Ranking severity of speech errors by their phonological impact in context

Sofia Strömbergsson[1], *Christina Tånnander*[2], *Jens Edlund*[1]

[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, edlund@speech.kth.se

Abstract

Children with speech disorders often present with systematic speech error patterns. In clinical assessments of speech disorders, evaluating the severity of the disorder is central. Current measures of severity have limited sensitivity to factors like the frequency of the target sounds in the child's language and the degree of phonological diversity, which are factors that can be assumed to affect intelligibility. By constructing phonological filters to simulate eight speech error patterns often observed in children, and applying these filters to a phonologically transcribed corpus of 350K words, this study explores three quantitative measures of phonological impact: Percentage of Consonants Correct (PCC), edit distance, and degree of homonymy. These metrics were related to estimated ratings of severity collected from 34 practicing clinicians. The results show an expected high correlation between the PCC and edit distance metrics, but that none of the three metrics align with clinicians' ratings. Although these results do not generate definite answers to what phonological factors contribute the most to (un)intelligibility, this study demonstrates a methodology that allows for large-scale investigations of the interplay between phonological errors and their impact on speech in context, within and across languages.

Index Terms: speech disorders, intelligibility, child speech

10.21437/Interspeech.2014-374

1. Introduction

Children with speech sound disorders (SSDs) often exhibit systematic errors, affecting groups of sounds or sound patterns. For example, a child with SSD can display patterns of stopping of fricatives, velar fronting, or consonant cluster reductions [1]. Children with SSD may exhibit only one error pattern, but more often, children display combinations of speech errors [2]. In speech-language therapy, these different types of errors need to be prioritized. Considering that the ultimate goal of speech-language therapy is to optimize communicative functioning, it makes sense to focus intervention on those speech error patterns that have the most detrimental effects on communication.

In clinical descriptions of SSDs, "severity" is a central concept, referring to the degree to which the speech disorder affects the child's communication skills [3]. The most commonly used metric of severity is the Percentage of Consonants Correct (PCC) [4]. However, there are several limitations associated with using the PCC as an index of severity. For example, the measure is based on binary evaluations of correct/incorrect production of consonants, and therefore not sensitive to different types and degrees of distortions/errors [5]. Although variants of the PCC metric addressing this limitation have been suggested [6], together with alternative measures like the Weighted Speech Sound Accuracy (WSSA) score [5], no measure to date is based on statistically motivated weighting of speech errors. For example, although it may make intuitive sense to weight deletions or additions of weak segments (e.g. glottals and glides) half as much as errors involving strong segments (e.g. orally articulated consonants and vowels) as in [5], both the classification and the weighting are arbitrary. Moreover, existing measures of severity have almost exclusively been evaluated with regards to English; hence, they are not necessarily directly applicable to other languages.

Misarticulation of sounds that occur frequently has more pervasive effects than misarticulation of sounds that are less frequent [7]. Moreover, the types of words that are affected also plays a role; misarticulation of content words has more detrimental effects on intelligibility than misarticulation of function words [8]. Hence, measures of severity should also be sensitive to what types of words are affected by speech errors. Another potential source of diminished intelligibility is neutralization of phonological contrasts, or homonymy [9]. This is apparent in cases where, for instance, a child's production of the word *key* is indistinguishable from his or her production of the word *tea*. Hence, a metric of severity should also be sensitive to the extent to which the speech error(s) cause homonymy. The phonological distance between a speech target and another version – e.g. a dialectal version, or a misarticulated version – of the same word may be measured by means of the Levenshtein metric [10]. Although it shares several features with the PCC metric, the interrelation between PCC and phonological distance as measured by the Levenshtein metric has not yet been described. By considering the relative impact of each of these factors, and of all of them combined, currently used measures of severity may be extended to better reflect aspects of the impact of specific speech errors in a communicative context.

The severity of speech disorders is closely related to intelligibility. Although the relation between different types of speech errors and their impact on intelligibility has been studied, it is not yet fully understood; suggestions of phonetic/phonological correlates to (un)intelligibility are often not based on empirical evidence [11]. However, by exploring the distribution of different speech errors across children grouped by level of intelligibility, indirect links between different speech errors and levels of intelligibility have been observed [12]. This way, omissions of speech sounds have been concluded as being more damaging to intelligibility than phonetic distortions [12]. More direct links between errors and (un)intelligibility may be described by simulating error patterns, and exploring outcomes with regards to intelligibility. A rare example of using this approach is described in [11], where three different speech error patterns – final consonant deletion, stopping, and velar fronting – were ranked with regards to their impact on intelligibility. By applying three different "phonological filters" (each representing one speech error pattern) to a phonetically transcribed text, speech errors were simulated and read aloud by an adult male speaker. Although the results from this study – e.g. that final consonant deletion has the most detrimental effects on intelligibility – have important clinical implications, the ecological validity in

having an adult male producing speech errors from transcribed texts may be questioned. However, using phonological filters to represent speech errors is still a viable approach to studying the impact of different speech errors in context, particularly for large-scale text-based studies.

There is an apparent value in knowing how specific speech error patterns contribute to decreased intelligibility, as therapy targeting those error patterns that are most detrimental to intelligibility will potentially be most rewarding in terms of functional gains. Moreover, by relating objective measures of severity to practicing clinicians' intuitive estimates of severity, the measures may be compared with regards to how they reflect clinical intuition on severity – thus examining their construct validity. This study constitutes the first report of such a comparison, by addressing the following questions:

1. How does the PCC relate to two other measures describing the impact of speech errors: the degree of homonymy and Levenshtein distance?
2. How are different speech errors ranked by their impact on severity by these different metrics?
3. How do these ranks compare to practicing clinicians' intuitive rankings of how different speech errors affect intelligibility?

2. Method

2.1. Data

With the aim of gathering text material as representative of children's speech production as possible, reflecting expected vocabularies in preschool-aged children, a corpus of children's books was collected from Språkbanken[1]. Two versions of this corpus were used: the full corpus ($Corpus_{full}$), and a version where all frequent function words (occurring among the 100 most common words) had been excluded ($Corpus_{content}$). Word statistics for both corpora are presented in Table 1.

Table 1. *Word statistics describing the two corpora.*

Corpus	# word tokens	# word types	Type/token ratio
$Corpus_{full}$	351 094	18 962	.054
$Corpus_{content}$	192 957	18 904	.098

Phonological transcriptions of all words in the corpora, available from the Swedish Agency for Accessible Media, were on a broad level, reflecting frequent reductions in colloquial speech. For example, the final consonant of the word "och" (Eng. *and*) and of adjectives ending in "-ig" (e.g. "gullig", Eng. *cute*) is omitted.

2.2. Procedure

Based on a description of speech error patterns observed in Swedish children with PD [13], eight error patterns were selected for analysis. This selection was restricted to context-independent speech error patterns that could be expressed as paradigmatic substitutions. Context-dependent patterns like assimilations and metatheses were not included in the analysis. Thus, the included patterns could all be straight-forwardly represented as phonological filters. Table 2 presents a list of the filters/error patterns examined.

Table 2. *The speech error patterns examined, as expressed as phonological filters.*

Speech error	Substitution(s)
Stopping	Fricatives substituted by stops, while retaining (approx.) place of articulation. E.g. /s/ → /t/, /f/ → /p/.
Cluster reductions	Syllable-initial consonant sequences reduced, a) In cases of CC, where one C is a plosive, the other is omitted. E.g. /pl/ → /p/, /st/ → /t/ b) In other cases of CC, where the first is /s/, /s/ will be omitted. Eg. /sn/ → /n/, /sl/ → /l/ c) In other cases of CC, where the second is /l/, /r/ or /j/, the second C will be omitted. E.g. /fl/ → /f/, /mj/ → /m/ d) In cases of CCC, the last C is omitted. E.g. /str/ → /st/
/h/-zation	Initial voiceless consonants (and /r/) substituted by /h/.
Backing	/t, d, n/ → /k, g, ŋ/
Fronting	/k, g, ŋ/ → /t, d, n/
Labialization	/ɧ/ → /f/
/r/-weakening	/r/ → /j/
/s/-distortion	/s/ → /θ/

Target transcriptions representing expected (correct) production of all words in the corpus were available in the lexicon. Different versions of these transcriptions, each representing the expected production assuming a specific speech error type, were generated by constructing "phonological filters" and applying these to the target transcriptions. For example, when the target transcription is passed through the phonological filter representing velar fronting (substitution of [t, d, n] for the targets [k, g, ŋ]), a filtered transcription is generated, representing the expected pronunciation of a child exhibiting consistent velar fronting. The eight filters were applied to both corpora.

2.3. Effect measures

Three different metrics were used, each representing a different aspect of the phonological effect of applying a specific phonological filter to the target transcription:

1. The Percentage of Consonants Correct (PCC): the proportion of consonants that are not produced in error. Here, this refers to the proportion of consonants that are unaffected by the application of the phonological filters.

[1] A subset of the corpus Läsbart (available from http://spraakbanken.gu.se/eng/resource/lasbart), where only children's literature was included.

2. Degree of homonymy (HOMONYMY): the proportion of words that share the same phonological transcription as at least one other word. A low degree of homonymy corresponds to high phonological diversity.
3. Overall phonological distance (EDITDIST): the average Levenshtein (or edit) distance (i.e. the minimum number of insertions, deletions or substitutions required to transform one string into another) between target transcriptions and error transcriptions, across all word tokens in the corpus.

2.4. Clinical survey

34 speech-language pathologists (SLPs) were recruited to provide their estimated evaluations of the impact of each of the error types on intelligibility. The participants all reported having at least 5 years of clinical experience with childhood speech and language disorders. Responses were collected by means of a web form, in which the SLPs rated the eight speech error types with regards to a 5-step scale, where 0 indicates "no impact" and 5 indicates "severe impact". Inter-rater reliability was estimated by means of an intra-class correlation, which revealed a high level of consistency: ICC(2,34) = .98 (95% CI: .962-.996).

3. Results

3.1. Relation between metrics

In order to explore the correlation between the three different effect metrics, three separate Pearson's correlations were conducted – one for each pair of metrics. When estimated on Corpus_{full}, these analyses showed a strong negative correlation between PCC and EDITDIST ($r = -.99$, $p < .001$), whereas HOMONYMY was found to correlate neither with PCC ($r = -.14$, $p = .74$), nor with EDITDIST ($r = .26$, $p = .53$). Although the numeric details were not identical, the same correlation pattern was observed also for $\text{Corpus}_{content}$. Table 3 displays the outcomes of the different effect measures for all speech error patterns, for both corpora. As indicated in the table, the speech errors are ranked similarly across the corpora, although the numeric details are not identical.

Table 3. *Phonological impact of the eight error patterns, as measured by the three different metrics.*

	Corpus_{full}			$\text{Corpus}_{content}$		
Error	PCC	EDIT	HOM	PCC	EDIT	HOM
Backing	66	1,3	0,02	69	1,0	0,02
Stopping	81	1,0	0,05	80	0,8	0,05
/h/-zation	86	0,9	0,11	86	0,8	0,11
/r/-weak.	88	0,8	0,01	86	0,7	0,01
/s/-dist.	90	0,8	0,01	89	0,7	0,01
Fronting	91	0,7	0,02	89	0,7	0,02
ClustRed.	92	0,7	0,04	90	0,7	0,04
Labial.	100*	0,5	0,01	99	0,5	0,01

* Here, the PPC value 99.598 is rounded up to 100.

In accordance with the results of the correlation analyses, the numbers in Table 3 illustrate the close inverse relation between PCC and EDITDIST, and that the HOMONYMY metric yields a more disparate ranking of speech errors. Three separate Pearson's correlation analyses (one per metric) showed strong significant correlations across the two corpora, all three being $r(8) > .98$, $p < .001$. However, three paired-samples t-tests (one per metric) revealed corpus-dependent differences for EDITDIST: $t(7) = 3.69$, $p < .01$ and for HOMONYMY: $t(7) = 17.36$, $p < .001$, such that phonological impact was generally higher in Corpus_{full} than in $\text{Corpus}_{content}$. For the PCC metric, no difference between the corpora were found: $t(7) = 1.22$, $p = .26$.

3.2. Clinical survey

Figure 1 displays the results of the clinical survey, where SLPs rated the different error patterns by their impact on intelligibility. A Kruskal-Wallis analysis exploring the variation in rating across the different speech error patterns showed a significant dependence: $\chi^2(6, N = 272) = 118.44$, $p < .001$. Multiple pairwise comparisons revealed significant differences between all error patterns except for those indicated with "n.s." in Figure 1.

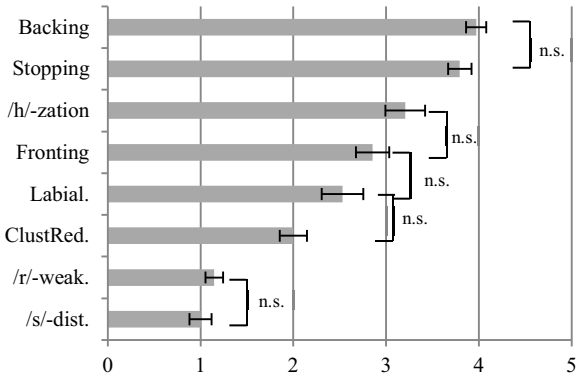

Figure 1: *Impact on intelligibility, for each speech error pattern, as estimated by practicing clinicians with reference to a scale from "no impact" (0) and "severe impact" (5). Error bars represent 1 S.E.*

3.3. Metrics vs. clinical survey

A linear regression was conducted in order to explore whether any of the three metrics could predict the clinical estimates of impact on intelligibility. This analysis showed that none of the three metrics significantly predicted the clinical estimates, neither alone, nor in combination with any of the other two. This was found for both corpora. Figure 2 illustrates how the rankings of the three effect metrics relate to the ranking of the clinical estimates, for $\text{Corpus}_{content}$. Clinicians' estimated ranking is indicated by the ordering of the error patterns along the y-axis, ranging from the most severe errors on top, to the least severe errors at the bottom. Ranking on the x-axis is given as the ratio to the maximum values observed for PCC, EDITDIST and HOMONYMY, respectively.

If an impact metric corresponded well to clinical estimates of severity, a decreasing pattern from top to bottom would be expected in figure 2. However, in congruence with the results of the linear regression, figure 2 shows that neither of the impact measures matches the clinical estimates of severity very closely. For example, in relation to the clinical estimates, both the PCC metric and the EDITDIST metric overestimate the severity of /r/-weakening and /s/-distortions. In the same vein, the HOMONYMY metric overestimates /h/-zation, while under-

estimating the effect of backing. Moreover, labialization is ranked by all three metrics as a less severe error patterns compared to clinicians' estimates.

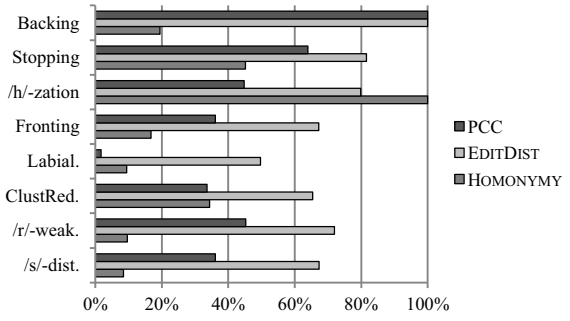

Figure 2: *Ranking of impact, for each speech error pattern, presented as the ratio to the maximum value for PCC, EDITDIST and HOMONYMY, respectively. Clinicians' estimates of severity are indicated by the ordering along the y-axis, from the highest impact on top, to the lowest impact at the bottom.*

4. Discussion

We have presented an investigation of the severity of different speech error patterns, as measured by their phonological impact on speech in context. By simulating speech error patterns often observed in children with speech sound disorders, and by investigating how these error patterns are ranked by three objective measures of phonological change, the interrelations between these metrics were explored, as well as how each of them relates to experienced clinicians' estimates of how the same speech errors affect intelligibility. Two of the measures, PCC and EDITDIST, were found to be highly intercorrelated, whereas the third, HOMONYMY, shared little resemblance with the other two. Considering that PCC and EDITDIST are both based on character-by-character comparisons between target transcriptions and error transcriptions, their close interrelation is not surprising. Hence, these measures both represent the average phonological distance between target and error productions, which has been suggested as one factor contributing to perceived (un)intelligibility [14]. However, the fact that neither PCC nor EDITDIST predict clinicians' intuitive ratings of how different speech errors affect intelligibility, accords with established knowledge that many other factors also affect intelligibility.

The observation that phonological impact was generally higher in Corpus_{full} than in $\text{Corpus}_{content}$ indicates that function words, by virtue of occurring frequently, pose a stronger influence on measures of phonological effects. However, although the effect measure values are generally higher in Corpus_{full}, their relative rankings of speech errors were observed not to be different from those found for $\text{Corpus}_{content}$. Hence, the examined speech errors affect function words very similarly to how they affect content words. However, considering that misarticulation of function words is less deteriorating for intelligibility than misarticulation of content words, the measures as applied to $\text{Corpus}_{content}$ can be assumed to better reflect effects on intelligibility.

To the authors' knowledge, the present study constitutes a first report of clinical estimates of how different speech errors affect intelligibility for Swedish. Considering the high agreement between raters, this may be considered a reliable benchmark to which quantitative measures of severity may be compared. In the present report, none of the three measures reflect clinical estimates very closely. For one, PCC and EDITDIST overestimate the effects of "light" problems like /r/-weakening and /s/-distortions. (This limitation of the PCC has been reported earlier, e.g. [6].) In this respect, the HOMONOMY metric reflects clinical estimates better. On the other hand, the HOMONYMY metric overestimates the severity of cluster reductions and /h/-zation. The fact that not even the combination of the three metrics predicts the clinical estimates indicates that other factors need to be included in the model.

Although all error patterns examined in the present have also been reported for other languages (e.g. English [15] and Dutch [16]), their phonological impact on speech in context will vary with phonotactic characteristics tied to different languages. Therefore, a ranking of speech errors by severity for one language will not necessarily be valid for other languages. However, the method described may easily be applied to other languages, allowing for exploration of cross-linguistic differences in phonological impact of speech errors.

Using children's literature as a proxy for children's expected speech production is a choice that deserves motivation. Not only are linguistic differences expected between spoken and written modes of language, one would also expect different linguistic features in texts written for children, compared to what the children themselves would produce. However, in lack of access to a similarly-sized corpus of children's spoken utterances, this option was considered the best available for the present investigation.

The idea of constructing phonological filters to simulate speech errors involves some limitations that should be considered. One aspect is that of describing speech errors with reference to categorical labels, even though acoustic evaluation of speech errors often reveals more gradual differences between targets and errors [17]. However, when aiming at ensuring a controlled and linguistically representative setting, analyses require large corpora of transcribed speech, where reduction of phonetic detail is inevitable. Another limitation regards the application of each phonological filter separately and across all word positions; clearly, this is a simplification of how speech errors actually appear in children's speech. Refining the phonological filters, allowing less consistent application, and exploring effects of applying multiple filters, are evident venues for future work.

This investigation constitutes a first step towards large-scale examination of the severity of different speech errors with reference to controlled linguistically representative contexts. Although the phonological filters described may be refined, and although additional measures of phonological impact may be introduced, the report demonstrates the potential of ranking speech errors by their impact in context. Clinically, such quantitative measures may extend current recommendations regarding what specific speech error patterns should be prioritized in therapy with information regarding how the errors affect children's communicative functioning.

5. References

[1] Nijland, L., ”Speech perception in children with speech output disorders,” Clinical Linguistics & Phonetics, 23(3): 222-239, 2009.

[2] Rvachew, S. and Nowak, M., ”The Effect of Target-Selection Strategy on Phonological Learning,” Journal of Speech Language and Hearing Research, 44(3), 610-623, 2001.

[3] Prezas, R. and Hodson, B., ”Diagnostic evaluation of children with speech sound disorders,” in S. Rvachew (Ed.) Encyclopedia of language and literacy development, London, Ontario, Canadian Language and Literacy Research Network, 2007, pp. 1-7.

[4] Shriberg, L. D. and Kwiatkowski, J., ”Phonological disorders III: A procedure for assessing severity of involvement,” Journal of Speech and Hearing Disorders, 47(3), 256-270, 1982.

[5] Preston, J. L., Ramsdell, H. L., Oller, D. K., Edwards, M. L. and Tobin, S. J., ”Developing a weighted measure of speech sound accuracy,” Journal of Speech Language and Hearing Research, 54(1), 1-18, 2011.

[6] Shriberg,, L. D., Austin, D., Lewis, B. A., McSweeny, J. L. and Wilson, D. L., ”The percentage of consonants correct (PCC) metric: Extensions and reliability data,” Journal of Speech Language and Hearing Research, 40(4), 708-722, 1997.

[7] Brown, A., ”Functional load and the teaching of pronunciation,” TESOL Quarterly, 22(4), 593-606, 1988.

[8] Lam, J. and Tjaden, K., ”Intelligibility of Clear Speech: Effect of Instruction,” Journal of Speech Language and Hearing Research, 56(5), 1429-1440, 2013.

[9] Ingram, D., Phonological disability in children, London, UK: Cole and Whurr Ltd, 1989.

[10] Heeringa, W., Measuring dialect pronunciation differences using Levenshtein distance, PhD Thesis, Groeningen, The Netherlands: University of Groeningen, 2004.

[11] Klein, E. S. and Flint, C. B., ”Measurement of intelligibility in disordered speech,” Language, Speech, and Hearing Services in Schools, 37(3), 191-199, 2006.

[12] Hodson, B. W. and Paden, E., Targeting intelligible speech: A phonological approach to remediation, San Diego, CA: College-Hill Press, 1983.

[13] Nettelbladt, U., ”Fonologiska problem hos barn med språkstörning,” in U. Nettelbladt & E.-K. Salameh (Eds.) Språkutveckling och språkstörning hos barn, Lund, Studentlitteratur, 2007, pp. 95-134.

[14] Gooskens, C., Heeringa, W. and Beijering, K., ”Phonetic and lexical predictors of intelligibility,” International Journal of Humanities and Arts Computing, 2(1-2), 63-81, 2008.

[15] Grunwell, P., Clinical Phonology, Baltimore, MD: Williams & Wilkins, 1987.

[16] Beers, M., ”Phonological processes in Dutch language impaired children,” Scandinavian Journal of Logopedics and Phoniatrics, 17, 9-16, 1992.

[17] Tyler, A. A., Figurski, G. R. and Langsdale, T., ”Relationships between acoustically determined knowledge of stop place and voicing contrasts and phonological treatment progress,” Journal of Speech and Hearing Research, 36(4), 746-759, 1993.