Speech intelligibility:

A generalized latent variable approach on utterances' entropies

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
[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

December 2, 2022

Corresponding author: Jose Rivera, Department of Training and Education Sciences, University of Antwerp, Antwerp, Belgium. e-mail: JoseManuel.RiveraEspejo@uantwerpen.be

Finantial support: The project was financed by the Flemish Government through the Research Fund of the University of Antwerp (BOF).

Competing interests: The authors declare they have no conflict of interest regarding this manuscript.

Keywords— intelligibilty, children with coclear implants, utterances' entropy, generalized linear latent and mixed model.

Abstract

Contents

List of Figures

List of Tables

1 Introduction

Intelligible spoken language requires all core components of speech perception, cognitive processing, linguistic knowledge, and articulation to be mastered (Freeman et al.; 2017). In that sense, its attainment carries an important societal value, as it 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 benefits of cochlear implantation (Chin et al.; 2012).

But what is speech intelligibility?. Intelligibility is usually conceptualized as the extent to which the elements in an acoustic signal generated by a speaker, e.g. phonemes or words, can be correctly recovered by a listener (Freeman et al.; 2017; Kent et al.; 1989; Munro and Tracey; 1999; van Heuven; 2008; Whitehill and Chau; 2004). The latter definition sets a clear contrast with comprehensibility, which involves the listener's ability to understand the sounds' message and its intent (Munro and Tracey; 1999; Smith and Nelson; 1985).

But the literature reveals that intelligibility is an intricate concept, with particular challenges to its assessment/measurement. The latter is because intelligibility can be affected by features of the communicative environment, such as noise (Munro; 1998); by features of the speaker, like speaking rate (Munro and Derwing; 1998) or accent (Jenkins; 2000; Ockey et al.; 2016); or features of the listener, like vocabulary mastery (Varonis and Susan; 1985). Moreover, this further emphasizes another aspect of the concept: its dynamic nature, where changes in intelligibility stem from the speaker's online adaptations to the listener and/or context.

Therefore, the literature suggests there are three aspects to the study of speech sounds, and therefore, three from which intelligibility can be asssessed: the acoustic, articulatory, and auditory aspects (Gudivada et al.; 2018). The first is focused on assessing the transmission and physical properties of speech sounds (Boonen et al.; 2020, 2021). The second is more concerned with the sounds' production (Rowe and Levine; 2018). While the last, center its attention on the speech sounds' perception, i.e. how the stimuli are perceived by a listener (Boonen et al.; 2020, 2021).

Focusing our attention on the last one, perceptual studies also use multiple approaches to measure intelligibility, but they can be largely grouped into two: subjective and objective ratings methods (Hustad et al.; 2020). In the former, listeners directly infer the intelligibility score of the speech samples through different procedures. While in the latter, listeners transcribe children's utterances orthographically (or phonetically), and use these as information to construct an entropy score that expresses the degree of (dis)agreement in the transcriptions (Boonen et al.; 2021; Shannon; 1948).

Consequently, objective ratings try to infer intelligibility from the extent to which a set of transcribers can identify the words contained in the utterances (Boonen et al.; 2021). While subjective ratings, try to directly produce a score based on a listener's perception of the sounds' intelligibility. In either case, the methods produce a proxy measure of the speaker's intelligibility as judged by a listener, a snapshot of his/her performance under a specific set of circumstances (Hustad et al.; 2020).

Moreover, the methods' validity, i.e. the extent to which scores are appropriate for their intended interpretation and use (Lesterhuis; 2018; Trochim; 2022), is founded on the idea that intelligibility is an intuitively understood notion, "something" that anyone can judge, but that can only be measured indirectly because of its entanglement with other features of the communication (Guilford; 1954; Stevens; 1946).

Recently in the literature, objective rating procedures applied on children's utterances recovered from spontaneous speech tasks have received special attention (Boonen et al.; 2021; Hustad et al.; 2020). The scores produced from these tasks are characterized by their clustered and bounded nature. The former happens because the data register multiple measurements per child; more specifically, one score per utterance, where multiple utterances are assessed. While the latter happens because the entropy score values are expresed in the continuum between zero and one (Shannon; 1948).

Although the literature has been clear on the benefits of the aforementioned method to (indirectly) quantify intelligibility (Boonen et al.; 2020, 2021; Hustad et al.; 2020), we notice the statistical procedures used to model such data have not been fully at par to the measurement procedure's sophistication.

First, previous research have dealt with the data clustering, but ignored its bounded nature. Specifically, Boonen et al. (2021) used multilevel linear models (MLM) to statistically model similar data, and test some research hypotheses of interest. But it is relevant to first highlight why the use of such models is a requirement when the data have this characteristic.

When more than one observation arises from the same individual, location, or time, then traditional (single-level) statistical models may mislead us (McElreath; 2020). The reason for this, is that one of the main assumptions of these models gets violated: the independence of errors (Finch et al.; 2019).

The latter is easier to understand with a thought experiment. Consider the scenario hinted in previous paragraphs: we observe one entropy score per utterance, for a total of ten utterances per child. In this scenario, it would be reasonable to believe the ten entropy scores observed for the same child would be more similar with one another, than what they are with the scores observed for other children. This within-child correlation would be due, for example, to having the same speech pattern or articulation, the same linguistic knowledge, among other reasons, perceived by the listener.

The presence of this within-child correlation in the data will, in turn, result in two know statistical issues (Finch et al.; 2019). On the one hand, the inapropriate estimation of the standard errors parameters of the model. This is important, as biased parameters might lead us to less appropriate statistical inferences, e.g. larger t-statistics with smaller p-values, that lead to the rejection of a "true" null hypothesis (Type I error). On the other hand, if the multilevel structure of the data is ignored, we may miss important relationships involving each level in the data. Following the previous scenario, we have two levels in our data: utterances (level 1) and children (level 2). Notice that we might have different information at each level that explains the data behavior, and by not including it, we will be using an incorrect model for understanding the outcome of interest.

As mentioned, previous literature have already considered the data clustering Boonen et al. (2021). However, we argue the latter practice is not sufficient, as with bounded data not only the location (average), but also the spread (variance), of the entropy distribution might inform about the speaker's intelligibility (McCullagh and Nelder; 1983), and we need to model both.

To understand the preceding statement, it is important to first highlight the main assumption of MLMs: the errors are normally distibuted, and by extension, our outcome of interest. But what the normality assumption implies?. First, it imply that we are only interested on modeling the outcome's location (average). Second, it implies that the outcome's average can take any value without constraint. And third, it implies that the distributon's average and variance are two independent parameters (McCullagh and Nelder; 1983).

However, a simple thought experiment can show us why this is a problem. Consider children with three different patterns for ten entropy measures, all reporting the same mean entropy of 0.5. The patterns are: (a) scores closely agglomerated around 0.5, (b) scores loosely aglomerated around 0.5, and finally, (c) half of the scores agglomerated around 0.1 and the other half around 0.9. From the mean score we can say that the three children have an "average" level of intelligibility. However, from the spread of the scores we can notice that more uncertainty (to the asesssment of "average") should be assigned to child (c), followed by (b) and finally (a). This just mean that we can be more confident that child (a) has an "average" level of intelligibility, than in the other two cases, where (c) represent one extreme example of uncertainty. In that sense, we can easily notice that not only the average but also the spread of the entropy's distribution informs intelligibility.

Therefore, under this scenario, it is clear that the estimation of the spread (variance) of the distribution should not take a secondary role, mainly justified by the need of appropriate inferences, but it should also take a center role. Moreover, because the entropy data is bounded, it cannot be modeled using the mnormality assumption. Finally, they don't provide information to one another, and therefore, they are estimated independently form each other.

Considering the previous, it is clear that more sophisticated statistical procedures, that integrates all these pieces of information, could improve our intelligibility estimates (McElreath; 2020).

Second, although the literature suggest the entropy scores capture the intelligibility of a child, it is clear that they can still be considered a surrogate of what it intends measure. Notice the previous can be stated because we observe multiple entropy scores per child, and that these scores are not "intelligibility" scores, but entropy scores. Therefore we can say that these multiple outcomes are a manifestation of a child's intelligibility, but such manifestation is measured with uncertainty (error), i.e. there is an unobserved (latent) intelligibility construct that is responsible for what it is observed on the entropy scores and their variation.

Therefore, if we hope to understand or intervene on the factors that drives speech intelligibility, first one needs to "construct" a children' *intelligibility* scale (Carroll; 2006), allowing us to test our research hypotheses at the appropriate level. Furthermore, the literature suggest that failing to model this phenomena as a "latent construct" would lead us to incorrect inferences (deHaan et al.; 2019).

Considering all of the above, the aim of this reasearch is to propose a novel analysis of the entropy data using a Bayesian implementation of the Generalized Linear Latent and Mixed Model (GLLAMM) (Rabe-Hesketh et al.; 2004a,c,b, 2012; Skrondal and Rabe-Hesketh; 2004). The statistical procedure offers four benefits. First, it allows to appropriately model the bounded nature of the entropy data. Second, it provides a way to "construct" the speaker's latent intelligibility scale. Third, it allow us to test our research hypothesis at the appropriate level. And fourth, as a result from the first two, we successfully avoid producing false confidence in the parameter estimates, which help us to produce better informed statistical inferences (McElreath; 2020).

We find that when the proposed method is used to investigate the speech intelligibility levels of normal hearing (NH) versus hearing-impaired children with cochlear implants (HI/CI), in a data composed of ten utterances recordings from thirty two NH and HI/CI children selected from a large corpus of spontaneously spoken speech collected by the CLiPS research center, it brings bring new insights about the use of replicated entropy scores to measure intelligibility. Furthermore, the method also provide a way to assess how some factors affect the (under)development of children's intelligibility.

Bibliography

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.

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.

Carroll, J. (2006). *Measurement error in nonlinear models: a modern perspective*, Chapman and Hall/CRC.
doi: https://doi.org/10.1201/9781420010138.

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.

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

Finch, W., Bolin, J. and Kelley, K. (2019). *Multilevel Modeling Using R (2nd edition)*, Chapman and Hall/CRC.
doi: https://doi.org/10.1201/9781351062268.

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.

Gudivada, A., Rao, D. L. and Gudivada, V. N. (2018). Linguistics: Core concepts and principles, *in* V. N. Gudivada and C. Rao (eds), *Computational Analysis and Understanding of Natural Languages: Principles, Methods and Applications*, Vol. 38 of *Handbook of Statistics*, Elsevier, chapter 1, pp. 3–14.
doi: https://doi.org/10.1016/bs.host.2018.07.005.
url: https://www.sciencedirect.com/science/article/pii/S0169716118300208.

Guilford, J. (1954). *Psychometric methods*, McGraw-Hill Book Company.

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.

Jenkins, S. (2000). Cultural and linguistic miscues: a case study of international teaching assistant and academic faculty miscommunication, *International Journal of Intercultural Relations* **24**(4): 477–501.
doi: https://doi.org/10.1016/S0147-1767(00)00011-0.
url: https://www.sciencedirect.com/science/article/pii/S0147176700000110.

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.

Lesterhuis, M. (2018). *The validity of comparative judgement for assessing text quality: An assessor's perspective*, PhD thesis, University of Antwerp.

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

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

Munro, M. (1998). The effects of noise on the intelligibility of foreign-accented speech, *Studies in Second Language Acquisition* **20**(2): 139–154.
doi: https://doi.org/10.1017/S0272263198002022.

Munro, M. and Derwing, T. (1998). The effects of speaking rate on listener evaluations of native and foreign-accented speech, *Language Learning* **48**(2): 159–182.
doi: https://doi.org/10.1111/1467-9922.00038.
url: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-9922.00038.

Munro, M. and Tracey, D. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners, *Language Learning* **49**(s1): 285–310.
doi: https://doi.org/10.1111/0023-8333.49.s1.8.
url: https://onlinelibrary.wiley.com/doi/abs/10.1111/0023-8333.49.s1.8.

Ockey, G., Papageorgiou, S. and French, R. (2016). Effects of strength of accent on an l2 interactive lecture listening comprehension test, *International Journal of Listening* **30**(1-2): 84–98.
doi: https://doi.org/0.1080/10904018.2015.1056877.

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.

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.

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.

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.

Rowe, B. and Levine, D. (2018). *A Concise Introduction to Linguistics*, Routledge.

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.

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

Smith, L. and Nelson, C. (1985). International intelligibility of english: directions and resources, *World Englishes* **4**(3): 333–342.
doi: https://doi.org/10.1111/j.1467-971X.1985.tb00423.x.
url: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-971X.1985.tb00423.x.

Stevens, S. (1946). On the theory of scales of measurement, *Science* **103**(2684): 677–680.
doi: 10.1126/science.103.2684.677.
url: https://www.science.org/doi/abs/10.1126/science.103.2684.677.

Trochim, W. (2022). The research methods knowledge base.
url: https://conjointly.com/kb/.

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.

Varonis, E. and Susan, G. (1985). Non-native/non-native conversations: A model for negotiation of meaning, *Applied Linguistics* **6**(1): 71–90.
doi: https://doi.org/10.1093/applin/6.1.71.
url: https://academic.oup.com/applij/article-pdf/6/1/71/9741729/71.pdf.

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.