A comparative intelligibility study of single-microphone noise reduction algorithms

Yi Hu and Philipos C. Loizou[a)]
The University of Texas at Dallas, Department of Electrical Engineering, P.O. Box 830688, Richardson, Texas 75083, USA

(Received 9 May 2007; revised 3 July 2007; accepted 4 July 2007)

The evaluation of intelligibility of noise reduction algorithms is reported. IEEE sentences and consonants were corrupted by four types of noise including babble, car, street and train at two signal-to-noise ratio levels (0 and 5 dB), and then processed by eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, sub-space, statistical model based and Wiener-type algorithms. The enhanced speech was presented to normal-hearing listeners for identification. With the exception of a single noise condition, no algorithm produced significant improvements in speech intelligibility. Information transmission analysis of the consonant confusion matrices indicated that no algorithm improved significantly the place feature score, significantly, which is critically important for speech recognition. The algorithms which were found in previous studies to perform the best in terms of overall quality, were not the same algorithms that performed the best in terms of speech intelligibility. The subspace algorithm, for instance, was previously found to perform the worst in terms of overall quality, but performed well in the present study in terms of preserving speech intelligibility. Overall, the analysis of consonant confusion matrices suggests that in order for noise reduction algorithms to improve speech intelligibility, they need to improve the place and manner feature scores. © *2007 Acoustical Society of America.*
[DOI: 10.1121/1.2766778]

PACS number(s): 43.72.Kb, 43.72.Qr, 43.72.Ar [DOS] Pages: 1777–1786

I. INTRODUCTION

The objective of noise reduction (also called speech enhancement) algorithms is to improve one or more perceptual aspects of noisy speech, most notably, quality and intelligibility. Improving quality, however, might not necessarily lead to improvement in intelligibility. In fact, in some cases improvement in quality might be accompanied by a decrease in intelligibility. This is due to the distortion imparted on the clean speech signal resulting from excessive suppression of the noisy signal.

In some applications, the main goal of speech enhancement algorithms is to improve speech quality, with a secondary goal to preserve, at the very least, speech intelligibility. Hence, much of the focus of most speech enhancement algorithms has been to improve speech quality. Only a small number of algorithms have been evaluated using formal intelligibility tests (Boll, 1979; Lim, 1978; Tsoukalas *et al.*, 1997; Arehart *et al.*, 2003), and in those studies, only a single speech enhancement algorithm was evaluated and in a limited number of noise conditions. Also, in most of these studies, no statistical tests were run to assess whether the differences among algorithms were statistically significant. It therefore remains unclear as to which of the many speech enhancement algorithms proposed in the literature performs well in terms of speech intelligibility. At the very least, we would like to know which algorithm(s) preserve or maintain speech intelligibility in reference to the noisy (unprocessed) speech, and which algorithm(s) impair speech intelligibility, particularly in extremely low signal-to-noise ratio (SNR) conditions. Given the absence of accurate and reliable objective measure to predict the intelligibility of speech processed by enhancement algorithms, we must resort to formal listening tests to answer the above questions.

In this paper, we report on the intelligibility evaluation of eight speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. Phonetically balanced sentences and consonants were corrupted by four different types of noise commonly encountered in daily life, and processed by the above enhancement algorithms. The enhanced speech files were presented to normal-hearing subjects for identification in a double-walled soundproof booth. The present intelligibility study is a followup study to the one we reported in Hu and Loizou (2007b), with two main differences. First, we increased the number of subjects who participated on the sentence recognition task from 24 to 40. This was done to increase the power of the statistical tests used to assess significant differences between algorithms. Second, we now test subjects on the consonant recognition task. While the sentence test is attractive and practical as it reflects real-world communicative situations, it cannot be used to understand why some algorithms do not perform well or understand how to design algorithms that would improve intelligibility. For that reason, we chose to complement the sentence recognition task with the consonant recognition task. The consonant task was included for its diagnostic value and is similar in many respects to the diagnostic rhyme test (Voiers, 1983). More specifically, informa-

[a)]Author to whom correspondence should be addressed. Electronic mail: loizou@utdallas.edu

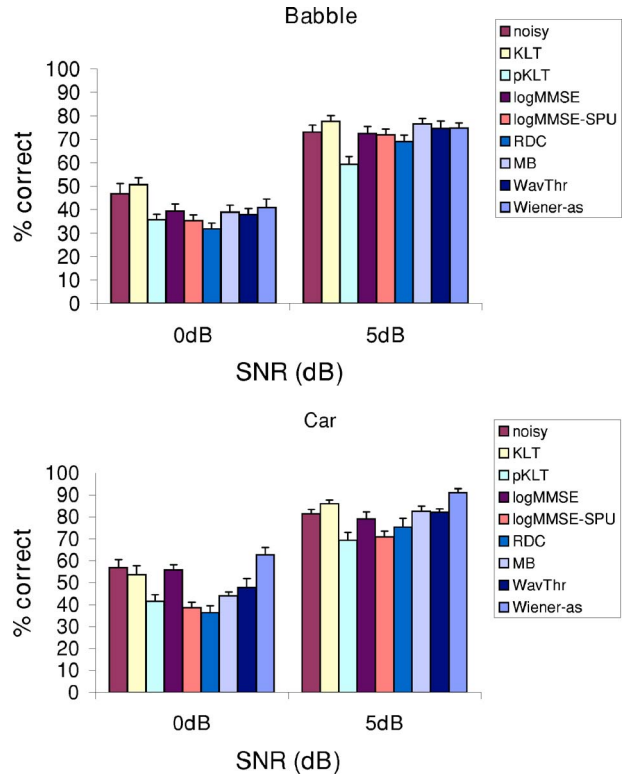

FIG. 1. (Color online) Mean sentence intelligibility scores for eight speech enhancement algorithms in the babble and car noise conditions at 0 and 5 dB SNR.

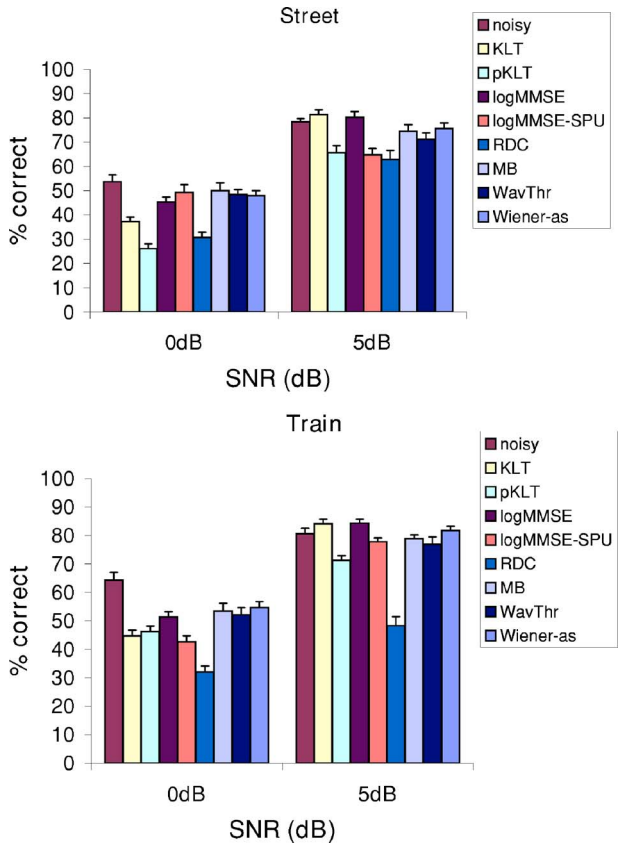

FIG. 2. (Color online) Mean sentence intelligibility scores for eight speech enhancement algorithms in the street and train noise conditions at 0 and 5 dB SNR.

tion transmission analysis, as per Miller and Nicely (1955), is used in the present study to analyze the consonant confusion matrices in terms of information transmitted for three articulatory features: place of articulation, manner of articulation and voicing. The information transmission analysis is important, because the detailed analysis of the feature scores can help us identify the limitations of existing algorithms as well as pinpoint the type of spectral/temporal distortions introduced by current enhancement algorithms. If, for instance, a particular algorithm yields low place feature scores, that would suggest that the spectral cues (e.g., formant transitions) are not adequately preserved and are perhaps distorted by the noise reduction algorithm. If a particular algorithm yields low manner or voicing feature scores, that would suggest that the gross temporal envelope cues (e.g., short-time energy, consonant/vowel energy ratio) are not adequately preserved by the noise reduction algorithm. The feature analysis of consonant confusion matrices thus provides valuable information that can help us identify the weaknesses of existing noise reduction algorithms and consequently help us design better noise reduction algorithms capable of improving speech intelligibility.

II. INTELLIGIBILITY STUDY

A. Methods

1. Materials and subjects

IEEE sentences (IEEE, 1969) and consonants in /a C a/ format were used for the intelligibility studies. The IEEE database was selected as it contains phonetically balanced sentences with relatively low word-context predictability. The consonant test included 16 consonants recorded in /a C a/ context, where C=/p, t, k, b, d, g, m, n, dh, l, f, v, s, z, sh, jh/. All consonants were produced by a female speaker, and all sentences were produced by a male talker. The IEEE sentences and consonants were recorded in a soundproof booth using Tucker Davis Technologies recording equipment. The sentences and consonants were originally sampled at 25 kHz and downsampled to 8 kHz. These recordings are available in Loizou (2007). To simulate the receiving frequency characteristics of telephone handsets, the speech and noise signals were filtered by the modified intermediate reference system (IRS) filters used in ITU-T P.862 (2000) for evaluation of the PESQ measure.

A total of 40 native speakers of American English were recruited for the sentence intelligibility tests. Ten additional listeners were recruited for the consonant tests. All subjects were paid for their participation.

2. Noise reduction algorithms

Noise was artificially added to the sentences as follows. The IRS filter was independently applied to the clean and noise signals to bandlimit the signals to 3.2 kHz. The active speech level of the filtered clean speech signal was first determined using the method B of ITU-T P. 56. A noise segment of the same length as the speech signal was randomly cut out of the noise recordings, was appropriately scaled to reach the desired SNR level and finally added to the filtered

TABLE I. Results from statistical comparisons between algorithms on sentence recognition. Algorithms indicated with asterisks performed equally well.

		Subspace		Statistical-model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	pKLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB			*					*
Babble		*							
Street				*	*		*	*	*
Train				*			*	*	*
Car	5 dB	*							*
Babble		*		*	*		*	*	*
Street		*		*			*		*
Train		*		*			*		*

clean speech signal. The noise signals were taken from the AURORA database (Hirsch and Pearce, 2000) and included the following recordings from different places: babble, car, street, and train. The noise signals were added to the speech signals at SNRs of 0 and 5 dB.

The noise-corrupted sentences were processed by eight different speech enhancement algorithms which included: the generalized KLT approach (Hu and Loizou, 2003), the perceptual KLT approach (pKLT) (Jabloun and Champagne, 2003), the Log Minimum Mean Square Error (logMMSE) algorithm (Ephraim and Malah, 1985), the logMMSE algorithm with speech presence uncertainty (logMMSE-SPU) (Cohen and Berdugo, 2002), the spectral subtraction algorithm based on reduced delay convolution (RDC) (Gustafsson *et al.*, 2001), the multiband spectral subtraction algorithm (MB) (Kamath and Loizou, 2002), the Wiener filtering algorithm based on wavelet-thresholded (WavThr) multitaper spectra (Hu and Loizou, 2004), and the Wiener algorithm based on *a priori* SNR estimation (Wiener-as) (Scalart and Filho 1996). With the exception of the logMMSE-SPU algorithm which was provided by the author (Cohen and Berdugo, 2002), all other algorithms were based on our own implementation. The parameters used in the implementation of these algorithms were the same as those published unless stated otherwise.

A statistical-model based voice activity detector (VAD) was used in all (but the subspace methods) algorithms to update the noise spectrum during speech-absent periods (Sohn *et al.*, 1999). The subspace methods used the VAD method proposed in Mittal *et al.* (2000) with the threshold value set to 1.2. Detailed description of the eight algorithms tested can be found in Hu and Loizou (2007a) and Loizou (2007). MATLAB implementations of all noise reduction algorithms tested in the present study are available in Loizou (2007).

3. Procedure

A total of 40 native speakers of American English were recruited for the sentence intelligibility tests. The 40 listeners were divided into four panels (one per type of noise), with each panel consisting of ten listeners. Each panel of listeners listened to sentences corrupted by a different type of noise. This was done to ensure that no subject listened to the same sentence twice. Each subject participated in a total of 19 listening conditions (=2 SNR levels $\times$8 algorithms +2 noisy references +1 quiet). Two sentence lists (ten sentences per list) were used for each condition. The presentation order of the listening conditions was randomized among subjects. Subjects were asked to write down all the words they heard.

An additional ten native speakers of American English were recruited for the consonant recognition task. There were six repetitions of each consonant. The presentation of the 16 consonants was completely randomized. The test session was preceded by one practice session in which the identity of the consonants (in quiet) was indicated to the listeners. To collect responses, a graphical interface was used that allowed the subjects to identify the consonants they heard by clicking on the corresponding button in the graphical interface.

TABLE II. Statistical comparisons between the intelligibility of noisy (unprocessed) sentences and enhanced sentences. Algorithms indicated with "E" were found to be equally intelligible to noisy speech, algorithms indicated with "L" yielded lower intelligibility scores and algorithms indicated with "B" improved intelligibility.

		Subspace		Statistical model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	E	L	E	L	L	L	L	E
Babble		E	L	E	L	L	E	L	E
Street		L	L	L	E	L	E	E	E
Train		L	L	L	L	L	L	L	L
Car	5 dB	E	L	E	L	E	E	E	B
Babble		E	L	E	E	E	E	E	E
Street		E	L	E	L	L	E	E	E
Train		E	L	E	E	L	E	E	E

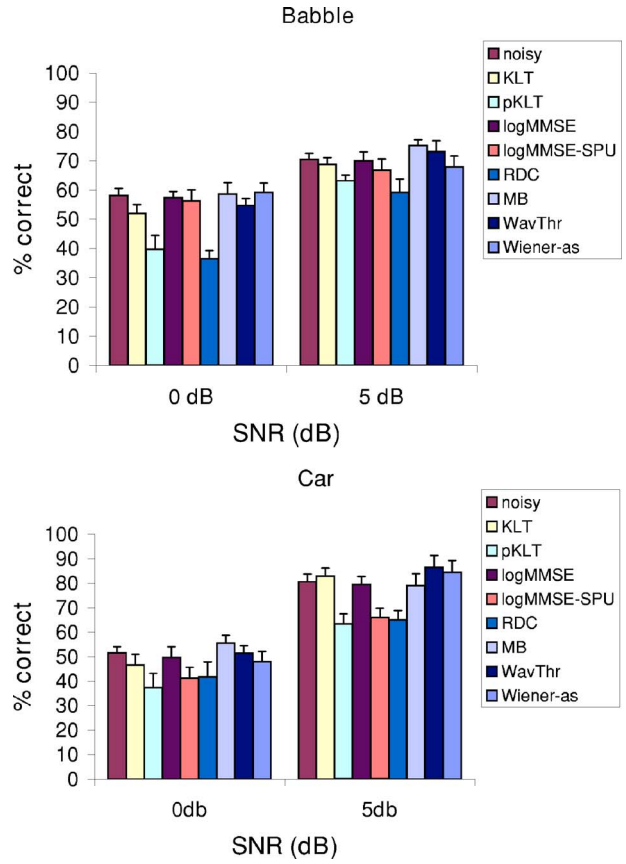

FIG. 3. (Color online) Mean consonant intelligibility scores for eight speech enhancement algorithms in the babble and car noise conditions at 0 and 5 dB SNR.

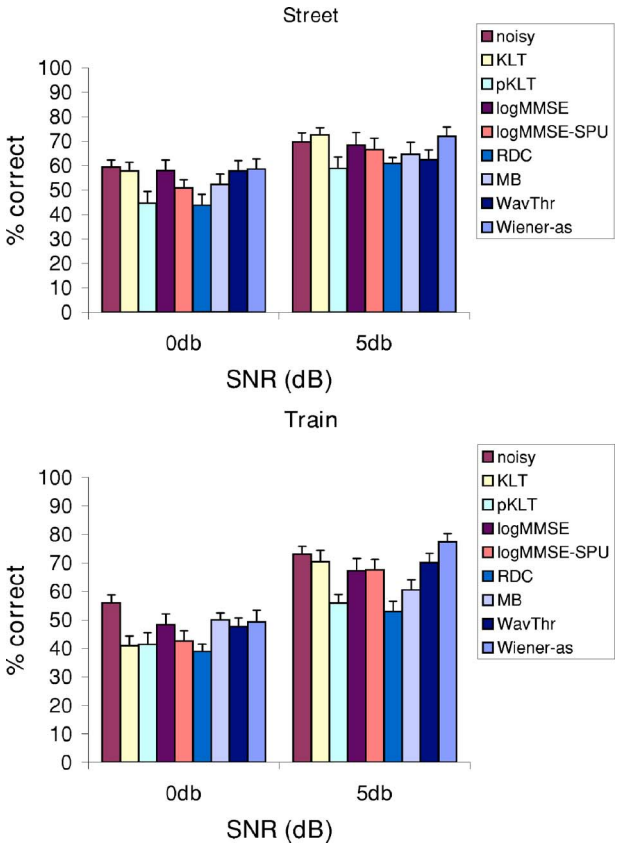

FIG. 4. (Color online) Mean consonant intelligibility scores for eight speech enhancement algorithms in the street and train noise conditions at 0 and 5 dB SNR.

The processed speech files (sentences/consonants), along with the clean and noisy speech files, were presented monaurally to the listeners in a double-walled soundproof booth via Sennheiser's (HD 250 Linear II) circumaural headphones at a comfortable level. Tests were conducted in multiple sessions with each session lasting no more than 2 h. The subjects were allowed to take a break during the listening session to reduce fatigue.

III. RESULTS

Listening tasks involved sentence and consonant recognition in noise. Speech intelligibility was assessed in terms of percentage of words identified correctly. All words were considered in the scoring. We report the results on sentence and consonant recognition separately.

A. Sentence recognition

Figure 1 shows the mean intelligibility scores for babble and car noises, and Fig. 2 shows the mean scores for street and train noises. The error bars in the figures give the standard errors of the mean. The intelligibility scores of noisy (unprocessed) speech are also given for comparative purposes. Of all the conditions tested, algorithms performed the worst (i.e., yielded lowest intelligibility scores) in multitalker babble.

We present comparative analysis at two levels. At the first level, we compare the performance of the various algorithms across all classes aiming to find the algorithm(s) that performed the best across all noise conditions. At the second level, we compare the performance of all algorithms in reference to the noisy speech (unprocessed). This latter comparison will provide valuable information as to which, if any, algorithm(s) significantly improve the intelligibility of noisy speech. If no improvement is obtained, we can learn at the very least which algorithm(s) maintain speech intelligibility and which algorithm(s) diminish speech intelligibility.

In order to assess significant differences between the intelligibility scores obtained from each algorithm, we subjected the scores of the 40 listeners to statistical analysis. Analysis of variance (ANOVA) indicated a highly significant effect ($p < 0.005$) of speech enhancement algorithms on speech intelligibility (a highly significant effect was found in all SNR conditions and types of noise). Following the ANOVA, we conducted multiple comparison statistical tests according to Fisher's LSD test to assess significant differences between algorithms. Differences between scores were deemed significant if the obtained p value (level of significance) was smaller than 0.05.

1. Intelligibility comparison among algorithms

Statistical analysis was performed to assess significant differences in performance between algorithms. Multiple paired comparisons (Fisher's LSD) were conducted between the algorithm with the highest score against all other algorithms. Table I reports the results from the statistical analysis.

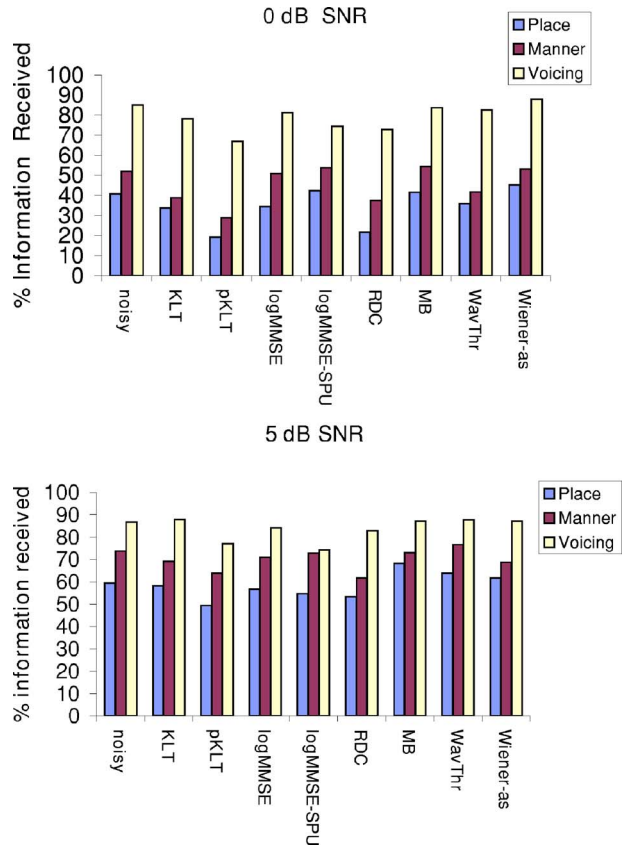

FIG. 5. (Color online) Mean percent information received for the multitalker babble conditions at 0 and 5 dB SNR.

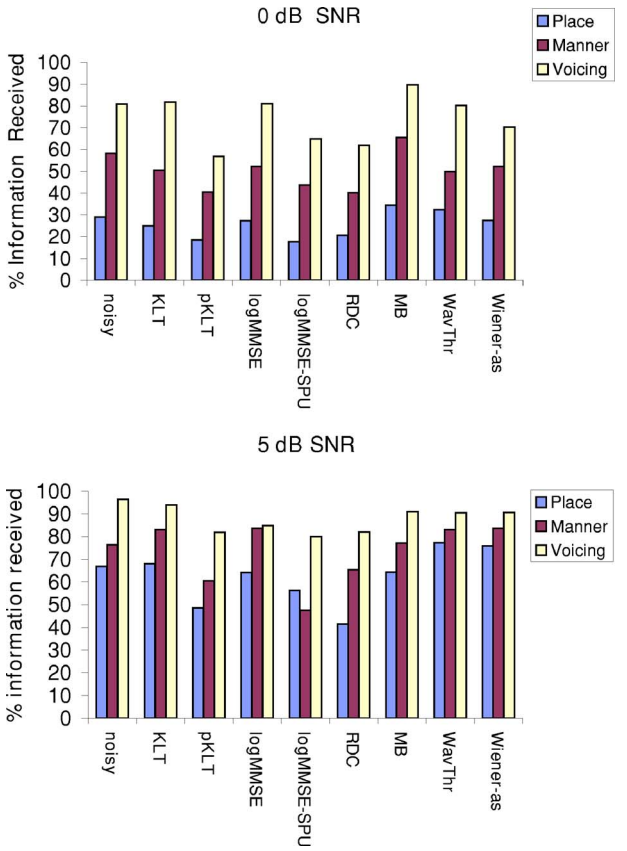

FIG. 6. (Color online) Mean percent information received for the car conditions at 0 and 5 dB SNR.

The algorithms denoted by asterisks in Table I performed equally well. At 5 dB SNR, the KLT and Wiener-as algorithms performed equally well in all conditions. This was followed by the logMMSE and MB algorithms. The pKLT, RDC, logMMSE-SPU and WavThr algorithms performed poorly. At 0 dB SNR, the Wiener-as and logMMSE algorithms performed equally well in most conditions. This was followed by the MB and WavThr algorithms. The KLT algorithm performed poorly except in the babble condition in which it performed the best among all algorithms. Considering all conditions, the Wiener-as algorithm performed consistently well in nearly all conditions, followed by the logMMSE algorithms which performed well in six of the eight noise conditions, followed by the KLT and MB algorithms which performed well in five conditions.

2. Intelligibility comparison against noisy speech

Further analysis was performed to find out whether intelligibility is improved or at least maintained (i.e., speech was equally intelligible) in reference to noisy (unprocessed) speech. Multiple paired comparisons (Fisher's LSD) were conducted between the intelligibility scores obtained with noisy speech (unprocessed) samples and the scores obtained with sentences enhanced by the various algorithms. The results from the statistical analysis are given in Table II. Algorithms indicated with "E" yielded equal intelligibility to noisy speech, algorithms indicated with "L" yielded lower intelligibility, and algorithms indicated with "B" improved speech intelligibility. The Wiener-as algorithm maintained speech intelligibility in six of the eight noise conditions tested, and improved intelligibility in 5 dB car noise. Good performance was followed by the KLT, logMMSE and MB algorithms which maintained speech intelligibility in six conditions. All algorithms produced a decrement in intelligibility in train noise at 0 dB SNR. The pKLT and RDC algorithms significantly reduced the intelligibility of speech in most conditions.

B. Consonant recognition

Figure 3 shows the mean consonant recognition scores for babble and car noises, and Fig. 4 shows the mean scores for street and train noises. The error bars in the figures give the standard errors of the mean. The intelligibility scores of noisy (unprocessed) consonants are also given for comparative purposes. Of all the conditions tested, the train condition at 0 dB SNR was the most challenging, with most algorithms performing worse than the unprocessed (noisy) consonants.

The consonant confusion matrices were subjected to information transmission analysis (Miller and Nicely, 1955) to obtain the percent information transmitted for three articulatory features: place of articulation, manner of articulation and voicing. The mean feature scores are plotted in Figs. 5–8 for the four different types of noise tested. Overall, the place scores were the lowest and the voicing scores were the highest. This outcome is consistent with the findings in Miller and Nicely (1955) with noisy (unprocessed) consonants presented at various SNR levels.

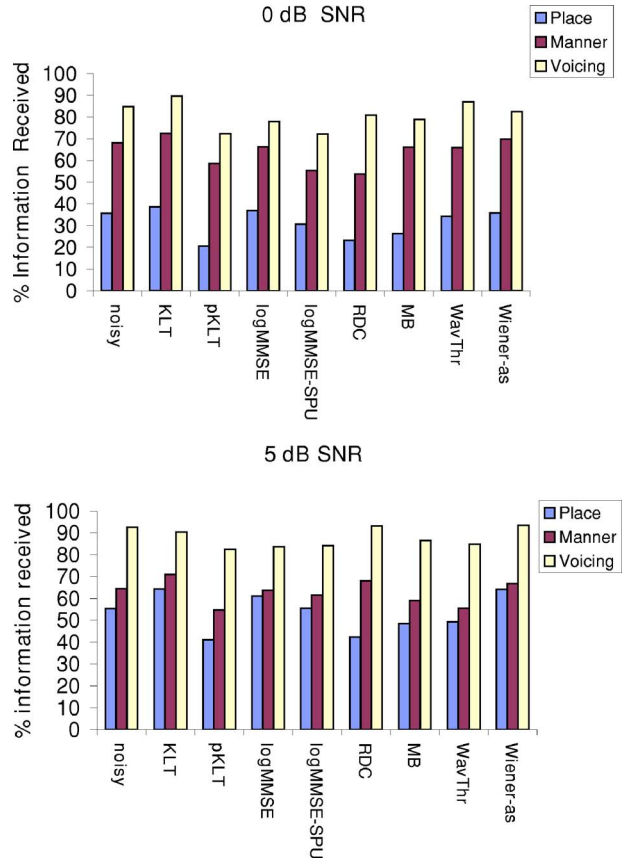

FIG. 7. (Color online) Mean percent information received for the street conditions at 0 and 5 dB SNR.

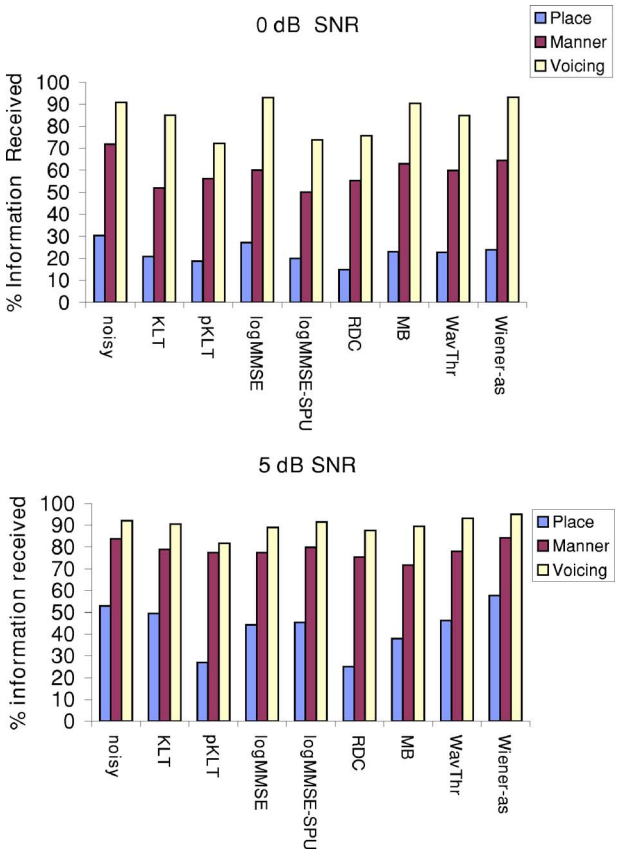

FIG. 8. (Color online) Mean percent information received for the train conditions at 0 and 5 dB SNR.

As before, we present comparative analysis at two levels. At the first level, we compare the performance of the various algorithms across all classes aiming to find the algorithm(s) that performed the best across all noise conditions. At the second level, we compare the performance of all algorithms in reference to the noisy consonants (unprocessed).

1. Consonant intelligibility comparison among algorithms

Statistical analysis was performed to assess significant differences in performance between algorithms. Multiple paired comparisons (Fisher's LSD) were conducted between the algorithm with the highest score against all other algorithms. Table III reports the results from the statistical analysis. Asterisks in the table indicate absence of statistically significant difference (i.e., $p > 0.05$) between the algorithm with the highest score and the denoted algorithm. That is, the algorithms denoted by asterisks in Table III performed equally well. At 5 dB SNR, with the exception of a few algorithms (pKLT and RDC), most algorithms performed equally well. A similar pattern was also observed at 0 dB SNR. The KLT, logMMSE, MB and Wiener-as algorithms performed equally well in most conditions. The logMMSE-SPU performed well in most conditions except in car noise. Overall, the Wiener-type algorithms (Wiener-as and WavThr) and the KLT algorithm performed consistently well in all conditions, followed by the logMMSE and MB algorithms. The RDC and pKLT algorithms performed poorly relative to the other algorithms, and the underlying causes are investigated in the next section.

2. Consonant intelligibility comparison against noisy consonants

Further analysis was performed to find out whether consonant intelligibility is improved or at least maintained in reference to the intelligibility of noisy (unprocessed) consonants. Multiple paired comparisons (Fisher's LSD) were conducted between the intelligibility scores obtained with noisy speech (unprocessed) samples and the scores obtained with consonants enhanced by the various algorithms. The results from the statistical analysis are given in Table IV. In this table, algorithms indicated with E yielded equal consonant intelligibility to noisy consonants and algorithms indicated with L yielded lower intelligibility. Statistical analysis revealed that the Wiener-type algorithms (Wiener-as and WavThr) and the logMMSE algorithm preserved consonant intelligibility in all eight conditions. That is, enhanced consonants were found to be as intelligible as that of noisy (unprocessed) consonants. This was followed by the KLT and MB algorithms, which maintained consonant intelligibility in seven of the eight conditions. The RDC and pKLT algorithms produced a decrement in consonant intelligibility in a number of conditions.

Next, we used the data from feature transmission analysis to examine why some algorithms performed poorly with

TABLE III. Results obtained form the comparative statistical analysis of intelligibility scores of /a*C*a/ syllables. Algorithms indicated by asterisks performed equally well, in terms of speech intelligibility. Algorithms with no asterisks performed poorly.

		Subspace		Statistical model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	*		*			*	*	*
Babble		*		*	*		*	*	*
Street		*		*	*		*	*	*
Train		*	*	*	*		*	*	*
Car	5 dB	*		*			*	*	*
Babble		*		*	*		*	*	*
Street		*		*	*		*	*	*
Train		*			*			*	*

TABLE IV. Statistical comparisons between the intelligibility of noisy (unprocessed) consonants and enhanced consonants. Algorithms indicated with "E" were found to be equally intelligible to noisy speech. Algorithms indicated with "L" obtained lower intelligibility scores than noisy consonants.

		Subspace		Statistical model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	E	L	E	E	E	E	E	E
Babble		E	L	E	E	L	E	E	E
Street		E	L	E	E	L	E	E	E
Train		L	L	E	L	L	E	E	E
Car	5 dB	E	L	E	L	L	E	E	E
Babble		E	E	E	E	L	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	L	E	E	L	L	E	E

TABLE V. Place feature transmission comparison with respect to noisy (unprocessed) consonants.

		Subspace		Statistical-model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	E	E	E	E	E	E	E	E
Babble		E	L	E	E	L	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	E	E	E	L	E	E	E
Car	5 dB	E	L	E	E	L	E	E	E
Babble		E	E	E	E	E	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	L	E	E	L	L	E	E

TABLE VI. Manner feature transmission comparison with respect to noisy (unprocessed) consonants.

		Subspace		Statistical model		Spectral subtractive		Wiener type	
Noise	SNR	KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	E	L	E	L	L	E	E	E
Babble		L	L	E	E	L	E	E	E
Street		E	E	E	L	L	E	E	E
Train		L	L	E	L	L	E	E	E
Car	5 dB	E	L	E	L	E	E	E	E
Babble		E	E	E	E	E	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	E	E	E	E	L	E	E

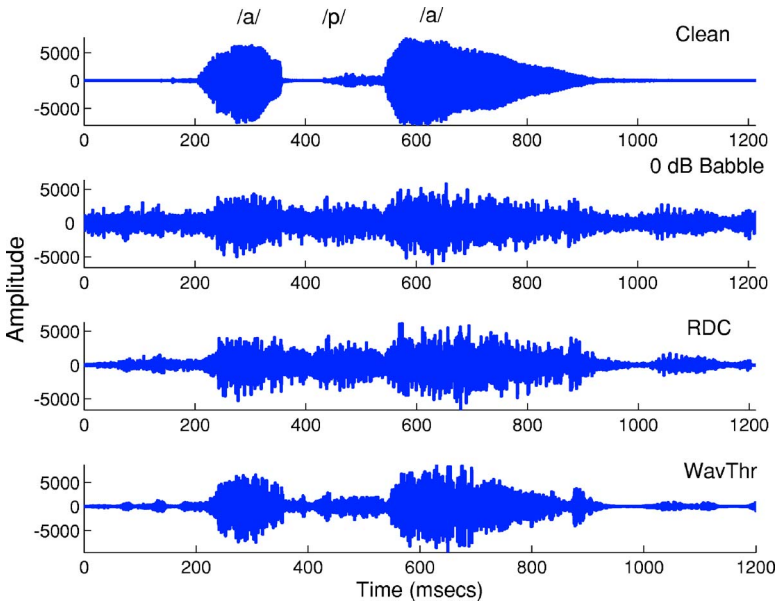

FIG. 9. (Color online) Example consonant waveforms processed by the RDC and WavThr algorithms. The top two panels show the /a p a/ waveform in quiet and in 0 dB babble, respectively.

respect to noisy speech. Statistical analysis was conducted separately for each of the three articulatory features (place, manner, and voicing). Multiple paired comparisons (Fisher's LSD) were conducted between the feature scores obtained with noisy consonant (unprocessed) samples and the feature scores obtained with consonants enhanced by the various algorithms. The comparative results (with respect to the feature scores obtained for noisy consonants) from the statistical analysis are given in Tables V–VII for place, manner and voicing respectively.

From Table IV, we see that the KLT algorithm performed poorly in only one condition, 0 dB SNR train noise. According to Table VI, this was attributed to manner confusion errors. The place and voicing features were preserved. The logMMSE-SPU algorithm performed poorly in the 0–dB SNR train condition and in the 5 dB SNR car condition. From Tables VI and VII, we observe that this was due to manner and voicing errors. Lastly, the RDC and pKLT algorithms performed poorly on consonant recognition (see Table IV) due to place, manner and voicing errors. The majority of the feature errors, however, made by the RDC and pkLT algorithms at 0 dB SNR were caused by manner confusion errors (Table VI).

The manner confusion errors were found to be quite common in most algorithms (Table VI), particularly at low SNR levels (0 dB). The manner errors are produced when enhancement algorithms do not adequately preserve the gross temporal envelope of speech. The stop consonant /p/, for instance, is characterized by a brief silence (closure) followed by a short burst and a low-energy aspiration segment (see example in Fig. 9, top panel). Preserving the silence in /p/ prior to the occurrence of the burst is critical for accurate perception of the stop consonant /p/. Any errors made by enhancement algorithms in preserving the low-energy characteristics of stop consonants will undoubtedly cause manner (or voicing) confusion errors. Figure 9 shows an example waveform of /a p a/ that caused a manner confusion error when the noisy signal was processed by the RDC algorithm at 0 dB SNR. As can be seen, the RDC algorithm enhanced poorly the /p/ segment of the word, as the energy (and gross envelope) of the enhanced /p/ segment was as large as the preceding vocalic segment /a/. Consequently, the enhanced consonant /p/ was not perceived as /p/ by listeners but as /f/, thus contributing to a manner of articulation error. In direct contrast, Fig. 9 (bottom panel) shows /a p a/ enhanced by an algorithm (WavThr) that did not produce as many manner confusion errors as the RDC algorithm. In this example, the WavThr algorithm preserved the low-energy characteristics of /p/.

IV. DISCUSSION AND CONCLUSIONS

This paper compared the intelligibility of speech produced by eight different enhancement algorithms operating in several types of noise and SNR conditions. Based on the statistical analysis of the sentence and consonant intelligibility scores, we can draw the following conclusions:

1. With the exception of a single noise condition (car noise at 5 dB SNR), no algorithm produced significant improvements in speech intelligibility. The majority of the algorithms (KLT, logMMSE, MB, WavThr, Wiener-as) tested were able to maintain intelligibility at the same level as that of noisy speech.
2. When comparing the performance of the various algorithms, we found that the Wiener-as algorithm performed consistently well in nearly all conditions for both sentence and consonant recognition tasks. Following the Wiener-as algorithm, the KLT (subspace), MB and logMMSE algorithms performed comparably well on sentence recognition. On consonant recognition, the Wiener-as, KLT, and WavThr algorithms performed equally well, followed by the logMMSE and MB algorithms.
3. The algorithms that were found in our previous study (Hu and Loizou, 2007a) to perform the best in terms of overall quality were not the same algorithms that performed the best in terms of speech intelligibility. The KLT (subspace) algorithm was found in Hu and Loizou (2007a) to perform the worst in terms of overall quality, but performed

TABLE VII. Voicing feature transmission comparison with respect to noisy (unprocessed) consonants.

Noise	SNR	Subspace		Statistical-model		Spectral subtractive		Wiener type	
		KLT	*p*KLT	logMMSE	logMMSE-SPU	RDC	MB	WavThr	Wiener-as
Car	0 dB	E	L	E	E	L	E	E	E
Babble		E	E	E	E	E	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	L	E	L	L	E	E	E
Car	5 dB	E	E	E	L	E	E	E	E
Babble		E	E	E	E	E	E	E	E
Street		E	E	E	E	E	E	E	E
Train		E	L	E	E	E	E	E	E

well in the present study in terms of preserving speech intelligibility. In fact, in babble noise (0 dB SNR), the KLT algorithm performed significantly better than the logMMSE algorithm (see Table I), which was found in Hu and Loizou (2007a) to be among the algorithms with the highest overall speech quality.

4. The Wiener-as algorithm performed the best in terms of preserving speech intelligibility (in one case, it improved intelligibility). We believe that this is due to the fact that it applies the least amount of attenuation to the noisy signal, and thus introduces negligible speech distortion. This is done, however, at the expense of introducing noise distortion (residual noise). At the other extreme, the pKLT approach significantly reduces the noise distortion but introduces a great deal of speech distortion, which in turn impairs speech intelligibility. In between the two extremes of speech/noise distortion lie the KLT and logMMSE algorithms. The WavThr and MB algorithms also fall between the two extremes of speech/noise distortion, and preserve consonant intelligibility in nearly all conditions.
5. Analysis of the consonant confusion matrices revealed that the majority of the confusions are due to place of articulation errors, followed by manner of articulation errors and voicing errors. The manner confusion errors were found to be quite common in most algorithms (Table VI), particularly at low SNR levels (0 dB). This suggests that most algorithms do not adequately preserve the low-energy characteristics of consonants at low-SNR environments.
6. The performance of speech enhancement algorithms, in terms of speech intelligibility, seems to be dependent on the temporal/spectral characteristics of the noise, and this dependence is more evident in the low-SNR conditions (0 dB in our case). In the 0 dB babble condition, for instance, the subspace algorithm performed the best but did not perform as well in the other conditions (car, street and train environments). In the 0 dB train condition, none of the evaluated speech enhancement algorithms preserved speech intelligibility (see Table II). The same algorithms, however, did preserve speech intelligibility in other noise conditions (same SNR).

The analysis of consonant confusion matrices provided some insight as to why none of the enhancement algorithms improved speech intelligibility. As shown in Figs. 5–8 (and Table V) none of the enhancement algorithms improved the place feature scores, which are critically important for conveying formant frequency information (Borden *et al.*, 1994). The place scores were found to be the lowest, about 30% at 0 dB SNR across the various types of noise. For an enhancement algorithm to improve intelligibility, it needs to significantly improve the place of articulation feature score. This can be done by designing algorithms that preserve spectral information, and particularly formant frequency information. The manner feature scores, which were the second lowest (about 50–70% at 0 dB SNR across conditions) also need to improve. This can be done by designing techniques that preserve the gross temporal envelope and low-energy characteristics of consonants.

Finally, it is important to point out that the disappointing conclusion drawn from this study that single-microphone enhancement algorithms do not improve speech intelligibility is only applicable to normal-hearing listeners and not necessarily to hearing-impaired listeners wearing hearings aids (Arechart *et al.*, 2003) or cochlear implants (Loizou *et al.*, 2005). In a different study (Loizou *et al.*, 2005), we showed that the KLT algorithm can significantly improve speech intelligibility in cochlear implant users. Further research is therefore needed to investigate the performance of existing speech enhancement algorithms in hearing-impaired listeners.

ACKNOWLEDGMENTS

The authors would like to thank Jessica Dagley for her help in collecting the data. The authors also would like to thank Dr. Israel Cohen for providing the MATLAB implementation of the log-MMSE method with speech presence uncertainty. Research was supported by Grant No. R01 DC007527 from the National Institute of Deafness and other Communication Disorders, NIH.

Arehart, K. H., Hansen, J. H., Gallant, S., and Kalstein, L. (**2003**). "Evaluation of an auditory masked threshold noise suppression algorithm in normal-hearing and hearing-impaired listeners," Speech Commun. **40**, 575–592.

Boll, S. F. (**1979**). "Suppression of acoustic noise in speech using spectral subtraction," IEEE Trans. Acoust., Speech, Signal Process. **ASSP-27**, 113–120.

Borden, G., Harris, K. and Raphael, L. (**1994**). *Speech Science Primer: Physiology, Acoustics, and Perception of Speech* (Williams and Wilkins, Baltimore).

Cohen, I., and Berdugo, B. (**2002**). "Noise estimation by minima controlled recursive averaging for robust speech enhancement," IEEE Signal Process. Lett. **9**, 12–15.

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

Gustafsson, H., Nordholm, S., and Claesson, I. (**2001**). "Spectral subtraction using reduced delay convolution and adaptive averaging," IEEE Trans. Speech Audio Process. **9**, 799–807.

Hirsch, 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 ASR2000*, Paris, France.

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

Hu, Y., and Loizou, P. C. (**2004**). "Speech enhancement based on wavelet thresholding the multitaper spectrum," IEEE Trans. Speech Audio Process. **12**, 59–67.

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

Hu, Y., and Loizou, P. C. (**2007b**). "A comparative intelligibility study of speech enhancement algorithms," in *Proc. IEEE int. Conf. Acoust., Speech, Signal Processing*, Honolulu, Hawaii, pp. 561–564.

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

ITU-T P.862 (**2000**). *Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs*, (ITU-T Recommendation P.862).

Jabloun, F., and Champagne, B. (**2003**). "Incorporating the human hearing properties in the signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **11**, 700–708.

Kamath, S., and Loizou, P. C. (**2002**). "A multi-band spectral subtraction method for enhancing speech corrupted by colored noise," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, Orlando, FL.

Lim, J. S. (**1978**). "Evaluation of a correlation subtraction method for enhancing speech degraded by additive white noise," IEEE Trans. Acoust., Speech, Signal Process. **26**, 471–472.

Loizou, P. (**2007**). *Speech Enhancement: Theory and Practice* (CRC, Boca Raton, FL).

Loizou, P. C, Lobo, A., and Hu, Y. (**2005**). "Subspace algorithms for noise reduction in cochlear implants," J. Acoust. Soc. Am. **118**, 2791–2793.

Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.

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.

Scalart, P, and Filho, J. (**1996**). "Speech enhancement based on a priori signal to noise estimation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Processing*, Atlanta, GA, pp. 629–632.

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

Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (**1997**). "Speech enhancement based on audible noise suppression," IEEE Trans. Speech Audio Process. **5**, 479–514.

Voiers, W. (**1983**). "Evaluating processed speech using the diagnostic rhyme test," Speech Technol. **Jan./Feb.**, 30–39.