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

Berken Tekin[1]

Supervisor(s): Jorge Martinez Castaneda[1]**, Dimme de Groot**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfillment of the Requirements
For the Bachelor of Computer Science and Engineering
January 26, 2025

Name of the student: Berken Tekin
Final project course: CSE3000 Research Project
Thesis committee: Jorge Martinez Castaneda, Dimme de Groot, Przemyslaw Pawelczak

Abstract

Objective speech intelligibility metrics (OIMs) are widely used in various fields, including public service announcements. These metrics do not directly predict the intelligibility of a speech (defined as the ratio of understandable words in an audio sample), but produce values that tend to monotonically increase with intelligibility. Several mapping functions, typically logistic models, are applied to raw objective scores to produce accurate predictions. However, there exists no standard methodology for choosing the best mapping curve, therefore, researchers tend to reuse curves originally meant for other datasets and OIMs. This research applies a method called Akaike Information Criterion (AIC), specifically developed for model selection, to existing candidate models as well as new ideas based on simple heuristics. Afterwards, the models are evaluated using AIC. The new criterion affirmed the logistic mapping functions chosen for the objective intelligibility metrics STOI and MIKNN, and highlighted alternative models for the SIIB and SIIB_{gauss}. However, with too few listening conditions on the dataset, strong inferences could not be easily made from the data.

I. Introduction

Speech intelligibility can be described as a measure of the understandable portion of a speech sample. High intelligibility of speech is especially important when there is little room for ambiguity and misunderstandings, such as critical emergency announcements.

Measurement of speech intelligibility can be done by conducting surveys with participants [1]. Listening tests are well studied to the point that official institutions provide guidelines to facilitate scientifically sound subjective measurements [2]. Usually, percent Word Correct Ratios (WCR), words understood by a listener as a percentage of all words in a speech sample, are used for scoring. Unfortunately, collecting subjective word-correct ratios from large groups are "time-consuming and expensive" [1], and speech intelligibility cannot be evaluated on the spot with this method.

Better computing capabilities allowed the development of algorithms that mathematically approximate speech intelligibility. These algorithms are known as *Objective Intelligibility Metrics (OIMs)*. Different OIMs utilize a variety of methods to predict subjective speech intelligibility metrics, from information-theoretic approaches [3] to (simpler) frequency domain comparisons [4]. Nonetheless, they all aim to follow a monotonic relationship with subjective data [5].

However, an ideal intelligibility metric would not just rank different samples correctly; it would be able to reliably estimate a word-correct ratio from a given speech sample. The capacity of an OIM to predict WCRs is commonly measured [4]–[13] by first fitting a monotonic function with free variables ("the model") to capture the relationship between objective and subjective metrics. Afterwards, the goodness of fit of the model is measured using several performance metrics (the procedure is detailed in Section II). However, a particular monotonic function is chosen based on heuristics in many cases, and it is not always clear from the research why a certain monotonic mapping function is preferable to others for a given OIM.

Then *how can researchers confidently use a provided mapping function for a given OIM with their own dataset?* If a provided model is biased towards trends in the provider's dataset, performance metrics for different datasets would also show a propensity towards rewarding features found in the provider's dataset, to the extent they are carried through the mapping function. Such unintended effects may be prevented with careful model selection.

In this paper, an information theoretical metric called Akaike's Information Criterion (AIC) [14] is explored for its utility in selecting the most suitable model that can map objective metrics to word-correct ratios. In Section II, the data set and the proposed criterion are introduced. The model selection criteria described in [6] are also compared with the proposed criterion. In Section III, more detail is given on the programming environment used to run the experiments. Two experiments are done in section IV, and the AIC best fits for given OIMs are compared to original mapping function recommendations for each OIM. Section V explains the results of the experiment. Section VI clarifies that all data is licensed, ethically sourced and reproducible. In Section VII, limitations related to the research are explained. Finally, VIII shortly summarizes the research question, recommended solution and the next steps.

II. Methodology

In this section, first, the contents of a data set with Word Correct Ratios (WCRs) are summarized. The contemporary methodology for selecting a function that maps objective intelligibility scores to WCRs is then explained, and a different methodology is proposed.

A. Dataset with subjective word correct ratios

An ideal data set for speech intelligibility evaluation would have a variety of listening conditions that could be used to perform more comprehensive tests on objective intelligibility metrics. However, the only available data set with subjective WCRs not used by other members of the project group is the ALLSSTAR corpus (L1 Native Speakers):

- Data obtained from English speakers [15].
- The portion used for the project contains HINT sentences uttered [16].
- 26 native English speakers with ages ranging from 18 to 23 years
- 120 sentences repeated by 23 native speakers, 110 sentences repeated by 2 more native speakers.
- 250 participants rate clean files and noised files with signal-to-noise ratios (SNR) ranging from -4 to 8. Only one type of noise is used (Speech-shaped white noise). In total, 29980 speech samples (each with one sentence) is rated.

Listening Conditions

The clean sentences are processed to create seven noisy samples with different Signal-to-Noise Ratios (SNRs) before being scored by 250 participants. As only one type of noise is used, 7 different listening conditions can be reproduced from ALLSSTAR, one for each level of SNR.

B. Current Model Selection Criteria

The correlation between objective intelligibility metrics (OIMs) and word correct ratios (WCRs) is determined by first fitting a monotonically increasing function ("the model") to the data with a nonlinear least squares (NLS) procedure (see Section III for the NLS procedure used for the experiments). Afterwards, OIM scores (P) are mapped to estimated word correct ratios ($\hat{P}$) using the fitted model. Finally, performance metrics such as the root mean square of prediction errors (RMSE) between $\hat{P}$ and actual word correct ratios (S):

$$\mathrm{RMSE}(S,\hat{P}) = \sqrt{\frac{\sum_{j=0}^{N-1}(s_j - \hat{p}_j)^2}{N}},$$

and the Pearson Correlation Coefficient (ρ), a linear correlation measurement[1]

$$\rho_{S,\hat{P}} = \frac{\sum_{j=1}^{n}\left(s_j - \overline{s}\right)\left(\hat{p}_j - \overline{\hat{p}}\right)}{\sqrt{\sum_{j=1}^{n}\left(s_j - \overline{s}\right)^2}\sqrt{\sum_{j=1}^{n}\left(\hat{p}_j - \overline{\hat{p}}\right)^2}}$$

are calculated to test the goodness of fit achieved by a given mapping function.

However, research does not always include a clear roadmap to choose a suitable monotonic fit. The same mapping function may be used for different metrics (see [4], [8], [13]), or a later paper may suggest a new mapping function for an earlier algorithm (see [4], [17]). However, the reasoning behind such modifications is not always elaborated.

As a notable exception, in [6] it is suggested that due to the logistic relation between signal-to-noise ratios (SNR) and word-correct ratios (WCR), this shape could be composed with the shape of the relationship formed by the SNR and the OIM, to create a suitable candidate model for an adequate fit. For example, OIMs that linearly correlate with SNR values could comfortably utilize a logistic model to test the OIM. However, the proposed guideline has certain drawbacks:

1) Although the logistic curve may be a good selection for WCR-SNR graphs with one type of noise, as more listening conditions are associated with each SNR level, other monotonic relationships may better explain the correlation. As an example: Speech intelligibility algorithms are known to correlate less with enhanced noisy speech [18], and if there is a monotonic relationship between the OIM and the SNR, this means that the WCR-SNR relationship would also be affected.
2) If such a transformation is applied to data with a single listening condition per SNR, this essentially defeats the purpose of OIMs, as the SNR measure itself would be just as good at predicting WCRs.
3) Even if the WCR-SNR relationship is indeed logistic, one would still need to select suitable candidate models for the OIM-SNR relationship for anything more complicated than a linear or exponential relationship. The shape of such a model would depend on the inner workings of the OIM.
4) When the WCR value per SNR level is utilized as a ground truth to derive accurate mapping functions, this requires the SNR values to always be within the set of listening conditions. This may not be too inconvenient for researchers, but it is an arbitrary constraint.

Thankfully, there exists a model selection method that makes no assumptions about any particular listening condition and provides an easy way to evaluate different candidate models.

C. Akaike Information Criterion (AIC)

Akaike Information Criterion [14] is one of the most well-known model selection criteria [19], and has been used in speech intelligibility research [20], as well as a variety of other applications [21]. AIC is developed to estimate the difference between a candidate model and a "true model" that is assumed to generate the data fitted by the candidate model. This difference is called the "Kullback-Leibler divergence", and AIC is able to estimate that parameter even without information about the true model [19]. For models fitted to the same data, the lowest AIC score indicates the model that best captures the true model. However, for small sample sizes, there exists a modified AIC with a corrective term to prevent overfitting toward models with more parameters (e.g. [22]):

$$\mathrm{AIC_c} = \mathrm{AIC} + \frac{2k^2 + 2k}{n - k - 1}$$

where k is the number of free parameters in the model and n is the number of data points [23]. If $n/k \leq 40$, $\mathrm{AIC_c}$ should be used instead of AIC [23]. For the database at hand, n is 7 (see II-A), therefore, $\mathrm{AIC_c}$ is used to evaluate the models in this paper instead of AIC. However, the terms will be used interchangeably for descriptions.

The value produced by AIC is not an absolute measure of quality, it changes based on factors not related to the model itself, including the sample size of the data [23]. The AIC values make sense only in relation to the AIC values of other candidate models fitted to the same data set. Therefore, for all candidate models, the AIC value is normalized to obtain a relative scoring w.r.t. the minimum AIC [23], [24]:

$$\triangle_i = \mathrm{AIC}_i - \mathrm{AIC}_{\min}$$

Each $\triangle_i$ is a measure of "information loss" about the underlying data compared to the best model among the candidates, with the best model having $\triangle_{\min} = 0$ [23]. A model selection "rule of thumb" has been recommended for AIC, where candidates with $\triangle_i \leq 2$ are strong alternatives to the best model, $4 \leq \triangle_i \leq 7$ implies less support for the candidates, and $\triangle_i > 10$ means the model is not a

[1]The mapping function linearizes the WCR-OIM graph.

suitable candidate [23]. However, the exact cut-off points are controversial [25], and therefore the focus will be on the "best model" (in the Kullback-Leibler sense) selected by AIC.

$\triangle_i$ may also be used to calculate Akaike weights w_i [23], [26]

$$w_i(\text{AIC}) = \frac{\exp\left\{-\frac{1}{2}\Delta_i(\text{AIC})\right\}}{\sum_{m=1}^{M} \exp\left\{-\frac{1}{2}\Delta_m(\text{AIC})\right\}}$$

which convert $\triangle_i$ values into probabilities. w_i weights add up to 1, and each w_i indicates the probability that the model i is the best model, in a Kullback-Leibler sense, among the candidate models given (e.g. [27]).

D. Description of experiments

1) The objective intelligibility metrics STOI [4] and MIKNN [8], despite using different theoretical frameworks to calculate speech intelligibility, are mapped using the same logistic curve. However, both authors utilize a two-parameter logistic curve (L_2) in their papers. In experiment IV-A, several different logistic functions with different parameters will be tested to see whether L_2 is indeed optimal.
2) There exist two related OIMs: SIIB and SIIB$_{gauss}$. Despite both being based on information theory, different models are suggested by the author to fit these OIMs to the WCR data. In experiment IV-B, the proposed criterion is used to examine how strongly each model is preferred by the respective OIM for the data set at hand.

III. Experimental Setup

Speech samples are processed with Praat software, version 6.4.25 (December 8, 2024) [28], and the objective metrics are calculated using Python 3.13. The fittings and plots of the model are made using R Statistical Software [29].

A. Praat

Praat is a free and open-source software that is used to manipulate and analyze audio files. During the project, Praat has been used for two purposes:

1) For ALLSSTAR database, a metadata file with a .TextGrid extension is distributed for each audio sample. The samples contain 60 sentences each, and the TextGrid file has timestamps for the beginning and end of each sentence. Using the scripting functionality of Praat, each file is split into 60 audio tracks, one per sentence. However, instead of scripting in Praat directly, the Parselmouth [30] API for Praat is used inside Python.
2) The ALLSSTAR maintainers provide several Praat scripts on their website. A script from the website that mixes a provided noise file with clean speech given at different SNR levels is used to generate degraded (noisy) audio samples.

B. Python

Python has been used primarily for calculating objective intelligibility metrics for clean and degraded speech samples. For each objective speech intelligibility metric, the following libraries are used:

- **STOI:** pystoi, based on MATLAB implementation of STOI [7],
- **MIKNN:** Executed within MATLAB engine [31], code available in an online repository, license permits academic use.
- **SIIB & SIIB$_{gauss}$:** pysiib, ported from MATLAB implementation [32]

For each calculated metric, a Polars [33] dataframe has been created with the file name, the SNR, and the objective intelligibility score as parameters. Afterwards, each dataframe is saved as a CSV file, before being merged with the word-correct ratios provided into one big dataset in R.

C. R Statistical Software

R Statistical Software, along with RStudio GUI [34], has been the main programming environment for the project. All performance metrics and graphs in this paper have been produced with R.

For fitting non-linear models into data, the `nlsLM` function inside `minpack.lm` package has been used [35]. R's built in `nls` uses Gauss-Newton algorithm by default, but the Levenberg-Marquardt algorithm used in `nlsLM` combines Gauss-Newton method with gradient descent, with more focus on gradient descent when the coefficients are far from the optimal parameters for the data [36]. The practical use for this property is that it is easier to find convergent starting parameters for `nlsLM`.

The graphs are drawn with ggplot2 [37].

IV. Experiments

A. Testing different numbers of parameters

For STOI [4] and MIKNN [11], the creators utilize a logistic function to measure the performance of their OIMs:

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

where d is the raw objective intelligibility score. Using the OIM and WCR values, the variables a and b are fitted with a least-squares method to generate the most performant logistic model $L_{\#\text{ parameters}}$ for the given dataset.

For some OIMs exponentially related to SNR, a modified logistic function is suggested that normalizes the OIM variable [6]:

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

One point of interest is that the subjective word correct ratios (at least for the database at hand) are not 100% even for clean speech, where OIMs get their theoretical maximum value. This may be due to issues with clean audio samples, momentary lapses of attention by the participants, or other random effect. Despite that, the asymptote of the

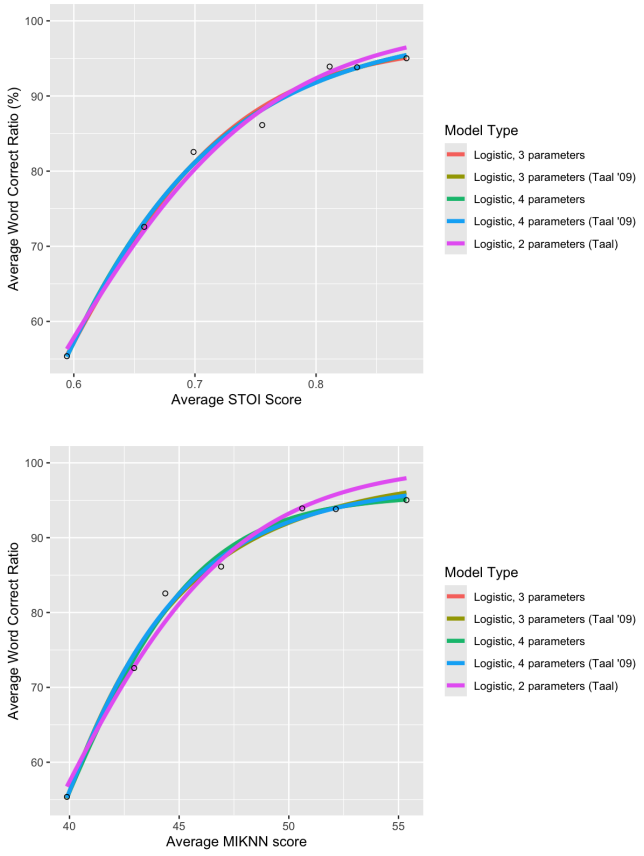

Fig. 1: A side by side view of models fitted for STOI and MIKNN objective intelligibility metrics in Experiment IV-A.

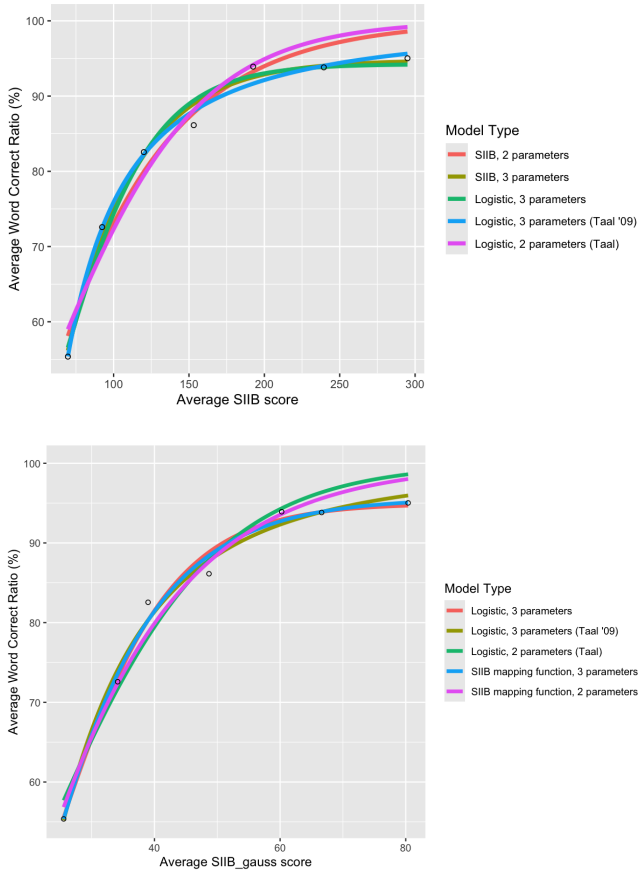

Fig. 2: Models in Experiment IV-B. The divergence between two and three parameter models is clearly visible.

given logistic function cannot be altered, and this limitation may cause suboptimal fits as more data points are added. Therefore, in this paper a different logistic model with three parameters is proposed, where a new h variable is added to allow calibrating the upper asymptote:

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

And the same idea is applied to L_3^{Taal}:

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

Finally, a four-parameter logistic function with an additional variable to calibrate the lower asymptote h^* is evaluated to observe in more detail how AIC_c scores models as the number of parameters increases:

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

The resulting fitted models are displayed in Figure 1, and their corresponding performance metrics and AIC_c values are shown in Table I.

B. Validating provided models

The SIIB metric estimates speech intelligibility by calculating how much information is shared between clean and degraded signals [32]. For the classical version where mutual information is estimated through a k-nearest neighbor algorithm, the author suggests a custom monotonic function:

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

Like in IV-A, a three-parameter version of the function with a variable for the upper asymptote is also examined:

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

However, there exists another version of SIIB (SIIB_{gauss}) in which a different method is used to estimate mutual information [5]. Although written by the same author as [32], a L_2 model is used for all OIMs in [5], including SIIB_{gauss}. To gain confidence in the choice of models, the S and L models are compared using AIC.

Figure 2 contains visualizations of the models fitted for SIIB and SIIB_{gauss}. Table I contains the measured model selection criteria.

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**

V. Results

Table I visualizes all performance metrics used to evaluate various mapping functions. RMSE and ρ values can be used as a measurement for the goodness of fit of a model to particular data points. However, such criteria would always prioritize models with more parameters to have the best possible fit for the specific dataset, with no penalty terms for overfitting. On the other hand, the (corrected) Akaike Information Criterion strikes a balance between parsimony and goodness of fit, and estimates the model that is the most generalizable to outside data.

In experiment IV-A, AIC_c prefers a logistic two-parameter fit for both STOI and MIKNN; however, L_2 is more likely to be the best fit for STOI. Examining Figure 1, one can observe that the differences in two- and three-parameter fits are more pronounced for MIKNN, which may explain the larger set of candidate models for MIKNN. However, since L_2 is also selected by the authors, L_2 can be thought of as the best predictor (in the Kullback-Leibler sense) for STOI and MIKNN among given models, despite yielding the largest RMSE and the smallest Pearson correlation coefficient ρ for both OIMs.

The results of experiment IV-B suggest that L_3^{Taal} does a particularly good job in preserving information on the underlying relationship between SIIB and WCR scores. However, that might also be the logarithmic transformation of SIIB scores by L_3^{Taal} resulting in a fitted model that simply lines up with this particular the WCR-SIIB relationship. According to the Akaike weight w_i, there is a 33.5% chance that L_3^{Taal} actually performs worse (in the Kullbach-Leibler sense) than a different candidate function. For SIIB$_{gauss}$, interestingly, the criterion highlights all models except the one chosen by the author for the metric. However, all values are within close vicinity, and no conclusions can be made about the best model for SIIB$_{gauss}$.

VI. Responsible Research

Objective intelligibility metrics aim to predict human responses. Both the voice sample datasets and subjective word correct ratios are, in essence, sensitive data. For this reason, it is important to ensure that all external data and code are licensed for at least research purposes.

The ALLSSTAR database is licensed under a Creative Commons Attribution 4.0 International License [38], allowing everyone to process and share the given speech samples. No other source for speech samples is used for this report. According to the project manual published by ALLSSTAR, all participants are paid, and are asked to sign a consent form.

The MATLAB code for MIKNN [8] has a license that permits use for research purposes.

Versions of all software used for processing and scoring audio files are shared in Section III. For both Python and R codes, virtual environments are used to guarantee reproducibility, and the code can run on any supported environment.

VII. Discussion

In Table I, only values with $\triangle_i \leq 2$ have been included in the set of candidate models, to only have models with "substantial support" that do not lose too much information about the underlying data, as proposed in [23]. However, a newer article by the same lead author actually relaxes this cut-off [25], and suggests that models with $\triangle_i \leq 7$ should also be considered. Of course, the best model among candidate models would not change; but researchers should not be too quick to disregard alternative models.

Interpolating candidate models from a psychometric WCR-SNR curve [6] is not the only model selection method in the literature. A more complicated statistical method to find a fitted model is described in [39]. However, such measures have not been researched for this project.

VIII. Conclusions and Future Work

Currently, there is no agreed upon methodology for selecting a model to linearize the relationship between an objective intelligibility metric and subjective word correct ratios. However, the Akaike Information Criterion (AIC) may be a good candidate.

In encountered research, AIC has never been used to choose the best model that combines objective intelligibility

scores with word correct ratios. And yet, it is a metric created exclusively for model selection with a strong theoretical basis, and it is used in a variety of fields. AIC and related model selection criteria are promising helpers in the pursuit of more accurate objective intelligibility metrics.

One of the first things that come to mind as future work is to run the experiments for multiple databases, more listening conditions, and combined datasets. If an OIM were particularly performant with one mapping function across different datasets, that would be an interesting observation.

References

[1] H.-T. Chiang, K.-H. Hung, S.-W. Fu, H.-C. Kuo, M.-H. Tsai, and Y. Tsao, "Study on the correlation between objective evaluations and subjective speech quality and intelligibility," 2023. [Online]. Available: https://dx.doi.org/10.48550/arxiv.2307.04517

[2] I. T. Union, "P.807," 2016.

[3] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An instrumental intelligibility metric based on information theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018. [Online]. Available: https://dx.doi.org/10.1109/lsp.2017.2774250

[4] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech." IEEE, 2010, Conference Proceedings, pp. 4214–4217. [Online]. Available: https://dx.doi.org/10.1109/icassp.2010.5495701

[5] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Evaluation of Intrusive Instrumental Intelligibility Metrics," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 11, pp. 2153–2166, 2018.

[6] C. H. Taal, R. C. Hendriks, R. Heusdens, J. Jensen, and U. Kjems, "An evaluation of objective quality measures for speech intelligibility prediction," in *Interspeech*, 2009, pp. 1947–1950.

[7] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[8] J. Taghia and R. Martin, "Objective Intelligibility Measures Based on Mutual Information for Speech Subjected to Speech Enhancement Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 6–16, 2014.

[9] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Communication*, vol. 52, no. 7, pp. 678–692, Jul. 2010.

[10] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI)," *Speech Communication*, vol. 65, pp. 75–93, Nov. 2014.

[11] J. Taghia and R. Martin, "Objective Intelligibility Measures Based on Mutual Information for Speech Subjected to Speech Enhancement Processing," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 1, pp. 6–16, 2014.

[12] J. M. Kates and K. H. Arehart, "The Hearing-Aid Speech Perception Index (HASPI) Version 2," *Speech Communication*, vol. 131, pp. 35–46, 2021.

[13] A. Edraki, W.-Y. Chan, J. Jensen, and D. Fogerty, "Speech Intelligibility Prediction Using Spectro-Temporal Modulation Analysis," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 210–225, 2021.

[14] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, Dec. 1974.

[15] A. R. Bradlow, "ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings," n.d., retrieved from the Northwestern University SpeechBox database. [Online]. Available: https://speechbox.linguistics.northwestern.edu/allsstar

[16] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise," *The Journal of the Acoustical Society of America*, vol. 95, no. 2, pp. 1085–1099, Feb. 1994.

[17] J. B. Boldt and D. P. W. Ellis, "A SIMPLE CORRELATION-BASED MODEL OF INTELLIGIBILITY FOR NONLINEAR SPEECH ENHANCEMENT AND SEPARATION."

[18] I. López-Espejo, A. Edraki, W.-Y. Chan, Z.-H. Tan, and J. Jensen, "On the deficiency of intelligibility metrics as proxies for subjective intelligibility," *Speech Communication*, vol. 150, pp. 9–22, 2023.

[19] J. E. Cavanaugh and A. A. Neath, "The Akaike information criterion: Background, derivation, properties, application, interpretation, and refinements," *WIREs Computational Statistics*, vol. 11, no. 3, p. e1460, May 2019.

[20] W. Hu, B. A. Swanson, and G. Z. Heller, "A Statistical Method for the Analysis of Speech Intelligibility Tests," *PLOS ONE*, vol. 10, no. 7, p. e0132409, 2015.

[21] P. Stoica and Y. Selen, "Model-order selection," *IEEE Signal Processing Magazine*, vol. 21, no. 4, pp. 36–47, Jul. 2004.

[22] N. Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike' s," *Communications in Statistics - Theory and Methods*, vol. 7, no. 1, pp. 13–26, Jan. 1978.

[23] K. P. Burnham and D. R. Anderson, "Multimodel Inference: Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, vol. 33, no. 2, pp. 261–304, Nov. 2004.

[24] N. Sugiura, "Further analysis of the data by Akaike's information criterion and the finite corrections: Further analysis of the data by akaike' s," *Communications in Statistics - Theory and Methods*, vol. 7, no. 1, pp. 13–26, Jan. 1978.

[25] K. P. Burnham, D. R. Anderson, and K. P. Huyvaert, "AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons," *Behavioral Ecology and Sociobiology*, vol. 65, no. 1, pp. 23–35, Jan. 2011.

[26] E.-J. Wagenmakers and S. Farrell, "AIC model selection using Akaike weights," *Psychonomic Bulletin & Review*, vol. 11, no. 1, pp. 192–196, Feb. 2004.

[27] K. P. Burnham and D. R. Anderson, "Kullback-Leibler information as a basis for strong inference in ecological studies," *Wildlife Research*, vol. 28, no. 2, p. 111, 2001.

[28] P. Boersma and D. Weenink, "Praat: doing phonetics by computer [computer program]," 2024, retrieved 8 December 2024 from http://www.praat.org/.

[29] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2024. [Online]. Available: https://www.R-project.org/

[30] Y. Jadoul, B. Thompson, and B. de Boer, "Introducing parselmouth: A python interface to praat," *Journal of Phonetics*, vol. 71, pp. 1–15, 2018.

[31] I. The MathWorks, *MATLAB Engine API for Python*, 2024. [Online]. Available: https://www.mathworks.com/help/matlab/matlab_external/get-started-with-matlab-engine-for-python.html

[32] S. Van Kuyk, W. B. Kleijn, and R. C. Hendriks, "An Instrumental Intelligibility Metric Based on Information Theory," *IEEE Signal Processing Letters*, vol. 25, no. 1, pp. 115–119, 2018.

[33] R. Vink and other contributors, "Polars: Lightning-fast dataframe library for rust and python," 2024. [Online]. Available: https://www.pola.rs/

[34] RStudio Team, *RStudio: Integrated Development Environment for R*, RStudio, PBC., Boston, MA, 2020. [Online]. Available: http://www.rstudio.com/

[35] Timur V. Elzhov, Katharine M. Mullen, Andrej-Nikolai Spiess, Ben Bolker, "Minpack.lm: R Interface to the Levenberg-Marquardt Nonlinear Least-Squares Algorithm Found in MINPACK, Plus Support for Bounds," pp. 1.2–4, Apr. 2022.

[36] H. P. Gavin, "The Levenberg-Marquardt algorithm for nonlinear least squares curve-fitting problems," 2024.

[37] H. Wickham, *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016. [Online]. Available: https://ggplot2.tidyverse.org

[38] A. R. Bradlow, "Speechbox," n.d., retrieved from https://speechbox.linguistics.northwestern.edu.

[39] K. Yamamoto, T. Irino, S. Araki, K. Kinoshita, and T. Nakatani, "GEDI: Gammachirp envelope distortion index for predicting intelligibility of enhanced speech," *Speech Communication*, vol. 123, pp. 43–58, Oct. 2020.