Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English[a)]

Junfeng Li,[b)] Lin Yang, Jianping Zhang, and Yonghong Yan
Institute of Acoustics, Chinese Academy of Sciences, Beijing, 100190, China

Yi Hu
Department of Electrical Engineering and Computer Science, University of Wisconsin–Milwaukee, Milwaukee, Wisconsin 53201

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

Philipos C. Loizou
Department of Electrical Engineering, The University of Texas–Dallas, Richardson, Texas 75083

(Received 6 May 2010; revised 6 March 2011; accepted 7 March 2011)

A large number of single-channel noise-reduction algorithms have been proposed based largely on mathematical principles. Most of these algorithms, however, have been evaluated with English speech. Given the different perceptual cues used by native listeners of different languages including tonal languages, it is of interest to examine whether there are any language effects when the same noise-reduction algorithm is used to process noisy speech in different languages. A comparative evaluation and investigation is taken in this study of various single-channel noise-reduction algorithms applied to noisy speech taken from three languages: Chinese, Japanese, and English. 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 five single-channel noise-reduction algorithms. The processed signals were finally presented to normal-hearing listeners for recognition. Intelligibility evaluation showed that the majority of noise-reduction algorithms did not improve speech intelligibility. Consistent with a previous study with the English language, the Wiener filtering algorithm produced small, but statistically significant, improvements in intelligibility for car and white noise conditions. Significant differences between the performances of noise-reduction algorithms across the three languages were observed. © *2011 Acoustical Society of America*.
[DOI: 10.1121/1.3571422]

PACS number(s): 43.71.Hw, 43.71.Es, 43.71.Gv, 43.72.Dv [SSN] Pages: 3291–3301

I. INTRODUCTION

In everyday listening conditions, speech signals are often corrupted by various types of background noise. In the past several decades, many studies on speech perception in noise have been conducted to examine the effects of noise on speech recognition (Gelfand, 1986; Leek *et al*., 1987; Parikh and Loizou, 2005). Speech recognition generally declines in conditions wherein background interference is present. Hearing loss further deteriorates speech understanding in noisy listening conditions (Helfer and Wilber, 1990).

In order to mitigate the effects of background noise and facilitate speech recognition in noise, a variety of single-channel noise-reduction algorithms have already been reported (Benesty *et al*., 2009; Chen *et al*., 2006; Loizou, 2007). Typical single-channel noise-reduction algorithms include subspace-based, statistical-model-based, spectral subtractive, and Wiener-type algorithms (Chen *et al*., 2007; Loizou, 2007). Most of these noise-reduction algorithms were designed upon certain mathematical criteria (or in some cases, on empirical and heuristic rules), and proven effective in suppressing background noise and in improving speech quality (Benesty *et al*., 2009; Hu and Loizou, 2007b). Speech quality is related to how natural (free of distortion) speech sounds, while speech intelligibility is related to the number of words that are recognized correctly by the listener. It is of great interest to know whether these existing noise-reduction algorithms improve or degrade speech intelligibility, something that is especially motivated by hearing aid and prostheses applications. Only a few of these algorithms were evaluated using intelligibility listening tests (Montgomery and Edge, 1988; Niederiohn and Groteluschen, 1976; Simpson *et al*., 1990). Among the few studies, the study by Hu and Loizou (2007a) investigated the ability of different noise-reduction algorithms in enhancing speech intelligibility using an English corpus. Their evaluation results showed that for the most part, no algorithm produced significant improvements in speech intelligibility compared with unprocessed noisy signals for English in all noisy conditions. One algorithm (Wiener filter) did produce significant improvements in intelligibility, but only in car noise

[a)]Part of this work was done when Dr. Junfeng Li was an assistant professor at Japan Advanced Institute of Science and Technology.
[b)]Author to whom correspondence should be addressed. Electronic mail: junfeng.li.1979@gmail.com

conditions. Another study demonstrated that it is possible to develop single-channel noise-reduction algorithms that can improve speech intelligibility, provided that these algorithms are optimized for a particular noisy background (Kim *et al.*, 2009).

Most of the above studies were performed using 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 structures (Trask, 1998). 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 (as carried in the F0 contour) 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 (Trask, 1998). In contrast, the 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. F0 information, however, can be used by listeners to segregate the target talker in competing-talker listening tasks.

The differences between languages have been extensively studied in the literature, particularly in the context of speech recognition in noise (Fu *et al.*, 1998; Xu *et al.*, 2005; Kang, 1998; Uchida *et al.*, 1999) and the contributions of envelope and fine-structure cues (Fu *et al.*, 1998; Shannon *et al.*, 1995; Xu and Pfingst, 2003). Shannon *et al.* (1995) examined the role of temporal envelope information in speech recognition by vocoding speech into a small number of channels so as to preserve temporal-envelope cues while discarding fine-structure cues. Following the same approach, Fu *et al.* (1998) investigated the role of temporal envelope information in Chinese and showed a high level of tone recognition (about 80% correct) despite the limited amount of spectral information available. A significant difference was found between English and Chinese sentence recognition when no-spectral information was available. Investigations of the relative contribution of temporal envelope and fine structure to speech intelligibility of English and Mandarin tone recognition demonstrated that the fine structure cues play a less important role than envelope cues in speech recognition of English (Drullman, 1995) and that fine-structure information is more important for Mandarin lexical-tone recognition, particularly in noise (Xu and Pfingst, 2003).

Concerning speech intelligibility of Japanese, the effect of filtering the time trajectories of spectral envelopes was examined, and was found that speech intelligibility was not severely impaired when the filtered cepstral coefficients fell between certain rates-of-change limits (Arai *et al.*, 1996, 1999). More worthwhile to note, when demonstrating the effectiveness of the rapid speech transmission index (RASTI), Houtgast and Steeneken (1984) tested the RASTI across ten Western languages and showed that language-specific effects could result in disparity among diverse tests. Concerning the comparisons of speech intelligibility among different languages, Kang (1998) performed a series of intelligibility tests in two different enclosures and reported that speech intelligibility of English is considerably better than intelligibility of Mandarin speech when speech was presented in noisy backgrounds. Moreover, Uchida *et al.* (1999) examined the effects of language characteristics on speech intelligibility of English and Japanese, and showed that speech intelligibility of Japanese is much higher than that of English under noisy conditions.

It is clear from the above-mentioned studies that speech recognition exhibits great variation across different languages. Therefore, it is important to examine the effects of language on noise-reduction algorithms by assessing the performance in speech intelligibility of noise-reduction algorithms for different languages in various noisy conditions. Accordingly, the present study first investigates the performance of four major classes of single-channel noise-reduction algorithms including subspace-based, statistical-model-based, spectral subtractive, and Wiener-type algorithms. Phonetically-balanced Chinese words and Japanese words are corrupted by three different types of noise (white, car, babble), and further processed by the above noise-reduction algorithms. The processed signals are presented to native Chinese and Japanese speakers, respectively, for word identification. Subsequently, the differences in the performance of noise-reduction algorithms, as applied to different languages, are examined in terms of their potential effects on the temporal envelope and F0 contour.

The contributions of this present research are as follows. First, to our knowledge it is the first comprehensive evaluation in speech intelligibility of noise-reduction algorithms for Chinese and Japanese languages. These evaluations will help us understand which algorithm(s) preserves or enhances speech intelligibility compared with that of unprocessed signals for Chinese and Japanese. Second, and more importantly, this study will assess performance differences, if any, in speech intelligibility of existing noise-reduction algorithms when applied to different languages. Third, it sets to examine the effects of language-specific features on the performance of noise-reduction algorithms for different languages. Finally, it will provide us with valuable information and insight regarding which algorithm is appropriate for each of the three languages examined.

II. EXPERIMENT I: INTELLIGIBILITY INVESTIGATION OF SINGLE-CHANNEL NOISE-REDUCTION ALGORITHMS FOR MANDARIN CHINESE

A. Methods

1. Subjects

Ten normal-hearing listeners (five females and five males) participated in this experiment. All subjects were native speakers of Mandarin Chinese, and were paid for their participation. The subjects' age ranged from 23 to 31 years old, with all being post-graduate students from Institute of Acoustics, Chinese Academy of Sciences.

2. Stimuli

In the intelligibility evaluations for Mandarin (a type of spoken Chinese called *Putonghua*), the syllable database for intelligibility test reported by Ma and Shen (2004), which has been established as the Chinese national standard (GB/T15508-1995) (Ma and Shen, 2004), were adopted as speech material in this experiment. This set of test material consists of 10 syllable tables, each of which contains 75 phonetically-balanced (PB) Mandarin syllables with Consonant–Vowel (CV) structure. In each table every three syllables are combined randomly to form nonsense sentences. Consequently, every table can produce the sentence lists with 25 nonsense sentences to be used in listening tests. The sentence lists were uttered by one female speaker and recorded in a sound-proof booth at a sampling rate of 16 kHz and stored in a 16-bit format, and then down-sampled to 8 kHz.

Three types of background noises, including white noise, car, and babble noises taken from AURORA (Pearce and Hirsch, 2000), were used in the listening experiments. To simulate the receiving frequency characteristics of telephone handsets, both speech and noise signals were filtered by the modified intermediate reference system (IRS) filters used in ITU-T P.862, which was similar to the study of Hu and Loizou (2007a). The noise segment of the same length as the speech signal was scaled accordingly to reach the desired signal-to-noise ratio (SNR) level and finally added to the filtered clean speech signal at SNRs of 0 and 5 dB.

3. Signal processing

The noise-corrupted signals were processed by five representative single-channel 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 estimate (Wiener-as) (Scalart and Filho, 1996). The reasons for selection of these five algorithms are twofold: (1) These approaches cover the state-of-the-art four major classes of noise-reduction algorithms mentioned above; (2) these algorithms yielded higher speech intelligibility of English compared with other algorithms (Hu and Loizou, 2007a).

Noise spectrum estimation was updated using a statistical-model based voice activity detector (VAD) (Sohn *et al.*, 1999), for all algorithms (except the KLT algorithm). In the KLT algorithm, the VAD algorithm given in Mittal and Phamdo (2000) was used for noise estimation. MATLAB implementations of the above noise-estimation and noise-reduction algorithms are available in Loizou (2007).

4. Procedure

Stimuli were presented to the subjects at a comfortable listening level through TDH-39 headphone and Madsen Iteral II audio meter in a sound-proof booth. Prior to the test, each subject listened to a set of sample sentences to get familiar with the testing procedure. During the test, the subjects were asked to write down the words that they heard, and tone recognition tests were conducted using a four-alternative forced-choice (4-AFC) task; no feedback was provided. Subjects participated in a total of 36 listening conditions [=2 SNR levels × 3 types of background noise × 6 algorithms (1 noisy reference + 5 noise-reduction algorithms)]. One list of sentences (i.e., 25 sentences) was used per condition, and none of the lists were repeated across conditions. Thus, each subject listened to 900 nonsense sentences (=25 sentences × 36 conditions) in the listening tests. According to the types of background noise, all listening conditions were divided into three sessions with 12 conditions per session. In each session, the presentation order of the stimuli and listening conditions was randomized for each subject.

B. Results

1. Speech recognition

The mean percentage of words correctly identified across ten subjects for five noise-reduction algorithms under three noise conditions at two SNRs is plotted in Fig. 1, in which the error bars represent the standard errors of the mean. The intelligibility scores of unprocessed noisy speech are also provided for comparison.

As shown in Fig. 1, the word recognition scores at 0 dB SNR levels were much lower than those at 5 dB SNR levels for all algorithms in the tested noise conditions. Of all the algorithms, the logMMSE-SPU algorithm provided the lowest intelligibility scores in all noise conditions. In most conditions, speech intelligibility scores of the processed signals by the tested algorithms (except for the Wiener-as algorithm) were much lower than that of the unprocessed noisy signals.

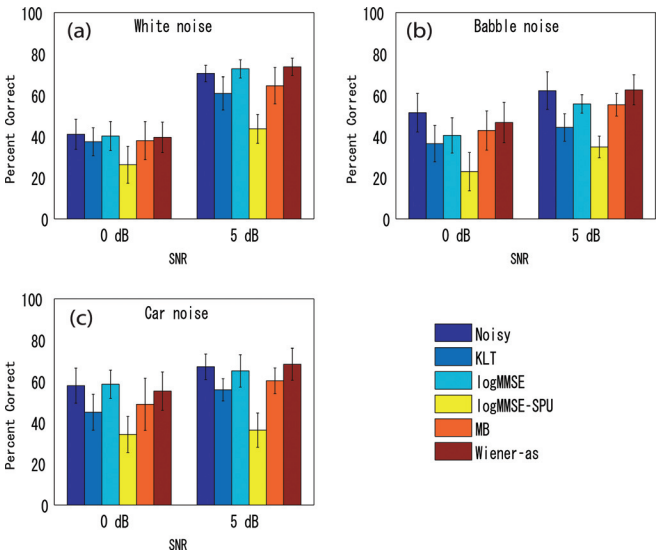

FIG. 1. (Color online) Mean percent correct scores for Mandarin Chinese under six processing conditions (5 single-channel noise-reduction algorithms +1 unprocessed noisy) in the white noise (a), babble noise (b), and car noise (c), at 0 and 5 dB SNRs.

TABLE I. Statistical comparison between the intelligibility of unprocessed noisy speech and that of processed speech by five noise-reduction algorithms for Japanese. Algorithms indicated with "E" were found to be equally intelligible to noisy speech and algorithms indicated with "L" provided lower intelligibility scores.

	0 dB			5 dB		
Algorithm	White	Babble	Car	White	Babble	Car
KLT	E	L	L	E	L	L
logMMSE	E	L	E	E	L	E
logMMSE-SPU	L	L	L	L	L	L
MB	E	L	L	L	L	L
Wiener-as	E	E	E	E	E	E

To examine the effects of SNR levels (5 and 10 dB) and processing conditions (5 noise-reduction algorithms + 1 unprocessed signal), the word intelligibility scores were subjected to statistical analysis using the score as the dependent variable, and the SNR level and processing condition as the two within-subjects factors. For white noise, two-way analysis of variance (ANOVA) with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 544.70$, $p < 0.001$] and processing conditions [$F(5, 45) = 58.28$, $p < 0.001$]. There was significant interaction between SNR level and processing conditions [$F(5, 45) = 10.24$, $p < 0.001$]. For babble noise, two-way ANOVA with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 72.71$, $p < 0.001$] and processing conditions [$F(5, 45) = 56.51$, $p < 0.001$]. There was no significant interaction between SNR level and processing conditions [$F(5, 45) = 1.90$, $p = 0.11$]. For car noise, two-way ANOVA with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 47.71$, $p < 0.001$] and processing conditions [$F(5, 45) = 89.37$, $p < 0.001$]. There was significant interaction between SNR level and processing conditions [$F(5, 45) = 3.13$, $p < 0.05$].

Following ANOVA, to further identify whether the tested algorithms significantly decreased or maintained Mandarin speech intelligibility, the *post-hoc* tests (multiple paired comparisons according to Ryan's method with appropriate correction) (Ryan, 1959, 1960) were done between the recognition score of unprocessed noisy speech and the score of processed speech by the tested noise-reduction algorithms. Difference between scores was treated as significant if the significance level $p < 0.05$. The analysis results are listed in Table I. No noise-reduction algorithms tested here could significantly improve word recognition scores for Mandarin in all conditions. Of the tested algorithms, the best performance in recognition scores was obtained by the Wiener-as algorithm which maintained the word intelligibility at the same level attained by the unprocessed noisy speech (i.e., yielded the same scores as the unprocessed speech). Considering the word recognition scores in all conditions, the logMMSE algorithm ranked second which yielded the same word intelligibility as the unprocessed noise speech under white and car noise conditions. Good performance was followed by the KLT algorithm which maintained the recognition score in white noise condition, and the MB algorithm which maintained the performance in 0 dB white noise condition. The logMMSE-SPU algorithm resulted in a significant deterioration in Mandarin speech intelligibility under all listening conditions.

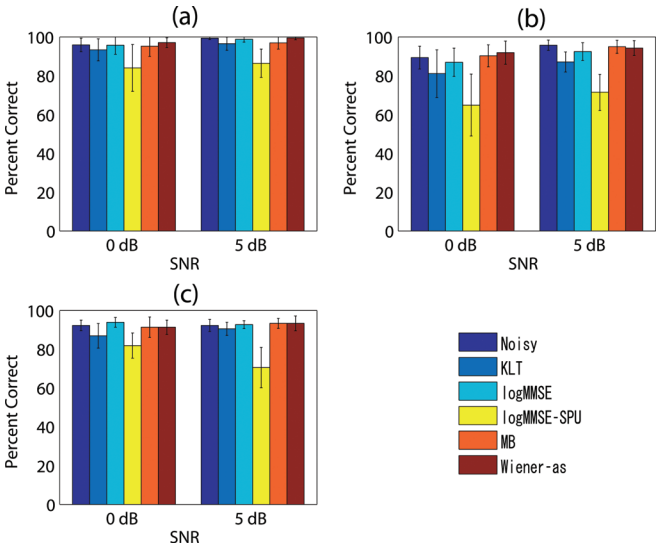

FIG. 2. (Color online) Mean percent correct scores for tone recognition in Mandarin Chinese under six processing conditions (5 single-channel noise-reduction algorithms +1 unprocessed noisy) in white noise (a), babble noise (b), and car noise (c), at 0 and 5 dB SNRs.

2. Tone recognition

The mean correct tone recognition results across subjects for five noise-reduction algorithms in different noisy conditions are shown in Fig. 2. Figure 2 demonstrates that tones in Mandarin Chinese were correctly recognized to a large degree in all tested conditions. Compared with the unprocessed noisy signal, the noise-suppressed speech signals, with the exception of the logMMSE-SPU algorithm, did not decrease severely the tone recognition scores. The logMMSE-SPU algorithm gave quite low tone recognition scores in all tested noise conditions relative to the other tested algorithms. This was consistent with the low intelligibility scores obtained with the logMMSE-SPU algorithm (see Fig. 1).

The effects of SNR level and processing conditions (noise-reduction algorithms and unprocessed signal) on tone recognition were investigated through statistical analysis. The tone recognition score was used as the dependent variable, and the SNR level and processing conditions as two within-subjects factors. Two-way ANOVA with repeated measures indicated: significant effects of SNR [$F(1, 9) = 6.15$, $p = 0.035$] and processing [$F(5, 45) = 19.18$, $p < 0.005$] in the white noise condition; significant effects of SNR [$F(1, 9) = 13.43$, $p = 0.005$] and processing [$F(5, 45) = 37.43$, $p < 0.005$] in the babble condition; non-significant effect of SNR [$F(1, 9) = 0.96$, $p = 0.353$], significant effect of processing [$F(5, 45) = 38.84$, $p < 0.0005$] and significant interaction [$F(5, 45) = 11.78$, $p < 0.005$] in the car noise condition. No significant interaction between SNR level and processing was observed in the babble and white noise conditions.

When the ANOVA revealed a significant difference, the Ryan's method with appropriate correction was used for *post hoc* pair-wise comparisons of the tone recognition scores between the unprocessed noisy speech and the processed speech by the tested algorithms. Difference between scores was treated as significant if the significance level $p < 0.05$. Results demonstrated that relative to the unprocessed noisy signals, only the logMMSE-SPU algorithm yielded significantly lower scores in tone recognition for all three types of noise tested. The other tested algorithms yielded statistically equivalent tone recognition performance as that of the unprocessed signals.

III. EXPERIMENT II: INTELLIGIBILITY INVESTIGATION OF SINGLE-CHANNEL NOISE-REDUCTION ALGORITHMS FOR JAPANESE

A. Methods

1. Subjects

Thirty normal-hearing listeners (27 males and three females) participated in this experiment. All subjects were native speakers of Japanese, and were paid for their participation. The subjects' age ranged from 23 to 36 years old, with all being post-graduate students from Japan Advanced Institute of Science and Technology.

2. Stimuli and signal processing

In the intelligibility evaluations of single-channel noise-reduction algorithms for Japanese, the words taken from the familiarity-controlled word lists 2003 (FW03) were used as speech material, which consisted of 80 lists with 50 phonetically-balanced words per list (Amano *et al.*, 2009). Because word familiarity has a strong effect on word recognition (the higher the word familiarity is, the higher the word recognition rate is), all word lists in FW03 were divided into four sets in four word-familiarity ranks. All words were recorded in a soundproof room at a 48 kHz sampling rate. In the present investigation, only the word lists with the lowest familiarity uttered by one male were used.

The noise signals in the intelligibility evaluations for Japanese were the same as those used in Mandarin intelligibility evaluations. As in the investigation of Mandarin, speech and noise stimuli were first down-sampled to 8 kHz and filtered by IRS filters, and then mixed to generate the noise-corrupted signal at SNRs of 0 and 5 dB. The noise-corrupted signals were finally processed by five single-channel noise-reduction algorithms as used in Experiment I.

3. Procedure

Thirty listeners were grouped into three panels (one panel per type of noise), with each panel consisting of ten listeners (one female and nine males). Each panel of listeners listened to words corrupted by a different type of noise, which ensures that each subject listened to each sentence once. The noisy and processed signals were presented to the subjects at a comfortable listening level through HDA-200 headphone in a sound-proof booth. Prior to the test, each subject went through a training session to become familiar with the testing procedure. In the tests, each subject participated in a total of 12 listening conditions [=2 SNR levels × 6 algorithms (5 noise-reduction algorithms + 1 unprocessed references)]. One word list of 50 words was used for each condition. Each subject listened to 600 low-familiarity words (=50 sentences × 12 conditions) in the listening test. For each panel, the presentation order of the stimuli and listening conditions were randomized across each subject. Subjects were asked to write down the words they heard.

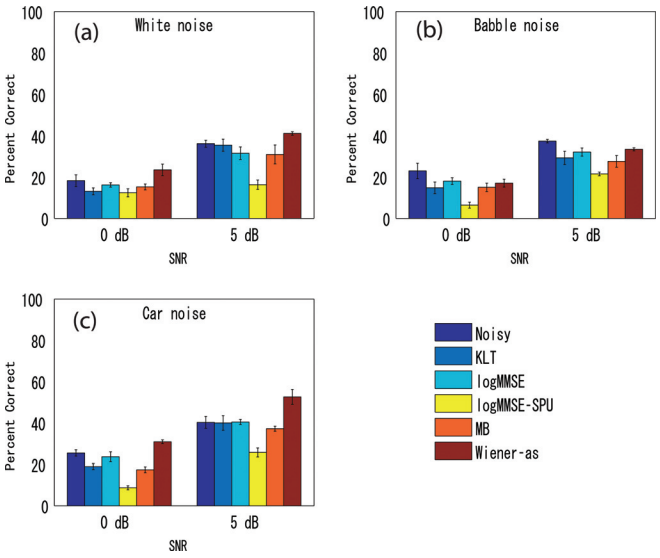

FIG. 3. (Color online) Mean percent correct scores for Japanese under six processing conditions (5 single-channel noise-reduction algorithms + 1 unprocessed noisy) in the white noise (a), babble noise (b), and car noise (c), at 0 and 5 dB SNRs.

B. Results

The mean percentage of Japanese words correctly identified across subjects for six processing conditions under three noise conditions at two SNRs is plotted in Fig. 3, in which the error bars represent the standard errors of the mean.

Similar to the Mandarin intelligibility scores, the word recognition scores of Japanese at 0 dB were also significantly lower than those at 5 dB for all processing conditions in the tested noise conditions, as indicated in Fig. 3. To examine the effects of SNR level and processing conditions, the word intelligibility scores were subjected to statistical analysis using the score as the dependent variable, and the SNR level and processing condition as the two within-subjects factors. For white noise, two-way ANOVA with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 2790.04$, $p < 0.001$] and processing conditions [$F(5, 45) = 163.28$, $p < 0.001$]. There was significant interaction between SNR levels and processing conditions [$F(5, 45) = 31.01$, $p < 0.001$]. For babble noise, two-way ANOVA with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 5691.92$, $p < 0.001$] and processing conditions [$F(5, 45) = 270.11$, $p < 0.001$]. There was significant interaction between SNR levels and processing conditions [$F(5, 45) = 3.49$, $p < 0.01$]. For car noise, two-way ANOVA

TABLE II. Statistical comparison between the intelligibility of unprocessed noisy speech and that of processed speech by five noise-reduction algorithms for Japanese. Algorithms indicated with "E" were found to be equally intelligible to noisy speech, algorithms indicated with "L" provided low intelligibility scores, and algorithms indicated with "B" improved intelligibility.

	0 dB			5 dB		
Algorithm	White	Babble	Car	White	Babble	Car
KLT	L	L	L	E	L	E
logMMSE	E	L	E	L	L	E
logMMSE-SPU	L	L	L	L	L	L
MB	L	L	L	L	L	L
Wiener-as	B	L	B	B	L	B

with repeated measures indicated significant effects of SNR levels [$F(1, 9) = 1749.75$, $p < 0.001$] and processing conditions [$F(5, 45) = 353.70$, $p < 0.001$]. There was significant interaction between SNR level and processing conditions [$F(5, 45) = 12.49$, $p < 0.001$].

To further examine whether the tested algorithms statistically improved, maintained, or reduced speech intelligibility of Japanese compared with the unprocessed noisy speech, the *post-hoc* tests (multiple paired comparisons according to Ryan's method with appropriate correction) (Ryan, 1959, 1960) were done between the intelligibility scores of unprocessed noisy speech and the scores of processed speech by the noise-reduction algorithms tested. Difference between scores was treated as significant if the significance level $p < 0.05$. The analysis results are listed in Table II. Of the tested algorithms, only the Wiener-as algorithm yielded the best performance. More precisely, statistically significant improvement was obtained in Japanese word recognition compared with the unprocessed noisy speech in white and car noise conditions. This was followed by the logMMSE algorithm that maintained word intelligibility (i.e., the same recognition score as the unprocessed noisy speech) under car noise and 0 dB white noise conditions. The KLT algorithm provided identical word intelligibility as the unprocessed speech in 5 dB white and car noises. The worst performance was obtained by the MB and logMMSE-SPU algorithms which yielded significant decreases in Japanese word recognition scores under all conditions.

IV. SUMMARY OF INTELLIGIBILITY EVALUATION OF SINGLE-CHANNEL NOISE-REDUCTION ALGORITHMS FOR ENGLISH

For completeness, we report a summary of the results taken from Hu and Loizou (2007a) on the intelligibility evaluation of single-channel noise-reduction algorithms for English. In that evaluation, the IEEE sentence database (IEEE, 1969) was used as test material and the masker signals were, among others, babble and car noise. The corrupted (at 0 and 5 dB SNR) and processed sentences were presented to native English listeners for identification. Only a subset of the conditions and algorithms tested in the study by Hu and Loizou (2007a) are reported here for comparative purposes. Because no intelligibility test was conducted in white noise for English, therefore, only intelligibility evaluation results in babble and car noise conditions are presented. The mean percentage of words identified correctly [as reported in Hu and Loizou (2007a)] is shown in Fig. 4.

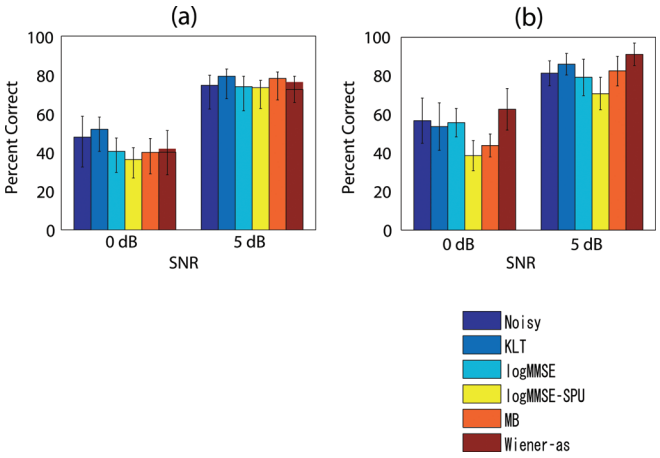

FIG. 4. (Color online) Mean percent correct scores for English [reproduced from Hu and Loizou (2007a)] for six processing conditions (5 single-channel noise-reduction algorithms +1 unprocessed noisy) in babble (a) and car noise (b), at 0 and 5 dB SNRs.

The Wiener-as algorithm maintained speech intelligibility in most test conditions and in fact improved intelligibility in the 5-dB car noise condition. Good performance was followed by the KLT, logMMSE, and MB algorithms. Lowest performance was obtained with the logMMSE-SPU algorithm. The intelligibility scores obtained with the other algorithms tested in Hu and Loizou (2007a) were comparable, or lower, to those obtained with the KLT and Wiener-as algorithms.

V. INTELLIGIBILITY COMPARISON OF NOISE-REDUCTION ALGORITHMS BETWEEN DIFFERENT LANGUAGES

It is of great interest to study the differences in performance of the various noise-reduction algorithms when applied to different languages (e.g., Chinese, Japanese, and English). Due to the fundamental differences across languages, it is unreasonable to directly compare the absolute word identification scores obtained in different languages (Kang, 1998). Alternatively, we considered comparing the performance (intelligibility score) of each algorithm in different languages relative to that of unprocessed speech.

The relative word recognition scores (difference between processed and unprocessed scores) across all tested noise-reduction algorithms are shown in Fig. 5 in two noise conditions (babble and car noises) at two SNRs (0 and 5 dB) and three languages (Chinese, Japanese, and English). Positive numbers in Fig. 5 indicate improvement in performance, while negative numbers indicate a decrement in performance relative to the baseline (unprocessed signals) performance. Because no intelligibility test was conducted in white noise for English, the intelligibility comparisons among three languages were only performed in babble and car noises. As can be seen from Fig. 5, there was a large variability in performance with the noise-reduction algorithms even when tested in the same noise conditions. In the 0 dB babble

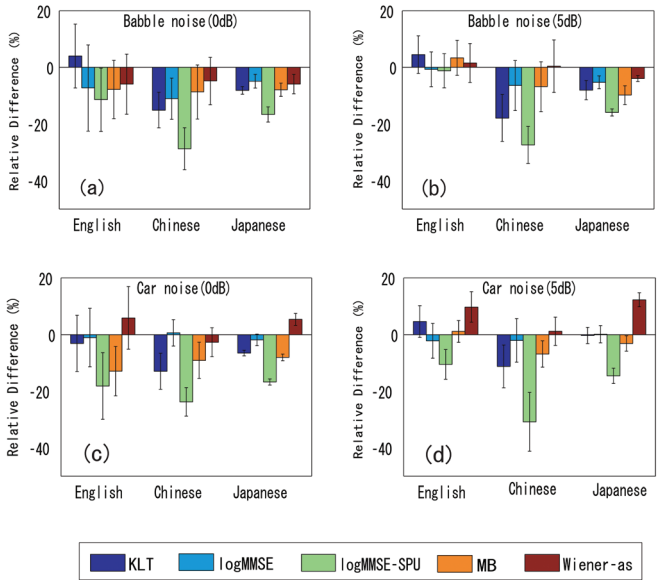

FIG. 5. (Color online) Difference scores between intelligibility of unprocessed noisy speech and intelligibility of speech processed by five noise-reduction algorithms in two types of background noise and two SNR levels. Positive numbers indicate relative improvement, and negative numbers indicate degradation.

condition, for example, the KLT algorithm showed an improvement, albeit small, in word identification compared with that of noise-corrupted speech for English. In contrast, no improvement in word recognition was obtained by the KLT algorithm in this condition for Chinese and Japanese. Moreover, although the logMMSE-SPU algorithm yielded a decrement in speech intelligibility in all tested conditions, the degree of degradation varied largely across different languages.

Three-way ANOVA was performed with the relative (difference) word identification scores as the dependent variable, and the noise type (babble and car noises), the SNR level (0 dB and 5 dB), and the processing condition (5 noise-reduction algorithms) as the within-subjects factors, and language (Chinese, Japanese, and English) as the between-subject factor. Results indicated significant effects of SNR level [$F(1, 27) = 16.7$, $p < 0.005$], noise-reduction algorithm [$F(4, 108) = 220.28$, $p < 0.005$], and noise type [$F(1, 27) = 4.96$, $p = 0.034$]. There were significant two-way interactions between language and SNR [$F(2, 27) = 4.85$, $p = 0.016$], between language and algorithm [$F(8, 108) = 22.43$, $p < 0.005$] and masker by algorithm [$F(8, 108) = 7.89$, $p < 0.005$]. The masker by language interaction was not significant [$F(2, 27) = 2.17$, $p = 0.133$]. There was a significant between-subject effect of language [$F(1, 27) = 286.12$, $p < 0.005$] on speech intelligibility scores.

The performance differences of the noise-reduction algorithms applied in the three languages were further investigated by *post-hoc* tests. The KLT and logMMSE-SPU algorithms showed significant differences ($p < 0.005$) in relative speech intelligibility among the three languages under all tested conditions. Significant differences were noted for the three languages for the MB and Wiener-as algorithms in certain conditions, but no significant differences were noted for the logMMSE algorithm in all tested conditions. Overall, these results suggest that the performance of the noise-reduction algorithms was affected by the characteristics of the language. This was confirmed with the significant language effects observed using ANOVA statistical analysis.

VI. GENERAL DISCUSSION

With the exception of the Wiener-as algorithm, the remaining noise-reduction algorithms resulted in decreased speech intelligibility (see Fig. 5) in all tested conditions for the three languages (Mandarin, Japanese, and English). The Wiener-as algorithm maintained for the most part speech intelligibility to the level attained in unprocessed (noisy) conditions in all three languages tested. Significant differences in intelligibility of noise-suppressed speech under each condition for different languages were clearly noted (see Fig. 5). Next, we discuss some factors that potentially influenced speech intelligibility. In particular, we will focus on the effects of temporal envelope and F0 contour, both of which are known to contribute to tonal language recognition (Fu *et al.*, 1998).

A. Temporal envelope

The information carried by the speech temporal envelope has been found to contribute to speech recognition not only in English (Drullman *et al.*, 1994; Shannon *et al.*, 1995) but also in Chinese (Fu *et al.*, 1998). In English, low-frequency (2–16 Hz) amplitude modulations, present in the temporal envelope, have been shown to carry important information about speech (Drullman *et al.*, 1994). In fact, some intelligibility measures [e.g., the speech transmission index (STI) measure] were designed to assess the reduction in amplitude modulations as introduced by noise and reverberation, because these reductions have been found to correlate with speech intelligibility (Houtgast and Steeneken, 1984). Consequently one would expect that corruption of the temporal envelope should produce reduction in speech intelligibility.

To investigate this, we computed the temporal envelopes of noise-suppressed signals. Because similar pattern across different utterances was observed, the example temporal envelopes of a Chinese utterance (shown at the band with center frequency of 500 Hz) are plotted in Fig. 6 for several conditions (clean, noisy, processed by the Wiener-as and logMMSE-SPU algorithms). The temporal envelope produced by the logMMSE-SPU was severely attenuated, suggesting a significant amount of distortion. On this regard, this explains the relatively low recognition scores obtained by the logMMSE-SPU algorithm. In contrast, the temporal envelope was preserved for the most part (except at $t = 0.5$ s) by the Wiener-as algorithm. This partly explains as to why the Wiener-as algorithm preserved speech intelligibility in most conditions. A similar pattern was observed for Japanese, as shown in Fig. 7. Thus, the degree to which the temporal envelope of speech is affected or modified due to masking noise reflects the degree that intelligibility is affected. This is consistent with previous results reported by

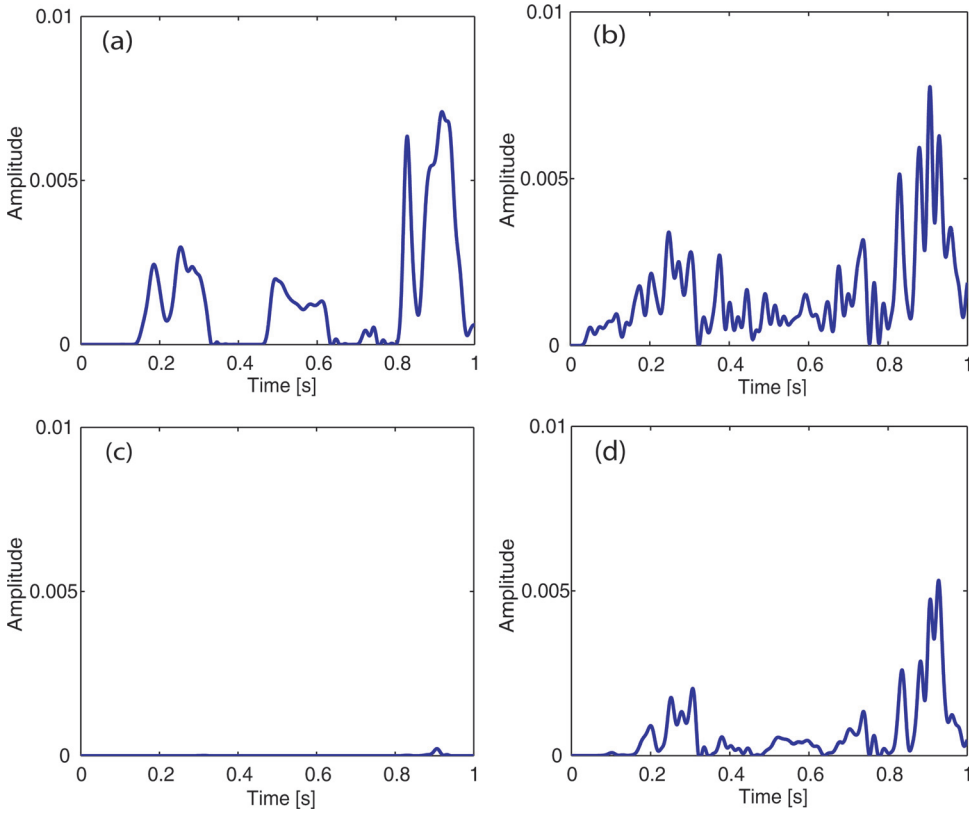

FIG. 6. (Color online) Temporal envelopes (at band with center frequency of 500 Hz) of a Chinese utterance "Zi Shi Ku" in the clean condition (a), unprocessed noisy signal in babble at 0 dB (b), processed signal by the logMMSE-SPU algorithm (c), and processed signal by the Wiener-as algorithm (d).

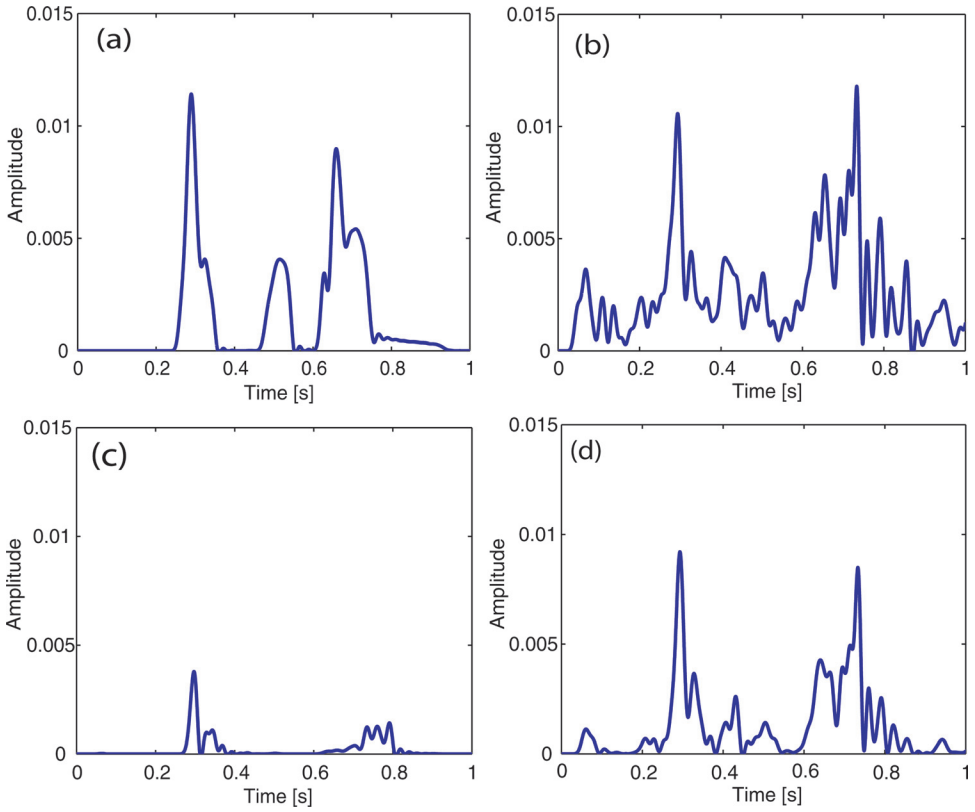

FIG. 7. (Color online) Temporal envelopes (at band with center frequency of 500 Hz) of the Japanese utterance "Ro Ku Ten" under clean condition (a), unprocessed noisy signal in babble at 0 dB (b), processed signal by the logMMESE-SPU algorithm (c), and processed signal by the Wiener-as algorithm (d).

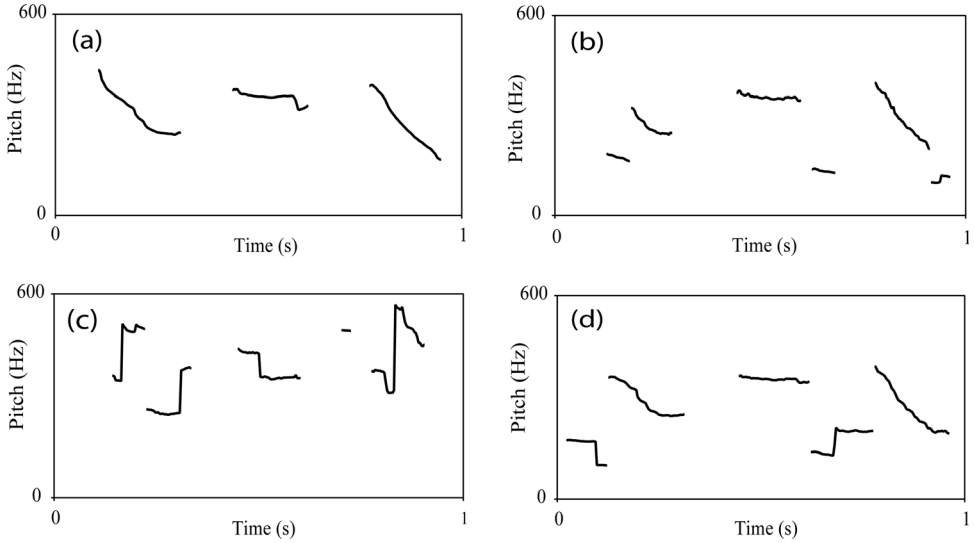

FIG. 8. F0 contours of the Chinese utterance "Zi Shi Ku" under clean condition (a), unprocessed noisy signal in babble at 0 dB (b), processed signal by the logMMSE-SPU algorithm (c), and processed signal by the Wiener-as algorithm (d).

Houtgast and Steeneken (1984); Drullman *et al*. (1994); Shannon *et al*. (1995); Fu *et al*. (1998).

B. F0 contour

Different languages are characterized by different features. For example, Mandarin Chinese utilizes pitch (F0 contour) to distinguish lexical items. English, on the other hand, does not use pitch distinctively to convey lexical information. In Japanese, the specification of some accent location(s) is sufficient to predict the tone configuration of the entire word. This inherent difference in the role of pitch information across languages can be physically described using the F0 contour.

We thus considered examining the F0 contours following noise reduction. A severe corruption of the F0 contour would suggest a reduction in tone recognition, and subsequently a reduction in word recognition for Chinese. As an example, the F0 contour of a Chinese utterance (same as in Fig. 6), is plotted in Fig. 8. The Wiener-as algorithm is able to recover part of the F0 contour of noise-corrupted speech. Hence, the F0 contour was preserved when the Wiener-as algorithm was used, suggesting that the subjects were able to identify accurately the four tones. This finding is consistent with the tone recognition data shown in Fig. 2. High tone recognition was maintained when the Wiener-as algorithm was used. In contrast, the logMMSE-SPU algorithm significantly damaged the F0 contour of speech signal. This is in line with the tone-recognition data shown in Fig. 2. A decrement in tone recognition was observed when the logMMSE-SPU algorithm was used, explaining the lower word identification scores obtained by the logMMSE-SPU algorithm (see Fig. 1).

The F0 contour of a Japanese utterance (same as in Fig. 7) is shown in Fig. 9. The background noise seems to damage the F0 contour of Japanese speech to a different degree compared with Chinese. Both the Wiener-as and logMMSE-SPU algorithms failed to recover the F0 contour of speech signal from the noise-corrupted signal. Despite this failure, the Wiener-as algorithm performed quite well and in some

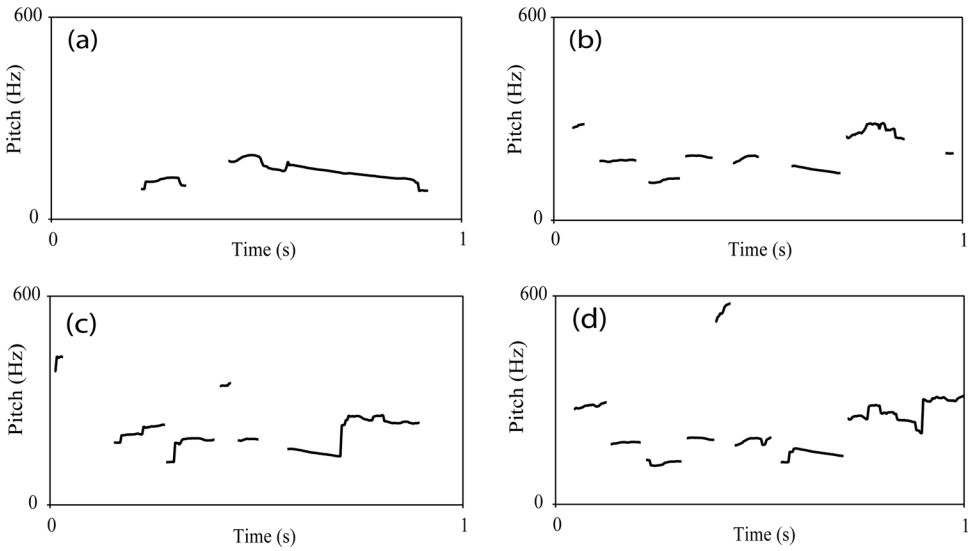

FIG. 9. F0 contours of the Japanese utterance "Ro Ku Ten" under clean condition (a), unprocessed noisy signal in babble at 0 dB (b), processed signal by the logMMSE-SPU algorithm (c), and processed signal by the Wiener-as algorithm (d).

cases improved intelligibility (see car noise condition in Fig. 3). We believe that this is possibly due to the fact that the role of F0 cue in Japanese speech recognition is much weaker than that in Chinese (Hasegawa, 1999).

VII. CONCLUSIONS

In this paper, two experiments were carried out in order to investigate the intelligibility of noise-corrupted signals processed by five conventional noise-reduction algorithms for Mandarin and Japanese, under three types of noise and two SNR levels. The noises used in the experiments were white noise, babble, and car noise. The results obtained for Mandarin and Japanese were compared with those obtained for English and reported by Hu and Loizou (2007a). Based on these results, the following findings were shared across the three different languages examined:

(1) In severe noise conditions, most single-channel noise-reduction algorithms were unable to recover or enhance the weak consonants (unvoiced speech), resulting in low word recognition scores. The weak consonants (e.g., /t/) in particular are known to be masked by noise more easily than the vowels (Parikh and Loizou, 2005), making it extremely challenging to enhance or fully recover them in noisy conditions. The inability of noise-reduction algorithms to recover (at least to some extent) the weak consonants contributes partially to the lack of intelligibility improvement.
(2) The word recognition scores of speech processed in babble by most noise-reduction algorithms were much lower than those in car noise for all languages. This is mainly due to the fact that babble, compared with car noise, is non-stationary and as such it is difficult to estimate or track its spectrum.
(3) No noise-reduction algorithm improved significantly word recognition scores for Chinese in any condition, and only the Wiener-as algorithm provided small, but statistically significant, improvements for Japanese in white and car noise conditions. This was consistent with the outcome in English.
(4) Detailed analysis was conducted to isolate the factors contributing to reduction in speech intelligibility obtained by some algorithms (e.g., logMMSE-SPU). Results indicated that the logMMSE-SPU algorithm is unable to recover accurately the temporal envelope of the target signal, showing evidence of severe attenuation (see Figs. 6 and 7). Further analysis indicated that the logMMSE-SPU algorithm does not preserve accurately the F0 contour, a finding consistent with the tone recognition scores (Fig. 2). In view of this, we conclude that the reduction in intelligibility by the logMMSE-SPU algorithm can be attributed to the corrupted envelopes and the lack of preservation of the F0 contour, which is needed for tonal language recognition. In contrast, the Wiener-as algorithm was able to recover, at least to some extent, the temporal envelope and preserved the F0 contour (see Figs. 6–9). This finding is consistent with the fact that the Wiener-as algorithm preserved speech intelligibility in most conditions (small improvements were noted in some conditions in Japanese).
(5) Considering the conditions examined, the Wiener-as algorithm performed the best in that it either maintained word recognition or provided improved word recognition scores compared with the unprocessed noisy speech in the tested conditions (except 0 dB for Chinese). The logMMSE algorithm ranked second in speech intelligibility, followed by the KLT and MB algorithms. The logMMSE-SPU algorithm yielded the worst performance as it decreased the speech recognition scores in all tested conditions. This ranking of different noise-reduction algorithm in terms of speech recognition score is consistent for Chinese, Japanese, and English
(6) The high speech recognition scores obtained by the Wiener-as algorithm in all three languages can be attributed to the fact that it produces little speech distortion at the cost of low noise suppression. On the other hand, the logMMSE-SPU algorithm attenuated a large amount of noise components and introduced severe speech distortion. It is believed that the speech distortions introduced severely damaged the obstruent consonants that are weak in intensity, making it difficult to recognize these sounds. Between these two extremes of noise suppression/speech distortion, the logMMSE, KLT, and MB algorithms provided moderate noise reduction and relative low speech distortion that placed their word recognition scores in between the two extremes.

The second primary focus of the present study was to identify differences in speech intelligibility of noise-reduction algorithms resulting from perceptual differences between the three languages. In summary, we can draw the following conclusions regarding the effects of language:

(1) The differences in relative word recognition score for different languages demonstrated that most noise-reduction algorithms (except for the logMMSE algorithm) were significantly affected by the characteristics of the language (see Fig. 5). This was confirmed by the ANOVA statistical analysis. In the extremely difficult conditions (e.g., the 0 dB babble), only the KLT and logMMSE-SPU algorithms showed significant difference among different languages. One possible reason for this is that the important perceptual speech cues (e.g., temporal envelope, formant information, F0 contour) for word recognition were too difficult to be extracted from the processed signal by most noise-reduction algorithms, and that was found to be consistent across all three languages.
(2) Significant differences in relative word recognition score between Chinese and Japanese were found for the KLT, logMMSE-SPU, and Wiener-as algorithms in most conditions, and significant differences in relative word recognition score between Japanese and English were found mainly in the 5 dB babble condition.

The outcomes from the present investigation of noise-reduction algorithms in different languages provided useful information about differences in speech intelligibility of noise-reduction algorithms in different languages. Such

knowledge can be used to develop more advanced noise-reduction algorithm capable of improving speech intelligibility for a specific language.

ACKNOWLEDGMENTS

This research was partially supported by the National Natural Science Foundation of China (Nos. 10925419, 90920302, 10874203, 60875014, 61072124, and 11074275); the China–Japan (NSFC-JSPS) Bilateral Joint Projects; the SCOPE (071705001) of Ministry of Internal Affairs and Communication (MIC), Japan; NIH/NIDCD Grant Nos. R03-DC008887 and R01-DC07527, USA.

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.

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

Arai, T., Pavel, M., Hermansky, H., and Avendano, C. (**1999**). "Syllable intelligibility for temporally filtered LPC cepstral trajectories," J. Acoust. Soc. Am. **105**, 2783–2791.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (**2009**). *Noise Reduction in Speech Processing* (Springer Press, New York), pp. 151–183.

Chen, J., Benesty, J., and Huang, Y. (**2007**). "On the optimal linear filtering techniques for noise reduction," Speech Commun. **49**, 305–316.

Chen, J., Benesty, J., Huang, Y., and Doclo, S. (**2006**). "New insights into the noise reduction wiener filter," IEEE Trans. Audio, Speech, Lang. Process. **14**, 1218–1234.

Cohen, I., and Berdugo, B. (**2001**). "Speech enhancement for non-stationary noise environments," Signal Process. **8**, 2403–2418.

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

Drullman, R., Festen, J., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **96**, 1053–1064.

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.

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.

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

Hasegawa, Y. (**1999**). "Pitch accent and vowel devoicing in japanese.," in *Proceedings of the 14th International Congress of Phonetic Sciences*, pp. 523–526.

Helfer, K., and Wilber, L. (**1990**). "Hearing loss, aging and speech perception in reverberation and noise," J. Speech Hear. Res. **33**, 149–155.

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

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

Hu, Y., and Loizou, P. (**2007b**). "Subjective evaluation and comparison of speech enhancement algorithms," Speech Commun. **49**, 588–601.

IEEE (**1969**). "IEEE recommended practice for speech quality measurements," IEEE Trans. Audio Electroacoust. **11**, 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.

Kang, J. (**1998**). "Comparison of speech intelligibility between English and Chinese," J. Acoust. Soc. Am. **103**, 1213–1216.

Kim, G., Lu, Y., Hu, Y., and Loizou, P. (**2009**). "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," J. Acoust. Soc. Am. **126**, 1486–1494.

Leek, M., Dorman, M., and Summerfield, Q. (**1987**). "Minimum spectral contrast for vowel identification by normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. **81**, 148–154.

Loizou, P. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Taylor Francis Group, Florida), pp. 97–394.

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

Mittal, U., and Phamdo, N. (**2000**). "Signal/noise KLT based approach for enhancing speech degraded by colored noise," IEEE Trans. Speech, Audio Process. **8**, 159–167.

Montgomery, A., and Edge, R. (**1988**). "Evaluation of two speech enhancement techniques to improve intelligibility for hearing-impaired adults," J. Speech, Hear. Res. **31**, 386–393.

Niederiohn, R., and Groteluesch en, J. (**1976**). "The enhancement of speech intelligibility in high noise levels by high-pass filtering followed by rapid amplitude compression," IEEE Trans. Acoust., Speech, Signal Process. **24**, 277–282.

Parikh, G., and Loizou, P. (**2005**). "The influence of noise on vowel and consonant cues," J. Acoust. Soc. Am. **118**, 3874–3888.

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

Ryan, T. (**1959**). "Multiple comparisons in psychological research," Psychol. Bull. **56**, 26–47.

Ryan, T. (**1960**). "Significance tests for multiple comparisons of proportions, variances, and other statistics," Psychol. Bull. **57**, 318–328.

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.

Simpson, A., Moore, B., and Glasberg, B. (**1990**). "Speech enhancement to improve the intelligibility of speech in noise for hearing-impaired listeners," Acta Otolaryngoal. Suppl. **469**, 101–107.

Sohn, J., Kim, N., and Sung, W. (**1999**). "A statistical model-based voice activity detection," IEEE Signal Process. Lett. **6**, 1–3.

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

Uchida, Y., Lilly, D., and Meikle, M. (**1999**). "Cross-language speech intelligibility in noise: the comparison on the aspect of language dominance," J. Acoust. Soc. Am. **106**, 2151–2151.

Xu, L., and Pfingst, B. (**2003**). "Relative importance of temporal envelope and fine structure in lexical-tone perception," J. Acoust. Soc. Am. **114**, 3024–3027.

Xu, L., Thompson, C., and Pfingst, B. (**2005**). "Relative contributions of spectral and temporal cues for phoneme recognition," J. Acoust. Soc. Am. **117**, 3255–3267.