



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 or superior performance on Voice Conversion Challenge 2016. We also conducted an ablation study to further investigate how each module benefits the task. The experiment results are implemented and reproducible with publicly available toolkits¹.

Index Terms: MOS prediction, speech quality assessment, self-supervised learning

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 syntheses such as text-to-speech and voice conversion, 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 [19] 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-

* equal contribution

¹ <https://github.com/s3prl/s3prl>

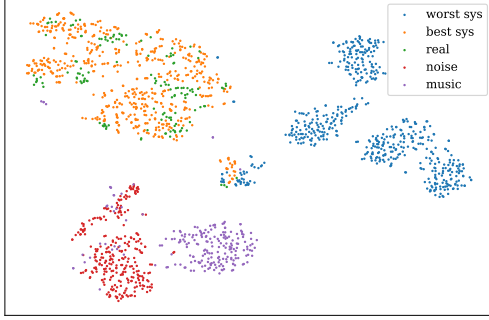


Figure 1: A visualization of representations of different kinds of audios extracted by wav2vec 2.0.

ing t-SNE [23]. Figure 1 illustrates the relationships among different kinds of audios. We can see that real clean speech (green) clusters together, and audios from the best system (orange) are mixed with them. On the other hand, audios from the worst system do not cluster with clean speech (blue far away from green), neither do noise (red) and music (purple). This suggests wav2vec 2.0 can classify different kinds of audios. Synthetic speech sounding more realistic tends to be closer to the cluster of real speech, and vice versa.

We extended the analysis by canonical correlation analysis [24] (CCA) on four pre-trained models: wav2vec 2.0 base, TERA base, APC with 3-layer GRU, and CPC. CCA finds a linear transform mapping the representations to scores, and the linear correlation between these estimated scores and ground truth is maximized. For each utterance, the model gives frame-level representations, which then form the utterance-level one by taking the average over all frames. These utterance-level representations were then used for CCA. We performed CCA on VCC 2018 training set and then applied the found transform on VCC 2018 testing set and VCC 2016 to see whether the representations are general. As the baseline, we included two conventional speech features in the analysis: MFCC and Mel-spectrogram.

Table 1 presents the linear correlation between estimated scores and ground truths on VCC 2018 testing set and VCC 2016 using four different self-supervised pre-trained models. We can see that the correlation on these representations is much higher than the baseline, which implies they contain information about audio quality. Also, when the found transform is applied on VCC 2016, the correlation remains much high, suggesting that these representations are general to evaluate audio quality. With both qualitative and quantitative analysis, we can know that it is useful and feasible to use self-supervised pre-trained models for MOS prediction.

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.

Model	Utterance-level	System-level	
	VCC 2018	VCC 2016	VCC 2018
wav2vec 2.0	0.734	0.966	0.99
TERA	0.727	0.943	0.987
APC	0.678	0.891	0.964
CPC	0.699	0.890	0.980
MFCC	0.183	0.196	0.326
Mel-spec.	0.215	0.487	0.618

3. Proposed Framework

In MOS test, an audio is evaluated by K subjects giving a sequence of scores y_1, y_2, \dots, y_K . These scores are averaged to obtain a mean score, denoted as y . Our goal is to predict y from \mathbf{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).

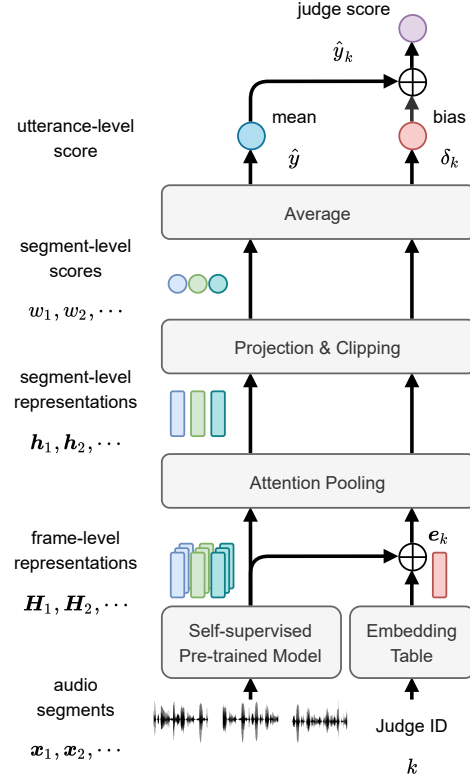


Figure 2: The proposed framework for MOS prediction.

3.1. Segmental embeddings

We start with a pre-trained self-supervised model denoted as $f(\cdot)$. An audio of T sample points $\mathbf{x} = [x_1, x_2, \dots, x_T]$ is first divided into several segments \mathbf{x}_{seg} ,

$$\mathbf{x}_{seg} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N], \quad (1)$$

$$\mathbf{x}_i = [x_{i \cdot S}, x_{i \cdot S + 1}, \dots, x_{i \cdot S + \ell}], \quad (2)$$

where ℓ is the length of each segment and S is the stride. Then the pre-trained model encodes these segments into a sequence of representations,

$$f(\mathbf{x}_{seg}) = [\mathbf{H}_1, \mathbf{H}_2, \dots, \mathbf{H}_N], \quad (3)$$

$$\mathbf{H}_i = [h_{i1}, h_{i2}, \dots, h_{iM}], \quad (4)$$

where N is the number of segments and M is the length of features of a segment. Here $h_{ij} \in \mathbb{R}^d$ is referred to as **frame-level** representations, and d is their dimensionality.

3.2. Attention pooling

Attention mechanism [29] has gained much success in several tasks. Here we use attention pooling [30] to obtain segment-

level representation from a sequence of frame-level representations. For each segment, the attention module first encodes frame-level representations into queries,

$$\mathbf{Q}_i = \text{softmax}(\mathbf{W}\mathbf{H}_i), \quad (5)$$

where $\mathbf{H}_i \in \mathbb{R}^{d \times M}$, $\mathbf{W} \in \mathbb{R}^{1 \times d}$, $\mathbf{Q}_i \in \mathbb{R}^{1 \times M}$. It then gives the **segment-level** representation by

$$\mathbf{h}_i = \mathbf{H}_i \mathbf{Q}_i^T. \quad (6)$$

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

$$w_i = g(\mathbf{h}_i). \quad (7)$$

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

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N w_i \quad (8)$$

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 \mathbf{e}_k by a trainable embedding table. It is added to frame-level representations $\mathbf{h}_{i,j}$ to form judge-biased representations $\hat{\mathbf{h}}_{i,j}$. Then, we obtain the bias of a judge δ_k from $\hat{\mathbf{h}}_{i,j}$ by a bias network with similar process described in (5) - (8). We can determine the score given by judge k by

$$\hat{y}_k = \hat{y} + \delta_k. \quad (9)$$

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:

$$\hat{w}_i = 2 \tanh g(\mathbf{h}_i) + 3. \quad (10)$$

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

$$\mathcal{L} = (\hat{y} - y)^2 + \frac{\alpha}{N} \sum_{i=1}^N (w_i - y)^2 + \frac{\beta}{K} \sum_{j=1}^K (\hat{y}_j - y)^2, \quad (11)$$

where α and β 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 (ℓ) and stride (S) are 1.0 and 0.5 seconds respectively. The raw representations $f(\mathbf{x}_{seg})$ 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 **with** 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 pairwise, 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

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

³<https://github.com/sky1456723/Pytorch-MBNet>

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

(a) utterance-level				(b) system-level						
Model	VCC 2018			VCC 2016			VCC 2018			
	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC	
with bias network				with bias network						
wav2vec 2.0	0.450	0.739	0.718	wav2vec 2.0	0.483	0.964	0.893	0.083	0.981	0.968
TERA	0.496	0.705	0.680	TERA	0.655	0.940	0.916	0.090	0.987	0.972
APC	0.425	0.685	0.659	APC	0.318	0.941	0.886	0.028	0.967	0.961
CPC	0.408	0.698	0.668	CPC	0.180	0.957	0.854	0.016	0.981	0.968
MBNet	1.134	0.628	0.592	MBNet	0.106	0.945	0.878	0.771	0.982	0.980
without bias network				without bias network						
wav2vec 2.0	0.496	0.722	0.698	wav2vec 2.0	0.615	0.961	0.868	0.104	0.991	0.981
TERA	0.426	0.692	0.661	TERA	0.407	0.942	0.896	0.023	0.988	0.977
APC	0.440	0.678	0.650	APC	0.238	0.938	0.887	0.046	0.971	0.949
CPC	0.423	0.686	0.651	CPC	0.170	0.955	0.844	0.016	0.983	0.960
MOSNet	0.471	0.639	0.604	MOSNet	0.336	0.901	0.850	0.054	0.960	0.918

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.

Model	original			- segmental embeddings			- attention pooling			- range clipping		
	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC	MSE	LCC	SRCC
wav2vec 2.0	0.483	0.964	0.893	0.414	0.958	0.854	0.499	0.962	0.859	0.623	0.967	0.860
TERA	0.655	0.940	0.916	0.660	0.887	0.785	0.637	0.941	0.920	0.509	0.934	0.905
APC	0.318	0.941	0.886	0.373	0.944	0.902	0.306	0.938	0.890	0.364	0.926	0.866
CPC	0.180	0.957	0.854	0.208	0.940	0.824	0.188	0.958	0.866	0.243	0.948	0.848

the performance than MSE does. This is because MSE cannot capture the positive/negative correlation between two sets of samples but only reflects the absolute difference. As in **with** scenario, MBNet achieved the best SRCC in VCC 2018 (even better than it was reported in the original paper though trained with less audios). However, when it comes to SRCC in VCC 2016, pre-trained models performed better except CPC. As for LCC, we see all pre-trained models are at least comparable to MBNet in both VCC 2016 and 2018, and wav2vec 2.0 even surpassed much (LCC = 0.964). Last, we can again find that **with** scenario is superior to **without** scenario, suggesting that using individual judge scores is an important key to achieving better MOS prediction.

5.2. Ablation study

We then inspected how each module works by removing them one at a time. In the case 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 2018 and achieved comparable or superior performance on VCC 2016. 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

- [1] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.
- [2] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2010, pp. 4214–4217.
- [3] R. Kubichek, "Mel-cestral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.
- [4] T. Yoshimura, G. E. Henter, O. Watts, M. Wester, J. Yamagishi, and K. Tokuda, "A hierarchical predictor of synthetic speech naturalness using neural networks," 2016.
- [5] S.-W. Fu, Y. Tsao, H.-T. Hwang, and H.-M. Wang, "Quality-net: An end-to-end non-intrusive speech quality assessment model based on blstm," *arXiv preprint arXiv:1808.05344*, 2018.
- [6] B. Patton, Y. Agiomyrgiannakis, M. Terry, K. Wilson, R. A. Saurous, and D. Sculley, "Automos: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv preprint arXiv:1611.09207*, 2016.
- [7] A. R. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 631–635.
- [8] T. H. Falk and W. . Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1935–1947, 2006.
- [9] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "Mosnet: Deep learning-based objective assessment for voice conversion," *Proc. Interspeech 2019*, pp. 1541–1545, 2019.
- [10] B. Efron and R. J. Tibshirani, *An introduction to the bootstrap*. CRC press, 1994.
- [11] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnnet: Mos prediction for synthesized speech with mean-bias network," *arXiv preprint arXiv:2103.00110*, 2021.
- [12] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
- [13] A. Baeviski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.
- [14] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *Proc. Interspeech 2019*, pp. 146–150, 2019.
- [15] A. T. Liu, S.-W. Li, and H.-y. Lee, "Tera: Self-supervised learning of transformer encoder representation for speech," *arXiv preprint arXiv:2007.06028*, 2020.
- [16] A. Baeviski and A. Mohamed, "Effectiveness of self-supervised pre-training for asr," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7694–7698.
- [17] W. Xia, C. Zhang, C. Weng, M. Yu, and D. Yu, "Self-supervised text-independent speaker verification using prototypical momentum contrastive learning," *arXiv preprint arXiv:2012.07178*, 2020.
- [18] Y. Y. Lin, C.-M. Chien, J.-H. Lin, H.-y. Lee, and L.-s. Lee, "Fragmentvc: Any-to-any voice conversion by end-to-end extracting and fusing fine-grained voice fragments with attention," *arXiv preprint arXiv:2010.14150*, 2020.
- [19] J. Lorenzo-Trueba, J. Yamagishi, T. Toda, D. Saito, F. Villavicencio, T. Kinnunen, and Z. Ling, "The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods."
- [20] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," *Interspeech 2016*, pp. 1632–1636, 2016.
- [21] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "Wham!: Extending speech separation to noisy environments," *Proc. Interspeech 2019*, pp. 1368–1372, 2019.
- [22] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.
- [23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.
- [24] B. Thompson, "Canonical correlation analysis," *Encyclopedia of statistics in behavioral science*, 2005.
- [25] M. T. Shami and M. S. Kamel, "Segment-based approach to the recognition of emotions in speech," in *2005 IEEE International Conference on Multimedia and Expo*, 2005, pp. 4 pp.–.
- [26] D. Rybach, C. Gollan, R. Schluter, and H. Ney, "Audio segmentation for speech recognition using segment features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4197–4200.
- [27] Y.-A. Chung, C.-C. Wu, C.-H. Shen, H.-Y. Lee, and L.-S. Lee, "Audio word2vec: Unsupervised learning of audio segment representations using sequence-to-sequence autoencoder," 2016.
- [28] Y.-H. Wang, H. yi Lee, and L. shan Lee, "Segmental audio word2vec: Representing utterances as sequences of vectors with applications in spoken term detection," 2018.
- [29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017.
- [30] P. Safari, M. India, and J. Hernando, "Self-attention encoding and pooling for speaker recognition," *Proc. Interspeech 2020*, pp. 941–945, 2020.
- [31] V. Panayotov, G. Chen, D. Povey, and S. 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.
- [32] J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux, "Libri-light: A benchmark for asr with limited or no supervision," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7669–7673, <https://github.com/facebookresearch/libri-light>.
- [33] A. T. Liu and Y. Shu-wen, "S3prl: The self-supervised speech pre-training and representation learning toolkit," 2020. [Online]. Available: <https://github.com/s3prl/s3prl>
- [34] K. Pearson, "Notes on the history of correlation," *Biometrika*, vol. 13, no. 1, pp. 25–45, 1920.
- [35] C. Spearman, "The proof and measurement of association between two things." *The American Journal of Psychology*, 1904.