UNSEEN BUT NOT UNKNOWN: USING DATASET CONCEALMENT TO ROBUSTLY EVALUATE SPEECH QUALITY ESTIMATION MODELS

Jaden Pieper and Stephen D. Voran

Institute for Telecommunication Sciences, Boulder, Colorado, USA
{jpieper, svoran}@ntia.gov

ABSTRACT

We introduce Dataset Concealment (DSC), a rigorous new procedure for evaluating and interpreting objective speech quality estimation models. DSC quantifies and decomposes the performance gap between research results and real-world application requirements, while offering context and additional insights into model behavior and dataset characteristics. We also show the benefits of addressing the corpus effect by using the dataset Aligner from AlignNet when training models with multiple datasets. We demonstrate DSC and the improvements from the Aligner using nine training datasets and nine unseen datasets with three well-studied models: MOSNet, NISQA, and a Wav2Vec2.0-based model. DSC provides interpretable views of the generalization capabilities and limitations of models, while allowing all available data to be used at training. An additional result is that adding the 1000 parameter dataset Aligner to the 94 million parameter Wav2Vec model during training does significantly improve the resulting model's ability to estimate speech quality for unseen data.

Index Terms— corpus effect, dataset alignment, no-reference estimator, speech quality, subjective test

1. INTRODUCTION

Objective estimation of speech quality is an important and enduring problem. Full-reference estimators (or models) are suitable for out-of-service applications, but only no-reference (NR) models can provide the real-time, in-service results needed for detecting, tracking, and diagnosing the effects of acoustic environments, network loading, and radio conditions in today's complex and dynamic telecommunications environments. NR estimation is a hard problem. A successful model must implicitly embody a very general and flexible yet highly detailed representation of what speech should sound like. Machine learning (ML) has made this hard problem somewhat easier — ML can leverage a sufficient quantity of speech-quality data (typically speech files and corresponding speech quality values from subjective tests) to train an algorithm so that it can then process previously unseen streams of speech and generate estimates of the corresponding speech quality values.

There is a gap between this very natural, dataset-driven training paradigm and real-world use cases involving arbitrary speech and impairments. ML-based models typically perform worse on data outside of their training datasets [1]. In order to improve model performance for real-world applications, we need a clear understanding of how and why these performance drops occur. We provide a path to developing more useful, robust models via a rigorous evaluation of the relationship between model architecture and training data. This approach quantifies and decomposes the gap between dataset specific research results and real-world applications.

The primary contribution of this paper is Dataset Concealment (DSC). DSC is a set of steps for training and evaluating a model that provides a truly interpretable view of its generalization capabilities. DSC trains with multiple datasets which increases the range of conditions seen at training and allows us to understand the interactions between datasets and a given model architecture. Increasing the range of conditions during training increases the ultimate performance on unseen data. But using multiple datasets for training comes with a cost due to the non-absolute nature of subjective test ratings: the corpus effect, or range-equalizing bias, as demonstrated in [2]. We address this by using the dataset Aligner module from the AlignNet architecture [3] to learn dataset alignments. A secondary contribution of this paper is demonstrating that Aligners improve a model's ability to estimate quality at inference for unseen data.

In Section 2 we catalog and discuss existing training and evaluation strategies for designing generalizable speech quality estimators. We also describe the corpus effect and an existing approach to mitigating its impact on model performance. In Section 3 we formally define DSC. Sections 4 and 5 describe 18 datasets, 3 well-known speech quality models, and the results obtained when DSC is applied using 9 of the datasets. Finally, in Section 6 we discuss quality estimation results on nine additional unseen datasets for models trained with and without dataset Aligners.

2. BACKGROUND AND MOTIVATION

Developing generalizable speech quality models requires robust and careful evaluation strategies. One existing strategy is to "hold out" randomly selected portions of data at training and to use these for testing. This is a good first step — any given test file is indeed unseen at training time. But the structure of many datasets means that random splitting of files will cause conditions (e.g., street noise at 10 dB SNR), talkers, or sentences to appear in both the training and testing sets. Cross-validation can be viewed as repeating this operation multiple times in a structured way. While this is a more robust approach, it does not account for the existence of overlap.

Another strategy is to curate splits within datasets so that specific conditions appear only during training and other conditions appear only at testing. This is sensitive to the way the conditions are chosen and begs questions like "is street noise at 0 and 20 dB SNR different conditions or the same condition?" An additional drawback is the possibility of other types of overlaps between the training and testing sets, e.g., talkers, speech activity structure, or recording conditions.

A third strategy is to hold out one or more entire datasets as unseen at training and then use model performance on the unseen dataset(s) as a measure of generalization. This should minimize any overlap, but meaningful interpretation of the results can be hindered by two unknowns. First, how "easy" or "hard" is an unseen dataset? A dataset can be easy when a large portion of the label variance is driven by an easily detected condition, such as the low-pass

filtering that is used to form anchor conditions in some subjective tests. A dataset can be hard when the majority of the label variance is driven by acoustic nuances such as those that distinguish neural codecs or voice conversion systems that are of similar quality. The second issue for interpreting results on held-out datasets is the question of how different a dataset is from those used in training. The terms "in/out of distribution" and "in/out of domain" [4, 5, 6] or "matched/mismatched case" [7] are used to describe dataset uniqueness in a very coarse and highly subjective way. A more objective and quantitative perspective on dataset uniqueness is needed.

As will be demonstrated, DSC addresses these issues to provide rich context for understanding how a model performs on a given dataset. This paper also addresses the corpus effect, which must be considered when multiple datasets are used during training. The corpus effect describes the non-absolute nature of MOS ratings that can cause inherent mismatches in subjective labels between different datasets. For example, a speech file that received an average rating of 3.0 in one listening experiment could have an average of 3.8 when included in a second experiment because of different contexts, listeners, and experimental designs. This mismatch between experiments becomes label noise when training a model that must be accurate across multiple datasets. Seeking to efficiently use as much data as possible while training, we explore addressing the corpus effect through the addition of a dataset Aligner [3] during training. We specifically aim to understand the effect of learning dataset alignments on the inference performance for truly unseen datasets.

3. THE DATASET CONCEALMENT PROCESS

DSC is a new combination of training, testing, and analysis steps for N datasets that yields meaningful insights on model performance as well as the datasets in use. We assume that each dataset D_j comprises a training set, a validation set, and a held-out test set. DSC applied to N datasets entails training a model multiple times to create the following:

- Individual Models: Train the model N times, using each dataset individually, $T_{I,j} = \{D_j\}$
- Global Model: Train the model one time using all N datasets together, $T_G = \{D_i\}_{i=1}^{N}$
- Concealed Models: Train the model N times, each time using all datasets except one, $T_{C,j} = \{D_i\}_{i \neq j}$

The final step of DSC is to apply three different versions of the model to each held-out test set, yielding $3N$ total test results. That is, for dataset D_j: (1) Test the Individual Model that was trained using only dataset D_j, (2) Test the Global Model that was trained using all datasets, (3) Test the Concealed Model that was trained using all datasets except D_j. Each testing step produces a figure of merit, and this could be any figure of merit that quantifies the model's agreement with the subjective labels, e.g., linear correlation coefficient (LCC) or Spearman's rank correlation coefficient (SRCC). Whenever multiple datasets are used, one must consider the corpus effect which can significantly limit the usefulness of RMSE, especially when a model must produce estimates for previously unseen datasets. Thus the following definitions are intended for LCC and SRCC and our examples use LCC only, without loss of generality.

We denote the dataset concealment Individual, Global and Concealed LCC values for dataset D_j with $\rho_{I,j}$, $\rho_{G,j}$, and $\rho_{C,j}$, respectively. $\rho_{I,j}$ is the correlation observed for the Individual Model that is trained using only dataset D_j. It describes how well the model can learn this dataset in isolation, which is typically the easiest way to learn a dataset. $\rho_{G,j}$ is the correlation observed from the Global Model that is trained with all N datasets T_G. This describes the broad performance of the model, or how well it can combine information from multiple datasets. The versatility gap of a model on dataset D_j is

$$v_j = |\rho_{I,j}| - |\rho_{G,j}|. \quad (1)$$

Currently, when models are trained on multiple datasets, they generally perform worse on any given dataset when compared to individual training. So, versatility gap values are generally positive, and the goal is to minimize this gap. Ideally, a model would be able to harmonize learning from multiple datasets to produce a more generalized and robust relationship between speech and labels, increasing model performance on any dataset, and giving negative values of v_j.

The correlation $\rho_{C,j}$ is observed from the Concealed Model for dataset D_j, where all the datasets except for D_j are used during training. This correlation describes the model's ability to generalize information from other training datasets for this specific dataset and make estimations on the truly unseen dataset D_j. The concealment gap of a model on dataset D_j is

$$c_j = |\rho_{G,j}| - |\rho_{C,j}|. \quad (2)$$

Models that generalize better have smaller concealment gaps.

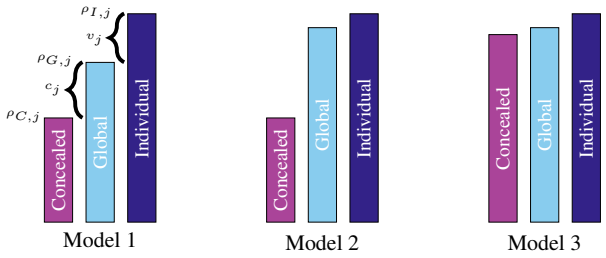

Fig. 1. Example dataset concealment results and interpretation.

Figure 1 provides a pictorial representation of the three correlations and the two gaps for three models and one dataset. The figure shows a progression of model improvement (Model 1 to Model 2 to Model 3). The three models show similar performance, $\rho_{I,j}$, when trained on the dataset individually. But Model 1 is not very versatile at learning from multiple datasets at once, as seen through its significant versatility gap; when trained with multiple datasets, it loses the ability to perform well on the considered dataset. Model 2 reduces the versatility gap, but does not generalize well. This is clear because it has a significant concealment gap — when the considered dataset is concealed during training, Model 2 does poorly on that dataset. Model 3 has minimal versatility and concealment gaps so we conclude that Model 3 is both versatile and generalizable.

4. DATASETS AND MODELS

Table 1 summarizes key attributes of the nine datasets used to demonstrate DSC and nine other datasets that are truly unseen and used for additional context on generalization ability of models. The NOIZEUS and PSTN datasets contain narrowband speech (nominal upper limit at 3.4 kHz) and the remaining datasets contain wideband speech (7 kHz limit) or fullband speech (20 kHz limit).

We demonstrate DSC for three different well-established and well-studied speech quality models. This allows us to compare what DSC tells us about these models with what is already understood about these models. The intended use case for DSC is the evaluation of proposed *new* models, thoroughly gathering rich, well-contextualized insights that enable us to see strengths and weaknesses relative to other models. It is our intent to do this going forward, and we invite other researchers to leverage DSC as well.

Dataset	Year	Files	Conditions	Votes per File	Use
FFTNet [8]	2018	1200	Neural codecs	8.9	DSC
NOIZEUS [9]	2007	1664	Noise & NR	8.0	DSC
VMC2022 [10]	2022	7106	Synthesized & natural speech	8.0	DSC
Tencent [11]	2022	11,563	Noise, NR, reverb, codecs, packet loss, PLC	20.0	DSC
NISQA SIM Val [12]	2021	12,500	Codecs, packet loss, noise, filtering, clipping	5.2	DSC
TMHINT-QI [13]	2022	14,915	Noise & NR	1.6	DSC
VCC2018†[14]	2018	19,670	Voice conversion algorithms	4.0	DSC
IU [15]	2020	36,000	Noise & reverb	5.0	DSC
PSTN [16]	2020	58,709	PSTN/VoIP calls, noise	4.6	DSC
NISQA Livetalk [12]	2021	232	Live calls (wired & wireless, PSTN & VoIP), talking levels, natural and recorded noise	24.0	Unseen
NISQA FOR [12]	2021	240	Codecs, noise, packet loss, clipping, live OTT calls	29.3	Unseen
NISQA NSC [12]	2021	240	Codecs, noise, packet loss, clipping & live calls	27.2	Unseen
NISQA P501 [12]	2021	240	Codecs, noise, packet loss, clipping & live calls	28.3	Unseen
Blizzard2021 SS1 Nat. [17]	2021	242	Synthesized & natural speech	16.0	Unseen
Blizzard2021 SH1 Nat. [17]	2021	338	Synthesized & natural speech	24.1	Unseen
Blizzard2021 SS1 Acc. [17]	2021	363	Synthesized & natural speech	16.1	Unseen
Blizzard2008 News Nat. [18]	2008	802	Synthesized & natural speech	10.5	Unseen
Blizzard2008 Novel Nat. [18]	2008	802	Synthesized & natural speech	10.1	Unseen

Table 1. Summary of datasets used. "Overall speech quality" was rated except where "Nat." indicates that "speech naturalness" was rated and "Acc." means that "speech acceptability" was rated. NR: noise reduction, PLC: packet loss concealment, OTT: over-the-top. †VCC2018 has been punctured to remove overlap with VMC2022.

The first model we consider is MOSNet [19]. This model represents an early approach to NR ML estimation that relies on a CNN and BLSTM based architecture. It is a medium-sized model with 1.4 million parameters and is the basis for subsequent work [20, 21]. MOSNet is known for poor generalization, so it provides a useful example model for DSC. We trained MOSNet as described in [3].

The second model we consider is a single-headed NISQA [22], as most of the datasets used here have labels for speech quality only. NISQA is an early attention-based approach to speech quality estimation that is extremely light-weight and successful and is known to generalize well [11], despite having only 218,000 parameters.

Our third example is a Wav2Vec2.0-based [23] speech quality model, often called SSL-MOS [1], but here referred to as Wav2Vec, for simplicity. This model is an early effort to leverage massive self-supervised speech models for speech quality estimation. This is done by attaching a single fully-connected layer to the feature output from the SSL model. The Wav2Vec model has 94 million parameters. Large SSL-based models are known to be extremely robust and generalizable at inference for unseen data.

We trained Global and Concealed Models in two different ways: conventionally and with the addition of an Aligner as in the AlignNet architecture [3]. Conventional training treats all datasets as one, despite the fact that their labels may not be compatible. The AlignNet architecture appends a small and effective dataset Aligner to the output of the model to mitigate the corpus effect between datasets. The corpus effect can cause incompatibility between labels from one subjective test and those from a different subjective test. The Aligner uses a dataset indicator to map an intermediate, reference dataset score to the appropriate target dataset scale, effectively learning the alignment functions between datasets. This reduces the dissonance involved in learning from multiple datasets, and allows the audio model to focus solely on accurately learning the audio space.

ML is a powerful tool and some models are able to do some learning through the label noise introduced by the corpus effect, without the use of an Aligner [24]. However, removing label noise during the training process allows for the learning of more robust relationships between speech and quality and thus improves model performance at inference on unseen data. We expect the magnitude of this effect from the Aligner to be model-dependent.

5. EXAMPLE DATASET CONCEALMENT RESULTS

We ran a variety of experiments to demonstrate the insights that DSC provides and the improvements that training with an Aligner can offer. We trained Global and Individual models 10 times each with randomized training/validation/testing splits, respecting curated splits where applicable. Due to the number of datasets used here, we ran two replications when training Concealed Models. We report "average" correlations (we average the Fisher z-transformation[25] of the per-replication correlations and then apply the inverse of the Fisher z-transformation). Statistical significance is determined by computing the standard error of the Fisher z-transformed average correlation and computing a 95% confidence interval of the difference of the transformed averages. Training details specific to dataset alignment include: NISQA Sim is the reference dataset; Tencent is the reference dataset when NISQA Sim is the concealed dataset; when training MOSNet and NISQA the Aligner is frozen until the validation correlation reaches a threshold of 0.6, so that the model can reach a meaningful representation before the learning of dataset alignments; when training Wav2Vec, due to the existing SSL-based speech representation and comparatively long training times, the Aligner is never frozen. Full details are provided at `https://github.com/NTIA/Dataset-Concealment`.

The DSC LCC results and the impact of adding an Aligner during training can be seen in Fig. 2. Figure 3 shows the DSC versatility gaps v_i and concealment gaps c_i, but only for models trained with an Aligner, due to space constraints. Insights about models, datasets, and interactions between the two can be easily seen in these two figures. For Individual Models, MOSNet performs worse than NISQA, which performs worse than Wav2Vec. This is well known and expected, and often where analysis of this sort starts and ends; but DSC provides much more information.

The Global Models show a different trend. Figure 3 makes it clear that, both conventionally and with an Aligner, NISQA has very low versatility gaps — training NISQA with multiple datasets has a very small impact on its ability to predict any individual dataset. Wav2Vec with an Aligner, on the other hand, has versatility gaps that are slightly larger than those of NISQA, indicating that Wav2Vec has slightly lower versatility. In fact, Fig. 2 shows that the NISQA Global Model has higher correlations than the Wav2Vec Global Model on the FFTNet, NOIZEUS, VMC22, TMHINT-QI, and IU datasets. MOSNet shows the largest versatility gaps for every dataset outside of VMC22, and the MOSNet ρ_G values in Fig. 2 show that MOSNet struggles to learn from multiple datasets at all. This is different behavior from what was observed in [3], which we posit is due to using better defined, more challenging splits for the datasets and the group of DSC training datasets being more challenging and diverse than the two groups used in that work.

Adding an Aligner does not produce any statistically significant differences in MOSNet Global Models. But adding an Aligner to NISQA Global Models during training produces statistically significant LCC increases for five out of nine datasets. And adding an Aligner to Wav2Vec Global Models gives statistically significant LCC increases for seven out of nine datasets. Wav2Vec, in many ways, can be considered a state-of-the-art speech quality model, and adding an Aligner improved its ability to predict speech quality in held-out test sets. The fact that adding a 1000-parameter Aligner to

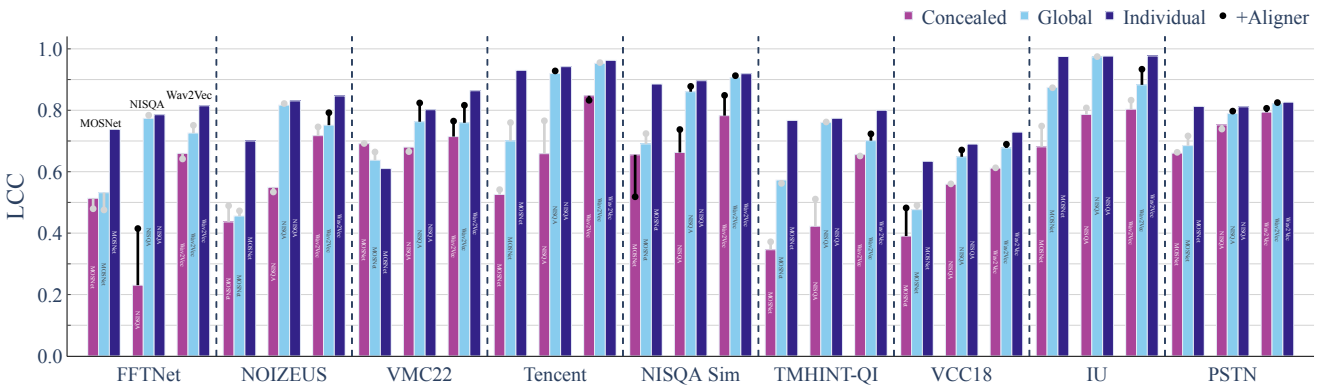

Fig. 2. DSC results for MOSNet, NISQA, and Wav2Vec models across nine datasets. Bars show performance when training conventionally and lines extending from bars show the effect of training with an Aligner. Black lines indicate statistically significant changes, grey otherwise.

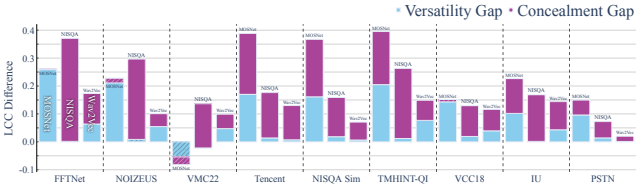

Fig. 3. DSC versatility and concealment gaps for models trained with an Aligner. Solid bars represent statistically significant gaps, otherwise bars are hatched.

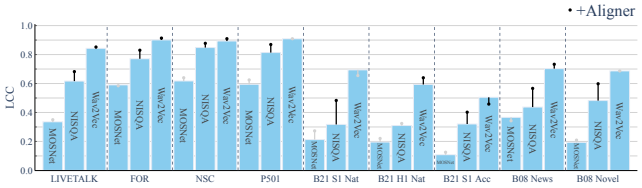

Fig. 4. Inference performance of MOSNet, NISQA, and Wav2Vec across nine unseen datasets. Bars show the performance when training conventionally. Lines extending from the bars show the effect of training with an Aligner. Black lines indicate statistically significant changes, grey otherwise.

a 94-million parameter model produces a statistically significant improvement in the majority of cases demonstrates that label noise due to the corpus effect is impacting the ability of a model to learn and this can be efficiently mitigated.

Concealed training results describe how well each model can predict speech quality in a dataset that has not been seen during training. Figures 2 and 3 show that the Wav2Vec concealment gaps are generally the smallest. The results for PSTN with Wav2Vec are particularly impressive, with Concealed training achieving a correlation of $\rho_{C,j} = 0.81$ compared to the Individual training result of $\rho_{I,j} = 0.83$. Figure 3 makes it clear that this difference is almost entirely due to the Concealment gap — the versatility gap is not statistically significant. NISQA models have larger Concealment gaps than Wav2Vec models for almost every dataset. So while NISQA proves itself to be more versatile than Wav2Vec in terms of training with multiple datasets, Wav2Vec shows a better ability to generalize to unseen data than NISQA. This is an example of the type of nuanced model comparison that only DSC can provide. Finally, Concealed MOSNet models have a very poor showing, reinforcing MOSNet's limitations as a model that tends to overfit to the data it is trained on. In general, adding an Aligner has lesser impact in Concealed training than it does in Global training.

DSC also yields insights about datasets. For example, the FFTNet dataset gives the largest and second largest concealment gap for NISQA and Wav2Vec respectively, consistent with the unique conditions in that dataset. The Tencent dataset gives the largest MOSNet concealment gap, second largest for Wav2Vec, and fifth largest for NISQA. While the Tencent dataset shares telecommunication conditions with the NISQA Sim and TMHINT-QI datasets, it apparently is novel in other relevant aspects (e.g., recording conditions, speech material, or talkers).

6. INFERENCE RESULTS

Figure 4 shows inference LCC values for globally-trained MOSNet, NISQA, and Wav2Vec, trained with and without an Aligner. These inference results are for nine unseen test datasets (datasets not used in DSC). The trends in Fig. 4 agree with what we have already learned from the DSC concealment gaps: for truly unseen datasets, Wav2Vec performs the best, NISQA is somewhat worse, and MOSNet estimates are unlikely to be useful. Training with an Aligner proves to be beneficial for the models at inference. The addition of an Aligner gives statistically significant improvements for eight of the nine datasets for NISQA and five of the nine for Wav2Vec. Wav2Vec has proved itself to be one of the more powerful current models for producing estimates for unseen data, so the fact that a 1000-parameter Aligner from the AlignNet architecture can improve its inference ability demonstrates the value of acknowledging and attempting to mitigate the impact of the corpus effect during training.

7. CONCLUSION

We have defined DSC, applied it to three speech quality models and nine datasets, and described the insights that it provides. DSC produces a versatility gap and a concealment gap for each model and dataset. These illuminate the limitations of a model architecture and add context that can help inform future architectural designs. We evaluated the models with nine additional truly unseen datasets, and those results affirm the insights derived from using DSC. In addition, both sets of experiments demonstrate that adding a lightweight Aligner when training with multiple datasets yields statistically significant improvements most of the time, both for held-out test sets and for truly unseen data.

8. REFERENCES

[1] Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi, "Generalization ability of MOS prediction networks," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 8442–8446.

[2] Erica Cooper and Junichi Yamagishi, "Investigating range-equalizing bias in mean opinion score ratings of synthesized speech," in *Proc. Interspeech*, 2023, pp. 1104–1108.

[3] Jaden Pieper and Stephen Voran, "Alignnet: Learning dataset score alignment functions to enable better training of speech quality estimators," in *Proc. Interspeech 2024*, pp. 82–86.

[4] Alessandro Ragano, Jan Skoglund, and Andrew Hines, "SCOREQ: Speech quality assessment with contrastive regression," in *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, Eds. 2024, vol. 37, pp. 105702–105729, Curran Associates, Inc.

[5] Imran E Kibria and Donald S Williamson, "AttentiveMOS: A Lightweight Attention-Only Model for Speech Quality Prediction," in *Proc. Interspeech 2025*, pp. 2340–2344.

[6] Yu-Fei Shi, Yang Ai, and Zhen-Hua Ling, "Universal Preference-Score-based Pairwise Speech Quality Assessment," in *Proc. Interspeech 2025*, pp. 2345–2349.

[7] Danilo de Oliveira, Julius Richter, Jean-Marie Lemercier, Simon Welker, and Timo Gerkmann, "Non-intrusive Speech Quality Assessment with Diffusion Models Trained on Clean Speech," in *Proc. Interspeech 2025*, pp. 2330–2334.

[8] Zeyu Jin, Adam Finkelstein, Gautham J. Mysore, and Jingwan Lu, "FFTNet: A real-time speaker-dependent neural vocoder," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[9] Yi Hu and Philipos C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7, pp. 588–601, 2007.

[10] Wen Chin Huang, Erica Cooper, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, "The VoiceMOS Challenge 2022," in *Proc. Interspeech 2022*, pp. 4536–4540.

[11] Gaoxiong Yi, Wei Xiao, Yiming Xiao, Babak Naderi, Sebastian Moller, Wafaa Wardah, Gabriel Mittag, Ross Cutler, Zhuohuang Zhang, Donald S. Williamson, Fei Chen, Fuzheng Yang, and Shidong Shang, "ConferencingSpeech 2022 Challenge: Non-intrusive objective speech quality assessment (NISQA) challenge for online conferencing applications," in *Proc. Interspeech 2022*, pp. 3308–3312.

[12] 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 *Proc. Interspeech*, 2021, pp. 2127–2131.

[13] Yu-Wen Chen and Yu Tsao, "InQSS: A speech intelligibility and quality assessment model using a multi-task learning network," in *Proc. Interspeech 2022*, pp. 3088–3092.

[14] Jaime Lorenzo-Trueba, Junichi Yamagishi, Tomoki Toda, Daisuke Saito, Fernando Villavicencio, Tomi Kinnunen, and Zhenhua Ling, "The Voice Conversion Challenge 2018: Promoting development of parallel and nonparallel methods," in *Proc. Speaker Odyssey*, 2018.

[15] Xuan Dong and Donald S. Williamson, "A pyramid recurrent network for predicting crowdsourced speech-quality ratings of real-world signals," in *Proc. Interspeech 2020*, pp. 4631–4635.

[16] Gabriel Mittag, R. Cutler, Yasaman Hosseinkashi, M. Revow, Sriram Srinivasan, Naglakshmi Chande, and R. Aichner, "DNN no-reference PSTN speech quality prediction," in *Proc. Interspeech 2020*, pp. 2867–2871.

[17] Zhen-Hua Ling, Xiao Zhou, and Simon King, "The Blizzard Challenge 2021," in *Proc. Blizzard Challenge Workshop*, 2021.

[18] Vasilis Karaiskos, Simon King, Robert A. J. Clark, and Catherine Mayo, "The Blizzard Challenge 2008," in *The Blizzard Challenge 2008*, 2008, pp. 1–18.

[19] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech*, 2019, pp. 1541–1545.

[20] Yichong Leng, Xu Tan, Sheng Zhao, Frank Soong, Xiang-Yang Li, and Tao Qin, "MBNET: MOS prediction for synthesized speech with mean-bias network," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 391–395.

[21] Wen-Chin Huang, Erica Cooper, Junichi Yamagishi, and Tomoki Toda, "LDNet: Unified listener dependent modeling in MOS prediction for synthetic speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, 2022, pp. 896–900.

[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," *Proc. Interspeech 2021*, pp. 2127–2131.

[23] 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.

[24] Wen-Chin Huang, Erica Cooper, and Tomoki Toda, "MOSbench: Benchmarking generalization abilities of subjective speech quality assessment models," 2024, arXiv:2411.03715 [cs].

[25] R. A. Fisher, "Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population," *Biometrika*, vol. 10, no. 4, pp. 507–521, 1915.