Evaluation of Objective Measures for Speech Enhancement

Yi Hu and Philipos C. Loizou

Department of Electrical Engineering
University of Texas at Dallas
Richardson, TX, USA.
(yihuyxy, loizou)@utdallas.edu

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 SNRs 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 speech quality along three dimensions: signal distortion, noise distortion and overall quality. This paper reports the correlations of five common objective measures with these three subjective measures. Improvements to the PESQ measure are reported along with new composite objective measures.

Index Terms: speech enhancement, noise reduction, ITU-T P.835, objective measures, subjective listening test, correlation analysis.

1. Introduction

Currently the most accurate method for evaluating speech quality is through subjective listening tests. Although subjective evaluation of speech enhancement algorithms is always accurate and preferable, it is time consuming and cost expensive. For that reason, much effort has been placed on developing objective measures that you 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 them, however, were developed for the purpose of evaluating the distortions introduced by speech codecs and/or communication channels [2]. To our knowledge, only a few, if any, of these measures have been formally evaluated with noisy speech enhanced by noise suppression algorithms.

In this paper, we report on the evaluation of common objective measures using a noisy speech corpus (NOIZEUS) developed in our lab that is suitable for evaluation of speech enhancement algorithms [1]. 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 type algorithms (*a priori* SNR estimation based method, audible noise suppression based method, and 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. The results of the subjective listening tests were reported in [3][4]. In this paper, we make use of the subjective test results to evaluate several widely used objective measures.

[1] Available at: http://www.utdallas.edu/~loizou/speech/noizeus/

2. Speech corpus and subjective quality evaluations

In our objective evaluations, we considered distortions introduced by speech enhancement algorithms and background noise. The list of speech enhancement algorithms considered in our study can be found in [4]. Noise was artificially added to the speech signal as follows. The Intermediate Reference System (IRS) filter used in ITU-T P.862 [5] for evaluation of the PESQ measures was independently applied to the clean and noise signals. The active speech level of the filtered clean speech signal was first determined using method B of ITU-T P.56 [6]. A noise segment of the same length as the speech signal was randomly cut out of the noise recordings taken from the AURORA database [7], appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal. A total of 16 sentences corrupted in four background noise environments (car, street, babble and train) at two SNR levels (5dB and 10dB) were processed by the 13 speech enhancement algorithms. These sentences were produced by two male and two female speakers.

2.1. Subjective tests

The subjective listening tests were designed according to ITU-T recommendation P.835 and were conducted by Dynastat, Inc. The P.835 methodology was designed to reduce the listener's uncertainty in a subjective 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:

1. the speech signal alone using a five-point scale of signal distortion (SIG) (Table 1),
2. the background noise alone using a five-point scale of background intrusiveness (BAK) (Table 2),
3. the overall effect using the scale of the Mean Opinion Score (OVRL) - [1=bad, 2=poor, 3=fair, 4=good, 5=excellent].

The process of rating the signal and background of noisy speech was designed to lead the listener to integrate the effects of

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

Table 1: Scale of signal distortion (SIG).

5 - Not noticeable
4 - Somewhat noticeable
3 - Noticeable but not intrusive
2 - Fairly conspicuous, somewhat intrusive
1 - Very conspicuous, very intrusive

Table 2: Scale of background intrusiveness (BAK).

both the signal and the background in making their ratings of overall quality. Each trial in a P.835 test involved a triad of speech samples, where each sample consisted of a single sentence recorded in background noise. For each sample within the triad, listeners successively used one of the three five-point rating scales (SIG, BAK, and OVRL) to register their judgments of the quality of the test condition. In addition to the experimental conditions, each experiment included a number of reference conditions designed to independently vary the listener's SIG, BAK, and OVRL ratings over the entire five-point range of the rating scales.

A total of 32 listeners were recruited for the listening tests. Listeners were recruited from Dynastat's database of native speakers of North American English. Listeners were between the ages of 18 and 50 years of age. No listener had participated in a listening test in the previous three months. The listening panels in the two experiments were independent, i.e., no listener participated in more than one experiment. The tests lasted approximately 1.25 hours. Listeners took short breaks (10 minutes) between sessions. At the beginning of Session 1, the listeners were presented with a practice block of 12 trials to familiarize them with the task and the timing in the trial presentation. The practice blocks were also designed to present the listeners with the range of conditions that would be involved in the tests on both the Signal and the Background scales. For each test, half the panels were presented with trials in which the rating scale order was SIG-BAK-OVRL for the first two sessions and BAK-SIG-OVRL for sessions 3 and 4. To train the listeners for the change in scale order, listeners were presented with the practice block again at the beginning of session 3. For the other half of the panels, the sessions and scale order was counter-balanced.

2.2. Contribution of speech and noise distortion to judgment of overall quality

The P.835 process of rating the signal and background of noisy speech was designed to lead the listener to integrate the effects of both the signal and the background in making their ratings of overall quality. Of great interest is finding out the individual contribution of speech and noise distortion to judgment of overall quality. Our previous subjective data [4] led us to believe that listeners were influenced more by speech distortion when making quality judgments. To further substantiate this, we performed multiple linear regression analysis on the ratings obtained for overall quality, speech and noise distortion. We treated the overall quality score as the dependent variable and the speech and noise distortion scores as the independent variables. Regression analysis revealed the following relationship between the three rating scales:

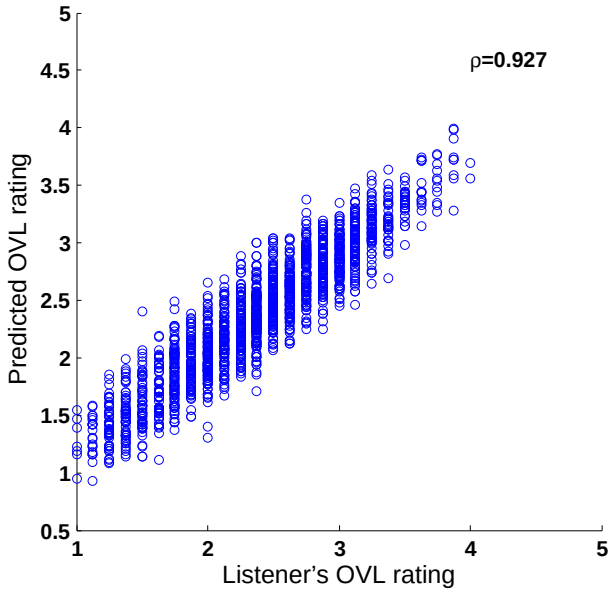

Figure 1: Regression analysis of listener's OVRL ratings, based on SIG and BAK ratings.

$$R_{\text{OVRL}} = -0.0783 + 0.571 \cdot R_{\text{SIG}} + 0.366 \cdot R_{\text{BAK}} \qquad (1)$$

where R_{OVRL} is the predicted overall (OVRL) rating score, R_{SIG} is the SIG rating and R_{BAK} is the BAK rating. The resulting correlation coefficient was $\rho = 0.927$ and the standard deviation of the error was 0.22. Figure 1 shows the scatter plot of the listener's overall quality ratings against the predicted ratings obtained from Eq. (1). The above equation confirms that listeners were indeed integrating the effects of both signal and background distortion when making their ratings. Different emphasis was placed, however, on the two types of distortion. Consistent with our observation, listeners seem to place more emphasis on the distortion imparted on the speech signal itself rather than on the background noise, when making judgments of overall quality.

3. Objective measures

Five widely used objective speech quality measures were evaluated: segmental SNR (segSNR), weighted-slope spectral (WSS) distance [8], perceptual evaluation of speech quality (PESQ) [9, 10], log likelihood ratio (LLR) and Itakura-Saito (IS) distance measure [1]. Correlations were run between the above objective measures and each of the three subjective rating scores (SIG, BAK, OVRL). A total of 1792 processed speech samples were included in the correlations encompassing two SNR levels, four different types of background noise and speech/noise distortions introduced by 13 different speech enhancement algorithms. A total of 43008 subjective listening scores for the three rating scales were used in the computation of the correlation coefficients.

The LLR measure is defined as [1] (pp. 48)

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

where $\vec{a}_c$ is the LPC vector of the original speech signal, $\vec{a}_p$ is the LPC vector of the enhanced speech, and $\mathbf{R}_c$ is the autocorrelation matrix of the original speech signal. The IS measure is defined as:

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

where σ_c^2 and σ_p^2 are the LPC gains of the clean and enhanced signals respectively. The segSNR measure was computed as per [11]. Only frames with segmental SNR in the range of -10dB to 35dB were considered in the average.

Among the five objective measures mentioned above, the PESQ measure is the most complex to compute, and it is recommended by ITU-T for speech quality assessment of 3.2 kHz (narrow-band) handset telephony and narrow-band speech codecs [9]. The final PESQ score is obtained by a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values as follows A_{ind} [5]:

$$pesq = a_0 - a_1 \cdot D_{ind} - a_2 \cdot A_{ind} \quad (2)$$

where $a_0 = 4.5$, $a_1 = 0.1$, and $a_2 = 0.0309$. Clearly we can not expect the PESQ measure to correlate highly with all three quality measures (speech distortion, noise distortion and overall quality). For that reason, 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 scale. The modified PESQ measures were obtained by treating a_0, a_1 and a_2 in Eq. (2) 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 Eq. (2) were treated as independent obtained variables in the regression analysis. The actual subjective scores for the three scales were used in the regression analysis. We obtained three different measures suitable for predicting signal distortion ($pesq_s$), noise distortion ($pesq_b$) and overall speech quality ($pesq_o$):

$$pesq_s = 4.754 - 0.186 \cdot D_{ind} - 0.008 \cdot A_{ind} \quad (3)$$

$$pesq_b = 5.611 - 0.070 \cdot D_{ind} - 0.068 \cdot A_{ind} \quad (4)$$

$$pesq_o = 4.906 - 0.148 \cdot D_{ind} - 0.021 \cdot A_{ind} \quad (5)$$

We refer to the modified PESQ measures as the mPESQ measures.

4. Evaluation results

Two figures of merit are computed for each objective measure. The first one is the correlation coefficient (Pearson's correlation) between the subjective quality measure S_d and the objective measure O_d, and is given by:

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

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:

$$\hat{\sigma}_e = \hat{\sigma}_s \sqrt{1 - \rho^2}$$

where $\hat{\sigma}_s$ 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.

	segSNR	WSS	PESQ	mPESQ	LLR	IS
SIG	0.19	0.50	0.57	0.65	0.64	0.12
BAK	0.42	0.37	0.48	0.56	0.23	0.07
OVRL	0.31	0.53	0.65	0.67	0.61	0.11

Table 3: Estimated correlation coefficients for six objective measures.

	segSNR	WSS	PESQ	mPESQ	LLR	IS
SIG	0.78	0.68	0.65	0.60	0.61	0.79
BAK	0.53	0.54	0.51	0.48	0.57	0.58
OVRL	0.58	0.52	0.46	0.45	0.49	0.61

Table 4: Standard deviations of the error ($\hat{\sigma}_e$) for the correlations with the six objective measures.

We computed the correlation coefficient (ρ) and $\hat{\sigma}_e$ of each objective measure with each of the three subjective measures (SIG, BAK, OVRL) across all conditions. Table 3 shows the correlation coefficients of the objective measures with the subjective scores, and Table 4 shows the corresponding standard deviations of the error $\hat{\sigma}_e$ obtained for each objective measure. From Table 3 and 4, we can see that of all the conventional objective measures, the LLR measure performed the best in terms of predicting signal distortion (SIG), followed by the PESQ, WSS, segSNR and IS measures. In terms of noise distortion (BAK), the PESQ measure performed the best, followed by the segSNR, WSS, LLR and IS measures. In terms of overall speech quality (OVRL), the PESQ measure performed the best, followed by the LLR, WSS, segSNR and IS measures. The proposed PESQ measures (Eqs. 3,4,5) were better than the conventional measures for all three subjective scales. Particularly large improvement was obtained in predicting signal and background distortions with the $pesq_s$ and $pesq_b$ measures (Eq. 3,4) respectively. Overall, the PESQ measure did not yield as high correlation with speech quality as found with speech transmitted through communication networks [10]. A similar finding was also reported in [12].

4.1. Composite measures

Aiming to improve further the correlation coefficients, we considered composite measures. Composite objective measures are obtained by linearly combining existing objective measures to form a new measure [1]. This can be done by utilizing linear regression analysis [1], or by applying nonlinear techniques (e.g. [12]). In this paper, we used multiple linear regression analysis to form the following composite measures: (a) a measure we call C_{sig} for signal distortion (SIG) formed by linearly combining the LLR, PESQ, and WSS measures; (b) a measure we call C_{bak} for noise distortion (BAK) formed by linearly combining the segSNR, PESQ, and WSS measures, and (c) a measure we call C_{ovl} for overall quality (OVRL) formed by linearly combining the PESQ, LLR, and WSS measures. The three new composite measures obtained from multiple linear regression analysis are given below:

	C_{sig}	C_{bak}	C_{ovl}
SIG	0.7 (0.56)		
BAK		0.58 (0.48)	
OVRL			0.73 (0.42)

Table 5: Correlation coefficients and standard deviations of the error (shown in parenthesis) for the new composite measures.

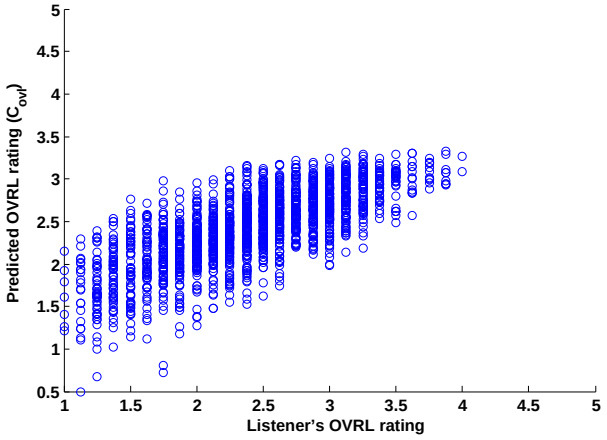

Figure 2: Scatter plot of C_{ovl} vs. the true subjective OVRL ratings.

$$
\begin{aligned}
C_{sig} &= 3.093 - 1.029 \cdot LLR + 0.603 \cdot PESQ - \\
&\quad 0.009 \cdot WSS \\
C_{bak} &= 1.634 + 0.478 \cdot PESQ - 0.007 \cdot WSS + \\
&\quad 0.063 \cdot segSNR \\
C_{ovl} &= 1.594 + 0.805 \cdot PESQ - 0.512 \cdot LLR - \\
&\quad 0.007 \cdot WSS
\end{aligned}
$$

Table 5 shows the correlation coefficients and standard deviations of the error for these new composite measures. As can be seen, these new composite measures show moderate improvements over the existing objective measures. Highest correlation ($\rho = 0.73$) was obtained with the C_{ovl} measure. Scatter plot of C_{ovl} vs. the true subjective OVRL ratings is shown in Fig. 2.

5. Conclusions

The present study evaluated several objective measures commonly used for evaluating speech quality. The test conditions included speech/noise distortions introduced by four real world noises at two SNR levels (5 and 10 dB), and 13 representative speech enhancement algorithms [4]. In contrast to the correlations obtained with speech codecs and communication channel distortions [1], our data shows that most of the current objective measures are not adequate in predicting the subjective quality of noisy speech enhanced by noise suppression algorithms. The segSNR measure, for instance, which is widely used for evaluating the performance of speech enhancement algorithms, yielded a very poor correlation coefficient ($\rho = 0.31$) with overall quality. Further research is needed to improve the correlations of current objective measures with subjective speech quality. Our data demonstrated that the use of composite measures can greatly improve the correlations of existing objective measures.

6. Acknowledgements

Research is supported in part by Grant No. R01 DC07527 from NIDCD/NIH. The authors would like to thank Dr. Alan Sharpley of Dynastat, Inc for all his help and advice throughout the project.

7. References

[1] S. Quackenbush, T. Barnwell, and Clements, *Objective measures of speech quality*, Englewood Cliffs, NJ: Prentice-Hall, 1988.

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

[3] Y. Hu and P. C. Loizou, "Subjective evaluations and comparisons of speech enhancement methods," submitted to *Speech Communication*, 2006.

[4] Y. Hu and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Vol. I, pp. 153-156, 2006.

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

[6] ITU-T P.56, *Objective measurement of active speech level.*, ITU-T Recommendation P.56, 1993.

[7] H. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions.," in *ISCA ITRW ASR2000*, Sept. 2000, Paris, France.

[8] D. Klatt, "Prediction of perceived phonetic distance from critical band spectra," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1982, pp. 1278–1281.

[9] ITU-T P.835, *Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm*, ITU-T Recommendation P.835, 2003.

[10] 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 Processing*, 2001, pp. 749–752.

[11] J. Hansen and B. Pellom, "An effective quality evaluation protocol for speech enhancement algorithms," in *ICSLP*, Dec. 1998, pp. 2819–2822, Sydney, Australia.

[12] B. Grundlehner, J. Lecocq, R. Balan, and J. Rosca, "Performance assessment method for speech enhancement systems," in *Proc. of the first annual IEEE BENELUX/DSP Valley Signal Processing Symposium*, Apr. 2005, Het Provinciehuis, Antwerp, Belgium.