Investigation of objective measures for intelligibility prediction of noise-reduced speech for Chinese, Japanese, and English

Junfeng Li,[a] Risheng Xia, Dongwen Ying, and Yonghong Yan
Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China

Masato Akagi
School of Information Science, Japan Advanced Institute of Science and Technology, Ishikawa, 923-1292, Japan

(Received 16 June 2014; revised 16 October 2014; accepted 23 October 2014)

Many objective measures have been reported to predict speech intelligibility in noise, most of which were designed and evaluated with English speech corpora. Given the different perceptual cues used by native listeners of different languages, examining whether there is any language effect when the same objective measure is used to predict speech intelligibility in different languages is of great interest, particularly when non-linear noise-reduction processing is involved. In the present study, an extensive evaluation is taken of objective measures for speech intelligibility prediction of noisy speech processed by noise-reduction algorithms in Chinese, Japanese, and English. Of all the objective measures tested, the short-time objective intelligibility (STOI) measure produced the most accurate results in speech intelligibility prediction for Chinese, while the normalized covariance metric (NCM) and middle-level coherence speech intelligibility index ($\mathrm{CSII_m}$) incorporating the signal-dependent band-importance functions (BIFs) produced the most accurate results for Japanese and English, respectively. The objective measures that performed best in predicting the effect of non-linear noise-reduction processing in speech intelligibility were found to be the BIF-modified NCM measure for Chinese, the STOI measure for Japanese, and the BIF-modified $\mathrm{CSII_m}$ measure for English. Most of the objective measures examined performed differently even under the same conditions for different languages.

© *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4901079]

PACS number(s): 43.71.Gv, 43.71.Hw, 43.72.Dv [TB] Pages: 3301–3312

I. INTRODUCTION

Speech can be assessed in terms of its quality and intelligibility. Speech quality is related to how natural speech sounds, while speech intelligibility is related to the number of speech items that are recognized correctly by the listener (Havelock *et al.*, 2009). The latter is the focus of the present study. Speech intelligibility can be measured in a subjective or objective way. The most accurate approach for speech intelligibility evaluation is through listening tests which involve panels of human subjects. Though subjective evaluation is accurate, it is inconvenient, expensive, and time consuming. In contrast, objective measurement of speech intelligibility is not only convenient and less expensive, but also yields more consistent results that are immune to subjects' biases (Liu *et al.*, 2006; Ma *et al.*, 2009; Taal *et al.*, 2011). Therefore, much effort has been devoted to developing objective measures that are able to predict speech intelligibility as accurately as possible.

A number of objective measures have been proposed in the literature to predict speech intelligibility in the presence of background noise. Among these measures, the articulation index (AI) (French and Steinberg, 1947; Kryter, 1962) and speech transmission index (STI) (Steeneken and Houtgast, 1980; Houtgast and Steeneken, 1985) are by far the most commonly used for predicting speech intelligibility in noisy and reverberant conditions. The STI is computed as a weighted average of metrics derived from envelopes of signals in multiple frequency bands (Houtgast and Steeneken, 1985). By incorporating the factors used in the computation of STI, the AI measure was further developed to create the speech intelligibility index (SII) (ANSI, 1997). The SII measure is based on the idea of estimating an effective amount of audible speech information in a number of frequency bands. The audible information is weighted by an empirically determined importance function that describes the relative importance of the individual frequency bands to intelligibility (ANSI, 1997). These objective measures have been found to perform poorly in predicting intelligibility of processed speech wherein non-linear operations (e.g., noise-reduction) are involved (Ma *et al.*, 2009; Taal *et al.*, 2011). In recent years, therefore, increased interest has been focused on objective measures able to predict the intelligibility of speech signals after non-linear noise-reduction processing.

A variety of objective speech quality measures have been assessed in predicting speech intelligibility in the context of additive noise as well as degradations introduced by non-linear noise-reduction processing (Liu *et al.*, 2006; Yamada *et al.*, 2006; Ma *et al.*, 2009; Taal *et al.*, 2011). No objective quality measures were consistently found to perform well for speech intelligibility prediction. For instance, Yamada *et al.* (2006) reported that a high correlation with subjective intelligibility ratings was obtained by the perceptual estimation of speech quality (PESQ) measure, while

[a] Author to whom correspondence should be addressed. Electronic mail: junfeng.li.1979@gmail.com

Taal *et al.* (2011) found the PESQ measure showed a low correlation. By extending the original SII concept, Kates and Arehart (2005) presented a coherence SII (CSII) measure to include broadband peak-clipping and center-clipping distortion. While the CSII measure yielded a modest correlation with subjective intelligibility ratings, three level CSII measures were further suggested that divided the speech segments into three level regions and computed the CSII index separately for each (Kates and Arehart, 2005); this processing yielded higher correlations with speech intelligibility ratings when hearing-aid type distortions were involved (Arehart *et al.*, 2007). Moreover, Hollube and Kollmeier (1996) presented the normalized covariance metric (NCM) that was computed as a weighted sum of the transmission index determined from the envelopes of the input and output signals in each frequency band. The NCM measure was shown to reliably predict the intelligibility of noise-reduced speech containing non-linear distortions (Hollube and Kollmeier, 1996; Goldsworthy and Greenberg, 2004). To further improve the predictive power of the CSII and NCM measures, Ma *et al.* (2009) suggested a set of signal-dependent band-importance functions (BIFs) for predicting the intelligibility of noise-reduced speech in situations where the target speech was corrupted by fluctuating maskers. The modified CSII and NCM measures incorporating the signal-dependent BIFs were found to perform the best for noise-reduced speech among various tested objective measures (Ma *et al.*, 2009). More recently, Taal *et al.* (2011) introduced a short-time objective intelligibility (STOI) measure that decomposes signals into short-time time-frequency regions, followed by the normalization and clipping procedures. The STOI measure accurately predicted speech intelligibility for speech signals distorted by non-stationary noises and non-linear processing, e.g., single-channel noise-reduction algorithms (Taal *et al.*, 2011).

The existing studies on objective measures for predicting speech intelligibility of noise-corrupted signals processed by non-linear processing were mainly performed using Western languages (e.g., English). However, different languages are characterized by diverse specific features at the acoustic and phonetic levels (Trask, 1998). For example, the tone information (as carried in F0 contour) in Chinese and the pitch accent information in Japanese are used to distinguish word meaning and thus contribute a great deal to Chinese and Japanese speech intelligibility (Arai *et al.*, 1996; Fu *et al.*, 1998). In contrast, F0 information in English is used primarily to emphasize or express emotion and convey intonation, among others, and thus contributes little to speech intelligibility, at least in quiet (Banziger and Scherer, 2005). F0 information, however, can be used by listeners to segregate the target talker in competing-talker listening tasks (Wang and Brown, 2006). Considering the possible influence of language structure and speech cues on objective measures, Houtgast and Steeneken (1984) examined the rapid speech transmission index (RASTI) across ten western languages and showed that language-specific effects could result in disparity among the fourteen conditions tested. Li *et al.* (2011) examined the effects of language on five single-channel noise-reduction algorithms in terms of speech intelligibility for three languages: Chinese, Japanese, and English. The results showed that no improvement in speech intelligibility was given by the majority of noise-reduction algorithms, and more importantly the performance of noise-reduction algorithms differed significantly across the three languages.

It is clear from the abovementioned studies that the performance in speech intelligibility of the noise-reduction algorithms varies significantly across different languages, and that the abilities of objective speech intelligibility prediction measures are influenced by non-linear noise-reduction processing. Therefore, an examination of the performance of various objective measures in predicting speech intelligibility across different languages is of great interest, especially after non-linear noise-reduction processing. To this end, the present study first selected six objective measures which showed high accuracy in predicting the intelligibility of noise-corrupted speech signals processed using the methods reported in many existing studies (Kates and Arehart, 2005; Hollube and Kollmeier, 1996; Ma *et al.*, 2009; Taal *et al.*, 2011). Due to their dependency on the BIF, the NCM and CSII measures were further evaluated by incorporating the different signal-dependent BIFs suggested in Ma *et al.* (2009). Subsequently, the potential effect of language on the BIF-modified objective measures in speech intelligibility prediction was evaluated for noise-corrupted signals and noise-reduced signals by non-linear noise-reduction processing. These evaluations were performed by correlating the objective prediction scores and subjective intelligibility ratings collected in two previous studies (Li *et al.*, 2011; Hu and Loizou, 2007) where speech signals in three languages (Chinese, Japanese, and English) were corrupted by two types of noise at two signal-to-noise ratios and then processed by five single-channel noise-reduction algorithms. The contributions of the present research are as follows: first, to our knowledge this research is the first evaluation of the objective measures in speech intelligibility prediction for Chinese and Japanese; second, the present study assesses the potential effect of language on the objective measures in predicting the speech intelligibility of noise-reduced signals by noise-reduction algorithms; third, it provides valuable information and insight regarding the objective measure(s) that is the most appropriate for predicting speech intelligibility after non-linear noise-reduction processing in different languages.

II. METHODS

The subjective intelligibility evaluation of noise-corrupted speech processed by five typical single-channel noise-reduction algorithms for Chinese and Japanese was reported in Li *et al.* (2011) and for English in Hu and Loizou (2007), which is summarized briefly below.

A. Materials

In the subjective evaluation of noise-reduction algorithms, the speech materials were taken from the following databases. For Chinese, the database for the intelligibility test reported by Ma and Shen (2004) was adopted. This database consists of ten tables, each of which contains 75 phonetically balanced Chinese words with consonant-vowel structure. In each table, every three words are combined randomly to form one nonsense sentence, producing a total of 25 sentences. For Japanese, the familiarity-controlled word lists (FW03) that consist of 80 lists with 50 phonetically balanced words per list were used (Amano *et al.*, 2009). Because word familiarity has a strong effect on word recognition, all word lists in FW03 are divided into four sets in four word-familiarity ranks. In the present investigation, only the word lists with the lowest familiarity were used. For English, the IEEE database was selected as it contains phonetically balanced words with relatively low word-context predictability (IEEE, 1969). In this database, there are 72 lists of sentences where each list contains 10 sentences and each sentence is composed of approximately 7–12 words. All these speech signals were downsampled to 8 kHz prior to being corrupted by babble and car noise signals taken from the AURORA database (Hisch and Pearce, 2000). Both clean speech and noise signals were processed by the IRS filter to simulate the receiving frequency characteristics of telephone handsets. The noise signals were added to the speech signals at signal-to-noise ratios (SNRs) of 0 and 5 dB.

B. Signal processing

The noise-corrupted signals were processed by five representative noise-reduction algorithms including the generalized Karhunen–Loeve-transform (KLT) approach (Hu and Loizou, 2003), the log minimum mean square error (logMMSE) algorithm (Ephraim and Malah, 1985), the log minimum mean square error with speech presence uncertainty (logMMSE-SPU) (Cohen and Berdugo, 2001), the multiband spectral subtraction algorithm (MB) (Kamath and Loizou, 2002), and the Wiener filter based on the *a priori* SNR estimation (Wiener-as) (Scalart and Filho, 1996), which cover the four major classes of state-of-the-art single-channel noise-reduction algorithms. MATLAB implementations of all these noise-reduction algorithms are available in Loizou (2007).

C. Procedure

The speech signals processed by the five noise-reduction algorithms, along with the noise-corrupted signals, were presented to listeners at a comfortable listening level through headphones for word identification. In the intelligibility evaluation for Chinese, ten Chinese listeners were recruited and participated in a total of 24 listening conditions [2 SNR levels × 2 types of background noise × 6 algorithms (1 noisy signal + 5 noise-reduction algorithms)]. One list of sentences (25 sentences) taken from the KXY database was used per condition. Each subject listened to 600 sentences (25 sentences × 24 conditions) in the listening tests (Li *et al.*, 2011). In the intelligibility evaluation for Japanese, 20 Japanese listeners were recruited and grouped into two panels (one panel per type of noise) with each panel consisting of ten listeners. Each subject participated in a total of 12 listening conditions [2 SNR levels × 6 algorithms]. One list of 50 words was used for each condition. Each subject listened to 600 words (50 words × 12 conditions) in the listening tests (Li *et al.*, 2011). In the intelligibility evaluation for English, 20 English listeners were recruited and divided into two panels with each panel of ten listeners. Each subject participated in 12 listening conditions [2 SNR levels × 6 algorithms]. Two sentence lists (ten sentences per list) were used for each condition. Each subject listened to 240 sentences (20 sentences × 12 conditions) in the listening tests (Hu and Loizou, 2007). The presentation order of the stimuli and listening conditions were randomized for each subject. Subjects were asked to write down the words they heard. The subjective intelligibility scores for Chinese, Japanese, and English obtained in Li *et al.* (2011) and Hu and Loizou (2007) were used in the present study to evaluate the predictive power of the objective speech intelligibility measures.

III. OBJECTIVE INTELLIGIBILITY PREDICTION MEASURES

Six objective measures were selected for evaluation in this study due to their high ability in predicting speech intelligibility when non-linear noise-reduction processing is involved (Hollube and Kollmeier, 1996; Kates and Arehart, 2005; Boldt and Ellis, 2009; Ma *et al.*, 2009; Taal *et al.*, 2011). For each objective measure, a general descriptive notation was adopted. The outcome of an objective measure is denoted by $d_s(x, \hat{x})$, where the subscript s indicates the name of objective measure, x denotes the clean speech signal, and $\hat{x}$ the noise-reduced speech signal by the single-channel noise-reduction algorithms. Let m, k, and l represent the time-frame index, frequency-bin or subband index, and time-sample index in a frame, respectively. The lth sample of the mth frame of x is denoted by $x(l, m)$ and its corresponding kth component $X(k, m)$ in the frequency domain. Similarly, the noise signal and its frequency-domain counterpart are denoted as $n(l, m)$ and $N(k, m)$. Let M, L, and K denote the total number of frames, the frame length and the total number of frequency bins (or subbands), respectively.

A. Coherence-based measure

The coherence-based (COH) measure investigated here is the magnitude-squared coherence (MSC), computed by dividing the input (clean) and output (noise-reduced) signals into a number of frames followed by computing the cross power spectrum in each frame and then averaging across all frames, that is,

$$\gamma(k) = \frac{\sum_m |X(k,m)\widehat{X}^*(k,m)|^2}{\sum_m |X(k,m)|^2 \sum_m |\widehat{X}^*(k,m)|^2}, \quad (1)$$

where an asterisk denotes the complex conjugate. The COH measure is eventually computed as

$$d_{\mathrm{COH}} = \frac{1}{K}\sum_k \gamma(k). \quad (2)$$

B. Short-time objective intelligibility measure

The short-time objective intelligibility (STOI) measure was originally proposed by Taal *et al.* (2011). Its essential idea is to compare the short-time temporal envelope of the clean signal and that of the noise-reduced signal in each one-third octave band by means of a correlation coefficient. The short-time temporal envelope of the signal in the k th one-third octave band is given by

$$\mathbf{P}_{\tilde{x}}(k,m) = [P_{\tilde{x}}(k,m-Q+1),\ P_{\tilde{x}}(k,m-Q+2), \ldots,\ P_{\tilde{x}}(k,m)]^T, \tag{3}$$

where $\tilde{x} \in \{x, \hat{x}\}$, Q is the number of frames used in computation of the short-time temporal envelope, $P_{\tilde{x}}(k,m)$ denotes the norm of the signal $\tilde{x}$ in the k th one-third octave band computed as

$$P_{\tilde{x}}(k,m) = \sqrt{\sum_j |\tilde{X}(j,m)|^2}, \tag{4}$$

where $j \in [B_k, B_{k+1}-1]$ is the frequency index, and B_k and B_{k+1} denote the boundary frequencies of the k th and $k+1$ one-third octave bands.

Prior to calculating the correlation, the short-time envelope of the noise-reduced signal, $\mathbf{P}_{\hat{x}}(k,m)$, is first normalized and clipped (Taal *et al.*, 2011), given by

$$\mathbb{P}_{\tilde{x}}(k,m) = \min\left(\frac{\|\mathbf{P}_x(k,m)\|}{\|\mathbf{P}_{\hat{x}}(k,m)\|}\mathbf{P}_{\hat{x}}(k,m),\ (1+10^{-\beta/20})\mathbf{P}_x(k,m)\right), \tag{5}$$

where β is the lower signal-to-distortion ratio (SDR) bound, which was empirically set to -15 dB. The correlation coefficient between the temporal envelopes of the clean and noise-reduced speech signals is computed as

$$\rho(k,m) = \frac{\left(P_x(k,m)-\bar{P}_x(k,m)\right)^T\left(\mathbb{P}_{\hat{x}}(k,m)-\bar{\mathbb{P}}_{\hat{x}}(k,m)\right)}{\|P_x(k,m)-\bar{P}_x(k,m)\| \times \|\mathbb{P}_{\hat{x}}(k,m)-\bar{\mathbb{P}}_{\hat{x}}(k,m)\|}, \tag{6}$$

where $\bar{P}_x$ and $\bar{\mathbb{P}}_{\hat{x}}$ denote the sample averages across the vector P_x and $\mathbb{P}_{\hat{x}}$, respectively. The STOI measure is calculated by averaging the correlation coefficients over all sub-bands and frames, given by

$$d_{\mathrm{STOI}} = \frac{1}{KM}\sum_{k,m}\rho(k,m). \tag{7}$$

C. Coherence speech intelligibility index

The coherence speech intelligibility index (CSII) was first presented by Kates and Arehart (2005) for assessing the effects of non-linear distortions (e.g., peak clipping) on speech intelligibility. The CSII measure improved the traditional SII measure by replacing the SNR in the computation of SII with the signal-to-distortion ratio (SDR), given by

$$\mathrm{SDR}(k,m) = 10\log_{10}\frac{\sum_j G_k(j)\gamma(j)|\hat{x}(j,m)|^2}{\sum_j G_k(j)[1-\gamma(j)]|\hat{x}(j,m)|^2}, \tag{8}$$

where j is the frequency index in the kth auditory band and G_k denotes the frequency weight by means of a ro-ex filter (Kates and Arehart, 2005), and $\gamma(k)$ is the MSC calculated as in Eq. (1). Consistent with the limitation applied in SII, the SDR is further confined to $[-15, 15]$ dB and mapped linearly between 0 and 1, that is,

$$\widehat{\mathrm{SDR}}(k,m) = \frac{\mathrm{SDR}(k,m)+15}{30}. \tag{9}$$

The CSII measure is then calculated as (Kates and Arehart, 2005)

$$d_{\mathrm{CSII}} = \frac{1}{M}\sum_m\frac{\sum_k W(k,m)\widehat{\mathrm{SDR}}(k,m)}{\sum_k W(k,m),} \tag{10}$$

where $W(k,m)$ denotes the band importance function (BIF). In addition to the fixed BIF used in the traditional SII computation (ANSI, 1997), the following four signal-dependent BIFs suggested by Ma *et al.* (2009) are also examined here:

$$W_1(k,m) = \begin{cases} 1 & \text{if } X(k,m) > D(k,m) \\ 0 & \text{else,} \end{cases} \tag{11}$$

$$W_2(k,m) = \begin{cases} (X(k,m)-D(k,m))^p & \text{if } X(k,m) > D(k,m) \\ 0 & \text{else,} \end{cases} \tag{12}$$

$$W_3(k,m) = \begin{cases} X^p(k,m) & \text{if } X(k,m) > D(k,m) \\ 0 & \text{else,} \end{cases} \tag{13}$$

$$W_4(k,m) = X^p(k,m), \tag{14}$$

where $D(k,m)$ denotes the critical-band spectrum of the masker signal, the power exponent p controls the emphasis or weight placed on spectral peaks and spectral valleys.

Furthermore, Kates and Arehart (2005) found that the speech segments could be divided into three level regions and the CSII measure could be computed separately in each region. Among these measures, the middle level CSII ($\mathrm{CSII_m}$) measure yielded the highest correlation for English, as reported by Arehart *et al.* (2007).

D. Normalized covariance metric

The normalized covariance metric (NCM) was first reported by Hollube and Kollmeier (1996), which was designed on the covariance between the envelopes of the clean and processed signals. The envelope in each subband was computed by the Hilbert transform followed by limiting the envelope modulation frequencies to 0–12.5 Hz. The normalized covariance in the k th subband of the envelope of the clean signal and that of the processed signal is then calculated as

$$\varrho(k) = \frac{\sum_t (e_x(k,t)-\bar{e}_x(k,t))(e_{\hat{x}}(k,t)-\bar{e}_{\hat{x}}(k,t))}{\sqrt{\sum_t (e_x(k,t)-\bar{e}_x(k,t))^2}\sqrt{\sum_t (e_{\hat{x}}(k,t)-\bar{e}_{\hat{x}}(k,t))^2},} \tag{15}$$

where e_x and $e_{\hat{x}}$ are the envelopes of the clean and processed signals, respectively. $\bar{e}_x$ and $\bar{e}_{\hat{x}}$ are the mean values of e_x and $e_{\hat{x}}$. Note that the values of $\varrho(k)$ are limited to $|r(k)| \leq 1$.

The SNR term at the kth subband is then computed as

$$SNR(k) = 10\log_{10}\left(\frac{\varrho^2(k)}{1-\varrho^2(k)}\right), \tag{16}$$

which is subsequently limited to the range of [−15, 15] dB and further mapped to $\widehat{\mathrm{SNR}}(k)$ in the range of [0,1] in the same way as given by Eq. (9). The NCM measure is finally computed by

$$d_{\mathrm{NCM}} = \frac{\sum_k W(k)\widehat{\mathrm{SNR}}(k)}{\sum_k W(k)}, \tag{17}$$

where $W(k)$ are the weights applied to the kth subband. The value of $W(k)$ can be the fixed weights (ANSI, 1997), or the following signal-dependent BIFs suggested by Ma *et al.* (2009):

$$W^{(1)} = \left(\sum_t x^2(k,t)\right)^p, \tag{18}$$

$$W^{(2)} = \left(\sum_t (\max[x(k,t) - n(k,t), 0])^2\right)^p. \tag{19}$$

E. Normalized subband envelope correlation

The normalized subband envelope correlation (NSEC) was suggested by Boldt and Ellis (2009). In the computation of NSEC, a gammatone filter bank is first applied to the clean and processed signals. The normalized, compressed and high-passed filtered intensity envelopes $\mathrm{E}(k,m)$ in the kth frequency band and the m th frame are then extracted (Boldt and Ellis, 2009). The NSEC measure is eventually determined as the normalized correlation over all time frames and frequency bins (Boldt and Ellis, 2009), given by

$$d_{\mathrm{NSEC}} = \frac{\sum_{k,m} \mathrm{E}_x(k,m)\mathrm{E}_{\hat{x}}(k,m)}{\sqrt{\sum_{k,m} \mathrm{E}_x^2(k,m) \sum_{k,m} \mathrm{E}_{\hat{x}}^2(k,m)}}. \tag{20}$$

IV. ANALYSIS AND RESULTS

Prior to examining the abilities of the objective measures in predicting speech intelligibility in different languages, the set of CSII and NCM measures were first investigated and improved by incorporating the signal-dependent band-importance functions (BIFs) due to their dependence on the BIFs as shown in Eqs. (10) and (17). Subsequently, two examinations were performed to assess the abilities of the objective measures in different languages. The first examination was to test the overall abilities of the objective measures in predicting speech intelligibility in all tested conditions for three languages. This examination helped to discover the potential effect of language on the overall predicting abilities of the objective measures in all conditions. The second examination was to further demonstrate the effect of language on the abilities of the objective measures in predicting the benefits of the non-linear noise-reduction algorithms in terms of speech intelligibility.

A. Improvement of the CSII and NCM measures by incorporating signal-dependent band-importance functions (BIFs)

The set of CSII and NCM measures were examined in predicting speech intelligibility for the three languages in terms of Pearson's correlation coefficient (r) between the objectively predicted scores and the subjective intelligibility ratings, and further improved by incorporating the signal-dependent band-importance functions (BIFs). The higher value of r indicates that the objective measure is better in predicting speech intelligibility. In the present examinations, the value of p used in the BIFs [Eqs. (11)–(14) and (18) and (19)] varied from 0.5 to 4.0, which controls the emphasis on spectral peaks and spectral valleys (Ma *et al.*, 2009).

1. Results

The results of the set of CSII measures (CSII, $\mathrm{CSII_m}$) and the NCM measure with various signal-dependent BIFs in terms of correlation coefficient (r) for Chinese, Japanese, and English are shown in Figs. 1–3. In all cases, the lowest correlation was obtained when the fixed BIF, taken from the (ANSI, 1997) standard, was used in the computation of the CSII and NCM measures. Significant improvements in the correlations were obtained with the CSII and NCM measures when applying the signal-dependent BIFs [Eqs. (11)–(14) and (18) and (19)]. The correlation of the CSII measure improved from $r = 0.62$ with the fixed weights (ANSI, 1997) to $r = 0.75$ with the signal-dependent BIF (W_3, $p = 4.0$) in Eq. (13) for Chinese, and from $r = 0.57$ to 0.76 for Japanese, and from $r = 0.81$ to 0.89 for English. With the signal-dependent BIF (W_4, $p = 4.0$) in Eq. (13), the correlation of the middle-level CSII ($\mathrm{CSII_m}$) measure increased from $r = 0.71$ to 0.82 for Chinese, from $r = 0.53$ to 0.81 for Japanese and from $r = 0.91$ to 0.94 for English. The correlations of the NCM measure with the fixed weights ($r = 0.75$ for Chinese, $r = 0.77$ for Japanese, and $r = 0.89$ for English) were also significantly improved when incorporating the signal-dependent BIF $W^{(2)}$ with $p = 4.0$ ($r = 0.87$ for Chinese, $r = 0.84$ for Japanese, and $r = 0.92$ for English).

2. Discussion

In comparison to the fixed BIF (ANSI, 1997), the signal-dependent BIFs [Eqs. (11)–(14) and (18) and (19)] significantly improved the correlation of the CSII and NCM measures for Chinese, Japanese, and English. This result is consistent with the results reported in Ma *et al.* (2009). In the three languages, it was observed that the improvements in predicting speech intelligibility of the CSII and NCM measures were influenced by the signal-dependent BIFs. Furthermore, the correlation improvements introduced by

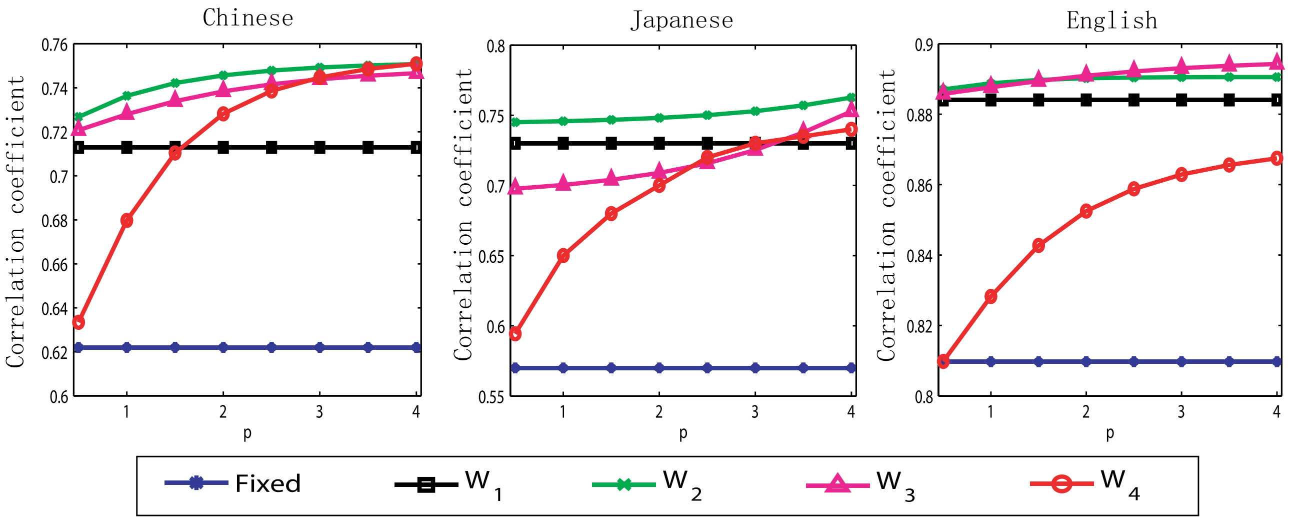

FIG. 1. (Color online) Pearson's correlation coefficients between subjective intelligibility ratings and the predicted scores by the CSII measure incorporating the different signal-dependent BIFs for Chinese (left), Japanese (middle), and English (right).

most of the signal-dependent BIFs were also dependent on the value of the power exponent p. The correlations of the CSII and NCM measures incorporating the signal- and p-dependent BIFs increased as the value of p increased, as shown in Figs. 1–3. The improvement in the correlation of the CSII and $\mathrm{CSII_m}$ measures could be attributed to the fact that the increased value of p placed more emphasis on the dominant spectral peaks. The increased correlation of the NCM measure was attributed to the fact that more emphasis was placed to each SDR (transmission index) value in proportion to the signal energy in each band [Eq. (18)] or to the excess masked signal [Eq. (19)].

The best performance in predicting speech intelligibility in terms of the correlation with the CSII measure was found for the signal-dependent BIF (W_2, $p = 4$) defined in Eq. (12) for Chinese and Japanese, and for the BIF (W_3, $p = 4$) defined in Eq. (13) for English. These two BIFs included only the bands with positive SDRs where the target signal was stronger than the masker, which contributed the most to speech intelligibility in noise. The signal-dependent BIF (W_4, $p = 4$) consistently yielded the highest correlation for the $\mathrm{CSII_m}$ measure in the three languages. This benefit was obtained because the BIF W_4 with $p = 4.0$ places more emphasis on envelope transients and spectral transitions in the computation of the $\mathrm{CSII_m}$ measure, which are critical for the transmission of the information regarding place of articulation (Furui, 1986). The highest correlation of the NCM measure was found for the signal-dependent BIF ($W^{(2)}$, $p = 4.0$) for Chinese and English, and for the BIF ($W^{(1)}$, $p = 4.0$) for Japanese. The NCM measure accounts for the average envelope power in each band as well as for the low-frequency envelope modulations, which are known to carry critically important information about speech (Luo and Fu, 2006; Brown and Bacon, 2010; Liu *et al.*, 2014). The evaluation results suggest that in the three different languages (Chinese, Japanese, and English), the set of CSII measures (CSII, $\mathrm{CSII_m}$), and the NCM measure could be greatly improved by incorporating the signal- and p-dependent BIFs. The CSII and NCM measures with the signal-dependent BIFs, which resulted in the highest correlation in different languages, will be further used in the following two examinations.

B. Evaluation of the objective measures in predicting speech intelligibility of noise-corrupted and noise-reduced signals for three languages

The overall performance of the objective measures in predicting speech intelligibility was evaluated for the three

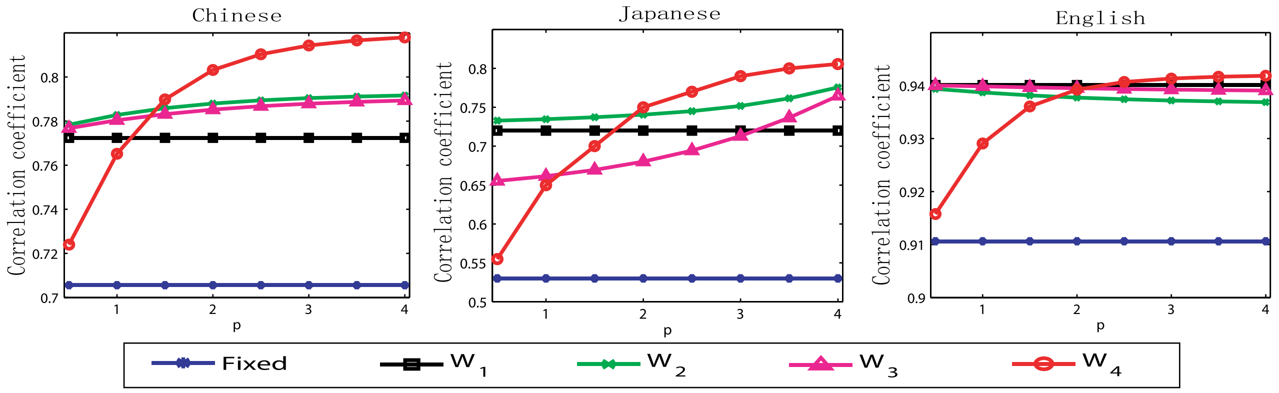

FIG. 2. (Color online) Pearson's correlation coefficients between subjective intelligibility ratings and the predicted scores by the $\mathrm{CSII_m}$ measure incorporating the different signal-dependent BIFs for Chinese (left), Japanese (middle), and English (right).

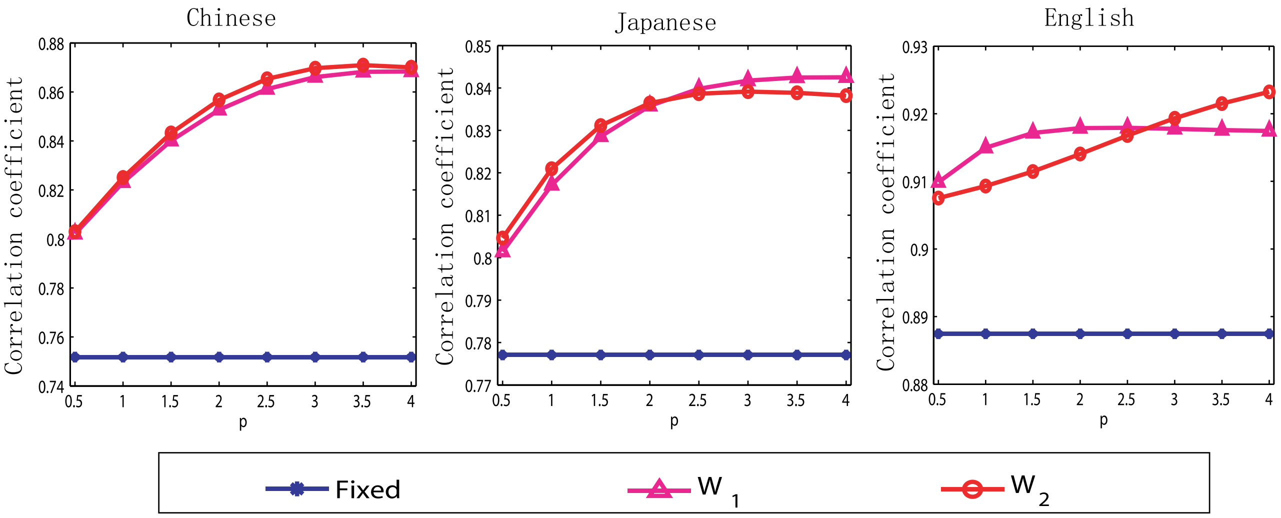

FIG. 3. (Color online) Pearson's correlation coefficients between subjective intelligibility ratings and the predicted scores by the NCM measure incorporating the different signal-dependent BIFs for Chinese (left), Japanese (middle), and English (right).

languages under all conditions, including the noise-corrupted signals and the noise-reduced signals by the non-linear noise-reduction algorithms. The evaluations were performed in terms of two objective measures. The first measure was the Pearson's correlation coefficient r, as used in the previous section; the second 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.

1. Results

The evaluation results of the objective intelligibility prediction measures in terms of the correlation coefficient (r) and the standard deviation of prediction error (σ_e) for Chinese, Japanese, and English are shown in Table I. From Table I, it can be seen most objective measures exhibited different in predicting speech intelligibility for the three different languages. For Chinese, the STOI measure yielded the highest correlation $(r = 0.90)$ and the lowest standard deviation of error $(\sigma_e = 4.73\%)$, corresponding to the highest ability in predicting the subjective intelligibility ratings. It was followed by the NCM measure $(r = 0.87, \sigma_e = 5.93\%)$. The lowest ability in predicting Chinese speech intelligibility was given by the COH measure $(r = 0.60, \sigma_e = 9.61\%)$. For Japanese, however, all the objective measures tested yielded modest ability in predicting speech intelligibility; the best intelligibility prediction ability was found as $(r = 0.84, \sigma_e = 6.07\%)$ with the NCM measure. The objective measures tested demonstrated quite similar performance in Japanese speech intelligibility prediction. The CSII measure had the lowest ability in predicting Japanese speech intelligibility (r = 0.75, σ_e = 7.47%). In comparison to the speech intelligibility prediction for Chinese and Japanese, most objective measures showed good ability in predicting English speech intelligibility, except for the COH measure that resulted in the lowest intelligibility prediction ability (r = 0.71, σ_e = 12.36%). The highest English intelligibility prediction ability was given by the CSII$_\mathrm{m}$ measure (r = 0.94, σ_e = 5.90%), which was followed by the STOI measure (r = 0.92, σ_e = 6.79%) and the NCM measure (r = 0.92, σ_e = 6.75%). It is unsurprising that most objective measures have good intelligibility prediction ability for English, because that these objective measures were originally designed and optimized for English. In contrast, these objective measures provided modest (even low) ability in speech intelligibility prediction for Chinese and Japanese.

TABLE I. The Pearson's correlation coefficients r and the standard deviations of the error σ_e, averaged across all signals under two noise conditions (babble noise and car noise) at two SNRs (0 and 5 dB), for six objective intelligibility prediction measures.

		COH	CSII	CSII$_\mathrm{m}$	NCM	NSEC	STOI
Chinese	r	0.60	0.75	0.82	0.87	0.78	0.90
	σ_e	9.61%	7.95%	6.92%	5.93%	7.25%	4.73%
Japanese	r	0.80	0.75	0.79	0.84	0.83	0.83
	σ_e	6.70%	7.47%	6.84%	6.07%	6.29%	6.40%
English	r	0.71	0.89	0.94	0.92	0.82	0.92
	σ_e	12.36%	7.86%	5.90%	6.75%	10.09%	6.79%

2. Discussion

The COH measure demonstrated the worst speech intelligibility prediction ability in all tested noise conditions especially for Chinese and English. This result may have arisen because speech intelligibility cannot be accurately predicted from the normalized short-time amplitude spectra of speech signals. The low ability of the COH measure in predicting speech intelligibility for English was also observed by Ma *et al.* (2009), especially when non-linear noise-reduction processing was involved.

Compared with the COH measure, the CSII measure improved the speech intelligibility prediction ability in all conditions. This benefit was partially because of the normalization and constraint operations of the SDR in the computation of the CSII measure, which accounted for the effects caused by non-linear processing, e.g., peak clipping as reported in Kates and Arehart (2005). The additional benefit of the CSII measure in predicting speech intelligibility may have come from the signal-dependent BIFs that emphasized the dominant spectral peaks. The CSII$_m$ measure consistently yielded higher speech intelligibility prediction over the CSII measure for all three languages. This result might be attributed to a high number of transitions between consonants and vowels in the middle-level range (Kates and Arehart, 2005; Chen and Loizou, 2012), which play an important role in speech understanding especially in noise (Furui, 1986).

The NSEC measure yielded modest performance in speech intelligibility prediction for Chinese, Japanese and English. This result might be partially due to the use of the spectral envelopes in the computation of the NSEC measure, and spectral envelopes play an important role in speech understanding (Drullman, 1995). The normalization, compression and high-pass filtering operations involved in the computation of the NSEC measure were originally developed for accounting for the artifacts introduced by the ideal binary mask (IBM) filter (Boldt and Ellis, 2009). These operations might be not appropriate for accounting for the effect of non-linear noise-reduction processing in predicting speech intelligibility.

The NCM and STOI measures showed much better ability in speech intelligibility prediction, which was mainly due to the employment of the temporal envelopes of the clean and noise-reduced signals that play a crucial role in speech recognition especially in noise (Drullman, 1995). The normalization and correlation of temporal envelopes in each short-time segment might account for the effect of non-linear noise-reduction processing to a great degree. The ability of the NCM measure to accurately predict speech intelligibility was also attributed to the use of the signal-dependent BIFs. While both the NCM and STOI measures only provided relatively modest speech intelligibility prediction for Japanese. This result might be related to the quite low subjective intelligibility ratings for Japanese as reported in Li *et al.* (2011).

As a result, the objective measure that provided good speech intelligibility prediction for certain languages (e.g., English) even after being processed by non-linear noise-reduction algorithms might be not appropriate for other languages (e.g., Chinese and Japanese). Therefore, there is a clear influence of language on the ability of the objective measures in predicting speech intelligibility, especially when the non-linear noise-reduction processing is involved.

C. Evaluation of the objective measures in predicting the performance in speech intelligibility of non-linear noise-reduction processing for three languages

In this section, the abilities of the objective measures in predicting the performance of non-linear single-channel noise-reduction algorithms were evaluated in terms of speech intelligibility for the three languages. Generally, only certain monotonic relationships are present in the intelligibility scores for noise-corrupted signals and noise-reduced signals by non-linear noise-reduction processing (Taal *et al.*, 2011). To further demonstrate the abilities of the objective measures in speech intelligibility prediction for the three languages before and after noise-reduction processing, a mapping was performed for each language to account for the non-linear relationship between the objective intelligibility prediction scores and the subjective intelligibility ratings. A widely used mapping is a logistic function, given by

$$f(d) = \frac{100}{1+\exp(ad+b)}, \tag{21}$$

where a and b are the parameters that were tuned with a non-linear least square procedure, and d denotes the objective prediction score. This logistic function was only fitted to the noise-corrupted conditions, which was then used to predict the intelligibility scores for the noise-reduced conditions processed by the noise-reduction algorithms. The performance of all objective measures was evaluated with the root mean square (RMS) of the prediction error (RMSE), defined as

$$\sigma = \sqrt{\frac{1}{Z}\sum_i (z_i - f(d_i))^2}, \tag{22}$$

where z_i refers to the intelligibility score obtained in the processing condition i and Z denotes the total number of processing conditions.

1. Results

The scatter plots of the subjective intelligibility ratings for Chinese, Japanese, and English against the intelligibility scores predicted by the objective measures are shown in Figs. 4–6, along with the fitting curves and the RMSE results. These results demonstrated that the lowest RMSEs were, respectively, given by the NCM measure for Chinese ($\sigma = 6.38\%$), by the STOI measure for Japanese ($\sigma = 6.24\%$), and by the CSII$_m$ measure for English ($\sigma = 6.00\%$), corresponding to the best ability in predicting the effect of the noise-reduction algorithms on speech intelligibility. The second lowest RMSEs were provided by the STOI measure ($\sigma = 8.82\%$) for Chinese, by the NCM measure ($\sigma = 8.09\%$) and the CII$_m$ measure ($\sigma = 8.78\%$) for Japanese, and by the NCM measure ($\sigma = 8.60\%$) and the STOI measure ($\sigma = 8.71\%$) for English. The worst performance was consistently introduced by the COH measure ($\sigma = 15.11\%$ for Chinese, $\sigma = 12.55\%$ for Japanese, $\sigma = 16.05\%$ for English) and the NSEC measure ($\sigma = 14.96\%$ for Chinese, $\sigma = 12.46\%$ for Japanese, $\sigma = 16.07\%$ for English).

More importantly, both the NCM and STOI measures yielded the much higher ability in predicting the performance of the noise-reduction algorithms for Chinese and Japanese, and the CII$_m$ measure did well for English. These findings were observed since the scattered points predicted by the NCM, STOI, and CSII$_m$ measures followed the general tendency of the mapping curves, as shown in Figs. 4–6. These findings were consistent with the results reported by Taal *et al.* (2011) in which the STOI measure showed the best ability in predicting the effect of non-linear noise-reduction

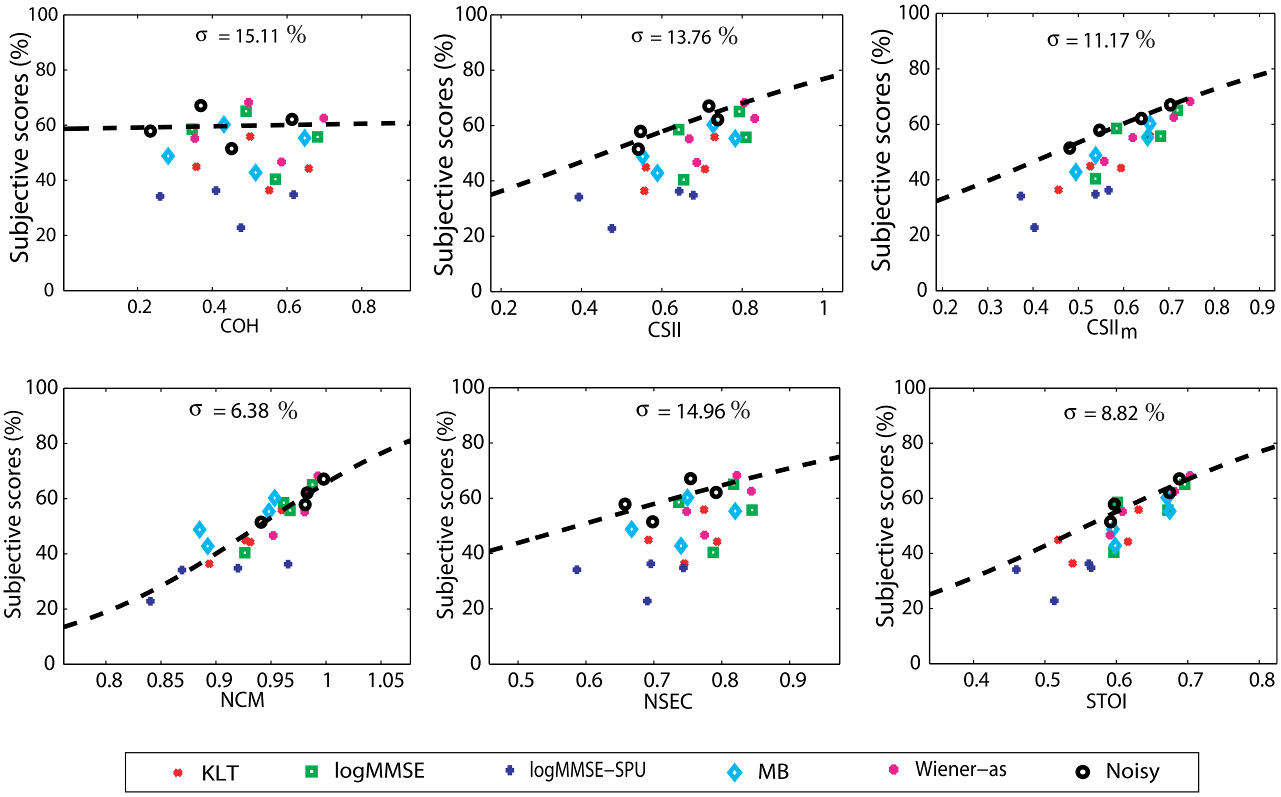

FIG. 4. (Color online) Scatter plots of the subjective intelligibility ratings against the objectively predicted scores for Chinese, along with the mapping results (dashed curves) and the RMSE results (σ). Each point on this figure represents one pair of the subjective intelligibility rating and the objectively predicted score in one tested condition.

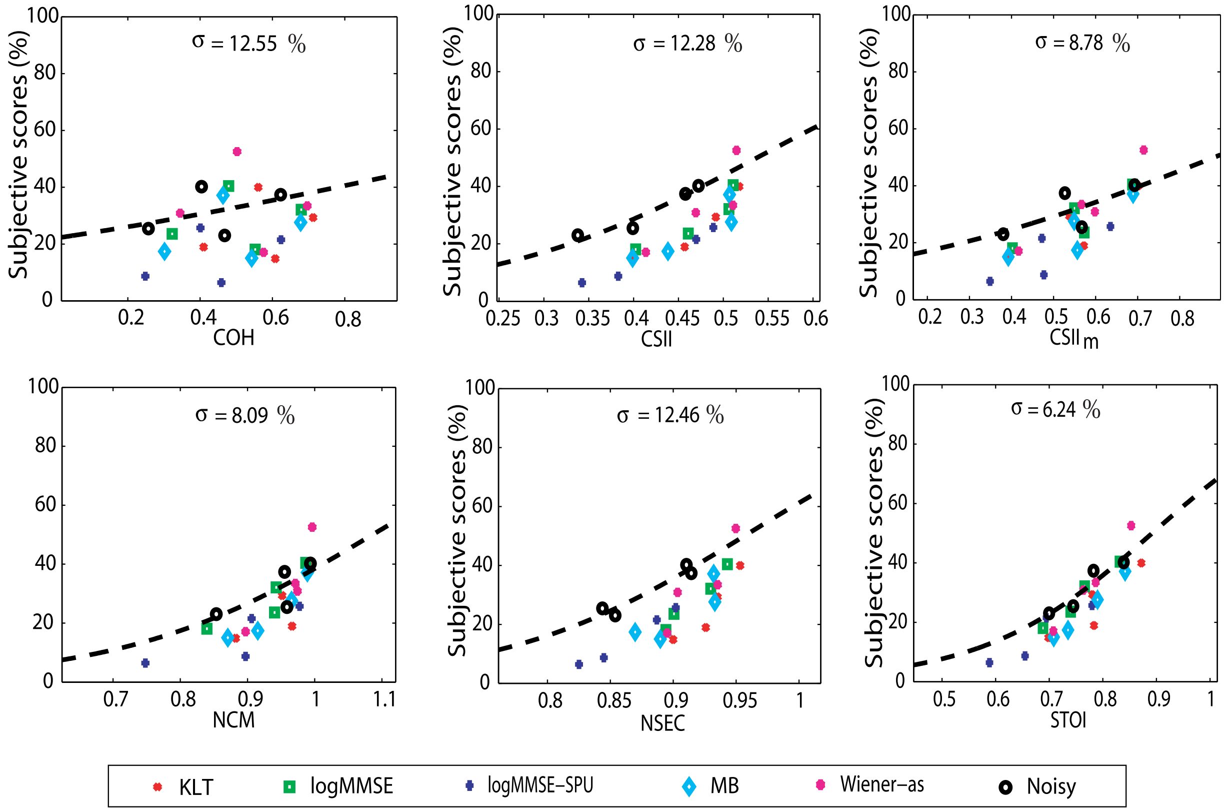

FIG. 5. (Color online) Scatter plots of the subjective intelligibility ratings against the objectively predicted scores for Japanese, along with the mapping results (dashed curves) and the RMSE results (σ). Each point on this figure represents one pair of the subjective intelligibility rating and the objectively predicted score in one tested condition.

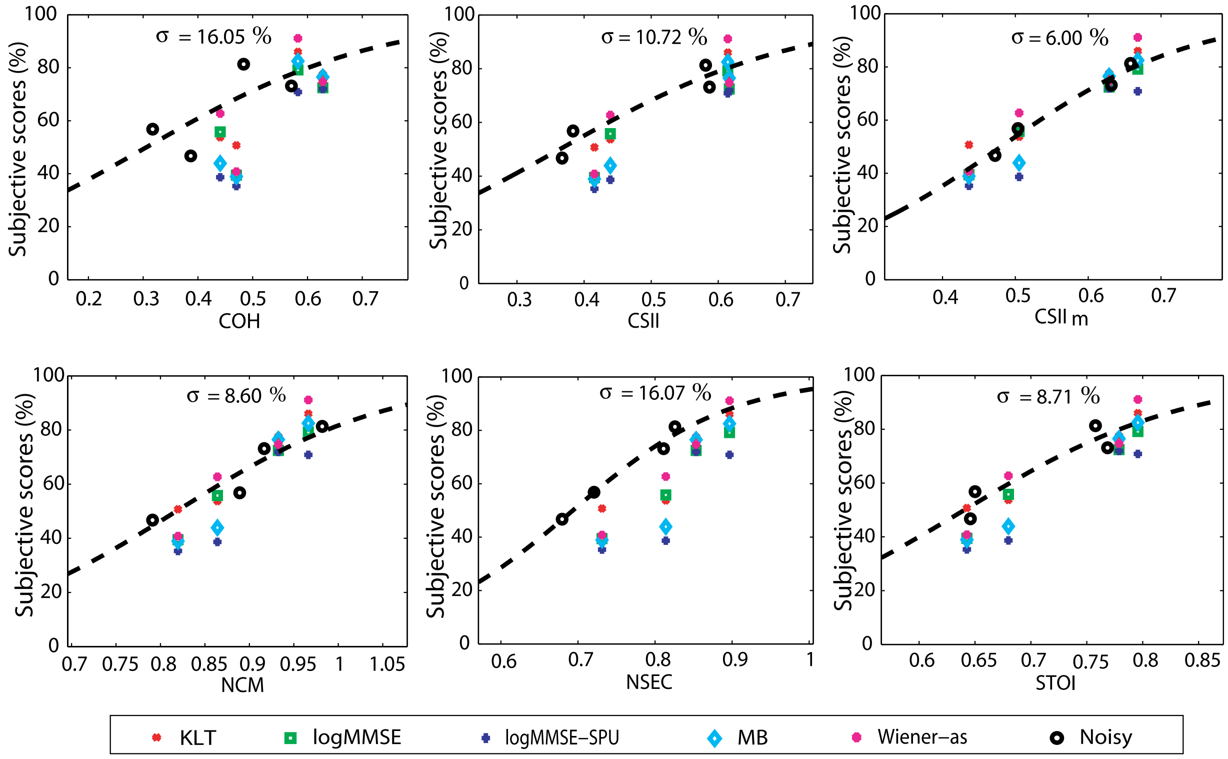

FIG. 6. (Color online) Scatter plots of the subjective intelligibility ratings against the objectively predicted scores for English, along with the mapping results (dashed curves) and the RMSE results (σ). Each point on this figure represents one pair of the subjective intelligibility rating and the objectively predicted score in one tested condition.

processing on English speech intelligibility. In contrast, the other objective measures (including the CSII, NSEC, COH measures) overestimated the effect of non-linear noise-reduction processing on speech intelligibility in most tested conditions for the three languages.

2. Discussion

Of the objective measures tested, three measures (NCM, STOI, and CSII$_m$) were much better at predicting the effects of the non-linear noise-reduction processing on speech intelligibility for the three languages. The increased ability might be due to the envelope cues involved in the computation of the NCM and STOI measures, and the normalization and constraint of the SDR (especially in the middle range of the SDR) and the use of the spectral transition and the signal-dependent BIF in the computation of the CSII$_m$ measure. These cues were important in predicting speech intelligibility even after non-linear noise-reduction processing (Drullman, 1995; Kates and Arehart, 2005; Ma *et al.*, 2009; Taal *et al.*, 2011). The NSEC and COH measures yielded the highest RMSEs between the objectively predicted intelligibility scores and the mapped subjective ratings for the three languages. That is, they were less able to predict the effect of noise-reduction processing. This result was partially attributed to their low correlation with the subjective intelligibility ratings. The performance of the other measures was found to lie in between.

For English, the CSII$_m$ measure demonstrated the lowest average RMSE. In its computation, the SDR was normalized and further constrained, and the amplitude range was also constrained to [0,10] dB below the overall RMS level as suggested for English (Kates and Arehart, 2005). These processes might not be appropriate for Chinese and Japanese where the RMSEs were shown to be merely modest. For Chinese, the NCM measure yielded the best ability (the lowest RMSEs) in predicting the effect of non-linear noise-reduction processing in speech intelligibility, which was followed by the STOI measure. The advantage of the NCM measure could be attributed to the signal-dependent BIF that places weight to each SNR value in proportion to the excess masked signal, and to that the NCM measure accounts for the low-frequency (<12.5 Hz) envelope modulations that are known to carry critically important information about speech (Drullman, 1995; Arai *et al.*, 1996). For Japanese, the best ability was introduced by the STOI measure, which was mainly due to the use of the temporal envelope in each subband (Arai *et al.*, 1996).

V. GENERAL DISCUSSION

The evaluation results of the present research indicated that among the objective measures tested, three measures (STOI, NCM, and CSII$_m$) consistently showed the best, at least good, ability in speech intelligibility prediction for three languages. The benefits of the NCM and STOI measures are mainly attributed to the temporal envelope information used in the calculation of these two measures, which has been found to play an important role in understanding speech for English (Drullman, 1995; Shannon *et al.*, 1995), Chinese (Fu *et al.*, 1998), and Japanese (Arai *et al.*, 1996). The good ability of the CSII$_m$ measure in speech intelligibility might be introduced by the normalization of the SDR and the constrained amplitude range. More importantly, it was found that most of the objective measures tested performed differently for different languages. The NCM measure exploits the low-frequency (<12.5 Hz) envelope modulations that are known to carry critically important information about speech (Drullman, 1995). The pitch information in Chinese, which mainly lies in the low frequencies, is robust against non-linear noise-reduction processing (Li *et al.,* 2011) and contributes to distinguish Chinese words (Fu *et al.,* 1998). The ability in speech intelligibility prediction of the NCM measure is further enhanced by the signal-dependent BIF ($W^{(2)}$, $p=4.0$) that places weight to each SNR value in proportion to the excess masked signal, especially when non-linear noise-reduction processing is involved. Though the CSII$_m$ measure showed the highest correlation with subjective intelligibility ratings for English, the constraint of amplitude range to [0 10] dB below the overall RMS level might not be appropriate for Chinese and Japanese. The CSII$_m$ measure was further enhanced by the signal-dependent BIF (W_4, $p = 4$) that accounts for a high number of transitions between consonants and vowels in the middle-level range (Kates and Arehart, 2005) which are crucial for understanding English speech especially in noise (Furui, 1986).

This study aims to examine the power of objective intelligibility prediction measures for noise-reduced speech across three different languages (Chinese, Japanese, and English). It is impossible to use the same database in different languages. Due to the different characteristics (e.g., syllable-based or mora-based, tonal or non-tonal) of each language, it is also difficult to exploit the databases that have the same type of speech materials. As described in Sec. II, therefore, three different databases were adopted for intelligibility testing in this present research for Chinese, Japanese, and English. Among these databases, there were some common features (e.g., the phonetically balanced materials, low word predictability, and the temporal and spectral cues of speech) as well as some different features, such as, the different words in the three languages. Though these features are related to speech intelligibility, most of them do not change after non-linear noise-reduction processing, such as word predictability and the material types. It is mainly the temporal and spectral cues of speech that are affected by non-linear noise-reduction processing. Furthermore, these temporal and spectral cues (e.g., pitch cues) that play different roles in understanding speech for different languages are all contained in the three databases adopted in this research. Using these three databases, therefore, it is possible to investigate the effect of languages (especially driven by the speech cues) on the power of the objective speech intelligibility prediction measures for non-linear noise-reduction processing. On the other hand, the strong context information of the conversational speech used in daily life helps listeners to understand speech even in noise and when processed by non-linear noise-reduction algorithms. While the three databases with low word predictability adopted in the present study did not account for the function of context information in understanding speech, it is possible for future to further examine objective speech intelligibility prediction measures for conversational speech across multiple languages.

VI. CONCLUSION

In the present study, the performance of six objective measures was evaluated in terms of predicting speech intelligibility for three languages (Chinese, Japanese, and English) before and after non-linear noise-reduction processing. For each language, the objective measures were tested under a total of 24 conditions which included noise-corrupted signals and noise-reduced signals produced by five typical single-channel noise-reduction algorithms. The performance of the objective measures was assessed by checking the relation of the objective prediction scores and the subjective intelligibility ratings in terms of the correlation coefficient and the standard deviation of the error in all conditions, and by checking their ability in predicting the effect of noise-reduction processing in speech intelligibility in terms of the RMSE. The distinct contributions of the present work include the following.

(1) The signal-dependent BIFs greatly improved the performance of both sets of CSII and NCM measures for the three languages, which was consistent with the results in Ma *et al.* (2009). This outcome clearly suggested that the traditional CSII measure as well as the NCM measure could benefit from the use of signal-dependent BIFs.

(2) The COH measure yielded the worst performance in predicting speech intelligibility for three languages. The best overall ability in predicting speech intelligibility was found as the STOI measure ($r=0.90$) for Chinese, the BIF-modified NCM measure ($r=0.84$) for Japanese, and the BIF-modified CSII$_m$ measure ($r=0.94$) for English.

(3) Of all the objective measures, those that performed best in predicting the effect of non-linear noise-reduction processing in speech intelligibility were the NCM measure incorporating the signal-dependent BIF ($\sigma=6.38\%$) for Chinese, the STOI measure ($\sigma=6.24\%$) for Japanese, and the BIF-modified CSII$_m$ measure ($\sigma=6.00\%$) for English. The other objective measures usually overestimated the effect of non-linear noise-reduction processing in most tested conditions for the three languages.

(4) Most of the objective measures examined performed differently for different languages. For example, the CSII$_m$ and NCM measures demonstrated exceptionally high accuracy in predicting speech intelligibility ($r>0.9$) for English, and was found to predict speech intelligibility for Chinese and Japanese ($r=0.79$–0.87) modestly well. The STOI measure performed well in speech intelligibility prediction for Chinese and English, but only modestly for Japanese. The performance differences in predicting the effect of non-linear noise-reduction processing on speech intelligibility were also observed across different languages for most objective measures. For instance, the objective measure performing best was given by the NCM measure for Chinese, the STOI measure for Japanese, and the CSII$_m$ measure for English.

ACKNOWLEDGMENTS

The authors thank Professor Y. Hu and Professor P. C. Loizou for their valuable discussions on this research and sharing the subjective intelligibility evaluation data for English. This work was partially supported by the National 973 Program (Grant No. 2013CB329302), the National Natural Science Foundation of China (Grant No. 11461141004), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500), and the National 863 Program (Grant No. 2012AA012503).

Amano, S., Sakamoto, S., Kondo, T., and Suzuki, Y. (**2009**). "Development of familiarity-controlled word lists 2003 (FW03) to assess spoken-word intelligibility in Japanese," Speech Commun. **51**, 76–82.

ANSI (**1997**). *Methods for Calculation of the Speech Intelligibility Index* (American National Standards Institute, New York).

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (**1996**). "Intelligibility of speech with filtered time trajactories of spectral envelopes," in *International Conference on Spoken Language (ICSLP)*, pp. 2490–2493.

Arehart, K., Kates, J., Anderson, M., and Harvey, L. (**2007**). "Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **122**, 1150–1164.

Banziger, T., and Scherer, K. (**2005**). "The role of intonation in emotional expressions," Speech Commun. **46**, 252–267.

Boldt, J., and Ellis, D. (**2009**). "A simple correlation-based model of intelligibility for nonlinear speech enhancement and separation," in *European Signal Processing Conference*, pp. 1849–1853.

Brown, C., and Bacon, S. (**2010**). "Fundamental frequency and speech intelligibility in background noise," Hear. Res. **266**, 52–59.

Chen, F., and Loizou, P. (**2012**). "Contribution of cochlea-scaled entropy versus consonant-vowel boundaries to prediction of speech intelligibility in noise," J. Acoustic. Soc. Am. **131**, 4104–4113.

Cohen, I., and Berdugo, B. (**2001**). "Speech enhancement for non-stationary noise environments," Sign. Process. **81**, 2403–2418.

Drullman, R. (**1995**). "Temporal envelope and fine structure cues for speech intelligibility," J. Acoust. Soc. Am. **97**, 585–592.

Ephraim, Y., and Malah, D. (**1985**). "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," IEEE Trans. Acoust. Speech Audio Process. **33**, 443–445.

French, N., and Steinberg, J. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.

Fu, Q., Zeng, F., Shannon, R., and Soli, S. (**1998**). "Importance of tonal envelope cues in Chinese speech recognition," J. Acoust. Soc. Am. **104**, 505–510.

Furui, S. (**1986**). "On the role of spectral transition for speech perception," J. Acoustic. Soc. Am. **80**, 1016–1025.

Goldsworthy, R., and Greenberg, J. (**2004**). "Analysis of speech-based speech transmission index methods with implications for nonlinear operations," J. Acoust. Soc. Am. **116**, 3679–3689.

Havelock, D., Kuwano, S., and Vorlander, M. (**2009**). *Handbook of Signal Processing in Acoustics* (Springer, New York), pp. 197–204.

Hisch, H., and Pearce, D. (**2000**). "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *ISCA Tutorial and Research Workshop ASR* 2000, pp. 29–32.

Hollube, I., and Kollmeier, K. (**1996**). "Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model," J. Acoust. Soc. Am. **100**, 1703–1715.

Houtgast, T., and Steeneken, H. (**1984**). "A multi-language evaluation of the RASTI method for estimating speech intelligibility in auditoria," Acta Acust. united Ac. **54**, 185–199.

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

Hu, Y., and Loizou, P. (**2003**). "A generalized subspace approach for enhancing speech corrupted by collored noise," IEEE Trans. Acoust. Speech Audio Process. **11**, 334–341.

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

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **17**, 225–246.

Kamath, S., and Loizou, P. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 4164–4167.

Kates, J., and Arehart, K. (**2005**). "Coherence and the speech intelligibility index," J. Acoust. Soc. Am. **117**, 2224–2237.

Kryter, K. D. (**1962**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698–1706.

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

Liu, C., Azimi, B., Bhandary, M., and Hu, Y. (**2014**). "Contribution of low-frequency harmonics to Mandarin Chinese tone identification in quiet and six-talker babble background," J. Acoust. Soc. Am. **135**, 428–438.

Liu, W., Jellyman, K., Evans, N., and Mason, J. (**2006**). "Assessment of objective quality measures for speech intelligibility estimation," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 1225–1228.

Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Taylor Francis Group, Boca Raton, FL), Chaps. 5–9.

Luo, X., and Fu, Q. (**2006**). "Contribution of low-frequency acoustic information to Chinese speech recognition in cochlear implant simulations," J. Acoustic. Soc. Am. **120**, 2260–2266.

Ma, D., and Shen, H. (**2004**). *Acoustic Manual* (Chinese Science Publisher, Beijing), Chap. 19.

Ma, J., Hu, Y., and Loizou, P. (**2009**). "Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions," J. Acoust. Soc. Am. **125**, 3387–3405.

Scalart, P., and Filho, J. (**1996**). "Speech enhancement based on *a priori* signal to noise estimation," in *International Conference on Acoustics, Speech, and Signal Processing*, pp. 629–632.

Shannon, R., Zeng, F., Kamath, V., Wygonski, J., and Ekelid, M. (**1995**). "Speech recognition with primarily temporal cues," Science **270**, 303–304.

Steeneken, H., and Houtgast, T. (**1980**). "A physical method for measuring speech transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Taal, C., Hendriks, R., Heusdens, R., and Jensen, J. (**2011**). "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," IEEE Trans. Audio Speech Lang. Process. **19**, 2125–2136.

Trask, R. (**1998**). *Key Concepts in Language and Lingustics* (Routledge, London), pp. 15–30.

Wang, D., and Brown, G. (**2006**). *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications* (Wiley/IEEE Press, Hoboken, NJ).

Yamada, T., Kumakura, M., and Kitawaki, N. (**2006**). "Word intelligibility estimation of noise-reduced speech," in *Interspeech*, pp. 169–172.