Comparative investigation of objective speech intelligibility prediction measures for noise-reduced signals in Mandarin and Japanese

Junfeng Li[1], *Fei Chen*[2], *Masato Akagi*[3], *Yonghong Yan*[1]

[1] Institute of Acoustics, Chinese Academy of Sciences
[2] Division of Speech and Hearing Sciences, The University of Hong Kong
[3] School of Information Science, Japan Advanced Institute of Science and Technology

Abstract

In this paper, eight state-of-the-art objective speech intelligibility prediction measures are comparatively investigated for noisy signals before and after noise-reduction processing between Mandarin and Japanese. Clean speech signals (Chinese words and Japanese words) were first corrupted by three types of noise at two signal-to-noise ratios and then processed by normal-hearing listeners for recognition, whose intelligibility was subsequently predicted by objective measures. Further investigations were conducted for objective measures in predicting speech intelligibility of noise-reduced signals between subjective evaluation scores and objective prediction results, and of noisy signals before and after noise-reduction processing, in terms of correlation analysis and prediction errors. Results showed that the majority of objective measures behave differently for Mandarin and Japanese in predicting the subjective ratings, and the STOI measure consistently provided the best ability in predicting the effect on speech intelligibility of the noise-reduction processing for both Mandarin and Japanese.

Index Terms: Speech intelligibility, objective intelligibility prediction, noise reduction

1. Introduction

Good objective estimators of speech intelligibility are of great use in designing speech signal processing algorithms for speech communication and hearing aids. In contrast with subjective tests, objective measurements of speech intelligibility are less expensive and time-consuming, give more consistent results and are not influenced to subjects' biases. In the past five decades, objective speech intelligibility prediction measures has received considerable attention [1]. Among the objective speech intelligibiilty prediction measures, the articulation index (AI) [2] and speech transmission index (STI) [3] 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) [4]. These objective measures were reported to be quite lowly correlated with the processed speech signal after non-linear noise-reduction processing [5]. In recent years, therefore, increased interests have been focused primarily on objectively predicting speech intelligibility, especially after being processed by non-linear processing [1, 5]. 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 a high correlation with the intelligibility ratings of noisy and noise-reduced signals [6].

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 [7]. Compared with English, for example, Chinese and Japanese contain much fewer vowels, which results in more severe phoneme and syllable confusions for Chinese and Japanese than for English in noise. Furthermore, the tone information in Chinese and the accent information in Japanese are used to distinguish word meaning and thus contribute a great deal to Chinese and Japanese speech intelligibility. In contrast, F0 contour in English is primarily used 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 [8].

Following the previous research [8], in this paper, we focus on comparative investigation of state-of-the-art objective measures in predicting speech intelligibility for Mandarin and Japanese. 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 speech intelligibility prediction for the unprocessed noisy and noise-reduced signals. Additional examination is given to the comparison of these objective speech intelligibility measures for Mandarin and Japanese.

2. Subjective evaluations and results

In this section, we report a summary of the results taken from [8] on the intelligibility evaluation of single-channel noise-reduction algorithms for Mandarin and Japanese. In the evaluation of single-channel noise-reduction algorithms for Mandarin, the syllable tables developed by Ma *et al.* was adopted as the speech materials that consist of 10 syllable tables, each of which produces enough lists consisting of 25 unmeaning sentences to fulfill general tests [9]. In the evaluation of single-channel noise-reduction algorithms for Japanese, the words taken from the familiarity-controlled word lists 2003 (FW03) with the lowest familiarity were used as speech material, which consisted of 20 lists with 50 phonetically-balanced words per list [10]. All these speech signals were down-sampled to 8 kHz before being presented to the listeners and used as test material. The masker signals were white, babble and car noises. The corrupted (at 0 and 5 dB SNRs) and processed sentences were presented to native Mandarin and

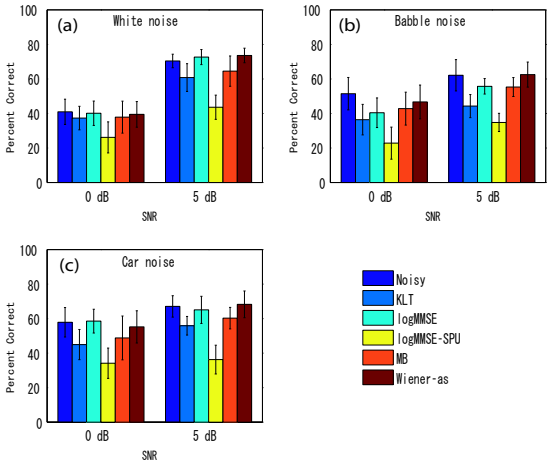

Figure 1: Mean recognition scores for five noise-reduction algorithms under three types of background noises with two SNRs for Mandarin.

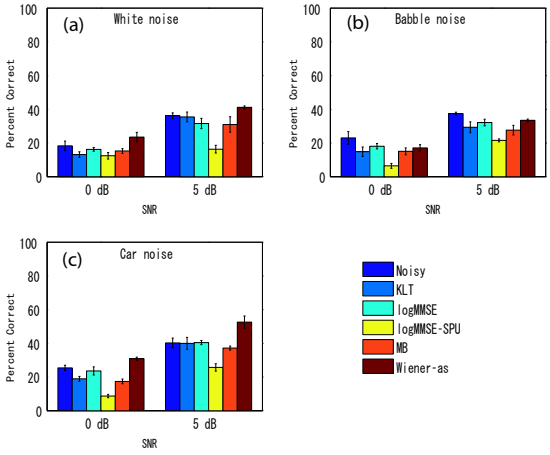

Figure 2: Mean recognition scores for five noise-reduction algorithms under three types of background noises with two SNRs for Japanese.

Japanese speakers for word identification. The mean percentage of words identified correctly [as reported in [8]] is shown in Fig. 1 and Fig. 2 for Mandarin and Japanese, respectively.

Figs. 1 and 2 show the mean recognition scores for five noise-reduction algorithms under three background noises at two SNRs. From the results, it is clear that, at the most cases, the speech intelligibility was decreased by the noise-reduction processing compared with that of the unprocessed speech for both Mandarin and Japanese. The negative effects of noise reduction on speech intelligibility was ascertained. Especially for the logMMSE-SPU algorithm, at various listening conditions it consistently yielded a severe damage for 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 SNR. In terms of the overall performance, the 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 [1, 6, 11, 12, 13], 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) [1]; while the correlation-based measures included the coherence-based measure (COH) [1], the short-time objective intelligibility measure (STOI) [6], the coherence SII (CSII) [11], the middle level CSII (CSII_m) [11], the objective intelligibility measure (I3) [11], the normalized covariance metric (NCM) [12] and the normalized subband envelope correlation (NSEC) [13]. 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 abilities of the objective measures mentioned above in predicting speech intelligibility for both Mandarin and Japanese. The first analysis was to show the overall abilities of the objective measures in predicting speech intelligibility averaged across all tested conditions, and the second one was to further demonstrate their intelligibility prediction abilities before and after non-linear noise-reduction processing for both Mandarin and Japanese.

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 [1]. 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) for Mandarin and Japanese are shown in Table 1. From Table 1, it is noted that most of eight objective measures exhibited different abilities in predicting speech intelligibility for Mandarin and Japanese. For Mandarin, the STOI measure yielded the highest correlation ($\rho = 0.9$), corresponding to 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. In contrast, the best prediction measure for Japanese was found as the CSII_m measure with ($\rho = 0.79$, $\sigma_e = 4.71\%$). It is followed by the NCM measure ($\rho = 0.82, \sigma_e = 7.65\%$) and the NSEC measure ($\rho = 0.81, \sigma_e = 7.93\%$) for Mandarin, and by the NCM measure ($\rho = 0.78, \sigma_e = 5.09\%$) and the I3 measure ($\rho = 0.77, \sigma_e = 5.33\%$) for Japanese. The COH measure consistently provided the lowest performance in predicting speech intelligibility for Mandarin ($\rho = 0.49$, $\sigma_e = 11.75\%$) and for Japanese ($\rho = 0.65$, $\sigma_e = 7.97\%$). The other objective measures fell between the

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
Mandarin	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%
Japanese	ρ	0.65	0.74	0.79	0.72	0.77	0.78	0.76	0.72
	σ_e	7.97%	7.50%	4.71%	11.08%	5.33%	5.09%	8.04%	9.26%

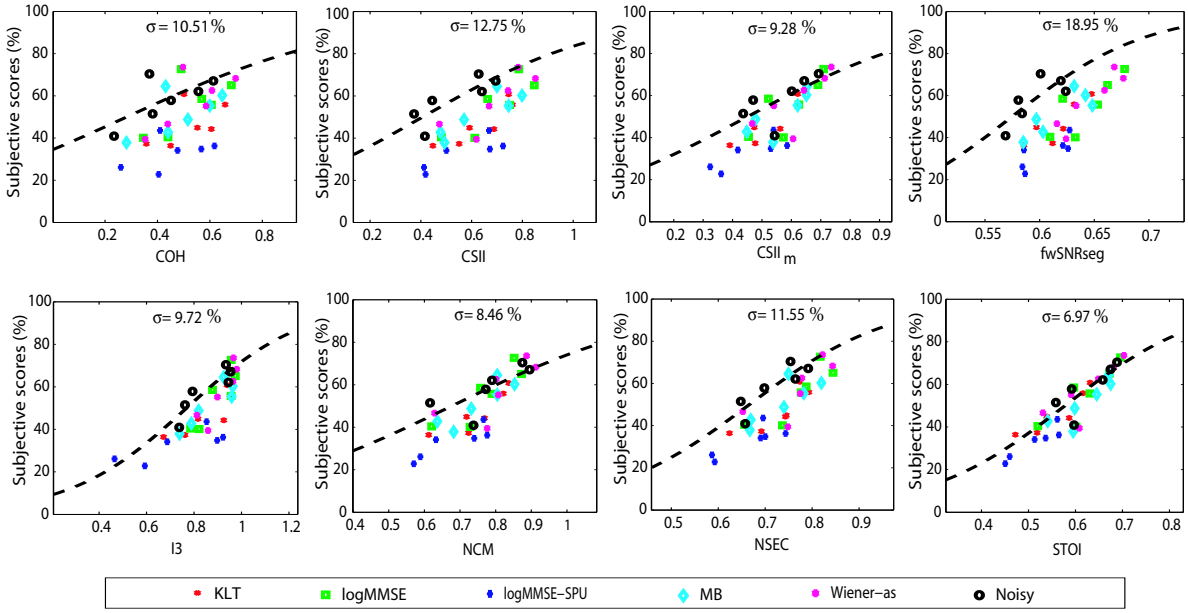

Figure 3: Scatter plots of the subjective intelligibility ratings against the objectively predicted scores for Mandarin, along with the mapping results (dashed curves) and the RMSE results (σ).

two extremes of the correlation coefficient (ρ) and the standard deviation of the error (σ_e).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 [6]. 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 [6], 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 for Mandarin and Japanese against the objective measures are shown in Figs. 3 and 4, respectively, along with the fitting curves and the RMSE results. Figs. 3 and 4 demonstrate that the STOI measure consistently provided the lowest RMSEs for Mandarin ($\sigma = 6.97\%$) and for Japanese ($\sigma = 8.02\%$), corresponding to the highest ability in prediction the effect of the noise-reduction algorithms on speech intelligibility for both Mandarin and Japanese. This is followed by the NCM measure ($\sigma = 8.46\%$ for Mandarin and $\sigma = 8.95\%$ for Japanese). The highest RMSEs were introduced by the fwSNRseg measure, that is, $\sigma = 18.95\%$ for Mandarin and $\sigma = 12.47\%$ for Japanese. 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, the STOI measure yielded the much more accurate intelligibility prediction for the signals after being processed by the noise-reduction algorithms.

Consistent with the results [6, 8], the STOI measure showed the best ability for predicting the effect on Mandarin and Japanese speech intelligibility of non-linear noise-reduction processing, that is, no significant improvement of speech intelligibility can be achieved by noise-reduction processing.

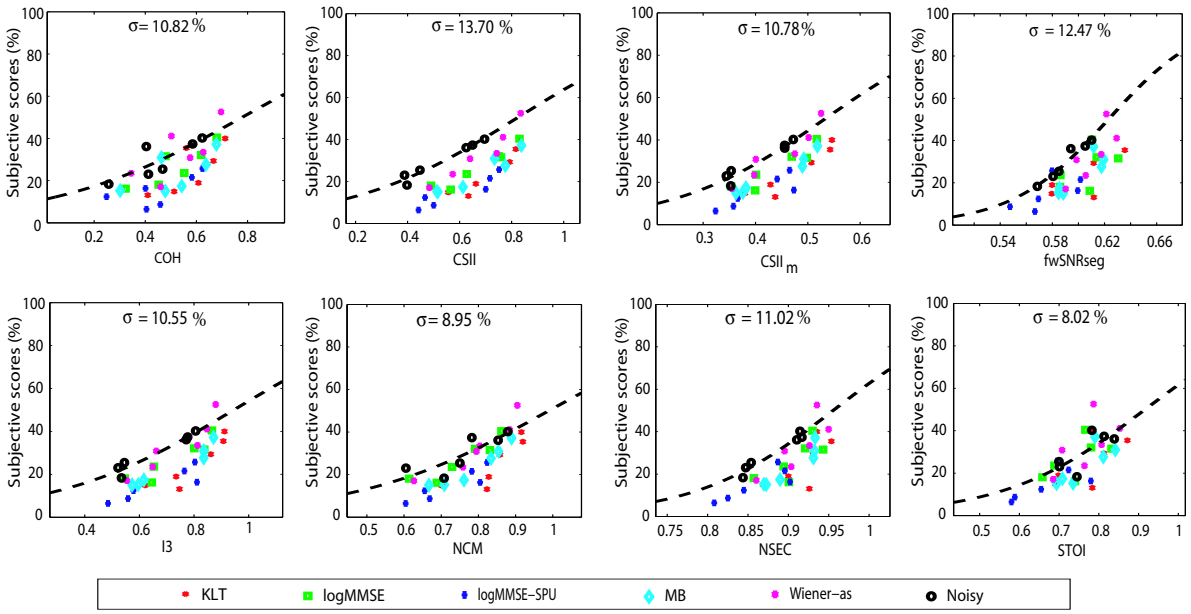

Figure 4: Scatter plots of the subjective intelligibility ratings against the objectively predicted scores for Japanese, along with the mapping results (dashed curves) and the RMSE results (σ).

5. Conclusions

In this paper, eight objective measures were comparatively evaluated in predicting speech intelligibility for Mandarin and Japanese before and after non-linear noise-reduction processing. Evaluation results showed that different objective measures exhibited different abilities in predicting the subjective ratings for Mandarin and Japanese in the tested conditions. In contrast, the STOI measure consistently provided the highest abilities in predicting the effect on speech intelligibility of the non-linear noise-reduction processing for both Mandarin and Japanese. 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 cues in its calculation.

6. Acknowledgements

This work is partially supported by the National 973 Program (2013CB329302), the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500) and the National 863 Program (No. 2012AA012503).

7. References

[1] 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.

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

[3] 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.

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

[5] W.M. Liu, *et al.*, "Assessment of objective quality measures for speech intelligibility estimation," in Proc. *ICASSP*, pp. 1225-1228, 2006.

[6] C. Taal, *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. ASLP*, pp. 2125-2136, September, 2011.

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

[8] J. Li, *et al.*, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese and English," *J. Acoust. Soc. Am.*, 129(5), pp. 3291-3301, 2011.

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

[10] S. Amano, *et al.*, "Development of familiarity-controlled word list 2003 (FW03) to assess spoken-word intelligibility in Japansese," *Speech Comm.*, 51, pp. 76-82, 2009.

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

[12] 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.

[13] 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.