# WhiSQA: Non-Intrusive Speech Quality Prediction Using Whisper Encoder Features

George Close[1], Kris Hong[1], Thomas Hain[2], and Stefan Goetze[2,3] ⋆

[1]ConnexAI, Manchester, UK,
[2]School of Computer Science, The University of Sheffield, Sheffield, UK,
[3]South Westphalia University of Applied Sciences, Iserlohn, Germany
{george.close,kris.hong}@connex.ai, t.hain@sheffield.ac.uk,
goetze.stefan@fh-swf.de

**Abstract.** There has been significant research effort developing neural-network-based predictors of speech quality (SQ) in recent years. While a primary objective has been to develop non-intrusive, i.e. reference-free, metrics to assess the performance of speech enhancement (SE) systems, recent work has also investigated the direct inference of neural SQ predictors within the loss function of downstream speech tasks. To aid in the training of SQ predictors, several large datasets of audio with corresponding human labels of quality have been created. Recent work in this area has shown that speech representations derived from large unsupervised or semi-supervised foundational speech models are useful input feature representations for neural SQ prediction. In this work, a novel and robust SQ predictor is proposed based on feature representations extracted from an automatic speech recognition (ASR) model, found to be a powerful input feature for the SQ prediction task. The proposed system achieves higher correlation with human mean opinion score (MOS) ratings than recent approaches on all NISQA test sets and shows significantly better domain adaption compared to the commonly used DNSMOS metric.

## 1 Introduction

To assess the performance of speech enhancement (SE) methods, there is a continuing interest in the development of metrics to assess the speech quality (SQ) of given input audio [1–6]. Such metrics allow for the automatic assessment and comparison of SE systems without the need for expensive and time-consuming human listening tests [7–10]. Many still commonly used metrics, such as the Perceptual Evaluation of SQ (PESQ) [11] or Short-Time Objective Intelligibilty (STOI) [12] are signal-processing-based *intrusive* metrics, i.e. are designed to operate over an input of clean reference audio and a *(typically artificially) corrupted* or *enhanced* version of that same audio, the latter being the signal under

⋆ This work was supported by the Centre for Doctoral Training in Speech and Language Technologies (SLT) and their Applications funded by UK Research and Innovation [grant number EP/S023062/1].

assessment. From a neural network perspective, intrusive metrics based on traditional signal processing have two major drawbacks. Firstly, many traditional metrics have stages to their computation which cannot be easily formulated in a differentiable way, which renders them difficult to optimise towards within a loss function for neural-network-based SE systems [13]. This limitation can partially be overcome by frameworks like MetricGAN [14–18], where an SE network and a neural *metric predictor* network are adversarially trained in a generative adversarial network (GAN) setting, but such networks might be prone to artifacts not properly assessed by the metric prediction [19, 20]. The second major drawback of most traditional metrics is their intrusive nature; the reliance on the existence of the reference signal usually requires that test data be *simulated* (i.e. as artificially corrupted versions of the reference audio) rather than *real* (i.e. gathered in the 'real world' from the target domain of the system under test).
To overcome these drawbacks, several datasets and network structures [5, 21–24] for the task of neural non-intrusive SQ prediction have been proposed. Datasets for the SQ prediction task typically consist of noisy audio with associated human MOS [7] quality labels that have been collected in listening tests conducted by human listeners. Neural networks can be trained with the noisy audio as input to predict the associated MOS label.

In parallel with the SQ prediction task is the related task of non-intrusive *intelligibility* prediction [12, 25–27]. As the datasets for this task are significantly smaller, much of the focus in this topic has been on finding powerful input feature representations rather than on designing large complex network structures. In particular, features derived from large, pre-trained *foundation models* have shown to be particularly useful for the intelligibility prediction task [28–30].

In this work, feature representations generated by a foundational model are analysed as input to a neural network for the SQ prediction task. Such features, which have primarily been developed as backbone models for ASR have proved to be useful feature representations for a number of speech related tasks [31, 32]. Experiments investigating different combinations of training data corpora with different score distributions are carried out, and the effects on test time performance are analysed. Although non-intrusive SQ prediction is the main aim of this work, the identified best-performing models are analysed as intrusive and multi-headed (i.e. predicting multiple labels at once) variants. State-of-the-art performance is achieved on common testsets using the proposed model. The implementation of the best performing model as a SQ metric is provided online[1].

The remainder of this work is structured as follows: Section 2 introduces the foundation model from which input feature representations for the model structure are extracted. Section 3 formally introduces the SQ prediction task and the proposed model structure. Section 4 describes and analyses the SQ datasets which are used to train, validate and test the proposed model. Section 5 details experiments in which the optimal training data setup and task variants are investigated before Section 6 concludes the paper.

[1] available at `https://github.com/leto19/WhiSQA`

## 2 Whisper Features

Whisper is a weakly supervised Transformer-based ASR system. It has shown state-of-the-art performance on a number of monolingual ASR benchmark datasets, as well as multilingual transcription and translation tasks [33].

It consists of several sequential Transformer-based encoder blocks $\mathcal{A}_{\mathrm{E}}(\cdot)$ followed by the same number of sequential Transformer-based decoder blocks $\mathcal{A}_{\mathrm{D}}(\cdot)$. The input to the encoder $\mathcal{A}_{\mathrm{E}}(\cdot)$ is a log-Mel spectrogram representation $\mathbf{X}_{\mathrm{MEL}}$ of the input audio $x[n]$ (padded to 30 seconds in length), which is processed by a 1-dimensional convolutional neural network (CNN) layer and a Gaussian Error Linear Unit (GELU) activation function, followed by a sinusoidal positional encoding before being processed by the first encoder Transformer block. The output of each encoder layer $\ell$ is denoted as $\mathbf{X}_{\mathrm{E}}^{(\ell)}$, a two-dimensional representation of dimension 768 by 1500 [33]. The Whisper decoder $\mathcal{A}_{\mathrm{D}}(\cdot)$ takes the form of a language model; the first Transformer block of the decoder takes as input a sequence of tokens which encode the language, task, timestamp in seconds, and the previously transcribed words of the utterance. Each Transformer block in the decoder has access to the output of the encoder via a cross-attention mechanism. The final output of the decoder (not used in this work) is a prediction of the next token (i.e. the next word) in the input sequence. The $T$ dimension of the output of each Whisper decoder layer is significantly smaller than any other feature used in this work.

In this work, the `whisper-small`[2] model, trained on 680k hours of labelled speech data is used. Recent work has found that features extracted from both the encoder [29] and decoder [30] layers of Whisper are useful for capturing *intelligibility*-related information. Hence, this work analyses their capability for *quality* prediction. The encoder $\mathcal{A}_{\mathrm{E}}(\cdot)$ and decoder $\mathcal{A}_{\mathrm{D}}(\cdot)$ of this model each have 12 transformer blocks; the set of outputs of each of the constituent transformer blocks are thus denoted as $\{\mathbf{X}_{\mathrm{E}}^{(0)}, ..., \mathbf{X}_{\mathrm{E}}^{(12)}\}$ and $\{\mathbf{X}_{\mathrm{D}}^{(0)}, ..., \mathbf{X}_{\mathrm{D}}^{(11)}\}$, respectively. The weighted sum of $\{\mathbf{X}_{\mathrm{E}}^{(0)}..\mathbf{X}_{\mathrm{E}}^{(12)}\}$ is defined as

$$\bar{\mathbf{X}}_{\mathrm{E}} = \sum_{\ell=0}^{12} \alpha_{\mathrm{E}}^{(\ell)} \cdot \mathbf{X}_{\mathrm{E}}^{(\ell)}, \tag{1}$$

where $\{\alpha_{\mathrm{E}}^{(0)}, .., \alpha_{\mathrm{E}}^{(12)}\}$ are parameter weights for each layer which are learned during prediction model training.

## 3 Speech Quality (SQ) Prediction Models

For non-intrusive speech quality prediction, the neural network $\mathcal{D}(\cdot)$ takes as input a feature representation

$$\mathbf{X}_{\mathrm{F}} = \mathcal{F}(x[n]) \tag{2}$$

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

of the speech or audio signal under test $x[n]$ and returns a predicted quality label $\hat{q}$. The operator $\mathcal{F}(\cdot)$ denotes the feature extraction process; for this work $\bar{\mathbf{X}}_{\mathrm{E}}$ is taken as input features. Typically, $\mathcal{D}(\cdot)$ is trained on data consisting of tuples $(x[n], q)$ where $q$ is the *true* MOS quality label of the audio $x[n]$ obtained from signal assessment by human listeners. The loss function used to train $\mathcal{D}(\cdot)$ is often a simple Mean Squared Error (MSE) between the model output i.e the *predicted* score $\hat{q} = \mathcal{D}(\mathbf{X}_{\mathrm{F}})$ and the *true* quality label $q$:

$$L_{\mathcal{D}} = (\mathcal{D}(\mathbf{X}_{\mathrm{F}}) - q)^2. \tag{3}$$

Note that while MOS labels are typically expressed in the range 1 to 5, higher being better, for the ease of training of neural SQ predictors, $q$ is typically normalised to a range between 0.2 and 1, which enables a sigmoid activation function on the final neural network layer to project to this label range [34]. SQ prediction models can be broadly classified into two types; *single-headed* models which predict only the MOS label and *multi-headed* models which predict MOS alongside some other label(s) of the input audio, e.g. Noisiness, Coloration, Discontinuity, etc.

The structure of the proposed SQ prediction models $\mathcal{D}(\cdot)$ is based on [35], and is shown in Figure 1. The model $\mathcal{D}_1(\cdot)$ (denoted as 'Single Head Prediction Model' in Figure 1) consists of 4 transformer layers, followed by an attention pooling mechanism with a sigmoid activation function, which returns the predicted MOS score $\hat{q}$ normalised between 0.2 and 1. The input dimension (and thus the parameter count) of the transformer stage depends on the feature dimension $F$ of the input feature, while the output dimension is fixed at 256. The attention pooling mechanism consists of two sequential linear layers, with a softmax function applied at the output and is multiplied by the output of the Transformer block. The result of this multiplication is further fed into a final linear layer with a sigmoid activation to a single output neuron. This single output neuron represents the predicted MOS label $\hat{q}$ of the input audio. A variant of this base model (denoted as 'Multi Head Prediction Model' in Figure 1) which incorporates multiple prediction 'heads' i.e. the three Linear layer structure is also proposed for multi-dimension speech quality prediction.

## 4 Datasets for Speech Quality Prediction

Datasets containing mean opinion score (MOS) scores $q$ obtained from listening test with humans for signals under test $x[n]$ have only been created during the last few years in quantities which allow training recent data-driven methods. Several SQ datasets are now available and briefly analysed in the following. It is important to consider several datasets to ensure that the SQ predictor has been exposed to a large variety of audio conditions during its training. For some datasets and subsets within datasets, further information is available such as a clean reference signal $s[n]$, the standard deviation of the MOS score, the raw scores assigned by each human evaluator or the number of human assessors.

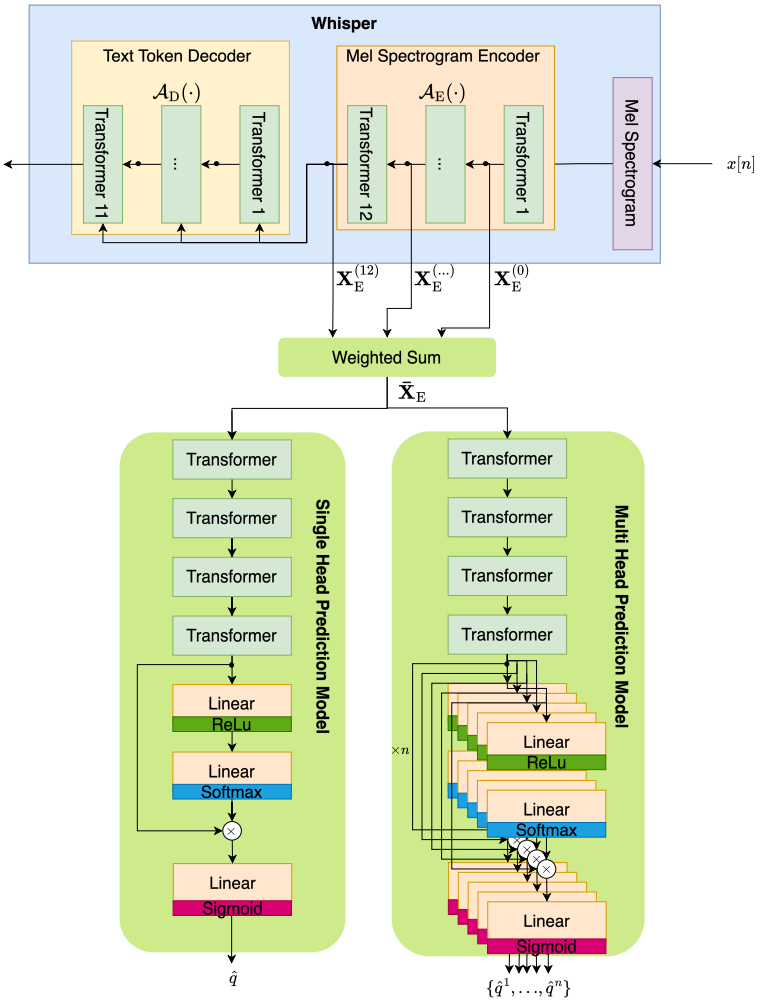


**Fig. 1.** Network structure of the proposed WhiSQA SQ predictor with Whisper Encoder feature extraction. Note that the 'Weighted Sum' block contains model parameters, i.e. layer weights $\{\alpha^{(0)}, .., \alpha^{(12)}\}$ from (1) which are updated during prediction model training.

### 4.1 NISQA Dataset

The Non-Intrusive SQ Assessment (NISQA) [5] dataset is an SQ assessment dataset, comprising of pre-defined train, validation and test sets. Each of these

are further divided into subsets, characterised by if the nature of the distortion in the speech signal is artificially simulated or occurring 'in the wild' as a real distortion. In addition to MOS scores of overall audio quality, the NISQA dataset also provides labels for other speech 'dimensions' [36] namely Noisiness, Coloration, Discontinuity and Loudness. It has three defined testsets, denoted as FOR, LIVETALK and P501. With the exception of the LIVETALK testset, clean reference signals $x[n]$ are available. The baseline NISQA model has single and multi-headed variants.

### 4.2 Tencent Dataset

The Tencent audio SQ dataset was released as part of the ConferencingSpeech 2022 challenge [23]. It consists of two artificially simulated training subsets, one with artificial reverberation added and one without.

### 4.3 Indiana University Bloomington (IUB) Dataset

The Indiana University Bloomington (IUB) [24] SQ dataset consists of two subsets. The first uses distorted audio sourced from the COnversational Speech In Noisy Environments (COSINE) [37] dataset, real multi party conversations captured using multi-channel wearable microphones recorded in noisy everyday environments. The second subset uses audio from the Voices Obscured in Complex Environmental Settings (VOiCES) [38] corpus where speech and noise were played aloud and recorded in two rooms of different sizes.
Unlike the other datasets used in this work, the MOS scores for this dataset were gathered using a Multiple Stimuli with Hidden Reference and Anchor (MUSHRA) [8] protocol, which is then transformed to a MOS scale between 0 and 10, rather than the 1 to 5 scale commonly used. The 1 - 5 MOS label is obtained via a fitting operation over the gathered MUSHRA ratings.

### 4.4 Public Switched Telephone Network (PSTN) Dataset

The Public Switched Telephone Network (PSTN) SQ dataset [39] consists of simulated 'real' phone calls, some with simulated background noise added to the transmitted signal. It follows a similar design to that of NISQA, but is significantly larger.

### 4.5 Overall MOS Distribution

To compare the available datasets and analyse prediction results later in this paper, the distributions of MOS scores in the training and validation subsets of the datasets (normalised between 0.2 and 1) are shown in Figure 2. The mean MOS value across the datasets is similar, at approximately 0.65. However, the datasets differ significantly in the shape of their distributions. Both NISQA and Tencent show a roughly uniform distribution of scores from 0.2 to 1, with

the ‘tail’ at the lower end of the Tencent distribution showing that that dataset contains a larger numbers of low scores. Conversely, the tapering in at the highest end in both NISQA and Tencent indicate that these datasets contain relatively few instances of very highly rated audio.

In contrast, the distribution of the PSTN dataset scores is generally normal, tailing off at the low and high end. Slightly more scores are above 0.5 than below, indicating that the audio in this dataset is generally of high quality.

The distribution of the MOS score in the IUB dataset is most different from the others, with very few points falling a the highest and lowest values. Further, it is significantly more erratic than the other datasets, with an extreme dearth in scores valued around 0.65. This can possibly be explained by the non-standard method that the MOS scores were gathered, as well as the differing range of the scores before normalisation.

The combined distribution across all the datasets is shown in purple at the top of Figure 2. It displays a similar normal-like distribution to that of the PSTN dataset, likely due to that dataset contributing roughly half of all samples. There are slightly more samples of low quality compared to high quality.

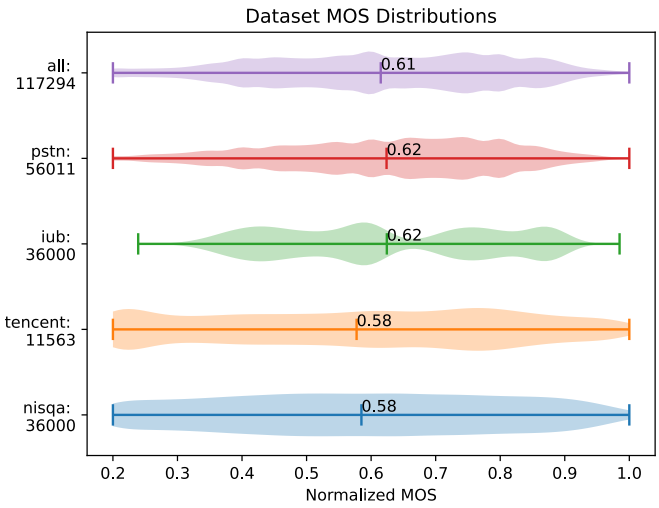


**Fig. 2.** Normalised MOS score distribution across SQ datasets with lines indicating minimum, mean and maximum MOS in each dataset.Numbers on y axis denote number of data points in each set.

## 5 Experiments

This experiment aims to find which training datasets have the greatest effect on test performance of the proposed SQ prediction networks, as well as enabling a fair comparison with other recently proposed SQ prediction systems.

### 5.1 Experiment Setup

All models are tested on each of the three NISQA test sets, i.e. FOR, LIVETALK and P501. Following [5], a training strategy where training stops only if the validation performance does not improve after 20 epochs is employed. The bias-aware loss function, scaling the contribution of the training samples in the loss computation based on the relative size of the training set/subset, as proposed in [5] is also used here. The Adam [40] optimiser is used with an initial learning rate of 0.00001, which is reduced by a factor of 0.1 if the validation loss does not improve after 15 epochs. All models are at first trained over a warmup epoch, where the learning rate increases up to the initial learning rate after each model update. A batch size $B$ of 128 is used. The best-performing epoch on the validation set in terms of validation loss is loaded at test time. Datasets other than NISQA do not have defined validation sets; for these, 10% of the training sets are partitioned for validation, following [41]. All possible permutations of the evaluated datasets are used. The proposed Multi Head model (right in Figure 1) is trained on the NISQA testset to predict the MOS as well as the Noisiness, Coloration, Discontinuity and Loudness labels.

Models are evaluated using Spearman correlation $r$ and MSE $e$, computed versus the true MOS value for each testset element.

### 5.2 Results

Table 1 shows the results for the training data ablation experiment for the three NISQA test sets. The overall (on average) best-performing combination of training datasets is "NISQA, Tencent and PSTN". By far the lowest-performing model is that trained solely on IUB; further, also any given combination of training datasets including IUB performs worse on average than that combination without IUB. As noted earlier in Section 4, this is likely due to the significantly different distribution of the MOS labels in this dataset relative to the others. The overall size of the training set has a smaller effect on performance - the inclusion of data more similar to the test sets (i.e. the NISQA training data) results in better performance. This can perhaps be attributed to the bias-aware loss function used, which attempts to control for the imbalance in size between the component datasets. It can be noted, that including the Chinese-language Tencent dataset in training generally improves performance on the German-language LIVETALK testset; this can be attributed to these models being better able to generalise to languages other than English.

Table 2 shows a comparison the proposed system with three state-of-the-art neural SQ predictor systems [5, 35, 41]. Results for the proposed system trained

**Table 1.** Training Data Ablation Study for best performing proposed single-head model. **Best** and second best shown in **bold** and underlined, respectively.

| **Training Data** | | | | | **FOR** | | **LIVETALK** | | **P501** | | **AVERAGE** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NISQA | Tencent | IUB | PSTN | **Train Points** | r ↑ | $e$↓ | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ |
| | ✓ | | | 9250 | 0.82 | 0.50 | 0.83 | 0.56 | 0.83 | 0.56 | 0.83 | 0.54 |
| ✓ | | | | 11020 | 0.92 | 0.35 | 0.82 | 0.54 | 0.93 | 0.37 | 0.89 | 0.44 |
| ✓ | ✓ | | | 20270 | 0.93 | 0.32 | 0.87 | 0.46 | **0.93** | 0.37 | 0.91 | 0.38 |
| | | ✓ | | 28800 | 0.27 | 0.84 | 0.42 | 0.85 | 0.41 | 0.92 | 0.37 | 0.87 |
| | ✓ | ✓ | | 38050 | 0.85 | 0.46 | 0.76 | 0.62 | 0.79 | 0.62 | 0.80 | 0.57 |
| ✓ | | ✓ | | 39820 | 0.93 | 0.32 | 0.83 | 0.52 | 0.92 | 0.40 | 0.89 | 0.41 |
| | | ✓ | | 44809 | 0.92 | 0.34 | 0.77 | 0.60 | 0.88 | 0.48 | 0.86 | 0.47 |
| ✓ | ✓ | ✓ | | 49070 | 0.93 | 0.32 | 0.86 | 0.48 | 0.91 | 0.42 | 0.90 | 0.41 |
| | ✓ | | ✓ | 54059 | 0.91 | 0.36 | 0.85 | **0.39** | 0.90 | 0.45 | 0.89 | 0.40 |
| ✓ | | | ✓ | 55829 | **0.94** | **0.29** | 0.83 | 0.51 | **0.94** | **0.35** | 0.90 | 0.38 |
| ✓ | ✓ | | ✓ | 65079 | 0.94 | 0.30 | **0.88** | 0.45 | 0.93 | 0.38 | **0.92** | **0.38** |
| | | ✓ | ✓ | 73609 | 0.89 | 0.40 | 0.72 | 0.65 | 0.76 | 0.39 | 0.79 | 0.48 |
| | ✓ | ✓ | ✓ | 82859 | 0.92 | 0.34 | 0.81 | 0.55 | 0.83 | 0.56 | 0.85 | 0.48 |
| ✓ | | ✓ | ✓ | 84629 | 0.94 | 0.30 | 0.87 | 0.46 | 0.93 | 0.39 | 0.91 | 0.38 |
| ✓ | ✓ | ✓ | ✓ | 93879 | 0.93 | 0.31 | 0.88 | 0.45 | 0.91 | 0.42 | 0.91 | 0.39 |

on the same combination of data are shown for a fair comparison. For all training data combinations, the proposed WhiSQA system outperforms the SOTA system.

**Table 2.** Comparison of WhisSQA with SOTA systems. **Best** and second best shown in **bold** and underlined, respectively.

| | | **FOR** | | **LIVETALK** | | **P501** | | **AVERAGE** | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Training Data** | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ |
| *NISQA Single Head* [5] | *NISQA* | *0.88* | *0.40* | *0.70* | *0.67* | *0.89* | *0.46* | *0.82* | *0.51* |
| Proposed WhiSQA | NISQA | 0.92 | 0.35 | 0.82 | 0.54 | 0.93 | 0.37 | 0.89 | 0.44 |
| MSQAT [41] | NISQA + Tencent + PSTN | 0.90 | 0.39 | 0.85 | 0.51 | 0.92 | 0.42 | 0.89 | 0.44 |
| Proposed WhiSQA | NISQA+ Tencent + PSTN | **0.94** | **0.30** | **0.88** | 0.45 | **0.93** | **0.38** | **0.92** | **0.38** |
| XLS-R SQA [35] | Tencent + PSTN | 0.90 | 0.38 | 0.83 | 0.52 | 0.89 | 0.46 | 0.82 | 0.51 |
| Proposed WhiSQA | Tencent + PSTN | 0.91 | 0.36 | 0.85 | **0.39** | 0.90 | 0.45 | 0.89 | 0.40 |

Table 3 compares the performance of the baseline NISQA model and the proposed model for multi-head / multi-label prediction. In both cases, the proposed system outperforms the NISQA baselines. For both systems, tasking the model with additionally predicting the other speech dimensions from the input audio slightly degrades the performance of the main task, i.e. quality MOS prediction.

Figure 3 shows a Spearman correlation matrix for the CHiME7-unsupervised domain adaptation speech enhancement (UDASE) listening test [42]. This listening test was designed to assess the enhancement performance of the entries to the UDASE challenge. Figure 3 compares human MOS (`SIG`, `BAK` and `OVRL`) with those predicted by the DNSMOS [22] metric (`DNSMOS_SIG`, `DNSMOS_BAK`, `DNSMOS_OVRL`) and by the proposed single head WhiSQA model. *The WhiSQA*

**Table 3.** MOS prediction results for Multi Headed (MH) $\mathcal{D}_1$ Models versus Single Head (SH) Prediction. **Best** shown in **Bold**.

| | **FOR** | | **LIVETALK** | | **P501** | | **AVERAGE** | |
|---|---|---|---|---|---|---|---|---|
| **Model** | r ↑ | $e$ ↓ | r ↑ | $e$ ↓ | **r** ↑ | $e$ ↓ | **r** ↑ | $e$ ↓ |
| *NISQA SH* | *0.88* | *0.40* | *0.70* | *0.67* | *0.89* | *0.46* | *0.82* | *0.51* |
| *NISQA MH* | *0.87* | *0.43* | *0.65* | *0.72* | *0.89* | *0.46* | *0.80* | *0.54* |
| WhiSQA SH | **0.92** | **0.35** | **0.82** | **0.54** | **0.93** | **0.37** | **0.89** | **0.42** |
| WhiSQA MH | 0.91 | 0.36 | 0.69 | 0.58 | 0.92 | 0.41 | 0.84 | 0.45 |

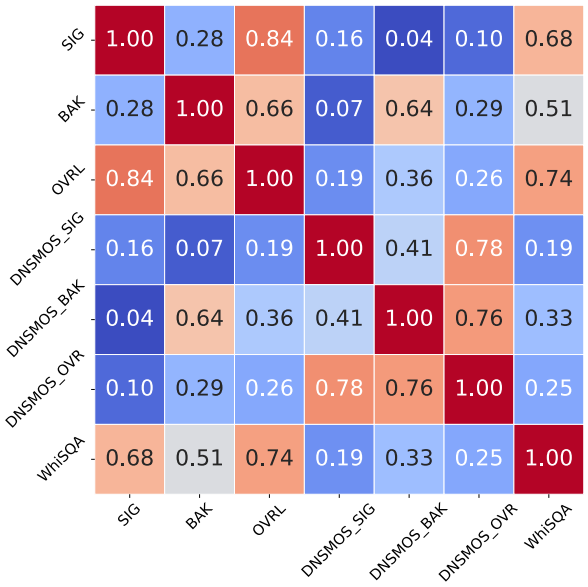


**Fig. 3.** Spearman Correlation Matrix for CHiME7-UDASE listening test data for DNSMOS and WhiSQA.

*score correlates significantly more strongly* with the true `SIG` and `OVRL` scores compared to the corresponding DNSMOS metric value, while showing similar correlation to the true `BAK` score that the `DNSMOS_BAK` metric does.

## 6 Conclusion and Future Work

This work introduces WhiSQA, a new SOTA system for speech quality prediction, as single- and multi-headed variants. Alayses for different datasets show improved performance over several baselines. Future work will explore further refinement of the system in the form of adaption to online 'in the wild' data as well as the applications of the Whisper encoder feature to other audio classification and evaluation tasks.

## References


1. S. Möller, W.-Y. Chan, N. Côté, T. H. Falk, A. Raake, and M. Wältermann, "Speech quality estimation: Models and trends," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 18–28, 2011.
2. T. Rohdenburg, S. Goetze, V. Hohmann, K.-D. Kammeyer, and B. Kollmeier, "Objective Perceptual Quality Assessment for Self-Steering Binaural Hearing Aid Microphone Arrays," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2008.
3. P. Loizou, *Speech Enhancement: Theory and Practice, Second Edition.* CRC Press, 2013.
4. S. Goetze, E. Albertin, J. Rennies, E. Habets, and K.-D. Kammeyer, "Speech Quality Assessment for Listening-Room Compensation," *J. Audio Eng. Soc.*, vol. 62, no. 6, 2014.
5. G. Mittag, B. Naderi, A. Chehadi, and S. Möller, "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech 2021*, Aug. 2021.
6. A. Kumar, K. Tan, Z. Ni, P. Manocha, X. Zhang, E. Henderson, and B. Xu, "Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio," 2023.
7. International Telecommunication Union, "Recommendation ITU-T P.800.2 Mean opinion score interpretation and reporting," ITU, ITU-T Recommendation, Jul. 2016.
8. ——, "Recommendation ITU-R BS.1534-3 Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," ITU, ITU-R Recommendation, Oct. 2015.
9. S. Goetze, A. Warzybok, I. Kodrasi, J. O. Jungmann, B. Cauchi, J. Rennies, E. A. P. Habets, A. Mertins, T. Gerkmann, S. Doclo, and B. Kollmeier, "A study on speech quality and speech intelligibility measures for quality assessment of single-channel dereverberation algorithms," in *Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2014.
10. A. Avila, B. Cauchi, S. Goetze, S. Doclo, and T. Falk, "Performance comparison of intrusive and non-intrusive instrumental quality measures for enhanced speech," in *Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, 2016.
11. A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE ICASSP*, 2001.
12. C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP 2010*.
13. J. Martín-Doñas, A. Gomez, J. Gonzalez Lopez, and A. Peinado, "A deep learning loss function based on the perceptual evaluation of the speech quality," *IEEE Signal Processing Letters*, vol. PP, pp. 1–1, 09 2018.
14. S.-W. Fu, C. Yu, T.-A. Hsieh, P. Plantinga, M. Ravanelli, X. Lu, and Y. Tsao, "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 201–205.
15. G. Close, T. Hain, and S. Goetze, "MetricGAN+/-: Increasing Robustness of Noise Reduction on Unseen Data," in *EUSIPCO 2022*, Belgrade, Serbia, Aug. 2022.
16. R. Cao, S. Abdulatif, and B. Yang, "CMGAN: Conformer-based Metric GAN for Speech Enhancement," in *Proc. Interspeech 2022*, 2022, pp. 936–940.

17. G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Multi-CMGAN+/+: Leveraging Multi-Objective Speech Quality Metric Prediction for Speech Enhancement," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'24)*, 2024.
18. Y. Mai and S. Goetze, "MetricGAN+KAN: Kolmogorov-Arnold Networks in Metric-Driven Speech Enhancement Systems," in *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP'25)*, 2025.
19. D. de Oliveira, S. Welker, J. Richter, and T. Gerkmann, "The pesqetarian: On the relevance of goodhart's law for speech enhancement," in *Interspeech 2024*, 2024, pp. 3854–3858.
20. G. Close, T. Hain, and S. Goetze, "Identifying hallucination in perceptually motivated speech enhancement networks," in *32nd European Signal Processing Conference (EUSIPCO24)*, Lyon, France, Aug. 2024.
21. B. Cauchi, K. Siedenburg, J. F. Santos, T. H. Falk, S. Doclo, and S. Goetze, "Non-Intrusive Speech Quality Prediction Using Modulation Energies and LSTM-Network," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 7, Jul. 2019.
22. C. K. A. Reddy, V. Gopal, and R. Cutler, "Dnsmos p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," 2022.
23. G. Yi, W. Xiao, Y. Xiao, B. Naderi, S. Möller, W. Wardah, G. Mittag, R. Cutler, Z. Zhang, D. S. Williamson, F. Chen, F. Yang, and S. Shang, "ConferencingSpeech 2022 Challenge: Non-intrusive Objective Speech Quality Assessment (NISQA) Challenge for Online Conferencing Applications," in *Proc. Interspeech 2022*, 2022, pp. 3308–3312.
24. X. Dong and D. S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," in *Interspeech*, 2020, pp. 4631–4635.
25. A. Warzybok, I. Kodrasi, J. Jungmann, E. Habets, T. Gerkmann, A. Mertins, S. Doclo, B. Kollmeier, and S. Goetze, "Subjective speech quality and speech intelligibility evaluation of single-channel dereverberation algorithms," in *Proc. Int. Workshop on Acoustic Signal Enhancement (IWAENC 2014)*, Sep. 2014.
26. M. Karbasi and D. Kolossa, "Asr-based speech intelligibility prediction: A review," *Hearing Research*, vol. 426, 2022.
27. J. Barker, M. Akeroyd, W. Bailey, T. J. Cox, J. F. Culling, J. Firth, S. Graetzer, and G. Naylor, "The 2nd Clarity Prediction Challenge: A machine learning challenge for hearing aid intelligibility prediction," in *ICASSP*, 2024.
28. G. Close, T. Hain, and S. Goetze, "Non intrusive intelligibility predictor for hearing impaired individuals using self supervised speech representations," in *Proc. Workshop on Speech Foundation Models and their Performance Benchmarks (SPARKS), ASRU sattelite workshop*, 2023.
29. Santiago Cuervo, Ricard Marxer, "Temporal-hierarchical features from noise-robust speech foundation models for non-intrusive intelligibility prediction," in *Clarity Workshop 2022*, 2022.
30. R. Mogridge, G. Close, R. Sutherland, T. Hain, J. Barker, S. Goetze, and A. Ragni, "Non-intrusive speech intelligibility prediction for hearing-impaired users using intermediate asr features and human memory models," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP'24)*, 2024.
31. A. Pasad, B. Shi, and K. Livescu, "Comparative Layer-Wise Analysis of Self-Supervised Speech Models," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

32. G. Close, W. Ravenscroft, T. Hain, and S. Goetze, "Perceive and predict: self-supervised speech representation based loss functions for speech enhancement," in *Proc. ICASSP 2023*, 2023.
33. A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision," 2022.
34. S.-W. Fu, C. Yu, K.-H. Hung, M. Ravanelli, and Y. Tsao, "Metricgan-u: Unsupervised speech enhancement/ dereverberation based only on noisy/ reverberated speech," 2021.
35. B. Tamm, R. Vandenberghe, and H. Van hamme, "Analysis of xls-r for speech quality assessment," in *Proc. WASPAA 2023*, 10 2023, pp. 1–5.
36. M. Wältermann, "Dimension-based quality modeling of transmitted speech," 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:63687570
37. A. Hashmi, "Perceptual evaluation of speech quality for inexpensive recording equipment," *Acoustics*, vol. 3, no. 1, pp. 200–211, 2021. [Online]. Available: https://www.mdpi.com/2624-599X/3/1/14
38. C. Richey, M. Barrios, Z. Armstrong, C. Bartels, H. Franco, M. Graciarena, A. Lawson, M. Nandwana, A. Stauffer, J. Hout, P. Gamble, J. Hetherly, C. Stephenson, and K. Ni, "Voices obscured in complex environmental settings (voices) corpus," 04 2018.
39. G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande, and R. Aichner, "DNN No-Reference PSTN Speech Quality Prediction," in *Proc. Interspeech 2020*, 2020.
40. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, 2014.
41. K. Shen, D. Yan, and L. Dong, "Msqat: A multi-dimension non-intrusive speech quality assessment transformer utilizing self-supervised representations," *Applied Acoustics*, vol. 212, p. 109584, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0003682X23003821
42. S. Leglaive, L. Borne, E. Tzinis, M. Sadeghi, M. Fraticelli, S. Wisdom, M. Pariente, D. Pressnitzer, and J. R. Hershey, "The CHiME-7 UDASE task: Unsupervised domain adaptation for conversational speech enhancement," 2023.