Combining Acoustic, Phonetic, Linguistic and Audiometric data in an Intrusive Intelligibility Metric for Hearing-Impaired Listeners

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 a system developed at UCL for the second Clarity Prediction Challenge. The system combines information about: signal acoustics from the STOI metric, phonetics from phone lattice comparison, linguistics from a language model, and audiometric data from pure-tone thresholds. A non-linear regression model based on these features showed an RMS prediction error of 22.4% on the training set compared to a baseline of 26.4% using the STOI metric alone. On the challenge evaluation set, the model had an error of 25.36%.

1. Introduction

The second Clarity Prediction Challenge [1, 2] was an open competition to compare the performance of speech intelligibility metrics on a common dataset. The materials for the prediction challenge were generated from previous enhancement challenges 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 held-out enhanced sentences by these listeners.

Our intelligibility prediction model builds on the success of our system entered for the first prediction challenge [3]. In this submission we continue to use the STOI metric broken out across frequency channels, a scene analyser for characterising the enhancement system, a language model for estimating the probability for sentences, and a speaker recogniser for characterising the talker. We continue to use a small neural network to train a non-linear regression model to predict intelligibility from combinations of the available features.

Innovations in this submission include: calculating the STOI best ear over time to allow for listener head rotation; the use of a phonetic recogniser to compare phone hypotheses between reference and target audio; and the direct use of pure-tone thresholds to characterise listeners.

Section 2 describes the methods used to extract the new features, while readers are referred to our previous paper [3] for descriptions of the features carried over to this study. Section 3 provides metric performance results for baseline measures and the features.

2. Methods

2.1. Two-Ear STOI

In the earlier study, we computed the STOI metric [4, 5] from the reference and target audio to find the "best ear" and represented the STOI outcome as a vector of 15 correlations, one per frequency channel. Since the second enhancement challenge allowed for listener head rotation, instead we computed the table of STOI correlations for each ear separately and then computed the "best ear over time" from the maximum in each time-frequency cell across the ears before the mean is taken over time within each channel.

2.2. Phone lattice comparison

To compare the phonetic properties of the reference and target sentence, we introduce a phonetic recogniser trained on British English to deliver a phone lattice for each signal and compute a correlation. The phone recogniser is based on a publicly available pre-trained DNN model WAV2VEC2-XLSR [6] which takes an input audio waveform and delivers feature vectors every 20ms. These feature vectors have been optimised for multi-lingual speech recognition. This model is then fine-tuned on the WSJCAM0 corpus of British English [7] to deliver softmax outputs over a 45-member phone set.

For use in the model, the frame logit scores for the phones are summed into 15 values representing Voice, Place and Manner (VPM) features (voice: 2 features, place: 6 features, manner: 6 features, silence: 1 feature). The correlations between the time series for each VPM feature across the reference and target sentences are then computed for use in the model.

2.3. Pure-tone thresholds

In the first prediction challenge, there was a closed set of listeners, so we were able to use the listener identity in the model. Since an aim in this challenge is to test generalisability, here we used instead the average pure-tone thresholds across left and right ears to characterise the listener.

3. Results

3.1. 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.

Average performance of these baseline models across the three training subsets is shown in Table 1. The RMS prediction error of 26.4% using STOI on the best ear provides a good estimate of the prediction error found using a current state of the art approach.

Table 1. *RMS error for baseline predictors*

Baseline method	RMS Prediction Error (%)
	Training Sets Average
NULL	38.307
LISTENER only	37.311
SYSTEM only	29.825
STOI best ear	26.369

3.2. Input Features

The following features were used to construct a regression model to predict percentage correct intelligibility:

STOI2EAR (15 features) – STOI correlations between source and processed audio per filter channel. The target and processed signals are first aligned by spectral cross-correlation [8] before calculation of the STOI correlations separately for each ear. The set of correlations is chosen from the best time-frequency cell correlations across the two ears.

LATTICE (15 features) – phone lattice correlations. The time aligned signals are processed into phone lattices by the recogniser, and the section containing speech is identified from the reference audio. The logit scores from that section are then summed into 15 VPM features and correlations between reference and target are computed for each feature.

SYSTEM (20 features) – predicted identity of the processing system found by the scene classifier in each training data subset, one probability per system. Note that only 17 systems are present in each training subset.

SPROB (11 features) – prompt sentence text probability and length. The probability 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 sentence length is represented in a unary code of 10 features.

PTA (8 features) – pure tone thresholds at 8 frequencies for the listener averaged over left and right ears. This is generated from the given metadata.

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

The regression model was implemented as a simple neural network with two hidden dense layers of 64 and 32 nodes. Input was a single vector of concatenated features taken from the sets above. Output was a single sigmoidal 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. Separate models were trained for each training subset and the held-out portion was used to terminate training.

3.3. 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) as each feature is introduced in turn.

Table 2. RMS error for non-linear regression model

Feature set	RMS Prediction Error (%)
	Training Sets Average
STOI2EAR alone	25.972
+ LATTICE	25.344
+ SYSTEM	23.758
+ SPROB	23.257
+ PTA	22.490
+ TALKER	22.399

On the training data, STOI2EAR provides a 0.4% improvement in RMSE over the standard STOI metric alone. The phonetic lattice features improves performance by 0.6% and the system prediction features improves performance by a further 1.6%. Adding the linguistic sentence probability features improves performance by 0.5%, while adding the audiometric PTA features gave a further improvement of 0.7%. The talker identity features only made a small improvement of about 0.1%.

In terms of computational load, the calculation of the STOI metric, system characterisation, sentence probability, pure-tone average and talker identity features takes less than 1s per file on a modern CPU. The largest computational load is the generation of the phone lattices, which requires a GPU. Using an NVIDIA Quadro P5000 GPU, lattices took about 1s each to generate. Overall the system is working at about 1x real-time.

4. Discussion

The model is operating about 0.4% RMSE worse than the model used in CPC1 [3], this is probably because of the increased variability in the CEC2 task. To reiterate a point made in [3], while this model has features which explicitly attempt to identify the processing system and the talker identity, the same information is undoubtedly available implicitly in CPC2 systems which map the audio to intelligibility score directly.

Further work could investigate how phonetic recognition is affected by individual listeners' impairments, as this might lead to a metric which is more sensitive to the particular listening problems of individuals.

5. 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].

6. 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] Clarity Prediction Challenge 2: https://claritychallenge.org/docs/cpc2/cpc2_intro

[3] Huckvale, M., Hilkhuysen, G., "ELO-SPHERES intelligibility prediction model for the Clarity Prediction Challenge 2022". *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH 1022*, Incheon, Korea, 2022.

[4] 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.

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

[6] HuggingFace WAV2VEC2-XLSR model: https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[7] WSJCam0 database of British English: https://catalog.ldc.upenn.edu/LDC95S24

[8] M. Brookes, "v_sigalign, from the VOICEBOX library". https://github.com/ImperialCollegeLondon/sap-voicebox