ELO-SPHERES intelligibility prediction model for the Clarity Prediction Challenge 2022

Mark Huckvale, Gaston Hilkhuysen

Speech, Hearing and Phonetic Sciences, University College London, UK

m.huckvale@ucl.ac.uk, g.hilkhuysen@ucl.ac.uk

Abstract

This paper describes and evaluates the ELO-SPHERES project sentence intelligibility model for the Clarity Prediction Challenge 2022. The aim of the model is to make predictions of the intelligibility of enhanced speech to hearing impaired listeners. Input to the model are binaural processed audio of short sentences generated in a simulated noisy and reverberant environment together with the original source audio. Output of the model is a prediction of the intelligibility of each sentence in terms of percentage words correct for a known hearing-impaired listener characterized by a pure-tone audiogram. Models are evaluated in terms of the root mean squared error of prediction. We approached this problem in three stages: (i) evaluation of the influences of the scene metadata on scores, (ii) construction of classifiers for estimation of scene metadata from audio, and (iii) training a non-linear regression model on the challenge data and evaluation using 5-fold cross validation. On the test data, a baseline system using only the standard short-time objective intelligibility metric on the better ear achieved a RMS prediction error of 27%, while our model that also took into account given and estimated scene data achieved an RMS error of 22%.

Index Terms: speech-in-noise, speech intelligibility, hearing aids, hearing loss, machine learning

1. Introduction

The Clarity Prediction Challenge 2022 [1] was an open competition to compare the performance of speech intelligibility metrics. The materials for the prediction challenge were generated from a previous enhancement challenge in which teams competed to process noisy speech for known hearing-impaired (HI) listeners. The goal of the prediction challenge was to predict the intelligibility of some of the enhanced sentences by these listeners.

For training the model, recordings of sentences containing 7-10 words collected in a set of 402 noisy scenes and enhanced with 10 enhancement systems were provided. There were seven interfering noises: (dishwasher, fan, hairdryer, kettle, microwave, vacuum, washing machine) mixed with the source sentence at seven signal-to-noise ratios (-6, -4, -2, 0, 2, 4, 6dB). The enhanced signals were presented to 27 HI listeners characterised by their audiometric data. Responses of each listener were recorded in terms of the target transcription, listener transcription, and number of words correct. Audio of both the source signal and processed signal are provided as two-channel files for the left and right ears of the simulated listener. In total there are 4863 training samples. More details can be found in [2].

Previous proposals for predicting the intelligibility of speech to HI listeners have adapted existing intrusive signal metrics for normal listeners by adding a correction to account for audibility by the HI listener. Examples are those of Magnusson [3] and Ching et al [4] which looked at modifications to the Speech Intelligibility Index (SII), while the HASPI metric [5] combines an intrusive metric with a model of impaired auditory processing. A comparison of some audibility-adapted metrics can be found in [6]. However, studies have shown that audibility alone is insufficient to explain the differences between impaired listeners. This inadequacy was shown clearly in the study by the authors [7] in which variations in audiograms only accounted for 40% of the variation in intelligibility performance across listeners. This has influenced our approach to the Clarity Challenge, in which we try to exploit all available factors that might influence intelligibility, not just changes to the signal and the listener audiogram.

Our approach to building an intelligibility prediction model for the challenge comprised three stages:

a. Explore which features of the speech, audio and listener metadata have impact on the speech intelligibility
b. Train classification models that attempt to recover metadata from the audio signals
c. Train a non-linear regression model that takes as input the source and processed audio, the given speech and listener metadata, and the estimated metadata to predict the intelligibility of a given speech recording to a given listener.

In the following sections we describe these stages and the performance of the model.

2. Metadata Exploration

To establish the size of the influence of each of the source factors, the mutual information (MI) between the factor and %Correct was measured using the MPMI toolbox [8]. Mutual information quantifies the "amount of information" obtained about one random variable from observation of another random variable. MI is preferred here to correlation because it does not presume a linear relationship between the two variables, and we expect to use non-linear regression for prediction. To gain better estimates of MI, a cross-validation process is used to correct for bias.

We investigate the following factors: the STOI metric value [9] calculated from the source and processed audio in the better of the two ears; the choice of processing system used to enhance the signal (from 10 choices); the azimuth of the source w.r.t the listener and the difference in azimuthal angle between the source and the interferer; the probability of the text in the prompt sentence; the signal-to-noise ratio of the source to the interferer; the pure-tone average of the listener in

each ear; the identity of the listener (from 27 choices); the choice of interfering noise type (from 7 choices); and the listener performance on the digit-triple test [10]. Table 1 provides the bias-corrected mutual information value between each factor and the intelligibility score.

To calculate the STOI metric value from the supplied reference and processed audio, the signals were first aligned with the `sigalign` function in VOICEBOX [11], then the STOI metric value was calculated for each ear independently using the reference implementation [12]. Finally, the larger of the two values was selected.

To calculate the probability of the prompt sentence, a simple trigram language model was built from texts in the British National Corpus [13]. The corpus was pre-processed to remove all punctuation except sentence mark-up. Upper case, lower case and mixed case words were merged. Each marked sentence in the BNC was divided into trigrams, which were sorted and counted. In total 97,881,081 trigrams were found, of 41,429,470 types. The source sentences in the Clarity data set were then divided into trigrams and the relative frequency of each trigram in the BNC was used to establish a log probability for each sentence. The average log probability per word in the sentence was then used as an influencing factor on intelligibility. The correlation between %Correct and mean log word probability was r=0.28.

Table 1. *Mutual intelligibility values for a number of possible factors influencing sentence intelligibility*

Influence on %Correct	MI (nats)
STOI value in better ear	0.267
Choice of processing system	0.222
Azimuth angle between source and interferer	0.098
Setting of listening volume	0.075
Sentence text probability	0.058
Signal-to-noise ratio of stimulus	0.029
Pure tone average in worse ear	0.026
Identity of listener	0.025
Identity of talker	0.019
Choice of interferer noise type	0.018
Pure tone average in both ears	0.017
Listener performance on digit triple test	0.017
Pure tone average in better ear	0.015
Azimuth of source	0.015

As expected, the STOI value and choice of processing system had a strong influence on % correct. Somewhat less useful are the azimuthal angle between target and interferer sources in the scene, the listening volume setting, and the sentence text probability. Plots of these are shown in Figures 1-3 where points are individual sentences, and the line a loess regression.

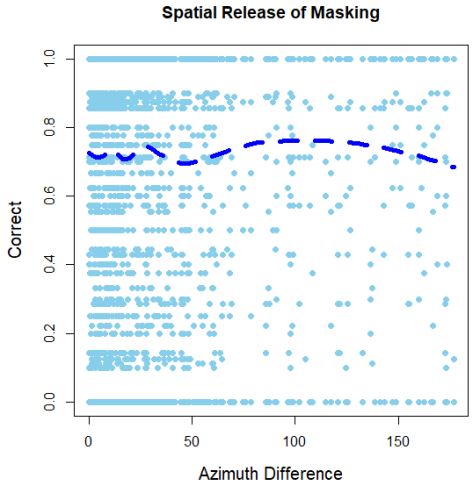

Figure 1. *Influence of difference in azimuth of target and interferer on proportion correct.*

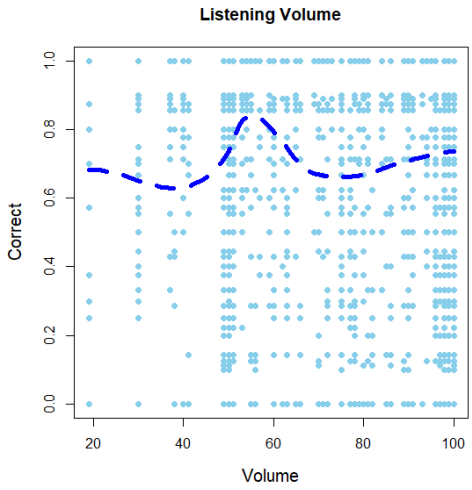

Figure 2. *Influence of listening volume on proportion correct.*

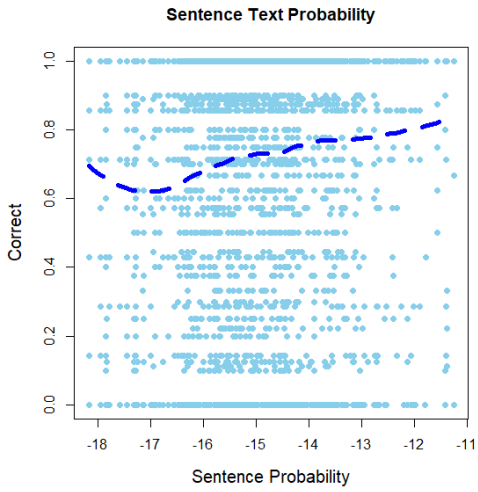

Figure 3. *Influence of prompt text probability (mean log probability per word) on proportion correct.*

It can be seen that the effects of azimuthal difference (Fig 1) is particularly due to the fact that most stimuli have angular separations of less than 90°. The influence of listening volume (Fig 2) is probably an artefact caused by the fact listening volume setting differed across listeners. The influence of sentence text probability (Fig 3) is small but shows increasing score with increasing probability, i.e. that more probable sentences were better recognized.

Unexpectedly, the influence of the listener was small, whether that was done by identifying the person, or through their pure-tone average, or through performance on the digit triple test. This may be because the enhancement systems had already modified their spectral characteristics for each listener. The identity of the talker producing the sentences had a small effect.

The type of noise interferer and the mixture SNR had very little effect, presumably because the processing systems were effective at removing noise.

3. Scene Classification

For evaluation of the prediction model on test data, the challenge only provides the target and processed audio, the identity of the listener, and the prompt sentence text. However the analysis in the previous section showed that factors like the identity of the processing system, the difference in angle between source and interferer, and the identity of the talker are also potentially important for predicting intelligibility. Thus we investigated the extent to which these factors could be predicted from the audio signals alone.

Our scene metadata classification system took as input 100 frames (2.56s) of filterbank energies selected from the filterbank used in the calculation of the STOI metric [12]. The STOI filterbank comprises 15 third-octave frequency bands from 150Hz to 4800Hz, with the amplitude measured in 51.2ms windows every 25.6ms.

For classification of the processing system, a convolutional neural network was used, with inputs from both the target and processed audio and filterbank channels interleaved. Two convolutional layers with an 8x8x4 kernel were fed into a max-pooling layer and a long short-term memory (LSTM) layer and finally into a dense layer with 10 softmax outputs. A validation set of 10% of the training set was used to halt training, and on the test set, the classifier was able to identify which processing system was used for a scene with an accuracy of 97%.

For classification of the talker, the same convolutional neural network was used, but with only the target audio as input. Output was a dense layer with 6 softmax outputs. Using 5-fold cross-validation on the training set, the classifier was able to identify which talker was used for a scene with an accuracy of 95%.

For prediction of the azimuthal angle difference the same convolutional network was used on both target and processed audio, but with a linear output. However no useful prediction of angle was obtained, this is likely because this information has little presence in the processed audio.

4. Baseline Models

To better understand the performance of our regression model we implemented four baseline models for predicting % correct from the supplied data:

NULL – a single % correct prediction based on the mean score over all scenes, listeners and systems.

LISTENER – a single % correct prediction for each listener, based on their mean performance over all scenes and systems.

SYSTEM – a single % correct prediction for each system, based on their mean performance over all scenes and listeners.

STOI – a regression model that predicted proportion of words correct from the reference and processed audio alone using the STOI metric (from the better ear). The STOI metric value was converted to a proportion correct score using logistic regression weighted by the number of words in each sentence. The regression model was fitted and tested on the training set using 5-fold cross-validation.

Performance of these baseline models on both training and test data is shown in Table 2. The RMS prediction error of 27% using STOI on the best ear provides a good estimate of the prediction error found using a current state of the art approach.

Table 2. *RMS error for baseline predictors*

Baseline method	RMS Prediction Error (%)	
	Train	**Test**
NULL	36.452	36.380
LISTENER only	35.584	35.375
SYSTEM only	27.402	27.173
STOI best ear	27.081	27.404

5. Regression Model

5.1. Input Features

Given the outcome of the metadata analysis and the outputs of the scene classification models, the following features were used to construct a regression model to predict percentage correct intelligibility:

STOIFILT (15 features) – STOI correlations between source and processed audio per filter channel. The target and processed signals are first aligned by spectral cross-correlation before calculation of the STOI correlations. The set of correlations is chosen from which ear delivered the better STOI value overall.

SYSTEM (10 features) – predicted identity of the processing system found by the scene classifier, one probability per system.

LISTENER (27 features) – identity of the hearing-impaired listener as one-hot vector. This is generated from the given metadata.

TALKER (6 features) – predicted identity of the talker of the sentence used found by the scene classifier, one probability per talker.

SPROB (1 feature) – Prompt sentence text probability. This is calculated from word trigram frequencies of the words in the prompt in the British National Corpus. The value is the mean log probability of the words in the prompt given their frequency of occurrence in trigrams that include the previous and following word. The SPROB value was z-score normalized before presentation to the model.

The regression model was implemented as a simple neural network with two hidden dense layers of 32 and 16 nodes. Input was a single vector of concatenated features taken from the sets above. Output was a single sigmodal node with an output between 0 and 1 representing the proportion of words correctly identified in the sentence. The model was trained using a binary cross-entropy loss function. A validation set based on 10% of the training data was used to terminate training.

5.2. Model Evaluation

To determine the relative importance of the feature sets, a greedy algorithm was used to find the first most useful, then the best two, the best three and so on. Table 3 shows how RMS prediction error reduces on the training data (with 5-fold cross-validation) and on the test data as each feature is introduced in turn.

Table 3. *RMS error for non-linear regression model*

Feature set	RMS Prediction Error (%)	
	Train	Test
STOIFILT alone	25.974	26.144
+ SYSTEM	24.068	23.818
+ LISTENER	22.525	22.705
+ SPROB	22.090	22.299
+ TALKER	22.039	22.421

On the training data, STOIFILT in which correlations are provided per filter channel provides a 1.1% improvement over the standard STOI metric alone. The SYSTEM prediction features, improved performance by a further 1.9%, while the LISTENER feature improved by a further 1.5%. Sentence text probability and TALKER prediction together only made a small further improvement of 0.5% RMS prediction error. A similar pattern was found on the test data, except that the introduction of the TALKER features slightly increased error on the test set. A graph of predictions of the full model compared to the actual % correct scores on the test data is shown in Figure 4.

6. Discussion

The analysis of the factors influencing % correct using mutual intelligibility revealed some unexpected results. The utility of the listener audiograms was quite small, possibly because the processing systems had already compensated for audibility by each listener. Listener performance on the digit triple test was also not helpful, possibly because those stimuli had not been corrected for audibility, and so did not provide information about supra-cochlear influences [7] on intelligibility.

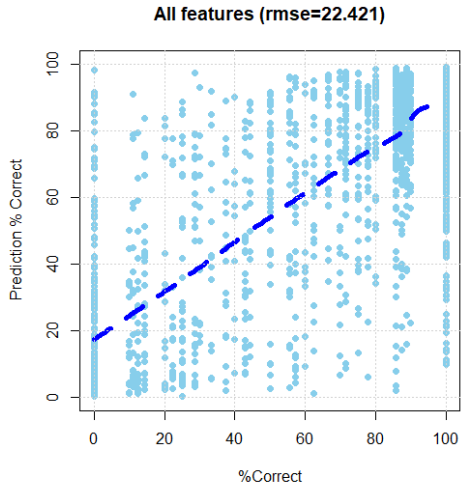

Figure 4. *Predictions of the non-linear regression model on the test data.*

Classification of the choice of processing system and classification of the talker from the supplied audio signals worked extremely well. It is possible that other models trained on the audio alone would pick up these particular influences even without explicit construction of a classifier.

The non-linear regression model showed useful improvements in prediction error on the training data. The switch from a single STOI value to one value per filter channel gave a reduction in error. This suggests there is information in the different frequency bands not yet being exploited in the metric. This idea has been previously explored in [14, 15]. Similarly the fact that the identity of the processing system is useful in addition to STOI suggests that there is information about the nature of processing not being yet captured by STOI. That the identity of the listener is more useful than their pure-tone average suggests there is still more to do to characterise listener performance.

The regression model here - trained for a closed set task in which the listener, talker and processing system came from a fixed set of choices - could easily be extended to an open set task. The system characterisation features could simply represent the type of processing system used, the listener identity could be replaced by audiogram data, and the talker characterisation features could represent the type of talker.

In this work we have used RMS error calculated on % correct as this is the requirement for the challenge. However we would note that in the prediction of a probability, a binomial distribution of scores would be expected, with error better measured in terms of log odds. A proposed method for this is described in [16].

7. Acknowledgements

The authors would like to thank the organisers of the Clarity Prediction Challenge for running the challenge and making the data available. The work described here was supported in part by the UK Engineering and Physical Sciences Research Council [grants: EP/S03580X/1 and EP/S035842/1].

8. References

[1] S. Graetzer, J. Barker, T. J. Cox, M. Akeroyd, J. F. Culling, G. Naylor, E. Porter, and R. Viveros Muñoz. "Clarity-2021 challenges: Machine learning challenges for advancing hearing aid processing". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*, Brno, Czech Republic, 2021.

[2] J. Barker, M. Akeroyd, T. Cox, J. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, G. Naylor, Z. Podwinska, E. Porter and R. Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction", *In Proceedings of the 23nd Annual Conference of the International Speech Communication Association (INTERSPEECH 2022). Incheon, Korea.*

[3] L. Magnusson, "Predicting the Speech Recognition Performance of Elderly Individuals with Sensorineural Hearing Impairment A Procedure Based on the Speech Intelligibility Index", *Scandinavian Audiology, 25:4, 215-222.*

[4] T. Ching, H. Dillon, D. Byrne, "Speech recognition of hearing-impaired listeners: predictions from audibility and the limited role of high-frequency amplification", *J Acoust Soc Am 103, 1128-40.*

[5] J. Kates, K. Arehart, "The hearing aid speech perception index (HASPI)", *Speech Communication 65, 75-93.*

[6] T. Falk, V. Parsa, J. Santos, K. Arehart, O. Hazrati, R. Huber, J. Kates, S.Scollie, "Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools". *IEEE Signal Processing Magazine, 2015;32(2):114-124.*

[7] M. Huckvale G. Hilkhuysen, "On the Predictability of the Intelligibility of Speech to Hearing Impaired Listeners", *1st International Workshop on Challenges in Hearing Assistive Technology (CHAT-2017)*, Stockholm 2017.

[8] C. Pardy, "MPMI package for calculating Mutual information" https://cran.r-project.org/web/packages/mpmi/mpmi.pdf

[9] C. Taal, R. Hendriks, R. Heusdens, J. Jensen. "An Algorithm for Intelligibility Prediction of Time-Frequency Weighted Noisy Speech". *IEEE Trans. Audio Speech Lang. Process*., vol. 19 (2011) 2125-2136.

[10] M. Vlaming, R. MacKinnon, M. Jansen, D. Moore. "Automated screening for high-frequency hearing loss". *Ear and Hearing*. 35(6) (2014) 667-79.

[11] M. Brookes, "VOICEBOX library". https://github.com/ImperialCollegeLondon/sap-voicebox

[12] C. Taal, "STOI – Short-Time Objective Intelligibility Measure". MATLAB implementation: https://ceestaal.nl/code/

[13] *British National Corpus*. http://www.natcorp.ox.ac.uk/

[14] L. Lightburn, M. Brookes, "A Weighted STOI Intelligibility Metric Based On Mutual Information", *Proc. ICASSP 2016*.

[15] A. Andersen, J, Mark de Haan, Z. Tan, J. Jensen, "On the use of Band Importance Weighting in the Short-Time Objective Intelligibility Measure", *Proc. Interspeech 2017*.

[16] G. Hilkhuysen, N. Gaubitch, M. Brookes, M. Huckvale, "Effects of noise suppression on intelligibility II: An attempt to validate physical metrics", *J.Acoust.Soc.Am.*, 135 (2014) 439-50.