Applied Acoustics 145 (2019) 144–148

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

A composite objective measure on subjective evaluation of speech enhancement algorithms

Zhibin Lin *, Lu Zhou, Xiaojun Qiu

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

ARTICLE INFO

Article history:
Received 30 March 2018
Received in revised form 27 August 2018
Accepted 1 October 2018

ABSTRACT

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.

1. 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, the subspace algorithms and machine learning based method [1–5].

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 [6].

Listening tests are usually time-consuming and expensive to conduct [7], 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 [8–11]. 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 [12–18], 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 is only consisted of five different objective measures. Subsequently, the measurement uncertainty of the proposed measure of different speech enhancement algorithms is investigated, and the reliability of the results is discussed.

2. A composite measure

Several existing objective measures have been combined to form a new measure by utilizing the linear regression analysis or nonlinear techniques [19]. 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

* Corresponding author.
E-mail addresses: zblin@nju.edu.cn (Z. Lin), xjqiu@nju.edu.cn (X. Qiu).

https://doi.org/10.1016/j.apacoust.2018.10.002
0003-682X/

(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 [20]. 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 measureO_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 $\bar{S}_d$ and $\bar{O}_d$ 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 [19].

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 5040 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].

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,

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

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

$$C_{\mathrm{OVL}} = -0.835 + 0.610PESQ_{\mathrm{OVL}} - 3.229LLR_{\mathrm{OVL}} + 0.804CEP_{\mathrm{OVL}} + 0.313fwSNRseg_{\mathrm{OVL}} - 0.008fwSNRsegVar_{\mathrm{OVL}}. \tag{5}$$

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 ($\rho = 0.674$) 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 [19].

3. Uncertainty of the proposed measure

3.1. 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) algo-

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	0.58 (0.64)	0.66 (0.58)	0.64 (0.60)	0.67 (0.57)	0.72 (0.55)	0.673 (0.253)
BAK	0.49 (0.50)	0.27 (0.57)	0.23 (0.58)	0.28 (0.58)	0.51 (0.51)	0.609 (0.308)
OVL	0.63 (0.44)	0.62 (0.45)	0.62 (0.50)	0.65 (0.47)	0.71 (0.43)	0.674 (0.298)