Word-level intelligibility model for the third Clarity Prediction Challenge

Mark Huckvale

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

m.huckvale@ucl.ac.uk

Abstract

This paper presents a speech intelligibility model for the third Clarity Prediction challenge based on an analysis of word-level intelligibility in the training dataset. Using the given test prompts, a word-level alignment was performed on the reference audio, and this was then used to extract information from the test audio, including word-level measures of acoustic and phonetic distortion. Lexical properties of the words were also obtained using other language resources, including phone count, syllable count, word frequency, trigram frequency and number of lexical neighbours. We present an analysis showing how the intelligibility of individual words relates to these properties and build a classification model that predicts word intelligibility. We show that sentence level intelligibility predictions derived from a word-level intelligibility prediction model gives better performance than a model based on whole sentences.

Index Terms: speech intelligibility model

1. Introduction

The third 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.

The work presented in this paper builds on the success of our systems entered for the first two prediction challenges [3,4]. In this submission we continue to use the STOI metric to create a measure of acoustic similarity between the test sentence and a clean reference, a phonetic recogniser to create measures of phonetic similarity, and a language model for estimating word sequence probability. The main innovation in this work is a focus on the intelligibility of individual words in the test sentences, which allows us to explore how word intelligibility is related to lexical properties of the word, such as phoneme count, syllable count, size of lexical neighbourhood, and position of word in the sentence. Using a model of word intelligibility based on these features we then predict sentence intelligibility to generate predictions for the challenge.

Section 2 describes the data and the methods used to extract the word features. Section 3 investigates the utility of the different features in predicting word intelligibility. Section 4 presents the accuracy of word-level and sentence level intelligibility predictions on the training and development data sets.

2. Data and Methods

2.1. Challenge data set

The challenge training data comprises 15520 different sentence intelligibility measurements collected from 26 different hearing-impaired listeners (9 Mild, 13 Moderate, 4 Moderately severe). 1047 different sentences were used (338 of length 7 words, 293 of length 8, 224 of length 9 and 192 of length 10). In these sentences there were 1781 different words. In total there were 128603 word intelligibility measurements, with 63.16% correctly identified.

2.2. Word segmentation

To extract word-level features from the supplied signals, we first compute a phonetic posteriorgram from the test and reference audio. This uses the phonetic recogniser described in [4] which is now openly available [5]. Using dictionary pronunciations of the words in the sentences, we then perform a dynamic-programming alignment between sentence transcription and posteriorgram to locate the start and end of each word in the reference signal. These word segmentations are then used to derive acoustic and phonetic distortion measures for each word in each sentence.

2.3. Signal features

The following features were extracted from the word-segmented signals and posteriorgrams:

STOI: The STOI metric [6] correlates the test and reference audio in 15 frequency bands and measures the degree of acoustic distortion present in the test signal compared to the reference. The target and processed signals are first aligned by spectral cross-correlation [7] before calculation of the STOI correlations separately for each ear and each word. The STOI value from the better ear is used in prediction.

Phonetic RMSE: This is a measure of the phonetic distortion present in the test signal compared to the reference. The phone posteriorgram is first reduced to 15 dimensions representing Voice, Place and Manner features (see [4]), and the RMS difference between the VPM features in the test compared to the reference is computed for each word.

Phonetic correlation: This is an alternative measure of phonetic distortion, computed in the same manner as for Phonetic RMSE, but using the correlation between the VPM features rather than the RMS difference.

2.4. Word features

The following features are calculated from dictionary and corpus properties of each word, independently from the audio.

#Words in sentence: the number of words in the prompt sentence containing this word.

Word position in sentence: the relative position of the word in the prompt sentence, expressed as a number between 0 and 1.

Phoneme count: the number of phonemes in the word's dictionary pronunciation

Syllable count: the number of syllables in the word's dictionary pronunciation

Neighbourhood size: the size of the lexical neighbourhood of the word [8]. This is computed by searching a pronunciation dictionary for all words which are one phoneme edit distance away from the word

Word frequency: the frequency of the word in the BNC corpus.

Trigram frequency: the frequency of the trigram made up from this word, the previous word and the following word in the BNC corpus.

3. Word intelligibility analysis

The relationships between each feature and the probability of the word being recognised correctly is shown in Table 1. The bootstrapped mutual information metric was calculated using the MPMI toolbox [9].

Table 1. Relationship between word-level features and word intelligibility in the training data

Feature	Correlation	Mutual Information
Neighbourhood size	-0.049	0.165
Audio STOI	0.525	0.163
Phonetic RMSE	-0.435	0.111
Phonetic correlation	0.369	0.107
Word frequency	0.161	0.060
Phoneme count	0.041	0.027
Trigram frequency	0.144	0.021
Syllable count	0.041	0.015
Word position in sentence	-0.057	0.001
# Words in sentence	-0.053	<0.001

The analysis shows that intelligibility increases with higher word frequency, higher trigram frequency, higher phoneme count, and higher syllable count, and reduces with increasing neighbourhood size, later word position, and number of words in sentence. For most words, neighbourhood size has little effect in these data, except for a small number of short words with large neighbourhoods, such as "pose", "says", "low", "raid" and "sigh", which are particularly poorly recognised.

4. Intelligibility models

4.1. Word intelligibility prediction

We use the features in Table 1 to build a model to make a binary prediction of whether the word would be recognised correctly. We use a Random Forest classifier, with 200 trees and a minimum leaf count of 5. To encourage generalisation, we first oversample the training data by synthesizing a further 128000 samples by linear interpolation using random mixing factors. Cross-validated accuracy and ROC area-under-curve on the training data are shown in Table 2.

Since there is considerable mutual information between the intelligibility of adjacent words in each sentence, we also trained the random forest classifier on a concatenation of three word-vectors representing the word and its immediate neighbours in the sentence. This slightly improves classification accuracy.

Table 2. Word intelligibility classifier accuracy

Feature vector	Accuracy	Area-under-curve
Single word	79.2%	0.851
Word in context	81.0%	0.869

4.2. Sentence intelligibility prediction

To compute sentence intelligibility, we use the random forest classifier to deliver a probability for each word to be correctly recognised and take the mean logit-transformed value. We then combine this with the hearing impairment severity for the listener in a logistic regression using a linear model. The performance of the sentence intelligibility prediction on the development data and cross-validated on the training data is given in Table 3. For reference, we include figures for a logistic regression model based on the acoustic and phonetic features calculated over the whole sentence, which is similar to the system in [4]. Results show that the model based on the word-level features has better performance, and that adding hearing severity information slightly improves results.

Table 3. Sentence intelligibility prediction performance on the challenge data sets

Model	Training set		Development set	
	Corr	RMSE	Corr	RMSE
Sentence only	0.749	26.404	0.764	26.491
Sentence and Severity	0.755	26.110	0.781	25.669
Word only	0.776	25.241	0.784	25.446
Word and Severity	0.782	24.798	0.799	24.638

5. Conclusions

In this paper, we have investigated factors affecting the intelligibility of individual words in the challenge data set. We have shown that we can build a successful model that predicts word intelligibility by combining acoustic and phonetic distortion measures computed over word regions in the signals with lexical features of the words themselves, like their frequency and the size of their lexical neighbourhood. We have shown that basing a sentence intelligibility prediction model from the word intelligibility predictions gives an improved accuracy of prediction over treating the sentence as a whole. Better modelling of the mutual intelligibility between words within a sentence is an opportunity for further work.

An interesting outcome of the word-level intelligibility analysis is the particular problems for intelligibility arising from short, relatively-infrequent words with large lexical neighbourhoods, probably because they can be readily confused with words with greater frequency.

Scripts to recreate the results presented in this paper will be made available on-line [10].

6. Acknowledgments

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

7. 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 3: https://claritychallenge.org/docs/cpc3/cpc3_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 2022*, Incheon, Korea, 2022.

[4] Huckvale, M., Hilkhuysen, G., "Combining Acoustic, Phonetic, Linguistic and Audiometric data in an Intrusive Intelligibility Metric for Hearing-Impaired Listeners", *4th Clarity Workshop on Machine Learning Challenges for Hearing Aids*, Trinity College Dublin, 2023.

[5] M.Huckvale, "GBPhone: WAV2VEC2-XLSR model adapted for British English". https://huggingface.co/mhuckvale/GBPhone

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

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

[8] P. Luce, D. Pisoni, "Recognizing spoken words: the neighborhood activation model". *Ear and Hearing*, 19 (1998) pp1–36.

[9] C. Pardy, "mpmi: Mixed-Pair Mutual Information Estimators", https://doi.org/10.32614/CRAN.package.mpmi

[10] M.Huckvale, https://github.com/mhuckvale