Evaluation of Objective Quality Measures for Speech Enhancement

Yi Hu and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—In this paper, we evaluate the performance of several objective measures in terms of predicting the quality of noisy speech enhanced by noise suppression algorithms. The objective measures considered a wide range of distortions introduced by four types of real-world noise at two signal-to-noise ratio levels by four classes of speech enhancement algorithms: spectral subtractive, subspace, statistical-model based, and Wiener algorithms. The subjective quality ratings were obtained using the ITU-T P.835 methodology designed to evaluate the quality of enhanced speech along three dimensions: signal distortion, noise distortion, and overall quality. This paper reports on the evaluation of correlations of several objective measures with these three subjective rating scales. Several new composite objective measures are also proposed by combining the individual objective measures using nonparametric and parametric regression analysis techniques.**

***Index Terms*—Objective measures, speech enhancement, speech quality assessment, subjective listening tests.**

I. Introduction

CURRENTLY, the most accurate method for evaluating speech quality is through subjective listening tests. Although subjective evaluation of speech enhancement algorithms is often accurate and reliable (i.e., repeatable) provided it is performed under stringiest conditions (e.g., sizeable listener panel, inclusion of anchor conditions, etc. [1]–[3]), it is costly and time consuming. For that reason, much effort has been placed on developing objective measures that would predict speech quality with high correlation. Many objective speech quality measures have been proposed in the past to predict the subjective quality of speech [1]. Most of these measures, however, were developed for the purpose of evaluating the distortions introduced by speech codecs and/or communication channels [4]–[9]. The quantization and other types of distortions introduced by waveform and linear predictive coding (LPC)-based speech coders [e.g., code excited linear prediction (CELP)], however, are different from those introduced by speech enhancement algorithms. As a result, it is not clear whether the objective measures originally developed for predicting speech coding distortions [1] are suitable for evaluating the quality of speech enhanced by noise suppression algorithms.

Manuscript received January 10, 2007; revised September 26, 2007. This work was supported in part by the National Institute on Deafness and other Communication Disorders/National Institute of Health (NIDCD/NIH) under Grant R01 DC07527. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Abeer Alwan.

The authors are with the Department of Electrical Engineering, University of Texas at Dallas Richardson, TX 75083-0688 USA (e-mail: loizou@utdallas.edu).

Digital Object Identifier 10.1109/TASL.2007.911054

The types of distortion introduced by speech enhancement algorithms can be broadly divided into two categories: the distortions that affect the speech signal itself (called speech distortion) and the distortions that affect the background noise (called noise distortion). Of these two types of distortion, listeners seem to be influenced the most by the speech distortion when making judgments of overall quality [10], [11]. Unfortunately no objective measure currently exists that correlates high with either type of distortion or with the overall quality of speech enhanced by noise suppression algorithms.

Compared to the speech coding literature [1], only a small number of studies examined the correlation between objective measures and the subjective quality of noise-suppressed speech [12]–[17]. Salmela and Mattila [13] evaluated the correlation of a composite measure with the subjective (overall) quality of noise-suppressed speech. The composite measure consisted of 16 different objective measures which included, among others, spectral distance measures, LPC measures (e.g., Itakura–Saito) and time-domain measures [e.g., segmental signal-to-noise ratio (SNR)]. The noisy speech samples were not processed by real enhancement algorithms, but rather by ideal noise-suppression algorithms designed to provide controlled attenuation to the background alone or to both background and speech signals. The resulting composite measure produced a high correlation of 0.95 with overall quality. Rohdenburg *et al.* [12] evaluated the correlation of several objective measures including LPC-based measures [e.g., log-area ratio (LAR)] and the perceptual evaluation of speech quality (PESQ) measure with speech enhanced by a single algorithm. The subjective listening tests were done according to the ITU-T P.835 methodology specifically designed to evaluate the distortions and overall quality of noise suppression algorithms. Correlations ranging from 0.7 to 0.81 were obtained with ratings of background distortion, signal distortion, and overall quality. Turbin and Fluchier [14] proposed a new objective measure for predicting the background intrusiveness rating scores obtained from ITU-T P.835-based listening tests. High correlation was found with the background noise ratings using a measure that was based on loudness density comparisons and coefficient of tonality. Turbin and Fluchier later extended their work in [18] and proposed an objective measure to estimate signal distortion (but not overall quality).

With the exception of [12], [14], and [18], most studies reported correlation of objective measures with only the overall quality of noise-suppressed speech. In those studies, only a small number (1–6) of noise suppression algorithms were involved in the evaluations. The study by Rhodenburg

et al. [12] evaluated the correlation of objective measures with speech/noise distortions and overall quality, but only for speech enhanced by a single statistical-model based enhancement algorithm, the minimum mean square error (mmse) algorithm. Other classes of algorithms (e.g., subspace and spectral subtractive), however, will likely introduce different types of signal/background distortion. Hence, the correlations reported in [12] are only applicable for distortions introduced by mmse-type of algorithms and not by other algorithms.

To our knowledge, no comprehensive study was done to assess the correlation of existing objective measures with the distortions (background and speech) present in enhanced speech and with the overall quality of noise-suppressed speech. Since different classes of algorithms introduce different types of signal/background distortion, it is necessary to include various classes of algorithms in such an evaluation. The main objective of the present study is to report on the evaluation of conventional as well as new objective measures that could be used to predict overall speech quality and speech/noise distortions introduced by representative speech enhancement algorithms from various classes (e.g., spectral-subtractive, subspace, etc) of algorithms. To that end, we make use of an existing subjective database that we collected for the evaluation of speech enhancement algorithms [10], [11]. The subjective quality ratings were obtained by Dynastat, Inc. using the ITU-T P.835 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion, and overall quality.

Preliminary evaluation of several objective measures with speech processed by enhancement algorithms was reported in [19]. In that study, we showed that the majority of the commonly used objective speech quality measures perform modestly well (but not exceeding 0.75) in terms of predicting subjective quality of noisy speech processed by enhancement algorithms. The correlations were performed using all speech samples (files) available without averaging the objective scores across conditions. The test chosen was undoubtedly stringent, resulting in only a few of the objective measures correlating high with speech and noise distortions introduced by speech enhancement algorithms. In this paper, we further extend the results reported in [19] and evaluate a larger set of objective speech quality measures after averaging the objective scores across conditions (SNR level, noise type, and algorithm). In addition, we propose several new composite objective measures derived using nonlinear and nonparametric regression models which are shown to provide higher correlations with subjective speech quality and speech/noise distortions than the conventional objective measures. The use of composite measures is necessary as we cannot expect the simple objective measures (e.g., LPC-based) to correlate highly with signal/noise distortions *and* with overall quality.

This paper is organized as follows. In Section II, we describe the NOIZEUS noisy speech corpus and the subjective quality evaluation protocols. In Section III, we present the objective measures evaluated and in Section IV we present the resulting correlation coefficients. The conclusions are given in Section V.

II. Speech Corpus and Subjective Quality Evaluations

In [19], we reported on the evaluation of several common objective measures using a noisy speech corpus (NOIZEUS[1]) developed in our lab that is suitable for evaluation of speech enhancement algorithms. This corpus was used in a comprehensive subjective evaluation of 13 speech enhancement algorithms encompassing four different classes of algorithms: spectral subtractive (multiband spectral subtraction, and spectral subtraction using reduced delay convolution and adaptive averaging), subspace (generalized subspace approach, and perceptually based subspace approach), statistical-model-based (mmse, log-mmse, and log-mmse under signal presence uncertainty) and Wiener-filtering type algorithms (the *a priori* SNR estimation based method, the audible-noise suppression method, and the method based on wavelet thresholding the multitaper spectrum). The enhanced speech files were sent to Dynastat, Inc. (Austin, TX) for subjective evaluation using the recently standardized methodology for evaluating noise suppression algorithms based on ITU-T P.835 [2].

The subjective listening tests were designed according to ITU-T recommendation P.835 and were conducted by Dynastat, Inc. (Austin, TX). The P.835 methodology was designed to reduce the listener's uncertainty in a subjective listening test as to which component(s) of a noisy speech signal, i.e., the speech signal, the background noise, or both, should form the basis of their ratings of overall quality. This method instructs the listener to successively attend to and rate the enhanced speech signal on:

- the speech signal alone using a five-point scale of signal distortion (SIG);
- the background noise alone using a five-point scale of background intrusiveness (BAK);
- the overall quality using the scale of the mean opinion score (OVRL)-[1 = bad, 2 = poor, 3=fair, 4=good, 5 = excellent].

The SIG and BAK scales are described in Table I. A total of 32 listeners were recruited for the listening tests. The results of the subjective listening tests were reported in [10] and [11]. In this paper, we make use of the subjective ratings along the three quality scales (SIG, BAK, OVRL) to evaluate conventional and new objective measures.

III. Objective Measures

Several objective speech quality measures were evaluated: segmental SNR (segSNR) [20], weighted-slope spectral distance (WSS) [21], PESQ [8], [22], LPC-based objective measures including the log-likelihood ratio (LLR), Itakura-Saito distance measure (IS), and cepstrum distance measures (CEP) [1], and frequency-weighted segmental SNR (fwsegSNR) [23]. Composite measures obtained by combining a subset of the above measures were also evaluated.

Two figures of merit are computed for each objective measure. The first one is the correlation coefficient (Pearson's

[1][Online]. Available: http://www.utdallas.edu/∼loizou/speech/noizeus/

TABLE I
DESCRIPTION OF THE SIG AND BAK SCALES USED IN THE SUBJECTIVE LISTENING TESTS

SIG scale	
Rating	Description
5	Very natural, no degradation
4	Fairly natural, little degradation
3	Somewhat natural, somewhat degraded
2	Fairly unnatural, fairly degraded
1	Very unnatural, very degraded

BAK scale	
Rating	Description
5	Not noticeable
4	Somewhat noticeable
3	Noticeable but not intrusive
2	Fairly conspicuous, somewhat intrusive
1	Very conspicuous, very intrusive

correlation) between the subjective quality ratings S_d and the objective measure O_d, and is given by

$$\rho = \frac{\sum_d (S_d - \bar{S}_d)(O_d - \bar{O}_d)}{\left[\sum_d (S_d - \bar{S}_d)^2\right]^{1/2} \left[\sum_d (O_d - \bar{O}_d)^2\right]^{1/2}} \quad (1)$$

where $\bar{S}_d$ and $\bar{O}_d$ are the mean values of S_d and O_d, respectively. The second figure of merit is an estimate of the standard deviation of the error when the objective measure is used in place of the subjective measure, and is given by

$$\hat{\sigma}_e = \hat{\sigma}_d \sqrt{1 - \rho^2} \quad (2)$$

where $\hat{\sigma}_d$ is the standard deviation of S_d, and $\hat{\sigma}_e$ is the computed standard deviation of the error. A smaller value of $\hat{\sigma}_e$ indicates that the objective measure is better at predicting subjective quality.

Two types of regression analysis techniques were used in this paper, namely parametric (linear regression) and nonparametric techniques. The nonparametric regression technique used was based on multivariate adaptive regression splines (MARS) [24] analysis. Unlike the linear and polynomial regression techniques, the MARS modeling technique is data driven and derives the best fitting function from the data. The basic idea of the MARS modeling is to use spline functions to locally fit the data in a region, and then generate a global model by combining the data regions using basis functions. One of the most powerful features of the MARS modeling is that it allows interactions between the predictor (independent) variables so that a better fit can be found for the target (dependent) variable.

A. PESQ

Among all objective measures considered, the PESQ measure is the most complex to compute and is the one recommended by ITU-T for speech quality assessment of 3.2 kHz (narrow-band) handset telephony and narrow-band speech codecs [7], [8]. As described in [8], the PESQ score is computed as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind} as follows:

$$\text{PESQ} = a_0 + a_1 D_{\text{ind}} + a_2 A_{\text{ind}} \quad (3)$$

where $a_0 = 4.5, a_1 = -0.1$, and $a_2 = -0.0309$. The parameters a_0, a_1 and a_2 in the above equation were optimized for speech processed through networks and not for speech enhanced by noise suppression algorithms. As we can not expect the PESQ measure to correlate highly with all three quality measures (speech distortion, noise distortion and overall quality), we considered optimizing the PESQ measure for each of the three rating scales by choosing a different set of parameters (a_0, a_1, a_2) for each rating scale. The modified PESQ measures were obtained by treating a_0, a_1 and a_2 in (3) as the parameters that need to be optimized for each of the three rating scales: speech distortion, noise distortion, and overall quality. Multiple linear regression analysis was used to determine the a_0, a_1 and a_2 parameters. The values of D_{ind} and A_{ind} in (3) were treated as independent variables in the regression analysis. The actual subjective scores for the three scales were used in the regression analysis. This analysis yielded three different modified PESQ measures suitable for predicting signal distortion noise distortion and overall speech quality. These measures will be described later in Section IV.

B. LPC-Based Objective Measures

Three different LPC-based objective measures were considered: the LLR, the IS, and the cepstrum distance measures.

The LLR measure is defined as [1]

$$d_{\text{LLR}}(\vec{a}_p, \vec{a}_c) = \log\left(\frac{\vec{a}_p \mathbf{R}_c \vec{a}_p^T}{\vec{a}_c \mathbf{R}_c \vec{a}_c^T}\right) \quad (4)$$

where $\vec{a}_c$ is the LPC vector of the original speech signal frame, $\vec{a}_p$ is the LPC vector of the enhanced speech frame, and $\mathbf{R}_c$ is the autocorrelation matrix of the original speech signal. Only the smallest 95% of the frame LLR values were used to compute the average LLR value [20]. The segmental LLR values were limited in the range of [0, 2] to further reduce the number of outliers.

The IS measure is defined as [1]

$$d_{\text{IS}}(\vec{a}_p, \vec{a}_c) = \frac{\sigma_c^2}{\sigma_p^2}\left(\frac{\vec{a}_p \mathbf{R}_c \vec{a}_p^T}{\vec{a}_c \mathbf{R}_c \vec{a}_c^T}\right) + \log\left(\frac{\sigma_c^2}{\sigma_p^2}\right) - 1 \quad (5)$$

where σ_c^2 and σ_p^2 are the LPC gains of the clean and enhanced signals, respectively. The IS values were limited in the range of [0, 100]. This was necessary in order to minimize the number of outliers.

The cepstrum distance provides an estimate of the log spectral distance between two spectra. The cepstrum coefficients can be obtained recursively from the LPC coefficients $\{a_m\}$ using the following expression:

$$c(m) = a_m + \sum_{k=1}^{m-1} \frac{k}{m} c(k) a_{m-k} \quad 1 \le m \le p \quad (6)$$

where p is the order of the LPC analysis. An objective measure based on cepstrum coefficients can be computed as follows [25]:

$$d_{\mathrm{CEP}}(\vec{c}_c, \vec{c}_p) = \frac{10}{\log 10}\sqrt{2\sum_{k=1}^{p}[c_c(k) - c_p(k)]^2} \quad (7)$$

where $\vec{c}_c$ and $\vec{c}_p$ are the cepstrum coefficient vector of the clean and enhanced signals, respectively. The cepstrum distance was limited in the range of [0, 10] to minimize the number of outliers.

C. Time-Domain and Frequency-Weighted SNR Measures

The time-domain segmental SNR (segSNR) measure was computed as per [20]. Only frames with segmental SNR in the range of -10 to 35 dB were considered in the average.

The frequency-weighted segmental SNR (fwSNRseg) was computed using the following equation:

$$\mathrm{fwSNRseg} = \frac{10}{M} \times \sum_{m=0}^{M-1} \frac{\sum_{j=1}^{K} W(j,m)\log_{10}\frac{|X(j,m)|^2}{(|X(j,m)|-|\hat{X}(j,m)|)^2}}{\sum_{j=1}^{K} W(j,m)} \quad (8)$$

where $W(j,m)$ is the weight placed on the jth frequency band, K is the number of bands, M is the total number of frames in the signal, $|X(j,m)|$ is the weighted (by a Gaussian-shaped window) clean signal spectrum in the jth frequency band at the mth frame, and $|\hat{X}(j,m)|$ in the weighted enhanced signal spectrum in the same band. For the weighting function, we considered the magnitude spectrum of the clean signal raised to a power, i.e.,

$$W(j,m) = |X(j,m)|^{\gamma} \quad (9)$$

where $|X(j,m)|$ is the weighted magnitude spectrum of the clean signal obtained in the jth band at frame m and γ is the power exponent, which can be varied for maximum correlation. In our experiments, we varied γ from 0.1 to 2 and obtained maximum correlation with $\gamma = 0.2$.

The spectra $|X(j,m)|$ in (8) were obtained by dividing the signal bandwidth into either 25 bands or 13 bands spaced in proportion to the ear's critical bands. The 13 bands were formed by merging adjacent critical bands. The weighted spectra used in (8) were obtained by multiplying the fast spectra with overlapping Gaussian-shaped windows [26, Ch. 11] and summing up the weighted spectra within each band. Prior to the distance computation in (8), the clean and processed FFT magnitude spectra were normalized to have an area equal to one. This normalization was found to be critically important.

The last conventional measure tested was the WSS measure [21]. The WSS distance measure [21] computes the weighted difference between the spectral slopes in each frequency band. The spectral slope is obtained as the difference between adjacent spectral magnitudes in decibels. The WSS measure evaluated in this paper is defined as

$$d_{\mathrm{WSS}} = \frac{1}{M}\sum_{m=0}^{M-1}\frac{\sum_{j=1}^{K} W(j,m)(S_c(j,m) - S_p(j,m))^2}{\sum_{j=1}^{K} W(j,m)} \quad (10)$$

where $W(j,m)$ are the weights computed as per [21], K and M are defined as in (8), and $S_c(j,m)$, $S_p(j,m)$ are the spectral slopes for jth frequency band at frame m of the clean and processed speech signals, respectively. In our implementation, the number of bands was set to $K = 25$.

Aside from the PESQ measure, all other measures were computed by segmenting the sentences using 30-ms duration Hamming windows with 75% overlap between adjacent frames. A tenth order LPC analysis was used in the computation of the LPC-based objective measures (CEP, IS, and LLR).

D. Composite Measures

Composite objective measures were obtained by combining basic objective measures to form a new measure [1]. As mentioned earlier, composite measures are necessary as we cannot expect the conventional objective measures (e.g., LLR) to correlate highly with speech/noise distortions *and* overall quality. The composite measures can be derived by utilizing multiple linear regression analysis or by applying nonlinear techniques (e.g., [5] and [27]). In this paper, we used both multiple linear regression analysis and MARS analysis to estimate three different composite measures: a composite measure for signal distortion (SIG), a composite measure for noise distortion (BAK), and a composite measure for overall speech quality (OVRL).

The task of forming a good composite measure by linearly combining basic objective measures is not an easy one. Ideally, we would like to combine objective measures that correlate highly with subjective ratings, and at the same time, capture different characteristics of the distortions present in the enhanced signals. There is no straightforward method of selecting the best subset of objective measures to use in the composite measure, other than by trying out different combinations and assessing the resulting correlation. Multidimensional scaling techniques [1 , Ch. 4], [13] may be used in some cases as a guide for the selection. The methodology used in [1, Ch. 9] was adopted in this study for selecting the individual objective measures. More specifically, we tested various combinations of basic objective measures to determine to what extent the correlation coefficient could be improved by combining them. Seven basic object measures were used in the analysis. We kept only the subset of measures for which the correlation of the composite measure improved significantly from the correlation coefficient of the individual measures.

IV. Results: Correlations of Objective Measures

Correlation coefficients ρ and estimates of the standard deviation of the error $\hat{\sigma}_e$ were computed for each objective measure and each of the three subjective rating scales (SIG, BAK, OVRL). Two types of correlation analysis were performed. The first analysis was done as in our previous study [19] and included all objective scores obtained for each speech sample (file). A

TABLE II
ESTIMATED CORRELATION COEFFICIENTS $|\rho|$ OF OBJECTIVE MEASURES WITH OVERALL QUALITY, SIGNAL DISTORTION, AND BACKGROUND NOISE DISTORTION. CORRELATIONS WERE OBTAINED USING THE OBJECTIVE SCORES OF ALL SPEECH SAMPLES

Objective measure	Overall quality	Signal distortion	Background distortion
SegSNR	0.31	0.19	0.42
Weighted spectral slope (WSS)	0.53	0.50	0.37
PESQ	0.65	0.57	0.48
Log-likelihood ratio (LLR)	0.61	0.64	0.24
Itakura-Saito distance (IS)	0.45	0.58	0.06
Cepstrum distance (CEP)	0.56	0.61	0.14
fwSNRseg (K=13 bands)	0.63	0.67	0.25
fwSNRseg (K=25 bands)	0.64	0.67	0.27
Modified PESQ (training)	0.66	0.62	0.55
Modified PESQ (testing)	0.67	0.66	0.56
Composite measures C (training)	0.71	0.71	0.59
Composite measures C (testing)	0.73	0.73	0.64
Composite measures C_MARS (training)	0.73	0.74	0.62
Composite measures C_MARS (testing)	0.73	0.75	0.64

total of 1792 processed speech samples were included in the correlations encompassing two SNR levels (5 and 10 dB), four different types of background noise, and speech/noise distortions introduced by 13 different speech enhancement algorithms. The ratings for each speech sample were averaged across all listeners involved in that test. A total of 43 008 ($= 1792$ files $\times 8$ listeners $\times 3$ rating scales) subjective listening scores were used in the computation of the correlation coefficients for the three rating scales. Acknowledging that the above correlation analysis is rather stringent (but perhaps more desirable in some applications), we considered performing correlation analysis using objective scores which were averaged across each condition. This analysis involved the use of mean objective scores and ratings computed across a total of 112 conditions ($= 14$ algorithms[2] $\times 2$ SNR levels $\times 4$ noise types). In order to cross-validate the composite measures (and any other measures requiring training), we divided our data set in half, with 50% of the data being used for training and the remaining 50% being used for testing. Of the 16 speech samples used for each condition, we used eight speech samples for training and eight for testing. So, in the first correlation analysis, we used the ratings and objective scores of $896 (= 1792/2)$ speech samples for training and the rest for testing. In the second correlation analysis, we used the ratings and objective scores averaged across eight (of 16) speech files for training. This yielded a total of 112 pairs of ratings and objective scores for training. For testing, we used the ratings and objective scores averaged across the remaining eight files in each condition. This yielded a total of 112 pairs of ratings and objective scores for testing. We also considered partitioning the data into training and testing sets according to the various classes of speech enhancement algorithms. In this setup, the composite measures were trained on data taken from a given set of algorithms and tested on data taken from the remaining algorithms. Resulting correlation coefficients of the composite measures were comparable (and remained robust) to those obtained with the aforementioned data partitioning. For that reason, we only report correlations with the former data partitioning.

We report separately the correlations (and errors $\hat{\sigma}_e$) obtained using the per speech sample analysis (Tables II and III) and the per condition analysis (Tables IV and V). Tables VI and VII provide the regression coefficients of the modified PESQ and the composite measures respectively obtained both using multiple linear regression analysis. Table VI tabulates the coefficients $\{a_k\}$ used in (3) for constructing the modified PESQ measures for the three rating scales. The assumed form of composite measures listed in Table VII is shown as follows:

$$C_X = \beta_0 + \sum_{k=1}^{5} \beta_k O_k \tag{11}$$

where C_X is the composite measure for rating scale x (signal distortion, background distortion, overall quality), $\{\beta_k\}$ are the regression coefficients given in Table VII, and $\{O_k\}$ are the corresponding objective measures. Empty entries in Table VII indicate that the corresponding objective measure was not included in the composite measure.

Comparing Tables II and IV, we see a large difference (of about 0.2) between the correlations obtained on a per sample basis and those obtained on a per condition basis. From Table II, we see that of the seven basic objective measures tested, the

[2]The noisy sentences (unprocessed) were also included.

TABLE III
STANDARD DEVIATION OF THE ERROR $\hat{\sigma}_e$ OF OBJECTIVE MEASURES WITH OVERALL QUALITY, SIGNAL DISTORTION, AND BACKGROUND NOISE DISTORTION. STANDARD DEVIATIONS OF ERROR WERE OBTAINED USING THE OBJECTIVE SCORES OF ALL SPEECH SAMPLES

Objective measure	Overall quality	Signal distortion	Background distortion
SegSNR	0.58	0.78	0.53
Weighted spectral slope (WSS)	0.52	0.68	0.54
PESQ	0.46	0.65	0.51
Log-likelihood ratio (LLR)	0.47	0.59	0.56
Itakura-Saito distance (IS)	0.54	0.64	0.58
Cepstrum distance (CEP)	0.49	0.60	0.57
fwSNRseg (K=13 bands)	0.47	0.58	0.56
fwSNRseg (K=25 bands)	0.47	0.56	0.59
Modified PESQ (training)	0.43	0.59	0.47
Modified PESQ (testing)	0.48	0.61	0.50
Composite measures C (training)	0.40	0.53	0.45
Composite measures C (testing)	0.44	0.56	0.46
Composite measures C_MARS (training)	0.39	0.51	0.44
Composite measures C_MARS (testing)	0.44	0.55	0.46

TABLE IV
ESTIMATED CORRELATION COEFFICIENTS $|\rho|$ OF OBJECTIVE MEASURES WITH OVERALL QUALITY, SIGNAL DISTORTION, AND BACKGROUND NOISE DISTORTION. CORRELATIONS WERE OBTAINED AFTER AVERAGING OBJECTIVE SCORES AND RATINGS ACROSS CONDITIONS

Objective measure	Overall quality	Signal distortion	Background distortion
SegSNR	0.36	0.22	0.56
Weighted spectral slope (WSS)	0.64	0.59	0.62
PESQ	0.89	0.81	0.76
Log-likelihood ratio (LLR)	0.85	0.88	0.51
Itakura-Saito distance (IS)	0.60	0.73	0.09
Cepstrum distance (CEP)	0.79	0.84	0.41
fwSNRseg (K=13 bands)	0.85	0.87	0.59
fwSNRseg (K=25 bands)	0.84	0.84	0.62
Modified PESQ (training)	0.89	0.89	0.75
Modified PESQ (testing)	0.92	0.89	0.76
Composite measures C (training)	0.90	0.91	0.81
Composite measures C (testing)	0.91	0.89	0.82
Composite measures C_MARS (training)	0.94	0.94	0.87
Composite measures C_MARS (testing)	0.91	0.90	0.80

PESQ measure yielded the highest correlation ($\rho = 0.65$) with overall quality, followed by the fwSNRseg measure ($\rho = 0.64$) and the LLR measure ($\rho = 0.61$). Compared to the PESQ measure, the LLR and fwSNRseg measures are computationally simpler to implement and yield roughly the same correlation coefficient. The lowest correlation ($\rho = 0.31$) was obtained with the SNRseg measure. The correlations with signal distortion were of the same magnitude as those of overall quality.

TABLE V
STANDARD DEVIATION OF THE ERROR $\hat{\sigma}_e$ OF OBJECTIVE MEASURES WITH OVERALL QUALITY, SIGNAL DISTORTION, AND BACKGROUND NOISE DISTORTION. STANDARD DEVIATIONS OF ERROR WERE OBTAINED AFTER AVERAGING OBJECTIVE SCORES AND RATINGS ACROSS CONDITIONS

Objective measure	**Overall quality**	**Signal distortion**	**Background distortion**
SegSNR	0.43	0.55	0.34
Weighted spectral slope (WSS)	0.36	0.46	0.33
PESQ	0.21	0.33	0.26
Log-likelihood ratio (LLR)	0.25	0.27	0.35
Itakura-Saito distance (IS)	0.37	0.38	0.41
Cepstrum distance (CEP)	0.29	0.31	0.38
fwSNRseg (K=13 bands)	0.24	0.28	0.33
fwSNRseg (K=25 bands)	0.25	0.31	0.32
Modified PESQ (training)	0.21	0.25	0.27
Modified PESQ (testing)	0.18	0.25	0.27
Composite measures C (training)	0.20	0.23	0.24
Composite measures C (testing)	0.19	0.25	0.24
Composite measures C_MARS (training)	0.16	0.20	0.20
Composite measures C_MARS (testing)	0.19	0.24	0.25

TABLE VI
REGRESSION COEFFICIENTS [SEE (3)] FOR THE MODIFIED PESQ MEASURES

Correlation analysis	**Rating**	a_0	a_1	a_2
Per speech sample	Overall quality	4.788	-0.152	-0.016
	Signal	4.959	-0.191	0.006
	Background	5.336	-0.082	-0.058
Per condition	Overall quality	5.413	-0.205	-0.016
	Signal	5.736	-0.25	0.003
	Background	5.758	-0.121	-0.057

This suggests that the same basic objective measure predicts equally well signal distortion and overall quality. This finding is consistent with our previous data [11] suggesting that listeners are more sensitive to signal distortion than background distortion when making judgments on overall quality. The correlations, however, with noise distortion were generally poorer suggesting that the basic objective measures are inadequate in predicting background distortion. A significant improvement in correlation with background distortion was obtained with the use of composite measures. Significant improvements were obtained in correlations with overall quality and signal distortion. Tables II and IV list separately the correlations obtained by the composite measures with training and testing data. The highest correlation coefficients with overall quality ($\rho = 0.73$), signal distortion ($\rho = 0.75$) and noise distortion ($\rho = 0.64$) were obtained with the MARS-based composite measure. The MARS composite measure improved particularly the background distortion correlation from 0.48 (obtained with PESQ) to 0.64.

Overall, the correlations obtained on a per condition basis (Table IV) were higher (by about 0.2) than the correlations obtained on a per speech sample basis and the standard deviations of the error were smaller (Table V). This is to be expected, given the smaller variance in objective scores following the averaging across files. The pattern of results, however, in terms of

TABLE VII
REGRESSION COEFFICIENTS [SEE (11)] AND OBJECTIVE MEASURES USED IN THE CONSTRUCTION OF THE COMPOSITE MEASURES

Correlation analysis	Rating	β_0	IS (β_1)	PESQ (β_2)	CEP (β_3)	LLR (β_4)	WSS (β_5)
Per speech sample	Overall quality	0.279	-0.011	1.137		0.041	-0.008
	Signal	2.164	-0.02	0.832	-0.494	0.352	
	Background	0.985		0.848	-0.319	0.295	-0.008
Per condition	Overall quality	-0.736	-0.012	1.5			
	Signal	-0.261	1.562	-0.02			
	Background	1.893	0.007	0.8	-0.468	0.291	-0.008

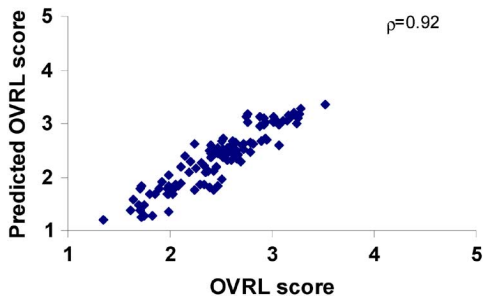


Fig. 1. Scatter plot of the modified PESQ measure against the true subjective ratings of overall speech quality (OVRL). The estimated correlation coefficient was 0.92.

which objective measures yielded the highest correlation was similar to that shown in Table II. Of the seven basic objective measures tested, the PESQ measure yielded the highest correlation ($\rho = 0.89$) on overall quality, followed by the fwSNRseg ($K = 13$) and LLR measures ($\rho = 0.85$). The composite measures further improved the correlation to overall quality to higher than 0.9. Highest correlation with overall quality was obtained with the modified PESQ measure ($\rho = 0.92$), and the highest correlation with background distortion was obtained with the composite MARS measure ($\rho = 0.8$).

Fig. 1 shows the scatter plot of the OVRL scores and predicted scores obtained by the modified PESQ measure, which yielded a correlation of $\rho = 0.92$ with overall quality. Table VIII shows the number of basis functions and objective measures involved in the composite MARS measures. Sample MATLAB code for the MARS composite measures used for predicting signal and overall quality is given in Appendix A. MATLAB code for the implementation of all objective measures tested are available from [26].

The cross-validation of the composite measures indicated that they are robust to new distortions. This was found to be true even when the data were partitioned into training and testing sets according to the various classes of algorithms.[3] The correlations obtained with the training data were comparable to those obtained with new and unseen data (Tables II and IV). Furthermore, the fact that these composite measures were tested on a publicly available speech corpus (NOIZEUS) makes these measures ideal for testing new enhancement algorithms.

TABLE VIII
NUMBER OF BASIS FUNCTIONS AND OBJECTIVE MEASURES USED IN THE CONSTRUCTION OF THE COMPOSITE MARS-BASED MEASURES

Correlation analysis	Rating	Number of basis functions	Objective measures
Per speech sample	Overall quality	8	PESQ, WSS, LLR, IS
	Signal	6	PESQ, CEP, IS, LLR
	Background	7	PESQ, CEP, IS, LLR, WSS
Per condition	Overall quality	5	PESQ, IS
	Signal	3	PESQ, IS
	Background	6	PESQ, WSS, CEP, LLR, IS

[3]After using the data from the first ten algorithms for training, and the data for the remaining four algorithms for testing, we obtained a correlation coefficient of 0.94 with the training data and a correlation coefficient of 0.96 with the test data for the proposed composite measures designed to predict overall quality. The corresponding correlation coefficients for training and testing data were (0.92, 0.96), respectively, for signal distortion, and (0.86, 0.82) for noise distortion. These data clearly demonstrate that the proposed composite measures are robust to alternate partitioning of the data.

V. Summary and Conclusion

The present study extended our previous evaluation of objective measures [19] and included a per condition correlation analysis. With this new type of analysis, the majority of the correlation coefficients improved by about 0.2. The correlation coefficient of the PESQ measure improved from 0.65 to 0.89.

Based on the correlation analysis reported above, we can draw the following conclusions: The segSNR measure, which is widely used for evaluating the performance of speech enhancement algorithms, yielded a very poor correlation coefficient ($\rho = 0.31 - 0.36$) with overall quality. This finding was consistent with both types of correlation analysis conducted, and thus makes this measure unsuitable for evaluating the performance of enhancement algorithms.

Of the seven basic objective measures tested, the PESQ measure yielded the highest correlation ($\rho = 0.89$) with overall quality and signal distortion. The LLR and fwSNRseg measures performed nearly as well ($\rho = 0.85$) at a fraction of the computational cost. Hence, the LLR and fwSNRseg measures are simpler alternatives to the PESQ measure.

The majority of the basic objective measures predict equally well signal distortion and overall quality, but not background distortion. This was not surprising, given that most measures take into account both speech-active and speech-absent segments in their computation. Measures that would place more emphasis on the speech-absent segments would be more appropriate and likely more successful in predicting noise distortion (BAK).

Appendix

This Appendix shows the MATLAB code for the implementation of the MARS-based composite measures for signal distortion and overall quality. These composite measures yielded correlations of 0.9 and 0.91 with signal distortion and overall quality, respectively.

```
function Y_sig = MARS_sig(PESQ, IS)
%
% composite measure for predicting SIG ratings
    BF1 = max(0, PESQ - 1.696);
    BF2 = max(0, IS - 8.745);
    BF3 = max(0, IS - 2.299);
    Y_sig = 2.463 + 1.557 * BF1 + 0.065 * BF2 -
    0.075 * BF3;
function Y_ovl = MARS_ovl(PESQ, IS)
%
% composite measure for OVRL ratings
    BF1 = max(0, PESQ - 1.696);
    BF2 = max(0, IS - 11.708);
    BF3 = max(0, IS - 3.559);
    BF4 = max(0, PESQ - 2.431);
    BF5 = max(0, PESQ - 2.564);
    Y_ovl = 1.757 + 1.740 * BF1 + 0.047 * BF2 -
    0.049 * BF3-
        2.593 * BF4 + 11.549 * BF5;
```

MATLAB code for the implementation of the PESQ, IS, and other measures tested in this paper, is available in [26].

Acknowledgment

The authors would like to thank Dr. A. Sharpley of Dynastat, Inc. for his help and advice throughout the project.

References

[1] S. Quackenbush, T. Barnwell, and M. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[2] "Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm," ITU-T, ITU-T Rec. P. 835, 2003.

[3] P. Kroon, , W. Kleijn and K. Paliwal, Eds., "Evaluation of speech coders," in *Speech Coding and Synthesis*. New York: Elsevier, 1995, pp. 467–494.

[4] L. Thorpe and W. Yang, "Performance of current perceptual objective speech quality measures," in *Proc. IEEE Speech Coding Workshop*, 1999, pp. 144–146.

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

[6] L. Malfait, J. Berger, and M. Kastner, "P.563-the ITU-T standard for single-ended speech quality assessment," *IEEE Trans. Audio, Speech,Lang. Process.*, vol. 14, no. 6, pp. 1924–1934, Nov. 2006.

[7] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2001, vol. 2, pp. 749–752.

[8] "Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," ITU, ITU-T Rec. P. 862, 2000.

[9] R. Kubichek, D. Atkinson, and A. Webster, "Advances in objective voice quality assessment," in *Proc. Global Telecomm. Conf.*, 1991, vol. 3, pp. 1765–1770.

[10] Y. Hu and P. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2006, vol. 1, pp. 153–156.

[11] Y. Hu and P. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, 2007.

[12] T. Rohdenburg, V. Hohmann, and B. Kollmeir, "Objective perceptual quality measures for the evaluation of noise reduction schemes," in *Proc. 9th Int. Workshop Acoust. Echo Noise Control*, 2005, pp. 169–172.

[13] J. Salmela and V. Mattila, "New intrusive method for the objective quality evaluation of acoustic noise suppression in mobile communications," in *Proc. 116th Audio Eng. Soc. Conv.*, 2004, preprint 6145.

[14] V. Turbin and N. Faucheur, "A perceptual objective measure for noise reduction systems," in *Proc. Online Workshop Meas. Speech Audio Quality Netw.*, 2005, pp. 81–84.

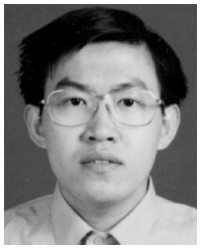
[15] E. Paajanen, B. Ayad, and V. Mattila, "New objective measures for characterization of noise suppression algorithms," in *IEEE Speech Coding Workshop*, 2000, pp. 23–25.

[16] V. Mattila, "Objective measures for the characterization of the basic functioning of noise suppression algorithms," in *Proc. Online Workshop Meas. Speech Audio Quality Netw.*, 2003 [Online]. Available: http://wireless.feld.cvut.cz/mesaqin2003/contributions.html

[17] A. Bayya and M. Vis, "Objective measures for speech quality assessment in wireless communications," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, vol. 1, pp. 495–498.

[18] V. Turbin and N. Faucheur, "Estimation of speech quality of noise reduced signals," in *Proc. Online Workshop Meas. Speech Audio Quality Netw.*, 2007 [Online]. Available: http://wireless.feld.cvut.cz/mesaqin2007/contributions.html

[19] Y. Hu and P. Loizou, "Evaluation of objective measures for speech enhancement," in *Proc. Interspeech*, 2006, pp. 1447–1450.
[20] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *Proc. Int. Conf. Spoken Lang. Process.*, 1998, vol. 7, pp. 2819–2822.
[21] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1982, vol. 7, pp. 1278–1281.
[22] "Application guide for objective quality measurement based on recommendations P.862, P.862.1 and P. 862.2," ITU-T Rec. P. 862. 3, 2005.
[23] J. 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, Signal Process.*, 1978, pp. 586–590.
[24] J. H. Friedman, "Multivariate adaptive regression splines (with discussion)," *Ann. Statist.*, pp. 1–141.
[25] N. Kitawaki, H. Nagabuchi, and K. Itoh, "Objective quality evaluation for low bit-rate speech coding systems," *IEEE J. Sel. Areas Commun.*, vol. 6, no. 2, pp. 262–273, Mar. 1988.
[26] P. Loizou, *Speech Enhancement: Theory and Practice*. Boca Raton, FL: CRC, 2007.
[27] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, "Performance assessment method for speech enhancement systems," in *Proc. 1st Annu. IEEE BENELUX/DSP Valley Signal Process. Symp.*, 2005.



Yi Hu received the B.S. and M.S. degrees in electrical engineering from the University of Science and Technology of China (USTC), Beijing, China, in 1997 and 2000, respectively, and the Ph.D. degree in electrical engineering from the University of Texas at Dallas, Richardson, TX.

He is currently a Research Associate at the University of Texas at Dallas. His research interests are in the general area of speech and audio signal processing and improving auditory prostheses in noisy environments.

Philipos C. Loizou (S'90–M'91–SM'04) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Arizona State University, Tempe, in 1989, 1991, and 1995, respectively.

From 1995 to 1996, he was a Postdoctoral Fellow in the Department of Speech and Hearing Science, Arizona State University, working on research related to cochlear implants. He was an Assistant Professor at the University of Arkansas, Little Rock, from 1996 to 1999. He is now a Professor in the Department of Electrical Engineering, University of Texas at Dallas. His research interests are in the areas of signal processing, speech processing, and cochlear implants. He is the author of the book *Speech Enhancement: Theory and Practice* (CRC, 2007).

Dr. Loizou was an Associate Editor of the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1999-2002) and is currently a member of the Speech Technical Committee of the IEEE Signal Processing Society and serves as Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.