Pre-Trained Intermediate ASR Features and Human Memory Simulation for Non-Intrusive Speech Intelligibility Prediction in the Clarity Prediction Challenge 2

Rhiannon Mogridge, George Close, Robert Sutherland, Stefan Goetze, Anton Ragni

Speech and Hearing Group, Dept. of Computer Science, Univeristy of Sheffield, UK

{rmogridge1, glclose1, rwhsutherland1, s.goetze, a.ragni}@sheffield.ac.uk

Abstract

This report describes an entry to the Clarity Prediction Challenge 2. Non-intrusive speech intelligibility neural networks are trained using the challenge data which make use of novel input feature representations sourced from the intermediate layers of a pre-trained automatic speech recognition (ASR) system. These are combined with an exemplar-based model of human memory to predict human intelligibility ratings. Performance improvement over the baseline system on disjoint validation sets is found, and a challenge entry using the proposed system is described.

Index Terms: speech recognition, human-computer interaction, computational paralinguistics

1. Introduction

Hearing loss is a widespread problem, affecting approximately 12 million people (1 in 5) in the United Kingdom, though this problem is only set to grow; by 2035 it is expected to impact 14.2 million people in the UK [1]. Age correlates with a person's chances of being affected by hearing impairment, and the population is expected to age. From the year 2015 to 2050, the proportion of the population aged over 60 is expected to nearly double, rising from 12% to 22% [2]. As hearing impairment typically worsens in an individual, the intelligibility of the speech that they hear decreases. Therefore, being able to predict the intelligibility of a speech source is vital in advancing assistive hearing technology.

The Clarity Prediction Challenge 2 (CPC2) data consists of speech signals $\hat{s}[n]$ and corresponding correctness values i, obtained from listening tests with hearing-impaired listeners. The signal generation process is shown in Figure 1. The signals $\hat{s}[n]$ are the enhanced outputs of hearing aid systems with binaural input $x[n]$, being an artificially corrupted versions of a clean reference audio $s[n]$ with additive noise $v[n]$. The correctness values i are the percentage of words which the listener was able to correctly reproduce from the speech signal $\hat{s}[n]$ they listened to. The challenge data also contains additional information such as left/right ear's representations of the listeners' hearing loss as audiograms $\mathbf{a}_l$ and $\mathbf{a}_r$, as well as a hearing loss simulation system $\mathcal{S}$ which can be used to further process $\hat{s}[n]$ based on the audiogram information $\{\mathbf{a}_l, \mathbf{a}_r\}$ to produce $\hat{s}'[n]$, an approximation of how $\hat{s}[n]$ would sound to a specific hearing impaired individual. All audio signals are stereo with a left and right channel.

The data is partitioned into three train sets each paired with a disjoint evaluation set. Each evaluation set contains listeners and hearing aid enhancement systems which are unseen in its corresponding training set, meaning that prediction models will need to generalise to unseen listeners and systems. Set 1 has training set of size 8599 and a test set of size 305. Set 2 has training set of size 8135 and a test set of size 294. Set 3 has training set of size 7896 and a test set of size 298.

Following findings in [3] where it was found that the use of $\hat{s}'[n]$ signals was not useful for the intelligibility prediction task, in this work we use $\hat{s}[n]$ only as the input to the proposed systems.

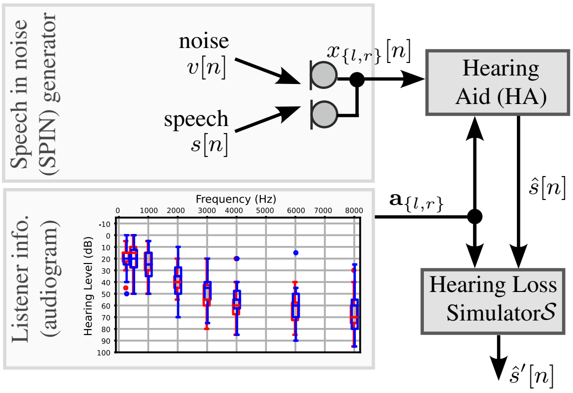

Figure 1: *Signal generation for Clarity Prediction Challenge.*

2. System architecture

2.1. Features

The input feature of our model is the outputs of the 12 decoder layers from a pre-trained Whisper [4] ASR model[1], given $\hat{\mathbf{S}}$, a spectrogram representation of $\hat{s}[n]$ as input. The audio signal $\hat{s}[n]$ is downsampled to 16kHz and padded to 30 seconds in length such that it can be input to Whisper. This spectrogram is an 80 channel log magnitude Mel Spectrogram with a window of 25ms and a stride of 10ms.

For each spectrogram $\hat{\mathbf{S}}$, the input to the proposed intelligibility prediction model is thus a representation with dimensions $W \times 768 \times 12$ where W is variable depending on the predicted number of words in the utterance represented by the input spectrogram, 768 is the feature dimension of the output of each decoder layer and 12 is the number of decoder layers in the pre-trained Whisper model. The parameters of the Whisper model are frozen and are not updated during the training of the metric prediction model described below.

[1] https://huggingface.co/openai/whisper-small

2.2. Model Structure

A system combination of two models is used, incorporating a base model and an exemplar-informed model, both shown in Figure 2. A model structure following work on the Clarity Prediction Challenge 1 (CPC1) in [3] is chosen for the base speech intelligibility (SI) prediction network, to the lower left in Figure 2.

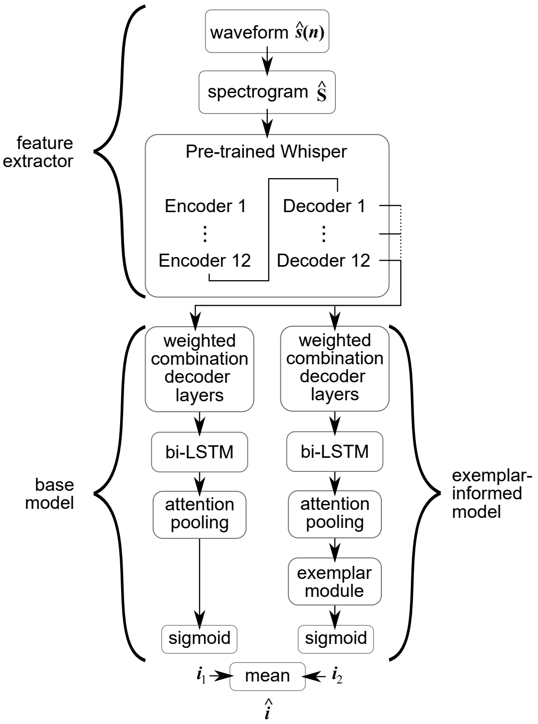

Figure 2: *Model architecture.*

Its input is the output of each of the 12 decoder layers from the pre-trained Whisper ASR system, as described above. A weighted sum of these representations is implemented as a learnable linear layer with 12 parameters, all initialised to 1, followed by a softmax. This representation is then processed by 2 bidirectional long short-term memory (BLSTM) layers with an input size of 768 and a hidden size of 384. Finally, an attention pooling feed-forward layer with a sigmoid activation outputs to a single neuron which represents the base predicted correctness value i_1 normalized between 0 and 1.

An exemplar-informed variation of the model described above is also trained. It differs from the base model in that the attention pooling output is fed into an exemplar-informed module based on a simplified theory of human memory [5]. The exemplar-informed module incorporates a set of labelled exemplars drawn from the training data. Let $\boldsymbol{y}$ be the output of the attention pooling for input waveform $\hat{s}[n]$. The exemplars, $\hat{s}_1^*[n], ..., \hat{s}_D^*[n]$, are processed in the same way as the input, producing exemplar outputs from the attention pooling $\boldsymbol{y}_1^*, ..., \boldsymbol{y}_D^*$. Let $i_1^*, ..., i_D^*$ be the exemplar correctness labels, scaled to lie between 0 and 1. The output, r, of the exemplar module is given by

$$a = \sum_{d=1}^{D} \frac{\boldsymbol{f}(\boldsymbol{y}) \cdot \boldsymbol{g}(\boldsymbol{y}_d^*)}{||\boldsymbol{f}(\boldsymbol{y})||\,||\boldsymbol{g}(\boldsymbol{y}_d^*)||} i_d^* \tag{1}$$

$$r = h(a) \tag{2}$$

The functions $\boldsymbol{f} : \mathbb{R}^{768} \to \mathbb{R}^{768}$, $\boldsymbol{g} : \mathbb{R}^{768} \to \mathbb{R}^{768}$ and $h : \mathbb{R} \to \mathbb{R}$ are all learned affine transformations, and r passes through a sigmoid activation to ensure the output falls between 0 and 1.

The output of the system combination $\hat{i}$ for a given input signal $\hat{s}[n]$ is the mean of the outputs of the base i_1 and exemplar-informed systems i_2.

2.3. Training

For each of the three splits, two listeners and two systems were randomly selected to form a disjoint validation set. All data with these listeners and systems were removed from the training set. A randomly selected non-disjoint validation set consisting of 10% of the remaining training data was also formed. The majority of model selection and hyperparameter tuning was performed using these validation sets, to test how well they generalised to unseen listeners and systems. For the final models, the disjoint validation set and all listeners/systems associated with it were merged back into the training data to make the best use of resources.

The base and exemplar-informed models are trained separately with mean squared error loss. The base system is trained for 25 epochs with batch size 8, learning rate 10^{-5} and weight decay 10^{-4}. The exemplar-informed system is trained for 50 epochs with learning rate 2×10^{-6} and weight decay 10^{-4}. During training and validation, $D = 8$ exemplars are chosen randomly from the training data for each minibatch.

3. Results and discussion

Table 1 shows the results for the final models on each of the data splits.

Table 1: *Validation and evaluation set results for the final models.*

Split	**RMSE**	
	validation	evaluation
1	21.6	
2	23.4	
3	22.7	

[insert section following analysis of results]

4. Conclusions

[insert section following analysis of results]

5. References

[1] RNID, "Prevalence of deafness and hearing loss," https://rnid.org.uk/get-involved/research-and-policy/facts-and-figures/prevalence-of-deafness-and-hearing-loss/, accesssed: 2023-07-26.

[2] World Health Organisation, "Ageing and health," https://www.who.int/news-room/fact-sheets/detail/ageing-and-health, accesssed: 2023-07-26.

[3] G. Close, T. Hain, and S. Goetze, "Non intrusive intelligibility predictor for hearing impaired individuals using self supervised speech representations," https://arxiv.org/abs/2307.13423, 2023.

[4] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[5] D. Hintzman, “Minerva 2: a simulation model of human memory,” *Behaviour Research Methods, Instruments & Computers*, vol. 16, pp. 96–101, 03 1984.