Evaluation of Objective Intelligibility Prediction Measures for Speech Enhancement in Mandarin

Junfeng Li*, Dongwen Ying*, Yonghong Yan* and Masato Akagi†
* Institute of Acoustics, Chinese Academy of Sciences
† School of Information Science, Japan Advanced Institute of Science and Technology

***Abstract*—In this paper, we evaluate the performance of several state-of-the-art objective measures in terms of predicting speech intelligibility in Mandarin of the processed noisy signals by speech enhancement algorithms. The speech signals were first corrupted by three types of noises at two signal-to-noise ratios, followed by four classes of speech enhancement algorithms. The objective intelligibility prediction measures were then performed. The subjective intelligibility ratings were obtained by performing a comprehensive investigation of the unprocessed noisy signals and the processed signals by various single-channel noise reduction algorithms through listening tests. Based on the subjective intelligibility scores, in this paper, we focus on examining the capability of objective intelligibility prediction measures using the correlation analysis and the standard deviation of error. The analysis results reported in this paper do provide valuable hints for analyzing and optimizing noise-reduction algorithms.**

I. Introduction

In everyday listening environments, speech signals are often corrupted by various kinds of background noise. In 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 and facilitate speech recognition in noise, a variety of single-channel noise-reduction algorithms have already been reported in the past decades [2]. Performance of these noise-reduction algorithms are generally examined in terms of speech quality and/or speech intelligibility. The most accurate evaluation approach for speech quality/intelligibility is through subjective listening tests. Although the subjective evaluation is normally accurate, it is costly and time consuming. Therefore, much effort has been placed on developing objective measures that were designed to be able to predict speech quality and/or speech intelligibility as accurate as possible [3], [4]. Among the researches in objective evaluations, most works were carried out in developing objective speech quality measures for the signal in noisy conditions, and few work on objective speech intelligibility prediction.

Concerning the researches on objective speech intelligibility prediction conducted in the past decades, the articulation index (AI) [5] and speech transmission index (STI) [6] were the most commonly used for predicting speech intelligibility in noisy and reverberant environments. By incorporating the factors used in the computation of STI, such as, relative importance of various frequencies to speech intelligibility and spread of masking, AI measure has been further evolved to speech intelligibility index (SII) [7]. Though these objective measures were reported to correlate well with the subjective intelligibility ratings by human but are rather limited to linear systems, and quite low correlated with the processed speech signal after non-linear processing, such as speech enhancement. In recent years, therefore, increased interests have been focused primarily on objectively predicting speech intelligibility of signals, especially processed by speech enhancement (non-linear processing) algorithms [8], [9], [10].

Liu *et al.* assessed the capability of a variety of objective speech quality measures in predicting speech intelligibility in the context of additive noises as well as degradations introduced by speech enhancement algorithms [8]. Their evaluation results showed that most quality measures correlated poorly with speech intelligibility especially when non-linear speech enhancement algorithms were concerned/invovled. Ma *et al.* comprehensively examined the performance of traditional as well as newly presented objective measures in terms of predicting speech intelligibility in realistic noisy conditions [9]. They showed that most of objective measures provide poor ability in predicting speech intelligibility, among others, the modified CSII and NCM measures were found to perform the best for speech intelligibility prediction in noise. Taal *et al.* evaluated five objective intelligibility measures in terms of predicting the difference in intelligibility before and after two single-channel noise reduction processing. It was found that only the short-time objective intelligibility measure (STOI) that the authors proposed gave relatively good intelligibility prediction performance, and other objective measures overestimated the intelligibility scores [10].

The studies on objective speech intelligibility prediction mentioned above were mainly performed 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 and syntax structure [11]. Mandarin is a tonal language and differs from the non-tonal language (e.g., English) in that different tones are used to express the lexical meaning of words. The four lexical tones in Mandarin are characterized by their patterns in fundamental frequency (F0) variation during voiced segments of speech. In contrast, the F0 contour in English is used primarily to emphasize or express emotion and convey intonation, among others, and thus contributes little to speech intelligibility. This difference of different languages, in our previous research,

were extensively examined in terms of intelligibility of speech signals processed by various of single-channel speech enhancement algorithms [12], [13].

Following our previous research [12], [13], in this paper, we focus on investigation of the capability of objective measures in predicting Mandarin Chinese intelligibility after single-channel noise reduction processing. Specifically, we first report on the subjective evaluation of four major classes of single-channel noise-reduction algorithms, including subspace-based, statistical-model-based, spectral subtractive, and Wiener-type algorithms, in terms of Mandarin speech intelligibility under three different types of noise at two signal-to-noise ratio (SNR) levels. Subsequently, nine objective intelligibility prediction measures to be tested in this paper are reviewed. Finally, the main focus is paid to the correlation analysis between subjective and objective evaluation results, followed by the general discussions.

II. Subjective Evaluations

A. 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 participations.

B. Materials

The syllable tables for articulation test reported by Ma *et al.* was adopted as the speech materials, which has been the national standard (GB/T15508-1995) [14]. This set of test materials consists of 10 syllable tables, each of which contains 75 PB Mandarin syllables with CV(Consonant-Vowel) structure. In each table every three syllables are combined randomly to form unmeaning sentences with the format "The ith sentence is word1, word2, word3". Thus every table can produce enough lists consisting of 25 unmeaning sentences to fulfil 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.

C. 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 one-channel speech enhancement algorithms, namely KLT, logMMSE, logMMSE-SPU, MB and Wiener-as, which cover the current four major classes of noise reduction. The implementations of these algorithms are derived from [2].

D. 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.

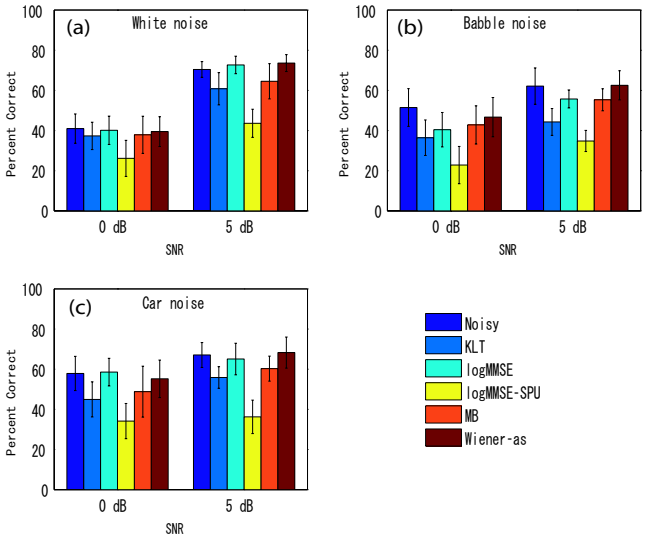

Fig. 1. *Mean correct syllable scores for five enhancement algorithms under three types of background noises with two SNR levels.*

E. Results

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

III. Objective Intelligibility Prediction Measures

Based on the researches on objective intelligibility prediction measures that previously reported in [9], [10], in this paper, in total eight different objective intelligibility measures are included in our evaluations. These are all a function of the

clean and the unprocessed/processed signal. All the objective intelligibility measures to be involved in the evaluations are briefly reviewed in what follows.

Coherence SII(CSII), which used the SII index as the base measure and replaced the SNR term with the signal-to-distortion ratio term, was computed using the coherence between the input and output signals [15]. Though an earlier study reported the CSII predicts an incorrect intelligibility improvement due to noise reduction in English [9], no examination has been conducted with various noise-reduction algorithms in Mandarin Chinese. Moreover, the mid-level CSII, CSII$_m$, which ranges from the overall rms level to 10 dB below, is also used for comparison [15].

The normalized covariance metric (NCM) was first presented by Hollube *et al.* [16]. Based on the covariance between the input and output envelope signals, the NCM was computed as a weighted sum of transmission index (TI) determined from the envelops of the input and output signals in each frequency band. It was reported that the NCM measure provided the high correlation for predicting sentence recognition [16].

The coherence-based measure (COH) is computed by dividing the clean and processed signals in a number of overlapping windowed segments, computing the cross power spectrum for each segment using the FFT, and then averaging across all segments. In our study, the MSC values averaged across all frequency bins was used as the objective measure. The COH objective measure demonstrated the relatively high correlation with the subjective recognition scores [9].

The frequency-weighted segmental SNR (fwSNRseg) was computed as first multiplying the spectra with overlapping Gaussian-shaped windows and summing up the weighted spectra within each band, the derived spectra are used for segmental SNR calculation in each frames, which is subsequently weighted by a weighting function that is considered as the magnitude spectrum of the clean signal raised to a power [9]. Of the conventional subjective quality measures, the fwSNRseg measure performed modestly well in terms of predicting both quality and intelligibility [9], [10].

The objective intelligibility measure, $I3$, was defined as a combination of the three CSII values (in the low, middle and high regions) followed by a logistic function transformation was subsequently used to model the intelligibility [9]. The $I3$ measure produced the highest correlation for consonants and sentence materials [9].

Based on the similarity between the time-varying spectral envelopes of target speech and system output, Boldt *et al.* presented the normalized subband envelope correlation (NSEC) which is defined as the correlation between the time-frequency representations of the target (reference) and the output of the time-frequency processing (e.g., noise-reduction) algorithm, after being normalized with Frobenius norm of the energy envelopes of the target signal and the system output [17].

A short-time objective intelligibility measure (STOI) is presented based on the correlation between temporal envelopes of the clean and degraded speech in short-time segments [10], [19]. Experimental results show that STOI is beneficial to take segment lengths of this order into account, and that STOI has high correlation with the speech intelligibility for three different listening tests.

IV. Results and discussions

Two figures of merit were used to assess the performance of the above objective measures in terms of predicting speech intelligibility. The first figure of merit was Pearson's correlation coefficient, r, and the second figure of merit was an estimate of the standard deviation of the error computed as $\sigma_e = \sigma_d\sqrt{(1-r^2)}$, 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. A higher r indicates that the objective measure is better at predicting speech intelligibility, while for σ_e, lower values represent better results.

The scatter plots of subjective rating scores against the examined objective measures are shown in Fig. 2, where their corresponding figures of merit are indicated at the top of each plot. Fig. 2 shows that objective measures do not directly predict an absolute intelligibility score but instead some monotonic relationship is present between the objective scores and the subjective recognition results from the listening experiments. More specifically, Fig. 2 indicates that of the eight objective measures tested, the STOI measure yielded the highest correlation ($r = 0.90$) and the lowest standard deviation ($\sigma_e = 5.84$) in terms of predicting speech intelligibility, followed by the NCM measure ($r = 0.82, \sigma_e = 7.65$) and the NSEC measure ($r = 0.81, \sigma_e = 7.93$). The lowest performance in objective intelligibility prediction was given by the COH measure characterized by the low correlation and the high standard deviation ($r = 0.49, \sigma_e = 11.75$).

In comparison with the results reported in previous studies [9], [10], [18] in which western language corpus were normally used. In line with the result in [10], of several tested objective measures, the STOI measure demonstrated the best ability in predicting speech intelligibility even after non-linear noise-reduction processing. The study in [9] demonstrated that the I3 measure provided the quite high ability in predicting speech intelligibility. In contrast, the results in our investigations showed that the I3 measure is less effective in speech intelligibility prediction, which is characterized by the relatively low correlation ($r = 0.79$) and the high deviation ($\sigma_e = 8.42$). Moreover, the NCM measure in our evaluations showed the very good speech intelligibility prediction ability, while it did not in the previous study [9]. The factors resulting in these differences are unknown and worthy to be further studied in the future.

V. Conclusions

In this paper, eight objective measures are evaluated in order to predict the speech intelligibility after non-linear noise-reduction processing. Five typical single-channel noise-reduction algorithms are considered and applied to babble and car noises at 0 and 5 dBs. Of all measures, the STOI measure provides highest correlation between objective prediction and

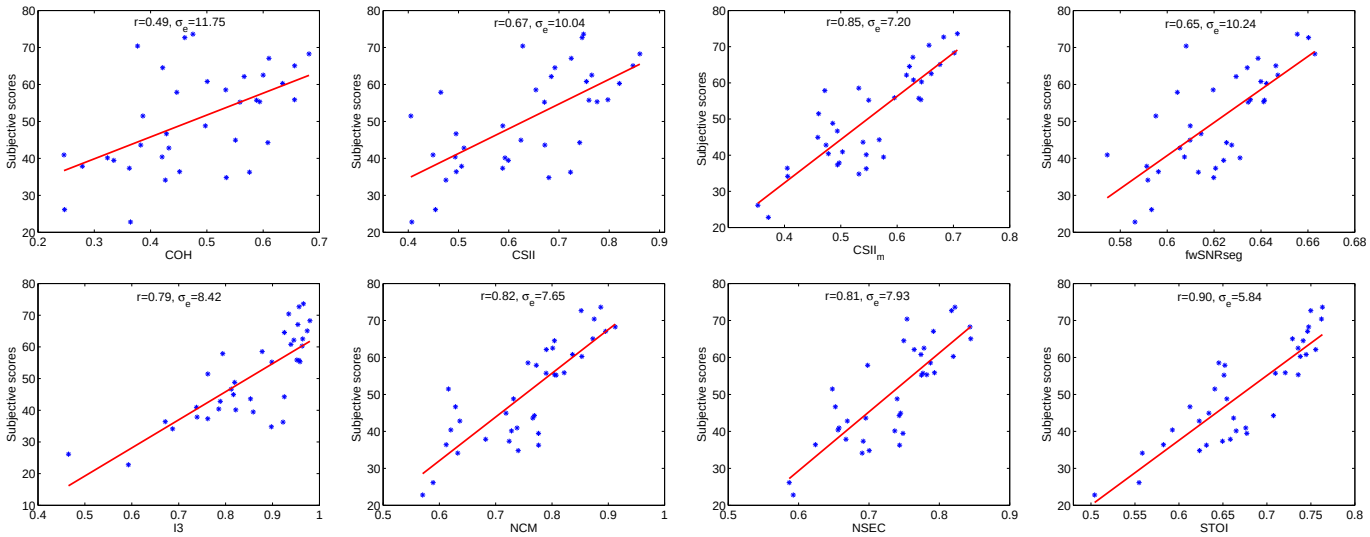

Fig. 2. Scatter plots of subjective scores against the tested objective measures in six conditions (unprocessed and processed by five single-channel noise-reduction algorithms), along with their linear regression results (solid line). The objective measures examined include CSII, COH, CSII$_w$, fwSNRseg, I3, NCM, NSEC and STOI. At the top of each plot, the correlation coefficient (r) and the standard deviation of the prediction error (σ_e) is denoted.

subjective listening scores, which corresponds to the highest ability in predicting speech intelligibility. This makes STOI a potential candidate for analysis and/or optimization of single-channel noise-reduction algorithms. The difference in speech intelligibility prediction ability between English(in [9]) and Mandarin (in this paper) is worthy to be further studied. The possible integration of language-specific cues in the calculation of objective speech intelligibility prediction is also quite promising to focus on in the near future.

ACKNOWLEDGMENT

This work is partially supported by the National Natural Science Foundation of China (No. 10574140, 10925419, 90920302, 10874203, 60875014, 61072124, 11074275), and the China-Japan-Korea A3 Foresight Program (11161140319).

REFERENCES

[1] S. Gelfand, "Consonant recognition in quiet and in noise with aging among normal hearing listeners," *J. Acoust. Soc. Am.*, 80, 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.*, vol. 34. 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.*, vol. 77, pp.1069-1077, 1985.

[7] Methods for Calculation of the Speech Intelligibility Index (American National Standards Institute, New York).

[8] W.M. Liu, K.A. Jellyman, N.W.D. Evans and J.S.D. Mason, "Assessment of objective quality measures for speech intelligibility," in Proc. *ICASSP2006*, 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, "Intelligibility prediction of single-channel noise-reduced speech," *ITF-Fachtagung Sprachkommunikation*, Oct. 2010.

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

[12] J. Li, C.D. Thauh, M. Akagi, L. Yang, J. Zhang and Y. Yan, "Intelligibility investigation of single-channel noise reduction algorithms for Chinese and Japanese," in Proc. *ISCSLP*, pp. 3-7, 2010.

[13] 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.

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

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

[16] 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, pp. 1703-1715, 1996.

[17] 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.

[18] C. Taal, R. Hendriks, R. Heusdens and J. Jensen, "On predicting the difference in intelligibility before and after single-channel noise reduction," *IWAENC*, September, 2010.

[19] 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*. (In Press)