A Composite Objective Measure on Subjective Evaluation of Speech Enhancement Algorithms

Zhibin Lin, Lu Zhou, and Xiaojun Qiu
zblin@nju.edu.cn, yxyyswy@163.com, xjqiu@nju.edu.cn

Key Laboratory of Modern Acoustics and Institute of Acoustics, Nanjing University, Nanjing, China.

The purpose of speech enhancement algorithms is to improve speech quality, naturalness and intelligibility by eliminating the background noise and improving signal to noise ratio. There are several objective measures predicting the quality of noisy speech enhanced by noise suppression algorithms, and different objective measures capture different characteristics of the degraded signal. In this paper, the multiple linear regression analysis is used to obtain a composite measure which has high correlation with subjective tests, and the performance of several speech enhancement algorithms under car noise conditions is compared. The uncertainty of the results of the proposed measures on different speech enhancement algorithms is analyzed, and the reliability of the results is discussed.

0 INTRODUCTION

Speech enhancement is concerned with improving perceptual aspects of speech that is degraded by background noise, and the main aim of speech enhancement is to improve speech quality and signal to noise ratio (SNR) level while preserving speech intelligibility. A large number of speech enhancement algorithms have been proposed such as the spectral-subtractive algorithms, the wiener algorithm, the minimum mean square error (MMSE) algorithms and the subspace algorithms [1].

Speech enhancement algorithms typically degrade the speech signal component while suppressing the background noise, particularly in low SNR conditions, which complicates the subjective evaluation of speech enhancement algorithms. It is not clear whether listeners evaluate their overall quality judgments basing on the signal distortion component, noise distortion component, or both, and this uncertainty decreases the reliability of the rating. Hence, ITU-T Rec. P.835 has been designed to lead the listeners to rate the speech signal, the background noise, and the overall effect of speech and noise separately [2].

Listening tests are usually time-consuming and expensive to conduct [3], so several objective measures have been proposed. However, most of these objective measures were developed for the purpose of evaluating the distortions introduced by speech codecs and communication channels, and it is not clear whether these objective measures are suitable for evaluating the speech quality enhanced by speech enhancement algorithm [4-5]. As a result, only a small number of studies were presented to examine the correlation between objective measure and the subjective quality evaluation of enhanced noise speech, such as the perceptual evaluation of speech quality (PESQ) for speech codec [6-11], the log likelihood ratio (LLR), the cepstrum (CEP) and segmental SNR (segSNR). However, the PESQ measure did not yield as high correlation coefficients with speech quality as that found with speech transmitted through network, whose correlation efficient was about 0.65 in term of signal distortion. The other conventional objective measures (CEP, LLR and segSNR) performed moderately well (by about 0.60) with overall quality whereas yielded poor correlation coefficient (by about 0.30) with ratings of background noise distortion [1].

Aiming to further improve the correlation coefficients for different types of distortion introduced by speech enhancement algorithms, a multiple linear regression analysis is used to obtain a new composite measure, which only consisted of five different objective measures. Then the measurement uncertainty of the proposed measure under different speech enhancement algorithms is investigated, and the reliability of the results is discussed.

1 A COMPOSITE MEASURE

Several existing objective measures have been combined to form a new measure by utilizing the linear regression analysis or nonlinear techniques [13]. Five widely used objective speech quality measures are selected in this paper, and

they are the perceptual evaluation of speech quality (PESQ), the log likelihood ratio (LLR), the cepstrum (CEP), the frequency-weighted segmental SNR (fwSNRseg) and the frequency-variant fwSNRseg with 25 bands (fwSNRsegVar). As mentioned above, these different objective measures only capture different characteristics of the distorted signal which is monotonous to rate different kind of distorted signal [1].

The PESQ measure described in the ITU-T P.862 is capable of performing reliably across a wide range of codecs and network conditions. However, the performance of PESQ is found to be sensitive to measurement noise when clean reference samples were used [14]. The range of PESQ score is [−0.5, 4.5]. The log likelihood ratio (LLR) measure and the cepstrum (CEP) measure are proposed based on the dissimilarity between all-pole models of the clean and enhanced speech signals, which assume that speech can be represented by a *p*-th order all-pole model over short time intervals. The LLR measure represents the ratio of the energies of the prediction residuals of the enhanced and clean signals. The range of LLR score is [0, 2]. The CEP measure provides an estimate of the log spectral distance between two spectra with a score range of [0, 10]. The advantage of using the fwSNRseg is the flexibility of assigning different weights for different frequency bands. The range of fwSNRseg score is [−10 dB, 35 dB]. Alternatively, the weights for each band can be obtained using the regression analysis to obtain fwSNRsegVar, which has a range of [−10 dB, 35 dB].

Various statistics have been used to evaluate interrater reliability. The most common statistic is the Pearson's correlation coefficient between the first and second ratings. To obtain the Pearson's coefficient, listeners are presented with the same speech samples at two different testing sessions, and the Pearson's correlation between the subjective quality measure S_d and the objective measure O_d, is given by [1]

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

where $\overline{S_d}$ and $\overline{O_d}$ are the mean values of S_d and O_d, respectively.

The standard deviation of the error when the objective measure is used in place of the subjective measure is given by [1]

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

where $\hat{\sigma}_s$ and $\hat{\sigma}_e$ are the standard deviation of S_d and error. A smaller value of $\hat{\sigma}_e$ indicates that the objective measure is better at predicting subjective quality [13].

The first five columns (excluding the title column) in Table 1 show the correlation coefficients and standard deviations of the error for the five objective measures above, where the correlations were run between the objective measures and the subjective rating scores. A total of 43008 subjective scores were included in the correlations computation, encompassing two SNR level (5 dB and 10 dB). And the noisy database contains 30 IEEE sentences, which were produced by three male and three female speakers and recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment, and sampled at 25 kHz and then down sampled to 8 kHz [1].

Table 1. Correlation coefficients and standard deviations of the error (shown in parenthesis) for the five objective measures and the proposed measure

	PESQ	LLR	CEP	fwSNRseg	fwSNRsegVar	proposed measure
SIG	0.57(0.65)	0.66(0.59)	0.65(0.60)	0.67(0.56)	0.73(0.54)	0.673(0.253)
BAK	0.48(0.51)	0.26(0.56)	0.22(0.57)	0.27(0.59)	0.51(0.50)	0.609(0.308)
OVL	0.65(0.46)	0.63(0.47)	0.60(0.49)	0.64(0.47)	0.70(0.43)	0.674(0.298)

From Table 1, it can be found that the fwSNRsegVar measure yields the highest correlation with the three subjective scales in terms of OVL (overall quality), SIG (signal distortion) and BAK (background distortion). The second best measure is the PESQ measure, and it is also found that the LLR, CEP and fwSNRseg measures performed best in terms of predicting overall quality and signal distortion, but with a large standard deviation.

In order to improve the correlation coefficients, a multiple linear regression analysis is used to obtain a new composite measure. Basing on the database mentioned above, a total of 14 listeners (22-50 years old) were recruited for the listening test. No listeners participated in a listening test in the previous 3 months before this test. Correlations are calculated between the objective measure and the three subjective rating scores. A total of 5040 subjective listening scores for three rating scales are obtained, including two SNR levels (5 dB and 10 dB) and two different types of background noise. The regression analysis is applied on the objective scores of five measures above and the subjective

scores for the three scales based on least square method by using the best fitting straight line. The weighting coefficients of each parameter are obtained, and the derived composite measures for signal distortion (C_{SIG}), noise distortion (C_{BAK}), and overall quality (C_{OVL}) are as follows,

$$C_{SIG} = 1.856 + 0.135PESQ_{SIG} - 1.569LLR_{SIG} + 0.338CEP_{SIG} + 0.044\,fwSNRseg_{SIG} + 0.224\,fwSNRsegVar_{SIG}, \quad (3)$$

$$C_{BAK} = -0.343 + 0.484PESQ_{BAK} - 2.548LLR_{BAK} + 0.646CEP_{BAK} - 0.049\,fwSNRseg_{BAK} + 0.520\,fwSNRsegVar_{BAK}, \quad (4)$$

$$C_{OVL} = -0.835 + 0.610PESQ_{OVL} - 3.229LLR_{OVL} + 0.804CEP_{OVL} + 0.313\,fwSNRseg_{OVL} - 0.008\,fwSNRsegVar_{OVL}. \quad (5)$$

where the PESQ, LLR, CEP, fwSNRseg and fwSNRsegVar indicate the objective scores, and the subscript indicates objective measure derived for signal distortion (SIG), background noise distortion (BAK) and overall quality (OVL).

The last column in Table 1 shows the correlation coefficients and standard deviations of the error for the proposed composite measures. Compared with other five objective measures, the proposed composite measures show moderate improvements over the existing objective measures in correlation, whereas the standard deviations of the error are smaller than other objective measures. The highest correlation (ρ =0.674) is obtained with the C_{OVL} measure. Being compared with the fwSNRsegVar method, the correlation of C_{SIG} and C_{OVL} declines slightly, however, smaller standard deviations of the error are obtained with the proposed measure. This property might be better for evaluating subjective quality of distorted speech [13].

2 UNCERTAINTY OF THE PROPOSED MEASURE

2.1 Selection of Experiment Conditions and Results

In order to evaluate the performance of the proposed composite measure for different speech enhancement algorithm, the same database mentioned above are selected, whose sentenses are corrupted only in car background noise environments.

In the tests, six different speech enhancement algorithms are adopted, i.e., the minimum mean square error (MMSE-SPU) algorithms [15], the logMMSE algorithm, the logMMSE algorithm with noise estimation (logMMSE_ne), the basic spectral-subtractive algorithms (Specsub), the subspace algorithm with embedded pre-whitening (Karhunen-Loeve Transform, KLT) and the wiener algorithm based on a priori SNR estimation (Wiener_as) [16].

The noise-corrupted sentences are processed by the speech enhancement algorithms mentioned above. Tables 2 and 3 present the objective scores obtained with the proposed measure, where the obtained average, the span of the objective values and the standard deviation are shown.

Table 2. The objective scores of the proposed measures for different algorithms (10 dB)

Algorithm	C_{SIG}			C_{BAK}			C_{OVL}		
	Avg	Span	σ	Avg	Span	σ	Avg	Span	σ
MMSE	3.932	0.593	0.1693	3.217	0.572	0.1421	3.263	0.816	0.2025
Specsub	3.886	0.542	0.1458	3.035	0.548	0.1428	3.018	0.760	0.1860
Wiener_as	3.949	0.602	0.1631	3.028	0.519	0.1342	3.099	0.564	0.1579
logMMSE	3.975	0.557	0.1654	3.158	0.427	0.1246	3.237	0.696	0.1799
KLT	3.646	0.721	0.2114	3.039	0.622	0.1581	2.937	0.817	0.2041
logMMSE_ne	3.891	0.646	0.1609	3.134	0.553	0.1399	3.162	0.755	0.1917

Table 3. The objective scores of the proposed measures for different algorithms (5 dB)

Algorithm	C_{SIG}			C_{BAK}			C_{OVL}		
	Avg	Span	σ	Avg	Span	σ	Avg	Span	σ
MMSE	3.661	0.512	0.1287	2.988	0.562	0.1441	2.899	0.779	0.1792
Specsub	3.473	0.535	0.1349	2.792	0.681	0.1640	2.580	0.891	0.2105
Wiener_as	3.649	0.505	0.1208	2.843	0.557	0.1471	2.761	0.654	0.1739
logMMSE	3.689	0.565	0.1296	2.953	0.552	0.1518	2.882	0.761	0.1876
KLT	3.357	0.547	0.1359	2.841	0.499	0.1313	2.610	0.568	0.1586
logMMSE_ne	3.606	0.646	0.1410	2.892	0.754	0.1890	2.790	0.905	0.2096

It can be found from the Tables that there is a large variability among different algorithms. The average objective score decreases for the low SNR condition, this is reasonable in terms of the perception of people under low SNR condition. The objective values vary significantly even under the same algorithm. For example, for the case of C_{OVL} at 10 dB SNR, the span of the KLT is as large as 0.817.

2.2 Statistical Analysis of the Uncertainty of the Proposed Measures Values

To analyze the probability distribution of the objective values, the histogram of the data obtained with the 6 adopted algorithms are processed. For the sake of brevity, Fig. 1 only shows the histogram related to the algorithm of the Wiener_as on C_{BAK}. It is obtained by dividing the horizontal axis into bins of constant width equal to 0.1 and by reporting on the vertical axis the frequencies of the whole objective results falling into each bin.

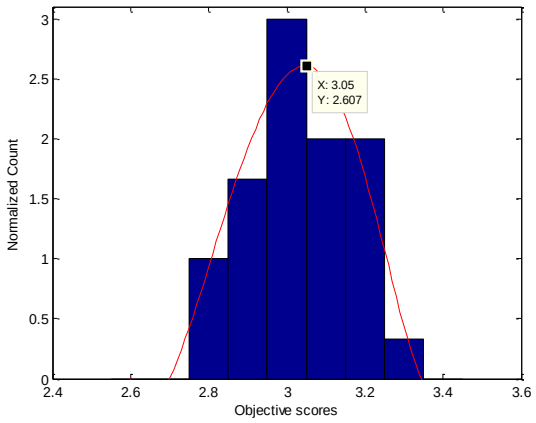

Fig. 1. Normalized histogram of the objective values for the Wiener_as algorithm (10 dB)

The mean and variance are 3.028 and 0.1342 respectively. By using a normal distribution to fit the distribution, the probability density function shows that the mean is 3.05 with a variance of 0.1308 [17]. The good agreement between the two curves suggests that the measurement results can be considered as normal distribution. The normal distribution assumption can be formally analyzed with the Chi-square test, and result shows that the assumption is true under the significance level of 0.05. In order to avoid Type II error (the error of failing to reject a null hypothesis when it is false) in the normality test, the Skewness and Kurtosis parameters of the measurement results are calculated, and the results are shown in Table 4.

Table 4. Statistical parameters of different speech enhancement methods

Algorithm	Skewness	Kurtosis
MMSE	-0.32	3.05
Specsub	-0.29	2.84
Wiener_as	0.06	2.93
KLT	0.02	2.90
logMMSE	-0.15	2.93
logMMSE_ne	-0.23	2.88

The Skewness parameter is a well-known indicator of probability density function symmetry with respect to its center value, whereas the Kurtosis indicates if the probability density function if peaked or flat with respective to a normal probability density function. In particular, null Skewness and Kurtosis equal to 3 are expected for normally distributed data [17]. As illustrated in Table 4, the measurement results of the Wiener_as algorithm can be approximately considered as normal distribution. Similar calculations (not presented in the paper) shows that the conclusion can be extended to the other algorithms too. Basing on this conclusion, the performance of different algorithms under car background noise environments is compared, and the reliability of the results is analyzed.

Table 5. Confidence interval of C_{OVL} values (10 dB)

Algorithm	CI 95%	Margin of error	CI 95%-lower limit	MOS (OVL)
MMSE	(3.127,3.259)	0.0662	3.138	3.25
Specsub	(2.949, 3.088)	0.0694	2.961	2.56
Wiener_as	(3.041, 3.158)	0.0590	3.050	2.81
KLT	(2.860, 3.013)	0.0762	2.873	2.49
logMMSE	(3.169, 3.304)	0.0672	3.181	3.37
logMMSE_ne	(3.091, 3.234)	0.0716	3.103	3.13

Table 6. Confidence interval of C_{OVL} values (5 dB)

Algorithm	CI 95%	Margin of error	CI 95%-lower limit	MOS (OVL)
MMSE	(2.8321, 2.9660)	0.0669	2.8435	2.77
Specsub	(2.5014, 2.6586)	0.0786	2.5147	2.13
Wiener_as	(2.6961, 2.8260)	0.0649	2.7071	2.40
KLT	(2.5503, 2.6687)	0.0592	2.5603	2.13
logMMSE	(2.8122, 2.9523)	0.0700	2.8241	2.90
logMMSE_ne	(2.7115, 2.8680)	0.0783	2.7247	2.65

For a random variable, the confidence coefficient (1-α) of the mean value can be calculated by [17],

$$\left(\bar{X} \pm s \cdot t_{\alpha/2}(n-1)/\sqrt{n}\right) \tag{6}$$

One-side confidence lower limit can be calculated by

$$\underline{\mu} = \bar{X} - s \cdot t_{\alpha}(n-1)/\sqrt{n} \tag{7}$$

where $\bar{X}$ and s represent the means and standard deviation, n corresponds to the number of samples, $t_{\alpha/2}(n-1)$ and $t_{\alpha}(n-1)$ are percentiles of the T distribution [17], and $s \cdot t_{\alpha/2}(n-1)/\sqrt{n}$ corresponds to margin of error at confidence coefficient (1−α).

For the sake of brevity, Table 5 only shows the confidence interval of C_{OVL} measure. In the second column of Table 5, the 95% confidence intervals of means are given. The third and fourth columns indicate the margin of error and the 95% confidence lower limit, whereas in the last column, the subjective MOS values in 10 dB car noise are obtained from Loizou [1]. Taking the MMSE algorithm for example, the average of the C_{OVL} values is between 3.127 and 3.259. If picking any value in this range as the approximation of the true value, the margin of error is not more than 0.1324 (two times of the margin of error). The reliability of the conclusion is 95%. Since the 95% confidence interval of the average estimate overlaps, z-test was performed to further investigate the performance of algorithms [17]. The results show that when the significance level is 0.05, there is no significant difference among the means of logMMSE, MMSE and logMMSE_ne; the same between the means of Specsub and KLT. Multiple C_{OVL} values of six algorithms indicate that the performance of different algorithms is significantly different. This conclusion is consistent to the 95% confidence lower limit.

As can be seen from the last column, subjective scores are consistent to the objective conclusion. The performances of the logMMSE, MMSE and logMMSE_ne algorithms are superior to others. The Wiener_as algorithm is better than the Specsub and KLT algorithms. The same conclusion can be drawn for the case with 5 dB SNR according to Table 6. Therefore, these testing results illustrate the differences between the speech enhancement algorithms are significant in both the objective and subjective testings. It should be noted here that the actual values of objective measures and subjective measures in Tables 5 and 6 do not line up well in some cases. The reason is that the correlations between objective and subjective measurements are low, especially for distorted speech.

3 CONCLUSIONS

A multiple linear regression analysis is used in this paper to obtain a new composite measure which has high correlation coefficients with small standard deviations. With the proposed composite measure, the majority of the correlation coefficients in terms of BAK are improved by about 0.2, and the standard deviations of the error are declined by about 0.2 to 0.4 in terms of OVL, SIG and BAK. Then, the uncertainty of the proposed measure under different test conditions is analyzed, and the values obtained by the proposed measure are shown to have almost normal distribution. Finally, 6 speech enhancement algorithms are investigated with the proposed measure, and the result shows that the differences between the speech enhancement algorithms are significant in both the objective and subjective testings. The composite objective measure can be regarded not only as subjective estimator but also as an overall system performance parameter for speech enhancement algorithms.

4 REFERENCES

[1] P. C. Loizou, Speech Enhancement: *Theory and Practice*, CRC Press, 2007.

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

[3] AES Staff Writer, "Measuring and predicting perceived audio quality," *Journal of the Audio Engineering Society*, vol. 53, pp.443-448, 2005.

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

[5] S. Möller and J. Berger, "Describing telephone speech codec quality degradations by means of impairment factors," *Journal of the Audio Engineering Society*, vol. 50, pp. 667-680, 2002.

[6] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, C. Colomes, M. Keyhl, G. Stoll, K. Brandenburg, and B. Feiten, "PESQ-The ITU-Standard for Objective Measurement of Perceived Audio Quality," *Journal of the Audio Engineering Society*, vol. 48, pp. 3-29, 2000.

[7] ITU-T Rec. P.862, "Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-to-end Speech Quality Assessment of Narrow-band Telephone Networks and Speech Codecs," International Telecommunication Union, Geneva, Switzerland (2001 Feb.).

[8] ITU-T Rec. P.862.1, "Mapping function for transforming P.862 raw result scores to MOS-LQO," Geneva, Switzerland (2003 Nov.).

[9] ITU-T Rec. P.862.2, "Wideband extension to Recommendation P.862 for the assessment of wideband telephone networks and speech codecs," Geneva, Switzerland (2005 Nov.).

[10] A. W. Rix, M. P. Hollier, A. P. Hekstra and J. G. Beerends, "PESQ, the new ITU standard for objective measurement of perceived speech quality, Part 1 - Time alignment," *Journal of the Audio Engineering Society*, vol. 50, pp. 755-764 (2002 Oct.).

[11] J. G. Beerends, A. P. Hekstra, A. W. Rix and M. P. Hollier, "PESQ, the new ITU standard for objective measurement of perceived speech quality, Part II - Perceptual model," *Journal of the Audio Engineering Society*, vol. 50, pp. 765-778 (2002 Oct.).

[12] A.W. Rix, J. G. Beerends, M. P. Hollier and A. P. Hekstra, "Perceptual Evaluation of Speech Quality (PESQ) - A New Method for Speech Quality Assessment of Telephone Networks and Codecs," in Proc. IEEE International, conference on Acoustics, Speech and Signal Processing, vol. 2, pp. 749 – 752, 2001.

[13] Y. Hu and P. C. Loizou, "Evaluation of objective measures for speech enhancement," in Proc. Interspeech, pp. 1447-1450, 2006.

[14] J. Salmela, O. Kirla, A. Lakaniemi, V. Mattila, and N. Zacharov, "Application and verification of the objective quality assessment method according to ITU Recommendation series ITU-T P.862," *Journal of the Audio Engineering Society*, vol. 54, pp. 1189-1202, 2006.

[15] D. E. Tsoukalas, J. N. Mourjopoulo, and G. Kokkinaki, "Perceptual filters for audio signal enhancement," *Journal of the Audio Engineering Society*, vol. 45, pp. 22-36, 1997.

[16] Y. Hu, and P. C. Loizou, "Subjective comparison of speech enhancement algorithms," IEEE International Conference Acoustic, Speech, and Signal Processing, pp. 153-156, 2006.

[17] R. A. Johnson, *Miller&Freund's Probability and Statistics for Engineers*, Prentice-Hall, 2005.