Audio Engineering Society

Convention Paper 9951

Presented at the 144th Convention
2018 May 23 – 26, Milan, Italy

Comparing the Effect of Audio Coding Artifacts on Objective Quality Measures and on Subjective Ratings

Matteo Torcoli[1] and Sascha Dick[2]

[1] *Fraunhofer Institute for Integrated Circuits IIS, Erlangen, Germany*
[2] *International Audio Laboratories Erlangen, a joint institution of Universität Erlangen-Nürnberg and Fraunhofer IIS*

Correspondence should be addressed to Matteo Torcoli (matteo.torcoli@iis.fraunhofer.de)

ABSTRACT

A recent work presented the subjective ratings from an extensive perceptual quality evaluation of audio signals, where isolated coding artifact types of varying strength were introduced. We use these ratings as perceptual reference for studying the performance of 11 well-known tools for objective audio quality evaluation: PEAQ, PEMO-Q, ViSQOLAudio, HAAQI, PESQ, POLQA, fwSNRseg, dLLR, LKR, BSSEval, and PEASS. Some tools achieve high correlation with subjective data for specific artifact types (Pearson's $r > 0.90$, Kendall's $t > 0.70$), corroborating their value during the development of a specific algorithm. Still, the performance of each tool varies depending on the artifact type and no tool reliably assesses artifacts from parametric audio coding. Nowadays, perceptual evaluation remains irreplaceable, especially when comparing different coding schemes introducing different artifacts.

1 Introduction

Audio coding has the goal of optimizing the quality that is perceived by a human listener when storing or transmitting audio at a given bit rate. Perceptual evaluation via formal listening tests (often referred to as subjective evaluation) is the most reliable method for audio quality evaluation [1]. This is however time-consuming and costly and cannot be easily carried out at each development stage. Therefore, objective evaluation measures are desired, i.e., computational methods that are able to assess the quality of audio as closely as possible to the human assessment.

Such methods were developed in the field of audio and speech coding as well as in related fields. In this paper, we test the quality evaluation tools listed in the following, grouped by their original application context.

Audio and Speech Coding:

- Perceptual Evaluation of Audio Quality (PEAQ);
- PErception MOdel-based Quality (PEMO-Q);
- ViSQOLAudio;
- Hearing-Aid Audio Quality Index (HAAQI);
- Perceptual Evaluation of Speech Quality (PESQ);
- Perceptual Objective Listening Quality Assessment (POLQA).

Speech Enhancement:

- Frequency-Weighted Segmental Signal to Noise Ratio (fwSNRseg);
- Log-Likelihood Ratio Distance (dLLR);
- Log Kurtosis Ratio (LKR).

Blind Source Separation (BSS):

- Blind Source Separation Evaluation (BSSEval);
- Perceptual Evaluation methods for Audio Source Separation (PEASS).

In [2], extensive subjective data was gathered via listening test. Subjects assessed the quality of signals that were distorted in a controlled fashion with different monaural coding artifacts. Here these ratings averaged over the subjects are used as perceptual reference in order to study the performance of 11 well-known objective evaluation tools on distinct, isolated artifact types.

To the best of our knowledge, this is the most extensive investigation of state-of-the-art objective measures published so far.

2 Related Works

An overview over typical audio coding artifacts is given in [3] and [4]; a selection of audio examples for educational purposes is also presented in [3].

Controlled degradation of audio signals is used in [5–7]. In [5] a set of basic distortions are used to simulate potential degradations in dialog enhancement services. Their effect on eight quality metrics is studied by means of a so-called response score and without comparing against subjective ratings. The performance of the objective measures are found to be highly dependent on the distortion type. In [6] a set of degradations is proposed, their implementation is made available, and they are used for studying the robustness of music processing algorithms. In [7] the focus is on BSS: different levels of interfering sources, additive Gaussian noise, and musical noise are simulated and evaluated via a listening test. The subjective results are compared with the objective metrics of BSSEval.

The correlation between objective measures and subjective ratings is studied by many authors, especially in the field of speech enhancement. In [8] PESQ is shown to yield good correlation for enhanced speech; dLLR and fwSNRseg perform nearly as well at a fraction of the computational cost. In [9] PESQ and PEASS are identified as the best tools for predicting separated speech quality. PESQ exhibits good correlation also with the speech recognition rate [10–12]. More recently, [13] shows that Cepstral Distance (CD) and fwSNRseg exhibit good correlation with the perceived amount of reverberation, while no objective measure is found to correlate well with the overall perceived quality of dereverberated speech. In the BSS community, [14] finds a combination of PEAQ features to be the best predictor for the subjective quality of output signals of BSS. For coded audio, [15] proposes another combination of PEAQ features (together with an external feature, i.e., the energy equalization threshold).

3 Objective Measures

This Section describes the investigated objective measures. Table 1 gives an overview of the considered metrics as well as the ranges and scales on which they are defined. All the considered measures are *intrusive* measures, i.e., they need as input a reference signal and the processed signal under test in order to quantify their (perceptual) difference.

Perceptual Evaluation of Audio Quality (PEAQ) [16] employs a peripheral ear model in order to calculate the basilar membrane representations of reference and test signal. Aspects of the difference between these representations are quantified by several features, i.e., the Model Output Variables (MOVs). Examples of MOVs are noise-to-mask ratio (NMR), bandwidth of the signal (BandwidthTest), average distorted block (ADB), average modulation difference (AvgModDiff), and RMS value of the averaged noise loudness (RmsNoiseLoud). By means of a neural network trained with subjective data, the MOVs are combined to give the main output that is referred to as Overall Difference Grade (ODG). The ODG estimates the Subjective Difference Grade (SDG) of a listening test carried out on coded audio signals via a five-grade impairment scale [17]. Hence, the ODG ranges from -4 (very annoying impairment) to 0 (imperceptible impairment). PEAQ can return ODG values slightly higher than 0, but they are here clipped to 0. A Basic and an Advanced version of PEAQ are defined. We compare the Basic version made available by the McGill University as MATLAB code [18] and both Basic and Advanced versions provided by gstPEAQ [19], referred to as gstBas and gstAdv, respectively. The individual MOVs of the Basic version are also investigated.

PErception MOdel-based Quality (PEMO-Q) [20] aims to be a general measure of audio quality effective for a wide range of types of signal and not only coded audio. Its design is founded on a single and coherent auditory model [21]. After time-alignment, the signals are transformed into the internal representations of the auditory model. The cross-correlation coefficient between the two representations is calculated and used as a measure of the perceived similarity, i.e., the Perceptual Similarity Measure (PSM). A regression function based on subjective data is then applied to map the PSM to the ODG. In [20], PEMO-Q is shown to outperform PEAQ. We use the demo version of this

Measure	Worst score	Best score	Scale
ODG (PEAQ and PEMO-Q)	-4.0	0 (*)	Five-grade [17]
ViSQOLAudio	0	1.0	-
HAAQI	0	1.0	-
PESQ	1.0	4.64	MOS [27]
POLQA	1.0	4.75	MOS [27]
fwSNRseg	-10	35	dB
dLLR	2.0	0	-
LKR	0.3 (*)	0 (*)	-
SAR	-10(*)	50(*)	dB
APS	0	100	MUSHRA [28]

Table 1: Measures' ranges and scales. Values limited in this work are indicated by (*).

measure [22]: higher accuracy should be achieved by the full version.

ViSQOLAudio [23] is a metric designed for music encoded at low bitrates developed from Virtual Speech Quality Objective Listener (ViSQOL). Both metrics are based on a model of the peripheral auditory system to create internal representations of the signals called neurograms [24]. These are compared via an adaptation of the structural similarity index, originally developed for evaluating the quality of compressed images.

Hearing-Aid Audio Quality Index (HAAQI) [25] is an index designed to predict music quality for individuals listening through hearing aids. The index is based on a model of the auditory periphery [26], extended to include the effects of hearing loss. This is fitted to a database of quality ratings made by listeners having normal or impaired hearing. The rated signals feature musical content, modified by different types of processing found in hearing aids. Some of these processes are common also in audio coding. The hearing loss simulation can be bypassed and the index becomes valid also for normal-hearing people; we use HAAQI in this normal-hearing mode. Based on the same auditory model, the authors of HAAQI also proposed a speech quality index (HASQI) and a speech intelligibility index (HASPI): references are given in [25].

Perceptual Evaluation of Speech Quality (PESQ) [29–31] was designed for speech transmitted over telecommunication networks. Hence, the method comprises a pre-processing that mimics a telephone handset. Measures for audible disturbances are computed from the specific loudness of the signals and combined in PESQ scores. From them a MOS score [27] is predicted by means of a polynomial mapping function. We use the wideband mode of the reference software [29].

Perceptual Objective Listening Quality Assessment (POLQA) [32–34] was developed as a "technology update" for PESQ and it was designed to predict the perceived overall speech quality of listening tests that comply with [27] or [35] (the test signals used in this work do not meet this requirement). POLQA operates in two modes: narrowband or superwideband. We use a proprietary implementation licensed from OPTICOM (Version 1.1 [32] and Version 2.4 [33, 34]) in the superwideband mode. Further improvements to POLQA are under development [36].

Frequency-Weighted Segmental Signal to Noise Ratio (fwSNRseg) [37] quantifies the ratio of the power of the reference signal and a noise signal that is obtained as the difference of the reference and the test signal. FwSNRseg is computed and weighted for each short time frame and for each subband of a filterbank with a critical-band spacing. The implementation in [38] is used, where the weights are computed from the subband-magnitude of the reference raised to the power of 0.2. This implementation limits the values in the range $[-10, 35]$ dB before the time average.

Log-Likelihood Ratio Distance (dLLR) [39] is based on the assumption that, over short time intervals, speech can be represented by an all-pole model. Hence, Linear Prediction Coefficients (LPC) are computed for the test signal and the reference; the two LPC sets predict the reference with certain residual energies. dLLR is defined as the logarithm of the ratio of these residual energies. We employ the implementation in [38], where the distance is limited to 2 before averaging over time.

Log Kurtosis Ratio (LKR) is a measure of perceived musical artifacts caused by spectral holes or islands. It is calculated as the logarithm of the ratio of the spectral kurtosis after and before processing. LKR was observed to be related to the perception of musical noise in [40]. We do not assume any particular value distribution of the signal power spectra. Instead, the kurtosis is calculated statistically as per [41].

Blind Source Separation Evaluation (BSSEval) [42, 43] is a multi-criteria performance evaluation toolbox. A target source signal is assumed to be estimated from a mixture of multiple sources. The estimated signal is decomposed by an orthogonal projection into target signal component, interference from other sources, artifacts, and spatial distortion. Metrics are computed as energy ratios of these components and expressed in dB. Herein, Source to Artifact Ratio (SAR) is of interest, i.e., the metric specific to introduced artifacts.

Perceptual Evaluation methods for Audio Source Separation (PEASS) [44] was designed as a perceptually motivated successor of BSSEval. The estimated target signal is decomposed by a projection that is carried out on time segments and with a gammatone filterbank. PEMO-Q [20] is used to provide multiple features. Estimates for four perceptual scores are obtained from these features using a neural network trained with subjective ratings. Herein, Artifact-related Perceptual Score (APS) is considered, i.e., the metric that evaluates the presence of computational artifacts.

Calculating PEASS takes exceptionally long: roughly 3 times as long as HAAQI, 10 times as long as PEMO-Q and BSSEval, 15 times as long as POLQA and PEAQ, and 40 – 100 times as long as the remaining tools.

4 Audio coding artifacts

The work in [2] presented methods on how to generate coding artifacts in an isolated and controllable fashion, by forcing audio encoders into controlled, sub-optimal operating modes. The following artifact types were proposed:

- Spectral holes or islands (SH)
  (also known as birdies or musical noise);
- Bandwidth limitation (BL);
- Pre-echoes (PE);
- Tonality or harmonicity mismatch (TM);
- Unmasked noise (UN)
  (i.e., noise substitution in high frequencies).

Out of these artifacts, SH and BL are caused by quantizing spectral parts to zero in transform based audio coders (e.g., MP3, AAC). Relatedly, PE are introduced by temporally smeared quantization noise.

Modern coders apply parametric coding for bandwidth extension (e.g., HE-AAC), which aims at recreating the perceptual properties of the higher frequencies, rather than waveform preserving coding. For a badly parametrized bandwidth extension, TM and UN artifacts can occur. For the generation of those artifacts, the high frequency part of the spectrum has been replaced either by a scaled copy of the lower part of the spectrum for TM or random noise of the same spectral envelope for UN.

Artifact	**Control**	**Quality Levels**				
Type	**Parameter**	**Q1**	**Q2**	**Q3**	**Q4**	**Q5**
SH	hole prob. [%]	70	50	30	20	10
BL	freq. [kHz]	3.5	7.0	10.5	12.0	15
PE	bitrate [kbps]	24	48		96	128
TM	freq. [kHz]	3.0	7.0	9.0	10.5	12
UN	freq. [kHz]	3.0	7.0	9.0	10.5	12

Table 2: Parameters used for artifact generation [2].

Five distinct quality levels for each distortion were selected: they are summarized in Table 2.

5 Subjective Data

As perceptual reference ratings we consider the average over the listeners of the ratings gathered via the listening test described in [2]. Eight items were generated in the five quality levels for each of the five artifact types. MUSHRA listening tests [28] were performed with 16 expert normal-hearing listeners (after post-screening). This resulted in 3200 individual item scores and 200 average scores, which showed a wide coverage of the quality scale, from poor to excellent. The test was divided into 4 to 5 sessions to avoid listener fatigue.

The test items were between 3 and 10 seconds long and contained mostly music. They featured excerpts of isolated instruments (e.g., solo violin, glockenspiel, castanets), instrumental ensembles (e.g., orchestral or pop music) and few pieces containing human voice, singing or talking, with or without accompaniment. Note that the focus on musical content is disadvantageous for the measures designed for the evaluation of speech quality.

As some artifacts affect differently certain types of signals, disjunct sets of items for different artifacts were selected. The set of test items included stereo recordings, however spatial artifacts concerning the stereo image were not included in the selected artifacts, so monaural quality measurements are applicable.[1] An analysis of variance (ANOVA) of the obtained subjective data is carried out in [2].

[1] Please note that independent generation of monaural artifacts (e.g. spectral holes) can still affect the perceived stereo image.

abs(r) | abs(t)	Birdies	Bandwidth limitation	Pre-echoes	Tonality mismatch	Unmasked noise	**Mean**
Select best MOV	0.92 | 0.75	0.98 | 0.77	0.97 | 0.78	0.81 | 0.74	0.81 | 0.68	0.90 | 0.74
ADB McGill	0.92 | 0.75	0.96 | 0.85	0.93 | 0.74	0.64 | 0.53	0.73 | 0.55	0.84 | 0.68
HAAQI	0.88 | 0.71	0.80 | 0.59	**0.92** | 0.73	0.68 | 0.42	0.76 | 0.51	**0.81** | **0.59**
ODG McGill	0.46 | 0.31	**0.95** | **0.84**	0.89 | 0.65	0.84 | **0.64**	0.73 | **0.53**	0.77 | **0.59**
ODG gstBas	0.47 | 0.33	**0.95** | **0.84**	0.89 | 0.64	0.84 | **0.64**	0.73 | **0.52**	0.77 | **0.59**
ODG gstAdv	0.49 | 0.30	0.91 | 0.78	**0.91** | 0.70	0.77 | 0.58	0.76 | **0.53**	0.77 | **0.58**
PESQ	0.74 | 0.58	0.67 | 0.62	0.82 | **0.75**	0.77 | 0.47	**0.78** | **0.53**	0.76 | **0.59**
ViSQOLAudio	0.70 | 0.52	0.79 | 0.61	0.88 | 0.62	0.75 | 0.50	0.69 | 0.46	0.76 | 0.54
APS (PEASS)	**0.95** | **0.86**	0.84 | 0.72	0.51 | 0.41	**0.87** | 0.59	0.59 | 0.41	0.75 | **0.60**
ODG PEMO-Q	0.93 | 0.77	0.82 | **0.85**	0.68 | 0.58	0.62 | 0.39	0.49 | 0.38	0.71 | **0.59**
POLQA V1.1	0.81 | 0.67	0.79 | 0.61	0.32 | 0.21	0.73 | 0.41	0.68 | 0.43	0.67 | 0.47
SAR (BSSEval)	0.34 | 0.28	0.91 | 0.77	0.66 | 0.55	0.53 | 0.36	0.52 | 0.37	0.59 | 0.47
POLQA V2.4	0.81 | 0.67	0.56 | 0.39	0.44 | 0.16	0.47 | 0.36	0.60 | 0.19	0.58 | 0.35
fwSNRseg	0.71 | 0.52	0.89 | 0.80	0.66 | 0.49	0.09 | 0.05	0.38 | 0.24	0.54 | 0.42
dLLR	0.02 | 0.03	0.67 | 0.58	0.85 | 0.57	0.13 | 0.02	0.12 | 0.03	0.36 | 0.25
LKR	0.31 | 0.09	0.41 | 0.33	0.04 | 0.05	0.13 | 0.01	0.22 | 0.13	0.22 | 0.12

Table 3: Absolute value of the Pearson's correlation r and of the Kendall's rank correlation t for the individual distortions and mean values. Measures are in order of decreasing mean $|r|$ from top to bottom.

6 Results

6.1 Correlation performance

The quality scores given by the objective measures are compared with the mean MUSHRA scores. Absolute Pearson's linear correlation coefficient r and Kendall's rank correlation coefficient t are used as performance criteria: they are listed in Table 3. The table also shows "Select best MOV", i.e., the MOV from Basic PEAQ (McGill implementation) that correlates best with subjective ratings for each distortion type. This MOV is shown in the upper-right subplot of Figs. 1 – 5. Moreover, the ADB is shown: this MOV correlates best with subjective ratings on average.

Figs. 1 – 5 depict the detailed results: the mean MUSHRA scores (x-axis) are plotted against the objective quality scores (y-axis). The colors of the circles depict the different quality levels (Table 2). Each dashed line connects the circles related to the same item, on which the different quality levels were applied. The title of each subplot reports the name of the measure together with the Pearson's and Kendall's correlation coefficients. Both correlation coefficients are calculated without considering the reference unprocessed signals (depicted with black circles).

6.2 Discussion

Considering the final output of the evaluation tools, the best average correlations are achieved by **HAAQI** ($|r| = 0.81$, $|t| = 0.59$) and the ODG as estimated by **PEAQ**[2] ($|r| = 0.77$, $|t| = 0.59$). HAAQI was specifically designed for music quality and PEAQ was designed for the assessment of audio coding, i.e., both fit well our test signals. Moreover, HAAQI and PEAQ perform particularly well for pre-echoes and bandwidth limitation, which are classical distortions both in hearing aids and audio coding. However, only modest correlation is exhibited by HAAQI and PEAQ in the case of tonality mismatch and unmasked noise, i.e., the distortions simulating suboptimal parametric coding of the higher frequency bands. While this could be not relevant in hearing aids, it is a critical aspect in audio coding. Only moderate correlation is exhibited even by the the best MOV of PEAQ, suggesting that a MOV reliably detecting these kinds of distortion is missing. In fact, no codec using this kind of technique was among the codecs used for the development of PEAQ, completed in 2001. Parametric bandwidth extension became more customary only later,

[2] The 3 implementations of PEAQ (gstBas, gstAdv, McGill) show very similar performance. They are here generally referred to as PEAQ. The coefficients obtained for McGill are reported in the text.

e.g., with HE-AAC (using Spectral Band Replication) that was standardized in 2003 [45]. Interestingly, the performance of PEAQ is underwhelming for birdies ($|r| = 0.46$, $|t| = 0.31$), while internal PEAQ MOVs such as ADB still show high correlation coefficients ($|r| = 0.92$, $|t| = 0.75$), suggesting that they are not sufficiently reflected in the combined ODG. In fact, a single PEAQ MOV, i.e., ADB shows higher average performance than the final ODG for our data. Moreover, mean correlation coefficients $|r| = 0.90$ and $|t| = 0.74$ can be achieved by just selecting the best MOV for each distortion. This can be obtained "manually" if a developer knows the system under test and the most relevant distortion that can appear. In this direction, Figs. 1-5 work as look-up tables for the best MOV in each case (upper-right subplots).

PESQ is the third best performing tool on average ($|r| = 0.76$, $|t| = 0.59$), in spite of the fact that it was designed for speech quality, while most of our test items consist of musical content. Still, Figs. 1 – 5 show a saturation effect, i.e., scores close to the maximum are often obtained already for the middle quality levels.

ViSQOLAudio achieves average Pearson's $|r| = 0.76$, similarly to PEAQ and PESQ, even if it does not exhibit the best $|r|$ for any of the artifact type. However, the absolute quality ranking across items is often wrong: similar scores are assigned to the low quality level of one item and the high quality level of a different item, resulting in the lower Kendall's $|t| = 0.54$.

APS shows moderate average performance. Especially good performance is observed for birdies ($|r| = 0.95$, $|t| = 0.86$), which is an important distortion in BSS. APS also exhibits the highest Pearson's $|r|$ for tonality mismatch. However, Pearson's $|r|$ shows here its limitations as it ignores the saturation effect exhibited by APS. Fig. 4 shows that the most promising measure in this case is NMR, as unveiled by Kendall's $|t| = 0.74$ and, in this regard, PEAQ has the best performing final output, albeit achieving only Kendall's $|t| = 0.64$.

PEMO-Q-based ODG performs well only for birdies ($|r| = 0.93$, $|t| = 0.77$) and bandwidth limitation ($|r| = 0.82$, $|t| = 0.85$). Yet, for bandwidth limitation PEMO-Q saturates on quality level 3, i.e., similar quality is measures for crossover frequencies $>=$ 10.5 kHz.

As far as **POLQA** is concerned, related literature shows that V2.4 improves on V1.1 [34] and that both versions improve on PESQ [32] for speech signals perceptually assessed in tests such as [27] and [35]. Interestingly, the opposite trend is observed for our data. We suspect that the measure evolved improving the accuracy for a specific case, while losing in generality and robustness if applied to another context. Moreover, the fact that PESQ and POLQA are designed for speech is evident considering that they saturate after the second quality level for bandwidth limitation. The crossover frequency of this level is 7.0 kHz, i.e., already including the most important frequency range of speech.

Finally, the lowest correlation coefficients are obtained for the simplest measures that comprise only little psycho-acoustical knowledge (**fwSNRseg**) or none at all (**SAR**, **dLLR**, and **LKR**). LKR is meant for birdies. A closer analysis of this case (bottom left subplot in Fig. 1) shows that LKR performs well in detecting the relative quality level ranking for individual items, but the absolute LKR value obtained across different items cannot be compared directly.

7 Conclusion

Eleven well-known tools for the objective evaluation of audio quality were applied to 200 signals created by introducing monaural coding artifacts with controlled strength within a large quality range. Reliable subjective ratings were available and used in order to assess the performance of the tools.

The results show that the performance of each tool depends on the artifact type, as also found in [5]. The quality of specific artifacts is estimated accurately by specific tools (Pearson's $|r| > 0.90$, Kendall's $|t| > 0.70$). On average, the highest correlation is exhibited by single MOVs of PEAQ, followed by HAAQI and the main output of PEAQ. Lower correlation is observed for artifacts simulating suboptimal parametric bandwidth extension, i.e., unmasked noise and tonality mismatch. This confirms that the measures can predict the quality of artifacts they have been modeled or trained towards, but fail to generalize and predict the quality of unknown artifacts. State-of-the-art objective measurements can be a powerful tool when the expected artifact characteristics are known and comparable for different working points, e.g., when developing a specific algorithm.

Still, objective measurements cannot be expected to produce reliable results when comparing different coding schemes, potentially introducing different artifacts. Nowadays, subjective evaluation remains the most reliable method for the assessment and comparison of the perceptual quality of generalized audio codecs.

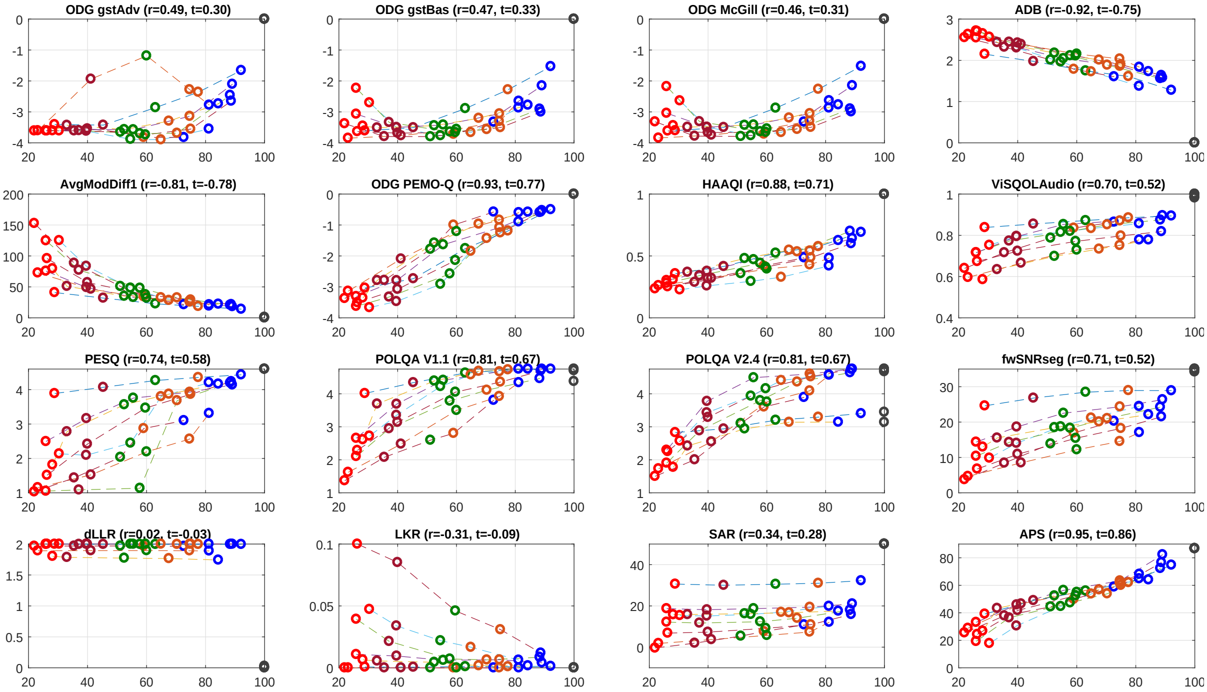

Fig. 1: Artifact type: birdies. Mean MUSHRA scores (x-axis) against objective measures (y-axis). The colors of the circles depict the different quality levels. Dashed lines connect the circles related to the same items.

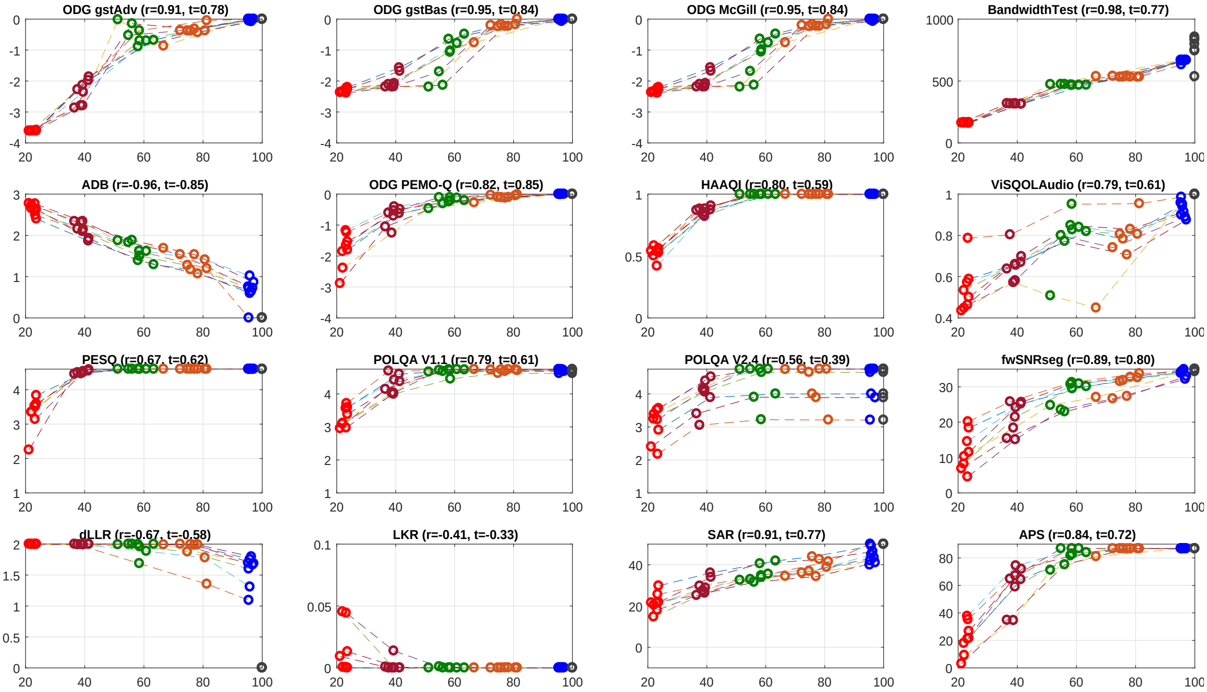

Fig. 2: Artifact type: bandwidth limitation.

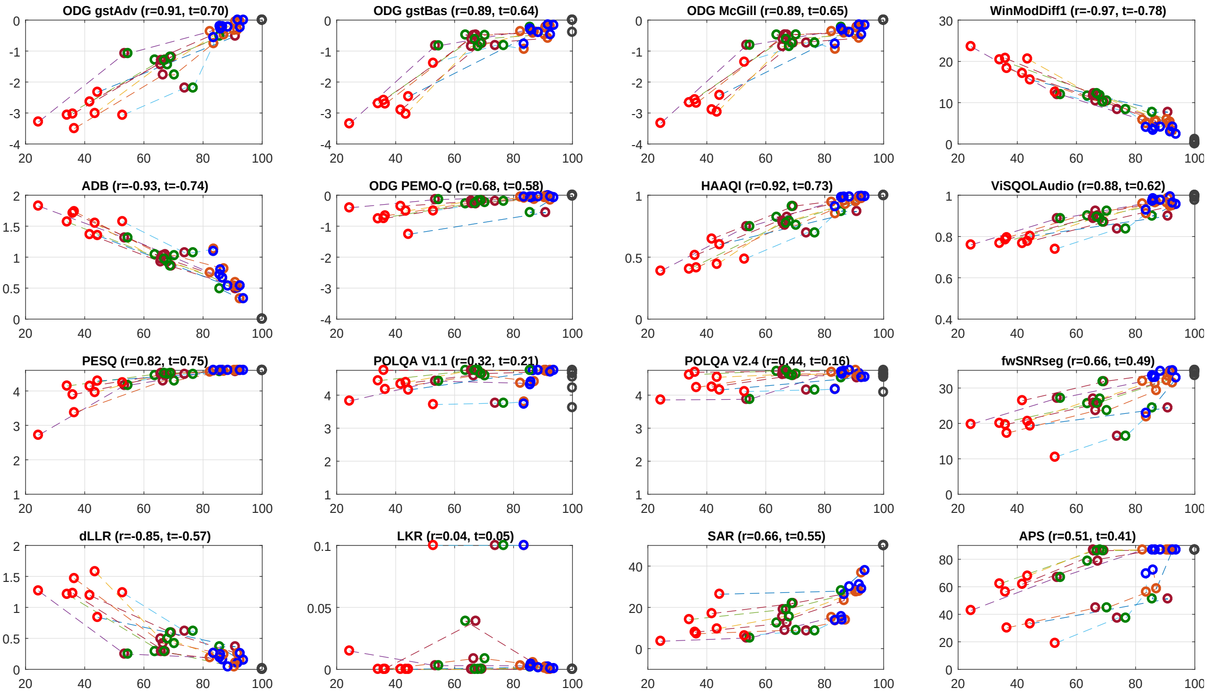

Fig. 3: Artifact type: pre-echoes. Mean MUSHRA scores (x-axis) against objective measures (y-axis). The colors of the circles depict the different quality levels. Dashed lines connect the circles related to the same items.

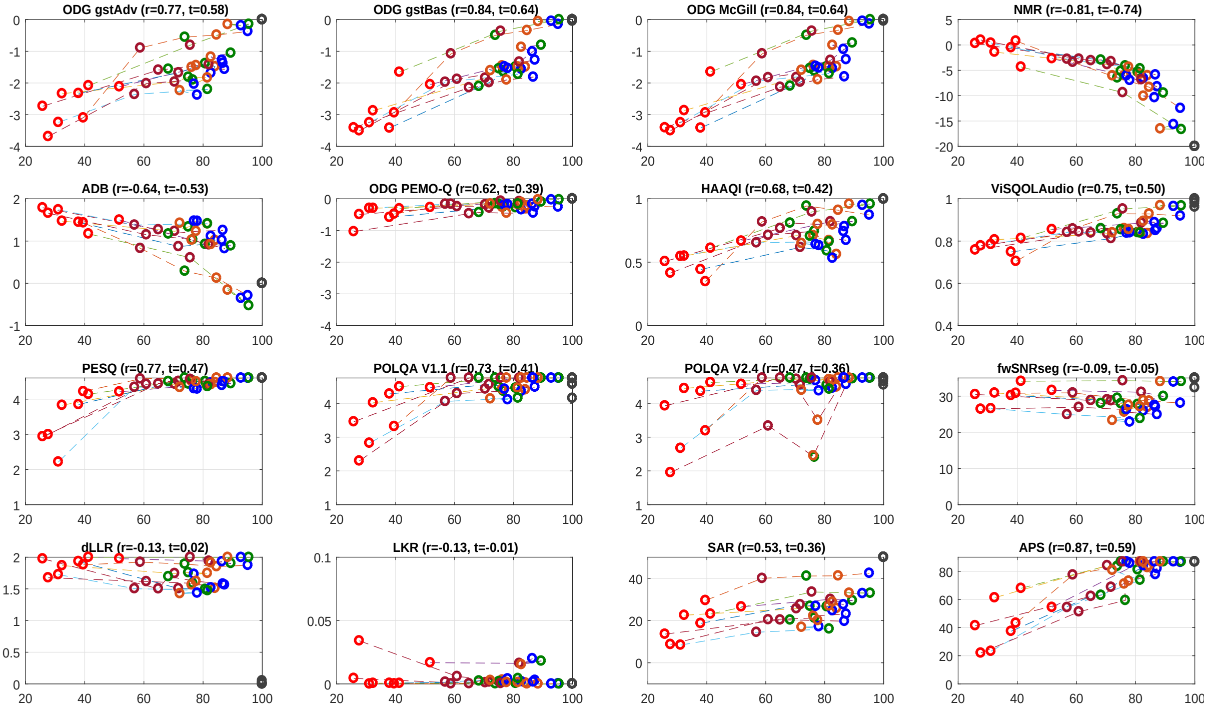

Fig. 4: Artifact type: tonality mismatch.

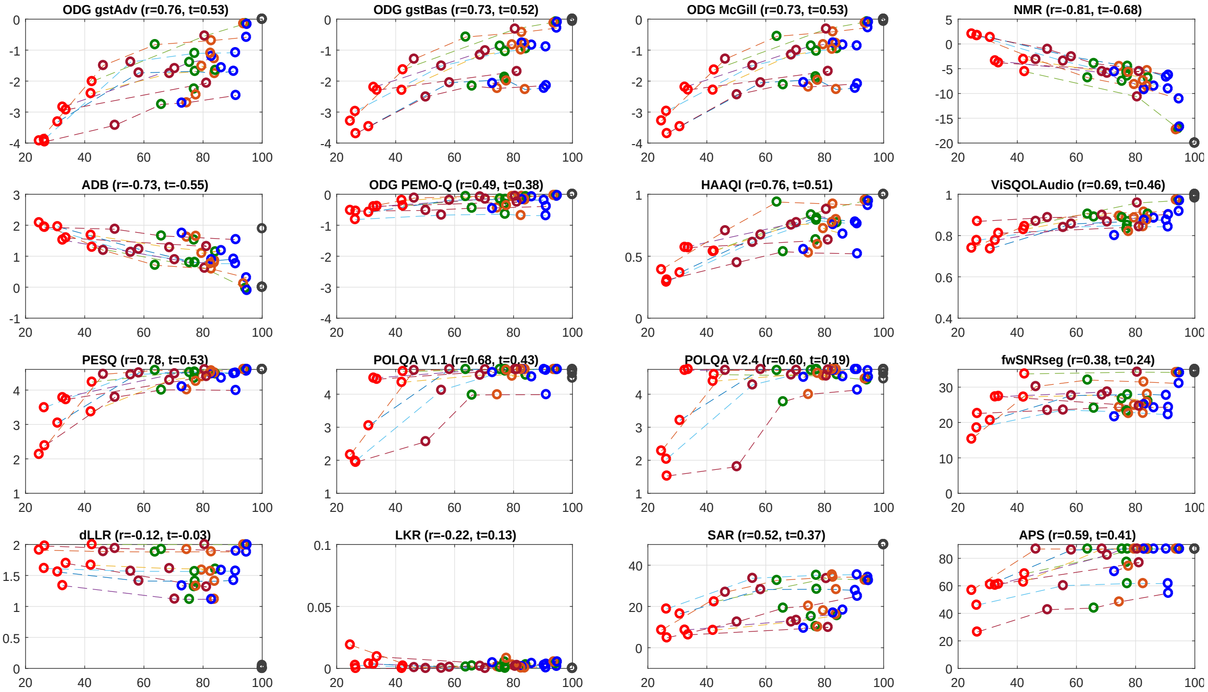

Fig. 5: Artifact type: unmasked noise. Mean MUSHRA scores (x-axis) against objective measures (y-axis). The colors of the circles depict the quality levels. Dashed lines connect the circles related to the same items.

References

[1] S. Bech and N. Zacharov, *Perceptual Audio Evaluation - Theory, Method and Application*. John Wiley & Sons, 2006.

[2] S. Dick, N. Schinkel-Bielefeld, and S. Disch, "Generation and Evaluation of Isolated Audio Coding Artifacts," in *Proc. AES 143rd Conv.*, 2017.

[3] M. Erne, "Perceptual Audio Coders "What To Listen For"," in *Proc. AES 111th Conv.*, 2001.

[4] C. Liu, H. Hsu, and W. Lee, "Compression Artifacts in Perceptual Audio Coding," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 4, 2008.

[5] M. Torcoli and C. Uhle, "On the Effect of Artificial Distortions on Objective Performance Measures for Dialog Enhancement," in *Proc. AES 141st Conv.*, 2016.

[6] M. Mauch and S. Ewert, "The Audio Degradation Toolbox and its Application to Robustness Evaluation," in *Proc. 14th ISMIR Conf.*, 2013.

[7] J. Kornycky, B. Gunel, and A. Kondoz, "Comparison of Subjective and Objective Evaluation Methods for Audio Source Separation," in *Proc. Mtgs on Acoust.*, 2008.

[8] Y. Hu and P. C. Loizou, "Evaluation of Objective Quality Measures for Speech Enhancement," *IEEE Trans. Audio, Speech and Language Process.*, vol. 16, no. 1, 2008.

[9] P. Mowlaee, R. Saeidi, *et al.*, "Subjective and Objective Quality Assessment of Single-channel Speech Separation Algorithms," in *Proc. IEEE ICASSP*, 2012.

[10] L. Di Persia, D. Milone, *et al.*, "Perceptual Evaluation of Blind Source Separation for Robust Speech Recognition," *Signal Process.*, vol. 88, no. 10, 2008.

[11] H. Sun, L. Shue, and J. Chen, "Investigations into the Relationship Between Measurable Speech Quality and Speech Recognition Rate for Telephony Speech," in *Proc. IEEE ICASSP*, 2004.

[12] T. Yamada, M. Kumakura, and N. Kitawaki, "Performance Estimation of Speech Recognition System Under Noise Conditions Using Objective Quality Measures and Artificial Voice," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 6, 2006.

[13] K. Kinoshita, M. Delcroix, *et al.*, "A Summary of the REVERB Challenge: State-of-the-art and Remaining Challenges in Reverberant Speech Processing Research," *EURASIP J. on Advances in Signal Process.*, vol. 2016, no. 1, 2016.

[14] T. Kastner, "Evaluating Physical Measures for Predicting the Perceived Quality of Blindly Separated Audio Source Signals," in *Proc. AES 127th Conv.*, 2009.

[15] C. D. Creusere, K. D. Kallakuri, and R. Vanam, "An Objective Metric of Human Subjective Audio Quality

Optimized for a Wide Range of Audio Fidelities," *IEEE Trans. Audio, Speech, and Language Process.*, vol. 16, no. 1, 2008.

[16] ITU-R Rec. BS.1387-1, "Method for Objective Measurements of Perceived Audio Quality," 2001.

[17] ITU-R Rec. BS.562-3, "Subjective Assessment of Sound Quality," 1990.

[18] P. Kabal, "An Examination and Interpretation of ITU-R BS.1387: Perceptual Evaluation of Audio Quality," tech. rep., TSP Lab, McGill University, 2002. Code available at `http://www-mmsp.ece.mcgill.ca/Documents/Software/`.

[19] M. Holters and U. Zölzer, "GstPEAQ - An Open Source Implementation of the PEAQ Algorithm," in *Proc. DAFx-15*, 2015. Code available at `https://github.com/HSU-ANT/gstpeaq`.

[20] R. Huber and B. Kollmeier, "PEMO-Q - A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, 2006.

[21] T. Dau, B. Kollmeier, and A. Kohlrausch, "Modeling Auditory Processing of Amplitude Modulation. I. Detection and Masking with Narrow-band Carriers," *J. Acoust. Soc. Am.*, vol. 102, no. 5, 1997.

[22] "PEMO-Q software: `http://www.hoertech.de/en/f-e-products/pemo-q.html`."

[23] A. Hines, E. Gillen, *et al.*, "ViSQOLAudio: An Objective Audio Quality Metric for Low Bitrate Codecs," *J. Acoust. Soc. Am.*, vol. 137, no. 6, 2015. Code available at `https://sites.google.com/a/tcd.ie/sigmedia/visqolaudio`.

[24] A. Hines and N. Harte, "Speech Intelligibility Prediction Using a Neurogram Similarity Index Measure," *Elsevier Speech Communication*, vol. 54, no. 2, 2012.

[25] J. M. Kates and K. H. Arehart, "The Hearing-Aid Audio Quality Index (HAAQI)," *IEEE Trans. Audio, Speech and Language Process.*, vol. 24, no. 2, 2016. Evaluation code kindly provided by Prof. J.M. Kates.

[26] J. M. Kates, "An auditory model for intelligibility and quality predictions," in *Proc. Mtgs on Acoustics ICA2013*, vol. 19, Acoust. Soc. Am., 2013.

[27] ITU-T Rec. P.800, "Methods for Subjective Determination of Transmission Quality," 1996.

[28] ITU-R Rec. BS.1534-3, "Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," 2015.

[29] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," 2001. Code available at `https://www.itu.int/rec/T-REC-P.862`.

[30] ITU-T Rec. P.862.1, "Mapping Function for Transforming P.862 Raw Results Scores to MOS-LQO," 2003.

[31] ITU-T Rec. P.862.2, "Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," 2007.

[32] ITU-T Rec. P.863, "Perceptual Objective Listening Quality Assessment," 2011.

[33] ITU-T Rec. P.863 Edition 2.4, "Perceptual Objective Listening Quality Assessment," 2014.

[34] "POLQA 2015 (V2.4) investigated - Technical White Paper," 2014. `http://www.polqa.info/download/WhitePaper_POLQA_V2\%204_2014-10-30.pdf`.

[35] ITU-T Rec. P.830, "Subjective Performance Assessment of Telephone-band and Wideband Digital Codecs," 1996.

[36] P. Počta and J. G. Beerends, "Subjective and Objective Assessment of Perceived Audio Quality of Current Digital Audio Broadcasting Systems and Web-casting Applications," *IEEE Trans. Broadcasting*, vol. 61, no. 3, 2015.

[37] J. M. Tribolet, P. Noll, *et al.*, "A Study of Complexity and Quality of Speech Waveform Coders," in *Proc. IEEE ICASSP*, 1978.

[38] P. C. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007. Includes CD with MATLAB code.

[39] F. Itakura, "Minimum Prediction Residual Principle Applied to Speech Recognition," *IEEE Trans. Acoust., Speech and Language Process.*, vol. 23, no. 1, 1975.

[40] Y. Uemura, Y. Takahashi, *et al.*, "Automatic Optimization Scheme of Spectral Subtraction Based on Musical Noise Assessment via Higher-order Statistics," in *Proc. 11th IWAENC*, 2008.

[41] H. Yu and T. Fingscheidt, "A Figure of Merit for Instrumental Optimization of Noise Reduction Algorithms," in *Proc. 5th Biennial Workshop on DSP for In-Vehicle Systems*, 2011.

[42] E. Vincent, R. Gribonval, and C. Févotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 14, no. 4, 2006. Code available at `http://bass-db.gforge.inria.fr/bss_eval/`.

[43] E. Vincent, H. Sawada, *et al.*, "First Stereo Audio Source Separation Evaluation Campaign: Data, Algorithms and Results," in *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation*, 2007.

[44] V. Emiya, E. Vincent, *et al.*, "Subjective and Objective Quality Assessment of Audio Source Separation," *IEEE Trans. Audio, Speech and Language Process.*, vol. 19, no. 7, 2011. Code available at `http://bass-db.gforge.inria.fr/peass/`, Version 2.0 is used.

[45] ISO/IEC 14496-3:2001/Amd 1:2003, "Bandwidth Extension," 2003.