Evaluating selection criteria for functions mapping objective speech intelligibility predictions to subjective scores

Author: Berken Tekin (B.Tekin@student.tudelft.nl)
Supervisors: Jorge Martinez Castaneda, Dimme de Groot

Introduction

Background

- Objective speech intelligibility metrics (OIMs) aim to predict how understandable the average listener would find a given speech. This prediction does not necessarily show linear correlation with subjective word correct ratio (WCR) scores [1][2].
- Mapping functions can be used to linearize the relationship, and measure the accuracy of OIMs. To do so, a suitable non-linear model needs to be selected.

Existing methodology

Models based on logistic functions are emphasized in literature, for cases where speech files are categorized by one type of noise, due to the sigmoid relationship between noise levels and subjective word correct ratios [3].

The Problem

Existing methodology treats speech intelligibility as a function of the signal-to-noise ratio (SNR). As a result, if noise levels are kept the same for all samples, but other scalar characteristics in speech are scored by participants and OIMs, one would need different methods to create models. Furthermore, the relationship between noise levels and an OIM has to be modelled separately for the logistic curve to be a viable starting point.

How can a model that maps objective speech intelligibility scores to word correct ratios be validated on different data sets?

Possible Solution

Akaike Information Criterion (AIC) [4] allows evaluating any candidate model mapping scores from an OIM to WCRs, without relying on a particular underlying listening condition (comparable or categorizable features of speech samples).

AIC employs information theoretical techniques to estimate how well a mapping function would model *outside data*. The aim is to find the model with the best chance of fitting unseen speech samples.

The Experiment

- ALLSSTAR Corpus [5]: speech sample data set categorized by 7 different signal-to-noise ratio (SNR) levels, with a total of 29980 provided ratings
- Scores from 4 different OIMs and subjective metrics are averaged per SNR, and 7 data points are created for each WCR-OIM relationship.
- New mapping functions with more variables are designed to follow the data more closely.
- A version of AIC adjusted for small sample sizes [6] (n=7 in this case) is used for the experiment
- The aim is to see if the new criterion will prefer the adjusted functions.

Mapping functions from research:

$$L_2(d;a,b) = \frac{1}{1+e^{ad+b}}$$

$$L_3^{\text{Taal}}(d;a,b,c) = \frac{1}{1+e^{a\log(d+c)+b}}$$

$$S_2(d;a,b) = (1-e^{-ad})^b$$

Adjusted mapping functions:

$$L_3(d;a,b,h) = \frac{1-h}{1+e^{ad+b}}$$

$$L_4(d;a,b,h,h^*) = h^* + \frac{1-h}{1+e^{ad+b}}$$

$$L_4^{\text{Taal}}(d;a,b,c,h) = \frac{1-h}{1+e^{a\log(d+c)+b}}$$

$$S_3(d;a,b,h) = (1-h)(1-e^{-ad})^b$$

Results

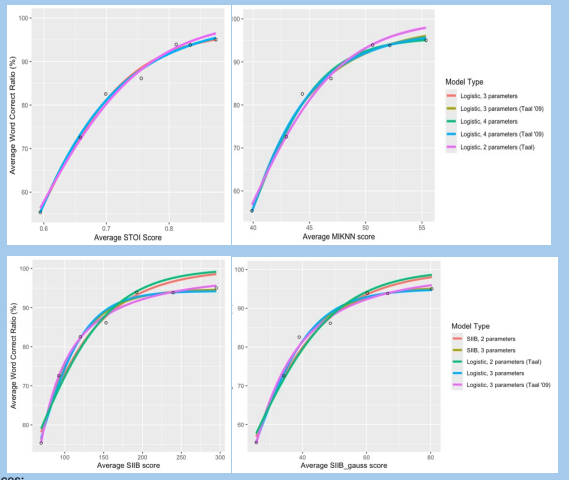
- For two objective speech intelligibility metrics, the functions recommended by the authors are validated
- The difference in two and three parameter fits are visible in graphs, but the effect of four parameters is not clear.
- For two other metrics, multiple mapping functions were found viable, but there is not strong enough evidence to reject recommended mapping functions for the given data set
- Additional evaluation criteria are unlocked with Akaike weights (See Table) that convert absolute metrics into a probability of being the best model among candidates.
- No specific preference towards adjusted functions; they never score higher than the ones in literature when they share the same number of free variables.

TABLE I: Root Mean Square Errors, Pearson Correlation Coefficients, △AICc values and Akaike weights for Experiments IV-A and IV-B. K is the number of free variables in the model. Models with ⋆ are suggested in the paper introducing the OIM, ✓ indicates the best model according to AIC_c. Bold results indicate selected models where $\triangle_i \leq 2$

Function	K	Method	RMSE	ρ	△AICc	wAICc
STOI						
L_2 ⋆ ✓	2	nlsLM	0,014	0,994	**0,000**	**0,788**
L_3	3	nlsLM	0,012	0,996	4,215	0,096
L_4	4	nlsLM	0,012	0,996	17,839	0,000
L_3^{Taal}	3	nlsLM	0,011	0,996	3,827	0,116
L_4^{Taal}	4	nlsLM	0,011	0,996	17,820	0,000
MIKNN						
L_2 ⋆ ✓	2	nlsLM	0,020	0,990	**0,000**	**0,400**
L_3	3	nlsLM	0,012	0,996	**0,116**	**0,377**
L_4	4	nlsLM	0,012	0,996	14,108	0,000
L_3^{Taal}	3	nlsLM	0,013	0,995	**1,170**	**0,223**
L_4^{Taal}	4	nlsLM	0,013	0,995	14,649	0,000
SIIB						
S_2 ⋆	2	nlsLM	0,026	0,984	4,192	0,082
S_3	3	nlsLM	0,014	0,995	2,641	0,178
L_2	2	nlsLM	0,032	0,978	7,115	0,019
L_3	3	nlsLM	0,016	0,993	4,935	0,056
L_3^{Taal} ✓	3	nlsLM	0,011	0,996	**0,000**	**0,665**
SIIB_{gauss}						
S_2 ✓	2	nlsLM	0,021	0,989	**0,000**	**0,329**
S_3	3	nlsLM	0,013	0,995	**0,685**	**0,234**
L_2 ⋆	2	nlsLM	0,026	0,984	2,907	0,077
L_3	3	nlsLM	0,014	0,994	**1,359**	**0,167**
L_3^{Taal}	3	nlsLM	0,014	0,995	**1,056**	**0,194**

Future Work

- The given database has too few listening conditions, testing evaluation methodologies on more data points would strengthen results.
- How different listening conditions (LCs), such as the average sound pitch or reverbation, affect SNR must be further investigated. If any LC is shown to react to changes in another LC, that should be considered when designing surveys to collect subjective measurements. As an example, rese
- The “holy grail” is to devise an objective intelligibility metric, able to estimate word correct ratios for any speech sample.



References:

1. Taal, Cees H., Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. ‘An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech’. IEEE Transactions on Audio, Speech, and Language Processing 19, no. 7 (2011): 2125–36. https://doi.org/10.1109/tasl.2011.2114881.
2. Van Kuyk, Steven, W. Bastiaan Kleijn, and Richard C. Hendriks. ‘An Instrumental Intelligibility Metric Based on Information Theory’. IEEE Signal Processing Letters 25, no. 1 (2018): 115–19. https://doi.org/10.1109/lsp.2017.2774250.
3. Taal, Cees H., Richard C. Hendriks, Richard Heusdens, Jesper Jensen, and Ulrik Kjems. ‘An Evaluation of Objective Quality Measures for Speech Intelligibility Prediction’, 1947–50, 2009.
4. Akaike, H. ‘A New Look at the Statistical Model Identification’. IEEE Transactions on Automatic Control 19, no. 6 (December 1974): 716–23. https://doi.org/10.1109/TAC.1974.1100705.
5. Bradlow, A. R. (n.d.) ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/allsstar
6. Burnham, Kenneth P., David Raymond Anderson, and Kenneth P. Burnham. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer, 2002.