This paper is the author accepted manuscript that was peer-reviewed as a complete manuscript and accepted for presentation at the 141st Convention of the Audio Engineering Society (AES), 2016 September 29 - October 2, Los Angeles, CA, USA, as paper number 9681. The full published version can be found in the AES E-Library: www.aes.org/e-lib/online/browse.cfm?elib=18485. The content of the present manuscript and the content of the AES version are identical. However, AES front matter, headers and footers are not shown here.

Please cite as: *M. Torcoli and C. Uhle, "On the Effect of Artificial Distortions on Objective Performance Measures for Dialog Enhancement." In Proc. 141st Audio Engineering Society Convention, 2016.*

On the Effect of Artificial Distortions on Objective Performance Measures for Dialog Enhancement

Matteo Torcoli[1] and Christian Uhle[1]

[1] *Fraunhofer Institute for Integrated Circuits IIS, Am Wolfsmantel 33, 91058 Erlangen, Germany*

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

ABSTRACT

The objective evaluation of dialog enhancement systems using computational methods is desired to complement the subjective evaluation using listening tests. It remains a challenge because for this application neither were performance measures specifically designed, nor were existing measures systematically analyzed. This work investigates eight objective performance measurement tools originally developed for audio and speech coding, speech enhancement, or source separation. To this end, a set of basic distortions is presented and used to simulate degradations that are common in dialog enhancement. The effect of the artificial distortions on the performance measures is quantified by means of a so-called response score that is proposed here.

1 Introduction

In broadcast material the preferred balance between dialog and all the other sound sources (background) depends on taste, listening environment, hearing, and mother language. It follows that the listener benefits from having the possibility to adjust this balance in addition to the overall volume. Dialog Enhancement (DE) addresses this challenge by enabling the control of the level of dialog and background. When they are not separately available, methods for decomposing the mixture signals need to be applied [6].

The evaluation of such a process can be categorized into subjective and objective methods. The subjective evaluation using listening tests is the most reliable procedure, but time-consuming and costly. Therefore, various computational methods for the assessment of sound quality, mixture decomposition performance, or intelligibility have been developed for the application to audio and speech coding, speech enhancement (SE), or blind source separation (BSS).

These applications can be similar to DE in some aspects, for example in the definition of the desired and interfering signal. SE deals with the same desired signal, i.e. speech, but the interfering sounds can be much more diverse in broadcast material than in telecommunication systems. BSS aims at removing the interferer instead of partially attenuating it, and the desired signal is not restricted to speech. In audio and speech coding, the quality of a wide range of signals can be deteriorated by quantization and other means to reduce the bit rate.

In this work, we investigate the performance tools listed in Table 1 that have been proposed for these applications. To this end, a set of basic distortions is presented and used to simulate degradations that are common in DE. We analyze how the performance measures respond to them using the “response score”. This is

Audio and speech coding	Perceptual Evaluation of Audio Quality (PEAQ)
	Perceptual Evaluation of Speech Quality (PESQ)
	Perceptual Objective Listening Quality Assessment (POLQA)
Speech Enhancement (SE)	Frequency Weighted Segmental Signal to Noise Ratio (fwSNRseg)
	Log-Likelihood Ratio Distance (LLRd)
	Short-Time Objective Intelligibility (STOI)
Blind Source Separation (BSS)	Blind Source Separation Evaluation (BSSEval)
	Perceptual Evaluation methods for Audio Source Separation (PEASS)

Table 1: Performance measurement tools and their original application context.

a novel metric with low cost and high reproducibility and we propose to use this analysis to complement investigations on objective measures that make use of subjective data.

The use of artificial distortions for other applications was proposed in [16] and [14]. The correlation between objective measures and subjective ratings was studied by many authors. Hu and Loizou [7] showed that PESQ yielded good correlation for enhanced speech; LLRd and fwSNRseg performed nearly as well at a fraction of the computational cost. Mowlaee et al. [17] identified PESQ and PEASS as the best tools for predicting separated speech quality and STOI for intelligibility. PESQ was also shown to have good correlation with the speech recognition rate in [4], [25], [30]. More recently, Kinoshita et al. [13] showed that Cepstral Distance (CD) and fwSNRseg exhibited good correlation with the perceived amount of reverberation, while no objective measure was found to correlate sufficiently well with the overall perceived quality of dereverberated speech. Kornycky et al. [14] related the performance criteria of BSSEval to subjective scores. Kastner [11] found a combination of PEAQ features (herein referred to as PEAQ4f) to be the best predictor for the subjective quality of output signals of BSS.

Measure	*Worst score*	*Best score*	*Scale*
fwSNRseg	-10	35	dB
LLRd	2.0	0	-
PEAQ	-4.0	0	Five-grade [3]
PEAQ4f	0	100	MUSHRA [2]
PESQ	1.0	4.6	MOS [18]
BSSEval	-10(*)	30(*)	dB
fwBSSEvalSeg	30	-10	dB
PEASS	0	100	MUSHRA [2]
STOI	0.5(*)	1.0	-
POLQA	1.0	4.75	MOS [18]

Table 2: Measures' ranges and scales. Values limited in this work are indicated by (*).

The remaining part of the paper is organized as follows. Sec. 2 introduces the objective measures. Sec. 3 defines the response score. Sec. 4 details the artificial distortions and Sec. 5 shows how the objective metrics respond to them. Conclusions are given in Sec. 6.

2 Objective Measures

This Section describes the investigated objective measures. As their ranges are different, a better comparison can be carried out after the following normalization to the common range $[0, 100]$.

$$q = 100 \frac{q^o - q_{min}}{q_{max} - q_{min}}, \tag{1}$$

with original measure q^o, normalized measure q, best score q_{max} and worst score q_{min} as given in Table 2.

2.1 Frequency Weighted Segmental Signal to Noise Ratio

FwSNRseg [27] 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 [15] is used, where the weights are computed from the subband-magnitude of the reference raised to the power of 0.2. In addition, the values are limited in the range $[-10, 35]$ dB before the time average is taken.

2.2 Log-Likelihood Ratio Distance

LLRd [9] 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. LLRd is defined as the logarithm of the ratio of these residual energies. We employ the implementation in [15], where the distance is limited to 2 before averaging over time.

2.3 Perceptual Evaluation of Audio Quality

PEAQ [1] 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, the Model Output Variables (MOVs). 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). A basic and an advanced version of PEAQ are available. The Matlab implementation of the basic version in [10] is used here.

Furthermore, PEAQ4f is investigated, i.e. the modified version proposed in [11] that combines four MOVs.

2.4 Perceptual Evaluation of Speech Quality

PESQ [21–23] was designed for speech transmitted over telecommunication networks. The method comprises a pre-processing that mimics a telephone handset. Hence, measures for audible disturbances are computed from the specific loudness of the signals and combined in PESQ scores. From them a MOS score [18] is predicted by means of a polynomial mapping function. We use the wideband mode of the reference software that comes as annex to [21].

2.5 Blind Source Separation Evaluation

BSSEval [28, 29] 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, additive noise, artifacts, and spatial distortion. Metrics are computed as energy ratios of these components and expressed in dB. Herein Source to Distortion Ratio (SDR), Source to Interference Ratio (SIR), and Source to Artifact Ratio (SAR) are considered. SDR is a measure of the total error, while SIR and SAR are specific criteria that have at the denumerator only the energy of the interference or of the artifacts.

BSSEval does not consider perceptual aspects and an augmented version was proposed in [11] for improving the correlation with subjective ratings. This is referred to as fwBSSEvalSeg. Similarly to fwSNRseg, fwBSSEvalSeg employs segmental calculation, critical-band spaced filterbank, and weighting. Here, the weights are obtained by raising to the power of 0.25 the subband-magnitude of a component that depends on the different measures. Furthermore, fwBSSEvalSeg calculates the inverse energy ratios with respect to BSSEval. In such a way, the lowest signal to distortion ratio has the strongest influence on the segmental calculation. Hence, fwDSRseg, fwISRseg, and fwASRseg are obtained, which are the perceptually improved versions of SDR, SIR, and SAR, respectively.

2.6 Perceptual Evaluation Methods for Audio Source Separation

PEASS [5] was designed as a perceptually motivated successor of BSSEval. It is based on a decomposition of the estimated target signal and on the use of PEMO-Q [8] to provide multiple features. Estimates for four perceptual scores are obtained from the features using a neural network trained with subjective ratings. The four scores are: Target-related Perceptual Score (TPS), Interference-related Perceptual Score (IPS), Artifact-related Perceptual Score (APS), and Overall Perceptual Score (OPS). TPS reflects how well the original source is preserved in the source estimate; IPS quantifies how intensely the interference from other sources is perceivable in the source estimate; APS quantifies the presence of computational artifacts in the estimated source; finally OPS expresses the overall quality. It is worth noting that the calculation of PEASS takes exceptionally long compared to the other tools.

2.7 Short-Time Objective Intelligibility

STOI [26] is a measure that is expected to have monotonic relation with the average intelligibility. It addresses especially speech processed by some type of time-frequency weighting. We consider only the higher

half of STOI's range, i.e. $[0.5, 1]$, as low intelligibility is out of the scope of this work.

2.8 Perceptual Objective Listening Quality Assessment

POLQA [24] was developed as a follow-up of PESQ and it was designed to predict the perceived overall speech quality of listening tests that comply with [18] or [20] (please note that the test signals used in this work do not meet this requirement). POLQA supports two operational modes, for narrowband speech signals and for superwideband. We employ a proprietary implementation of the latter one.

3 Response Analysis

This Section introduces a metric for analyzing the response of an objective measure q to a distortion d applied on different signals. The metric is named response score and denoted by $\rho_{q,d}$. It takes three desired properties into account, namely monotonicity, inter-item deviation, and range spanning.

3.1 Monotonicity

Monotonicity refers to the property that increasing the intensity of an artificial distortion results in monotonic change of a performance measure. Here, it is quantified using the Kendall's rank correlation coefficient $\tau_{q,d}$ [12]. This is a coefficient that quantifies the concordance between two sets of ranked data, which in our case are p_d and $\bar{q}(p_d)$: p_d denotes the parameter that controls the intensity with which d is applied and $\bar{q}(p_d)$ symbolizes the performance measure averaged over all test signals. For independent sets, $\tau_{q,d}$ is close to 0, while $|\tau_{q,d}| = 1$ if they relate strictly monotonically.

3.2 Inter-Item Deviation

Inter-item deviation refers to the similarity between the performance metrics measured for the same distortion (and same intensity) applied to different signals. This can be quantified by the averaged normalized standard deviation $\zeta_{q,d}$.

$$\zeta_{q,d} = 1 - \frac{\overline{\sigma_{q,d}(p_d)}}{r/2}, \tag{2}$$

where $\sigma_{q,d}(p_d)$ is the standard deviation from $\bar{q}(p_d)$, $\overline{\sigma_{q,d}(p_d)}$ is its average over all p_d, and r is the range on which q is defined. Thus, $\zeta_{q,d}$ is 0 for maximum inter-item deviation and 1 if $\sigma_{q,d}(p_d) = 0$ for all p_d.

3.3 Range spanning

Range spanning is the property of a performance measure to span most of its range when computed for distortions whose intensities range from hardly noticeable to severe. We propose to compute a 80% spanned range coefficient $\omega_{q,d}$ according to

$$\omega_{q,d} = \frac{\min(|\max(\bar{q}(p_d)) - \min(\bar{q}(p_d))|, 0.8r)}{0.8r}. \tag{3}$$

This quantity ranges between 0 (if $\bar{q}(p_d)$ is constant) and 1 (when 80% or more of the range is spanned).

3.4 Response score

Finally, we propose to compute the response score $\rho_{q,d}$ as the geometric mean of the previous coefficients,

$$\rho_{q,d} = \sqrt[3]{|\tau_{q,d}|\zeta_{q,d}\omega_{q,d}}. \tag{4}$$

Eq. (4) is a first heuristic attempt to combine the introduced properties in one score. Further works will be carried out studying the relationship between coefficients describing the subjective correlation (e.g., Pearson correlation) and $\tau_{q,d}$, $\zeta_{q,d}$, $\omega_{q,d}$, and combinations of them.

Moreover, the proposed properties are necessary but not sufficient conditions for good subjective correlation. It follows that $\rho_{q,d}$ can only complement subjective correlation studies by giving a detailed understanding of the response of the objective measures to isolated well-defined distortions.

Some measures are not supposed to be responsive to some artificial distortions, e.g., SAR or APS are insensitive to interfering noise. In these cases, $\rho_{q,d}$ is not reported herein.

4 Distortions

In the following the artificial distortions and the parameters for controlling their intensity are presented.

Additive noise: Pink noise is added to the input with varying Signal to Noise Ratio (SNR).

Modulated Noise Reference Unit (MNRU): It simulates speech distortion that can be caused by companding, a recurring process in telecommunications. White noise with unit variance $w(n)$ is

modulated by the input signal $x(n)$ and added to it, i.e. $mnru(n) = x(n)[1+10^{-Q/20}w(n)]$, where n is the time index and Q is the control parameter expressed in dB [19].

Musical noise: It is often described as warbling or having a tonal quality and is caused by narrow peaks in the time-frequency domain that result from manipulations of time-frequency representation. Here, it is simulated by setting to zero a controlled percentage of the Short-Time Fourier Transform (STFT)[1] tiles of the input signal.

Low-pass filter: As it is often the case that the background extends to higher frequencies than speech, a frequent observed impairment after DE is the loss of energy in the higher frequencies. To simulate this, a Butterworth low-pass filter of order 3 is employed and controlled via the cut-off frequency that is varied linearly between 3 kHz and 15 kHz. The low-passed signal is then rescaled so to have the original energy.

Clipping: This is a non-linear distortion that occurs when the range of an audio signal is limited and the signal is truncated. As simulation, we normalize the audio signal such that a controlled percentage of samples is outside the interval $[-1,1]$ and so clipped to -1 or 1.

Spatial image distortion: Background signals are typically stereophonic, and their attenuation results in a reduction of spatial information. We simulate this by mixing the channels of the stereo background with increasing cross-talking factor until a double mono signal is obtained (cross-talking factor equals 1).

Ideal background scaling: A DE process with ideal quality is simulated, where dialog and background are separately available. Mixtures of them with varying background levels are created and clean speech is taken as reference.

Oracle time-frequency weighting: Given the STFT magnitude of the mixture $X(m,k)$ and the background $B(m,k)$, an ideal weighting matrix $G(m,k)$ for DE is computed as follows.

$$G(m,k) = 1-(1-\alpha)\frac{|B(m,k)|}{|X(m,k)|}, \qquad (5)$$

where the DE output is derived as $Y(m,k) = G(m,k)X(m,k)$, α controls the background level, and

[1] STFT with zero-padding is used. The length of the STFT is 2048 samples and squared Hann window with 50% overlap is employed.

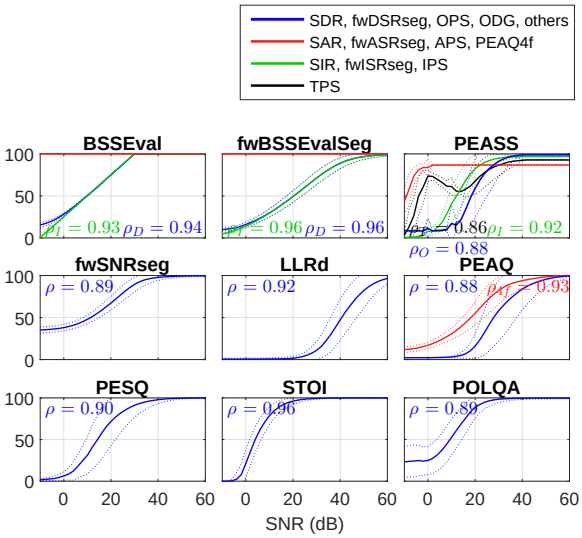

Fig. 1: Responses to additive noise. When more than one color per subplot is used, please refer to the legend. This holds also in the following.

m and k are time and frequency indices. This process introduces artifacts when $\alpha \neq 1$ due to the fact that $Y(m,k)$ is computed using the ideal spectral magnitudes, but the noisy phase spectra.

Reduced time-frequency resolution: The oracle weighting matrix $G(m,k)$ has a resolution of 21 ms and 23 Hz. In order to simulate the common case of poor resolution in the STFT domain, we decrease the time resolution with 23 linear steps of 21 ms down to 504 ms; at the same time, frequency resolution is decreased with 23 linear steps of 94 Hz down to 2.185 kHz. This causes artifacts that are often described as reverberant, pre-echo, ghost voice, or double voice.

Robustness with respect to delay and scaling: Constant delays and amplitude scaling are distortions that are not perceived as quality impairments and are used for testing the robustness of the performance metrics.

5 Evaluation

The test signals are 15 mixtures of stereo background signals and speech recordings (7 female and 8 male speakers in turn) panned to the center. The backgrounds comprise classical music, environmental recordings (e.g., rain, traffic, restaurant, applause), and panned direct signals (e.g., gunshots, helicopter flight, pitched

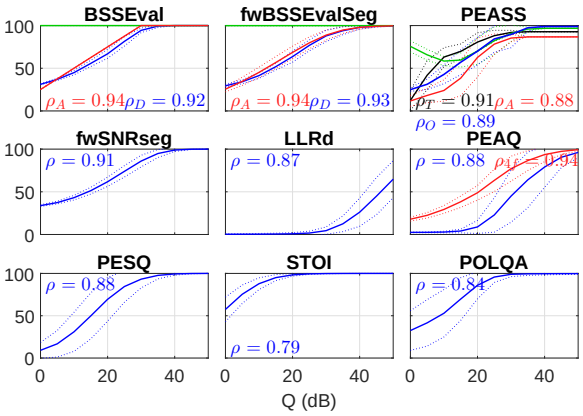

Fig. 2: Responses to MNRU.

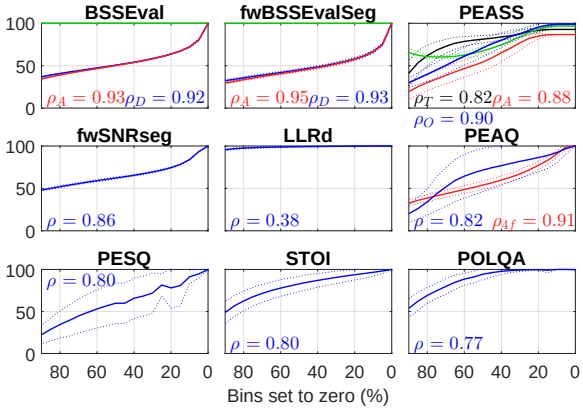

Fig. 3: Responses to musical noise.

pipe). Sampling frequency is 48 kHz and length is between 5 and 9 seconds for each signal.

Figs. 1-9 depict the response of the performance measures to the distortions via the mean over all test signals (solid lines) plus/minus the standard deviation (dotted lines) and the response score $\rho_{q,d}$ (simply denoted by ρ). The scores are also reported in Tables 3 and 4.

As shown by Fig. 1, all the measures but TPS respond monotonically to **additive noise** and span most of the range with different standard deviations. Even if APS, SAR, and fwASRseg are not supposed to respond to additive interferer, only SAR and fwASRseg are constant.

Fig. 2 illustrates the responses to **MNRU**. SIR and fwISRseg are constant as expected, while IPS is not. All the other measures respond monotonically but with different deviation, particularly high for POLQA.

The responses to **musical noise** are depicted in Fig. 3. BSSEval and fwBSSEvalSeg achieve the highest response scores for this distortion and behave almost identically (SAR and SDR overlap as well as fwASRseg and fwDSRseg, while SIR and fwISRseg are constant as expected). High scores are also realized by PEAQ4f, OPS, and fwSNRseg. LLRd spans only a very small portion of its range, and once more IPS is not constant as we would expect.

The objective measures responding to **low-pass filtering** are depicted in Fig. 4. SAR, fwASRseg, PESQ, and POLQA do not drop significantly even for low cut-off frequencies (so $\rho \simeq 0$). While the intelligibility

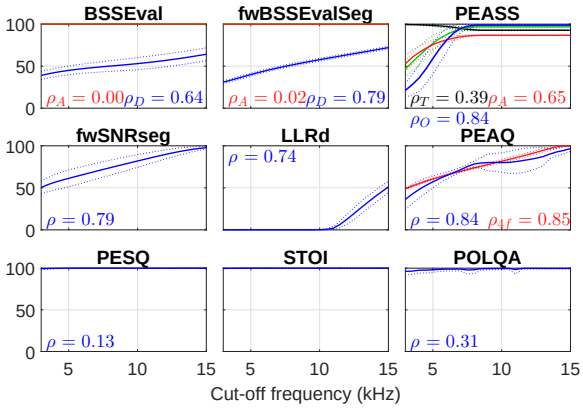

Fig. 4: Responses to low-pass filter.

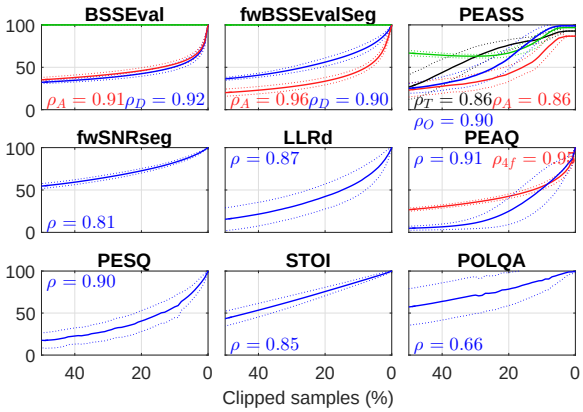

Fig. 5: Responses to clipping.

may not change (in fact STOI is constant), the overall quality is certainly degraded.

SIR, fwISRseg, and IPS are not supposed to be responsive to **clipping**. As can be observed in Fig. 5, SIR and fwISRseg are indeed constant, but IPS is not. All the other measures are sensitive to clipping with high response scores with the exception of POLQA that has once more high deviation.

Fig. 6 indicates that only SDR, fwDSRseg, OPS, fwSNRseg, PEAQ, and PEAQ4f are significantly sensitive to distortions to the **spatial image**.

The responses to the **ideal background scaling** are illustrated in Fig. 7. It can be observed that the multi-criteria evaluation tools (BSSEval, fwBSSEvalSeg, and PEASS) are particularly useful as they assess overall quality and background scaling separately. In particular, while APS stays on high values, IPS and OPS drop with a sigmoid shape. This sigmoid shape is followed also by the other perceptually motivated measures because they are affected by the background scaling, despite the ideal quality of the process.

DE via the **oracle time-frequency weighting** is examined by Fig. 8. It can be noted an overall decrease of the measured quality with respect to Fig. 7 due to imperfect decomposition. The difference is particularly highlighted by fwBSSEvalSeg, fwSNRseg, PEAQ, PEAQ4f, and PESQ. Also the background attenuation is less effective as measured by SIR, fwISRseg, and IPS. Stronger attenuation only comes at the cost of higher distortion. This double-faced nature can be described only by the multi-criteria tools.

Fig. 9 presents the measured performance as a function of the steps simulating **reduced time-frequency resolution** (the x-axis is simply labeled with step indices). Generally lower values of ρ are expected, still the relative values of ρ are of interest. BSSEval and fwBSSEvalSeg mostly ascribe this distortion to the interferer. Furthermore, it is peculiar that APS is not monotonic and that PEAQ, PEAQ4f, and PESQ do not vary much with the decreasing resolution.

Finally, regarding the **robustness with respect to delay and scaling**, all measures were found to ignore small time delays on the signal under test. On the other hand, SDR, fwDSRseg, PEAQ, and PEAQ4f tell apart the reference from scaled versions of itself even by small factors, e.g., 1 dB. If such a sensitivity is not desired, an additional level adaptation should be considered before these measures are computed.

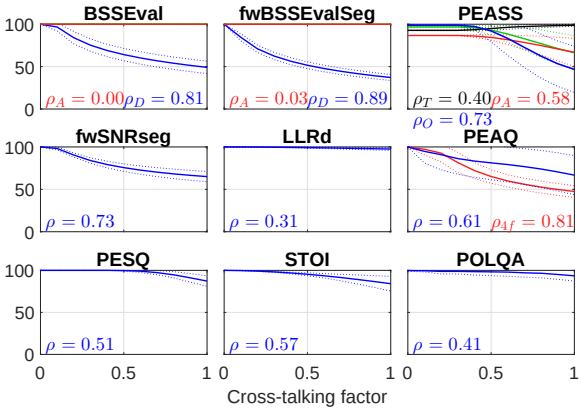

Fig. 6: Responses to spatial image distortion.

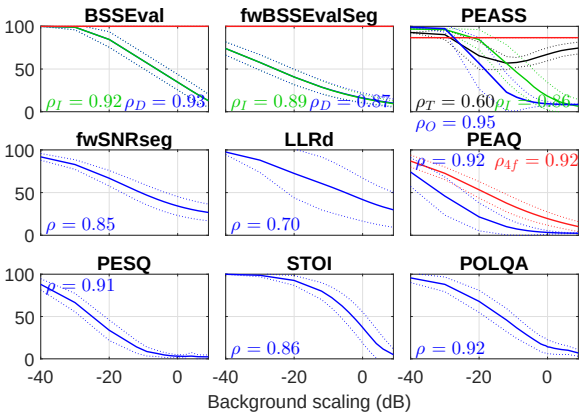

Fig. 7: Responses to ideal background scaling.

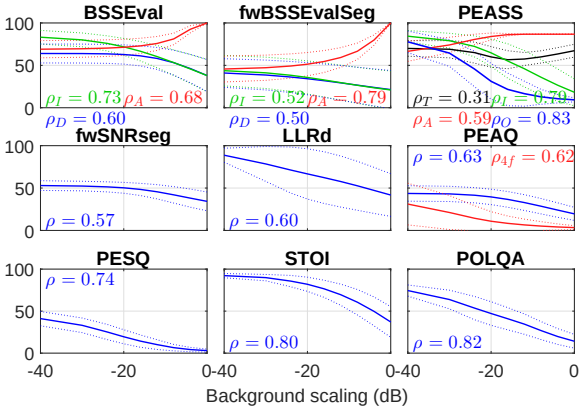

Fig. 8: Responses to oracle time-frequency weighting.

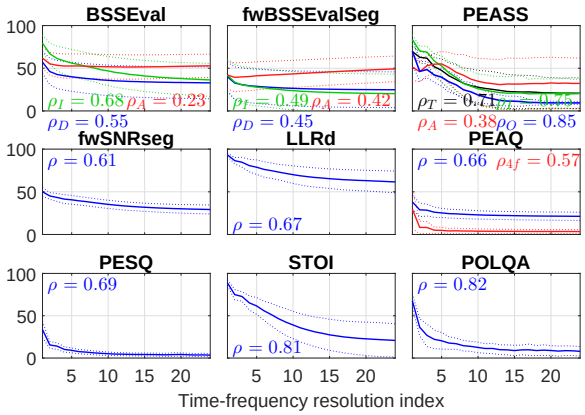

Fig. 9: Responses to reduced STFT resolution.

6 Conclusions

In this paper, state-of-the-art objective measures for the application in DE have been analyzed using artificial distortions. We have formulated and quantified three desired properties for objective measure, i.e. that the measures are monotonic functions of the intensity of the distortion, have low inter-item deviation, and span most of the range for which they are defined, and we have proposed a combined response score.

The presented analysis is proposed as an additional tool that complements the use of reference data obtained from listening tests. In future works, the subjective assessment of the distorted signals will be studied and compared with the objective criteria's responses shown in this work. Also, thanks to this data, an improved formulation of the response score is conceivable.

Depending on the system under test, only a subset of the proposed distortions may be relevant. Assuming that all of them are equally important, the mean values of the response score given in the last row of Tables 3 and 4 are of interest. For the overall quality, the highest mean values were achieved by OPS and PEAQ4f. It is important to note that the twofold aspect of background scaling against overall quality is assessed only by the multi-criteria tools such as BSSEval, fwBSSEvalSeg, and PEASS.

Even if not always monotonic, PEASS revealed to be a complete set of objective performance measures for DE. A reliable alternative set that comes at lower complexity is represented by PEAQ4f complemented by fwISRseg and fwASRseg or by SIR and SAR. If intelligibility is also of interest, STOI can be taken into account.

7 Acknowledgments

The authors sincerely thank Prof. Emanuël A.P. Habets, Dr.Sc. Jouni Paulus, M.Sc. Alexandra Craciun, and Dr.Sc. Jonathan Driedger for the fruitful discussions that led to substantial improvements of this paper.

References

[1] ITU-R Rec. BS.1387-1. Method for objective measurements of perceived audio quality, 2001.

[2] ITU-R Rec. BS.1534-3. Method for the subjective assessment of intermediate quality levels of coding systems, 2015.

[3] ITU-R Rec. BS.562-3. Subjective assessment of sound quality, 1990.

[4] L. Di Persia, D. Milone, H.L. Rufiner, and M. Yanagida. Perceptual evaluation of blind source separation for robust speech recognition. *Signal Processing*, 88(10), 2008.

[5] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann. Subjective and objective quality assessment of audio source separation. *IEEE Trans. Audio, Speech and Language Process.*, 19(7), 2011.

[6] H. Fuchs, S. Tuff, and C. Bustad. Dialogue enhancement - technology and experiments. *EBU Technical Review*, Q2, 2012.

[7] Yi Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *IEEE Trans. Audio, Speech and Language Process.*, 16(1), 2008.

[8] R. Huber and B. Kollmeier. PEMO-Q - a new method for objective audio quality assessment using a model of audatory perception. *IEEE Trans. Audio, Speech and Language Process.*, 14, 2006.

[9] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Trans. Acoust., Speech and Language Process.*, 23(1), 1975.

[10] P. Kabal. An examination and interpretation of ITU-R BS.1387: Perceptual evaluation of audio quality. Technical report, McGill University, 2002.

[11] T. Kastner. Evaluating physical measures for predicting the perceived quality of blindly separated audio source signals. In *Proc. AES 127th Conv.*, 2009.

[12] M.G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2), 1938.

[13] K. Kinoshita, M. Delcroix, S. Gannot, E.A.P. Habets, R. Haeb-Umbach, W. Kellermann, V. Leutnant, R. Maas, T. Nakatani, B. Raj, A. Sehr, and T. Yoshioka. 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.*, 2016(1), 2016.

[14] J. Kornycky, B. Gunel, and A. Kondoz. Comparison of subjective and objective evaluation methods for audio source separation. In *Proc. Meetings on Acoust.*, 2008.

[15] P.C. Loizou. *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.

[16] M. Mauch and S. Ewert. The audio degradation toolbox and its application to robustness evaluation. In *Proc. 14th Int. Soc. for Music Information Retrieval Conf.*, 2013.

[17] P. Mowlaee, R. Saeidi, M.G. Christensen, and R. Martin. Subjective and objective quality assessment of single-channel speech separation algorithms. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 2012.

[18] ITU-T Rec. P.800. Methods for subjective determination of transmission quality, 1996.

[19] ITU-T Rec. P.810. Modulated noise reference unit (MNRU), 1996.

[20] ITU-T Rec. P.830. Subjective performance assessment of telephone-band and wideband digital codecs, 1996.

[21] 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.

[22] ITU-T Rec. P.862.1. Mapping function for transforming P.862 raw results scores to MOS-LQO, 2003.

[23] ITU-T Rec. P.862.2. Wideband extension to recommendation P.862 for the assessment of wideband telephone networks and speech codecs, 2007.

[24] ITU-T Rec. P.863. Perceptual objective listening quality assessment, 2011.

[25] 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 Int. Conf. Acoust., Speech and Signal Process.*, 2004.

[26] C.H. Taal, R.C. Hendriks, R. Heusdens, and J. Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Audio, Speech and Language Process.*, 19(7), 2011.

[27] J.M. Tribolet, P. Noll, B. McDermott, and R.E. Crochiere. A study of complexity and quality of speech waveform coders. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, 1978.

[28] E. Vincent, R. Gribonval, and C. Févotte. Performance measurement in blind audio source separation. *IEEE Trans. Audio, Speech and Language Process.*, 14(4), 2006.

[29] E. Vincent, H. Sawada, P. Bofill, S. Makino, and J.P. Rosca. First stereo audio source separation evaluation campaign: data, algorithms and results. In *Proc. 8th Int. Conf. on Independent Component Analysis and Signal Separation*, 2007.

[30] 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.*, 14(6), 2006.

$\rho_{q,d}$	OPS	PEAQ4f	fwDSRseg	SDR	ODG	fwSNRseg	PESQ	POLQA	LLRd	**Mean**
Additive noise	0.88	0.93	0.96	0.94	0.88	0.89	0.90	0.89	0.92	**0.910**
MNRU	0.89	0.94	0.93	0.92	0.88	0.91	0.88	0.84	0.87	**0.896**
Musical noise	0.90	0.91	0.93	0.92	0.82	0.86	0.80	0.77	0.38	**0.810**
Low-pass	0.84	0.85	0.79	0.64	0.84	0.79	0.13	0.31	0.74	**0.659**
Clipping	0.90	0.95	0.90	0.92	0.91	0.81	0.90	0.66	0.87	**0.869**
Spatial image	0.73	0.81	0.89	0.81	0.61	0.73	0.51	0.41	0.31	**0.648**
Ideal scaling	0.95	0.92	0.87	0.93	0.92	0.85	0.91	0.92	0.70	**0.885**
Oracle weighting	0.83	0.62	0.50	0.60	0.63	0.57	0.74	0.82	0.60	**0.656**
Reduced resolution	0.85	0.57	0.45	0.55	0.66	0.61	0.69	0.82	0.67	**0.652**
Mean	**0.863**	**0.832**	**0.803**	**0.803**	**0.795**	**0.781**	**0.720**	**0.717**	**0.672**	

Table 3: Response score $\rho_{q,d}$ for overall quality measures. Measures are in order of decreasing mean $\rho_{q,d}$ (from left to right).

$\rho_{q,d}$	IPS	SIR	STOI	fwISRseg	APS	TPS	fwASRseg	SAR	**Mean**
Additive noise	0.92	0.93	0.96	0.96		0.86			**0.926**
MNRU			0.79		0.88	0.91	0.94	0.94	**0.890**
Musical noise			0.80		0.88	0.82	0.95	0.93	**0.875**
Low-pass					0.65	0.39	0.02	0	**0.265**
Clipping			0.85		0.86	0.86	0.96	0.91	**0.887**
Spatial image			0.57		0.58	0.40	0.03	0	**0.315**
Ideal scaling	0.86	0.92	0.86	0.89		0.60			**0.827**
Oracle weighting	0.79	0.73	0.80	0.52	0.59	0.31	0.79	0.68	**0.652**
Reduced resolution	0.75	0.68	0.81	0.49	0.38	0.71	0.42	0.23	**0.556**
Mean	**0.831**	**0.816**	**0.804**	**0.715**	**0.687**	**0.649**	**0.586**	**0.526**	

Table 4: Score $\rho_{q,d}$ for specific criteria. Empty entries if $\rho_{q,d}$ is not computed, i.e. when the measure is expected to be insensitive to the distortion. Measures are in order of decreasing mean $\rho_{q,d}$ (from left to right).