ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10447687

CORN: CO-TRAINED FULL- AND NO-REFERENCE SPEECH QUALITY ASSESSMENT

Pranay Manocha[1], *Donald Williamson*[2], *Adam Finkelstein*[1]

[1]Department of Computer Science, Princeton University, Princeton, NJ, USA
[2]Department of Computer Science and Engineering, The Ohio State University, Columbus, OH, USA

ABSTRACT

Perceptual evaluation constitutes a crucial aspect of various audio-processing tasks. Full reference (FR) or similarity-based metrics rely on high-quality reference recordings, to which lower-quality or corrupted versions of the recording may be compared for evaluation. In contrast, no-reference (NR) metrics evaluate a recording without relying on a reference. Both the FR and NR approaches exhibit advantages and drawbacks relative to each other. In this paper, we present a novel framework called CORN that amalgamates these dual approaches, concurrently training both FR and NR models together. After training, the models can be applied independently. We evaluate CORN by predicting several common objective metrics and across two different architectures. The NR model trained using CORN has access to a reference recording during training, and thus, as one would expect, it consistently outperforms baseline NR models trained independently. Perhaps even more remarkable is that the CORN FR model also outperforms its baseline counterpart, even though it relies on the same training data and the same model architecture. Thus, a single training regime produces two independently useful models, each outperforming independently trained models.

Index Terms— perceptual similarity, speech quality, deep metric, full-reference metric, no-reference metric

1. INTRODUCTION

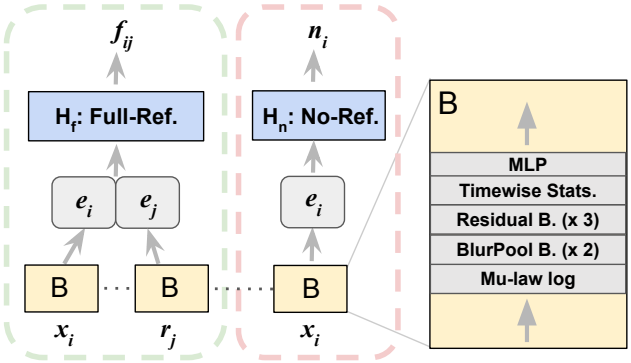
Audio quality assessment plays a significant role across a variety of applications. Human judgment indicating how good or bad a clip sounds serves as the "gold standard" method for such evaluations. However, obtaining these judgments is resource-intensive due to the associated time and cost factors. Mean Opinion Score (MOS) [1], a widely used technique to gauge sound quality, demands substantial resources, especially when repeated many times per recording, and is therefore not scalable. Additionally, ensuring controlled listening conditions further compounds the challenges of conducting MOS evaluations. Consequently, there exists a compelling impetus to explore alternative methodologies for quantifying sound quality.

Full-reference metrics, also known as intrusive or similarity metrics (e.g., PESQ [2], POLQA [3], VISQOL [4], DPAM [5], CDPAM [6]), require a clean reference to which a corrupted signal can be compared as the basis for a quality rating. Researchers commonly rely on full-reference metrics as a proxy for audio quality, because they were introduced earlier – consider, e.g. SNR. One of the most impactful is PESQ [2], introduced decades ago for telephony and still used today across tasks. These methods have been shown to correlate well with human perceptual judgments across tasks [5, 6]. However, a recent study by Manocha et al. [7] highlighted several real-world situations where established full-reference metrics (also known as similarity metrics) face discrepancies when compared with human perception. Specifically, their findings underscore that these metrics are unable to effectively account for the diverse range of audio quality variations in relation to "clean" recordings created under distinct environments. Additionally, these metrics tend to accentuate differences that are virtually imperceptible. This phenomenon arises due to the metrics' training on pairs of recordings with identical speech content, resulting in models that lack robustness in distinguishing between alterations in content and variations in quality.

To ameliorate the reliance on clean reference, *No-reference* methods rate quality on an absolute scale. Traditional methods like ITU standard P.563 [8] and SRMRnorm [9] involve complicated hand-crafted features. State of the art approaches rely on deep learning [10–15]. Earlier learning methods trained models on objective scores (e.g. PESQ) [11], while more recent approaches discover a mapping between noisy audio signals and MOS in a supervised learning fashion [10, 14–17]. However, as observed by Manocha et al. [18], no-reference metrics learn an implicit distribution of clean references, which suffers from both (a) high variance due to factors like mood and past experience; and (b) substantial label variance from human annotations. For example, in DNSMOS [10], almost half of the recordings have ratings with standard deviation > 1. Such label noise poses challenges in training robust models. Given the pronounced variance within the training labels, the task of training robust models is further complicated, leading to instability and difficulties in achieving robust performance. Moreover, the progress in refining no-reference models consistently lags behind the advancements observed in full-reference model development, contributing to the intricate landscape of audio quality assessment.

This paper proposes to learn a model of speech quality that combines multiple tasks. We call it CORN for *Co-trained Full-Reference and No-reference audio metrics*. CORN learns from different types of tasks (FR and NR), and produces speech quality scores, together with usable latent features and informative auxiliary outputs. Scores and outputs are concurrently optimized in a multi-task setting by all the different speech quality assessment tasks, with the idea that each type of model outputs the same score irrespective of its handicap (with or without reference).

As expected, the CORN NR model demonstrates superior performance compared to an independently trained NR model. This advantage arises from co-training, which provides access to the reference during training and ensures stable training with consistent gradients. However, more remarkable is that the CORN FR model surpasses its independently trained counterpart despite having the same architecture and training data. This outcome suggests that incorporating the NR loss during training assists the FR model in preventing over-generalization from the observed training content, thus enhancing its content-invariance. By flowing information through a shared latent space bottleneck, the considered objectives learn to cooperate and promote better and more robust representations while discarding non-essential information (especially speech content information) [19].



(a) FR: Full-Reference (b) NR: No-Reference (c) B: Base model

Fig. 1: Proposed CORN training framework with (a) Full-reference (FR, in green) and (b) No-Reference (NR, in red) models. Co-training (a) and (b) together – the network architecture (c) of the base model $\mathbf{B}$ is identical in each instance in the FR and NR models, and has shared weights indicated by the dotted lines. In (a) and (b), task-specific output heads $\mathbf{H}_f$ and $\mathbf{H}_n$ predict the FR and NR scores f_{ij} and n_i. In FR the embedding e_i of recording x_i is identical to its counterpart in NR; however only in FR it is concatenated with the embedding e_j of a reference recording r_j before passing along to the output head (Sections 3.1 and 3.2).

The *key contributions* of this paper are: (1) we propose a novel framework for speech quality assessment that produces *both* FR and NR models, each capable of assessing sound quality independently of the other; (2) we propose methods to train neural networks within this framework that are capable of predicting SI-SDR [20], SNR [21] and PESQ [2] scores both with and without reference recordings; and (3) we evaluate our framework through several objective evaluations and show that the FR and NR models trained via CORN outperform identical networks trained independently.

2. RELATED WORK

2.1. Full-reference metrics

Early models (PESQ [2], VISQOL [4]) mimicked human audio quality perception, but had drawbacks: sensitivity to changes, narrow focus (e.g., telephony), and non-differentiability for deep learning. Researchers then trained differentiable models to imitate PESQ [22], using GANs or gradients. Yet, these methods had optimization and generalization issues. Instead of using conventional metrics (e.g. PESQ) as a proxy, Manocha et al. [5] proposed DPAM that was trained directly on a new dataset of human just-noticeable difference (JND) judgments. DPAM correlates well with human judgment for small perturbations, but requires a large set of annotated labels to generalize well across unseen perturbations. Similarly, Serra et al.'s SESQA [23] employed the JND dataset [5], adding objectives like PESQ. However, the effects of adding different types of tasks remain unexplored.

2.2. No-reference metrics

Some of the earliest non-intrusive methods were based on complex hand-crafted, rule-based systems [8, 24, 25]. Although they are automatic and interpretable, they tend to be task-specific, and do not generalize well. Moreover, these methods are non-differentiable which limits their uses within deep learning frameworks. To overcome the last concern, various neural network-based methods have been developed [10–13, 26, 27]. However, the issue of task-specificity and generalization remains. To overcome this, researchers proposed to train models directly on a dataset of human judgment scores [10, 16, 27, 28]. Reddy et al. [10] used a multi-stage self-teaching model [29] to learn quality in the presence of noisy ratings. Nonetheless, no-reference metrics lag behind full-reference metrics in terms of correlation to human listening evaluations and adoption in practical cases.

3. THE CORN FRAMEWORK

Our framework, CORN, is designed to assess the quality of a given speech recording ($\mathbf{x}_\mathrm{i}$), an optional reference recording ($\mathbf{r}_\mathrm{j}$), and output a measure of FR f_{ij} and NR n_i quality. We propose a deep neural network for CORN and represent it by the function $\text{CORN} = \mathcal{N}(x_i, r_j)$. Given that we do not rely on any human-labeled data, the crucial components of the framework include designing tasks and objective functions that can help learn a quality score. Fig 1 is a simple illustration of the framework. In our current approach, CORN has the property of being monotonic (by design): if $\mathcal{N}(x_a, r_j) \geq \mathcal{N}(x_b, r_j)$, then $\mathtt{m}(x_a) \leq \mathtt{m}(x_b)$, where $\mathtt{m}$ is any quality assessment measure as defined in Section 3.2. We do not enforce other metric properties [30, 31] to allow flexibility in defining tasks and objectives for training the neural networks. Moreover, even human judgment of similarity may not constitute a metric [32], and hence there is no pertinent reason which necessitates CORN to have metric properties.

3.1. Framework Design and Model Architectures

CORN architecture (Fig 1) comprises two modules: a *base model* block $\mathbf{B}$, and task specific *output heads* $\mathbf{H}_f$ and $\mathbf{H}_n$.

Base model block B: We adopt a pre-existing model architecture inspired by SESQA [33], and illustrated in Figure 1(c). The model comprises of four primary stages. Initially, we pass the input x through a μ-law operation (without quantization) using a trainable μ parameter, which is initialized to 4. Following this, we utilize two pooling blocks, each comprising of convolution, batch normalization (BN), rectified linear unit (ReLU) activation, and BlurPool. These blocks employ 128 and 256 filters with a kernel width of 4, downsampling the input by a factor of 4. Subsequently, we employ three residual blocks. Each block consists of a BN preactivation followed by three stages of ReLU, convolution, and BN. These stages employ 512, 512, and 256 filters with kernel widths of 1, 3, and 1, respectively. A parametric linear averaging technique is employed to create a residual connection, expressed as follows: $h_0 = a_0 h + (1 - a_0)F(h)$, where a_0 is a vector of adjustable parameters bounded within the range of 0 to 1, F denotes the residual network, h signifies the input to the residual layers, and a represents the weight associated with the output and the residual layer. The components of a are initialized to 6, ensuring an initial emphasis on a direct path from h to h_0. After the residual blocks, temporal statistics are computed on a per-time basis, involving channel-wise mean and standard deviation calculations. This process consolidates all temporal information into a singular vector of dimensions 2×256. The vector undergoes BN before being fed into a multi-layer perceptron (MLP) comprising two linear layers with BN, and a ReLU activation in between. This MLP consists of 1024 and 200 units. We further show that our results/hypothesis do not change if the architecture changes (refer to Sec 5.1). The next blocks consist of output heads for the training tasks and are described below, along with the training loss functions.

3.2. Training Tasks and Loss Functions

FR Block: As shown in Figure 1(a), this block is designed such that the base network takes the two inputs (x_i and r_j), concatenates their embeddings (e_i and e_j) and feeds it further onto two shallow linear layers H_f that predict f_{ij}, the Scale-Invariant Signal to Distortion Ratio (SI-SDR) for the entire recording.

The goal of this task is to predict SI-SDR. Let $\hat{s} = f_{ij}$ be the recording level SI-SDR predicted by this output head. We then use the *Smoothed-L1* loss between $\hat{s} = f_{ij}$ and the target SI-SDR s to train the network:

$$L_Q(s, \hat{s}) = \begin{cases} (\hat{s} - s)^2/\beta, & |\hat{s} - s| \leq \beta \\ 2|\hat{s} - s| - \beta, & \text{otherwise} \end{cases} \quad (1)$$

NR Block: As shown in Figure 1(b), this block is designed such that the base network takes a single input x_i, passes it through the same base model B to produce the same embedding e_i as found by the FR path. Next it feeds further onto two shallow linear layers H_n that predict $\mathbf{n}_i$, the SI-SDR for the entire recording. Using $\hat{s} = n_i$ (predicted SI-SDR) we apply the *Smoothed-L1* loss between predicted and target SI-SDR s to train the network, as in equation (1).

Objective Metrics: Since we do not have any perceptual labels, CORN relies on a signal processing measure SI-SDR, to compare the quality of the two inputs. We consider SI-SDR objective metric as a proxy for human quality evaluation because it helps us train on limitless amounts of unlabeled or programmatically generated data and outputs quality scores that are consistent, unlike MOS. SI-SDR [20] is a measure that was introduced to evaluate performance of speech-processing algorithms. It is invariant to the scale of the processed signal and can be used to quantify quality in diverse cases, including additive background noises as well as other distortions. Additionally, we also show that our framework is invariant to the target label we use, so we show its performance on other objective metrics like SNR and PESQ (refer to Sec 5.1).

SNR is measured as the ratio of the signal power to the noise power and is primarily meant only for additive noises. Consider a mixture signal x, $x = r + \delta \in \mathbb{R}^L$ where r is the clean signal and δ is the noise signal, then

$$\text{SNR} = 10 \log_{10} \left(\frac{||r||^2}{||r - x||^2} \right) \quad (2)$$

$10log_{10}()$ factor measures SNR in dB-scale, and a higher SNR implies better signal quality. Yuan et al. [21] also showed that SNR as a distance metric had better properties than conventional metrics (like Euclidean distance).

PESQ [2], which stands for *perceptual evaluation of speech quality,* is an impactful objective metric used by many researchers to evaluate the sound quality of their model output with respect to a given reference. It was introduced decades ago for telephony and still used today for a wide variety of tasks including enhancement [34–37], vocoders [38], and transmission codecs [39, 40].

3.3. Training procedure

We now describe our training procedure. We assume the availability of a clean speech database $\mathcal{D}_{\text{clean}}$. The training inputs $\mathbf{x}_i$ and $\mathbf{r}_j$ are created by sampling a clean recording r_j from $\mathcal{D}_{\text{clean}}$. r_j is corrupted to produce x_i. The degradations we use can be largely grouped under two categories (a) additive noise degradations, and (b) speech distortions based on signal manipulations - *Clipping*, *Frequency Masking*, and *Mu-law compression*. For additive noise, we sample noise recordings, δ_i, from a noise database (Section 4.1) and add them to r_j at SI-SDR levels uniformly sampled from the range -40 dB to +40 dB. Once we have the degraded signal (x_i), its clean-reference counterpart (r_j), and their quality score $\mathbf{f}_{ij}$ and $\mathbf{n}_i$(=$\mathbf{f}_{ij}$), we can train the network as described in previous sections.

3.4. Inference

Once the network is trained, we can predict the quality score of a test input x_{i} along with the option of an accompanying reference input, r_{j}. Within this framework, our Full-Reference (FR) branch accepts two inputs, namely, x_{i} and r_{j}, and generates an output quality score, f_{ij}. Likewise, our No-Reference (NR) branch is designed to process a solitary input, x_{i}, and produces an NR quality score, n_{i}. The selection between these branches hinges on whether our task incorporates a reference, allowing us to determine the appropriate branch for utilization.

4. EXPERIMENTAL SETUP

4.1. Datasets and training

For training ($\mathcal{D}_{\text{train}}$), the clean set $\mathcal{D}_{\text{clean}}$ comes from the DAPS dataset [41], and the noise set $\mathcal{D}_{\text{noise}}$ comes from DNS Challenge [42] dataset. Along with additive noise, clipping, and frequency masking distortions are used during training. For robustness and better generalization to realistic conditions, we also add reverberation using room impulse responses from the DNS Challenge dataset. For the test-set ($\mathcal{D}_{\text{test}}$), we use TIMIT [43] as the source for clean speech, and ESC-50 [44] dataset for noise recordings. The test set also includes Gaussian noise addition and Mu-law compression as unseen degradations. The inputs to our model are 3-second waveform excerpts. We use the Adam optimizer with a learning rate of 10^{-4} with a batch size of 64. We train the network for 1000 epochs. Smoothed L1 parameter β=1 for all experiments.

4.2. Baselines

We undertake a comparative analysis involving our proposed methodology, CORN, in contrast to standalone models that were trained without adopting a multi-task framework, encompassing individual Full-Reference (FR) and No-Reference (NR) models. This endeavor aims to ascertain the potential superiority of our amalgamated model, CORN.

5. RESULTS

5.1. Performance across metrics and architectures

These aim to assess CORN as a proxy for subjective judgments by humans. More specifically, we evaluate how well CORN correlates

Name	SI-SDR		SNR		PESQ		New B Arch.	
	FR	NR	FR	NR	FR	NR	FR	NR
Indiv.	96.5	110.3	98.0	99.9	0.9	1.3	124.9	134.9
CORN	**85.9**	**92.9**	**79.5**	**82.7**	**0.7**	**0.9**	**103.2**	**108.2**

Table 1: **Evaluations**: Refer to Sec 5.1. Models include: CORN and individual FR and NR prediction models. The numbers show SI-SDR as a metric, unless specified otherwise. ↓ is better.

with ground truth target objective measures. We first hold out a subset of $\mathcal{D}_{\text{train}}$ and evaluate the performance of the models on that set. Next, to show generalization to unseen conditions (room environments, listeners, etc.), we also evaluate models across the unseen test dataset $\mathcal{D}_{\text{test}}$. We evaluate various models based on the output from the model, compared to the ground truth noise level using mean square error (MSE) between noise level differences.

Results are displayed in Table 1. The proposed CORN framework demonstrates superior performance compared to the individually trained baseline models. When using CORN for training, the no-reference (NR) model shows an improvement of around 16% over the independently trained NR model, suggesting that co-training stabilizes the NR model. Perhaps more remarkably, the full-reference (FR) model trained with CORN also exhibits a sizeable gain, surpassing the individually trained FR model by around 11%, despite the fact that the independent and co-trained FR models share the *exact* same architecture and training data. Thus, the co-training approach enhances the quality of each of the respective NR and FR models that were trained together.

Objectively, we also compare various metrics on (i) invariance to base model architecture; and (ii) invariance to target objective.

Invariance to base model To demonstrate the generality of our proposed framework, transcending reliance on any specific architectural paradigm, we effectuate a substitution of the base model $\mathcal{B}$ with an architecture introduced by Manocha et al. [18]. This alternative model employs a composite of both magnitude and phase spectra as input. Please refer to Table 1 for detailed insights. It becomes evident from the outcomes illustrated in Table 1 that the performance of this model aligns coherently with those derived from the model trained using the base model architecture described in Sec 3.1.

Invariance to choice of target objective In order to establish the universality of our proposed framework, devoid of any reliance on specific target objectives, we enact a substitution of the initial SI-SDR training objective with PESQ and SNR. For a fair comparison, given that SNR is confined to linear degradations, such as background noise, we adapt our approach to only introduce diverse background noise types during training and evaluation. For comprehensive insights, kindly refer to Table 1. The observations from the metrics presented therein harmonize with outcomes originating from the SI-SDR paradigm, underscoring its efficacy across various objectives. It can be noted that conducting training using the PESQ metric results in reduced error rates. This outcome may be attributed to the fact that PESQ has a scale ranging from 1 to 5, whereas SNR and SI-SDR exhibit a broader range spanning from -40 dB to +40 dB.

5.2. Evaluation of the embedding

Content invariance To assess the robustness of the system to content variations, we generate a test dataset encompassing two distinct groups: the first group comprises pairs of recordings characterized by identical background noise but different speech content, while the latter group comprises recordings featuring different noise and speech content. Embeddings e are extracted from the base model $\mathbf{B}$, and subsequently, the cosine similarity between recording pairs from both groups is computed. A Gaussian distribution is then fitted to the resultant samples drawn from these groups.

The calculation of the common region between these normalized Gaussian distributions is performed to gauge the extent of their overlap. A smaller common area is indicative of a more robust model. In CORN it is noteworthy that the common area is observed to be the lowest across both scenarios, namely with and without content variations (as delineated in Table 2).

Moreover, it is worth highlighting that a diminishing common area is associated with enhanced performance on the held-out dataset, as presented in Table 1. This finding suggests that the challenge of distinguishing between these two distribution groups may exert a significant influence on the process of acquiring a robust audio quality assessment metric.

Name	Invar. to content		Small signal shifts	
	FR	NR	FR	NR
Indiv.	0.5	0.8	1.4	2.0
CORN	**0.3**	**0.3**	**1.2**	**1.9**

Table 2: **Evaluations**: Refer to Sec 5.2. Models include: CORN and individual FR and NR prediction models. ↓ is better.

Small shifts in signal To evaluate the robustness to small (imperceptible) signal shifts, we create a test dataset of pairs of recordings with clean references and small noise-added signals. It is anticipated that the FR outputs shall closely approximate the highest attainable scores across a majority of the test instances. Conversely, in the context of NR scores, minimal disparities between the two input signals are sought, with the aim of these disparities approaching proximity to zero. Refer to Table 2. Here, we show the difference between the maximum score and the model FR outputs for the FR case and the magnitude difference of the respective scores for the NR case. In all cases, we see that our model has the lowest scores, showing that the model is robust to small, imperceptible changes.

Quality based retrieval: Here, we consider the outputs after the base model block as the quality embeddings, and use it for quality-based retrievals. Similar to Manocha et al. [18], we first create a test dataset of 1000 recordings at 10 discrete quality levels. We take randomly selected queries and calculate the number of correct class instances in the top K retrievals. We report the mean of this metric over all queries (MP^k). CORN gets $\text{MP}^{k=10} = 0.87$, as compared to $\text{MP}^{k=10} = 0.75$ for the FR model, and $\text{MP}^{k=10} = 0.80$ for the NR model. This suggests that our approach better clusters quality-level groups in this learned space.

6. CONCLUSIONS AND FUTURE WORK

This paper presents CORN – a novel approach that co-trains FR and NR models. We find that incorporating the NR loss during training assists the FR model in preventing over-generalization from the observed training content, thus enhancing its content-invariance. On the other hand, incorporating the FR loss during training assists the NR model by providing stable gradients during training.

In the future, we would like to apply this framework to a broader set of objectives and quality metrics. For example, we believe it would be valuable to collect a large dataset of human subjective ratings like MOS in a format suitable for training FR and NR models with such data. Likewise, the framework could be extended to learn from non-scalar data such as pairwise preference or triplet judgments.

7. REFERENCES

[1] R. C. Streijl, S. Winkler, et al., "Mean opinion score (MOS) revisited," *Multimedia Sys.*, vol. 22, pp. 213–227, 2016.

[2] A. W. Rix, J. G. Beerends, et al., "Perceptual evaluation of speech quality (PESQ)," in *ICASSP*, 2001, vol. 2, pp. 749–752.

[3] J. G. Beerends, C. Schmidmer, et al., "Perceptual objective listening quality assessment (POLQA)," *Journal of the AES*, 2013.

[4] A. Hines, J. Skoglund, et al., "ViSQOL: An objective speech quality model," *EURASIP*, vol. 2015, 2015.

[5] P. Manocha, A. Finkelstein, et al., "A differentiable perceptual audio metric learned from just noticeable differences," *Interspeech*, 2020.

[6] P. Manocha, Z. Jin, et al., "CDPAM: Contrastive learning for perceptual audio similarity," ICASSP 2021.

[7] P. Manocha, Z. Jin, et al., "Audio similarity is unreliable as a proxy for audio quality," in *Interspeech*, 2022.

[8] L. Malfait, J. Berger, et al., "P. 563—the itu-t standard for single-ended speech quality assessment," *IEEE TASLP*, vol. 14, no. 6, pp. 1924–1934, 2006.

[9] J. F. Santos, M. Senoussaoui, et al., "An improved non-intrusive intelligibility metric for noisy and reverberant speech," in *IEEE IWAENC*, 2014, pp. 55–59.

[10] C. K. Reddy, V. Gopal, et al., "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *ICASSP*, 2020.

[11] S.-W. Fu, Y. Tsao, et al., "Quality-Net: end-to-end non-intrusive speech quality assessment model," *Interspeech*, 2018.

[12] H. Gamper, C. K. Reddy, et al., "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *WASPAA*, 2019, pp. 85–89.

[13] M. Yu, C. Zhang, et al., "MetricNet: Improved modeling for non-intrusive speech quality assessment," *Interspeech*, 2021.

[14] Z. Zhang, P. Vyas, et al., "An end-to-end non-intrusive model for subjective and objective real-world speech assessment using a multi-task framework," in *ICASSP*, 2021, pp. 316–320.

[15] G. Mittag, B. Naderi, et al., "NISQA: A deep CNN-self-attention model for multidimensional speech quality prediction with crowdsourced datasets," in *Interspeech*, 2021.

[16] 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.

[17] A. A. Catellier and S. D. Voran, "Wawenets: A no-reference convolutional waveform-based approach to estimating narrow-band and wideband speech quality," in *ICASSP*, 2020.

[18] P. Manocha, B. Xu, et al., "NORESQA: A framework for speech quality assessment using non-matching references," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[19] S. Pascual, M. Ravanelli, et al., "Learning problem-agnostic speech representations from multiple self-supervised tasks," *arXiv preprint arXiv:1904.03416*, 2019.

[20] J. Le Roux, S. Wisdom, et al., "SDR–half-baked or well done?," in *ICASSP*. IEEE, 2019, pp. 626–630.

[21] T. Yuan, W. Deng, et al., "Signal-to-noise ratio: A robust distance metric for deep metric learning," in *CVPR*, 2019, pp. 4815–4824.

[22] S.-W. Fu, C.-F. Liao, et al., "MetricGAN: Generative adversarial networks based black-box metric scores optimization for speech enhancement," in *ICML*, 2019.

[23] J. Serrà, J. Pons, et al., "SESQA: semi-supervised learning for speech quality assessment," *arXiv preprint arXiv:2010*, 2020.

[24] M. Narwaria, W. Lin, et al., "Nonintrusive quality assessment of noise suppressed speech with mel-filtered energies and support vector regression," *IEEE TASLP*, vol. 20, no. 4, pp. 1217–1232.

[25] D. Sharma, L. Meredith, et al., "A non-intrusive pesq measure," in *GlobalSIP*. IEEE, 2014, pp. 975–978.

[26] A. H. Andersen, J. M. De Haan, et al., "Nonintrusive speech intelligibility prediction using convolutional neural networks," *IEEE/ACM TASLP*, vol. 26, pp. 1925–1939, 2018.

[27] C.-C. Lo, S.-W. Fu, et al., "MOSNet: Deep learning based objective assessment for voice conversion," *Interspeech*, 2019.

[28] B. Patton, Y. Agiomyrgiannakis, et al., "AutoMOS: Learning a non-intrusive assessor of naturalness-of-speech," *arXiv*, 2016.

[29] A. Kumar and V. Ithapu, "A sequential self teaching approach for improving generalization in sound event recognition," in *ICML*. PMLR, 2020, pp. 5447–5457.

[30] M. Li, X. Chen, et al., "The similarity metric," *IEEE transactions on Information Theory*, vol. 50, no. 12, pp. 3250–3264, 2004.

[31] S. Chen, B. Ma, et al., "On the similarity metric and the distance metric," *Theoretical Computer Science*, vol. 410, pp. 2365–2376, 2009.

[32] A. Tversky, "Features of similarity.," *Psychological review*, vol. 84, no. 4, pp. 327, 1977.

[33] J. Serrà, J. Pons, et al., "Sesqa: semi-supervised learning for speech quality assessment," in *ICASSP*, 2021, pp. 381–385.

[34] J. Su, Z. Jin, et al., "HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks," *Interspeech*, 2020.

[35] A. Defossez, G. Synnaeve, et al., "Real time speech enhancement in the waveform domain," in *INTERSPEECH*, 2020.

[36] R. E. Zezario, S.-W. Fu, et al., "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *arXiv preprint arXiv:2111.02363*, 2021.

[37] J. Su, Z. Jin, et al., "HiFi-GAN-2: Studio-quality speech enhancement via generative adversarial networks conditioned on acoustic features," in *WASPAA 2021*, Oct. 2021.

[38] M. Cernak and M. Rusko, "An evaluation of synthetic speech using the PESQ measure," in *Proc. European Congress on Acoustics*, 2005, pp. 2725–2728.

[39] J. G. Beerends, E. Larsen, et al., "Measurement of speech intelligibility based on the pesq approach," in *MESAQIN*, 2004.

[40] P. Paglierani and D. Petri, "Uncertainty evaluation of speech quality measurement in voip systems," in *2007 IEEE International Workshop on Advanced Methods for Uncertainty Estimation in Measurement*. IEEE, 2007, pp. 104–108.

[41] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE SPS*, vol. 22, no. 8, 2014.

[42] C. K. Reddy, E. Beyrami, et al., "The interspeech 2020 deep noise suppression challenge: Datasets, subjective speech quality and testing framework," in *INTERSPEECH*.

[43] J. S. Garofolo, L. F. Lamel, et al., "DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA Technical Report*, vol. 93, pp. 27403, 1993.

[44] K. J. Piczak, "ESC: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.