INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Intelligibility of Disordered Speech: Global and Detailed Scores

Mario Ganzeboom, Marjoke Bakker, Catia Cucchiarini, Helmer Strik

CLS/CLST, Radboud University, Nijmegen, The Netherlands
{m.ganzeboom, m.bakker, c.cucchiarini, w.strik }@let.ru.nl

Abstract

Measuring the intelligibility of disordered speech is a common practice in both clinical and research contexts. Over the years various methods have been proposed and studied, including methods relying on subjective ratings by human judges, and objective methods based on speech technology. Many of these methods measure speech intelligibility at the speaker or utterance level. While this may be satisfactory for some purposes, more detailed evaluations might be required in other cases such as diagnosis and measuring or comparing the outcomes of different types of therapy (by humans or computer programs). In the current paper we investigate intelligibility ratings at three different levels of granularity: utterance, word, and subword level. In a web experiment 50 speech fragments produced by seven dysarthric speakers were rated by 36 listeners in three ways: a score per utterance on a Visual Analogue and a Likert scale, and an orthographic transcription. The latter was used to obtain word and subword (grapheme and phoneme) level ratings using automatic alignment and conversion methods. The implemented phoneme scoring method proved feasible, reliable, and provided a more sensitive and informative measure of intelligibility. Possible implications for clinical practice and research are discussed.

Index Terms: intelligibility measures, dysarthric speech, speech therapy, automated speech analysis.

1. Introduction

In the clinical practice of speech therapy it is often necessary to establish to what degree a patient's speech is intelligible. According to Hustad [1] "Intelligibility refers to how well a speaker's acoustic signal can be accurately recovered by a listener". Assessments of intelligibility can be used for diagnostic purposes, but also to determine the degree of progress a patient has made. Similarly, in many lines of research on pathological speech it is necessary to assess patients' speech intelligibility, for instance to gauge the effectiveness of a specific treatment. Speech intelligibility has been studied not only in speech pathology research, but also in many other fields, such as second language (L2) pronunciation [2], speech synthesis evaluation [3, 4] and speech perception in adverse conditions [5, 6]. In spite of the considerable attention intelligibility scoring has received, many aspects are still unclear. In this paper we try to gain more insight into speech intelligibility scoring by investigating measures with different degrees of granularity.

Speech intelligibility is usually measured by collecting subjective judgments by human raters. Because these are by definition subjective, they should preferably be collected from multiple raters, after which average ratings and reliability measures are calculated. Subjective ratings of intelligibility can take many different forms [7, 8, 9]. A common practice is to ask raters to indicate the degree of intelligibility on a scale, such as an equal-appearing interval scale (or Likert scale; e.g. [9]), or a visual analogue scale (VAS), (placing a point on a horizontal line to indicate the degree of intelligibility; e.g. [10]). Although this procedure may provide reliable ratings, it is not clear to what extent these ratings are valid representations of intelligibility, because there is no ground truth. Second, scale ratings are generally collected at the speaker or utterance level, and thus provide relatively broad measures of intelligibility.

An alternative, and in a sense more valid procedure to measure intelligibility, consists in asking subjects to make orthographic transcriptions, i.e. to listen to speech fragments and write down what they hear (e.g., [11, 2, 12]). For this form of intelligibility measurement, different types of speech material can be used, including isolated words or pseudowords, whole sentences, and Semantically Unpredictable Sentences (SUS) [3]. All these types of materials have their own pros and cons. The advantage of using isolated words and pseudowords is that in this case the effect of context can be minimized. Isolated words and pseudowords have also been used to obtain more detailed scores at the word and even the phoneme level, e.g. by having experts write down or select the phoneme that was heard in a specific position in a certain (pseudo)word [13, 14, 15]. However, isolated words and pseudowords constitute a rather unnatural context, and it is unclear how the identification of specific phonemes in isolated (pseudo)words relates to speech intelligibility in a more natural context. In a sense, ratings of phonemes in isolated (pseudo)words are comparable to phonemic or phonetic transcriptions, where expert transcribers are supposed to indicate, as much as possible, how speech sounds have been realized, thus approximating an articulatory description of the sounds. However, it is questionable whether discrepancies observed between such phonetic transcriptions of the realized utterances and the corresponding canonical or reference transcriptions can be taken as measures of speech intelligibility, which is supposed to indicate to what extent a given utterance has been understood by a listener [1]. A similar discussion has been going on in the field of L2 pronunciation instruction, where a distinction has been made between measures of accentedness (as opposed to nativeness) and measures of intelligibility [2, 16]. Although accentedness and intelligibility appear to be related, they are distinct, partly independent dimensions. A relevant finding in this respect is that speech that is rated as heavily accented can still be intelligible [2].

Having listeners orthographically transcribe whole sentences instead of isolated words, seems preferable, because sentences constitute more natural speech material. Yet, sentences have the disadvantage that the contextual

http://dx.doi.org/10.21437/Interspeech.2016-1448

information they contain may facilitate comprehension. According to Yorkston et al. [17], in this case we would be measuring comprehensibility instead of intelligibility. To circumvent this problem, Semantically Unpredictable Sentences (SUS) have been proposed, which are syntactically correct, but semantically incoherent sentences (e.g. [3, 18]).

In general, orthographic transcriptions of regular or SUS sentences are scored at the word level: each word is scored as either correct or incorrect [11, 18]. Yet, both Hustad et al. [1] and Beijer et al. [19] argue that such word level scoring may still be quite broad, suggesting that it may be necessary to also collect judgments at even finer levels of granularity, i.e. the subword level. Intelligibility judgments on the subword level might indeed provide more detailed information about specific speech errors and may be more sensitive to changes within patients, enabling easier detection of treatment effects.

Next to human ratings of intelligibility, attempts have been made at developing objective measures of intelligibility that do not rely on human judgments. Many have employed ASR algorithms to obtain automatic measures of pathological speech quality [20, 21, 22, 23, 24]. It is, however, unclear to what extent such ASR-based metrics are valid representations of intelligibility. Firstly, they are often evaluated through comparison with benchmarks formed by human scale ratings or phonemic annotations. As explained above, it is not clear whether these benchmarks are themselves valid indicators of intelligibility. Secondly, while the automatic scoring methods that have been proposed so far are very interesting from a research point of view, they do not yet provide easy to use tools for clinical practice.

To summarize, in spite of the large body of research that has addressed intelligibility scoring of pathological speech, various issues still need clarification. The research reported in this paper aimed to contribute to this debate by investigating intelligibility measures with different degrees of granularity. We propose a procedure to automatically derive subword level intelligibility scores, i.e. scores at the phoneme and grapheme level, from orthographic transcriptions. The question we address is to what extent these subword intelligibility scores are reliable and how they relate to word level measures and utterance level ratings of intelligibility. In the following, we describe the procedures we used in collecting intelligibility evaluations of pathological speech on different levels of granularity (Section 2), we present the results (Section 3) and we discuss our findings (Section 4).

2. Method

An online listening experiment was set up to compare evaluations of speech intelligibility of dysarthric speech on three different levels of granularity: utterance level, word level, and subword level. Utterance level evaluations were obtained using subjective rating scales (VAS and Likert scale); word and subword level evaluations were obtained using orthographic transcriptions, which were scored on both word and subword level.

2.1. Speakers and speech material

The speech material used in the study was selected from the recordings collected by Beijer [18], from dysarthric speakers prior to speech therapy. To avoid speaker familiarity influencing the evaluation procedure, materials from seven different speakers were used. These were all male and suffered from hypokinetic dysarthria caused by Parkinson's disease.

To investigate the different levels of granularity in intelligibility evaluation for a broad range of speech material, four different types of recordings were used: lists of single words, declarative SUS sentences, interrogative SUS sentences, and regular sentences. All samples consisted of existing Dutch words. The word lists contained three or five words, the SUS sentences all contained six words, and the length of the regular sentences varied between five and eight words. varied between five and eight words

Table 1: *Overview of speech material used.*

Type of speech material	Speaker	Speech fragments
Word lists	S1	5 word lists (5 words each)
	S2	5 word lists (3 words each)
Declarative SUS sentences	S3	6 sentences
	S4	6 sentences
Interrogative SUS sentences	S5	6 sentences
	S6	6 sentences
Regular sentences	S7	8 sentences
	S1	8 sentences

Speech fragments with different levels of intelligibility, from low to high, were selected based on annotations by two listeners who did not participate in the current experiment.

2.2. Raters

Participants were invited by email or via Facebook. They filled in a questionnaire asking about mother tongue, gender, age, and familiarity with dysarthric speech. In total 36 listeners participated, 8 male and 28 female (age range 19-73). All listeners were native speakers of Dutch. Of the listeners, 31 had no experience with dysarthric speech and 5 had had the opportunity of listening to dysarthric speech before.

2.3. Procedure

The listening experiment was set up as an online experiment using the LimeSurvey application [25]. Listeners could participate by accessing the experiment through a link. At the beginning of the experiment the listeners filled in a questionnaire (see section 2.2), and were informed about the task and the types of speech material to be rated. In addition, they were told that they would hear only real Dutch words. Then they had to rate three example speech fragments, aimed to familiarize them with dysarthric speech and the rating procedures. These examples were specially selected to contain low and high intelligible speech, in order to give raters an idea of the intelligibility range they could expect and to stimulate them to use the whole range of the rating scales.

The raters had to evaluate each of the 50 speech fragments in three different ways: two subjective sentence level ratings – a Likert scale and a Visual Analogue Scale (VAS) – and an orthographic transcription. Every screen presented to the listeners contained one speech fragment and the accompanying three evaluation methods. Orthographic transcription was done by typing in a textbox what was heard. The Likert scale ranged from 1 ("very low intelligibility") to 7 ("very high intelligibility"). The VAS was implemented as a slider that could be positioned on any number between 0 ("very low

intelligibility") and 100 ("very high intelligibility"). The order of the questions on the screen varied: for half of the fragments the orthographic transcription task was placed at the top, for the other half of the fragments, the two subjective rating scales were placed at the top. However, since all three scales were presented simultaneously on one screen, they could be answered in any order.

The raters could listen to the speech fragments multiple times before scoring or transcribing them. The 50 speech fragments (screens) were presented in a random order. On average it took the raters 20 minutes to rate all the material.

2.4. Calculating intelligibility scores

This subsection describes how we calculated the intelligibility scores from the raw judgments and transcriptions of the raters

2.4.1. Intelligibility scores on utterance level

Intelligibility scores on utterance level were calculated as scores representing a percentage of intelligibility, ranging from 0 to 100. The VAS scores were already on a 0-100 scale, and were left unchanged. The scores on the Likert scale, ranging from 1 to 7, were transformed to percentage scores by first subtracting 1 and then multiplying by 16.67 (i.e. 1=0%, 2=16.67%, 3=33%, …, 7=100%).

2.4.2. Intelligibility scores on word level

The raters' orthographic transcriptions were compared to the reference transcriptions and the number of identical word matches was counted. Subsequently, a percentage correct score was calculated.

2.4.3. Intelligibility scores on subword level

Intelligibility scores at the grapheme and phoneme level were automatically obtained from the orthographic transcriptions. For both the phoneme and grapheme level the Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT) [26] was used. ADAPT computes the optimal alignment between two strings of phonetic symbols using a matrix that contains distances between the individual phonetic symbols. These distances are defined in terms of articulatory features and result in a distance measure expressing the phonetic similarity between the aligned transcriptions.

Listeners were instructed to not use any punctuation characters in their transcriptions. The punctuation characters we did find were removed and numerals were written out in words, which resulted in corrected orthographic transcriptions.

For the intelligibility scores on phoneme level, the orthographic transcriptions were converted to their phonemic equivalent using the canonical pronunciation variants from the lexicon of the Spoken Dutch Corpus [27]. Words that were not contained in the lexicon were manually added. As spelling errors complicated the lookup in the lexicon, those that did not affect the phonemic transcription were manually corrected. The resulting phonemic transcriptions were converted to the ADAPT symbol set (see appendix A in [26]). The ADAPT alignment algorithm and distance matrix were applied unchanged.

For the intelligibility scores on grapheme level, the phonetic symbols in the ADAPT distance matrix were replaced by the Dutch graphemes. The values of the graphemes 'articulatory feature' columns were all set to 0.0 except for the diagonals, which were set to 1.0. Using this matrix, the algorithm aligned the orthographic transcription with the reference transcription and calculated the distance between them. Each insertion and deletion was graded with distance 2.0 and substitution with distance 3.0.

3. Results

In total, five measures of intelligibility were collected for each speech fragment: two scale ratings on utterance level (Likert scale and VAS), a word level scoring of the orthographic transcription (OTW), and two subword level scorings of the orthographic transcriptions, at phoneme level (OTP) and at grapheme level (OTG). In this section we present the results regarding the reliability of these measures, and their relations.

3.1. Reliability

The reliability of each of the five intelligibility measures was calculated using Intraclass Correlation Coefficients (ICC) based on groups of raters, as we do not intend to devise an intelligibility measure relying on the judgment of a single rater. The ICC values for all 36 raters together were very high, ranging from .95 (OTP, OTG) to .97 (Likert, VAS, OTW). As such a large number of raters may not always be achievable, we also calculated ICCs based on smaller samples of raters, randomly drawn from our sample of 36 (for each sample size 10 random samples were drawn, and average ICCs were calculated). On average, for the utterance and word level scorings sufficient reliability is obtained with four raters (resulting in mean ICC values ranging from .79 to .84), while for subword scorings at least six raters are required (resulting in mean ICC values ranging from .79 to .80).

3.2. Intelligibility scores and correlations

Intelligibility scores for each fragment were calculated by averaging over the 36 raters. Mean scores for the five intelligibility measures, and the correlations between them are shown in Table 2. Correlations between all measures were significant ($p < .01$). The two utterance level measures were very highly correlated ($r = .998$), and the correlation between the two subword level measures was also very high ($r = .954$).

Table 2: *Means (SDs) and correlations of the five intelligibility measures (n = 50 speech fragments). VAS: Visual Analogue Scale, OTW: Orthographic Transcription at Word level, OTP: Orthographic Transcription at Phoneme level, OTG: Orthographic Transcription at Grapheme level. For Likert, VAS and OTW, higher scores correspond to higher intelligibility (higher percentage correct); for OTP and OTG higher scores correspond to lower intelligibility (higher distance). All correlations were significant (p < .01).*

	M (*SD*)	Correlations (Pearson *r*)			
		VAS	OTW	OTP	OTG
Likert	63.1 (21.1)	.998	.733	-.763	-.773
VAS	63.2 (19.0)		.732	-.755	-.764
OTW	78.3 (16.1)			-.805	-.869
OTP	8.0 (6.5)				.954
OTG	8.9 (7.4)				

To be able to directly compare subword level scores to word and utterance level percentage scores, we transformed the subword scores to percentage correct scores (phonemes or graphemes) in the utterance. Using an ANOVA with the five intelligibility measures as a within subject factor and percentage score as the dependent variable, we found significant differences between the different measures ($F(4,46) = 80.57$, $p < .01$). The percentage scores were significantly higher ($p < .01$) on the word level ($M = 78.3$) than on the utterance level (Likert: $M = 63.1$, VAS: $M = 63.2$), while the percent correct scores on the subword level were significantly higher ($p < .01$) than scores on the word level: 87.3 ($SD = 10.2$) for the phoneme level and 85.5 ($SD = 11.8$) the grapheme level. The differences between the Likert and VAS scores were not significant ($p > .05$).

4. Discussion and conclusions

4.1. Feasibility and reliability of subword scoring

In this paper, we calculated subword level scores by automatically processing the orthographic transcriptions. The ADAPT algorithm [26] used for this purpose only requires two text files as input, i.e. all orthographic transcriptions and the reference transcriptions. Both are formatted to contain a single transcription per line. The results of the alignments are stored in a comma-separated text file that allows easy viewing and import into spreadsheet software, making it feasible and relatively easy to use in clinical and research contexts.

Results showed that subword level intelligibility scorings are slightly less reliable than scorings at the utterance or word level. This can be explained by the effect of chance agreement [28]: since phoneme and grapheme scorings have a higher level of detail, they allow for more variation. Yet, when using at least six (unexperienced) raters, sufficiently reliable phoneme and grapheme scorings can be obtained. A sample of six raters is quite feasible in most research, and the fact that orthographic transcription tasks can easily be performed online makes it less problematic to involve multiple raters. With expert raters (e.g., speech-language pathologists), one would expect higher reliability, and the required number of raters might be lower. This should be verified in future research.

4.2. Comparisons between scores at different levels of granularity

The results show that intelligibility measures of different levels of granularity are fairly highly correlated, which is a reassuring outcome. When comparing the percentage scores obtained for the different measures, results show, first, that word level scoring produces higher intelligibility scores than utterance level scoring. This is in line with earlier research [11] suggesting that when using subjective rating scales, raters tend to underestimate the extent to which speakers are intelligible. A rater may, for example, understand every word, but still judge intelligibility as less than perfect when higher-than-normal listening effort is required because of articulatory irregularities. This again suggests that orthographic transcriptions may be more objective or valid measures of intelligibility while subjective rating scales indicate comprehensibility as defined by [2]: the difficulty a listener experiences in understanding an utterance.

A second finding is that percentage scores obtained with subword scorings are higher than those obtained with word level scoring. This is a natural consequence of the fact that with subword scoring, words that are not understood correctly can still be scored as partly correct: the transcription <stad> of the prompt <stap> would be incorrect at word level while at subword level only 1 of the 4 graphemes would be incorrect. In fact, subword level scores measure a different dimension of intelligibility, focusing on intelligibility of parts of words (graphemes, phonemes) rather than entire words.

4.3. Phoneme vs. grapheme scoring

We investigated two types of subword scorings, phoneme and grapheme scoring, in an attempt to obtain a more fine-grained measure of intelligibility. Clearly, phoneme scoring can provide more accurate indications of specific articulation problems of speakers, as phonemes directly represent speech sounds, whereas the connection between graphemes and speech sounds is blurred by spelling conventions (e.g. when two graphemes correspond to one phoneme, such as <oe> - /u/, <ng> - /ŋ/).

However, phoneme scoring is less easily performed, as it requires that a grapheme string is converted to a phoneme string, e.g. by means of a lexicon look-up (as we did in the current study) or a grapheme-to-phoneme conversion algorithm. The high correlation between the overall distance scores obtained from phoneme and grapheme scorings suggests that, at least when focusing on intelligibility per se, and not on specific articulation problems, grapheme scoring is a suitable alternative for phoneme scoring in cases where phoneme scoring is less feasible.

4.4. Conclusions

We conclude that automatic scoring of orthographic transcriptions on the subword level is a feasible way of obtaining a more fine-grained measure of intelligibility. While scoring at the phoneme level is more informative with respect to identifying specific articulation problems, scoring at the grapheme level is a reasonable alternative in cases where phoneme scoring is not possible, i.e. in clinical practice.

One can argue that utterance, word, and subword level scorings of intelligibility each measure a different dimension of intelligibility, with subjective utterance level ratings measuring comprehensibility as defined in [2], word level scorings of orthographic transcriptions measuring actual intelligibility of words, and subword level scorings measuring intelligibility of parts of words. As each of these dimensions of intelligibility are relevant in both clinical practice and research contexts, we suggest the use of subword scorings as a supplement to utterance level and word level scorings, not as a replacement. While subword scorings may require a few extra raters to achieve reliable measures, they provide worthwhile information at a finer level of granularity.

5. Acknowledgements

The authors would like to thank Lilian Beijer of the St. Maartenskliniek rehabilitation centre for collecting the dysarthric speech fragments used in this research. Thanks also extend to Ilona van der Linden and Sandra Schoemaker for collecting the listener judgments and their initial analyses. This research is funded by the NWO research grant with Ref. no. 314-99-101 (CHASING).

6. References

[1] K. C. Hustad, "The relationship between listener comprehension and intelligibility scores for speakers with dysarthria," *Journal of Speech, Language, and Hearing Research*, vol. 51, 562–573, 2008.

[2] M. J. Munro and T. M. Derwing, "Foreign Accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning*, vol. 45, no. 1, pp. 73–97, 1995.

[3] C. Benoit, M. Grice, and V. Hazan, "The SUS test: A method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences," *Speech Communication*, vol. 18, pp. 381–392, 1996.

[4] D. Gibbon, R. Moore, and R. Winski, *Handbook of standards and resources for spoken language systems.* Berlin: Mouton de Gruyter, 1997.

[5] A. Cutler, M. L. Garcia Lecumberri, and M. Cooke, "Consonant identification in noise by native and non-native listeners: Effects of local context," *Journal of the Acoustical Society of America*, vol. 124, no. 2, 1264–1268, 2008.

[6] M. Cooke, C. Mayo, C. Valentini-Botinhao, Y. Stylianou, B. Sauert, and Y. Tang, "Evaluating the intelligibility benefit of speech modifications in known noise conditions," *Speech Communication*, vol. 55, no. 4, pp. 572–585, 2013.

[7] S. S. Barreto and K. Z. Ortiz, "Intelligibility measurements in speech disorders: a critical review of the literature," *Pró-Fono Revista de Atualização Científica*, vol. 20, no. 3, pp. 201–206, 2008.

[8] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013.

[9] K. M. Yorkston and D. R. Beukelman, "A comparison of techniques for measuring intelligibility of dysarthric speech," *Journal of Communication Disorders*, vol. 11, pp. 499–512, 1978.

[10] C. Finizia, J. Lindstrom, and H. Dotevall, "Intelligibility and perceptual ratings after treatment for laryngeal cancer: laryngectomy versus radiotherapy", *Laryngoscope*, vol. 108, no. 1, pp. 138–143, 1998.

[11] K. C. Hustad, "Estimating the intelligibility of speakers with dysarthria," *Folia Phoniatrica et Logopaedica*, vol. 58, no. 3, pp. 217–228, 2006.

[12] J. S. Laures and G. Weismer, "The effects of a flattened fundamental frequency on intelligibility at the sentence level," *Journal of Speech Language and Hearing Research*, vol. 42, no. 5, pp. 1148–1156, 1999.

[13] G. van Nuffelen, C. Middag, M. de Bodt, and J.-P. Martens, "Speech technology-based assessment of phoneme intelligibility in dysarthria," *International Journal of Language & Communication Disorders*, vol. 44, no. 5, pp. 716–730, 2009.

[14] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbeck, "Toward phonetic intelligibility testing in dysarthria," *Journal of Speech and Hearing Disorders*, vol. 54, no. 4, pp. 482–499, 1989.

[15] W. Ziegler and A. Zierdt, "Telediagnostic assessment of intelligibility in dysarthria: A pilot investigation of MVP-online," *Journal of Communication Disorders*, vol. 41, no. 6, pp. 553–577, 2008.

[16] J. Levis, "Changing contexts and shifting paradigms in pronunciation teaching," *TESOL Quarterly*, vol. 39, no. 3, pp. 369–378, 2005.

[17] K. M. Yorkston, E. A. Strand, and M. R. Kennedy, "Comprehensibility of dysarthric speech: Implications for assessment and treatment planning," *American Journal of Speech-Language Pathology*, vol. 5, no. 1, pp. 55–66, 1996.

[18] L. J. Beijer, *E-learning based Speech Therapy (EST): Exploring the potentials of e-health for dysarthric speakers,* Phd Thesis. Nijmegen, Netherlands: Radboud University Nijmegen, 2012.

[19] L. J. Beijer, A. C. M. Rietveld, M. B. Ruiter, and A. C. H. Geurts, "Preparing an E-learning-based Speech Therapy (EST) efficacy study: Identifying suitable outcome measures to detect within-subject changes of speech intelligibility in dysarthric speakers," *Clinical Linguistics & Phonetics*, vol. 28, no. 12, pp. 927-950, 2014.

[20] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Noth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[21] C. Middag, J.-P. Martens, G. van Nuffelen, and M. de Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, no. 1, pp. 1–9, 2009.

[22] V. Berisha, R. Utianski, and J. Liss, "Towards a clinical tool for automatic intelligibility assessment," in *2013 IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP), May 26–31, Vancouver, BC, Canada, Proceedings*, 2013, pp. 2825–2828.

[23] T. Pellegrini, L. Fontan, J. Mauclair, J. Farinas, C. Alazard-Guiu, M. Robert, and P. Gatignol, "Automatic assessment of speech capability loss in disordered speech," *ACM Transactions on Accessible Computing*, vol. 6, no. 3, pp 8:1–8:14, 2015.

[24] M. J. Kim, Y. Kim, and H. Kim, "Automatic intelligibility assessment of dysarthric speech using phonologically-structured sparse linear model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 4, pp. 694–704, 2015.

[25] LimeSurvey Project Team and C. Schmitz (2015), *LimeSurvey: An open source survey tool, LimeSurvey Project.* Hamburg, Germany. http://www.limesurvey.org

[26] B. Elffers, C. van Bael, and H. Strik, *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions.* Internal report. Nijmegen, Netherlands: Department of Language & Speech, University of Nijmegen, 2013.

[27] N. Oostdijk, "The Spoken Dutch Corpus: Overview and first evaluation," in *LREC 2000*, Athens, Greece, Proceedings, 2000, pp 886–894.

[28] H. E. A. Tinsley and D. J. Weiss, "Interrater reliability and agreement of subjective judgments," *Journal of Counselling Psychology*, vol. 22, 358–376, 1975.