RHO-PERFECT: CORRELATION CEILING FOR SUBJECTIVE EVALUATION DATASETS

Fredrik Cumlin

School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Sweden

ABSTRACT

Subjective ratings contain inherent noise that limits the model-human correlation, but this reliability issue is rarely quantified. In this paper, we present ρ-Perfect, a practical estimation of the highest achievable correlation of a model on subjectively rated datasets. We define ρ-Perfect to be the correlation between a perfect predictor and human ratings, and derive an estimate of the value based on heteroscedastic noise scenarios, a common occurrence in subjectively rated datasets. We show that ρ-Perfect squared estimates test-retest correlation and use this to validate the estimate. We demonstrate the use of ρ-Perfect on a speech quality dataset and show how the measure can distinguish between model limitations and data quality issues.

***Index Terms*—** Subjective assessment, reliability measure, speech quality assessment, recommendation systems

1. INTRODUCTION

Developing objective emulators for subjective opinions is a rich research field with various applications. Examples include speech quality assessment [1], image aesthetics [2], and recommendation systems [3]. When developing these models, a fundamental question arises: what is the highest correlation any model can achieve with human ratings? Ratings contain inherent noise, which creates an upper bound on model-human correlation that is less than 1.0. Quantifying this bound is both challenging and important for effective model development.

Despite the widespread use of subjective ratings in ML evaluation, the reliability of the ratings is often overlooked. Recent surveys on speech and image aesthetics [1, 2, 4] extensively cover evaluation methodologies but omit reliability measures and correlation ceilings. Ignoring rating reliability might lead to misguided conclusions about model performance. For example, an objective model might perform well overall but poorly under certain conditions, but this could reflect poor rating reliability rather than model limitations.

Existing reliability measures such as Pearson's correlation ratio (η^2) [5], the family of intraclass correlations (ICC) [6], and Cronbach's alpha [7] have limitations in this setting, since they assume homoscedastic noise (equal noise variance across items) [8] or are difficult to interpret with respect to model performance. Subjectively rated datasets in, for example, speech quality assessment and recommendation systems typically have an uneven number of ratings per item, and different items can have larger or smaller disagreement among raters. Generalizability theory provides measures under heteroscedastic noise assumptions, but has a theoretical disposition and is not discussed with respect to model performance [9, 10], which might be the reason that they are not used in the context of model performance in these scenarios.

In this paper, we introduce ρ-Perfect, a practical upper bound of model-human correlation for subjectively rated datasets. The estimate can be calculated from a single evaluation knowing only the distribution of the individual ratings per item. Moreover, squaring ρ-Perfect approximates the correlation between two independent subjective evaluations, enabling empirical validation. In short, ρ-Perfect provides a principled way to interpret model performance relative to the limits of human ratings, enabling differentiation between model limitations and data quality issues. Our contributions are as follows: (1) ρ-Perfect, a practical upper bound on model-human correlation for unbalanced subjectively rated datasets; (2) a theoretical link to test-retest correlation and empirical validation thereof; and (3) case studies showing how ρ-Perfect estimates can separate model limitations from data quality.

2. THE ρ-PERFECT METRIC

Consider a subjectively rated dataset $\mathcal{D}$. It consists of items which we denote by $\mathcal{X} = \{x_i\}_{i=1}^{n}$ where n is the number of items. It also consists of a set of $m \in \mathbb{N}$ human raters that provide subjective ratings on the items. Since human annotation is costly, the most common rating strategy is to have $m_i \ll m$ raters provide with ratings on item x_i. Thus, a subjectively rated dataset consists of pairs $(x_i, r_i^{(j)})$, where x_i is an item and $r_i^{(j)} \in \mathbb{R}$ is a rating provided by rater (j). We define the average rating to be given by $y_i = \frac{1}{m_i} \sum_{j=1}^{m_i} r_i^{(j)}$.

2.1. Mathematical Derivation of ρ-Perfect

We first formalize the notion of a perfect predictor, which serves as the basis for defining ρ-Perfect.

Let X denote a random variable representing the items.

Let Y denote a random variable such that $Y|X = x_i$ represents the distribution of the average rating y_i.

We define the perfect predictor as

$$\hat{f} \triangleq \arg\min_f \mathbb{E}[(f(X) - Y)^2]. \tag{1}$$

It is a standard result that the minimizer is the regression function $\hat{f}(X) = \mathbb{E}[Y|X]$ [11, Th. 1.4.1]. For brevity, we denote $\hat{Y} \triangleq \hat{f}(X) = \mathbb{E}[Y|X]$.

We are interested in the Pearson Correlation Coefficient of the perfect predictor and the average ratings given. This can be viewed as a correlation ceiling. We have a Lemma.

Lemma 2.1. *Let X, Y be two random variables and $\hat{Y} = \mathbb{E}[Y|X]$. Then the correlation of $\hat{Y}$ and Y is given by*

$$\text{Corr}(Y, \hat{Y}) = \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y)\text{Var}(\hat{Y})}} = \sqrt{\frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}}. \tag{2}$$

Proof. If we show that $\text{Cov}(Y, \hat{Y}) = \text{Var}(\hat{Y})$, the second equality in Eq. 2 follows and we are done.

From the law of total expectation, we have

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[\hat{Y}].$$

Further, we have

$$\mathbb{E}[\hat{Y}Y] = \mathbb{E}[\mathbb{E}[Y\hat{Y}|X]] = \mathbb{E}[\hat{Y}\mathbb{E}[Y|X]] = \mathbb{E}[\hat{Y}^2].$$

It now follows that

$$\text{Cov}(Y, \hat{Y}) = \mathbb{E}[\hat{Y}Y] - \mathbb{E}[\hat{Y}]\mathbb{E}[Y] = \mathbb{E}[\hat{Y}^2] - \mathbb{E}[\hat{Y}]^2 = \text{Var}(\hat{Y}).$$

□

Estimating $\text{Var}(Y)$ can be done using the unbiased variance estimate given the data, and is given by

$$\text{Var}(Y) = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \tag{3}$$

where $\overline{y} = 1/n \sum_i y_i$. Estimating $\text{Var}(\hat{Y})$ is more difficult as neither the distribution nor the perfect predictor is accessible in closed form. We estimate it in the following way. From the law of total variance, conditioning on X, we have

$$\begin{aligned} \text{Var}(Y) &= \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X]) \\ &= \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\hat{Y}), \end{aligned} \tag{4}$$

hence, $\text{Var}(\hat{Y}) = \text{Var}(Y) - \mathbb{E}[\text{Var}(Y|X)]$. Now, if we assume that the raters are uncorrelated given an item, the variance of the average rating $Y|X = x_i$ can be estimated using the standard result for sample means; if individual ratings for item i have sample variance $s^2_{\text{rating}} = \frac{1}{m_i - 1}\sum(r_i^{(j)} - y_i)^2$, then:

$$\text{Var}(Y|X = x_i) = \frac{s^2_{\text{rating}}}{m_i} = \frac{1}{m_i(m_i - 1)}\sum_{j=1}^{m_i}(r_i^{(j)} - y_i)^2. \tag{5}$$

Therefore,

$$\mathbb{E}\left[\text{Var}(Y|X)\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i - 1)}\sum_{j=1}^{m_i}(r_i^{(j)} - y_i)^2. \tag{6}$$

We conclude with the definition of ρ-Perfect.

Definition 2.1 (ρ-Perfect)**.** *Given a subjectively rated dataset $\mathcal{D} = \left\{\left(x_i, r_i^{(j)}\right)\right\}$, the ρ-Perfect metric is given by*

$$\rho\text{-Perfect} \triangleq \sqrt{\frac{\text{Var}(\hat{Y})}{\text{Var}(Y)}} \tag{7}$$

where Var$(\hat{Y})$ is the variance of a perfect predictor, and Var(Y) is the variance of the average ratings per item. They are estimated by the following formulae:

$$\begin{aligned} \text{Var}(Y) &= \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2 \\ \text{Var}(\hat{Y}) &= \text{Var}(Y) - \frac{1}{n}\sum_{i=1}^{n}\frac{1}{m_i(m_i - 1)}\sum_{j=1}^{m_i}(r_i^{(j)} - y_i)^2 \end{aligned} \tag{8}$$

Formally, the definition coincides with Pearson's correlation ratio (η^2) but is extended to a heteroscedastic noise scenario [8].

From the definition of ρ-Perfect, we suggest that there should be at least 50 items (as it is correlation-based), and that each item has at least 3 ratings (as the variance of the average is estimated). The computational cost to estimate ρ-Perfect is $\mathcal{O}(M)$, where $M = \sum_i m_i$ is the total number of ratings.[1]

3. EXPERIMENTAL VALIDATION

Direct validation of ρ-Perfect is difficult since the distribution $Y|X$ is unknown. However, we can do indirect verification by the following.

Assume we have done two subjective evaluations on the same items. We model the evaluations by Y_1 and Y_2. Now, we assume that both evaluations yield the same perfect predictor, which is reasonable if the subjective evaluations are done using a similar group of raters with similar biases. This means $\hat{Y} = \mathbb{E}[Y_1|X] = \mathbb{E}[Y_2|X]$. Now, we have

$$\begin{aligned} \text{Cov}(Y_1, Y_2) &= \mathbb{E}[Y_1Y_2] - \mathbb{E}[Y_1]\mathbb{E}[Y_2] \\ &= \mathbb{E}[\mathbb{E}[Y_1Y_2|X]] - \mathbb{E}[\hat{Y}]^2 \\ &= \mathbb{E}[\text{Cov}(Y_1, Y_2|X) + \mathbb{E}[Y_1|X]\mathbb{E}[Y_2|X]] - \mathbb{E}[\hat{Y}]^2 \\ &= \underbrace{\mathbb{E}[\text{Cov}(Y_1, Y_2|X)]}_{\approx 0} + \mathbb{E}[\hat{Y}^2] - \mathbb{E}[\hat{Y}]^2 \\ &= \text{Var}(\hat{Y}) \end{aligned} \tag{9}$$

[1]Implementation of ρ-Perfect can be found at `https://github.com/fcumlin/rho-perfect`.

We justify $\mathbb{E}[\text{Cov}(Y_1, Y_2|X)] \approx 0$ empirically in Section 3.1 and provide intuition here. Given item X, the ratings thereof are the inherent noise of subjectivity, which could be perceptual differences (different views of the rating scale), personal differences (individual taste and standards), or simply noise in the human judgment (variability by external factors, such as time of day, when last meal occurred, etc.). Assuming the ratings are provided independently, the noises thereof are also expected to be independent. Note that $\mathbb{E}[\text{Cov}(Y_1, Y_2|X)] \approx 0$ is only needed to validate ρ-Perfect, not to calculate it.

If we assume that the two subjective evaluations are similar, we have $\text{Var}(Y_1) \approx \text{Var}(Y_2)$. It follows that

$$\text{Corr}(Y_1, Y_2) = \frac{\text{Var}(\hat{Y})}{\sqrt{\text{Var}(Y_1)\text{Var}(Y_2)}} \approx \frac{\text{Var}(\hat{Y})}{\text{Var}(Y_1)} = \rho\text{-Perfect}^2 \tag{10}$$

where ρ-Perfect^2 is the squared value of the ρ-Perfect measure using only the first subjective evaluation. This means that ρ-Perfect^2 is an estimate of the correlation between two similar subjective evaluations on the same items. Note that ρ-Perfect^2 is measured using **only** the first evaluation.

We conclude that if $\mathbb{E}[\text{Cov}(Y_1, Y_2|X)] \approx 0$, ρ-Perfect is an appropriate measure of the expected PCC of a perfect predictor on the data, if ρ-Perfect^2 is equal to the PCC of two similar subjective evaluations on the same data.

3.1. Validating ρ-Perfect on real datasets

We now validate ρ-Perfect on real datasets. To the author's knowledge, there are no subjectively rated datasets publicly available that have been rated twice. Thus, we simulate two subjective assessments from one assessment. We propose two methods. The first method is to extract all the raters and split them into two equally sized sets. The scores from the raters in the first set constitute the first assessment, and the scores from the raters in the second set constitute the second assessment. We call this simulation **Split-Raters**. The other method is to split the ratings per item into two equally sized sets. We call this simulation **Split-Ratings**. Note that for the second method, some raters might have been provided with ratings in both assessments. However, the second method ensures that we have an equal number of ratings per item.

The datasets used are BVCC, MovieLens, SOMOS, and MERP.

BVCC: The BVCC dataset consists of $4,974$ speech clips with exactly 8 speech quality ratings per clip [12]. The ratings are given on a discrete scale from 1 to 5.

MovieLens: The MovieLens dataset is a movie dataset that human raters have provided with recommendation ratings [13]. It consists of $1,349$ movies with an average of 74 ratings per movie; the least rated movie has 5 ratings, and the most rated movie has 583 ratings. The ratings are given on a discrete scale from 1 to 5.

SOMOS: The SOMOS dataset is a speech dataset that human raters have annotated based on speech quality [14]. It consists of $20,100$ speech clips with an average of 18 ratings per speech clip; the least rated clip has 17 ratings, and the most rated clip has 146 ratings. The ratings are given on a discrete scale from 1 to 5.

MERP: The MERP dataset is a music dataset that humans have annotated based on emotional experience [15]. In this experiment, we only extract the 'arousal' ratings, which are ratings based on the intensity of emotion, a continuous value from -1 to 1. For a song, a rater provides one rating per second. We consider the average of these ratings to be the arousal rating of the song. There are 60 songs in total, with an average of 57 ratings per song; the least rated song has 6 ratings, and the most rated song has 100 ratings.

We do 10 iterations with different seeds when splitting the raters/ratings, and report the mean and standard deviation. The results can be seen in Table 1.

Dataset	$\mathbb{E}[\text{Cov}(Y_1, Y_2 \mid X)]$	ρ-Perfect^2	Target $\text{Corr}(Y_1, Y_2)$
Split-Raters			
BVCC	0.0^*	$0.798_{\pm 0.001}$	$0.801_{\pm 0.001}$
MovieLens	0.0^*	$0.734_{\pm 0.001}$	$0.728_{\pm 0.001}$
SOMOS	0.0^*	$0.258_{\pm 0.002}$	$0.297_{\pm 0.001}$
MERP	0.0^*	$0.499_{\pm 0.020}$	$0.502_{\pm 0.008}$
Split-Ratings			
BVCC	0.0^*	$0.800_{\pm 0.001}$	$0.800_{\pm 0.001}$
MovieLens	0.0^*	$0.710_{\pm 0.001}$	$0.701_{\pm 0.001}$
SOMOS	0.0^*	$0.281_{\pm 0.001}$	$0.281_{\pm 0.001}$
MERP	0.0^*	$0.478_{\pm 0.009}$	$0.502_{\pm 0.007}$

Table 1. Comparison of the squared ρ-Perfect measure of a subjectively rated dataset and the correlation of the subjectively rated dataset to a similar one. Validation of the ρ-Perfect metric is done by comparison of the squared value to the correlation. * All values below 10^{-18}; effectively zero within numerical precision.

As can be seen in Table 1, the term $\mathbb{E}[\text{Cov}(Y_1, Y_2 \mid X)]$ is effectively zero, hence justified by both empirical investigation and intuitive reasoning. This means that ρ-Perfect squared should be a good estimator of the correlation between two subjective ratings. In both split-raters and split-ratings methods, we find that ρ-Perfect squared is a suitable estimator of this correlation. This validates the use of ρ-Perfect as an estimator of the correlation between a subjectively rated dataset and a perfect predictor thereof; informally, the highest achievable correlation on that dataset.

3.2. Comparison to existing measures

We now compare with existing reliability measures. We report the test-retest experimental correlation of the split-raters

Table 2. Comparison of ρ-Perfect with existing reliability metrics on subjective datasets

Dataset	$\mathrm{Corr}(Y_1, Y_2)$	ICC(2, k)	Subsampling reliability	ρ-Perfect^2
BVCC	$0.801_{\pm 0.001}$	$0.822_{\pm 0.001}$	$0.893_{\pm 0.001}$	$0.796_{\pm 0.001}$
MovieLens	$0.728_{\pm 0.001}$	$0.898_{\pm 0.001}$	$0.879_{\pm 0.001}$	$0.719_{\pm 0.001}$
SOMOS	$0.297_{\pm 0.002}$	$0.326_{\pm 0.001}$	$0.716_{\pm 0.001}$	$0.269_{\pm 0.001}$
MERP	$0.502_{\pm 0.010}$	$0.554_{\pm 0.001}$	$0.807_{\pm 0.001}$	$0.483_{\pm 0.011}$

methodology for comparison. Note that this means we compute the reliability of 'half' the datasets; only the first evaluation is used for reliability computation.

ICC(2, k): Following the guideline in Table 3 and Figure 1 in [16], we implement ICC(2, k), which measures test-retest reliability of the average ratings (k ratings per item) in a two-way random effects model with absolute agreement. The measure assumes that for an item, only a subset of the raters rate the item, and that there is an equal number of raters per item. The last assumption is violated for most of the datasets we test on, which ρ-Perfect is designed to address.

Subsampling reliability: The subsampling reliability measure was used in [17] and [14] for measuring the reliability of ratings in the speech quality assessment task. It is done by randomly drawing half of the raters from the full rater pool and computing the correlation between the average ratings of the two sets. This is done over several iterations, and the average correlation is reported as the reliability.

As can be seen in Table 2, ICC(2, k) and ρ-Perfect^2 generally agree, except for MovieLens, where ICC(2, k) counterintuitively suggests higher reliability than BVCC despite lower rater test-retest correlation. A likely culprit is that MovieLens has a large, varying number of raters per clip, hence violating the assumptions of the ICC(2, k) measure. ρ-Perfect does not suffer from this limitation.

The subsampling reliability measure seems to consistently overestimate the reliability, but follows the overall trend of reliability when studying the test-retest correlation. A likely reason for the overestimation is that the ratings from the subset are already present in all the ratings, resulting in dependent realisations which does not reflect a truly independent evaluation.

4. INTERPRETING MODEL PERFORMANCE WITH ρ-PERFECT

To demonstrate practical utility, we evaluate the objective speech quality model DNSMOS Pro [18] on the NISQA validation dataset [19]. DNSMOS Pro predicts the overall speech quality given a speech clip, and the NISQA dataset consists of 2500 speech clips, each of which is rated by 5 raters according to the overall quality. Extracting subsets of the dataset and computing model performance can give insights into problematic speech distortion scenarios for DNSMOS Pro.

Table 3. Model performance relative to ρ-Perfect upper bounds on NISQA validation data and different subsets.

Condition	Model PCC	ρ-Perfect
All data	0.873	0.954
Bandpass filtered	0.934	0.969
Clean conditions	0.621	0.816
Bursty distortions	0.392	0.701

The results are shown in Table 3. The PCC on all data is 0.873 with a ρ-Perfect of 0.954, which suggests strong model performance and reliable ratings. However, the performance varies significantly for the different subsets presented. For bandpass filtered clips, the PCC is high, close to ρ-Perfect. For clean clips and high-burst distorted clips, the model PCC is significantly poorer (0.621 and 0.392, respectively), and ρ-Perfect suggests that the reliability of the data is also lower. For the clean condition, part of the performance degradation can be explained by a lower ρ-Perfect, but this could also constitute an improvement area for DNSMOS Pro. For the high-burst scenario, while the PCC is the lowest (0.392), the moderate ρ-Perfect (0.701) indicates that both model improvements and potentially more reliable subjective evaluation will benefit a more accurate diagnosis.

This analysis demonstrates a practical use case of ρ-Perfect to distinguish between dataset limitations and model deficiencies. When model correlation drops, ρ-Perfect can aid in determining whether this reflects model weaknesses or limitations in subjective evaluation reliability.

5. CONCLUSION

In this paper, we have derived ρ-Perfect, an estimation of the model-human correlation ceiling. We have shown that ρ-Perfect squared approximates the correlation between two independently, but similar, subjective evaluations. We have shown that this theoretical result holds on real datasets across different subjectively rated datasets and different reliabilities of the ratings thereof. We have compared ρ-Perfect to other reliability measures and shown that ρ-Perfect has a clear interpretation. Finally, we have shown how to use ρ-Perfect in practice by comparing the measure to model-human correlation on the NISQA dataset.

6. ACKNOWLEDGEMENT

This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

7. REFERENCES

[1] Erica Cooper, Wen-Chin Huang, Yu Tsao, Hsin-Min Wang, Tomoki Toda, and Junichi Yamagishi, "A review on subjective and objective evaluation of synthetic speech," *Acoustical Science and Technology*, vol. 45, no. 4, pp. 161–183, 2024.

[2] Maedeh Daryanavard Chounchenani, Asadollah Shahbahrami, Reza Hassanpour, and Georgi Gaydadjiev, "Deep learning based image aesthetic quality assessment- a review," *ACM Comput. Surv.*, vol. 57, no. 7, Feb. 2025.

[3] Bushra Alhijawi, Arafat Awajan, and Salam Fraihat, "Survey on the objectives of recommender systems: Measures, solutions, evaluation methodology, and new perspectives," *ACM Comput. Surv.*, vol. 55, no. 5, Dec. 2022.

[4] Gabriel Mittag, *Deep Learning Based Speech Quality Prediction*, T-Labs Series in Telecommunication Services. Springer, 2022.

[5] Karl Pearson, *On the General Theory of Skew Correlation and Non-linear Regression*, Dulau and Co., 1905.

[6] Patrick E Shrout and Joseph L Fleiss, "Intraclass correlations: uses in assessing rater reliability," *Psychological Bulletin*, vol. 86, no. 2, pp. 420–428, 1979.

[7] Lee J. Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika*, vol. 16, no. 3, pp. 297–334, 1951.

[8] Frederic M. Lord and Melvin R. Novick, *Statistical theories of mental test scores*, Addison-Wesley Publishing Company, Reading, Massachusetts, 1968, With contributions by Allan Birnbaum.

[9] Robert L. Brennan, *Generalizability Theory*, Statistics for Social and Behavioral Sciences. Springer-Verlag, New York, 2001.

[10] Seonghoon Kim and Robert L. Brennan, "A note on the reliability coefficients for item response model-based ability estimates," Technical report, Iowa Testing Programs, University of Iowa, Iowa City, IA, 2007.

[11] Peter J. Bickel and Kjell A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I, Second Edition*, Chapman & Hall / CRC Texts in Statistical Science, Boca Raton, FL, 2 edition, 2015.

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

[13] F. Maxwell Harper and Joseph A. Konstan, "The movielens datasets: History and context," *ACM Trans. Interact. Intell. Syst.*, vol. 5, no. 4, Dec. 2015.

[14] Georgia Maniati and Alexandra Vioni and Nikolaos Ellinas and Karolos Nikitaras and Konstantinos Klapsas and June Sig Sung and Gunu Jho and Aimilios Chalamandaris and Pirros Tsiakoulis, "SOMOS: The Samsung Open MOS Dataset for the Evaluation of Neural Text-to-Speech Synthesis," in *Interspeech 2022*, 2022, pp. 2388–2392.

[15] En Yan Koh, Kin Wai Cheuk, Kwan Yee Heung, Kat R Agres, and Dorien Herremans, "Merp: A music dataset with emotion ratings and raters' profile information," *Sensors*, vol. 23, no. 1, pp. 382, 2023.

[16] Terry K Koo and Mae Y Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *Journal of Chiropractic Medicine*, vol. 15, no. 2, pp. 155–163, 2016.

[17] 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 *Interspeech 2019*, 09 2019, pp. 1541–1545.

[18] Fredrik Cumlin, Xinyu Liang, Victor Ungureanu, Chandan KA Reddy, Christian Schüldt, and Saikat Chatterjee, "Dnsmos pro: A reduced-size dnn for probabilistic mos of speech," in *Proc. Interspeech 2024*, 2024, pp. 4818–4822.

[19] 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 *Interspeech 2021*. Aug 2021, ISCA.