Utilizing Self-supervised Representations for MOS Prediction

Wei-Cheng Tseng*, Chien-yu Huang*, Wei-Tsung Kao, Yist Y. Lin, Hung-yi Lee

College of Electrical Engineering and Computer Science, National Taiwan University, Taiwan
{r09942094, r08921062, r09942067, r08922048, hungyilee}@ntu.edu.tw

Abstract

Speech quality assessment has been a critical issue in speech processing for decades. Existing automatic evaluations usually require clean references or parallel ground truth data, which is infeasible when the amount of data soars. Subjective tests, on the other hand, do not need any additional clean or parallel data and correlates better to human perception. However, such a test is expensive and time-consuming because crowd work is necessary. It thus becomes highly desired to develop an automatic evaluation approach that correlates well with human perception while not requiring ground truth data. In this paper, we use self-supervised pre-trained models for MOS prediction. We show their representations can distinguish between clean and noisy audios. Then, we fine-tune these pre-trained models followed by simple linear layers in an end-to-end manner. The experiment results showed that our framework outperforms the two previous state-of-the-art models by a significant improvement on Voice Conversion Challenge 2018 and achieves comparable results on VCC 2016. In the challenge, participants submit the audios generated by their voice conversion systems. Then, to evaluate these systems, subjects were asked to score the synthesized audios, enhancing the generalizability to unseen systems. However, these previous works only rely on scarce human-labeled data, which could limit the performance.

Self-supervised learning enables the model to learn meaningful representations from large-scale unlabeled data. CPC [12] predicts the future representation in a contrastive learning manner. Wav2vec2.0 [13] learns the representation by masking the latent space. APC [14] autoregressively generates the input feature in the next time step. TERA [15] reconstructs the original complete input from a corrupted one. Self-supervised learning has achieved remarkable performance in several tasks including automatic speech recognition [16], speaker verification [17], voice conversion [18], and so on. However, their potential for speech quality assessment has not been explored yet. This paper presents the first use of self-supervised representations pre-trained on large-scale unlabeled data for automatic MOS prediction. We first showed that self-supervised pre-trained models can cluster audios in various types, and then proposed a new framework in which the model is fine-tuned with some simple yet effective modules. Our framework surpasses the two previous state-of-the-art models on Voice Conversion Challenge (VCC) 2018 while being at least comparable on VCC 2016 [20]. We also conducted an ablation study to see how each module affects performance.

1. Introduction

Speech quality assessment is to evaluate the quality of audios, and it has been an important part of speech processing to measure the performance of a system for decades. Several assessment metrics were used to evaluate different aspects of audio quality. In speech enhancement, perceptual evaluation of speech quality [1] (PESQ) and short-time objective intelligibility [2] (STOI) are widely used to measure the noise reduction. In speech synthesis, mel cepstral distance [3] (MCD) is used to measure the distortion of synthesized speech. These metrics require reference audio, which implies the need for clean, parallel data.

When reference audios are not available, the most common way to evaluate speech quality is the mean opinion score (MOS). Each subject is asked to give the audios opinion scores, integers ranging from 1 to 5, and the MOS is the mean score of several subjects. A higher score indicates better quality and vice versa. MOS correlates better to human perception compared to automatic assessments above. However, such measurement usually requires a great number of humans to involve, making it time-consuming. Several machine-learning-based models [4, 5, 6, 7, 8] were thus proposed for automatic speech quality assessment. Lo et al. [9] verified the predictability of MOS with statistical method [10] and proposed MOSNet for MOS prediction. MBNet [11] further utilizes the judge identities in the training dataset for modeling the bias of subjects. The bias modeling improves the correlation between the predicted and real scores, enhancing the generalizability to unseen systems. However, these previous works only rely on scarce human-labeled data, which could limit the performance.

Self-supervised learning enables the model to learn meaningful representations from large-scale unlabeled data. CPC [12] predicts the future representation in a contrastive learning manner. Wav2vec2.0 [13] learns the representation by masking the latent space. APC [14] autoregressively generates the input feature in the next time step. TERA [15] reconstructs the original complete input from a corrupted one. Self-supervised learning has achieved remarkable performance in several tasks including automatic speech recognition [16], speaker verification [17], voice conversion [18], and so on. However, their potential for speech quality assessment has not been explored yet. This paper presents the first use of self-supervised representations pre-trained on large-scale unlabeled data for automatic MOS prediction. We first showed that self-supervised pre-trained models can cluster audios in various types, and then proposed a new framework in which the model is fine-tuned with some simple yet effective modules. Our framework surpasses the two previous state-of-the-art models on Voice Conversion Challenge (VCC) 2018 while being at least comparable on VCC 2016 [20]. We also conducted an ablation study to see how each module affects performance.

2. Preliminary Analysis

Intuitively, if the representations from a self-supervised pre-trained model are more discriminative between the high- and low-quality audios than conventional features are, they are more likely beneficial to MOS prediction. We thus first explore to what extent these models can evaluate audio quality without fine-tuning. Two datasets are involved in this paper: VCC 2016 and VCC 2018. In the challenge, participants submit the audios generated by their voice conversion systems. Then, to evaluate these systems, subjects were asked to score the synthesized audios from different systems based on the audio quality. We refer to the mean scores from all subjects of an utterance as utterance-level score, and the average of utterance-level scores of a system is called system-level score. In VCC 2016 dataset, only the system-level MOS is available. VCC 2018 dataset, in contrast, provides detailed scoring information, where the score of each subject for each utterance is available. Following previous work [11], we divided VCC 2018 into training, validation, and testing set with 13580, 3000, and 4000 utterances respectively. For VCC 2016 dataset, since only the system-level MOS score is available, we only used it in the testing stage.

We sampled synthetic data from the best and the worst systems as well as ground truth in VCC 2018. Additionally, we also collected real-world noises from WHAM! [21] and music from MUSAN [22]. Then we extracted their representations with wav2vec 2.0 base [13] and projected them to 2D space us-
Table 1: The linear correlation coefficients between estimated scores and ground truths. Estimated scores were obtained using the linear transform found with CCA on VCC 2018 training set.

<table>
<thead>
<tr>
<th>Model</th>
<th>Utterance-level</th>
<th>System-level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VCC 2018</td>
<td>VCC 2016</td>
</tr>
<tr>
<td>wav2vec 2.0</td>
<td>0.734</td>
<td>0.966</td>
</tr>
<tr>
<td>TERA</td>
<td>0.727</td>
<td>0.943</td>
</tr>
<tr>
<td>APC</td>
<td>0.678</td>
<td>0.891</td>
</tr>
<tr>
<td>CPC</td>
<td>0.699</td>
<td>0.890</td>
</tr>
<tr>
<td>MFCC</td>
<td>0.183</td>
<td>0.196</td>
</tr>
<tr>
<td>Mel-spec.</td>
<td>0.215</td>
<td>0.487</td>
</tr>
</tbody>
</table>

3. Proposed Framework

In MOS test, an audio \( x \) is evaluated by \( K \) subjects giving a sequence of scores \( y_1, y_2, \ldots, y_K \). These scores are averaged to obtain a mean score, denoted as \( y \). Our goal is to predict \( y \) from \( x \) with a pre-trained model. Figure 2 explains our framework, which consists of segmental embeddings [25, 26, 27, 28] of pre-trained model (Sec. 3.1), attention pooling (Sec. 3.2), bias network (Sec. 3.3), and range clipping (Sec. 3.4).

![Figure 1: A visualization of representations of different kinds of audios extracted by wav2vec 2.0.](image)

![Figure 2: The proposed framework for MOS prediction.](image)
level representation from a sequence of frame-level representations. For each segment, the attention module first encodes frame-level representations into queries,

\[ Q_i = \text{softmax}(W H_i), \]

where \(H_i \in \mathbb{R}^{d \times M}, W \in \mathbb{R}^{d \times q}, Q_i \in \mathbb{R}^{1 \times M}.\) It then gives the segment-level representation by

\[ h_i = H_i Q_i^T. \]

We then obtain the segment-level score \(w\) by \(g(\cdot): \mathbb{R}^d \rightarrow \mathbb{R},

\[ w_i = g(h_i). \]

Finally, the utterance-level score \(\hat{y}\) is determined by taking the average of segment-level scores,

\[ \hat{y} = \frac{1}{N} \sum_{i=1}^{N} w_i. \]

### 3.3. Bias network

Utilizing individual judge’s scores in training has been shown to improve the performance [11]. Here we also adopt a bias network to make use of them. A judge ID \(k\) is first transformed into an embedding \(e_k\) by a trainable embedding table. It is added to frame-level representations \(h_{ij}\) to form judge-biased representations \(\hat{h}_{ij}\). Then, we obtain the bias of a judge \(\delta_k\) from \(\hat{h}_{ij}\) by a bias network with similar process described in (5) - (8). We can determine the score given by judge \(k\) by

\[ \hat{y}_k = \hat{y} + \delta_k. \]

In the test stage, we use only the mean score and discarded the bias network as was done in MBNet for better performance.

### 3.4. Range clipping

While the model is trained to fit the human scoring distributions, the existence of outliers is inevitable. To reduce such distortion, we can apply hyperbolic tangent function to ensure a fixed range of segment-level score. In this way, (7) above becomes:

\[ \hat{w}_i = 2 \tanh g(h_i) + 3. \]

The range of \(\hat{w}_i\) is constrained between 1 and 5, and therefore the model is always guaranteed to give reasonable scores. This clipping is only applied in the mean score prediction but not the bias ones because the bias is not necessarily a positive number.

### 3.5. Training objectives

During the training, we minimize utterance- and segment-level mean squared error (MSE) for mean score. As for the biased score for each judge, we minimize the MSE on utterance-level only. The training objective is thus expressed as

\[ \mathcal{L} = (\hat{y} - y)^2 + \frac{\alpha}{N} \sum_{i=1}^{N} (w_i - y)^2 + \frac{\beta}{K} \sum_{j=1}^{K} (\hat{y}_j - y_j)^2, \]

where \(\alpha\) and \(\beta\) are hyperparameters balancing the losses.

Although Lo et al. [9] showed frame-level MSE loss mitigates the variance of predicted scores within an utterance, here we resort to the segment-level MSE instead. We believe that it is more reasonable to score segments of utterance but not every frame, as the former one correlates better to human perception.

### 4. Experimental Settings

Following Section 2, we used four self-supervised pre-trained models in the experiments: wav2vec 2.0, CPC, TERA, and APC. These models were pre-trained with large-scale unlabeled data such as LibriSpeech [31] and Libri-Light [32] and can be accessed via publicly-available toolkits S3PRL [33].

For segmental embeddings, the duration \((\ell)\) and stride \((S)\) are 1.0 and 0.5 seconds respectively. The raw representations \(f(x, y)\) from the pre-trained model were projected to 256-dim space for each frame. Then, the projection \(g(\cdot)\) from segment-level representations to scores was simply a linear layer. For the bias network, we adopted similar architecture.

Two baseline models were included in the experiments for comparison: MOSNet and MBNet. For MOSNet, we trained the model with the official implementation². As for MBNet, we resorted to an unofficial implementation³ because the official one is not available. These models were trained with hyperparameters from the original papers.

We fine-tuned pre-trained models for 20k steps in three learning rates: 1e-4, 5e-5 and 1e-5 with a warm-up in the first 500 steps and linear decay in the remaining steps. For baseline models, we followed the settings in the original papers. We performed testing on the validation set every 250 steps, and the checkpoint with the best system-level performance (in terms of Spearman’s rank correlation coefficient) was then used for evaluation on the testing set.

### 5. Results

#### 5.1. Quantitative results

We evaluated the performance on VCC 2016 and VCC 2018 test set in three metrics: mean squared error (MSE), linear correlation coefficient [34] (LCC), and Spearman’s rank correlation coefficient [35] (SRCC). MSE measures the absolute difference between predicted scores and ground truths, while the latter two tell how correlated the predictions and ground truths are.

Table 2 lists the performances of the proposed framework using different pre-trained models along with two baselines. We compared our framework with the baselines in two scenarios: with and without bias network. In scenario, our framework is trained with the help of a bias network and compared to MBNet, which also utilizes the individual judge scores. In without scenario, we train the model without bias network \((\beta = 0\) in (11)) and take MOSNet as the baseline, which simply uses the mean score of each utterance. From Table 2a, we see in utterance-level MOS, all pre-trained models outperformed the baseline models in both LCC and SRCC significantly. In terms of MSE, only wav2vec 2.0 in without scenario slightly fell behind MOSNet. We can also observe that the models in with scenario outperformed those in without scenario pairwisely, though the bias network was not used in the testing.

On the other hand, Table 2b presents the system-level performance for different models on VCC 2016 and VCC 2018 testing set. We see that the LCC and SRCC of all models including baselines are higher than those in utterance-level, which means system-level scores are more predictable than those in utterance-level. In without scenario, all the pre-trained models surpassed MOSNet in terms of LCC and SRCC on VCC 2016 and 2018, while wav2vec 2.0 and TERA fell behind MOSNet in MSE. However, we believe that LCC and SRCC better evaluate

---

2https://github.com/lochenchou/MOSNet

3https://github.com/skyl1456723/Pytorch-MBNet

2783
Table 2: The performances of our frameworks (trained with/without bias network) and baselines on VCC 2016 and VCC 2018 test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>VCC 2016</th>
<th></th>
<th></th>
<th>VCC 2018</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>LCC</td>
<td>SRCC</td>
<td>MSE</td>
<td>LCC</td>
<td>SRCC</td>
</tr>
<tr>
<td>wav2vec 2.0</td>
<td>0.430</td>
<td>0.739</td>
<td>0.718</td>
<td>0.483</td>
<td>0.964</td>
<td>0.893</td>
</tr>
<tr>
<td>TERA</td>
<td>0.496</td>
<td>0.705</td>
<td>0.680</td>
<td>0.655</td>
<td>0.940</td>
<td>0.916</td>
</tr>
<tr>
<td>APC</td>
<td>0.452</td>
<td>0.685</td>
<td>0.659</td>
<td>0.318</td>
<td>0.941</td>
<td>0.886</td>
</tr>
<tr>
<td>CPC</td>
<td><strong>0.408</strong></td>
<td>0.698</td>
<td>0.668</td>
<td>0.180</td>
<td>0.957</td>
<td>0.854</td>
</tr>
<tr>
<td>MBNet</td>
<td>1.134</td>
<td>0.628</td>
<td>0.592</td>
<td><strong>0.106</strong></td>
<td>0.945</td>
<td>0.878</td>
</tr>
</tbody>
</table>

Table 3: The system-level performance on VCC 2016 in with scenario when one of the modules is removed. The notation “- X” means module X is removed from the framework.

<table>
<thead>
<tr>
<th>Model</th>
<th>original</th>
<th>- segmental embeddings</th>
<th>- attention pooling</th>
<th>- range clipping</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE</td>
<td>LCC</td>
<td>SRCC</td>
<td>MSE</td>
</tr>
<tr>
<td>wav2vec 2.0</td>
<td>0.483</td>
<td>0.964</td>
<td>0.893</td>
<td>0.414</td>
</tr>
<tr>
<td>TERA</td>
<td>0.655</td>
<td>0.940</td>
<td>0.916</td>
<td>0.660</td>
</tr>
<tr>
<td>APC</td>
<td>0.318</td>
<td>0.941</td>
<td>0.886</td>
<td>0.373</td>
</tr>
<tr>
<td>CPC</td>
<td>0.180</td>
<td>0.957</td>
<td>0.854</td>
<td>0.208</td>
</tr>
</tbody>
</table>

5.2. Ablation study

We then inspected how each module works by removing them one at a time. In the case of segmental embedding is not used, we directly calculate the utterance score from frame-level representations. Also, following the previous work [9], the segment-level MSE in (11) becomes frame-level MSE, which enforces all frames produce the same score. When attention pooling is removed, a simple mean pooling is adopted for segment-level scores. Last, without the presence of range clipping, the range of w is no longer limited and therefore can be arbitrary real number.

Table 3 lists system-level performances on VCC 2016 in with scenario when each of the module is removed from our framework. Due to the space limit, we do not present the results in without scenario, where a similar trend was observed. We see as segment embedding is not used, the performance dropped for almost all models, showing that segment is a better unit for quality assessment than the frame is, as we mentioned in Sec 3.5. Then, when attention pooling is replaced, there is little change. This is probably because we adopted segment-level MSE in the training objective, which enforces all segments in an utterance give similar or same scores and therefore weaken the power of attention. Last, we can observe that without range clipping, MSE and LCC worsened for almost all models. This means the clipping works well for stabilizing model output within a reasonable range. Meanwhile, we also see that SRCC was not affected much because SRCC only considers the ranking order of two sets, not the exact values in them.

6. Conclusions

This paper presents the first use of self-supervised pre-trained models for MOS prediction. We show that these models can measure audio quality without fine-tuning in both qualitative and quantitative aspects, and then propose a framework in which the model is fine-tuned with simple yet effective modules. Our framework outperformed the previous works on VCC 2016 and achieved comparable or superior performance on VCC 2018. Experimental results showed that wav2vec 2.0 achieved the best or competitive performance in both utterance- and system-level. We also conducted a thorough ablation study to explore how each module in the framework works. The use of segmental embeddings boosts the performance the most, confirming that it is more reasonable to score a segment of utterance instead of every frame in the previous work.

7. Acknowledgements

We acknowledge the support of AWS Machine Learning Research Awards program.
8. References


