Objective Quality Evaluation of Noise-suppressed Speech: Effects of Temporal Envelope and Fine-structure Cues

Fei Chen [1 2], *Yi Hu* [3]

[1] Division of Speech and Hearing Sciences, The University of Hong Kong
[2] Dept. of Electrical & Electronic Engineering, South Univ. of Science and Technology of China
[3] Department of Electrical Engineering & Computer Science, University of Wisconsin-Milwaukee

feichen1@hku.hk, huy@uwm.edu

10.21437/Interspeech.2014-467

Abstract

While temporal envelope and fine-structure cues are known to be good predictors for speech intelligibility, it is not clear how well they are correlated with subjective quality ratings, particularly those using noise-suppressed speech. The present work evaluated the performance of two objective measures (i.e., NCM and TFSS), which were originally developed with primarily envelope or fine-structure cue as speech intelligibility indices, when they were applied for predicting the subjective quality ratings of noise-suppressed speech along three dimensions of signal distortion, noise distortion and overall quality. We considered a wide range of distortion introduced by four types of real-world noises at two signal-to-noise-ratio levels and by four classes of noise-suppression algorithms. This work finds that the present envelope- and fine-structure-based measures poorly predict the subjective quality ratings of noise-suppressed speech. The PESQ measure is so far the best choice in terms of objectively evaluating both subjective quality ratings and intelligibility scores of noise-suppressed speech.

Index Terms: PESQ, temporal envelope, fine structure, speech quality, noise suppression

1. Introduction

Quality is an important attribute of human speech that needs be preserved by speech enhancement algorithms [1-3]. Subjective listening tests provide perhaps the most reliable method for assessing speech quality and examining the performance of speech enhancement algorithms. However, these tests are normally costly, time-consuming, and laborious. Alternatively, many studies developed objective measures for predicting subjective ratings of speech quality [4-7].

Hu and Loizou evaluated objective quality prediction for speech enhancement by a number of quality measures [8]. They found that most of the current objective measures were not adequate in predicting the subjective quality ratings of noisy speech enhanced by speech enhancement algorithms. The time-domain segmental signal-to-noise-ratio (SNR) measure (segSNR) [4], for instance, which is being widely used for evaluating the performance of speech enhancement algorithms, yielded a very poor correlation (r=0.36) with the overall quality, and the perceptual evaluation of speech quality (PESQ) measure led to a high correlation (r=0.89) with the overall quality [8]. Further research is still needed to improve the correlations of current objective measures with subjective quality ratings of noise-suppressed speech.

Intelligibility is another important attribute for speech communication. The relationship between speech intelligibility and speech quality is not fully understood, and this is in part due to the fact that we have not yet identified the acoustic correlates of quality and intelligibility [1]. Hence, different models (or measures) were developed to assess the quality and intelligibility of processed (e.g., noise-suppressed) speech. Given that many objective measures are available for evaluating speech quality/intelligibility, a question is often raised as whether those objective measures could be used interchangeably for predicting both speech quality ratings and intelligibility scores, or which measure could best account for the variances of both speech quality and intelligibility. Ma et al. evaluated the performance of a number of objective measures (i.e., both quality and intelligibility measures) to predict the intelligibility of noise-suppressed speech [9]. They found that, of all the conventional quality measures, the frequency-weighted segmental SNR (fwSNRseg) and PESQ measures performed modestly well in terms of objectively predicting the intelligibility scores of noise-suppressed speech.

Envelope and temporal fine-structure (TFS) have been long identified as two important acoustic cues for speech intelligibility [10]. Many envelope- and TFS-based intelligibility indices have been developed to successfully predict the intelligibility of noise-distorted and noise-suppressed speech. For instance, the envelope-based normalized covariance metric (NCM) [11] was shown to well predict the intelligibility of speech in a number of conditions (e.g., noise-distortion and vocoder processing) [9, 11-12]. Recently, a TFS spectrum based index (i.e., TFSS) was designed, and it showed good performance to predict the intelligibility of noise-suppressed speech [13]. Tiago et al. proposed a non-intrusive quality and intelligibility metric for reverberated and deverberated speech [14]. The metric was developed with primarily temporal envelope cue, and was found to be well correlated with the quality and intelligibility of reverberated and deverberated speech.

While the contributions of envelope and fine-structure for subjective speech perception and objective speech intelligibility prediction have been investigated in many studies, little is known on how well these envelope and fine-structure based indices (e.g., NCM and TFSS) can predict the subjective quality ratings of noise-suppressed speech. This knowledge is of significance to understand the contribution of acoustic cues for predicting speech quality, and consequently guides our design of new objective quality measures. The purpose of this paper is to assess the contribution of envelope and fine-structure cues for predicting the subjective quality ratings of noise-suppressed speech. More specifically, we would evaluate the performance of two intelligibility measures (i.e., NCM

and TFSS), which were originally developed with primarily envelope or fine-structure cue, when they were used for objective quality prediction of noise-suppressed speech.

2. Subjective quality ratings of noise-suppressed speech

Twenty sentences from the NOIZEUS corpus were corrupted by four background noises (i.e., car, street, babble and train) at two signal-to-noise-ratio (SNR) levels (i.e., 5 and 10 dB), processed by four classes (i.e., spectral subtractive, subspace, statistical-model-based, and Wiener filtering) of speech enhancement algorithms, and presented to 32 listeners for subjective evaluation of speech quality. The four classes of speech enhancement algorithms included 13 different algorithms, i.e., multiband spectral subtraction, spectral subtraction using reduced delay convolution and adaptive averaging, generalized subspace approach, perceptually based subspace approach, MMSE (i.e., minimum mean squared error estimate), log-MMSE, and log-MMSE under signal presence uncertainty, and Wiener-filtering algorithms with the *a priori* SNR estimation based method, the audible-noise suppression method, and the method based on wavelet thresholding the multitaper spectrum [15-18].

The subjective quality evaluation tests were designed according to ITU-T recommendation P.835 [19]. The P.835 methodology was designed to reduce listener's uncertainty in a subjective test as to which component(s) of a noisy speech signal (i.e., speech signal, background noise, or both) should affect their ratings of overall quality. This testing method instructed listeners to successively rate the processed (e.g., noise-suppressed) speech signal on:

1) The speech signal alone using a five-point scale of signal distortion (SIG) [1=very unnatural, very degraded; 2=fairly unnatural, fairly degraded; 3=somewhat natural, somewhat degraded; 4=fairly natural, little degradation; 5=very natural, no degradation];
2) The background noise alone using a five-point scale of background conspicuous/intrusiveness (BAK) [1=very conspicuous, very intrusive; 2=fairly conspicuous, somewhat intrusive; 3=noticeable but not intrusive; 4=somewhat noticeable, 5=not noticeable]; and
3) The overall effect using the five-point scale of the Mean Opinion Score (OVRL) [1=bad; 2=poor; 3=fair; 4=good; 5=excellent]

More details on noise-suppression algorithms, testing methodology and subjective quality ratings can be found in [18].

3. Objective measures for quality evaluation

3.1. The envelope-based measure

The NCM measure is computed as follows [11]. The stimuli are first bandpass filtered (via fourth-order Butterworth filters) into N bands spanning the signal bandwidth (300–3400 Hz in this study). The cutoff frequencies of the N band-pass filters are computed according to the cochlear frequency-position mapping function $f = 165.4 \times (10^{0.06 \cdot d} - 1)$ in this study, where f is the –3 dB cut-off frequencies and d denotes the distance along the cochlea partition [20]. The envelope of each band is computed using Hilbert transform, and then down-sampled to $2 \times f_{cut}$ Hz, thereby limiting the envelope modulation rate to f_{cut} Hz. Let $x_i(t)$ and $y_i(t)$ be the down-sampled envelope in the ith band of the clean signal and the processed signal, respectively. The normalized covariance in the ith frequency band is computed as:

$$\rho_i = \frac{\sum_t (x_i(t) - \mu_i)(y_i(t) - \nu_i)}{\sqrt{\sum_t (x_i(t) - \mu_i)^2} \sqrt{\sum_t (y_i(t) - \nu_i)^2}}, \quad (1)$$

where μ_i and ν_i are the mean values of the $x_i(t)$ and $y_i(t)$, respectively. The apparent SNR in each band is computed as:

$$\mathrm{SNR}_i = 10 \log_{10} \left(\frac{\rho_i^2}{1 - \rho_i^2} \right), \quad (2)$$

and subsequently limited to the range of [−15,15] dB. The transmission index (TI) in each band is computed by linearly mapping the SNR values between 0 and 1 as:

$$\mathrm{TI}_i = (\mathrm{SNR}_i + 15)/30. \quad (3)$$

Finally, the transmission indices are averaged across all bands to produce the NCM index, as:

$$\mathrm{NCM} = \frac{\sum_{i=1}^{N} w_i \times \mathrm{TI}_i}{\sum_{i=1}^{N} w_i}, \quad (4)$$

where w_i denotes the weight applied to each band. This study used the most common articulation index (AI) weight [21].

3.2. The fine-structure-based measure

The TFSS measure is computed as follows [13]. As in computing the NCM measure, the stimuli are first band-pass filtered into N bands spanning the signal bandwidth. Let $x_i(t)$ denote the TFS waveform of the clean speech signal in the ith frequency band, and let $y_i(t)$ be the TFS waveform of the processed speech signal in the ith frequency band. The magnitude-squared coherence (MSC) between $x_i(t)$ and $y_i(t)$ is computed by first dividing these signals into M overlapping windowed frames, computing the cross power spectrum for each frame using the fast Fourier Transform (FFT), and then averaging across all frames. The MSC of the TFS spectra at frequency f is given as:

$$\mathrm{MSC}_i(f) = \frac{\left| \sum_{m=1}^{M} X_i^{(m)}(f) Y_i^{(m)*}(f) \right|^2}{\sum_{m=1}^{M} \left| X_i^{(m)}(f) \right|^2 \cdot \sum_{m=1}^{M} \left| Y_i^{(m)}(f) \right|^2}, \quad (5)$$

where asterisk denotes the complex conjugate, and $X_i^{(m)}(f)$ and $Y_i^{(m)}(f)$ are the FFT spectra of signals $x_i(t)$ and $y_i(t)$, respectively, computed in the mth frame. Subsequently, a coherence index (coh_i) is computed in the ith band as:

$$\mathrm{coh}_i = \frac{1}{N_i} \sum_{j=b_i}^{e_i} \mathrm{MSC}_i(j), \quad (6)$$

where b_i and e_i are the indices of frequency bins spanning the bandwidth of the ith band (i.e., $[b_i, e_i]$), and $N_i = e_i - b_i + 1$. Finally, the coherence indices are accumulated across all bands to produce the TFSS index, as:

$$\mathrm{TFSS} = \frac{\sum_{i=1}^{N} w_i \times \mathrm{coh}_i}{\sum_{i=1}^{N} w_i}, \quad (7)$$

where w_i denotes the weight applied to each band (i.e., AI weight).

The coherence index (coh_i) in Eq. (6) takes values between 0 and 1, and assesses the degree to which the TFS of the clean reference speech signal is preserved following the processing by noise-reduction algorithms. A coh_i value near 1, for instance, suggests that the TFS of the input signal is preserved by the processing; while a value near 0 indicates significant degradation/distortion in the TFS in the ith band.

Table 1. Correlation coefficients (|*r*|) between the NCM, TFSS and PESQ values and the subjective quality ratings of noise-suppressed speech. The modulation rate f_{cut} is 12.5 Hz and the frame is 16 ms in computing the NCM and TFSS measures, respectively.

	Measure	The number of channels (N)						PESQ
		1	2	4	8	16	20	
OVRL	NCM	0.63	0.42	0.35	0.03	0.11	0.04	0.89
	TFSS	0.62	0.39	0.16	0.43	0.22	0.01	
SIG	NCM	0.55	0.37	0.38	0.07	0.06	0.10	0.81
	TFSS	0.54	0.40	0.07	0.41	0.22	0.00	
BAK	NCM	0.61	0.32	0.15	0.06	0.20	0.04	0.76
	TFSS	0.61	0.22	0.25	0.32	0.17	0.00	

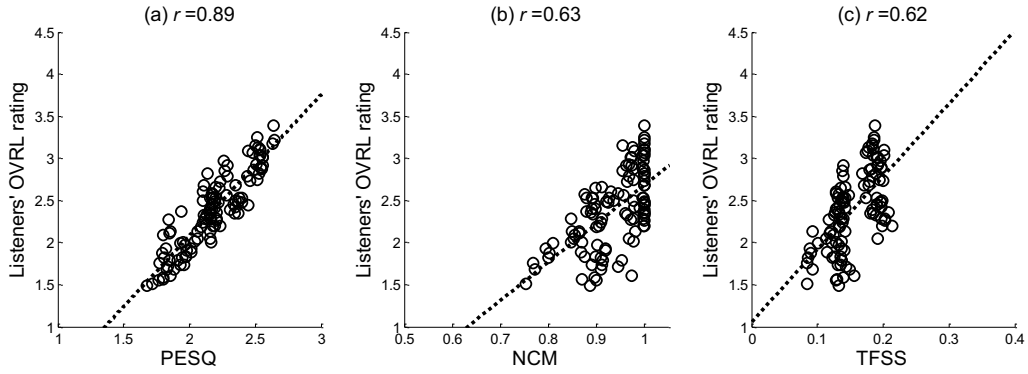

Figure 1. *Scatter plots of listeners' OVRL ratings against the (a) PESQ, (b) NCM, and (c) TFSS values.*

4. Results

The subjective quality ratings of noise-suppressed speech in a total of 112 conditions [=4 maskers × 2 SNR levels × 13 noise-reduction algorithms + 4 maskers × 2 noisy references (i.e., at 2 SNR levels)] were subjective to the correlation analysis with the corresponding average values (across all sentences tested) obtained in each condition by the NCM and TFSS measures. The Pearson's correlation coefficient (*r*) was used to assess the performance of the above quality evaluation.

Table 1 shows the correlation coefficients of the NCM and TFSS measures with the subjective quality ratings of noise-suppressed speech in this study. For the purpose of comparison, the correlation coefficient computed with the PESQ measure is also listed in Table 1, as previous study reported that the PESQ measure yielded the highest correlation coefficients amongst the present objective measures when applied to evaluate the same set of subjective quality ratings [8]. We changed the number of channels (N) used to compute the NCM and TFSS measures (i.e., from 1 to 20) in order to see whether the evaluation performance would be affected by the usage of different N value. Table 1 shows that both NCM and TFSS measures poorly predict the quality ratings of the three types of distortion (i.e., OVRL, SIG and BAK). The highest correlation coefficients for the NCM measure (computed with N=1 and f_{cut} =12.5 Hz) are 0.63, 0.55 and 0.61 for evaluating the OVRL, SIG and BAK distortion, respectively; while those for the TFSS measure (computed with N=1 and frame 16 ms) are 0.62, 0.54 and 0.51 for evaluating the OVRL, SIG and BAK distortion, respectively. All these coefficients are much smaller than those obtained with the PESQ measure, i.e., 0.89, 0.81 and 0.76 for evaluating the OVRL, SIG and BAK distortion, respectively. This suggests that the NCM and TFSS measures do not outperform the PESQ measure in evaluating the subjective quality ratings of noise-suppressed speech. Table 1 also shows that the maximum correlations for both NCM and TFSS measures are obtained when they are computed with N=1. Further increasing the number of channels N may worsen the correlation results. For instance, when N=20 is used in computing the NCM and TFSS measures, their correlations with the OVRL ratings are 0.04 and 0.01, respectively.

Table 2. Correlation coefficients (|*r*|) between the NCM values and the subjective quality ratings of noise-suppressed speech. The number of channels is N=1.

	f_{cut} =10 Hz	20	50	100
OVRL	0.64	0.62	0.61	0.61
SIG	0.55	0.54	0.53	0.52
BAK	0.62	0.60	0.60	0.61

Figure 1 shows the scatter plots of the OVRL ratings against the PESQ, NCM and TFSS values (with the highest correlation in Table 1). It is seen in Fig. 1 (b) that, due to the high SNR level used in quality evaluation experiment (i.e., 5 and 10 dB), the NCM values approach to 1.

Table 2 shows the correlation coefficients when the modulation rate f_{cut} in computing the NCM measure changes from 10 to 100 Hz. It is seen that using high modulation does not improve the correlation between NCM values and subjective quality ratings.

5. Summary and conclusion

Following previous studies of evaluating objective quality measures for speech enhancement [8-9], the present work

further assessed the performance of predicting subjective quality ratings by using envelope or fine-structure based measures (i.e., NCM and TFSS). There two measures, particularly the TFSS measure, carry much information important for objective intelligibility prediction [13]. However, when applied to predict the subjective quality ratings (i.e., OVRL, SIG, and BAK) of noise-suppressed speech in this study, it is seen that they are poorly correlated with subjective quality ratings. The correlation coefficients in quality prediction are much smaller than those obtained with the PESQ measure (see Table 1). This indicates that, although highly correlated with subjective intelligibility scores, the distortion of temporal envelope or fine-structure waveform between the processed and clean signals may carry limited information for predicting the subjective quality ratings of noise-suppressed speech.

Previous studies showed that, with the exception of fwSNRseg and PESQ measures, most objective quality measures poorly predicted the intelligibility of noise-suppressed speech [9]. This study further showed that present intelligibility measures could not well account for the variance of the subjective quality ratings of noise-suppressed speech. Taken together, this study suggests that most of the present objective measures may not well predict *both* the subjective quality ratings and intelligibility scores of noise-suppressed speech. Amongst the all present objective measures examined, the PESQ measure is so far the best choice in terms of predicting subjective quality ratings and intelligibility scores of noise-suppressed speech.

6. Acknowledgements

This research was supported by Faculty Research Fund (Faculty of Education) and Seed Funding for Basic Research, The University of Hong Kong.

7. References

[1] P.C. Loizou, *Speech Enhancement: Theory and Practice,* Taylor & Francis Group, Boca Raton, 2007.

[2] Y. Hu and P.C. Loizou, "A generalized subspace approach for enhancing speech corrupted by colored noise," *IEEE Trans. Speech Audio Process.*, vol. 11, no. 4, pp. 334–341, Jul. 2003.

[3] F. Chen and P.C. Loizou, "Speech enhancement using a frequency-specific composite Wiener function," in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2010, pp. 4726–4729.

[4] 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.

[5] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual valuation 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.

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

[7] S.R. Quackenbush, T.P. Barnwell and M.A. Clements, *Objective Measures of Speech Quality*, Prentice-Hall, Englewood Cliffs, 1988.

[8] Y. Hu and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Speech Audio Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.

[9] J. Ma, Y. Hu, and P. Loizou, "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3387–3405, May 2009.

[10] Z.M. Smith, B. Delgutte, and A.J. Oxenham, "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, vol. 416, no. 6876, pp. 87–90, Mar. 2002.

[11] R. Goldsworthy and J. Greenberg, "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," *J. Acoust. Soc. Am.*, vol. 116, no. 6, pp. 3679–3689, Dec. 2004.

[12] F. Chen and P.C. Loizou, "Predicting the intelligibility of vocoded and wideband Mandarin Chinese," *J. Acoust. Soc. Am.*, vol. 129, no. 5, pp. 3281–3290, May 2011.

[13] F. Chen, L.N. Wong, and Y. Hu, "A Hilbert-fine-structure-derived physical metric for predicting the intelligibility of noise-distorted and noise-suppressed speech," *Speech Commun.*, vol. 55, no. 10, pp. 1011–1020, Dec. 2013.

[14] T.H. Falk, C.X. Zheng, and W.Y. Chan, "A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 7, pp. 1766–1774, Sep. 2010.

[15] Y. Ephraim and D. Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 6, pp. 1109–1121. Dec. 1984

[16] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 33, no. 2, pp. 443–445, Apr. 1985.

[17] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 27, no. 2, pp. 113–120, Apr. 1979.

[18] Y. Hu and P.C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, no. 7–8, pp. 588–601, Jul.–Aug., 2007.

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

[20] D. Greenwood, "A cochlear frequency-position function for several species – 29 years later," *J. Acoust. Soc. Amer.*, vol. 87, no. 6, pp. 2592–2605, Jun. 1990.

[21] ANSI, "Methods for calculation of the speech intelligibility index," American National Standards Institute, New York, S3.5–1997.