Evaluating a 3-factor listener model for prediction of speech intelligibility to 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

A speech intelligibility prediction model for hearing impaired listeners would be useful in the development of better signal enhancement methods and for the fitting of hearing aids. Most current prediction models use only information from a pure-tone audiogram to characterise impaired listeners, although evidence suggests that listeners vary in ways not captured by pure-tone thresholds. In this paper we evaluate a model in which each listener is described by three factors: average pure-tone thresholds, sensitivity to phonetic distortion and sensitivity to word likelihood. We build and evaluate the model using the corpus collected by the second Clarity Prediction Challenge, which contains over 13,000 intelligibility judgments by 31 hearing impaired listeners. We describe how the factors were estimated and test their independence. We show that incorporating the listener-dependent factors into an existing intelligibility metric can improve the accuracy of prediction on held-out test data with a 9.8% relative improvement in prediction error.

Index Terms: speech intelligibility, hearing impairment, speech enhancement, hearing aids

1. Introduction

Speech intelligibility prediction metrics provide an automated assessment of the likely intelligibility of a speech audio signal. Such metrics would be practically useful if they could be extended to make predictions for hearing-impaired listeners. These measures could then be used to choose the best signal enhancement strategy for an impaired listener in a given setting, or find the settings of a hearing aid which might maximise speech intelligibility.

Existing speech intelligibility prediction models for hearing impaired listeners, such as HASPI [1], focus on the auditory level processing of the signal related to pure-tone sensitivity and selectivity, and amplitude compression. In HASPI an auditory front-end delivers a frequency-time representation of the signal amplitude of a target sentence through the impaired ear which can be compared to the same representation of a clean reference passed through a non-impaired ear. Because hearing-impaired listeners are typically only assessed in terms of pure-tone thresholds, the whole model is predicated on the assumption that pure-tone thresholds alone are sufficient to characterise an impaired listener.

Analysis of databases of hearing-impaired listeners allow the measurement of the degree to which pure-tone thresholds can account for variation in speech intelligibility performance. For example, [2] show that pure-tone thresholds alone only account for 40% of the variance in performance of listeners on speech-in-noise intelligibility performance. This is not unexpected, as there is evidence that listeners with similar audiograms can vary in terms of other cochlear dysfunctions [e.g. 3]. Listeners also vary in many other ways unrelated to hearing, such as language experience and working memory capacity which are known to modulate the performance of speech perception [e.g. 4]. However measures of cochlear function, language experience and memory are not usually available for a given listener, and so far have not yet been be incorporated into speech intelligibility metrics for the hearing-impaired.

In this study we look into incorporating pragmatic listener factors into an intelligibility prediction metric - factors which can be estimated from empirical performance of a listener on existing intelligibility tests. In particular we develop and evaluate a 3-factor model that follows a highly simplified model of speech perception shown in Fig.1.

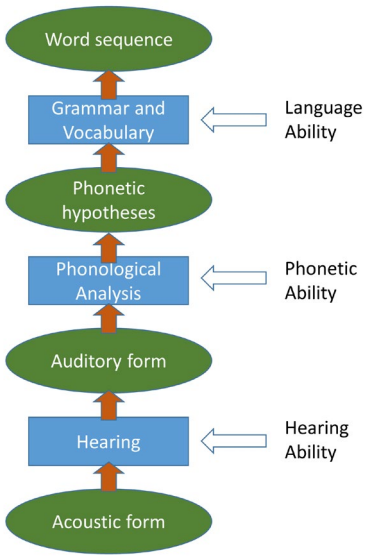

Figure 1. Three factor model of sources of variation in speech perception

In this 3-factor model we describe each listener in terms of three abilities: for hearing, for phonetic perception, and for language. To keep the model simple, we reduce each of these factors to a single parameter that describes the relative ability of the listener w.r.t. a population of other listeners. The factors are estimated from results of a speech intelligibility test for the listener. These three numeric factors can then be incorporated into a logistic regression model to predict the proportion of words recognised by a given listener from a given spoken sentence given some acoustic properties of the sentence and its orthographic transcription.

In this paper we first describe the data set used to estimate the factors of the model for a group of listeners. Next we describe how the factors were estimated and show that these factors are to some extent independent of one another. We then evaluate an intrusive speech intelligibility model based on the STOI metric [5] using the 3 factors to optimise the performance function relating STOI values to proportion of words correctly recognised by a given listener. Finally we discuss how well the factors improved the model and what further work is required.

2. Data Set

For this modelling work, we are using the corpus collected for the second Clarity Prediction Challenge (CPC2) [6]. CPC2 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. Results of the challenge are available on the workshop website[1].

The CPC2 corpus consists of reference audio recordings of English sentences spoken by 6 different speakers, which have been subsequently corrupted by added noise and reverberation before being enhanced by a number of competitor systems and then presented binaurally to a group of moderately hearing-impaired listeners. Intelligibility models can be constructed which are either: *intrusive* and based on comparing the enhanced signal with the reference signal, or *non-intrusive* and based on the enhanced signal only.

The CPC2 training partition consists of 12243 intelligibility measurements built from 921 different sentences/auditory scenes, enhanced by 20 different systems and presented to 31 listeners. The test partition consists of 897 intelligibility measurements built from 200 different sentences/scenes, enhanced by a subset of 9 systems and presented to a subset of 15 listeners.

In this study we use 80% of the training partition as the training set for computing the listener factors, and 20% as the development set for building models that predict proportion correct from STOI values and listener factors. Final evaluation is performed on the test partition. Although the same systems and listeners were used in both training and testing, we performed our final evaluation using leave-one-subject-out cross validation to estimate the performance of the model on an unseen listener.

3. Listener Factors

3.1. Hearing Sensitivity

We chose to express the sensitivity of the listener to sound in terms of the Pure-Tone Average (PTA) calculated as the mean of the pure-tone thresholds at 500, 1000, 2000 & 4000Hz thresholds expressed in dBHL.

Fig 2 shows how the mean listener score in terms of proportion of words correctly recognised varies with PTA in the training set.

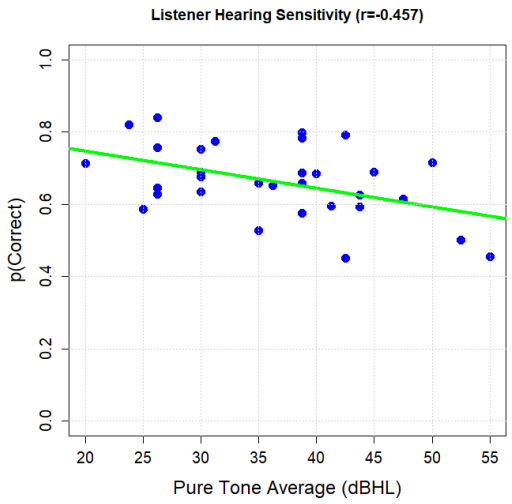

Figure 2. Effect of pure-tone average on word intelligibility

Note that only 20% of the variance in performance in these data are captured by differences in PTA. This is less than the 40% found in [2], possibly because enhancement systems were allowed to take the listeners' thresholds into account during processing. Thus the CPC2 data could be considered "aided listening", and any remaining variation in performance with PTA must be due to weaknesses in the equalisation, or because of other variations in hearing ability correlated with PTA such as loudness recruitment.

3.2. Phonetic Distortion Sensitivity

To establish a listener's sensitivity to distortions in the phonetic properties of the signal, we first introduce a phonetic distortion score computed by comparing the enhanced corrupted audio signal with the reference audio within a phonetic feature space. We then investigate how listener performance varies as a function of the degree of phonetic distortion.

To compare the phonetic properties of the reference and target sentence, we built a phonetic recogniser trained on British English to deliver a lattice of phone posterior probabilities for each signal. The phone recogniser is based on a publicly available pre-trained DNN model WAV2VEC2-XLSR [7] 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 [8] to deliver softmax outputs over a 45-member phone set.

For use in the distortion model, the posterior scores for the 45 phones in each frame are summed into 15 values representing Voice, Place and Manner (VPM) features (voice: 2 features, place: 6 features, manner: 6 features, silence: 1 feature). This mapping to VPM reduces the dimensionality of the feature space and decreases the proportion of zero-valued cells. The distortion score is then calculated from the mean correlations of the VPM feature time series between the reference and target sentences. The correlations are performed over non-silent regions in the reference signal only. Thus the

[1] https://claritychallenge.org/clarity2023-workshop/

distortion score is a correlation between -1 and 1, with 1 corresponding to an unchanged phonetic representation.

Fig.3 shows the results of a logistic regression between phonetic distortion scores of the target sentences and word recognition performance on them for each of the listeners in the training set.

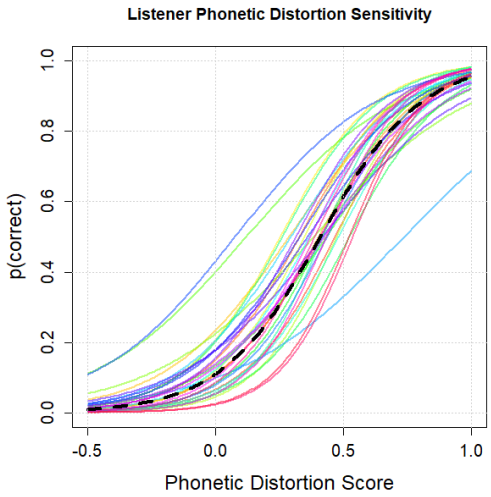

Figure 3. Effect of phonetic distortion score on word intelligibility. Coloured lines are individual listeners, dashed line is the group mean.

The analysis shows that listeners vary considerably in their sensitivity to the phonetic distortion score, with two listeners achieving 50% correct for distortion scores of 0.1, while one listener requiring a score of 0.7 to achieve the same performance.

3.3. Word Probability Sensitivity

To establish the sensitivity of each listener to word frequency, we first estimated the prior probability of each word found in each target sentence from a language model. The model is built using word trigram frequencies found in the British National Corpus [9]. The log probability of each word in the sentences is calculated from the frequency of its occurrence in trigrams that include the previous and following word in the BNC.

To analyze the sensitivity of each listener to word probability, we perform a logistic regression that relates the word probability as found in the language model to whether the word was recognised correctly by that listener in their response to the target sentence. This is computed over all responses in the training set. Fig.4 shows the variation in sensitivity to prior word probability across the listeners.

The analysis shows that intelligibility is better for more frequent words, as expected, but also shows considerable variation in performance across listeners.

3.4. Independence of Factors

From the preceding analysis, we end up with three numeric factors that describe each listener: (i) their pure-tone average, (ii) the phonetic distortion score which shows a correct probability of 0.5, and (iii) the log word probability which shows a correct probability of 0.5.

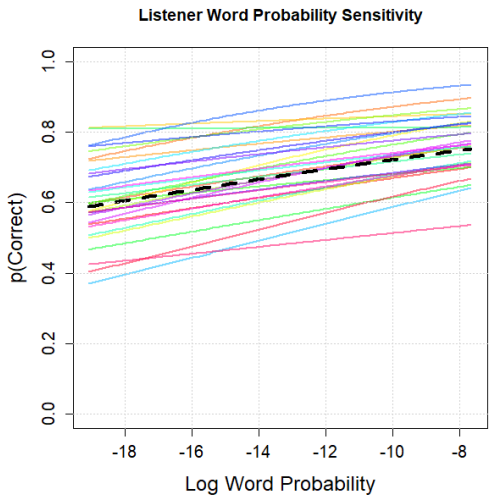

Figure 4. Effect of word probability of correct score. Coloured lines are individual listeners, dashed line is the group mean.

To investigate whether these three factors are providing independent evidence for listener performance, we perform a principal components analysis (PCA) and plot the listeners on a two–dimensional plot, see Fig.5. The arrows show the directions of the loadings on the input factors. The first two dimensions capture 84% of variance in the data.

The PCA shows that the phonetic and language factors are to a degree independent of the hearing factor. There is some evidence that the phonetic factor and the hearing factor are more closely related than the language factor and the hearing factor, which is not unexpected.

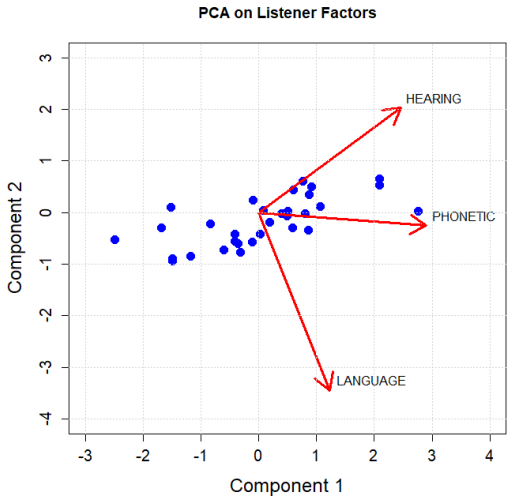

Figure 5. Principal Components Analysis of the three listener factors, showing directions associated with each factor independently.

4. Model Evaluation

4.1. Methods

To evaluate the utility of the 3 listener factors in a speech intelligibility metric, we constructed a baseline model using the standard implementation of the STOI metric [10]. This metric

takes as input the processed corrupted audio and compares it to the reference audio. The comparison is performed in the spectral domain in short time windows across 15 frequency bands. The output is the mean correlation between the amplitude envelopes found in each window over all speech-active regions and all frequency bands. Thus the STOI value provides an acoustic measure of the distortion caused by the signal corruption and subsequent enhancement.

For this study we first time aligned the reference audio and the target audio using spectral cross-correlation [11], then computed the STOI score for the left ear and right ear independently. We then chose the better of the ears as our reference STOI value for the model.

To fit a performance function that relates the better-ear STOI value to the proportion of words correct, we compute a logistic regression using the listeners' scores for each sentence. This regression is trained using the development set, and the performance on both the development set and test set is measured in terms of RMS prediction error expressed in percent (this is the performance measure used in CPC2). Cross-validation is applied by building multiple models with each listener left out in turn.

We first model the performance function using the STOI values alone and no listener dependent factors using a generalised linear model, and then incorporate combinations of the 3 listener-dependent factors to determine the influence of listener specific information.

Table 1. Intelligibility Metric Performance

Listener Factors Included	**Development**		**Test**	
	RMSE	Change	RMSE	Change
STOI alone	26.806		26.978	
+PTA	26.641	-0.165	26.551	-0.427
+PHONETIC	25.013	-1.793	24.386	-2.592
+WORD	25.839	-0.967	26.964	-0.014
+PTA+PHONETIC	25.006	-1.800	**24.344**	**-2.634**
+PHONETIC+WORD	**24.726**	**-2.080**	24.656	-2.322
+PTA+WORD	25.869	-0.937	26.971	-0.007
+PTA+PHONETIC+WORD	24.752	-2.054	24.688	-2.290
HASPI	29.566	+2.760	28.400	+1.422

4.2. Results

A summary of results is shown in Table 1. Using the STOI metric alone as the basis for the performance function, generates an RMSE value of 27.0% and a correlation of 0.737 on the prediction of proportion of words correctly recognised in sentences of the test set. Incorporating the PTA factor into the model, reduces the RMSE by 0.43%, while incorporating the phonetic factor reduces the RMSE by 2.59%. The word factor reduces the RMSE by 0.97% on the development set, but only by 0.01% on the test set. Overall the best performing measure on the development set used the phonetic and word factors, but the best measure on the test set used the PTA and phonetic factors, which showed an absolute RMSE reduction of 2.6% equivalent to a relative reduction of 9.8% over the model that did not include listener factors. This system also showed an improved correlation of r=0.793. For comparison, Table 1 also includes the performance of a logistic regression based on HASPI calculated using the CPC2 baseline implementation. The HASPI metric scores worse than STOI alone on these data despite the fact that it incorporates audiometric thresholds of the listeners.

5. Discussion

In this study we investigated the utility of three listener-specific factors on the performance of a speech intelligibility metric. The hearing sensitivity factor, estimated from the pure-tone audiogram only showed a small improvement, as might have been expected given that the audio stimuli had been equalised for the listeners. This might also have contributed to the weaker performance of the HASPI metric. The phonetic distortion sensitivity factor showed the greatest improvement in RMSE, suggesting it is capturing some of the interesting variation across listeners not yet captured by the PTA. The word probability factor showed a useful improvement on the development set, but had no effect on the test set. This is probably because the development set contained the same prompt sentences as the training set, while the test set had different ones. This difference probably made the estimates of listener sensitivity to words in the test set sentences poorly estimated. The word probability factor might be improved by incorporating the "recognisability" of words from a model trained on a large population of listeners.

There are a number of limitations in the current study. The listeners all have moderate levels of impairment, with the most impaired speaker having a PTA of 55dBHL. The PTA might be a better predictor for intelligibility at greater levels of impairment, where loudness recruitment may play a more significant role [12]. Finally, the range of audio qualities found in CPC2 were all relatively good, with most listeners achieving about 70% words correct overall. A database that included more challenging materials might have given more representative results.

In summary, this study has shown one simple way in which listener-specific factors could be included in a speech intelligibility metric for hearing-impaired listeners. Analyzing responses from a speech intelligibility experiment can deliver information about the listener's sensitivity to phonetic distortions and to word probability. We have shown that incorporating these factors into an intelligibility metric can deliver improved metric performance.

The relative independence of the phonetic and language factors from the pure-tone average makes this approach promising for building better speech intelligibility prediction models for a given hearing-impaired listener. Future work is required to investigate whether there are interactions between these factors and the best signal enhancement methods for the listener.

6. Acknowledgements

The authors would like to thank the organisers of the Clarity Prediction Challenge for running the challenge and making the data available.

7. References

[1] J. Kates, K. Arehart, K.H., “The hearing-aid speech perception index (HASPI) version 2”. *Speech Communication*, 131, pp.35-46, 2021

[2] 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.

[3] M. Fereczkowski T. Dau, E. MacDonald, “Comparison of Behavioral and Physiological Measures of the Status of the Cochlear Nonlinearity”. *Trends in Hearing*, 25, 2021.

[4] J. van Rooij, R. Plomp; “Auditive and cognitive factors in speech perception by elderly listeners. II: Multivariate analyses”. *J. Acoust. Soc. Am.*, 88, pp2611–2624, 1990.

[5] 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*., 19, pp2125-2136, 2011

[6] T. Cox, M. Akeroyd, J. Barker, J. Culling, J. Firth, S. Graetzer, H. Griffiths, L. Harris, M., R. Viveros Munoz, G. Naylor, Z. Podwinska, E. Porter, "Predicting Speech Intelligibility for People with a Hearing Loss: The Clarity Challenges", *InterNoise22*, Glasgow , 2023.

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

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

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

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

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

[12] A. Vermiglio, S. Soli, D. Freed, X. Fang, “The Effect of Stimulus Audibility on the Relationship between Pure-Tone Average and Speech Recognition in Noise Ability”. *J Am Acad Audiol.* 31, pp224-232, 2020.