
Preprint:
Close, G., Hain, T. and Goetze, S. (Submitted: 2023) Non intrusive intelligibility predictor for hearing impaired individuals using self supervised speech representations. [Preprint - arXiv] (Submitted)

https://doi.org/10.48550/arxiv.2307.13423

NON-INTRUSIVE SPEECH INTELLIGIBILITY PREDICTION FOR HEARING IMPAIRED INDIVIDUALS USING SELF-SUPERVISED SPEECH REPRESENTATIONS

George Close, Thomas Hain, Stefan Goetze

Speech and Hearing Group, Department of Computer Science, University of Sheffield, United Kingdom

ABSTRACT

Self-supervised speech representations (SSSRs) have been successfully applied to a number of speech-processing tasks, e.g. as feature extractor for speech quality (SQ) prediction, which is, in turn, relevant for assessment and training speech enhancement systems for users with normal or impaired hearing. However, exact knowledge of why and how quality-related information is encoded well in such representations remains poorly understood. In this work, techniques for non-intrusive prediction of SQ ratings are extended to the prediction of intelligibility for hearing-impaired users. It is found that self-supervised representations are useful as input features to non-intrusive prediction models, achieving competitive performance to more complex systems. A detailed analysis of the performance depending on Clarity Prediction Challenge 1 listeners and enhancement systems indicates that more data might be needed to allow generalisation to unknown systems and (hearing-impaired) individuals.

Index Terms: hearing loss, metric prediction, neural networks, self-supervised speech representations

1. INTRODUCTION

Age-related hearing loss (HL) is an increasingly prevalent problem in countries with ageing populations worldwide. In the United Kingdom, for example, approximately 12 million people suffer from HL of greater than 25 decibels hearing level (dBHL); by 2035 this is predicted to increase to 14.2 million [1]. Hearing loss can often impede an individual's ability to participate in a spoken conversation, especially in noisy environments, as parts of the speech can become unintelligible [2]. As such, the development of methods to increase speech intelligibility (SI) in assistive listening devices to alleviate this is of paramount importance [3]. While there have been large improvements in speech enhancement technology thanks to neural network-based approaches [4–6] these can often be challenging to implement in small form factor hearing aid (HA) hardware. Furthermore, given that the exact severity and nature of hearing loss differs greatly between individuals, a 'one size fits all' approach is not viable. The Clarity Project [7] aims to improve the design of hearing aids via two alternating challenges and related datasets [8], the Clarity Enhancement Challenge (CEC) and the Clarity Prediction Challenge (CPC). The CEC is concerned with the design of the actual enhancement algorithm, while the CPC involves the prediction of the intelligibility of the enhanced speech for hearing-impaired listeners. The overall aim of the CPC is to produce systems that mimic the behaviour of hearing-impaired listeners, reducing the need for expensive and time-consuming human assessment by listening tests, while also providing a training target or metric for enhancement systems.

Self-supervised speech representations (SSSRs) have been found to be useful either as pretrained layers or feature transformations in many speech-related applications [9–11]. It is understood that SSSRs are able to encode and predict the context of the speech content in the input audio, and thus model the patterns of spoken language. Recent work [12–16] has found that in addition to speech content, SSSRs are also able to capture information on potentially corrupting noise and distortion in the input audio.

In this work, SSSRs are used as a feature transformation for non-intrusive neural speech intelligibility prediction networks, trained on the CPC1 challenge dataset. Non-intrusive metric prediction networks using SSSRs are proposed to serve as feature extractors and analysed for different latent or output SSSR layers to predict SI for hearing-impaired users.

The remainder of the paper is structured as follows. In Section 2, the SSSRs used in this work are briefly introduced and Section 3 reviews the use of SSSRs in related tasks. Section 4 introduces the dataset as used in this work and in Section 5 the relationships between this dataset and SSSR distance measures are examined. Finally, in Sections 6 and 7, experiments involving the use of SSSR as feature representations in non-intrusive intelligibility prediction networks are described, and the results analysed.

THIS WORK WAS SUPPORTED BY [redacted]

2. SELF SUPERVISED SPEECH REPRESENTATIONS

Generally, SSSRs are neural networks that, given a waveform representation of speech audio $\mathbf{s}[n]$, produce a final output which expresses the *context* of that input speech audio. As the name suggests, they are trained in a self-supervised way, typically by *masking* some segment of the input audio representation and tasking the network in training to recreate the masked segment [9]. Structurally, the networks consist of two distinct stages as shown in Fig. 1.

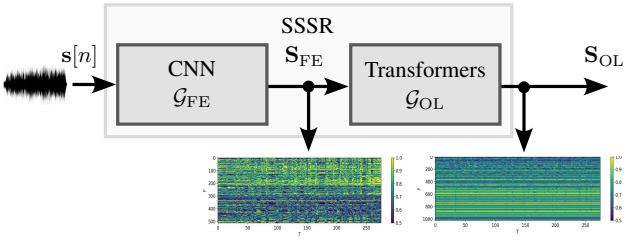

Fig. 1. Representations extracted from SSSR model with time-domain input signal $\mathbf{s}[n]$. Feature channels are sorted [17] and values normalised for clarity.

The input waveform $\mathbf{s}[n]$ with discrete time index n is first processed by a number of 1-D convolutional layers, resulting in a two-dimensional feature encoding representation. In the second stage, this representation is processed by a number of Transformer [18] layers, to give the final two-dimensional output. In both stages, one of the dimensions represents time, while the other represents features. For a time domain signal $\mathbf{s}[n]$, the output of the initial convolutional neural network (CNN) encoder stage $\mathcal{G}_{\mathrm{FE}}$ of a SSSR is

$$\mathbf{S}_{\mathrm{FE}} = \mathcal{G}_{\mathrm{FE}}(\mathbf{s}[n]), \tag{1}$$

where operator $\mathcal{G}_{\mathrm{FE}}$ denotes the 1-D convolutional layers encompassing the encoder. The subsequent processing by the Transformer based stage can be defined by an additional operator $\mathcal{G}_{\mathrm{OL}}$ denoting the Transformer layers encompassing the final output stage, i.e.

$$\mathbf{S}_{\mathrm{OL}} = \mathcal{G}_{\mathrm{OL}}(\mathcal{G}_{\mathrm{FE}}(\mathbf{s}[n])). \tag{2}$$

Both signal representations $\mathbf{S}_{\mathrm{FE}}$ and $\mathbf{S}_{\mathrm{OL}}$ have two dimensions: time T, depending on the length of the input audio in block time, and feature dimension F.

The *Cross-Lingual Speech Representation (XLSR)* [19] is one of the SSSR representations used in this work. It is a variant of the Wav2Vec2 [9] structure. It is trained on $436,000$ hours of audio data from a number of languages, including the BABEL dataset which contains potentially noisy telephone conversation recordings. Its network is structured in the way described above, with the outputs of $\mathcal{G}_{\mathrm{FE}}$ having a feature dimension of 512 and the final (transformer) outputs of $\mathcal{G}_{\mathrm{OL}}$ having a feature dimension of 1024. In this work, the smallest version of XLSR sourced from the HuggingFace[1] is used.

Hidden Unit Bidirectional Encoder Representations from Transformers (HuBERT) [10] differ w.r.t. the general training of a SSSR described above in that the training target is a cluster of masked frames similar to BERT [20] rather than the masked frame itself. However, its network structure follows the same pattern, with the outputs of $\mathcal{G}_{\mathrm{FE}}$ having a feature dimension of 512 and the final output representation a feature dimension of 768. In this work, we use the HuBERT Large model trained on 960 hours of clean English speech sourced from the LibriSpeech [21] dataset, from the Fairseq GitHub repository[2].

3. SSSRS FOR METRIC PREDICTION

SSSRs have been applied to metric prediction tasks, typically to quality prediction [22, 23]. In [13], XLSR representations are used as feature extraction in a non-intrusive human MOS prediction network. Similarly, in [24] SSSRs are used for the same quality prediction task, but they are fine-tuned with a mean pooling layer rather than being used simply as feature extraction. SSSRs were also applied to the CPC1 challenge in [25], where they were used as feature extractors alongside spectrograms and learnable filter banks.

In all these cases, only the final SSSR output $\mathcal{G}_{\mathrm{OL}}$ was considered. However, findings in [12] suggest that the output of the initial encoding stage $\mathcal{G}_{\mathrm{FE}}$ better captures quality-related information. As such, in this work, both representations stages are considered and compared as feature transformations for speech intelligibility prediction.

4. CLARITY CHALLENGE 1 DATA

The dataset for the first Clarity Prediction Challenge 1 (CPC1) [26] as used in this work can be expressed as a series of sequences: $(\hat{\mathbf{s}}[n], \{\mathbf{a}_l, \mathbf{a}_r\}, i)$, which is generated as visualised in Fig. 2. $\hat{\mathbf{s}}[n]$ represents the output of a hearing aid system for some noisy speech input $\mathbf{x}[n]$, containing some clean speech $\mathbf{s}[n]$. $\{\mathbf{a}_l, \mathbf{a}_r\}$ are the left and right audiogram representations of a particular listener's hearing loss. Blue and red box plots in Fig. 2 illustrate the HL distribution in the CPC1 dataset from which the individual audiograms are sampled. Finally, i represents the intelligibility of the audio $\hat{\mathbf{s}}[n]$ for that listener, defined as the percentage of words they were able to reproduce by speaking aloud immediately after hearing the audio, compared to a ground truth transcription of the speech also denoted as the *correctness* of the listener's response. Additionally, audio $\hat{\mathbf{s}}'[n]$ is defined as the output of the baseline Cambridge MSBG hearing loss simulator (HLS),

[1] https://huggingface.co/facebook/wav2vec2-xls-r-300m

[2] https://github.com/facebookresearch/fairseq

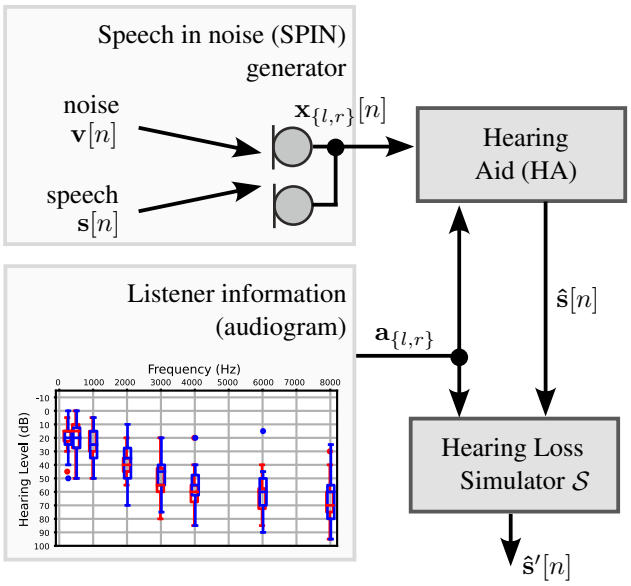

Fig. 2. Signal generation for Clarity Prediction Challenge.

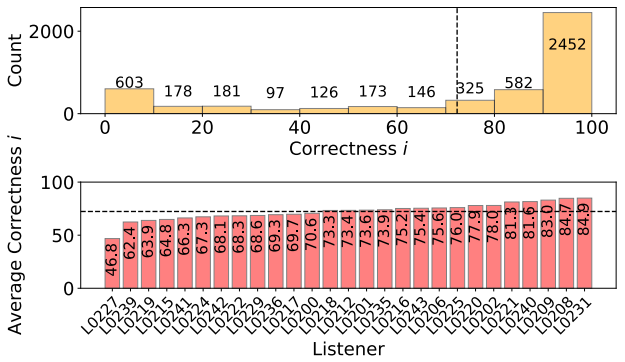

Fig. 3. Histogram showing the distribution of ground truth correctness i in CPC1 training set (top) and a bar chart showing average correctness i per listener in the CPC1 training set (bottom). Dotted lines are respective overall average values.

denoted here by operator $\mathcal{S}$, cf. [27] for additional details on the clarity system.

$$\hat{\mathbf{s}}'[n] = \mathcal{S}(\hat{\mathbf{s}}[n], \{\mathbf{a}_l, \mathbf{a}_r\}) \tag{3}$$

The signal $\hat{\mathbf{s}}'[n]$ is an approximation of the audio that is perceived by the hearing-impaired listener. This can be thought of as encoding the hearing characteristics of the specific listener (audiogram) within the signal. Note that all signals in the dataset are binaural i.e. consist of left and right channels, denoted by l and r, respectively.

The upper plot in Fig. 3 shows the distribution of correctness i in the CPC1 training set. From this, it can be observed that in the majority of cases, the listener was able to fully reproduce the speech in the audio they heard, i.e. $i = 100$ for $\approx 50\%$ of the assessed files. The next largest class is where $i = 0$, meaning that the listener was not able to reproduce any words in the audio. This distribution is due to the more realistic *in-the-wild* SI measurement strategy for the Clarity dataset [26] which is in contrast to lab-based SI matrix tests [28]. The lower panel of Fig. 3 shows the average correctness i for each listener in the CPC1 training set. With the exception of listener L0227, all of the listeners achieve similar performance.

5. ANALYSING RELATIONSHIPS BETWEEN SSSRS AND HUMAN SPEECH INTELLIGIBILITY

In order to express the relationship between SSSRs and correctness i in the dataset, two distance measures are defined in a mean squared error (MSE) sense:

$$d_{\mathrm{FE}} = \frac{1}{TF}\sum_{t}^{T}\sum_{f}^{F}(\mathbf{S}_{\mathrm{FE}}[t,f] - \mathbf{P}_{\mathrm{FE}}[t,f])^2 \tag{4}$$

$$d_{\mathrm{OL}} = \frac{1}{TF}\sum_{t}^{T}\sum_{f}^{F}(\mathbf{S}_{\mathrm{OL}}[t,f] - \mathbf{P}_{\mathrm{OL}}[t,f])^2 \tag{5}$$

The distance d_{FE} in (4) expresses the MSE distance between the SSSR *feature encoding* layer representations $\mathbf{S}_{\mathrm{FE}}[t,f]$ of the clean reference audio $\mathbf{s}[n]$ and the representations $\mathbf{P}_{\mathrm{FE}}[t,f]$ of the test signal $\mathbf{p}[n]$, while (5) expresses the MSE distance between the SSSR *output layer* representations $\mathbf{S}_{\mathrm{OL}}[t,f]$ and $\mathbf{P}_{\mathrm{OL}}[t,f]$, with t and f denoting block time and feature index, respectively. Note that $\mathbf{p}[n]$ and is a placeholder for either the speech signal after HA enhancement $\hat{\mathbf{s}}[n]$ or this signal after HLS processing $\hat{\mathbf{s}}'[n]$ as shown in Fig. 2. Distances in (4), (5) are designed to express the distortion captured by the SSSR due to the transformations which have been applied to $\mathbf{s}[n]$ to produce e.g. $\hat{\mathbf{s}}'[n]$, i.e. the artificial distortion/reverb added to create $\mathbf{x}[n]$, enhancement by the hearing aid system (in $\hat{\mathbf{s}}[n]$) and finally the HLS. In addition to distances (4) and (5) the MSE distance between spectrogram representations of $\mathbf{s}[n]$ and $\mathbf{p}[n]$,

$$d_{\mathrm{SG}} = \frac{1}{TF_{\mathrm{Hz}}}\sum_{t}^{T}\sum_{f_{\mathrm{Hz}}}^{F_{\mathrm{Hz}}}(\mathbf{S}_{\mathrm{SG}}[t,f_{\mathrm{Hz}}] - \mathbf{P}_{\mathrm{SG}}[t,f_{\mathrm{Hz}}])^2, \tag{6}$$

will be analysed, with f_{Hz} and F_{Hz} denoting the technical frequency and the highest frequency analysed, respectively. In the following, the left (first) channel of the audio is used to compute the distance measures (4), (5) and (6).

Table 1 shows the Spearman and Pearson correlations of the MSE distances (4)-(6) with the correctness values i for the CPC1 training set. Absolute correlations are low, but this

Table 1. Spearman and Pearson correlations between distance measures and correctness values i in the CPC1 training set, strongest correlations in bold.

Representation, Distance	$\mathbf{p}[n]$	Spearman	Pearson
SPEC, d_{SG}, (6)	$\hat{\mathbf{s}}[n]$	-0.10	-0.18
SPEC, d_{SG}, (6)	$\hat{\mathbf{s}}'[n]$	-0.09	-0.07
XLSR, d_{FE}, (4)	$\hat{\mathbf{s}}[n]$	-0.13	-0.16
XLSR, d_{FE}, (4)	$\hat{\mathbf{s}}'[n]$	-0.24	-0.28
XLSR, d_{OL}, (5)	$\hat{\mathbf{s}}[n]$	-0.26	-0.27
XLSR, d_{OL}, (5)	$\hat{\mathbf{s}}'[n]$	-0.24	-0.24
HuBERT, d_{FE}, (4)	$\hat{\mathbf{s}}[n]$	$\mathbf{-0.38}$	$\mathbf{-0.47}$
HuBERT, d_{FE}, (4)	$\hat{\mathbf{s}}'[n]$	-0.23	-0.29
HuBERT, d_{OL}, (5)	$\hat{\mathbf{s}}[n]$	-0.10	-0.17
HuBERT, d_{OL}, (5)	$\hat{\mathbf{s}}'[n]$	-0.28	-0.32

is expected for the Clarity dataset (cf. [26] and Section 4). Comparing the distances between the feature representations in (4)-(6) and the intelligibility scores i allows for an expression of how distortion in the signal, which might affect intelligibility, is captured by that feature representation. Interestingly, applying the hearing loss simulation $\mathcal{S}$ in (3) does not uniformly improve the correlation with i across all distances in Table 1; only for the XLSR encoder output representation distance d_{FE} and the HuBERT final output representation distance d_{OL} does using $\hat{\mathbf{s}}'[n]$ lead to higher correlation than using $\hat{\mathbf{s}}[n]$.

6. SSSR-BASED INTELLIGIBILITY PREDICTION

This section proposes the use of SSSRs as features in non-intrusive neural intelligibility prediction networks. Following the findings from Table 1, both the hearing aid output signal $\hat{\mathbf{s}}[n]$ and that signal processed by the hearing loss simulation $\hat{\mathbf{s}}'[n]$ are used as the input audio to the models, as no conclusive best representation is indicated by these results.

6.1. Dataset

Models are trained on both the open and closed training and test sets detailed in the CPC1 description [8]. The closed set has the same listeners and systems for both train and testsets, while the open set has 5 unseen listeners and 1 unseen system in its testset. For more detail as to how these sets are differentiated, see [8]. A validation set is created using 10% of the available training data. As we are interested in non-intrusive predictors, either the hearing aid output signal $\hat{\mathbf{s}}[n]$ or the hearing loss simulated audio signal $\hat{\mathbf{s}}'[n]$ are used to predict the Correctness label i. For the closed set, the training set contains 4376 utterances, the validation set 487 and the test set 2421. The training set contains 3222 utterances for the open set, 358 for the validation set, and 632 for the test set.

6.2. Model Structure and Experiment Setup

A model structure inspired by [13] is chosen for the SI prediction network. Five feature extraction methods are used; outputs of $\mathcal{G}_{\mathrm{FE}}$ and $\mathcal{G}_{\mathrm{OL}}$ for both, HuBERT and XLSR representations, as well as a spectrogram representation denoted as SPEC. After the feature extraction, the resultant representation is processed by 2 bidirectional long short-term memory (BLSTM) layers with an input size equal to the feature dimension F of the input and a hidden layer size of $F/2$. The final layer is an attention pooling feed-forward layer, similar to that in NISQA [22] with a single output neuron and a sigmoid activation to output the predicted correctness $\hat{i}$ (normalised between 0 and 1). Note that due to different dimensions F of different feature representations, the number of parameters in each network varies from $923,906$ for the models using spectrogram representations to as many as $14,701,570$ for the models using the XLSR output layer, i.e. $\mathcal{G}_{\mathrm{OL}}$.

The two input audio representations $\hat{\mathbf{s}}[n]$ or $\hat{\mathbf{s}}'[n]$ are used, i.e. the output of the hearing aid systems and the enhanced audio processed by the hearing loss simulation, as in (3). As these audio representations have two channels, each channel is processed by the model separately; during training, the loss for each channel is computed and then summed before being back-propagated to the model. During validation and testing, the maximum value between each channel is taken as an approximation of the *better ear effect* [29].

The spectrogram representation is created by a short time Fourier transform (STFT) with a window length of 20 ms, a hop length of 10 ms and an FFT size of 1024. All audio is re-sampled to 16 kHz such that it can be used as inputs to the SSSR models.

7. RESULTS

In addition to the intrusive (reference-signal-based) challenge baseline, the best-performing non-intrusive entries to the challenge are reported in this section as additional baselines, as the proposed system is also non intrusive. Challenge entry E23 [30] makes use of contrastive predictive coding and vector quantisation features. E06 [31] is similar to the proposed system, denoted by SPEC in the following, but uses a CNN based network structure. E33 [25] also utilises SSSRs as feature extraction, but spectrogram and learnable filterbank features are also used as model inputs. E29 [32] makes use of an information-theory-inspired approach, wherein the difference between internal representations in neural automatic speech recognition (ASR) systems is used to approximate human intelligibility, and was the overall best non-intrusive challenge entry.

Table 2. Non-Intrusive Prediction Performance on the CPC1 closed set. Best performances for baselines and proposed methods in boldface font.

Model Name	RMSE	Var	Spearman	Pearson
CPC1 Baseline	*28.50*	–	*0.62*	–
E23 [30]	*41.50*	–	*0.07*	–
E06 [31]	*32.00*	–	*0.43*	–
E33 [25]	*24.10*	–	*0.75*	–
E29 [32]	***23.30***	–	***0.77***	–
SPEC $\hat{\mathbf{S}}_{\mathrm{SPEC}}$	25.45	0.52	0.59	0.72
SPEC $\hat{\mathbf{S}}'_{\mathrm{SPEC}}$	25.45	0.52	0.58	0.72
HuBERT $\hat{\mathbf{S}}_{\mathrm{FE}}$	30.82	0.61	0.44	0.56
HuBERT $\hat{\mathbf{S}}'_{\mathrm{FE}}$	26.64	0.53	0.56	0.70
HuBERT $\hat{\mathbf{S}}_{\mathrm{OL}}$	**24.76**	**0.50**	0.59	**0.74**
HuBERT $\hat{\mathbf{S}}'_{\mathrm{OL}}$	24.82	**0.50**	**0.61**	**0.74**
XLSR $\hat{\mathbf{S}}_{\mathrm{FE}}$	25.01	**0.50**	0.60	**0.74**
XLSR $\hat{\mathbf{S}}'_{\mathrm{FE}}$	25.33	0.51	0.60	0.72
XLSR $\hat{\mathbf{S}}_{\mathrm{OL}}$	28.42	0.58	0.47	0.66
XLSR $\hat{\mathbf{S}}'_{\mathrm{OL}}$	30.20	0.61	0.52	0.64

7.1. Results on CPC1 Closed set

Table 2 shows the performance of the proposed systems for the CPC1 closed set. All proposed systems show comparable performance with the best-performing challenge entries, although, none of the proposed systems outperforms system E29. It should be noted, however, that the computation overhead to implement system E29 is significantly greater than any of the proposed systems here, as several state-of-the-art ASR systems must be trained and fine-tuned for E29. Of the proposed systems trained on the outputs of the hearing loss simulation $\hat{\mathbf{s}}'[n]$, the best performing is the model which uses HuBERT output representations $\hat{\mathbf{S}}'_{\mathrm{OL}}$ as features. This is consistent with the findings in Table 1 which shows that the distance measure using this representation had the highest correlation with i of those distances computed using $\hat{\mathbf{s}}'[n]$. Of those trained using the hearing aid outputs $\hat{\mathbf{s}}[n]$, HuBERT's output $\hat{\mathbf{S}}_{\mathrm{OL}}$ is also the best performing achieving near identical performance to the $\hat{\mathbf{s}}'[n]$ model. In terms of the difference in performance between using earlier SSSR representations $\mathcal{G}_{\mathrm{FE}}$ or output representations $\mathcal{G}_{\mathrm{OL}}$ as features, this seems to depend on the SSSR used; for HuBERT the output layers perform best, while for XLSR the feature encoder layers show better performance.

7.2. Results of CPC1 Open set

Table 3 shows the performance of the proposed systems on the more challenging CPC1 open set (cf. Section 4). Performance of the proposed systems is significantly worse than that of the closed set for all systems, with a much larger variance in MSE in all cases, but all proposed systems still outperform the baseline. The poorer performance might be due to over-

Table 3. Non Intrusive Prediction Performance on the CPC1 open set.

Model Name	RMSE	Var	Spearman	Pearson
CPC 1 Baseline	*36.50*	–	*0.53*	–
E23 [30]	*43.70*	–	*0.05*	–
E33 [25]	*28.9*	–	*0.65*	–
E29 [32]	***24.60***	–	***0.73***	–
SPEC $\hat{\mathbf{S}}_{\mathrm{SPEC}}$	32.84	1.29	0.35	0.50
SPEC $\hat{\mathbf{S}}'_{\mathrm{SPEC}}$	29.16	1.15	0.57	0.60
HuBERT $\hat{\mathbf{S}}_{\mathrm{FE}}$	33.69	1.30	0.27	0.45
HuBERT $\hat{\mathbf{S}}'_{\mathrm{FE}}$	35.31	1.40	0.19	0.24
HuBERT $\hat{\mathbf{S}}_{\mathrm{OL}}$	32.43	1.22	0.47	0.54
HuBERT $\hat{\mathbf{S}}'_{\mathrm{OL}}$	29.66	1.14	0.60	0.61
XLSR $\hat{\mathbf{S}}_{\mathrm{FE}}$	31.83	1.26	0.49	0.52
XLSR $\hat{\mathbf{S}}'_{\mathrm{FE}}$	30.86	1.19	0.56	0.56
XLSR $\hat{\mathbf{S}}'_{\mathrm{OL}}$	31.85	1.25	0.42	0.49
XLSR $\hat{\mathbf{S}}'_{\mathrm{OL}}$	34.54	1.36	0.26	0.37

fitting of the models to the training data, (in particular to the enhancement systems in the training set) as the test data contains unseen enhancement systems and listeners. All of the models here perform similarly poorly.

7.3. System and Listener-wise Analysis

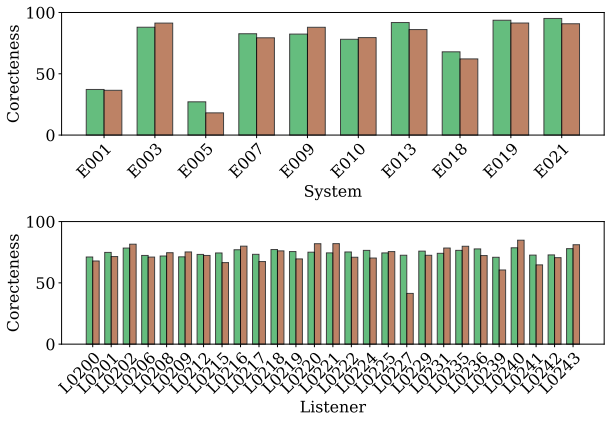

Fig. 4. System (top) and listener-wise (bottom) correctness prediction $\hat{i}$ (l./green) vs. true i (r./brown) using HuBERT output for $\hat{\mathbf{S}}_{\mathrm{OL}}$ model on CPC1 closed set.

For further analysis, Figs. 4 and 5 show ground truth and predicted correctness across the hearing aid systems and across the listeners in the CPC1 open testset for the HuBERT $\hat{\mathbf{S}}_{\mathrm{OL}}$ and HuBERT $\hat{\mathbf{S}}'_{\mathrm{OL}}$ models, respectively. Both models show similar performance across the different hearing aid systems, both successfully assigning low scores to the audio enhanced by the E005 hearing aid system. This indicates that the models are able (at some level) to detect the distortions introduced

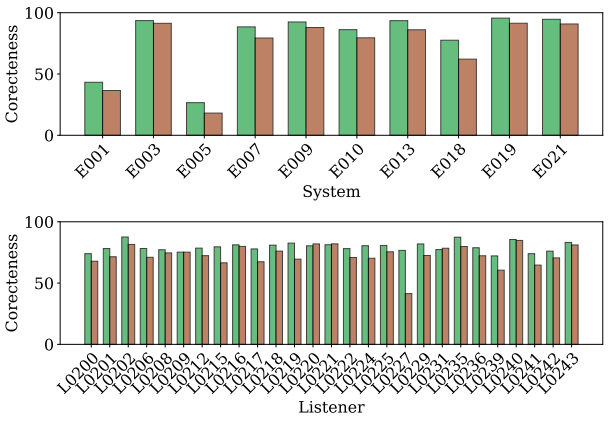

Fig. 5. System (top) and listener-wise (bottom) correctness prediction $\hat{i}$ (l./green) vs. true i (r./brown) using HuBERT outpur for $\hat{\mathbf{S}}'_{\mathrm{OL}}$ model on CPC1 closed set.

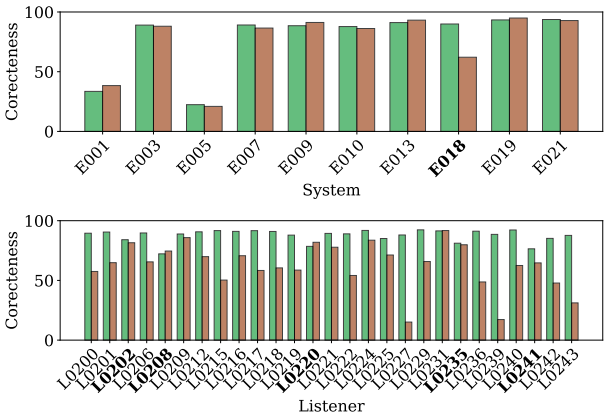

Fig. 6. System (top) and listener-wise (bottom) correctness prediction $\hat{i}$ (l./green) vs. true i (r./brown) using HuBERT output for $\hat{\mathbf{S}}_{\mathrm{OL}}$ model on CPC1 open set. Listeners and Systems unseen during training are bold.

by this enhancement. Similarly, there is little difference in performance across the subset of listeners for the two models; this suggests that the listener-specific hearing loss information which the $\hat{\mathbf{S}}'_{\mathrm{OL}}$ model has access to (encoded in the audio) does not aid in the intelligibility prediction performance. It should be noted that already the enhancement system (hearing aid) has (implicitly) access to the hearing loss information and is expected to process its input signal accordingly (cf. Fig. 2). Interestingly, both models overestimate the intelligibility ratings of speaker L0227 who performs worse than average at the intelligibility task (cf. Fig. 3). This suggests that L0227's lower performance is not due to their hearing loss but rather other unknown factor(s); audiogram information for this listener does not show that they have particularly severe hearing loss.

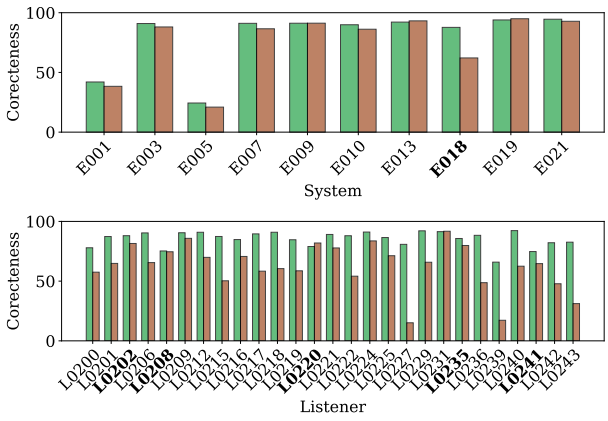

Fig. 7. System (top) and listener-wise (bottom) correctness prediction $\hat{i}$ (l./green) vs. true i (r./brown) using HuBERT output for $\hat{\mathbf{S}}'_{\mathrm{OL}}$ model on CPC1 open set. Listeners and Systems unseen during training are bold.

Figs. 7 and 6 show ground truth and predicted correctness across the hearing aid systems and across the listeners for the more challenging CPC1 closed testset for the HuBERT output for $\hat{\mathbf{S}}_{\mathrm{OL}}$ and HuBERT output for $\hat{\mathbf{S}}'_{\mathrm{OL}}$ models, respectively. Systems and listeners which are unseen during the training of the models are highlighted by bold-font. Here, the overfitting of the proposed system to the hearing aid systems during training can be observed by the poor performance on the unseen hearing aid system in the testset, E018. The overall lower performance of the proposed systems on the closed set is shown by the listener-wise plots, with both systems significantly overestimating the correctness versus the true value; however the encoding of the hearing loss information in $\hat{\mathbf{S}}'_{\mathrm{OL}}$ does appear to have some positive effect here.

8. CONCLUSION

This work explores the use of SSSR models as feature extraction for non-intrusive speech intelligibility prediction networks in comparison to traditional, spectrogram-based input. Both, the final SSSR representation and the intermediate output of the SSSR feature encoder are compared for the first time for an SI prediction task for hearing-impaired users. Results indicate that encoding the hearing loss of a particular listener via (an additional) hearing loss simulation does not typically improve performance. Additionally, models tend to overfit to specific hearing aid systems, as demonstrated by the results on the open set which might be alleviated by larger datasets released in the future. The upcoming CPC2 challenge data, which includes twice the number of enhancement systems, could mitigate this issue.

9. REFERENCES

[1] Neil Park, "Population estimates for the UK, England and Wales, Scotland and Northern Ireland, provisional: mid-2019," *Hampshire: Office for National Statistics*, 2020.

[2] Christoph Völker, Anna Warzybok, and Stephan M.A. Ernst, "Comparing Binaural Pre-processing Strategies III: Speech Intelligibility of Normal-Hearing and Hearing-Impaired Listeners," *Trends in Hearing*, vol. 19, 2015.

[3] Stefan Goetze, Feifei Xiong, Jan Rennies, Thomas Rohdenburg, and Jens-E. Appell, "Hands-free telecommunication for elderly persons suffering from hearing deficiencies," in *IEEE Int. Conf. on E-Health Networking, Application and Services (Healthcom'10)*, 2010.

[4] Harishchandra Dubey, Ashkan Aazami, Vishak Gopal, Babak Naderi, Sebastian Braun, Ross Cutler, Alex Ju, Mehdi Zohourian, Min Tang, Hannes Gamper, Mehrsa Golestaneh, and Robert Aichner, "Icassp 2023 deep noise suppression challenge," 2023.

[5] Szu-Wei Fu, Cheng Yu, Kuo-Hsuan Hung, Mirco Ravanelli, and Yu Tsao, "Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," 2021.

[6] George Close, Thomas Hain, and Stefan Goetze, "MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data," in *EUSIPCO 2022*, Belgrade, Serbia, Aug. 2022.

[7] Simone Graetzer, Michael Akeroyd, Jon P. Barker, Trevor J. Cox, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros Muñoz, "Clarity: Machine learning challenges to revolutionise hearing device processing," 2020.

[8] Simone Graetzer, Jon Barker, Trevor J. Cox, Michael Akeroyd, John F. Culling, Graham Naylor, Eszter Porter, and Rhoddy Viveros Muñoz, "Clarity-2021 Challenges: Machine Learning Challenges for Advancing Hearing Aid Processing," in *Proc. Interspeech 2021*, 2021, pp. 686–690.

[9] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, Eds. 2020, vol. 33, pp. 12449–12460, Curran Associates, Inc.

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," 2021.

[11] Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Proc. Interspeech 2021*, 2021, pp. 1194–1198.

[12] George Close, William Ravenscroft, Thomas Hain, and Stefan Goetze, "Perceive and predict: self-supervised speech representation based loss functions for speech enhancement," in *Proc. ICASSP 2023*, 2023.

[13] Bastiaan Tamm, Helena Balabin, Rik Vandenberghe, and Hugo Van hamme, "Pre-trained speech representations as feature extractors for speech quality assessment in online conferencing applications," in *Interspeech 2022*. Sep 2022, ISCA.

[14] Tsun-An Hsieh, Cheng Yu, Szu-Wei Fu, Xugang Lu, and Yu Tsao, "Improving perceptual quality by phone-fortified perceptual loss using wasserstein distance for speech enhancement," 2020.

[15] Amitay Sicherman and Yossi Adi, "Analysing discrete self supervised speech representation for spoken language modeling," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] Ankita Pasad, Bowen Shi, and Karen Livescu, "Comparative layer-wise analysis of self-supervised speech models," 2023.

[17] William Ravenscroft, Stefan Goetze, and Thomas Hain, "Att-TasNet: Attending to Encodings in Time-Domain Audio Speech Separation of Noisy, Reverberant Speech Mixtures," *Frontiers in Signal Processing*, vol. 2, 2022.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, L ukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., 2017, vol. 30.

[19] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh,

Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli, "XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale," in *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. of ACL 2019*, Minneapolis, Minnesota, June 2019, pp. 4171–4186, Association for Computational Linguistics.

[21] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[22] Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech 2021*. aug 2021, ISCA.

[23] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of mos prediction networks," 2021.

[24] Helard Becerra, Alessandro Ragano, and Andrew Hines, "Exploring the influence of fine-tuning data on wav2vec 2.0 model for blind speech quality prediction," in *Proc. Interspeech 2022*, 2022, pp. 4088–4092.

[25] Ryandhimas Edo Zezario, Fei Chen, Chiou-Shann Fuh, Hsin-Min Wang, and Yu Tsao, "MBI-Net: A Non-Intrusive Multi-Branched Speech Intelligibility Prediction Model for Hearing Aids," in *Proc. Interspeech 2022*, 2022, pp. 3944–3948.

[26] Jon Barker, Michael Akeroyd, Trevor J. Cox, John F. Culling, Jennifer Firth, Simone Graetzer, Holly Griffiths, Lara Harris, Graham Naylor, Zuzanna Podwinska, Eszter Porter, and Rhoddy Viveros Munoz, "The 1st Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *Proc. Interspeech 2022*, 2022, pp. 3508–3512.

[27] Michael Anthony Stone and Brian C. J. Moore, "Tolerable hearing aid delays. i. estimation of limits imposed by the auditory path alone using simulated hearing losses.," *Ear and hearing*, vol. 20 3, pp. 182–92, 1999.

[28] Birger Kollmeier, Anna Warzybok, Sabine Hochmuth, Melanie A. Zokoll, Verena Uslar, Thomas Brand, and Kirsten C. Wagener, "The multilingual matrix test: Principles, applications, and comparison across languages: A review," *International Journal of Audiology*, vol. 54, no. sup2, pp. 3–16, 2015.

[29] Patrick M Zurek and GA Studebaker, "Binaural advantages and directional effects in speech intelligibility," *Acoustical factors affecting hearing aid performance*, vol. 2, pp. 255–275, 1993.

[30] Alex F. McKinney and Benjamin Cauchi, "Non-intrusive binaural speech intelligibility prediction from discrete latent representations," *IEEE Signal Processing Letters*, vol. 29, pp. 987–991, 2022.

[31] George Close, Samuel Hollands, Stefan Goetze, and Thomas Hain, "Non-intrusive Speech Intelligibility Metric Prediction for Hearing Impaired Individuals," in *Proc. Interspeech 2022*, 2022, pp. 3483–3487.

[32] Zehai Tu, Ning Ma, and Jon Barker, "Unsupervised Uncertainty Measures of Automatic Speech Recognition for Non-intrusive Speech Intelligibility Prediction," in *Proc. Interspeech 2022*, 2022, pp. 3493–3497.