Audio Engineering Society

Convention Paper 10580

Presented at the 152nd Convention
2022 May, In-Person and Online

Assessing the relevance of perceptually driven objective metrics in the presence of handling noise

Céline Angonin[1], Emmanouil Theofanis Chourdakis[1], and Ruben Andre Åeng[1]

[1]*Nomono AS*

Correspondence should be addressed to Céline Angonin (celine.angonin@nomono.co)

ABSTRACT

This paper examines how perceptually driven objective metrics found in the speech enhancement and separation literature react when adding handling noise to speech corrupted with environmental noise. Identifying sensitive metrics will inform us which metrics are appropriate for the development or evaluation of speech enhancement techniques when dealing with handling noise. Using an in-house synthetic dataset and paired sample tests, we examine how nine different perceptual metrics behave on audio mixtures containing both handling and background noise. We show that eight of them react to handling noise but only when the handling to background noise power ratio is over a specific threshold which we identify using logistic regression.

1 Introduction

As its name suggests, handling noise is extraneous sound introduced to a microphone recording due to inappropriate microphone handling. Examples of handling noise include unintentionally tapping on the microphone mesh or rubbing on the user's clothes in the case of lapel microphones. In various ways, it differs from sound caused by the user's environment, referred to as background or environmental noise. Handling noise tends to be very short in duration, spanning a couple of milliseconds. Environmental noise will have a more varied duration, from a couple of milliseconds if caused by individual events in the user's surroundings to spanning the whole recording. Additionally, it might be desirable to keep environmental sound in the recording as ambience while handling noise is usually undesired and should be removed.

Speech Enhancement and Separation methods exist to improve the quality of a speech recorded in a noisy environment. Those usually are developed by minimizing or maximizing an objective measure, called a metric, assuming that a significant change in this metric corresponds to a significant change in the perception of the quality of the enhanced speech. As humans, we consider handling noise to degrade the audio quality more than environmental noise when listening to podcasts [1]. Therefore, we would like metrics used in speech enhancement to reflect that perceptual observation. This paper assesses the ability of various metrics found in the speech enhancement and separation literature to react to the addition of handling noise and quantify the extent of this ability. The findings will tell us whether the currently used metrics are good enough for evaluating enhancement methods that consider handling noise, which is a first step towards developing new, more appropriate metrics and methods for speech enhancement. For this reason, we use paired sample tests to examine several well-known metrics on a synthetic dataset. For the metrics that show differences in

Metric	Val.	Eval.	Optim.
PEAQ [2]	$[-4,0]\uparrow$	[3]	-
SDR [4]	$(-\infty,\infty)\uparrow$	[5, 6]	-
WB-PESQ [7]	$[0,5]\uparrow$	[5, 6] [8–12]	[6, 10] [13]
CSIG [14]	$[1,5]\uparrow$	[11]	-
CBAK [14]	$[1,5]\uparrow$	[11]	-
COVL [14]	$[1,5]\uparrow$	[11]	-
STOI [15]	$[0,100]\uparrow$	[8–10] [16]	[10, 13]
VISQOL [17]	$[1,5]\uparrow$	[12]	-
MOSNET [18]	$[1,5]\uparrow$	[19]	-
CDPAM [20]	$[0,1]\downarrow$	[20]	[20]

Table 1: Table of metrics and information about them. The Val. column indicates the range of the metric, and the arrow next to them indicates whether a higher (↑) or a lower (↓) value indicates better quality. Eval. shows works where the metric has been used for evaluation and Optim. shows works where the metric has been used in the optimization process.

the presence of handling noise, we calculate the threshold the handling to background noise power ratio must surpass for handling noise to be detectable.

The rest of the paper is structured as follows: Section 2 introduces the metrics examined in this work. Section 3 gives a summary of existing work. Section 4 describes the method followed for assessing the metrics. Section 5 discusses results. Sections 6 and 7 discuss the findings and limitations of this study and provide directions for future work. Finally, Section 8 concludes the paper with a summary.

2 Objective metrics for speech enhancement and separation

Speech separation and enhancement methods use metrics for two purposes: to evaluate the quality of an algorithm or to integrate them into the optimization process when training a machine learning model. Table 1 presents the metrics employed in separation and enhancement tasks and which are designed to correlate to human perception. Perceptual Evaluation of Audio Quality (PEAQ) [2] was introduced as a standard for the rapid evaluation of codecs. It tries to predict the difference in quality according to the ITU-R five-grade impairment scale between the degraded speech signal and a clean reference. Since the impairment scale takes values between 1 and 5, PEAQ gives values between 0 and -4. It is also suitable for 48KHz sampled audio. Signal-to-Distortion Ratio (SDR) [4] was introduced in the Blind Source Separation Evaluation (BSS_EVAL) package of speech quality metrics for source separation and measures the energy of the clean signal versus the energy of the introduced distortions. While it has a low correlation with perception [21], it is still useful as a baseline. In this work, the only distortions introduced come from the addition of noise, and therefore its definition is equal to Signal-to-Noise Ratio (SNR). Sometimes, variants of SNR (SSNR, SI-SNR) [12] are also reported. Perceptual Evaluation of Speech Quality (PESQ) [22] is a perception-driven metric initially developed for assessing speech quality in telephone systems. However, it has also been used to evaluate denoising [5, 8, 10, 11]. Here, Wide-Band PESQ (WB-PESQ) [7], an extension of PESQ to 16KHz, is used instead. CSIG, CBAK, and COVL [14] are composite metrics created by predicting a value for signal distortion (SIG), background intrusiveness (BAK), and overall quality and weighting each by the value given by PESQ. Short-Time Objective Intelligibility (STOI) [15] is a metric predicting intelligibility in audio recordings. The Virtual Speech Quality Objective Listener (VISQOL) [17] is a metric designed to assess quality in Voice-over-IP communication. MOSNET [18] is a neural-network architecture for predicting speech quality on a Mean-Opinion-Score (MOS) rating scale. Finally, CDPAM [20] improves upon DPAM [23], a neural network trained on just noticeable differences which can be used as a loss function when optimizing a neural network.

3 Related Work

In this section, we list previous studies connecting handling noise to perception. In [24], the authors use multiple linear regression to model perceived audio quality scores from listening tests, using a number of factors. They found that handling noise is the third most important factor for perceiving changes in audio quality after the presence of environmental and wind noise. The authors in [12] calculate Pearson's correlation coefficient between MOS scores and various metrics, including PESQ, POLQA [25], VISQOL, and SNR. They find that SNR has the lowest correlation (0.56) and POLQA has the highest (0.78). They also specify that these

correlation values are too low to consider that these metrics are a good representation of the perceptual impression. In [1], the authors added rubbing and tapping handling noise to three different podcasts and conducted an internet-based psychoacoustic experiment to quantify how such noise affects the perception of audio quality. They found that an A-weighted SNR, with the signal level calculated across the whole signal, correlated best with perceived changes in audio quality. They also calculated the threshold where 50% of the experiment's subjects could perceive a change in audio quality at $4.2 \pm 0.2 dBA$. Handling noise has also been the objective of event detection and denoising work. In [26], the authors train a decision tree classifier with MFCC features to identify handling noise. However, their target classes are handling noise levels, and their evaluation was done solely on classification performance with no reported audio quality metrics. In [16], the authors develop methods based on Recurrent Neural Networks for removing microphone rustle noise from recordings, and they evaluate those methods using SDR and STOI.

4 Method

We aim at evaluating the addition of handling noise on the scores of the audio quality metrics in Table 1. Therefore, we calculated these scores for clean speech corrupted with background noise and clean speech corrupted with both background noise and handling noise. Then, we compared the scores between the two cases to see whether they were statistically different under various conditions and precisely determined those conditions.

Using AUDIOMENTATIONS[27], 1536 noisy mixtures were created by adding handling and background noise to clean speech at various SNRs. Half of these mixtures are speech corrupted with background noise, and the other half are the same files with handling noise added. We used 1280 mixtures for the analysis and held out 256 for validation. Each mixture has a sampling rate of 48KHz, which is the standard for broadcasting. The parameters sampled when generating the mixtures are the background and handling noise categories, the clean speech-to-background noise ratio $SNR(s, bn)$, and the ratio $SNR(s+bn, hn)$, which measures the level of the background noise added to the clean speech, over the level of the handling noise. Those parameters can be seen in Table 2. $SNR(s, bn)$ ranges from -6 to 35 dB.

Parameter	Domain
Background noise category	*{cafe, fan}*
Handling noise category	*{tapping, rustle}*
$SNR(s, bn)$ (dB)	$[-6, 35]$
$SNR(s+bn, hn)$ (dB)	$[-6, 12]$

Table 2: Parameters used to generate the dataset. Each mixture in the dataset has been generated by sampling uniformly from the domain of those parameters. The handling noise category parameter is sampled only when handling noise exists in that mixture.

An $SNR(s, bn)$ of -6 dB corresponds to a recording in a very noisy environment, and 35 dB is typical for a recording studio. $SNR(s+bn, hn)$ of -6 and 12 dB represent respectively very loud and almost inaudible handling noise. Clean speech was taken from the DNS Challenge 2021 full band dataset [28]. We only picked files that did not contain too much silence, to reflect realistic conditions and because the metric scores cannot be computed if the target audio is silent. Background noise was sourced from DEMAND [29] and FREESOUND[1]. The majority of handling noise files were sourced from the dataset in [1]. We discarded four of them because of the presence of background noise. The rest of the files came from in-house recordings of handling noise. Handling noise can be very diverse in terms of duration and spectral content. We aimed to cover this diversity as much as possible by selecting handling noise using different microphones and cloth types. We added handling noise every 1 to 2 seconds in the noisy mixtures. We selected 7 female and 10 male clean speech extracts as well as 5 cafe and 4 fan background noise files. For handling noise, we used 96 tapping and 170 rustle noise files. Among them, 5 tapping and 72 rustle extracts come from in-house recordings.

To compute SDR, STOI, WBPESQ, and MOSNET, we use the SPEECHMETRICS library for PYTHON[2]. For CSIG, CBAK and COVL, we use the SEMETRICS PYTHON library [3]. For VISQOL[4] and CDPAM[5], we

[1] https://www.freesound.org
[2] https://github.com/aliutkus/speechmetrics
[3] https://github.com/usimarit/semetrics
[4] https://github.com/google/visqol
[5] https://github.com/pranaymanocha/PerceptualAudio

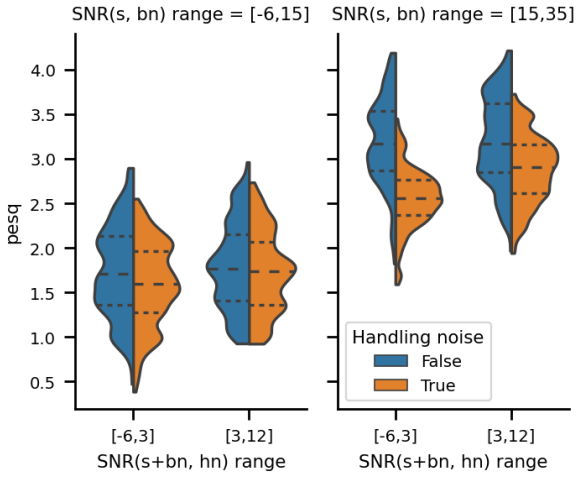
use the official implementations. For PEAQ, we use the implementation of GSTPEAQ [30]. Although the latter does not pass conformance testing, it is sufficiently close to the official implementation.

5 Results

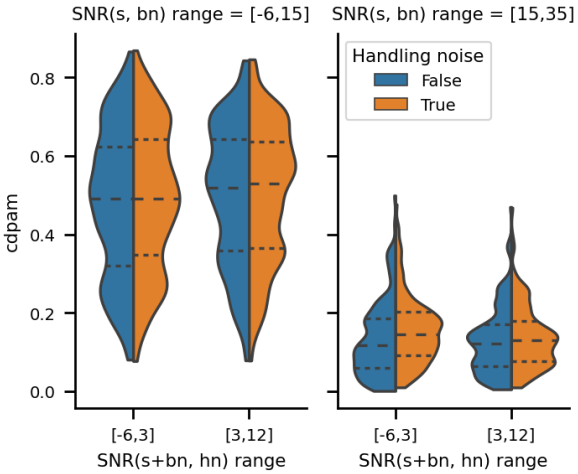
We computed the metrics from Section 4 for files containing speech and background noise and the same files with added handling noise. Figure 1 shows violin plots of the distribution of PESQ and CDPAM scores for two ranges of $SNR(s,bn)$ and $SNR(s+bn,hn)$. We observe that PESQ and CDPAM display very different behavior. As seen in Figure 1a, the PESQ scores are affected by the presence of handling noise, especially for low background noise levels, corresponding to a high $SNR(s,bn)$. We also observe that for the higher $SNR(s,bn)$ range, the quartiles of the PESQ scores distributions are more disjointed when $SNR(s+bn,hn)$ values are low. That suggests that PESQ is more sensitive to louder handling noise. However, this observation does not apply to the lower $SNR(s,bn)$ range. These findings also stand for all the other metrics, except CDPAM. Figure 1b indeed shows that the distribution of the CDPAM scores seems similar independent of the presence of handling noise for any range of $SNR(s+bn,hn)$ or $SNR(s,bn)$. This suggests that this metric is not very sensitive to the addition of handling noise. A Z-test, as well as a Wilcox Rank Sum test, verify that the difference in CDPAM is indeed zero in both cases ($p < \alpha = 0.05$). We observe that, for STOI, CSIG, CBAK, COVL, PESQ, PEAQ, VISQOL, and MOSNET, the differences increase with the increase of the Signal-to-Background noise ratio. We refer to those differences by ΔM, where M is one of the above metric names. Such an increase is not surprising; if we take SDR as an example, its difference is given by:

$$\begin{aligned}\Delta SDR &= \Delta SNR \\ &= \log\frac{P_s}{P_{bn}} - \log\frac{P_s}{P_{bn+hn}} \\ &= \log P_s - \log P_{bn} - \log P_s + \log P_{bn+hn} \\ &= \log\frac{P_{bn+hn}}{P_{bn}} \end{aligned} \tag{1}$$

where P_s, P_{bn}, and P_{bn+hn} are the powers of speech, background noise, and the combination of background noise with handling noise, respectively. We observe



(a) PESQ



(b) CDPAM

Fig. 1: Split violin plots representing respectively the distribution of PESQ and CDPAM scores with and without handling noise for different $SNR(s,bn)$ and $SNR(s+bn,hn)$ ranges. Both SNRs are in dB. The dashed lines in the violin plots represent the quartiles of the underlying distribution.

that the difference ΔSDR depends on the power of the background noise and the power of the combination of background noise and handling noise. When we increase $SNR(s,bn)$, we either increase the speech power or decrease the background noise power. The first case does not affect ΔSDR, as seen from Eq. 1. When

the background noise power decreases, the ratio in Eq. 1 increases, asymptotically diverging to infinity near zero. We have found that most of the perceptual metric differences correlate with ΔSDR (Pearson's correlation coefficient with ΔPEAQ : 0.77, ΔWB-PESQ : 0.86, ΔCSIG : 0.78, ΔCBAK : 0.90, ΔCOVL : 0.84, ΔSTOI : 0.54, ΔVISQOL : 0.68, ΔMOSNET : 0.57) and therefore expect this phenomenon to appear to a degree for those as well. Given the above observation and Eq. 1, it makes sense to figure out a relation between P_{bn} and P_{hn} where, for each metric, we detect an observable difference.

We arbitrarily define *observable* to mean a score difference of more than 5%. While this threshold should be more clearly chosen to reflect perception using listening tests, we leave this for future work. We use the constant $thr_{M,5\%}$ for each of the thresholds, where M is a metric name. The exact values are presented in Table 3. We calculate the power ratio of the handling noise over the background noise that we would need to have an observable difference for each metric as a Noise-to-Noise (NNR) ratio $NNR_{M,type} = \log \frac{P_{hn}}{P_{bn}}$ for metric M and noise type $type$. To find NNR, we fit a logistic regression model for each metric:

$$\log\left(\frac{p}{1-p}\right) = c_0 + c_1 \cdot NNR_{M,type} \tag{2}$$

where c_0 and c_1 are model coefficients, and p is the probability of detecting a 5% increase. If we set $p = 0.5$ (the probability of detecting a 5% increase, half of the time), we get a threshold for NNR:

$$NNR_{M,type} = -\frac{c_0}{c_1} \tag{3}$$

The coefficients and thresholds can be seen in Tables 4 and 5 for tapping and rustle noise. We also validate our results by showing the accuracy of detecting a 5% increase in each metric for a held-out dataset of 128 mixture pairs in Table 6. The handling noise dataset and the data used to calculate the NNR thresholds are provided in [31].

6 Discussion

Our experiment found that eight out of the nine metrics under consideration showed sensitivity to handling noise addition to various degrees. The ones that exhibited such a sensitivity displayed it only when the power ratio of the handling noise over the background noise exceeded the thresholds calculated in Section 5. We can also examine this finding in a denoising context: calculating the score of each metric as we did in Section 4 is equivalent to evaluating a very naive denoiser that does not remove any noise. If we use a metric not sensitive to handling noise, this hypothetical denoiser will be rated similarly irrespective of the presence or absence of handling noise. Since it lets more interference pass through when handling noise is present, we would want it to be rated worse. This indifference to handling noise hinders a correct evaluation of this denoiser's performance. We expect to observe the same behavior for more practical denoisers as well. However, this requires a study that focuses on denoising. In addition,

Metric diff.	$thr_{M,5\%}$
ΔSDR	1
ΔPEAQ	0.2
ΔWB-PESQ	0.25
ΔC{SIG, BAK, OVL}	0.2
ΔSTOI	5
ΔVISQOL	0.2
ΔMOSNET	0.2

Table 3: Thresholds for a 5% value difference for all metrics except ΔSDR. For ΔSDR, since it can take any value from $(-\infty, \infty)$, we arbitrarily set that threshold to 1 dB.

Metric diff.	c_0	c_1	$NNR_{M,tap}$
ΔSDR	−2.2304	0.7593	2.94
ΔPEAQ	−4.3046	0.2153	20.00
ΔWB-PESQ	−4.7410	0.2710	17.49
ΔCSIG	−4.2157	0.2550	16.53
ΔCBAK	−5.8651	0.3409	17.21
ΔCOVL	−4.0746	0.2715	15.01
ΔSTOI	−3.8453	0.0847	45.39
ΔVISQOL	−5.6277	0.1625	34.63
ΔMOSNET	−2.4630	0.0872	28.23

Table 4: Background-to-Tapping noise ratios $NNR_{M,tap}$ in dB needed for the perceptually-derived metrics to show a difference. c_0 and c_1 are the coefficients for the intercept and variable of the linear regression model. ΔSDR is included for comparison.

Metric diff.	c_0	c_1	$NNR_{M,rustle}$
ΔSDR	-	-	-
ΔPEAQ	−4.6376	0.2821	16.44
ΔWB-PESQ	−2.5935	0.2106	12.32
ΔCSIG	−0.1802	0.2201	0.82
ΔCBAK	−2.8434	0.4072	6.98
ΔCOVL	−0.9172	0.2567	3.57
ΔSTOI	−1.8389	0.0921	19.97
ΔVISQOL	−1.4927	0.1610	9.27
ΔMOSNET	−1.4147	0.1276	11.08

Table 5: Background-to-Rustle noise ratios $NNR_{M,rustle}$ in dB needed for the perceptually-derived metrics to show a difference. c_0 and c_1 are the coefficients for the intercept and variable of the linear regression model. SDR shows a 5% increase for rustle noise irrespective of the NNR value.

Metric diff.	Tapping Noise	Rustle Noise
ΔPEAQ	0.81	0.91
ΔWB-PESQ	0.91	0.92
ΔCSIG	0.94	0.89
ΔCBAK	0.94	0.98
ΔCOVL	0.97	0.88
ΔSTOI	0.97	0.69
ΔVISQOL	0.92	0.88
ΔMOSNET	0.73	0.78

Table 6: Accuracies of chosen thresholds in detecting a 5% difference in each respective metric for each noise type.

this sensitivity could be due either to the slight increase in noise power as a result of handling noise addition or to the presence of handling noise itself. We did not quantify the influence of each of these observations on the metrics scores in our article and leave it for future work.

We saw that CDPAM did not display a significant difference which may happen because it has been trained to detect just noticeable differences and then fails to evaluate files that are too noisy. The NNR thresholds we derived for the other metrics are lower for rustle noise than for tapping noise. We attribute this to the fact that rustle noise events last longer than tapping noise events, so they disturb the audio more and give worse results. We also observed that these thresholds are similar for the two background noise categories, suggesting that the background noise category does not have a significant influence. However, this finding needs further investigation as only two categories have been considered. Finally, Figure 1 hints that the metrics are more sensitive to handling noise addition for a high speech to background noise ratio.

7 Future work

In [32], the authors examine correlations between STOI, PEAQ, speech intelligibility, and perceived quality of speech enhancement, using a listening test focused on speech recognition. They found that neither STOI nor PEAQ correlate with the listening test results. Similarly, [33] tested correlations between listening test scores and a large number of state-of-the-art objective metrics in the context of musical signal source separation. They also did not find correlations with listening scores, except for a new metric derived from PEAQ output variables. Possible future work could include replicating such procedures for the findings in Section 5 in order for them to be validated perceptually. Such a study will allow us to confirm the NNR thresholds we found using logistic regression, and evaluate whether the background noise category influences this threshold and the degree of that influence. It would also be helpful to examine how our thresholds correlate with the signal-to-handling noise ratio threshold of $4.2dB$ found in [1]. A listening test will also determine if rustle noise is perceptually worse than tapping noise, as hinted in our experiment, and if handling noise is indeed more detectable for high speech-to-background noise ratio. Furthermore, future work could vary the total duration of handling noise to investigate its influence on the results and could also include extra background noise categories.

8 Summary

This work examined the suspicion that metrics developed to emulate perception can be overwhelmed by background noise in speech recordings and cannot register the presence of handling noise. We examined nine different perceptually-driven metrics whether that was the case: PEAQ, WB-PESQ, CSIG, CBAK, COVL, CSTOI, VISQOL, MOSNET, and CDPAM. We found that, apart from CDPAM, the rest could register the

presence of handling noise. However, this ability depends on the power ratio of the handling noise over the background noise. Logistic regression found the thresholds of the ratios over which handling noise leads to an observable difference in the metrics, which is different between Tapping and Rustle noise. We observe that the ratios for tapping noise are much higher than the ratios for rustle noise.

9 Acknowledgements

We thank the Research Council of Norway (RCN) for their funding through project '321604 - ANGAS'. We would also like to thank Iver Jordal and the contributors of AUDIOMENTATIONS for their excellent audio augmentation software.

References

[1] Kendrick, P. et al., "Microphone handling noise: Measurements of perceptual threshold and effects on audio quality," *PloS one*, 10(10), 2015.

[2] Recommendation, I.-R., "Method for objective measurement of perceived audio quality," *Rec. ITU-R BS.1387.1*, 1998.

[3] Porov, A. et al., "Music Enhancement by a Novel CNN Architecture," in *145th Audio Engineering Society Convention*, US, 2018.

[4] Vincent, E. et al., "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language*, 14(4), p. 1462–1469, 2006.

[5] Luo, Y. and Mesgarani, N., "Conv-TasNet: Surpassing Ideal Time–Frequency Magnitude Masking for Speech Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, 27(8), pp. 1256–1266, 2019.

[6] Martin-Doñas, J. M. et al., "A Deep Learning Loss Function Based on the Perceptual Evaluation of the Speech Quality," *IEEE Signal Processing Letters*, 25(11), pp. 1680–1684, 2018.

[7] Recommendation, I.-T., "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," *Rec. ITU-T P. 862.2*, 2007.

[8] Xia, Y. et al., "Weighted Speech Distortion Losses for Neural-Network-Based Real-Time Speech Enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Spain, 2020.

[9] Arian, A. and Nasser, K., "A review of multi-objective deep learning speech denoising methods," *Speech Communication*, 122, pp. 1–10, 2020.

[10] Zhao, Y. et al., "Perceptually Guided Speech Enhancement Using Deep Neural Networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Canada, 2018.

[11] Fu, S.-W. et al., "MetricGAN+: An Improved Version of MetricGAN for Speech Enhancement," in *2021 Conference of the International Speech Communication Association*, Czechia, 2021.

[12] Reddy, C. K. A. et al., "A Scalable Noisy Speech Dataset and Online Subjective Test Framework," in *2019 Conference of the International Speech Communication Association*, Austria, 2019.

[13] Kolbæk, M. et al., "On Loss Functions for Supervised Monaural Time-Domain Speech Enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 28, pp. 825–838, 2020.

[14] Hu, Y. and Loizou, P. C., "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, 16(1), pp. 229–238, 2007.

[15] Taal, C. H. et al., "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Transactions on Audio, Speech and Language Processing*, 19(7), pp. 2125–2136, 2011.

[16] Wichern, G. and Lukin, A., "Removing lavalier microphone rustle with recurrent neural networks," in *21st International Conference on Digital Audio Effects*, Portugal, 2018.

[17] Hines, A. et al., "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1), pp. 1–18, 2015.

[18] Lo, C.-C. et al., "MOSNet: Deep Learning-Based Objective Assessment for Voice Conversion," in *20th Annual Conference of the International Speech Communication Association*, Austria, 2019.

[19] Fu, S.-W. et al., "Utilizing Self-supervised Representations for MOS Prediction," in *2021 Conference of the International Speech Communication Association*, Czechia, 2021.

[20] Manocha, P., "CDPAM: Contrastive learning for perceptual audio similarity," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, Canada, 2021.

[21] Cano, E., Fitzgerald, D., and Brandenburg, K., "Evaluation of quality of sound source separation algorithms: Human perception vs quantitative metrics," in *24th IEEE European Signal Processing Conference*, Hungary, 2016.

[22] Recommendation, I.-T., "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," *Rec. ITU-T P. 862*, 2001.

[23] Manocha, P. et al., "A Differentiable Perceptual Audio Metric Learned from Just Noticeable Differences," in *21st Annual Conference of the International Speech Communication Association*, China, 2020.

[24] Fazenda, B. et al., "Perception and Automated Assessment of Audio Quality in User Generated Content," in *139th Audio Engineering Society Convention*, US, 2015.

[25] Recommendation, I.-T., "Perceptual objective listening quality assessment," *Rec. ITU-T P. 863*, 2011.

[26] Kendrick, P. et al., "Automatic detection of microphone handling noise," in *4th International IEEE Workshop on Cognitive Information Processing*, Denmark, 2014.

[27] Jordal, I. et al., "iver56/audiomentations: v0.22.0," 2022, doi:10.5281/zenodo.6142831.

[28] Reddy, C. K. et al., "2021 Deep Noise Suppression Challenge," in *2021 Conference of the International Speech Communication Association*, Czechia, 2021.

[29] Thiemann, J., Ito, N., and Vincent, E., "The Diverse Environments Multi-channel Acoustic Noise Database (DEMAND): A database of multichannel environmental noise recordings," in *21st International Conference on Acoustics*, Canada, 2013.

[30] Holters, M. and Zölzer, U., "GSTPEAQ–an open source implementation of the PEAQ algorithm," in *18th International Conference on Digital Audio Effects*, Czechia, 2015.

[31] Angonin, C., Chourdakis, E. T., and Åeng, R. A., "Dataset of handling noise for "Assessing the relevance of perceptually driven objective metrics in the presence of handling noise"," 2022, doi: 10.5281/zenodo.6414673.

[32] Gelderblom, F. B., Tronstad, T. V., and Viggen, E. M., "Subjective evaluation of a noise-reduced training target for deep neural network-based speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, 27(3), pp. 583–594, 2018.

[33] Torcoli, M., Kastner, T., and Herre, J., "Objective Measures of Perceptual Audio Quality Reviewed: An Evaluation of Their Application Domain Dependence," *IEEE Transactions on Audio, Speech, and Language Processing*, 29, pp. 1530–1541, 2021.