeners who do not click very often are given more weight than clicks from listeners who click more frequently.

The distribution of the weighted clicks was analyzed by means of Kernel Density Estimations (KDEs). The analysis resembles a histogram, but the produced curve is continuous and smooth. For each recording, the distribution of clicks was linked to the manually annotated speech errors; if a KDE peak was found within an interval of 500-1400 ms after an annotated speech error, this assembly of listener clicks was considered to reflect a reaction to that specific speech error.

Results

Potential differences between the two listener groups in their clicking behavior were explored by means of a one-way ANOVA, with total number of clicks as the dependent variable, and listener group (SLPs vs. untrained lis-

teners) as an independent variable. This analysis revealed no difference between the groups: $F(1,28) = .13$, $p = .72$.

A Pearson correlation analysis was used to explore the relationship between the PCC and the number of (weighted) clicks per recording, revealing a strong negative correlation between the two: $r(10) = -.91$, $p < .001$. The inverted correlation between PCC and the number of (weighted) clicks per recording is indicated by the figures in Table 2. In order to explore the relation between intelligibility ratings as assessed by the ICS and the PCC on the one hand, and between ICS and the number of (weighted) clicks per recording on the other hand, two separate Pearson correlation analyses were performed. No correlation was found between the children's ICS scores and the PCC scores [$r(10) = .55$, $p = .10$], nor between ICS scores and the number of (weighted) clicks per recording [$r(10) = -.38$, $p = .29$].

Table 2. Evaluation results for all recordings, with regards to the PCC and (weighted) listener clicks.

Rec	Child	ICS	PCC	Clicks
1	A	3.7	66%	10.5
2	A	3.7	61%	11.7
3	B	3.6	77%	3.7
4	B	3.6	83%	2.4
5	C	4.4	89%	1.1
6	D	4.7	96%	1.1
7	D	4.7	91%	2.9
8	E	4.0	88%	3.3
9	F	3.6	99%	0.7
10	G	3.3	70%	5.4

The extent to which different speech error types evoked listener reactions are listed, for all error types, in Table 3. As revealed in the table, assimilation and added voicing often evoke reactions in listeners, whereas errors like metathesis and syllable omission do not appear as destructive to intelligibility.

Table 3. The speech error types evidenced in the recorded data, together with the number of times they are followed by a KDE peak (interpreted as a listener reaction).

Speech error	Freq.	Evoked reactions	% of instances followed by reaction
Assimilation	10	5	50%
Unintelligible*	16	8	50%
Added voicing	4	2	50%
/r/-weak. + final cons. deletion	3	1	33%
Velar fronting	26	7	27%
Stopping	27	7	26%
/r/-weakening	79	19	24%
Final consonant deletion	19	4	21%
Omission	26	5	19%
Cluster reduction	27	5	19%
Other	33	6	18%
Cluster red. + velar fronting	6	1	17%
Vowel error	26	4	15%
Assimilation + devoicing	1	0	0%
Syllable omission	3	0	0%
Cluster reduction + /r/-weak.	1	0	0%
Metathesis	9	0	0%
/ɕ/-error	1	0	0%
Devoicing	1	0	0%

* Stretches of speech labeled as unintelligible by the annotator.

Discussion

This study has presented an application of an ARS-based method of evaluating intelligibility to the conversational samples of misarticulated child speech. The comparison between listeners with professional experience with misarticulated child speech and untrained listeners, with regards to their reactions to unintelligibility, revealed no difference between the groups. The results of the ARS-evaluation were validated against a standard clinical measure of severity (the PCC), revealing a strong correlation between the two. By linking annotated misarticulations to the distribution of listener reactions, we have demonstrated the potential in the ARS-based method to rank different types of speech errors (or, for that matter, any episodic speech phenomena) by their impact on, in this case, intelligibility.

Given observations that the correlation between the severity of the speech impairment and intelligibility is weak (e.g. Shriberg & Kwiatkowski, 1982), the lack of correlation between the ICS and the PCC measures is not surprising. However, the lack of correlation between the results of the listener evaluation and the ICS measure requires commenting. This may reflect the fact that perceived intelligibility varies across situations, and that a one-minute recording of a one-to-one conversation in a quiet room risks not being representative of general everyday situations.

Just as in (Strömbergsson & Tånnander, 2013), no difference was found between experienced clinicians and untrained listeners in their evaluations of intelligibility. This contradiction to earlier findings (Lundeborg & McAllister, 2007; McGarr, 2011), may be due to the nature of the speech material (conversational speech vs. isolated words), or to the nature of the misarticulations (primarily phonological errors vs. the speech of a child with apraxia of speech or deaf children).

A limitation concerns the uncertainty tied to the determination of the time window where listener reactions are sought, in the process of linking speech events to listener reactions. By using a relatively long time window, and by allowing listener reactions to be interpreted as having been evoked by more than one speech event, the risk of overlooking existing connections is minimized. This, however, is at the expense of precision, which may lead to the identification of links that are not actually there. In future work, these decisions may need refinement.

Many factors contribute to variations in intelligibility. Focus in the present study has been specific segmental speech errors, whereas other aspects of the speech material have been disregarded. In order to control for the influence of other factors – e.g. lexical, syntactic, prosodic or pragmatic factors – using a more restricted speech material, and/or including more material, from more speakers, should be considered.

Much work remains to arrive at firm conclusions on how specific speech errors contribute to decreased intelligibility. However, the present study constitutes an important step, in describing a viable method for collecting norms in this area. By integrating such information in the prioritizing of clinical targets, intervention may be better directed at those patterns that cause the most problems for children in their everyday lives.

Conclusions

We have demonstrated the potential of applying an ARS-based method to the domain of misarticulated child speech, to explore the relative contribution of different speech errors to perceived (un)intelligibility. Although more data – in terms of more speakers and broader coverage of speech error types – is required to allow general conclusions regarding the impact of different speech errors on intelligibility, the paucity of established norms in this area strongly motivates continued efforts in this direction.

Acknowledgements

The web-based platform for the ARS-test was provided by Södermalms Talteknologiservice (STTS). Jens Edlund produced the KDE curves.

References

Blumenthal, C. & Lundeborg Hammarström, I. (2013). *LINUS preliminärmanual från februari 2014*. Sweden: Dept. of Neuroscience and Locomotion/Speech Pathology, Linköping University.

Coplan, J. & Gleason, J. R. (1988). Unclear speech: Recognition and significance of unintelligible speech in preschool children. *Pediatrics*, *82*(3), 447–452.

Edlund, J., Hjalmarsson, A., & Tånnander, C. (2012). Unconventional methods in perception experiments. In *Proceedings of Nordic Prosody XI*. Tartu, Estonia.

Kwiatkowski, J. & Shriberg, L. D. (1992). Intelligibility assessment in developmental phonological disorders: Accuracy of caregiver gloss. *Journal of Speech and Hearing Research*, *35*(5), 1095–1104.

Lundeborg, I. & McAllister, A. (2007). Treatment with a combination of intra-oral sensory stimulation and electropalatography in a child with severe developmental dyspraxia. *Logopedics Phoniatrics Vocology*, *32*(2), 71–79.

McGarr, N. S. (1983). The intelligibility of deaf speech to experienced and inexperienced listeners. *Journal of Speech and Hearing Research*, *26*(3), 451–458.

McLeod, S., Harrison, L. J., & McCormack, J. (2012). The Intelligibility in Context Scale: Validity and reliability of a subjective rating measure. *Journal of Speech Language and Hearing Research*, *55*(2), 648–656.

Shriberg, L. D. & Kwiatkowski, J. (1982). Phonological disorders III: A procedure for assessing severity of involvement. *Journal of Speech and Hearing Disorders*, *47*(3), 256–270.

Sjölander, K. & Beskow, J. (2000). WaveSurfer - an open source speech tool. In B. Yuan, T. Huang, & X. Tang (Eds.), *Proceedings of ICSLP 2000, The 6th Intl Conf on Spoken Language Processing* (pp. 464–467). Beijing, China.

Strömbergsson, S. & Tånnander, C. (2013). Correlates to intelligibility in deviant child speech – comparing clinical evaluations to audience response system-based evaluations by untrained listeners. In *Proceedings of Interspeech 2013* (pp. 3717–3721). Lyon, France.