The University Of Sheffield.

SLT Centre for Doctoral Training

UK Research and Innovation

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

20 August 2023 Dublin

Rhiannon Mogridge, George Close, Robert Sutherland, Stefan Goetze and Anton Ragni
(University of Sheffield, UK)

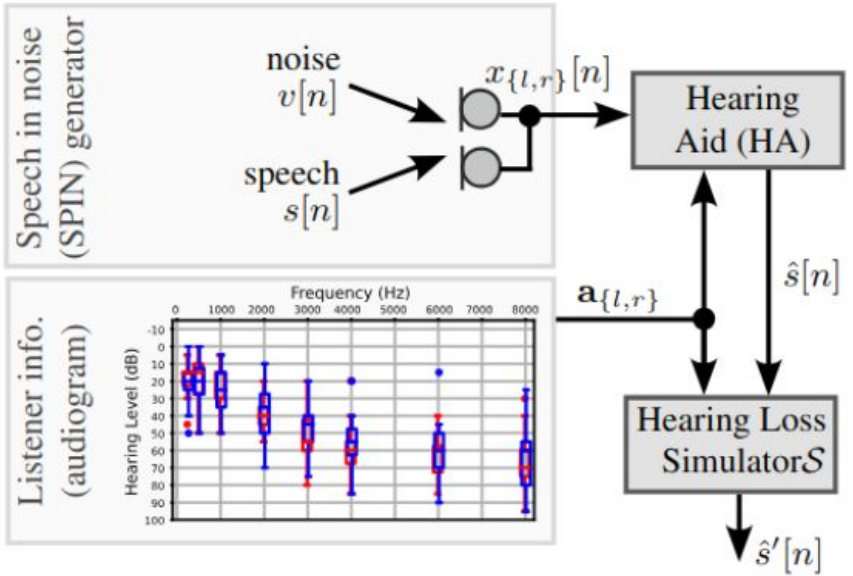
Motivation

- Hearing loss affects approximately 12 million people (1 in 5) in the UK, with the number expected to grow
- Automatic evaluation of speech intelligibility can help with the development of hearing aids
  - Typically, testing of hearing aid systems is expensive and time consuming
- Human audio rating prediction is an emerging area of research
- **Goal**: **Non-intrusive speech intelligibility prediction**

Clarity Prediction Challenge

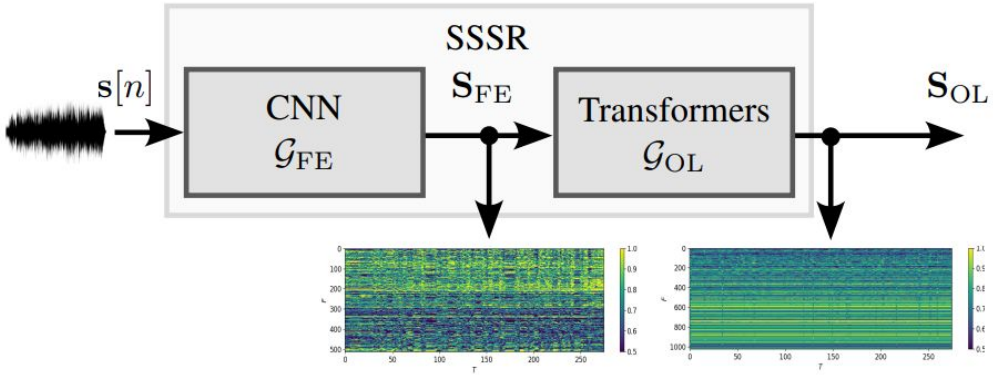
- Predict the number of words correctly identified by hearing impaired listeners
- Generalise to unseen listeners and systems

- Our system:
  - Non-intrusive
  - Uses only the output of the hearing aid enhancement
  - No explicit listener information used



Prior Work

- Recent work [1] has found that Self Supervised Speech Representations (SSSRs) are useful feature representations for speech **quality** estimation
- In our prior work [2], SSSRs are used successfully for in non-intrusive **intelligibility** prediction for the CPC1 challenge data
- Specifically, the **CNN Encoder** representations are useful
- However, such models may generalize poorly to unseen enhancement systems and listeners



[1]*"Pre-trained Speech Representations as Feature Extractors for Speech Quality Assessment in Online Conferencing Applications." B. Tamm, H. Balabin, R. Vandenberghe, H. Van hamme Interspeech 2022,*
[2] *"Non Intrusive Intelligibility Predictor for Hearing Impaired Individuals using Self Supervised Speech Representations" G. Close, T. Hain and S. Goetze, 2023*

WHISPER ASR System

- Weakly supervised ASR model [3] trained on 680,000 hours of data
- Time-domain signal is down-sampled to 16 kHz and padded to 30 seconds
- Input is the 80-channel log Mel Spectrogram, with a window of 25ms, stride of 10ms
- 12 encoder layers, 12 decoder layers
- Decoder layers appeared to be more useful than encoder layers in our experiments

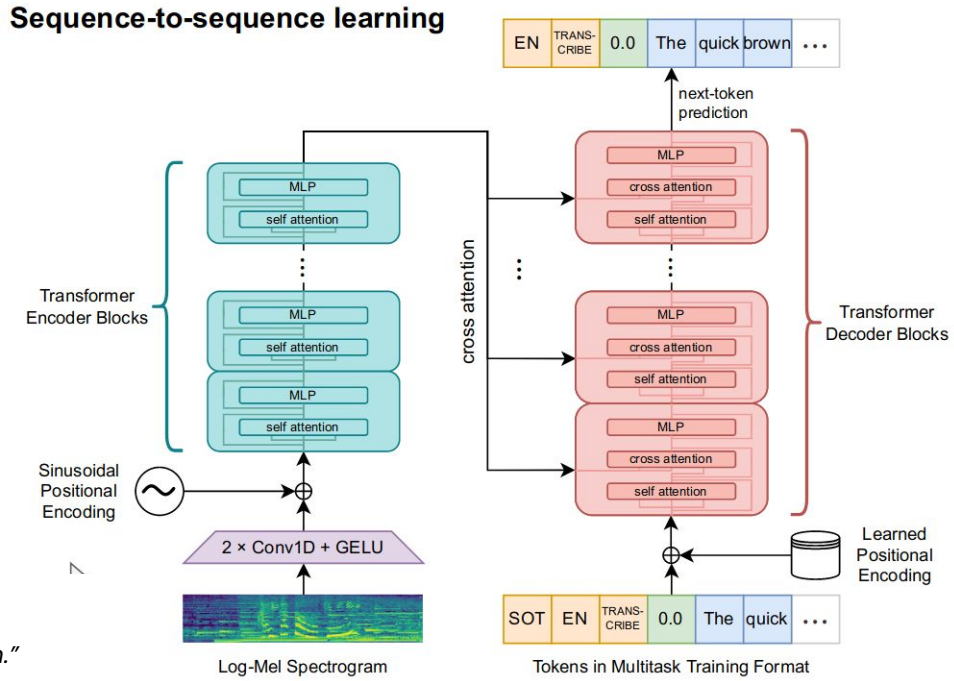
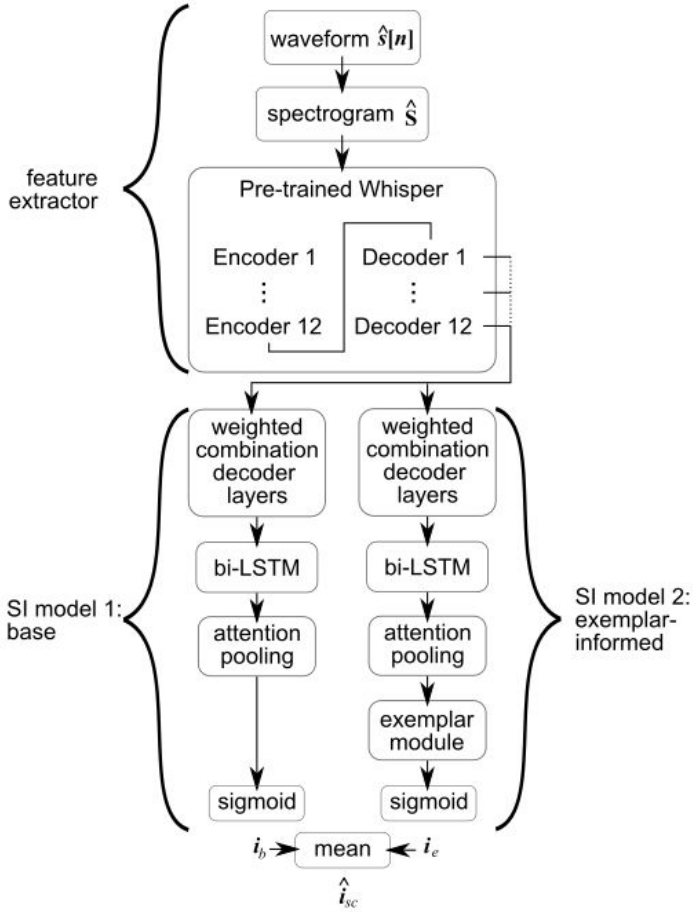


Figure taken from [3]

[3] *"Robust speech recognition via large-scale weak supervision." Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, CHristine McLeavey, Ilya Sutskever, 2022.*

Proposed Framework

- The model takes as input the output of each Whisper decoder layer
- These are weighted and passed through a BLSTM layer followed by a single attention head to a single output neuron with a sigmoid activation representing the predicted intelligibility
- We also use a a model feeding the output of the attention head into an exemplar-informed module
- The predicted intelligibility is the mean of the two model outputs.



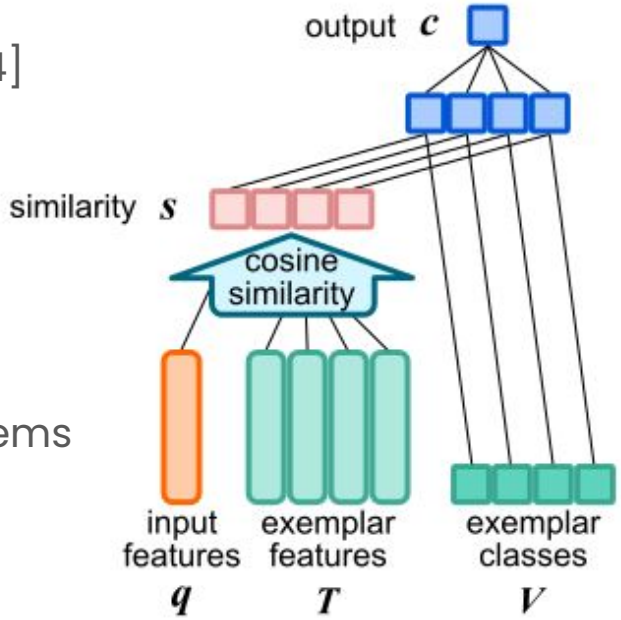
Exemplar-Informed Module

- Based on simulated human memory model [4]
- Incorporates a set of “exemplars”
  - Labelled examples from the training data
- Output is a weighted combination of the exemplar labels

Benefits:

- Potential to easily adapt to new listeners/systems

Although it looked promising on the validation set, there was little benefit on the evaluation set



[4]*”Minerva 2: a simulation of human memory.” D. Hintzman, 1986*

Experiment Setup

- Validation sets for model selection:
  - Disjoint validation sets: 2 listeners and 2 systems selected randomly from each training split
  - Non-disjoint validation set: 10% of remaining training data (used for best epoch)
- For the final models, the disjoint validation sets were folded back into the training data

The base and exemplar-informed systems are trained separately

- SI model 1: Base:
  - Trained for 25 epochs with batch size 8 and learning rate 10-5
- SI model 2: Exemplar-informed:
  - Trained for 50 epochs with batch size 8 and learning rate 2x10-6
  - 8 exemplar randomly selected from the training data for each minibatch

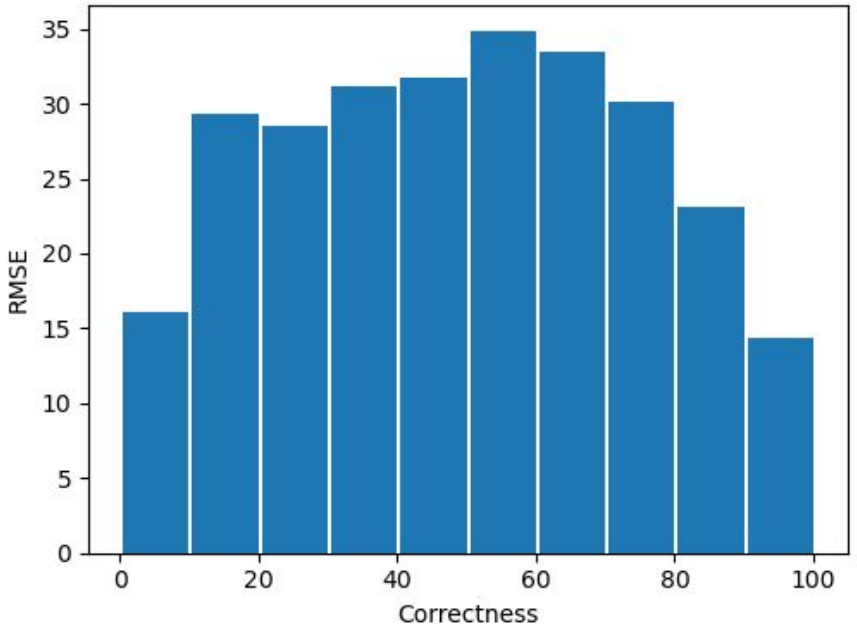
Results

- Lower performance on Split 1
  - Enhancement System E001 very poor, outlier
- Generalizes to unseen enhancement systems and listeners in both validation and evaluation

Split	RMSE	
	validation	evaluation
1	21.6	28.60
2	23.4	23.88
3	22.7	23.15
Overall	22.5	**25.32**

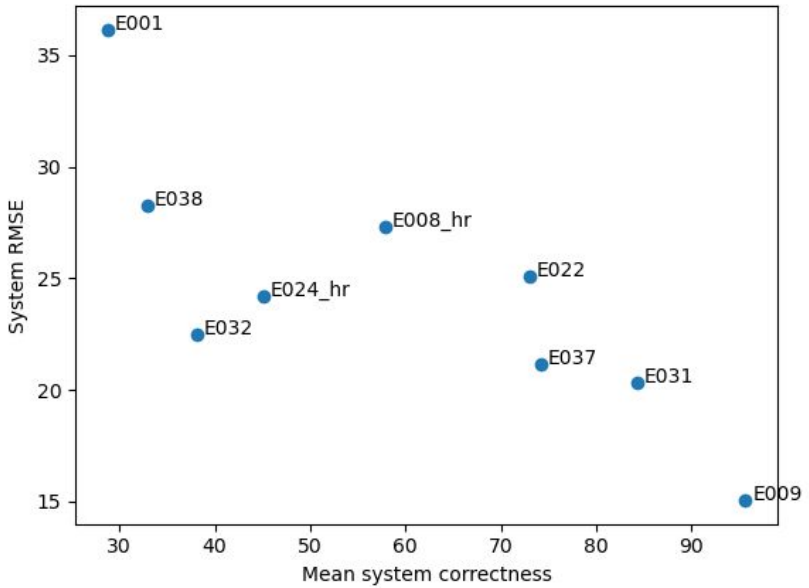
Analysis - RMSE by Correctness

- Good performance for very high- and low-intelligibility speech
- Poorer performance for medium-intelligibility speech
- Matches data availability



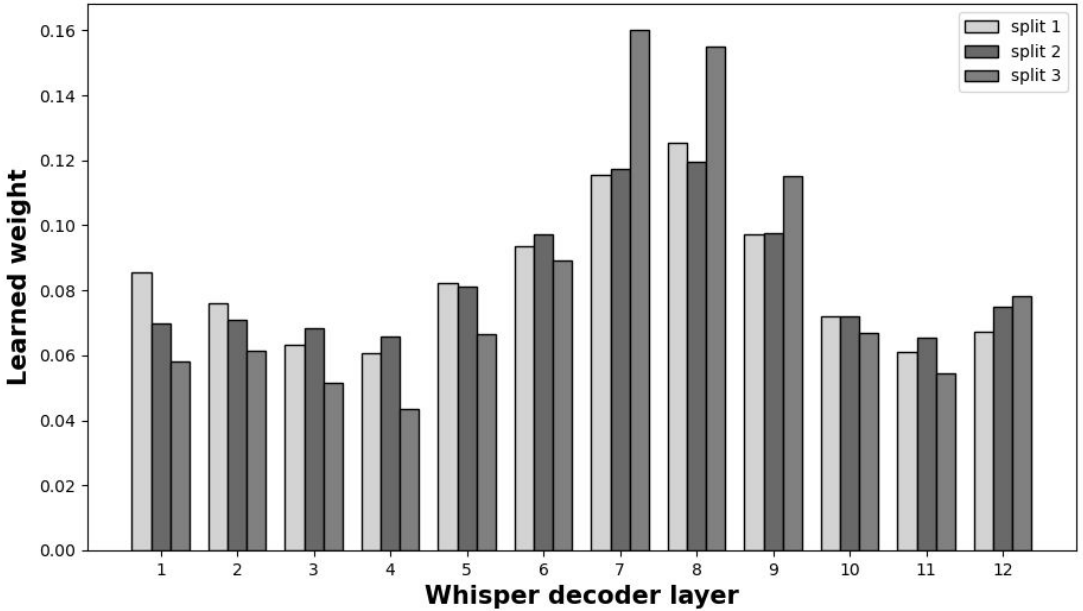
Analysis – RMSE by System

- Poor performance on systems with low mean correctness
- Performance on system E001 is particularly poor



Analysis - Whisper Features

- Layers 7 and 8 are preferred
- Similar weights learned for different splits
- This is consistent with other work which use intermediate features



Conclusions

- Pre-trained **WHISPER decoder layers** are a useful feature representation for speech intelligibility prediction

- While the proposed system does **generalize to unseen enhancement systems**, badly performing enhancement systems are more difficult to predict accurately.
  - This is an improvement over our prior work which tended to overfit to the enhancement system

The University Of Sheffield.

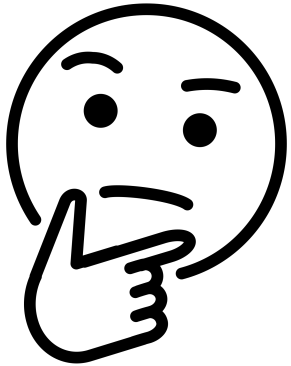
SLT Centre for Doctoral Training

UK Research and Innovation

Thank You!

Any Questions?

rmogridge1@sheffield.ac.uk



Analysis - Split 1

- Our model overestimates the scores of E001
- Excluding E001 from evaluation set results in similar performance to the other splits