EVALUATION OF OBJECTIVE INTELLIGIBILITY PREDICTION MEASURES FOR NOISE-REDUCED SIGNALS IN MANDARIN

Risheng Xia[1], *Junfeng Li*[1], *Masato Akagi*[2], *Yonghong Yan*[1]

[1]Institute of Acoustics, Chinese Academy of Sciences
[2]School of Information Science, Japan Advanced Institute of Science and Technology

ABSTRACT

In this paper, the performance of eight state-of-the-art objective measures is evaluated in terms of predicting speech intelligibility in Mandarin of the processed signals by noise-reduction algorithms. The speech signals were first corrupted by three types of noises at two signal-to-noise ratios and subsequently processed by four classes of noise reduction algorithms, followed by objective intelligibility prediction. The subjective intelligibility ratings were obtained through a set of listening tests. Further investigation was conducted for objective measures in predicting speech intelligibility of noisy signals before and after noise-reduction processing in terms of correlation analysis and prediction errors. The analysis results reported here do provide valuable hints for analyzing and optimizing noise-reduction algorithms for Mandarin.

Index Terms— Mandarin speech intelligibility, Objective intelligibility prediction, Noise reduction.

1. INTRODUCTION

In everyday listening environments, speech signals are often corrupted by various kinds of background noise. In the past several decades, many studies on speech perception in noise demonstrated that speech recognition in noise is much lower than that in quiet [1]. In order to deal with the effects of background noise, a variety of single-channel noise-reduction algorithms have already been reported in the past decades [2]. Their performance was generally examined in terms of speech quality and/or speech intelligibility. The most accurate evaluation approach is through subjective listening tests, however, it is costly and time consuming. Therefore, much effort has been placed on developing objective measures that are able to predict speech quality and/or speech intelligibility as accurate as possible [3, 4]. Among the researches on designing objective measures, most studies were carried out on developing objective speech quality prediction measures, while few work on objective speech intelligibility prediction.

This work is partially supported by the National Natural Science Foundation of China (No. 10574140, 10925419, 90920302, 10874203, 60875014, 61072124, 11074275, 11161140319).

Concerning the researches on objective speech intelligibility prediction, the articulation index (AI) [5] and speech transmission index (STI) [6] are the most commonly used measures for predicting speech intelligibility. By incorporating the factors used in the computation of STI, AI measure was further evolved to speech intelligibility index (SII) [7]. These objective measures were reported to be quite low correlated with the processed speech signal after non-linear noise-reduction processing [2, 8]. In recent years, therefore, increased interests have been focused primarily on objectively predicting speech intelligibility, especially after being processed by non-linear processing [8, 9]. In a recent study, Ma *et al.* suggested a set of new band-importance weighting functions (BIF) to be used in speech intelligibility prediction and found that some objective measures could benefit from the use of BIF. More recently, Taal *et al.* presented a short-time objective intelligibility measure (STOI) based on shorter time segments, which showed high correlation with the intelligibility ratings of noisy and noise-reduced signals [10].

The studies on objective speech intelligibility prediction mentioned above were mainly conducted using western language (e.g., English) speech materials. The field of linguistics, however, suggests that different languages are generally characterized by diverse specific features at the acoustic and phonetic levels due to their distinctive production manner, perceptual mechanism and syllable structure [11]. Mandarin is a tonal language and different tones characterized by F0 contour are used to express the lexical meaning of words. In contrast, F0 contour in English is primarily to emphasize or express emotion and convey intonation. The effects of these differences among different languages on performance of noise-reduction algorithms, in our previous research, were extensively examined in terms of speech intelligibility [12].

Following the previous research [12], in this paper, we focus on investigation of the ability of objective measures to predict Mandarin speech intelligibility of noisy and processed signals by noise-reduction processing. Specifically, eight objective measures are examined through investigating the relationship between the objective prediction scores and the subjective intelligibility ratings, and comparatively analyzing their ability in Mandarin speech intelligibility prediction for the unprocessed noisy and noise-reduced signals.

2. SUBJECTIVE EVALUATIONS

2.1. Subjects

Ten native Mandarin speakers (five females and five males) with normal hearing, aged from 23–31 years old, participated in our experiment. They were paid for their participation.

2.2. Materials

The syllable tables reported by Ma *et al.* was adopted as the speech materials, which has been the national standard (GB/T15508-1995) [13]. This set of test materials consists of 10 syllable tables, each of which contains 75 phoneme-balanced (PB) Mandarin syllables with Consonant-Vowel (CV) structure. In each table every three syllables are combined randomly to form nonsense sentences with the format "The ith sentence is word1, word2, word3". Thus every table can produce enough lists consisting of 25 unmeaning sentences to fulfill general tests. The sentence lists were recorded in a sound-treated booth at a sampling rate of 16 kHz and stored in a 16-bit format, and then down-sampled to 8 kHz before presented to the listeners.

2.3. Signal processing

The clean and noise signals were processed by the IRS filter to simulate the receiving frequency characteristics of telephone handsets. Then the noise signals were added to the clean speech at 0 dB and 5 dB SNRs respectively. We selected three types of background noise: white noise, babble noise and car noise. The noisy signals were enhanced by five representative single-channel noise-reduction algorithms, namely KLT, logMMSE, logMMSE-SPU, MB and Wiener-as, which cover the current four major classes of noise-reduction algorithms. The implementations of these algorithms are derived from [2].

2.4. Procedure

The noisy and enhanced signals were presented to the subjects at a comfortable level through TDH-39 headphone and Madsen Iteral II audio meter in a sound-treated booth. All subjects went through a training procedure to be familiar with the testing environment. In the formal test there were 36 listening conditions, including 3 types of background noises (white noise, babble noise and car noise) $\times$ 2 SNR levels (0 dB and 5 dB) $\times$ 6 enhancement types (noisy reference and five noise-reduction algorithms). Every subject would listen to 900 (25 $\times$ 36) unmeaning short sentences. All the listening conditions were divided into three sessions according to the background noise type, and in each listening session the sentences were presented to the subjects in a random order. Listeners were asked to write down the key words that she or he heard in every sentence as many as possible.

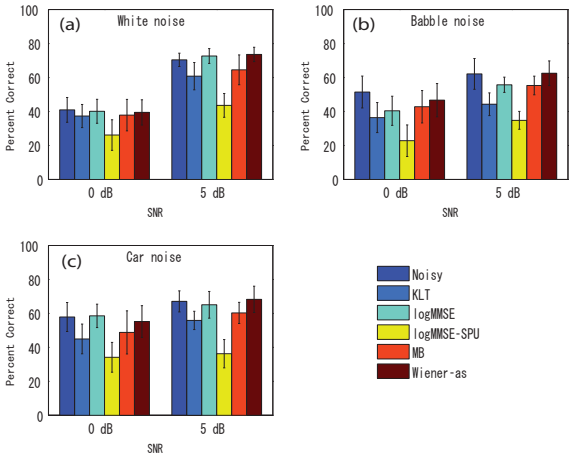

Fig. 1. Mean recognition scores for five noise-reduction algorithms under three types of background noises with two SNR levels.

2.5. Results

Figure 1 shows the mean recognition scores averaged across ten subjects for five noise-reduction algorithms under three background noises at two SNR levels. From this figure, it is clear that at the most cases the Mandarin speech intelligibility was decreased by the noise-reduction processing compared with the unprocessed speech. The negative effects of noise reduction on Mandarin speech intelligibility was ascertained. Especially for logMMSE-SPU, at various listening conditions it consistently yielded a severe damage for Mandarin speech intelligibility. Only the Wiener-as algorithm maintained the intelligibility to a large extent and even provided a slight improvement under white noises at 5 dB. In terms of the overall performance, logMMSE algorithm ranked the second since its performance was comparable to the unprocessed speech in white and car noise conditions.

3. OBJECTIVE MEASURES

Based on the previous researches [9, 10, 14, 15, 16], in this paper, two major classes of objective intelligibility prediction measures (i.e., SNR-based and correlation-based) were examined. Specifically, the SNR-based measure was the frequency-weighted segmental SNR (fwSNRseg) [9]; the correlation-based measures included the coherence-based measure (COH) [9], the short-time objective intelligibility measure (STOI) [10], the coherence SII (CSII) [14], the middle level CSII (CSII_{m}) [14], the objective intelligibility measure (I3) [14], the normalized covariance metric (NCM) [15] and the normalized subband envelope correlation (NSEC) [16]. All these objective measures are a function of the clean signal and the unprocessed/processed signal. The definitions and detail implementations of these objective measures are given in the corresponding references.

4. ANALYSIS AND RESULTS

In this section, two examinations were performed to assess the ability of the objective measures mentioned above for Mandarin speech intelligibility prediction. The first analysis was to show the overall ability of the objective measures in predicting speech intelligibility averaged across all tested conditions, and the second one was to further demonstrate their ability in Mandarin speech intelligibility prediction before and after non-linear noise-reduction processing.

4.1. Analysis of objective prediction scores and subjective intelligibility ratings

To examine the overall performance of the objective measures in speech intelligibility prediction, two figures of merit were used [9]. The first figure of merit was Pearson's correlation coefficient, ρ, between the objectively predicted scores and the subjective intelligibility scores, and the second figure of merit was an estimate of the standard deviation of the error computed as $\sigma_e = \sigma_d\sqrt{(1-\rho^2)}$, where σ_d is the standard deviation of the speech recognition scores in a given condition and σ_e is the computed standard deviation of the error. The higher ρ indicates that the objective measure is better at predicting speech intelligibility, while for σ_e, the lower values represent the better results.

The analysis results of the eight objective intelligibility prediction measures in terms of the correlation coefficient (ρ) and the standard deviation of prediction error (σ_e), averaged across all tested conditions, are shown in Table 1. From Table 1, it is noted that of the eight objective intelligibility prediction measures, the STOI measure yielded the highest correlation ($\rho = 0.90$), corresponding the highest ability in predicting the subjective intelligibility ratings, and the lowest standard deviation of the error ($\sigma_e = 5.84\%$) in predicting Mandarin speech intelligibility. It is followed by the NCM measure ($\rho = 0.82, \sigma_e = 7.65\%$) and the NSEC measure ($\rho = 0.81, \sigma_e = 7.93\%$). The lowest performance in predicting Mandarin speech intelligibility was given by the COH measure with the low correlation ($\rho = 0.49$) and the high standard deviation of the error ($\sigma_e = 11.75\%$). The other objective measures fell between the two extremes of the correlation coefficient(ρ) and the standard deviation of the error (σ_e).

In comparison with the results in [9, 10] where western language speech materials were used, the STOI measure demonstrated the best ability in predicting Mandarin speech intelligibility, which was in line with the result reported in [10]. In addition, some inconsistent findings were observed. For example, the study in [9] demonstrated that the I3 measure provided the quite high ability in predicting English speech intelligibility, however, it was less effective for Mandarin speech intelligibility prediction as shown Table 1. The factors resulting in these differences might come from the differences in languages to a certain extent.

4.2. Analysis of objective intelligibility prediction before and after noise-reduction processing

Generally, only certain monotonic relationship is present of the intelligibility scores before and after non-linear noise-reduction processing [10]. To further demonstrate the ability of the objective measures in speech intelligibility prediction before and after noise-reduction processing, therefore, a mapping was performed to account for the non-linear relationship between the objective and subjective scores. To do this, a logistic function $f(d) = 100/(1 + \exp(ad + b))$ was used as in [10], where a and b are the parameters that were tuned with a nonlinear least square procedure, and d denotes the objective score. This logistic function was only fitted to the unprocessed conditions, which was then used to predict the intelligibility scores for the noise-reduced conditions. The performance of all objective measures was evaluated with the root mean square (RMS) of the prediction error (RMSE), defined as $\sigma = \sqrt{\frac{1}{S}\sum_i (s_i - f(d_i))^2}$, where s_i refers to an intelligibility score obtained in the processing condition i and S denotes the total number of processing conditions.

The scatter plots of subjective ratings against the objective measures are shown in Fig. 2, along with the fitting curves and the RMSE results. Fig. 2 demonstrates that the STOI measure again provided the lowest RMSE ($\sigma = 6.97\%$), corresponding to the highest ability in predicting the effect on Mandarin speech intelligibility of the noise-reduction algorithms. The highest RMSE ($\sigma = 18.95\%$) was introduced by the fwSNRseg measure. More importantly, it is noted that most of objective measures overestimated the speech intelligibility for the noise-reduced signals, and that of the tested objective measures, only the STOI measure yielded the accurate intelligibility prediction for the signals after being processed by the noise-reduction algorithms.

Consistent with the previous results [10], the STOI measure showed the best ability for predicting the effect on Mandarin speech intelligibility of non-linear noise-reduction processing, that is, no significant improvement of speech intelligibility can be achieved by noise-reduction processing.

5. CONCLUSIONS

In this paper, eight objective measures were evaluated for predicting the speech intelligibility before and after non-linear noise-reduction processing. Of all tested objective measures, the STOI measure provided the highest abilities in predicting Mandarin speech intelligibility in all conditions and in predicting the effect on speech intelligibility due to non-linear noise-reduction processing. This indicates that the STOI measure is quite promising for analyzing and/or optimizing noise-reduction algorithms. The remaining RMSEs of STOI for Mandarin imply that there is still a large room to improve, which possibly can be done by integrating the language-specific (e.g., F0 contour) cues in its calculation.

Table 1. The Pearson's correlation coefficients ρ and the standard deviations of the error σ_e, averaged across all signals under three noise conditions at two SNRs, for eight objective intelligibility prediction measures.

	COH	CSII	CSIIm	fwSNRseg	I3	NCM	NSEC	STOI
ρ	0.49	0.67	0.81	0.65	0.79	0.82	0.81	0.90
σ_e	11.75%	10.04%	7.20%	10.24%	8.42%	7.65%	7.93%	5.84%

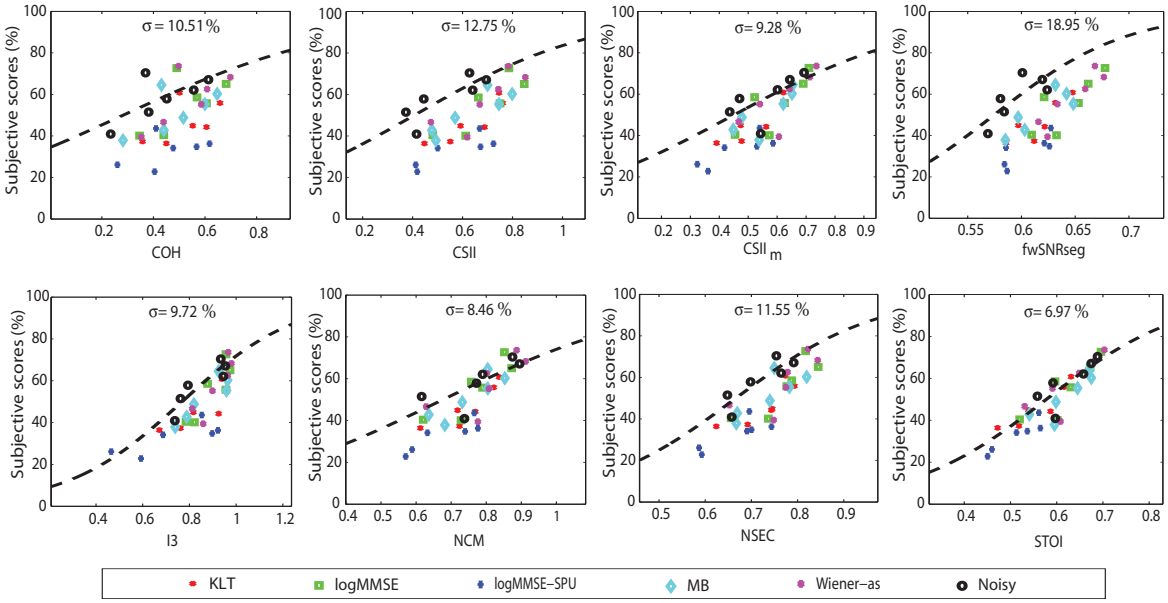

Fig. 2. Scatter plots of the subjective intelligibility ratings against the objectively predicted scores, along with the mapping results (dashed curves) and the RMSE results (σ).

6. REFERENCES

[1] S. Gelfand, "Consonant recognition in quiet and in noise with aging among normal hearing listeners," *J. Acoust. Soc. Am.*, 80(6), pp. 1589-1598, 1986.

[2] P.C. Loizou, Speech Enhancement: Theory and Practice (CRC Press, Taylor Francis Group, Florida), pp. 97-394, 2007.

[3] Y. Hu and P. Loizou, "A comparative intelligibility study of single-microphone noise reduction algorithms," *J. Acoust. Soc. Am.*, 122(3), pp. 1777-1786, 2007.

[4] Y. Hu and P. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 1, pp. 229-238, 2008.

[5] K.D. Kryter, "Validation of the articulation index," *J. Acoust. Soc. Am.*, 34(11), pp.1698-1706, 1962.

[6] T. Houtgast and H. Steeneken, "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.*, 77(3), pp.1069-1077, 1985.

[7] ANSI S3.5-1997, "Methods for Calculation of the Speech Intelligibility Index," (American National Standards Institute, New York), 1997

[8] W.M. Liu, K.A. Jellyman, N.W.D. Evans and J.S.D. Mason, "Assessment of objective quality measures for speech intelligibility estimation," in Proc. *ICASSP*, pp. 1225-1228, 2006.

[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.*, 125(5), pp. 3387-3405, 2009.

[10] C. Taal, R. Hendriks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Audio, Speech and Language Processing*, pp. 2125-2136, September, 2011.

[11] R. Trask, Key Concepts in Language and Lingustics (Routledge, London), pp. 15-30, 1998.

[12] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi and P. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese and English," *J. Acoust. Soc. Am.*, 129(5), pp. 3291-3301, 2011.

[13] D. Ma and H. Shen, Acoustic Manual (Chinese Science Publisher, Beijing), Chap. 20, 2004.

[14] J. Kates and K. Arehart, "Coherence and the speech intelligibility index," *J. Acoust. Soc. Am.*, 117(4), pp. 2224-2237, 2005.

[15] I. Hollube and K. Kollmeier, "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," *J. Acoutic. Soc. Am.*, vol. 100(3), pp. 1703-1715, 1996.

[16] J.B. Boldt and D.P.W. Ellis, "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in Proc. *EUSIPCO*, pp. 1849-1853, 2009.