# AN EFFECTIVE QUALITY EVALUATION PROTOCOL FOR SPEECH ENHANCEMENT ALGORITHMS

*John H.L. Hansen* and *Bryan L. Pellom*
Robust Speech Processing Laboratory
Duke University, Box 90291, Durham, NC 27708-0291
http://www.ee.duke.edu/Research/Speech jhlh@ee.duke.edu bp@ee.duke.edu

## ABSTRACT

Much progress has been made in speech enhancement algorithm formulation in recent years. However, while researchers in the speech coding and recognition communities have standard criteria for algorithm performance comparison, similar standards do not exist for researchers in speech enhancement. This paper discusses the necessary ingredients for an effective speech enhancement evaluation. We propose that researchers use the evaluation core test set of TIMIT (192 sentences), with a set of noise files, and a combination of objective measures and subjective testing for broad and fine phone-level quality assessment. Evaluation results include overall objective speech quality measure scores, measure histograms, and phoneme class and individual phone scores. The reported results are meant to illustrate specific ways of detailing quality assessment for an enhancement algorithm.

## 1. Introduction

Enhancement of speech in the presence of additive continuous broadband noise remains a challenging task, especially in moderate to high noise levels (SNRs -5 to 10 dB). A speech enhancement algorithm can be viewed as successful if it (i) suppresses perceivable background noise, and (ii) preserves or enhances perceived signal quality. Several surveys of speech enhancement exist[4, 2, 5, 6], though most traditional algorithms are based on optimizing mathematical criteria, which in general are not well correlated with speech perception. In general, these have not been as successful in preserving or improving quality in all regions of speech, especially transitional and unvoiced. It has also been difficult to compare the performance of speech enhancement algorithms, since most papers consider example evaluations on a few sentences degraded usually with white Gaussian noise (WGN). Performance is influenced by the (1) specific type of noise (2) specific SNR, (3) noise estimate updates and (4) algorithm parameter settings. In this paper, we propose to establish a test evaluation protocol which researchers in the field of speech enhancement can use to evaluate their algorithms. This protocol includes a standard set of test sentences, noise files, and software which will be made available via WWW for all to access and use. Because of the number of issues to be considered, this paper will highlight some of the main assessment steps (a more complete presentation is found in [1]).

## 2. Speech Quality vs. Time

When we consider noise reduction, we normally think of improving a signal-to-noise ratio (SNR). This may not be the most appropriate performance criterion for speech enhancement. All listeners have an intuitive understanding of speech quality, intelligibility and listener fatigue. However, these areas are not easy to quantify. A good overview of subjective quality testing methods and objective speech quality measures can be found in [14, 2]. For example, in an evaluation by Quackenbush [14], 322 types of distortion were considered in the evaluation of over 2000 objective measures of speech quality. Although SNR was evaluated only over waveform coder distortions, it's ability to predict subjective speech quality was very poor (correlation coefficient of 0.24 with DAM). It has also been shown in a number of studies that the impact of noise on degraded speech quality is non-uniform[12, 8] (see Fig. 1). An objective speech quality measure shows the level of distortion for each frame across time. Since speech frequency content varies across time due to the sequence of phonemes needed to produce the sentence, the impact of background distortion will also vary (note increased distortion levels for /n/, /r/, and stop closure /tcl/). This variable impact of noise on speech quality leads us to conclude that some phone classes are effected more than others when produced in a noisy environment. Our goals therefore are to emphasize the importance of employing a systematic diagnostic test procedure which would allow one to assess the impact of background noise or speech enhancement performance across individual phonemes or phoneme classes. We suggest that a combination of subjective and objective speech quality measures, applied to an easily accessible speech corpus test set, represents the most effective way to assess the impact of background noise, and quantify quality improvement for speech enhancement algorithms.

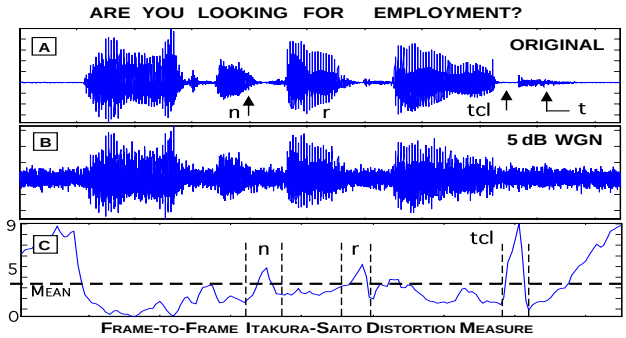


Figure 1: Variable impact of noise on speech quality across phonemes. Speech waveforms of (A) original, (B) degraded with white Gaussian Noise, and (C) IS quality measure versus time are shown. Overall mean IS and sample phone distortion are highlighted in the IS plot.

## 3. Objective Quality Assessment

Objective methods rely on a mathematically based measure between the original and coded/degraded speech signal. The success of these measures rests with their correlation with subjective quality.

**Itakura–Saito Distortion Measure:** For an original clean frame of speech with linear prediction (LP) coefficient vector, $\vec{a}_\phi$, and processed speech coefficient vector, $\vec{a}_d$, the Itakura–Saito distortion measure is given by,

$$d_{IS}(\vec{a}_d, \vec{a}_\phi) = \left[\frac{\sigma_\phi^2}{\sigma_d^2}\right]\left[\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right] + \log\left(\frac{\sigma_d^2}{\sigma_\phi^2}\right) - 1 \quad (1)$$

where $\sigma_d^2$ and $\sigma_\phi^2$ represent the all–pole gains for the processed and clean speech frame respectively.

**Log–Likelihood Ratio Measure:** The LLR measure is also refered to as the Itakura distance (note that the IS measure incorporates the gain estimate using variance terms, while the LLR does not; this influences how each measure emphasizes differences in general spectral shape versus an overall gain offset). The LLR measure is found as follows,

$$d_{LLR}(\vec{a}_d, \vec{a}_\phi) = \log\left(\frac{\vec{a}_d R_\phi \vec{a}_d^T}{\vec{a}_\phi R_\phi \vec{a}_\phi^T}\right). \quad (2)$$

**Log–Area–Ratio Measure:** The LAR measure is also based on dissimilarity of LP coefficients between original and processed speech signals. The log-area-ratio parameters are obtained from the $P^{th}$ order LP reflection coefficients for the original $r_\phi(j)$ and processed $r_d(j)$ signals for frame $j$. The objective measure is formed as follows,

$$d_{LAR} = \left|\frac{1}{M}\sum_{i=1}^{M}\left[\log\frac{1+r_\phi(j)}{1-r_\phi(j)} - \log\frac{1+\hat{r}_d(j)}{1-\hat{r}_d(j)}\right]^2\right|^{\frac{1}{2}} \quad (3)$$

**Segmental SNR Measure:** Since the correlation of SNR with subjective quality is so poor, it is of little interest as a general objective measure of speech quality[14]. Instead, we choose the frame-based segmental SNR which is a reasonable measure of speech quality. It is formed by averaging frame level SNR estimates as follows,

$$d_{SEGSNR} = \frac{10}{M}\sum_{m=0}^{M-1}\log\frac{\sum_{n=Nm}^{Nm+N-1} s_\phi^2(n)}{\sum_{n=Nm}^{Nm+N-1}\left[s_d(n) - s_\phi(n)\right]^2} \quad (4)$$

Frames with SNRs above 35dB do not reflect large perceptual differences, and generally can be replaced with 35dB in Eq. 4. Likewise, during periods of silence, SNR values can become very negative since signal energies are small. These frames do not truly reflect the perceptual contributions of the signal. Therefore, a lower threshold is often set to provide a bound on frame based SNR (we select -10dB, but the range of (0,-20dB) has been suggested[13]).

**Weighted Spectral Slope Measure:** The WSS measure by Klatt (1982) is based on an auditory model in which 36 overlapping filters of progressively larger bandwidth are used to estimate the smoothed short-time speech spectrum. The measure finds a weighted difference between the spectral slopes in each band. The magnitude of each weight reflects whether the band is near a spectral peak or valley, and whether the peak is the largest in the spectrum. A per-frame measure in decibels is found as,

$$d_{WSS}(j) = K_{spl}(K - \hat{K}) + \sum_{k=1}^{36} w_a(k)\left(S(k) - \hat{S}(k)\right)^2. \quad (5)$$

where $K, \hat{K}$ are related to overall sound pressure level of the original and enhanced utterances, and $K_{\text{spl}}$ is a parameter which can be varied to increase overall performance.

## 4. Subjective Quality Assessment

The range of subjective testing schemes include those early methods which focused on speech intelligibility, and those which focus on overall quality. Intelligibility tests include the *modified rhyme test (MRT)*, and the *diagnostic rhyme test (DRT)*. Isometric absolute judgement tests attempt to evaluate more than just intelligibility, such as aspects of overall quality. Some of these tests include; the 'Goodness' test, *mean opinion score (MOS)* tests, and the *paired acceptability rating (PAR)* method. Another test which evaluates speech and background signal quality across multiple scales is the *diagnostic acceptability measure (DAM)*[14]. While these tests have been used to evaluate many voice communication systems, it is important to note that they may be valid only for restricted distortions.

We emphasize that quality assessment of speech enhancement algorithms can not be achieved *without* formal subjective testing. It is therefore proposed that a combination of two test procedures be used. First, in order to assess perceived quality, a subjective MOS test be performed. This allows for an overall assessment of quality. However, since many enhancement algorithms also introduce processing artifacts, listeners may prefer one processing artifact more than another. To assess this, a subjective Pairwise Preference Test (PPT) should be employed. A series of pairwise randomized processed sentences are presented and listeners simply select the one they prefer. For the PPT, testing should also include the original degraded speech.

## 5. Enhancement Algorithms

The focus of this paper is on an assessment protocol for speech enhancement algorithms. In our study[1], we consider ten enhancement algorithms. Because of space limitations, only three of the methods are discussed here.

**Nonlinear Spectral Subtraction:** The NSS [11] algorithm takes into account frequency-dependent SNR. Using a nonlinear subtractor, the subtraction factor is reduced for spectral components of high SNR and increased for spectral components of low SNR. In addition, the noise model is extended by utilizing both an averaged noise spectrum and an overestimated noise spectrum. The NSS enhancement can be expressed in terms of a filtering operation,

$$|\hat{X}_i(\omega)| = H_i(\omega) \cdot |Y_i(\omega)|, \quad (6)$$

where $H_i(\omega)$ depends on a smoothed estimate of the noisy speech magnitude spectrum, $|\ddot{Y}_i(\omega)|$, and nonlinear subtraction term, $\Phi_i(\omega)$,

$$H_i(\omega) = \frac{|\ddot{Y}_i(\omega)| - \Phi_i(\omega)}{|\ddot{Y}_i(\omega)|}. \quad (7)$$

The subtraction term, $\Phi_i(\omega)$, is given by,

$$\Phi_i(\omega) = \frac{\max\limits_{i-M\le\tau\le i}(|N_\tau(\omega)|)}{1+\gamma\rho_i(\omega)} \quad \text{with} \quad \rho_i(\omega) = \frac{|\ddot{Y}_i(\omega)|}{|\ddot{N}_i(\omega)|} \quad (8)$$

where $\gamma$ is a constant scaling factor dependent on the range of $p_i(\omega)$. The dynamic range of $\Phi_i(\omega)$ is limited to $|\ddot{N}_i(\omega)| \le \Phi_i(\omega) \le 3|\ddot{N}_i(\omega)|$.

**MMSE Estimator:** Ephraim and Malah [7] proposed this estimator for the short-time spectral amplitude component of speech in noise. Here, the speech and noise spectral components are modeled as Gaussian random variables. The algorithm estimates the $k^{th}$ spectral magnitude component using a filter of the form,

$$H_i(\omega_k) = \left(\frac{\sqrt{\pi}}{2}\right)\left(\frac{\sqrt{v_k}}{\gamma_k}\right)\exp\left(-\frac{v_k}{2}\right) \cdot \left[(1+v_k)\,I_o\left(\frac{v_k}{2}\right) + v_k I_1\left(\frac{v_k}{2}\right)\right] \quad (9)$$

where $I_o(\cdot)$ and $I_1(\cdot)$ represent modified Bessel functions of the zero and first order. Furthermore, $v_k$ is given by,

$$v_k = \frac{\xi_k}{1+\xi_k}\gamma_k \quad (10)$$

where $\xi_k$ and $\gamma_k$ represent the *a priori* and *a posteriori* signal-to-noise ratios for the $k^{th}$ spectral component.

**Noise–Adaptive Auto-LSP:** The Constrained Iterative Wiener Filter (Auto–LSP) approach [8] is based upon a two-step maximum *a posteriori* (MAP) estimation of the all-pole speech parameters and noise-free speech. Between MAP estimation steps, spectral constraints are applied in order to ensure (i) stability of the all-pole model, (ii) that it possesses speech-like characteristics, and (iii) frame-to-frame continuity in vocal tract characteristics. *Inter-frame* constraints are applied to the Line Spectrum Pair (LSP) parameters while *intra-frame* constraints are applied across iterations to the autocorrelation lag sequence. For slowly varying colored noise, it is useful to adapt the constraints to the frequency range dominated by the noise. An extension was therefore developed, referred to here as *Noise Adaptive (Auto:I,LSP:T)*, which operates on subbanded signal components in which the terminating iteration is adjusted based on the *a posteriori* estimate of the signal-to-noise ratio in each signal subband. The enhanced speech is formulated as a combined estimate from individual signal subband estimators.

# 6. Results

In the evaluation protocol, we propose using the 192 sentence core TIMIT test set, and a collection of ten noise sources[10] (results are presented here for white Gaussian noise: WGN, and automobile highway noise: HWY). The evaluation begins by degrading and processing each sentence at SNRs of [0, 5, 10, 15dB]. We note at this time that source code to degrade speech, many of the enhancement routines, and all objective quality routines are available from our web site in 'C'. The intent here is to provide a common evaluation test platform for developers of speech enhancement algorithms[1]. Our evaluations here focus on the type of results which are important in assessing enhancement performance.

## 6.1. Objective Quality Results

Objective quality measure results are presented in four areas. We note that there are several ways to obtain overall quality scores. For most measures, finding a mean across a large test set is reasonable. If users want a general measure of performance the median of the resulting frame-level scores is more useful (a mean quality measure is typically biased by a few frames in the tails of the quality measure distribution). Another way to get a reasonable overall measure is to either (i) find the mean with outliers greater than $5\sigma$ removed from mean calculation $m_{5\sigma}$, or (ii) find the mean using the first 95% of the frames, $m_{95\%}$. This allows for the removal of a fixed number of frames which may have unrealistically high distortion levels (this is equivalent to limiting SegSNR to a perceptual meaningful range). All mean values here (except SegSNR) use the $m_{95\%}$ mean.

| Highway Noise, 5 dB SNR | | | | | |
|---|---|---|---|---|---|
| | IS | LLR | LAR | SegSNR | WSS |
| Degraded | 0.50 | 0.33 | 5.14 | -0.88 | 41.9 |
| MMSE | 0.39 | 0.25 | 3.60 | +3.55 | 39.8 |
| NA-AutoLSP | 0.47 | 0.27 | 3.94 | +3.59 | 40.1 |
| NSS | 0.42 | 0.23 | 3.21 | +4.15 | 39.1 |
| **White Gaussian Noise, 5 dB SNR** | | | | | |
| | IS | LLR | LAR | SegSNR | WSS |
| Degraded | 2.76 | 1.23 | 6.81 | -1.30 | 42.0 |
| MMSE | 1.63 | 0.91 | 5.71 | +2.39 | 40.7 |
| NA-AutoLSP | 1.71 | 1.02 | 6.10 | +1.76 | 43.9 |
| NSS | 1.61 | 0.82 | 5.39 | +3.04 | 47.4 |

Table 1: $m_{95\%}$ objective speech quality scores across speech sections for TIMIT core test set (192 sentences).

**Overall Performance:** Table 1 summarizes the five objective measures for two noise sources for original degraded and with each of the three enhancement algorithms across the 192 TIMIT core sentence set. We see that all three enhancement routines provide quality improvement. Since WGN degrades the entire frequency band (whereas HWY is mostly low-freq.), the starting distortion level is higher.

| Sound Type | Segmental SNR (dB) | | | |
|---|---|---|---|---|
| | (a) | (b) | (c) | #frames |
| Silence | -9.75 | -9.27 | -8.46 | 11,127 |
| Vowel | 3.38 | 8.26 | 8.17 | 22,471 |
| Nasal | -5.15 | 0.06 | 1.64 | 4,784 |
| Stop | -6.13 | -2.89 | -1.87 | 12,474 |
| Fricative | -6.42 | -2.09 | -0.42 | 12,429 |
| Semivowel | 3.50 | 8.56 | 8.75 | 6,237 |
| Diphthongs | 6.07 | 10.79 | 10.5 | 6,514 |
| Total (- Silence) | -0.88 | 3.55 | 4.15 | 67,154 |

Table 2: Segmental SNR across broad phoneme classes for TIMIT core test set (192 sentences) degraded by additive automobile highway noise at 5 dB SNR for (a) original degraded (b) MMSE enhanced, and (c) NA-AutoLSP enhanced.

**Phone-Class Performance:** Another way to explore performance is across phone-classes as summarized in Table 2. Such a comparison allows us to identify where an enhancement algorithm is functioning well and where further improvement is needed. Since these are SegSNR scores, larger values reflect higher quality (for IS and other measures, values closer to 0.0 are better). Here we see that MMSE does better for vowels and semivowels, while NA-AutoLSP does better for nasals, fricatives, and stops.

| ITAKURA-SAITO OBJECTIVE SPEECH QUALITY ACROSS AMERICAN PHONEMES | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Phoneme* | | Deg. | MMSE | NA-AutoLSP | NSS | *Phoneme* | | Deg. | MMSE | NA-AutoLSP | NSS |
| *CONSONANTS – nasal, fricatives, stops* | | | | | | *VOWELS, DIPHTHONGS, SEMIVOWELS* | | | | | |
| /m/ | m̲e | 0.348 | 0.319 | 1.003 | 0.617 | /iy/ | h̲eed | 0.162 | 0.112 | 0.128 | 0.265 |
| /s/ | s̲ip | 0.740 | 0.653 | 0.452 | 0.762 | /ao/ | a̲ll | 0.138 | 0.089 | 0.079 | 0.200 |
| /z/ | z̲ip | 0.786 | 0.648 | 0.464 | 0.748 | /ix/ | deb̲it | 0.254 | 0.185 | 0.245 | 0.356 |
| /k/ | k̲ey | 0.365 | 0.351 | 0.254 | 0.409 | /ow/ | c̲ode | 0.112 | 0.061 | 0.086 | 0.196 |
| /b/ | b̲e | 0.141 | 0.265 | 0.110 | 0.216 | /r/ | r̲an | 0.168 | 0.115 | 0.145 | 0.210 |
| /pcl/ | mo p̲ | 2.295 | 1.712 | 1.333 | 1.409 | overall - /h#/ | | 0.772 | 0.546 | 0.418 | 0.471 |

Table 3: A sample set of IS measures for original degraded (5dB HWY), and three enhancement methods across phonemes.

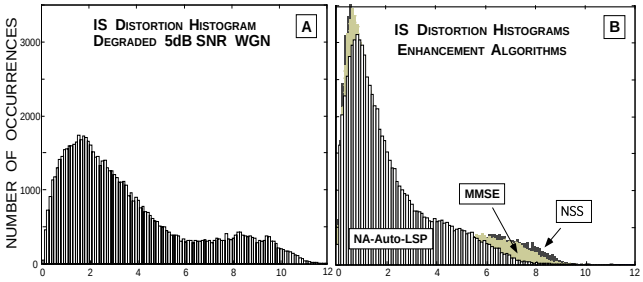


Figure 2: Histograms of frame–based Itakura–Saito (IS) distortion measures over 192 sentence TIMIT core test set: (A) speech degraded with additive white Gaussian noise, and (B) enhanced using NA-Auto-LSP, MMSE, NSS algorithms.

**Quality Measure Histograms:** Another way to compare performance is using quality measure histograms as shown in Fig. 2. For the IS measure distribution here, we see that after processing, all three enhancement algorithms move degraded frames closer to noisefree '0' distortion. The important aspect here is to compare the number of frames in the tails of the distribution, thus reflecting the consistency of the enhancement algorithm.

**Phone-level Quality Performance:** The fourth approach for illustrating quality performance is at the individual phone level. Since phone-level transcriptions exist for TIMIT, it is easy to group frames for a particular phoneme. Table 3 lists 10 of the 61 phonemes used in TIMIT, with IS measures for degraded, and the three enhancement methods. This allows the user to compare performance within individual phone classes (note similar performance for the enhancement algorithms for voiced /z/ and unvoiced /s/ fricatives). Using the overall score for comparison, it is easy to rank individual phones above and below the mean.

## 7. Conclusions

In this study, we have considered factors important for effective evaluation of speech enhancement algorithms. We proposed that researchers use the evaluation core test set of TIMIT (192 sentences), with a set of noise files, and a combination of objective measures and subjective testing for broad and fine phone-level quality assessment. We recommend that subjective MOS testing be done to identify broad quality, and Pairwise Preference Testing to determine listener preference to algorithm processing artifacts. Five objective speech quality measures were considered, since it is known that SNR is a poor predictor of speech quality. We illustrated four methods of demonstrating objective speech quality performance based on (i) overall quality measures, (ii) phone-class level scores, (iii) quality measure histograms, and (iv) phone-level evaluation. The reported results are meant to illustrate specific ways of detailing quality assessment for an enhancement algorithm. We emphasize that good objective quality measures exist which clearly outperform SNR, and that a combination of subjective and objective testing allows researchers to carefully identify algorithm performance. Interested readers are encouraged to check our web page for available enhancement evaluation tools.

## 8. REFERENCES


[1] J.H.L. Hansen, B.L. Pellom, "Speech Enhancement and Quality Assessment: A Survey," submitted to *IEEE Sig. Proc. Mag.* Nov. 1998.

[2] J. Deller, J. Proakis, J.H.L. Hansen, *Discrete-Time Processing of Speech Signals,* MacMillan Series for Prentice-Hall, New York, NY, 1993.

[3] J.S. Lim, A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech," *Proc. of IEEE,* **67**:1586-1604, 1979.

[4] *Speech Enhancement,* Editor: J. Lim, Prentice-Hall, Englewood Cliffs, N.J., 1983.

[5] D. O'Shaughnessy, "Enhancing speech degraded by additive noise or interfering speakers," *IEEE Comm. Mag.,* **27**(2):46-52, Feb. 1989.

[6] Y. Ephraim,"Statistical-Model-Based Speech Enhancement Systems," *Proc. of IEEE,* **80**(10):1526-1555, 1992.

[7] Y. Ephraim, D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. ASSP,* **32**(6):1109-21, 1984.

[8] J.H.L. Hansen, M. Clements, "Constrained Iterative Speech Enhancement with Application to Speech Recognition," *IEEE Trans. ASSP,* **39**(4):795-805, 1991.

[9] J.H.L. Hansen, S. Nandkumar, "Objective Speech Quality Assessment and the RPE-LTP coding algorithm in different noise and language conditions," *J. Acoust. Soc. Am.,* **97**(1):609-627, Jan. 1995.

[10] J.H.L. Hansen, L.M. Arslan, "Robust Feature Estimation and Objective Quality Assessment for Noisy Speech Recognition using the Credit Card Corpus," *IEEE Trans. SAP,* **3**(3):169-184, 1995.

[11] P. Lockwood, J. Boudy, "Experiments with NSS, HMMs and the projection, for Robust Speech Recognition in Cars," *Speech Comm.,* **11**:215-28, 1992.

[12] S. Nandkumar, J.H.L. Hansen, "Dual-Channel Iterative Speech Enhancement with Constraints Based on an Auditory Spectrum," *IEEE Trans. SAP,* **3**(1):22-34, Jan. 1995.

[13] P. Papamichalis, *Practical Approaches to Speech Coding,* Prentice-Hall, NJ, 1987.

[14] S.R. Quackenbush, T.P. Barnwell, M.A. Clements, *Objective Measures of Speech Quality,* Prentice-Hall, NJ, 1988.