2021 Thirteenth International Conference on Quality of Multimedia Experience (QoMEX)

Measuring Speech Quality of System Input while Observing only System Output

Stephen Voran
Institute for Telecommunication Sciences
Boulder, Colorado, USA
svoran@ntia.gov

***Abstract*—We present a set of relatively small-scale proof-of-concept experiments where we construct no-reference (NR) speech quality estimators that give reliable values of system-under-test (SUT) input speech quality in spite of the fact that NR estimators can only access SUT output speech. We then explain why this success is not as counter-intuitive as it might initially seem. Next we demonstrate that this advance can be used to adjust NR relative speech quality values to obtain the much more desirable and useful NR absolute speech quality values. The experiments start with over seven hours of studio-quality speech. A processor adds filtering, reverberation, and noise to simulate the somewhat lower quality speech that often must be used to test systems. Four different established full-reference speech quality estimators provide ground-truth values for these experiments.**

***Index Terms*—full-reference, machine learning, no-reference, speech quality, subjective testing**

I. Experiment and Results

Objective estimators of speech quality are extremely useful in numerous situations where subjective testing is impractical. Full-reference (FR) estimators compare perceptually-motivated spectral representations of the system-under-test (SUT) input (reference) and output (test) speech signals in a manner that seeks to emulate human cognition. No-reference (NR) estimators observe only the SUT output (test) speech signal, so NR estimators can be applied in many cases where FR estimators cannot.

We construct NR estimators that measure the speech quality of the SUT input in spite of the fact that NR estimators can only see the SUT output. We describe our experimental set-up, data, and results, and explain why the results are not as counter-intuitive as they may seem. We then show how to build on these results to create improved NR estimators that track absolute quality rather than relative quality.

Fig. 1 shows the experiment set-up. We use FR estimates as ground-truth. The notation $Q(t, r)$ indicates an FR estimate of the quality of t (test speech) with respect to r (reference speech). For best generality we repeat all experiments with four different wideband (WB) FR speech quality estimators: WB-PESQ [1], POLQA [2], PEMO [3], and ViSQOL [4]. To allow convenient and equitable comparisons we normalize the FR estimates Q_{raw} to the range [0,1] using

$$Q(y, x) = \frac{Q_{raw}(y, x) - Q_{min}}{Q_{max} - Q_{min}}. \quad (1)$$

U.S. Government work not protected by U.S. copyright

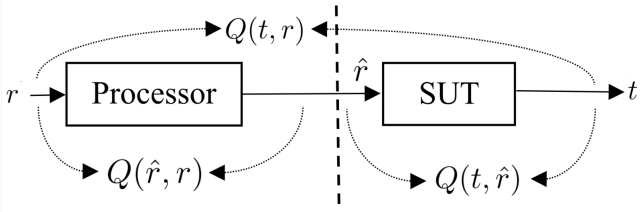

Fig. 1. Experiment set-up: r represents studio-quality speech, processor adds adjustable amounts of filtering, reverberation, and noise, $\hat{r}$ represents speech available in practice which may be studio-quality or lower quality. Three FR quality estimates used in the experiments are shown. Only entities to the right of the vertical dashed line are available in practical applications.

The functional maximum Q_{max} is found from a study of $Q(r, r)$. The functional minimum Q_{min} is found from a study of $Q(n, r)$ where n is white Gaussian noise with variance that matches the variance of r. Values are given in Table I.

We use WB speech (16,000 smp/s). The reference speech r comes from English-language studio-quality recordings from 24 talkers [5] segmented into 2907 different 3-second files that have a speech activity factor of 50% or greater.

The processor shown in Fig. 1 converts studio-quality signals to the type of signals that are often recorded in the field and available in speech databases. The amount of processing is adjustable from none to extreme. The processor can apply cut or boost at the low- and high-frequency ends of the spectrum to simulate imperfect microphones or microphone placement. It invokes one of nine different room impulse responses [6], [7] to produce reverberation at a selected ratio of direct-to-reverberant sound. It adds heating/cooling system noise at a selected SNR. We use ranges of all processor settings with a goal that the input quality $Q(\hat{r}, r)$ be roughly uniform on [0.5, 1.0] and we retain only files in this range. The number of files retained is different for each estimator as shown in Table I.

We use nine WB speech codec modes (codec/rate combinations) as SUTs in this work: EVS at 5.9, 9.6, and 16.4 kbps [8], AMR-WB at 6.6, 12.65, and 23.85 kbps [9], and Opus at 8, 12, and 16 kbps [10]. We randomly select one of the nine to process each $\hat{r}$ file. $Q(t, r)$ is the absolute quality and it reflects impairments from the processor and the SUT. $Q(t, \hat{r})$ is the relative quality and it responds primarily to the SUT.

We performed analysis-of-variance on the absolute quality $Q(t, r)$ for each FR estimator and calculated the variance due to the processor divided by the variance due to the SUT, shown as $\sigma^2_{Proc}/\sigma^2_{SUT}$ in Table I. POLQA finds that

TABLE I
FR ESTIMATOR AND DATASET PROPERTIES.

	WB-PESQ	POLQA	PEMO	ViSQOL
Q_{min}	1.04	1.01	0.39	1.00
Q_{max}	4.64	4.75	1.00	5.00
Number of Files	9077	13,682	14,251	13,660
Hours of Speech	7.6	11.4	11.9	11.4
$\sigma^2_{Proc}/\sigma^2_{SUT}$	0.27	1.15	1.91	1.75

our processor settings and SUT selections produce similar variance, WB-PESQ reports that the SUTs strongly dominate the total variance, while PEMO and ViSQOL find that the processor dominates. In spite of these expected individualized behaviors, the results that follow are fairly consistent across the four FR estimators.

We trained an NR estimator to emulate each of the three FR quality values in spite of the fact that NR estimators can only access t. The three NR estimators are distinguished by subscripts:

$$Q_{t,\hat{r}}(t) \approx Q(t,\hat{r}) \quad \text{(Relative Quality)}, \tag{2}$$

$$Q_{t,r}(t) \approx Q(t,r) \quad \text{(Absolute Quality)}, \tag{3}$$

$$Q_{\hat{r},r}(t) \approx Q(\hat{r},r) \quad \text{(Input Quality)}. \tag{4}$$

Note that the problem (2) and similar problems have been studied in [11]–[25] but the problems (3) and (4) are new. The importance of distinguishing relative quality $Q_{t,\hat{r}}(t)$ from absolute quality $Q_{t,r}(t)$ and their connection through input quality $Q_{\hat{r},r}(t)$ is addressed in Section II.

The NR estimators start by extracting the first eight mel-frequency cepstral coefficients (MFCCs) [26] from each frame (20 ms length, 50% overlap) of the three-second speech signal t. Single-frame differences (delta MFCCs) of these 8 are appended to give 16 values per frame. For each of these 16 we calculate 3 statistics across frames to arrive at a total of 48 features per speech signal. Those statistics are the 5%, 50%, and 95% values across the 298 frames. Our motivation is to extract robust measurements of the extreme and central values. Each NR estimator then uses the neural network shown in Table II to map these 48 features to a quality estimate. The network has 929 learnable parameters (mn weights and n biases in each $m \times n$ fully-connected layer). We trained the network with 90% of the files and validated its operation with the remaining 10%. The resulting Pearson correlations and root mean squared errors (RMSEs) are given in Table III.

These NR estimators use just 48 features that were selected for simplicity, not optimality. Those features are processed by relatively small, simple networks. We present them only as proof-of-concept results. Previous work [11]–[25] may address the problem (2) more thoroughly and effectively. But the simplistic, unoptimized approach we have used here produces surprisingly high correlations and low RMSEs, especially for the problem (4) — NR estimation of input speech quality. This suggests that further development might produce very rewarding results. Table III shows only slight drops in correlation and negligible RMSE increases between training and validation, suggesting this approach has the potential to generalize.

TABLE II
NETWORK THAT MAPS 48 MFCC FEATURES TO 1 NR QUALITY ESTIMATE.

Layer	Description
1	48 × 16 fully-connected
2	ReLU
3	16 × 8 fully-connected
4	ReLU
5	8 × 1 fully-connected

TABLE III
PEARSON CORRELATION (ρ) AND RMSE (ξ) SHOWING AGREEMENT OF NR ESTIMATOR AND FR GROUND-TRUTH FOR TRAINING / VALIDATION.

	WB-PESQ	POLQA	PEMO	ViSQOL
$\rho(Q_{t,\hat{r}}(t), Q(t,\hat{r}))$	.71 / .66	.72 / .69	.78 / .77	.74 / .73
$\rho(Q_{t,r}(t), Q(t,r))$	.74 / .69	.77 / .76	.83 / .80	.83 / .80
$\rho(Q_{\hat{r},r}(t), Q(\hat{r},r))$	.85 / .84	.94 / .94	.92 / .90	.92 / .91
$\xi(Q_{t,\hat{r}}(t), Q(t,\hat{r}))$	.12 / .13	.11 / .11	.06 / .06	.05 / .05
$\xi(Q_{t,r}(t), Q(t,r))$	.11 / .11	.10 / .11	.07 / .07	.05 / .05
$\xi(Q_{\hat{r},r}(t), Q(\hat{r},r))$	.07 / .07	.05 / .05	.04 / .04	.06 / .06

Successfully measuring input speech quality using only output speech may appear counter-intuitive at first. The key is that the impairments in the input speech (filtering, reverberation, and noise) are largely distinguishable from the impairments caused by the SUT (coding artifacts). We argue that this situation is realistic and even common (but not ubiquitous) when measuring multimedia QoE. The SUT significantly modifies the input speech and any impairments that it carries but input speech impairments are sufficiently present and recognizable in the output speech signal that $Q_{\hat{r},r}(t)$ can detect, quantify, and map them to levels of input speech quality.

II. APPLICATION: ABSOLUTE SPEECH QUALITY

We have demonstrated the ability to measure input speech quality using only the SUT output speech signal. Input speech quality values are intrinsically useful but we focus now on exploiting them as a bridge from relative to absolute NR estimates of SUT output speech quality.

Users experience absolute speech quality $Q(t,r)$, not relative speech quality $Q(t,\hat{r})$, and we naturally seek to measure what users experience. The difference can be easily appreciated by comparing the squares with the circles in Fig. 2. Absolute speech quality includes all impairments while relative speech quality includes only those added by the SUT.

Machine learning is extremely useful for creating NR estimators but it requires large amounts of speech and ground-truth quality values. Absolute category rating subjective test scores are a desired ground-truth, but amassing sufficient quantities of them remains a challenge, even when crowd-sourcing is exploited. Thus augmenting [18] or substituting [20], [24] subjective scores with FR estimates may be necessary. In principle one could use the framework of Fig. 1. In practice, available databases are often recorded in less-than-ideal field conditions which means that $\hat{r}$ is available but r is not. Thus the absolute quality $Q(t,r)$ is not available.

When $\hat{r}$ is studio-quality, $\hat{r} = r$ and absolute and relative quality are the same. But when $\hat{r}$ is imperfect they differ, and using relative quality as ground-truth leads to NR estimators that produce relative quality values rather than the desired

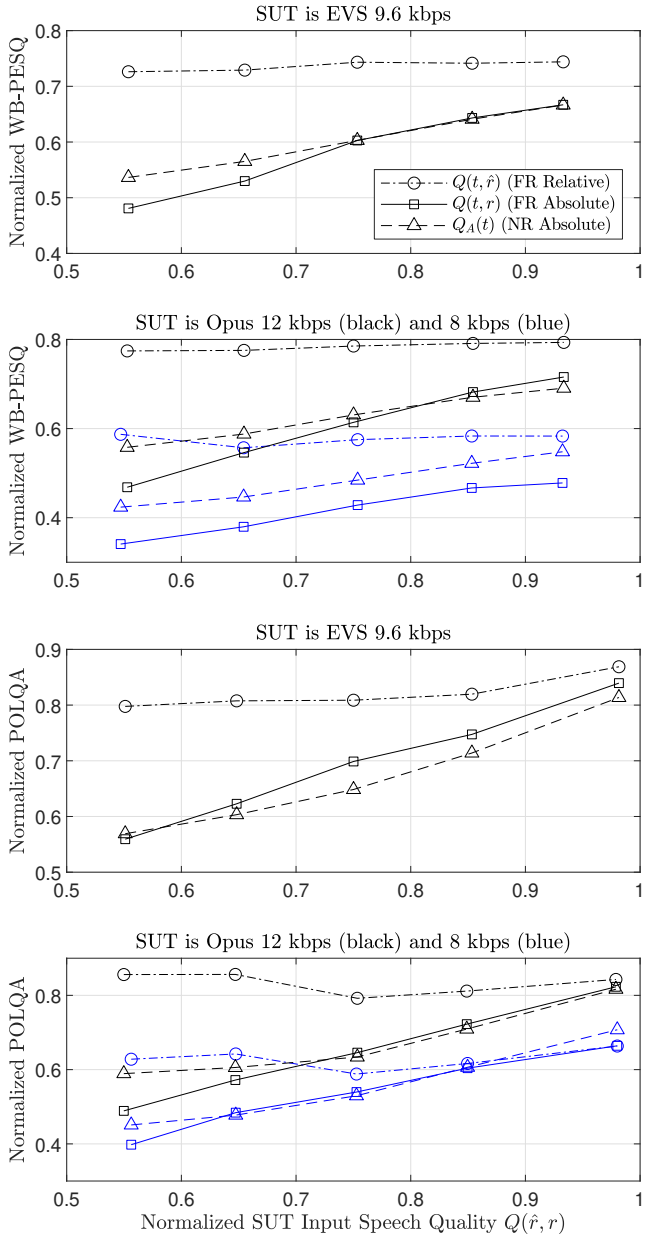

Fig. 2. Example normalized speech quality results: SUT output as a function of SUT input for three SUTs and two estimators. FR absolute quality (squares) responds to SUT and input speech quality but FR relative quality (circles) is largely invariant to input speech quality. NR absolute quality (triangles) achieves the goal of tracking FR absolute quality. Input speech quality values have been binned using five uniform bins between 0.5 and 1.0, then averaged. Average number of files per marker is 182 for WB-PESQ and 274 for POLQA.

absolute quality values. We now propose a way to build NR estimators of absolute quality, even when $\hat{r}$ is imperfect.

The E-Model is a successful and enduring additive model for speech impairments [27]–[29] and this motivates us to explore the additivity of impairments. We define "impairment" to be the decrease from the top of the quality scale, we weight and add the impairments of the processor and the SUT, and we include a constant term. This gives

$$1 - Q(t,r) \approx w_0 + w_1(1 - Q(t,\hat{r})) + w_2(1 - Q(\hat{r},r)), \quad (5)$$

which implies

$$Q(t,r) \approx w_1 Q(t,\hat{r}) + w_2 Q(\hat{r},r) + w_3, \quad (6)$$

where $w_3 = 1 - (w_0 + w_1 + w_2)$. That is, absolute quality may be approximated using relative quality and input quality. To test these assumptions and approximations we apply least-squares fitting to the measured values of $Q(t,r)$, $Q(t,\hat{r})$, and $Q(\hat{r},r)$ to find values for w_1, w_2, and w_3. Results are shown in Table IV. In the present experiment this simple model consistently explains significant amounts of variance in absolute quality.

TABLE IV
LEAST-SQUARES VALUES FOR WEIGHTS IN (6) AND VARIANCE EXPLAINED.

	WB-PESQ	**POLQA**	**PEMO**	**ViSQOL**
w_1	0.741	0.533	0.560	0.675
w_2	0.508	0.677	0.667	0.246
w_3	-0.341	-0.288	-0.300	-0.033
Variance Explained	90%	92%	93%	79%

This result motivates us to use NR estimates of relative quality and input quality to calculate NR absolute quality $Q_A(t)$. Combining (2), (4), and (6) gives

$$Q(t,r) \approx Q_A(t) = w_1 Q_{t,\hat{r}}(t) + w_2 Q_{\hat{r},r}(t) + w_3. \quad (7)$$

Fig. 2 shows that $Q_A(t)$ (triangles) does indeed follow the trends of absolute quality (squares). Space limitations prevent the display of detailed results for all 9 codec modes and 4 FR estimators, but they can be summarized by correlations across 45 data points (5 input speech quality levels $\times$ 9 codec modes) as shown in Table V. Our goal is absolute quality values and Table V allows us to compare three achievable alternatives to that goal. In this set of experiments, calculated NR absolute quality $Q_A(t)$ given in (7) is consistently by far the best option.

TABLE V
PEARSON CORRELATION (TRAINING / VALIDATION) BETWEEN FR ABSOLUTE QUALITY $Q(t,r)$ (UNAVAILABLE IN PRACTICE) AND THREE PRACTICAL ALTERNATIVES: FR RELATIVE QUALITY $Q(t,\hat{r})$, NR RELATIVE QUALITY $Q_{t,\hat{r}}(t)$, AND NR ABSOLUTE QUALITY $Q_A(t)$.

	WB-PESQ	**POLQA**	**PEMO**	**ViSQOL**
$Q(t,\hat{r})$	.81 / .81	.55 / .53	.81 / .76	.75 / .76
$Q_{t,\hat{r}}(t)$	.78 / .75	.46 / .43	.84 / .77	.82 / .80
$Q_A(t)$	.96 / .94	.91 / .92	.93 / .88	.89 / .90

Our proposal is to train NR estimators to estimate both input and output speech quality. In use, such estimators can invoke (7) to give absolute speech quality, $Q_A(t)$. The prime obstacle is the lack of input speech quality targets $Q(\hat{r},r)$ for training the NR estimator of input speech quality $Q_{\hat{r},r}(t)$. We have listened to speech from numerous databases and suggest that many databases can be characterized by a single quality value (or in some cases one quality value per talker) associated with the recording environment and equipment. We propose assigning this single informal subjective quality value to $Q(\hat{r},r)$ for each applicable group of files. This would provide targets for training $Q_{\hat{r},r}(t)$ which in turn would enable the no reference absolute speech quality estimator $Q_A(t)$.

REFERENCES

[1] Recommendation ITU-T P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," International Telecommunication Union, Geneva, 2007.

[2] Recommendation ITU-T P.863, "Perceptual objective listening quality prediction," International Telecommunication Union, Geneva, 2018.

[3] V. Emiya, E. Vincent, N. Harlander, and V. Hohmann, "Subjective and objective quality assessment of audio source separation," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2046–2057, Sep. 2011. Code at: http://bass-db.gforge.inria.fr/peass/

[4] A. Kokaram A. Hines, J. Skoglund and N. Harte, "ViSQOL: An objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, May 2015. Code at http://www.mee.tcd.ie/~sigmedia/Resources/ViSQOL

[5] P. Kabal, "Telecommunications and signal processing laboratory speech database," Database at http://www-mmsp.ece.mcgill.ca/documents/data/

[6] M. Jeub, M. Schäfer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Proc. Int. Conf. on Digital Signal Processing*, Santorini, Greece, July 2009, IEEE, IET, EURASIP.

[7] M. Jeub, M. Schäfer, H. Krüger, C. Nelke, C. Beaugeant, and P. Vary, "Do we need dereverberation for hand-held telephony?," in *Proc. Int. Congress on Acoustics*, Sydney, Australia, Aug. 2010, Australian Acoustical Society.

[8] Technical Specification ETSI 26.442, "Codec for enhanced voice services (EVS); ANSI C Code (fixed-point)," European Telecommunications Standards Institute, Sophia-Antipolis, Cedex, France, 2014.

[9] Technical Specification ETSI 26.171, "Speech codec speech processing functions; Adaptive Multi-Rate - Wideband (AMR-WB) speech codec; General description," European Telecommunications Standards Institute, Sophia-Antipolis, Cedex, France, 2011.

[10] IETF RFC 6716, "Definition of the Opus audio codec," Internet Engineering Task Force, 2012.

[11] T. Falk and W. Chan, "Single-ended speech quality measurement using machine learning methods," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, no. 6, pp. 1935–1947, Nov. 2006.

[12] S. Fu, Y. Tsao, H. Hwang, and H. Wang, "Quality-Net: An end-to-end non-intrusive speech quality assessment model based on BLSTM," in *Proc. Interspeech*, 2018, pp. 1873–1877.

[13] H. Salehi, D. Suelzle, P. Folkeard, and V. Parsa, "Learning-based reference-free speech quality measures for hearing aid applications," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 26, no. 12, pp. 2277–2288, Dec. 2018.

[14] C. Spille, S. Ewert, B. Kollmeier, and B. Meyer, "Predicting speech intelligibility with deep neural networks," *Computer Speech & Language*, vol. 48, pp. 51 – 66, Mar. 2018.

[15] R. Huber, M. Krüger, and B. T. Meyer, "Single-ended prediction of listening effort using deep neural networks," *Hearing Research*, vol. 359, pp. 40 – 49, Mar. 2018.

[16] P. Seetharaman, G. Mysore, P. Smaragdis, and B. Pardo, "Blind estimation of the speech transmission index for speech quality prediction," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2018, pp. 591–595.

[17] A. Avila, H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Non-intrusive speech quality assessment using neural networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 7125–7129.

[18] G. Mittag and S. Möller, "Non-intrusive speech quality assessment for super-wideband speech communication networks," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 7125–7129.

[19] J. Ooster and B. Meyer, "Improving deep models of speech quality prediction through voice activity detection and entropy-based measures," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2019, pp. 636–640.

[20] A. Catellier and S. Voran, "WEnets: A convolutional framework for evaluating audio waveforms," *arXiv e-prints*, arXiv:1909.09024, Sep. 2019.

[21] H. Gamper, C. Reddy, R. Cutler, I. Tashev, and J. Gehrke, "Intrusive and non-intrusive perceptual speech quality assessment using a convolutional neural network," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2019, pp. 85–89.

[22] J. Santos and T. Falk, "Towards the development of a non-intrusive objective quality measure for dnn-enhanced speech," in *Proc. Eleventh Int. Conf. on Quality of Multimedia Experience*, 2019.

[23] G. Mittag and S. Möller, "Quality degradation diagnosis for voice networks — Estimating the perceived noisiness, coloration, and discontinuity of transmitted speech," in *Proc. Interspeech*, 2019, pp. 3426–3430.

[24] A. Catellier and S. Voran, "WAWEnets: A no-reference convolutional waveform-based approach to estimating narrowband and wideband speech quality," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2020, pp. 331–335.

[25] G. Mittag, R. Cutler, Y. Hosseinkashi, M. Revow, S. Srinivasan, N. Chande, and R. Aichner, "DNN no-reference PSTN speech quality prediction," in *Proc. Interspeech*, 2020.

[26] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[27] N. Johannesson, "The ETSI computation model: A tool for transmission planning of telephone networks," *IEEE Communications Magazine*, vol. 35, no. 1, pp. 70–79, Jan. 1997.

[28] Recommendation ITU-T G.107, "The E-model: A computational model for use in transmission planning," International Telecommunication Union, Geneva, 2015.

[29] D. Rodríguez, D. Carrillo, M. Ramírez, P. Nardelli, and S. Möller, "Incorporating wireless communication parameters into the E-model algorithm," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 29, pp. 956–968, 2021.