# Developing Valid Measurement Procedures for Speech Intelligibility of Pathological Speech

*Wei Xue*[1]

[1]Center for Language and Speech Technology (CLST), Radboud University
w.xue@let.ru.nl

## Abstract

Speech disorders in general, and dysarthria especially, lead to decreased speech intelligibility. This can have a severe impact on the patients' quality of life because they can lose social contact and eventually become isolated from society. Most of the time, degraded speech intelligibility can be improved through speech therapy. However, monitoring the effectiveness of speech therapy requires clear definition and operationalization of speech intelligibility. Until now, little is known about which deviations in pathological speech most affect intelligibility, and how intelligibility can best be improved. Furthermore, measuring intelligibility is complex, and it is time-consuming as it requires a lot of manual work. Therefore, I will study the following novel aspects in this project: which deviations in pathological speech have the most impact on intelligibility, what are good procedures for measuring intelligibility, and how can the workload in measuring intelligibility be reduced by making use of software tools.

**Index Terms**: speech intelligibility, speech disorder, acoustic features, automatic speech recognition

## 1. Introduction

Speech disorders in general, and dysarthria especially [1], lead to decreased speech intelligibility. This can have a severe impact on the patients' quality of life because they can lose social contact and eventually become isolated from society. Most of the time, degraded speech intelligibility can be improved through speech therapy. However, the effects of intensive therapy are not always evident. Therefore, it is necessary to monitor a possible evolution, pre- and post-therapy evaluations in which intelligibility scores play an important role. Thus, intelligibility requires a clear definition and a robust operationalization.

A clear definition has been proposed by Hustad [2] “Intelligibility refers to how well a speaker’s acoustic signal can be accurately recovered by a listener”. In line with this definition, intelligibility can be measured in various subjective ways. One of them is based on orthographic transcriptions [1,3,4]. The percentage of words or phonemes correctly identified is employed as a measure of intelligibility [4]. Besides, intelligibility has also been measured by collecting scalar ratings from human judges [7, 8, 9] through an equal-appearing interval scales like the Likert scale [9], or by placing a point on a horizontal line like the Visual Analogue Scale (VAS) [10].

Since subjectively measuring intelligibility is very time-consuming, several investigations have been carried out to develop methods for the objective measurement of speech intelligibility. These investigations are normally based on Automatic Speech Recognition (ASR) or machine learning algorithms and have been shown to produce outcomes that correlate with subjective human ratings [14-18]. So far, the relation between the automatic approaches’ outcomes and the properties of pathological speech is far from clear. The question of the possibility of using the measures as easy-to-use tools for clinical practice has not been well addressed.

Therefore, in this project, I want to study the following novel aspects: which deviations in pathological speech have the most impact on intelligibility in terms of acoustic features, what are good procedures for measuring intelligibility, and how can the workload in measuring intelligibility be reduced by making use of software tools.

## 2. Acoustic correlates of speech intelligibility of phonemes

To explore what deviations in pathological speech have the most impact on intelligibility, I studied the relation between acoustic features and intelligibility for pathological and control speech [5], specifically using a standardized feature set called eGeMAPS [20] and complementary features related to speech rate. The eGeMAPS feature set has 88 acoustic features related to e.g. loudness, fundamental frequency, and formant frequencies, which might be correlated with intelligibility.

In this study, the COPAS database [19] was involved and the included phoneme intelligibility obtained by the Dutch Intelligibility Assessment (DIA) task was used as the subjective speech intelligibility [6]. The database contains speech samples collected from a large number of control speakers and pathological speakers with different speech disorders. The contained speech materials include isolated words used in the DIA task, isolated sentences and short passages such as a commonly used phonetically-balanced Dutch text “Papa en Marloes” (TM). To explore the relation using different speech materials, i.e. isolated words and passages, speakers were selected based on participation in both the DIA and TM tasks. In total, 49 dysarthric and 81 control speakers were selected with 20 female and 29 male dysarthric speakers, and 48 female and 33 male control speakers.

The correlation between the features and the phoneme intelligibility was studied. The highest ten correlates for the dysarthric, control and combined speech on two tasks varied and the strengths were from 0.21 to 0.53. The moderate to medium correlations between phoneme intelligibility and the acoustic features in the eGeMAPS seems promising for automatic speech intelligibility prediction. A stepwise linear multiple regression (SLMR) algorithm was applied for predicting phoneme intelligibility using the eGeMAPS features on different tasks and different speech, i.e. dysarthric and control. The analyses revealed important differences between dysarthric speech and control speech, and between different types of speech material (isolated words and running text).

In conclusion, this study has shown that speech intelligibility is a complex construct. Further research is needed to get a better understanding of both the human-generated measures of intelligibility and their more objective acoustic correlates. Besides, it is important to include different types of speech materials, to explore the substantial differences between pathological and control speech in more detail, and to employ various rating procedures (i.e., Likert, VAS, and orthographic transcriptions). The important insights that derive from this more comprehensive research will not be limited to the clinical domain but may help us analyze speech intelligibility in the field of e.g. second language pronunciation.

## 3. Towards a comprehensive assessment of speech intelligibility

To explore the relation between measures obtained from various intelligibility measurement procedures, a semi-automatic approach was proposed in [11]. In this study, a set of intelligibility ratings of disordered speech assigned by lay listeners were investigated to obtain measures at three different levels of granularity: utterance, word, and subword level (grapheme and phoneme). Utterance level evaluations were obtained using subjective rating scales (i.e. VAS and Likert scale) while the word and subword-level evaluations, i.e. distance scores, were obtained automatically from human-generated orthographic transcriptions using automatic alignment and grapheme-phoneme conversion algorithms. The results indicated that the distance measure at the phoneme level was feasible and reliable.

Therefore, my colleagues and I extended this semi-automatic approach and its automatically derived metrics as measures of pathological speech intelligibility on several important points [24]. We collected measures for a larger number of samples, including both pathological and control speech, covering different speech materials, i.e. meaningful sentences from TM, word lists in DIA and semantically unpredictable sentences (SUS). Intelligibility measures were collected from experts as opposed to lay listeners. The experts were asked to provide two types of transcriptions, ‘Word’ in terms of existing words and ‘Literal’ in terms of literal or perceived segments. More detailed automated measures were explored concerning speaker types, speech materials and levels of granularity. The reliability and validity of measures obtained from transcriptions are evaluated in relation to other measures such as VAS and severity level of dysarthria.

In the end, eight measures were calculated for each speech sample: VAS at the utterance level, accuracy (Acc) at the word level (W), and accuracy (Acc), distance (Dist) and the number of changes (Ch) at the grapheme (G) and phoneme level (P). The reliability and validity of the measures of intelligibility were investigated for both control and pathological speech. The analyses reveal that for the eight measures we acquired, the reliability coefficients were very high using different speech materials. This supports the usability of these measures since a limited number of raters might be sufficient to obtain highly reliable ratings, which is very important in a clinical setting.

At the subword level, the mean values of Acc at the phoneme-level results always slightly lower than those at the grapheme level, which showed a similar pattern to Dist and Ch. This is understandable because a phoneme may be associated with more than one grapheme and then its overall correctness requires correctness in its associated graphemes. The six automatically calculated subword-level measures are strongly correlated with each other, which could be explained by the fact that they are all based on the same orthographic transcriptions. However, it is worth noting that they are strongly related and that using one or the other does not make much difference. For instance, a grapheme-level measure may be easier to apply than a phoneme-level one in clinical practice, but both will yield accurate results.

To test the external validity of the subword-level measures, we included the independent measure VAS, the accuracy at the word level (Acc-W) and the severity level of dysarthria (SL) at the speaker level as evaluation criteria. The correlations between the three evaluation measures are moderate to strong, presumably showing that evaluative components involved in estimating intelligibility are different.

Concerning the correlations between these three measures and the six subword-level measures, the Acc measures outperform the Dist and Ch measures. Within the Acc measures, Acc at phoneme level (Acc-P) performs better than Acc at grapheme level (Acc-G). More detailed results can be found in [24] in Table 3. The results indicate that the phoneme measures outperformed the grapheme measures and that the best phoneme measure seems to be accuracy. This suggests that the investigated orthography-based subword-level measures are not only reliable indicators of speech intelligibility but that they can also be considered as valid descriptors of speech intelligibility in pathological speech.

In conclusion, the results show the possibility of using orthographic transcriptions and the automated phoneme measures to determine which mispronounced phonemes cause decreased speech intelligibility comparing with other automatic measures [14, 21]. In other words, these measures have potentially additional diagnostic value and can, therefore, be applied in speech therapy.

## 4. Future work

Again, as I mentioned above, further exploring how the acoustic features are correlated with speech intelligibility is needed and may help researchers to understand the underlying processes and mechanisms that affect speech intelligibility. Future work will also explore the possibility to fully automate intelligibility evaluation without any human-generated orthographic transcriptions. This could be achieved with the help of Automatic Speech Recognition (ASR) technology and the increasing availability of dysarthric speech data [12, 13, 22, 23]. Another option for prompted speech would be to use ASR in forced alignment mode, which is one of the methods I intend to investigate in future research.

## 5. Acknowledgements

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 766287. I would like to thank my supervisors Helmer Strik, Catia Cucchiarini, and Roeland van Hout, as well as a master student Wieke Harmsen for her contributions, and our colleagues from the Department of Otorhinolaryngology, Head and Neck Surgery and Communication Disorders of the University Hospital of Antwerp, Gwen Van Nuffelen, Marc De Bodt and Viviana Mendoza Ramos, for their cooperation and feedback.

# 6. References

[1] D. Kempler, D. Van Lancker, "Effect of speech task on intelligibility in dysarthria: a case study of Parkinson's Disease," Brain Lang, vol. 80, 2002, pp.:449-64.

[2] K. C. Hustad, "The Relationship Between Listener Comprehension and Intelligibility Scores for Speakers with Dysarthria," J. Speech Lang. Hear. Res., vol. 51, no. 3, 2008, pp. 562.

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

[4] K. Yorkston, D. R. Beukelman, and R. Tice, "Sentence intelligibility test [Measurement instrument]," Lincoln, NE: Tice Technologies, 1996.

[5] W. Xue, C. Cucchiarini, R. van Hout, H. Strik, "Acoustic correlates of speech intelligibility: the usability of the eGeMAPS feature set for atypical speech," in *Proc. SLaTE 2019*: 8th ISCA Workshop on Speech and Language Technology in Education, 2019, pp.: 48-52, DOI: 10.21437/SLaTE.2019-9.

[6] M. de Bodt, C. Guns, G. van Nuffelen, NSVO: handleiding. Vlaamse Vereniging voor Logopedie: Herentals, 2006.

[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, 2008, pp.: 201–206.

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

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

[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, 1998, pp.: 138–143.

[11] M. Ganzeboom, M. Bakker, C. Cucchiarini, and H. Strik, "Intelligibility of Disordered Speech: Global and Detailed Scores," in INTERSPEECH, 2006, pp.: 2503–2507.

[12] E. Yilmaz, M. Ganzeboom, C. Cucchiarini and H. Strik, "Combining Non-pathological Data of Different Language Varieties to Improve DNN-HMM Performance on Pathological Speech," in INTERSPEECH, 2016, pp.: 218-222.

[13] E. Yilmaz, M. Ganzeboom, C. Cucchiarini and H. Strik, "Multistage DNN training for Automatic Recognition of Dysarthric Speech," in *Proc. INTERSPEECH*, 2017, pp.: 2685-2689.

[14] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "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, 2006, pp.: 1741–1747.

[15] 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, 2009, pp.: 1–9.

[16] 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.

[17] 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, 2015, pp.: 8:1–8:14.

[18] 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, 2015, pp.: 694–704.

[19] C. Middag, "Automatic analysis of pathological speech," Dissertation. Ghent University, 2012.

[20] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, J. André et al., "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, 2015, pp.: 190-202.

[21] A. Maier, M. Schuster, A. Batliner, E. Nöth, and E. Nkenke, "Automatic scoring of the intelligibility in patients with cancer of the oral cavity," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[22] D. Martínez, P. Green, and H. Christensen, "Dysarthria intelligibility assessment in a factor analysis total variability space," In *Proceedings of INTERSPEECH*, 2013.

[23] D. Martínez, E. Lleida, P. Green, H. Christensen, A. Ortega, and A. Miguel, "Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace," *ACM Transactions on Accessible Computing (TACCESS)*, vol. 6, no. 3, 2015, pp.: 1-21.

[24] W. Xue, V. Mendoza Ramos, W. Harmsen, C. Cucchiarini, R. van Hout and H. Strik, "Towards a comprehensive assessment of speech intelligibility for pathological speech," in *Proc. INTERSPEECH*, 2020.