Speech intelligibility measurement

A latent variable approach on utterances' transcriptions

Jose Rivera[1], Sven de Maeyer[2], and Steven Gillis[3]

[1] Department of Training and Education Sciences,
University of Antwerp, Antwerp, Belgium
E-mail: JoseManuel.RiveraEspejo@uantwerpen.be
(corresponding author)

[2] Department of Training and Education Sciences,
University of Antwerp, Antwerp, Belgium
E-mail: sven.demaeyer@uantwerpen.be

[3] Computational Linguistics, and Psycholinguistics Research Centre
University of Antwerp, Antwerp, Belgium
E-mail: steven.gillis@uantwerpen.be

August 16, 2022

Abstract

Contents

List of Figures

List of Tables

1 Introduction

Intelligible speech can be defined as the extent in which the elements in a speaker's acoustic signal, e.g. phonemes or words, can be correctly recovered by a listener [17, 21, 39, 40]. Intelligible spoken language carries an important societal value, as its attainment requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered [17]. In that sense, *speech intelligibility* is considered a milestone in children's language development, and more practically, it is qualified as the ultimate checkpoint for the success of speech therapy, and the 'gold standard' for assessing the benefit of cochlear implantation [6].

Multiple approaches can be taken to quantify *speech intelligibility* [2, 3, 14, 20], but among them, *objective rating* methods on stimuli recovered from spontaneous speech tasks have received special attention [3, 20]. In objective rating methods, listeners transcribe children's utterances orthographically (or phonetically), and use such information to construct an intelligibility score. The construction of the score can be done in several ways, e.g. counting the number of (un)intelligible syllables or words in the utterances [14, 23], or calculating an entropy score, a measure that expresses the degree of (dis)agreement in the transcriptions [3, 36]. In that sense, the method tries to infer intelligibility from the extent in which a set of transcribers, can identify the words contained in multiple utterances [3].

As the literature suggests, objective rating procedures produce more valid[1] and reliable[2] scores than any other available procedure [3, 11], as the method does not hinge in the use or production of a *subjective rating scale*, i.e. a scale based on a personal perception of the child's intelligibility. Moreover, the previous advantages are further emphasized by the use of stimuli gathered from spontaneous speech tasks, as they have a greater level of ecological validity, especially compared to contextualized utterances or reading at loud tasks [14, 9].

However, although the literature is clear on the method's benefits to quantify (measure) *speech intelligibility* [2, 3, 20], we notice the statistical approaches used to model such data still face three important issues, and these come to the detriment of the measurement procedure's sophistication.

First, as previous paragraphs reveal, the intelligibility scores are 'complex' in nature, however, such 'complexity' is rarely fully considered in the statistical modeling procedure. The problem with the later is that, because the data does not fulfill the typical assumptions, e.g. normality, its analysis under such models might lead us to erroneous conclusions [citation]. On the one hand, outcomes such as the number of (un)intelligible words are discrete, while the entropy scores are continuous in nature. In addition, there is the consideration that both scores are constraint in specific bounds, i.e. the number of (un)intelligible words cannot be negative, while the entropy scores are in the bounds between zero and one. Finally, given the measurement procedure's nature, the scores are produced in a clustered manner or with replication, i.e. we observe several score measurements per child.

So far the literature shows the applied statistical procedures have always assumed 'normality', examples of this can be seen in Boonen et al. [3], Flipsen and Colvard [15] and Hustad et al. [20]. In addition, some papers in the literature have even used multilevel modeling to deal with the clustered nature of the data, e.g. Boonen et al. [3]. However, to the authors knowledge, no paper have dealt with all of the data 'complexity' at once, which leads us to believe that, by using more sophisticated statistical models to account for all of these nuances, we could improve our statistical inferences.

Second, although the literature suggest the number of (un)intelligible words or the entropy of transcriptions are scores that capture the level of intelligibility in a child, it is easy to notice these two can still be considered surrogate measures of it, i.e. scores that indirectly reflect what is intended to be measured. The latter is important because it implies these outcomes are 'measured with error', resulting from considering that there is an unobserved 'construct' that is responsible for the variation observed in them, i.e. the *speech intelligibility*. Moreover, is important to recognize that this 'measurement error' is of a different kind that the one produced by the clustered nature of the data, and that by failing to account for it, would also lead us to produce incorrect inferences [8].

To the authors knowledge, no attempt to create such intelligibility 'construct' have been made. Therefore, we believe the literature could benefit from showing how to implement such procedure in a statistical model, in combination with the other procedures needed to account for all of the aforementioned nuances in the data.

[1] validity is understood as the extent to which scores are appropriate for their intended interpretation and use [25, 38].

[2] reliability is though as the extend to which a measure would give us the same result over and over again [38], i.e. measure something, free from error, in a consistent way.

Third, even though the literature supplies a myriad of factors that are thought to contribute to the (under)development of intelligible spoken language [4, 18, 12, 29], no transparent framework of analysis is used to determine which factors are relevant, or conforms to valid and actionable causal hypothesis. The lack of such framework not only makes the assessment and selection of relevant factors harder, but also hinders the researcher's ability to avoid facing some common statistical issues related to such selection procedure, e.g. determine which factors can be analyzed in tandem without facing collinearity problems, which ultimately affects our inference capabilities [13].

As it was suggested, several factors are proposed by the literature, but these can be largely grouped into three categories: audiology, child and environmental related factors. For the first, they are the chronological age, age at implantation, the duration of device use, 'hearing' age, bilateral or contralateral cochlear implantation, and the children's preoperative and postoperative hearing levels. For the second, there is the etiology or the cause of the hearing impairment (e.g. genetic, infections), additional disabilities (e.g. mental retardation, speech motor problems), and gender. Finally for the last, there is the communication modality. Therefore, considering the aforementioned variables, and the relation's complexity with themselves and the outcome, we believe that using a causal framework would provide a more transparent way of state and analyze our research hypothesis.

Considering all of the above, we believe this paper make three specific contributions to the field. First, we develop a novel analysis using a Generalized Linear Latent and Mixed Model (GLLAMM) [31, 33, 32, 34, 37]. More specifically, we model *speech intelligibility* as a latent variable [10], that can be inferred from the replicated entropy scores. Moreover, the entropy measures are then modeled under a Generalized Linear Mixed Model (GLMM) [5, 24, 27].

This method offers three specific benefits. On the one hand, it allow us to consider all of the data nuances at once, i.e. we can model our 'non normal' data, and control for the different sources of variation (error) observed in it. The latter is particularly important because, by failing to account for these sources, we could be 'manufacturing' false confidence in the parameter estimates, leading us to incorrect inferences [28]. On the other hand, the method provides a way to 'construct' an intelligibility scale. This in turn, allow us to test different hypothesis, and even make individual comparisons at the children level. Finally, resulting from the statistical procedure sophistication, the method also provides a 'criterion' on how reliable are the entropy replicates to measure speech intelligibility.

Second, we use Directed Acyclic Graph (DAG) [30, 7] to depict all the relevant variables though to influence *speech intelligibility*. We describe in detail our causal and non-causal hypothesis, and supplement our description with a causal diagram. The benefit of the method lies, not only, in that it makes the assumptions of our hypothesis more transparent, but also allow us to derive statistical procedures from the aforementioned causal assumptions [28, 42, 35].

Third and final, given the statistical procedure complexity, we wrap the analysis under the Bayesian framework. Bayesian statistics can handle all kinds of data-generating processes [16], and it lends itself easily to complex and over-parameterized models [1, 22], as the one we are trying to implement in our current research. Furthermore, although the framework have similar estimation capabilities as its frequentist counterpart [1, 19, 41], the specific set of scenarios in our current research also favors its use. For example, the Bayesian analysis shines when we are faced with scenarios with small samples sizes [16, 28, 37], and when parameters need to be confined in a permitted space [26], e.g. variances have to be confined in the positive space. Finally, since the main output of Bayesian statistics are not point estimates, but rather the posterior distribution of the parameters' possible values [28], Bayesian modeling allow us to have a more nuanced view of our inferences and conclusions.

Bibliography

[1] Baker, F. [1998]. An investigation of the item parameter recovery characteristics of a gibbs sampling procedure, *Applied Psychological Measurement* **22**(22): 153–169.
doi: https://doi.org/10.1177/01466216980222005.

[2] Boonen, N., Kloots, H. and Gillis, S. [2020]. Rating the overall speech quality of hearing-impaired children by means of comparative judgements, *Journal of Communication Disorders* **83**: 1675–1687.
doi: https://doi.org/10.1016/j.jcomdis.2019.105969.

[3] Boonen, N., Kloots, H., Nurzia, P. and Gillis, S. [2021]. Spontaneous speech intelligibility: early cochlear implanted children versus their normally hearing peers at seven years of age, *Journal of Child Language* pp. 1–26.
doi: https://doi.org/10.1017/S0305000921000714.

[4] Boons, T., Brokx, J., Dhooge, I., Frijns, J., Peeraer, L., Vermeulen, A., Wouters, J. and van Wieringen, A. [2012]. Predictors of spoken language development following pediatric cochlear implantation, *Ear and Hearing* **33**(5): 617–639.
doi: https://doi.org/10.1097/AUD.0b013e3182503e47.

[5] Breslow, N. and Clayton, D. [1993]. Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**(421): 9–25.
doi: https://doi.org/10.2307/2290687.
url: http://www.jstor.org/stable/2290687.

[6] Chin, S., Bergeson, T. and Phan, J. [2012]. Speech intelligibility and prosody production in children with cochlear implants, *Journal of Communication Disorders* **45**: 355–366.
doi: https://doi.org/10.1016/j.jcomdis.2012.05.003.

[7] Cinelli, C., Forney, A. and Pearl, J. [2022]. A crash course in good and bad controls, *SSRN* .
doi: http://dx.doi.org/10.2139/ssrn.3689437.
url: https://ssrn.com/abstract=3689437.

[8] deHaan, E., Lawrence, A. and Litjens, R. [2019]. Measurement error in dependent variables in accounting: Illustrations using google ticker search and simulations, *Workingpaper*.

[9] Ertmer, D. [2011]. Assessing speech intelligibility in children with hearing loss: Toward revitalizing a valuable clinical tool, *Language, Speech, and Hearing Services in Schools* **42**(1): 52–58.
doi: https://doi.org/10.1044/0161-1461(2010/09-0081).

[10] Everitt, B. [1984]. *An Introduction to Latent Variable Models*, Monographs on Statistics and Applied Probability, Springer Dordrecht.
doi: https://doi.org/10.1007/978-94-009-5564-6.

[11] Faes, J., De Maeyer, S. and Gillis, S. [2021]. Speech intelligibility of children with an auditory brainstem implant: a triple-case study, pp. 1–50. (submitted).

[12] Fagan, M., Eisenberg, L. and Johnson, K. [2020]. Investigating early pre-implant predictors of language and cognitive development in children with cochlear implants, *in* M. Marschark and H. Knoors (eds), *Oxford handbook of deaf studies in learning and cognition*, Oxford University Press, pp. 46–95.
doi: https://doi.org/10.1093/oxfordhb/9780190054045.013.3.

[13] Farrar, D. and Glauber, R. [1967]. Multicollinearity in regression analysis: The problem revisited, *Review of Economics and Statistics* **49**(1): 92–107.
doi: https://doi.org/10.2307/1937887.
url: https://www.jstor.org/stable/1937887.

[14] Flipsen, P. [2006]. Measuring the intelligibility of conversational speech in children, *Clinical Linguistics & Phonetics* **20**(4): 303–312.
doi: https://doi.org/10.1080/02699200400024863.

[15] Flipsen, P. and Colvard, L. [2006]. Intelligibility of conversational speech produced by children with cochlear implants, *Journal of Communication Disorders* **39**(2): 93–108.
doi: https://doi.org/10.1016/j.jcomdis.2005.11.001.
url: https://www.sciencedirect.com/science/article/pii/S0021992405000614.

[16] Fox, J. [2010]. *Bayesian Item Response Modeling, Theory and Applications*, Statistics for Social and Behavioral Sciences, fienberg, s. and van der linden, w. edn, Springer Science+Business Media, LLC.

[17] Freeman, V., Pisoni, D., Kronenberger, W. and Castellanos, I. [2017]. Speech intelligibility and psychosocial functioning in deaf children and teens with cochlear implants, *Journal of Deaf Studies and Deaf Education* **22**(3): 278–289.
doi: https://doi.org/10.1093/deafed/enx001.

[18] Gillis, S. [2018]. Speech and language in congenitally deaf children with a cochlear implant, *in* E. Dattner and D. Ravid (eds), *Handbook of Communication Disorders: Theoretical, Empirical, and Applied Linguistic Perspectives*, De Gruyter Mouton, chapter 37, pp. 765–792.
doi: https://doi.org/10.1515/9781614514909-038.

[19] Hsieh, M., Proctor, T., Hou, J. and Teo, K. [2010]. A comparison of bayesian mcmc and marginal maximum likehood methods in estimating the item parameters for the 2pl irt model, *International Journal of Innovative Management, Information and Production* **1**(1): 81–89.
url: http://ismeip.org/IJIMIP/contents/imip1011/10IN15T.pdf.

[20] Hustad, K., Mahr, T., Natzke, P. and Rathouz, P. [2020]. Development of speech intelligibility between 30 and 47 months in typically developing children: A cross-sectional study of growth, *Journal of Speech, Language, and Hearing Research* **63**(6): 1675–1687.
doi: https://doi.org/10.1044/2020$_JSLHR-20-00008$.
url :$https://pubs.asha.org/doi/abs/10.1044/2020_JSLHR-20-00008$.

[21] Kent, R., Weismer, G., Kent, J. and Rosenbek, J. [1989]. Toward phonetic intelligibility testing in dysarthria, *Journal of Speech and Hearing Disorders* **54**(4): 482–499.
doi: https://doi.org/10.1044/jshd.5404.482.

[22] Kim, S. and Cohen, A. [1999]. Accuracy of parameter estimation in gibbs sampling under the two-parameter logistic model, *Annual Meeting of the American Educational Research Association*, American Educational Research Association.
url: https://eric.ed.gov/?id=ED430012.

[23] Lagerberg, T., Asberg, J., Hartelius, L. and Persson, C. [2014]. Assessment of intelligibility using childrens spontaneous speech: Methodological aspects, *International Journal of Language and Communication Disorders* **49**: 228–239.
doi: https://doi.org/10.1111/1460-6984.12067.

[24] Lee, Y. and Nelder, J. A. [1996]. Hierarchical generalized linear models, *Journal of the Royal Statistical Society: Series B (Methodological)* **58**(4): 619–656.
doi: https://doi.org/10.1111/j.2517-6161.1996.tb02105.x.
url: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1996.tb02105.x.

[25] Lesterhuis, M. [2018]. *The validity of comparative judgement for assessing text quality: An assessors perspective*, PhD thesis, University of Antwerp.

[26] Martin, J. and McDonald, R. [1975]. Bayesian estimation in unrestricted factor analysis: A treatment for heywood cases, *Psychometrika* (40): 505517.
doi: https://doi.org/10.1007/BF02291552.

[27] McCullagh, P. and Nelder, J. [1983]. *Generalized Linear Models*, Monographs on Statistics and Applied Probability, Routledge.
doi: https://doi.org/10.1201/9780203753736.

[28] McElreath, R. [2020]. *Statistical Rethinking: A Bayesian Course with Examples in R and STAN*, Chapman and Hall/CRC.

[29] Niparko, J., Tobey, E., Thal, D., Eisenberg, L., Wang, N., Quittner, A. and Fink, N. [2010]. Spoken language development in children following cochlear implantation, *JAMA* **303**(15): 1498–1506.
doi: https://doi.org/10.1001/jama.2010.451.

[30] Pearl, J. [2009]. *Causality: Models, Reasoning and Inference*, Cambrige University Press.

[31] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004a]. Generalized multilevel structural equation modeling, *Psychometrika* **69**(2): 167–190.
doi: https://www.doi.org/10.1007/BF02295939.

[32] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004b]. *GLLAMM Manual*, UC Berkeley Division of Biostatistics.
url: http://www.biostat.jhsph.edu/ fdominic/teaching/bio656/software-gllamm.manual.pdf.

[33] Rabe-Hesketh, S., Skrondal, A. and Pickles, A. [2004c]. Maximum likelihood estimation of limited and discrete dependent variable models with nested random effects, *Journal of Econometrics* **128**(2): 301–323.
doi: https://www.doi.org/10.1016/j.jeconom.2004.08.017.
url: http://www.sciencedirect.com/science/article/pii/S0304407604001599.

[34] Rabe-Hesketh, S., Skrondal, A. and Zheng, X. [2012]. Multilevel structural equation modeling, *in* R. Hoyle (ed.), *Handbook of Structural Equation Modeling*, The Guilford Press, chapter 30, pp. 512–531.

[35] Rohrer, J., Schmukle, S. and McElreath, R. [2021]. The only thing that can stop bad causal inference is good causal inference, *PsyArXiv* .
doi: https://doi.org/10.31234/osf.io/mz5jx.

[36] Shannon, C. [1948]. A mathematical theory of communication, *The Bell System Technical Journal* **27**(3): 379–423.
doi: https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[37] Skrondal, A. and Rabe-Hesketh, S. [2004]. *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*, Interdisciplinary Statistics, Chapman Hall/CRC Press.

[38] Trochim, W. [2022]. The research methods knowledge base.
url: https://conjointly.com/kb/.

[39] van Heuven, V. [2008]. Making sense of strange sounds: (mutual) intelligibility of related language varieties. a review, *International Journal of Humanities and Arts Computing* **2**(1-2): 39–62.
doi: https://doi.org/10.3366/E1753854809000305.

[40] Whitehill, T. and Chau, C. [2004]. Single-word intelligibility in speakers with repaired cleft palate, *Clinical Linguistics and Phonetics* **18**: 341–355.
doi: https://doi.org/10.1080/02699200410001663344.

[41] Wollack, J. A., Bolt, D. M., Cohen, A. S. and Lee, Y.-S. [2002]. Recovery of item parameters in the nominal response model: A comparison of marginal maximum likelihood estimation and markov chain monte carlo estimation, *Applied Psychological Measurement* **26**(3): 339–352.
doi: https://www.doi.org/10.1177/0146621602026003007.

[42] Yarkoni, T. [2020]. The generalizability crisis, *The Behavioral and brain sciences* **45**(e1).
doi: https://doi.org/10.1017/S0140525X20001685.