Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 49 (2007) 588–601
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Subjective comparison and evaluation of speech enhancement algorithms

Yi Hu, Philipos C. Loizou *

Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX 75083-0688, USA

Received 2 March 2006; received in revised form 10 July 2006; accepted 18 December 2006

Abstract

Making meaningful comparisons between the performance of the various speech enhancement algorithms proposed over the years has been elusive due to lack of a common speech database, differences in the types of noise used and differences in the testing methodology. To facilitate such comparisons, we report on the development of a noisy speech corpus suitable for evaluation of speech enhancement algorithms. This corpus is subsequently used for the subjective evaluation of 13 speech enhancement methods encompassing four classes of algorithms: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. The subjective evaluation was performed by Dynastat, Inc., using the ITU-T P.835 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion and overall quality. This paper reports the results of the subjective tests.

Keywords: Speech enhancement; Noise reduction; Subjective evaluation; ITU-T P.835

1. Introduction

Over the past three decades, various speech enhancement algorithms have been proposed to improve the performance of modern communication devices in noisy environments. Yet, it still remains unclear as to which speech enhancement algorithm performs well in real-world listening situations where the background noise level and characteristics are constantly changing. Reliable and fair comparison between algorithms has been elusive for several reasons, including lack of common speech database for evaluation of new algorithms, differences in the types of noise used and differences in the testing methodology. Without having access to a common speech database, it is nearly impossible for researchers to compare at very least the objective performance of their algorithms with that of others. Subjective evaluation of speech enhancement algorithms is further complicated by the fact that the quality of enhanced speech has both signal and noise distortion components, and it is not clear as to whether listeners base their quality judgments on the signal distortion, noise distortion or both. This concern was recently addressed by a new ITU-T standard (P.835) that was designed to lead the listeners to integrate the effects of both signal and background distortion in making their ratings of overall quality.

In this paper, we report on the subjective comparison and evaluation of 13 speech enhancement algorithms using the ITU-T P.835 methodology. The speech enhancement algorithms were chosen to encompass four different classes of noise reduction methods: spectral subtractive, subspace, statistical-model based and Wiener-type algorithms. These algorithms were evaluated using a newly developed noisy speech corpus (NOIZEUS) suitable for evaluation of speech enhancement algorithms and available from our website. The enhanced speech files were sent to Dynastat, Inc (Austin, TX) for subjective evaluation using the recently standardized methodology for evaluating noise suppression algorithms based on ITU-T P.835 (2003). This paper presents the results from the comparative analysis of the subjective tests.

* Corresponding author. Tel.: +1 972 883 2710.
E-mail address: loizou@utdallas.edu (P.C. Loizou).

0167-6393/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2006.12.006

2. NOIZEUS: a noisy speech corpus for evaluation of speech enhancement algorithms

NOIZEUS[1] is a noisy speech corpus recorded in our lab to facilitate comparison of speech enhancement algorithms among research groups. The noisy database contains 30 IEEE sentences (IEEE Subcommittee, 1969) produced by three male and three female speakers, and was corrupted by eight different real-world noises at different SNRs. The noise was taken from the AURORA database (Hirsch et al., 2000) and includes suburban train noise, multi-talker babble, car, exhibition hall, restaurant, street, airport and train-station noise. The list of sentences used in NOIZEUS are given in Tables 1 and 2, and the broad phonetic class distribution is shown in Fig. 1.

2.1. Speech material

Thirty sentences from the IEEE sentence database were recorded in a sound-proof booth using Tucker Davis Technologies (TDT) recording equipment. The sentences were produced by three male and three female speakers (five sentences/speaker). The IEEE database was used as it contains phonetically balanced sentences with relatively low word-context predictability. The thirty sentences were selected from the IEEE database so as to include all phonemes in the American English language (see Fig. 1). The sentences were originally sampled at 25 kHz and downsampled to 8 kHz.

2.2. Noise

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. The frequency response of the filter is shown in Fig. 2.

Noise was artificially added to the speech signal as follows. The IRS filter was independently applied to the clean and noise signals. The active speech level of the filtered clean speech signal was first determined using method B of (ITU-T P.56, 1993). A noise segment of the same length as the speech signal was randomly cut out of the noise recordings, appropriately scaled to reach the desired SNR level and finally added to the filtered clean speech signal.

Noise signals were taken from the AURORA database (Hirsch et al., 2000) and included the following recordings from different places: babble (crowd of people), car, exhibition hall, restaurant, street, airport, train station, and train. The noise signals were added to the speech signals at SNRs of 0 dB, 5 dB, 10 dB and 15 dB.

[1] Available at: http://www.utdallas.edu/~loizou/speech/noizeus/.

Table 1
List of sentences used in NOIZEUS

Filename	Speaker	Gender	Sentence text
sp01.wav	CH	M	The birch canoe slid on the smooth planks
sp02.wav	CH	M	He knew the skill of the great young actress
sp03.wav	CH	M	Her purse was full of useless trash
sp04.wav	CH	M	Read verse out loud for pleasure
sp05.wav	CH	M	Wipe the grease off his dirty face
sp06.wav	DE	M	Men strive but seldom get rich
sp07.wav	DE	M	We find joy in the simplest things
sp08.wav	DE	M	Hedge apples may stain your hands green
sp09.wav	DE	M	Hurdle the pit with the aid of a long pole
sp10.wav	DE	M	The sky that morning was clear and bright blue
sp11.wav	JE	F	He wrote down a long list of items
sp12.wav	JE	F	The drip of the rain made a pleasant sound
sp13.wav	JE	F	Smoke poured out of every crack
sp14.wav	JE	F	Hats are worn to tea and not to dinner
sp15.wav	JE	F	The clothes dried on a thin wooden rack

The sentences used in the subjective evaluation are underlined.

Table 2
List of sentences used in NOIZEUS

Filename	Speaker	Gender	Sentence text
sp16.wav	KI	F	The stray cat gave birth to kittens
sp17.wav	KI	F	The lazy cow lay in the cool grass
sp18.wav	KI	F	The friendly gang left the drug store
sp19.wav	KI	F	We talked of the sideshow in the circus
sp20.wav	KI	F	The set of china hit the floor with a crash
sp21.wav	SI	M	Clams are small, round, soft and tasty
sp22.wav	SI	M	The line where the edges join was clean
sp23.wav	SI	M	Stop whistling and watch the boys march
sp24.wav	SI	M	A cruise in warm waters in a sleek yacht is fun
sp25.wav	SI	M	A good book informs of what we ought to know
sp26.wav	TI	F	She has a smart way of wearing clothes
sp27.wav	TI	F	Bring your best compass to the third class
sp28.wav	TI	F	The club rented the rink for the fifth night
sp29.wav	TI	F	The flint sputtered and lit a pine torch
sp30.wav	TI	F	Let us all join as we sing the last chorus

The sentences used in the subjective evaluation are underlined.

3. Algorithms evaluated

A total of 13 different speech enhancement methods were evaluated based on our own implementation (see list in Table 3). Representative algorithms from four different classes of enhancement algorithms were chosen: three spectral subtractive algorithms, two subspace algorithms, three Wiener-type algorithms and five statistical-model based algorithms. The Wiener-type algorithms were grouped separately since these algorithms estimate the complex spectrum in the mean square sense while the statistical-model algorithms estimate the magnitude spectrum. The parameters used in the implementation of these algorithms were the same as those published unless stated otherwise. No adjustments were made for the algorithms (e.g., Cohen,

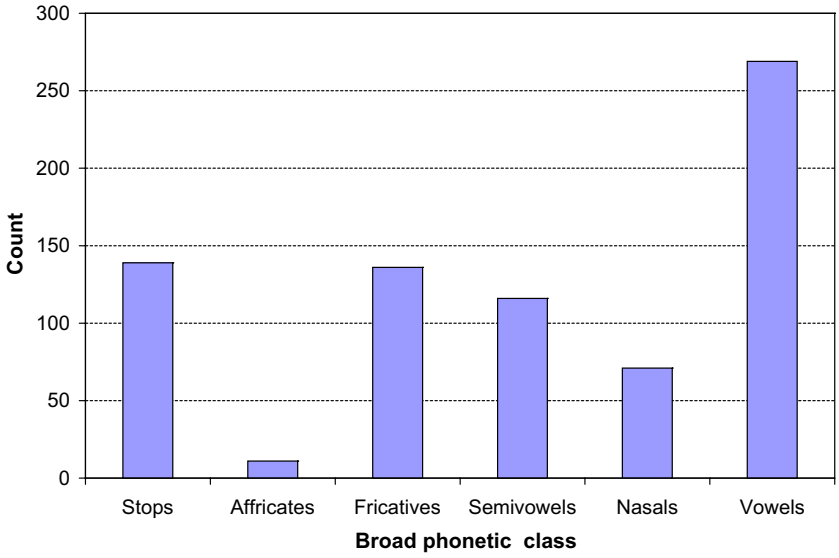

Fig. 1. Broad phonetic class distribution of the NOIZEUS corpus.

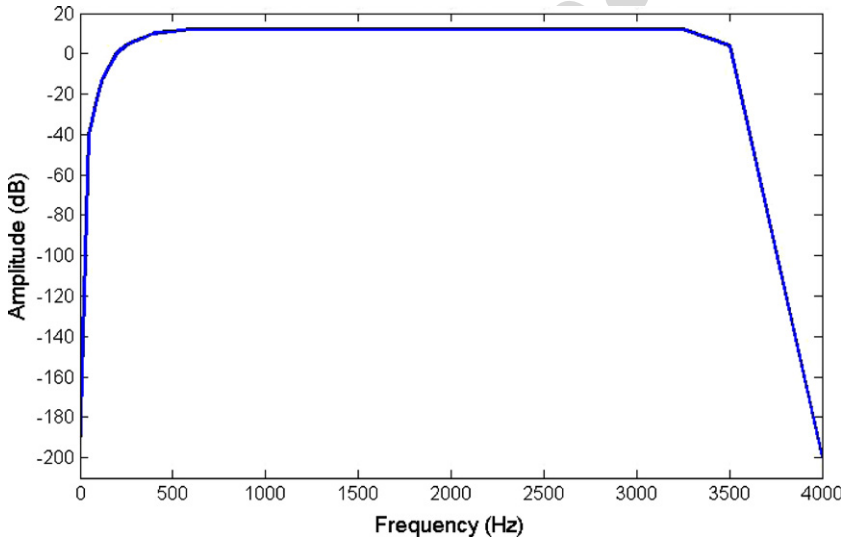

Fig. 2. Frequency response of IRS filter.

2002) originally designed for a sampling rate of 16 kHz. To assess the merit of noise-estimation algorithms, two speech-enhancement algorithms (denoted in Table 3 with the suffix -ne) were also implemented with a noise-estimation algorithm (Rangachari and Loizou, 2006). That is, a noise-estimation algorithm was used in the speech enhancement algorithms indicated in Table 3 with -ne, to estimate and update the noise spectrum. The majority of the algorithms tested updated the noise spectrum using a voice activity detector (more on this later).

With the exception of the multi-band (MB) spectral subtraction algorithm developed in our lab (Kamath and Loizou, 2002; Kamath, 2001), the remaining algorithms have been well documented and referenced in the literature. Next, we provide a brief description of the MB algorithm.

The spectrum is first divided into a number of frequency bands, from which *the posteriori* segmental SNR is estimated in each band. A subtraction factor is derived according to the segmental SNR in each band. The estimate of the clean speech spectrum $\widehat{S}_i(k)$ at frequency bin k and band i is obtained as follows:

$$|\widehat{S}_i(k)|^2 = |\widehat{Y}_i(k)|^2 - \alpha_i\delta_i|\widehat{D}_i(k)|^2, \quad b_i \leqslant k \leqslant e_i \tag{1}$$

where $\widehat{Y}_i(k)$ is the pre-processed noisy speech spectrum (see Eq. (5)), $\widehat{D}_i(k)$ is the noise spectrum estimate, b_i and e_i are the beginning and ending frequency bins of band i, α_i is the over-subtraction factor and δ_i is a tweaking factor that can be individually set for each frequency band to customize noise removal. Negative values in Eq. (1) are spectrally floored to: $0.002 \cdot |\widehat{Y}_i(k)|^2$. To further mask any remaining musical noise, a small amount of the noisy spectrum is

Table 3
List of the 13 speech enhancement algorithms evaluated

Algorithm	Equation/parameters	Reference
KLT	Eqs. (14), (48)	Hu and Loizou (2003)
*p*KLT	Eq. (34), $v = 0.08$	Jabloun and Champagne (2003)
MMSE-SPU	Eqs. (7) and (51), $q = 0.3$	Ephraim and Malah (1984)
log-MMSE	Eq. (20)	Ephraim and Malah (1985)
logMMSE-ne	Eq. (20)	Ephraim and Malah (1985)
logMMSE-SPU	Eqs. (2), (8), (10), (16)	Cohen (2002)
pMMSE	Eq. (12)	Loizou (2005)
RDC	Eqs. (6), (7), (10), (14) and (15)	Gustafsson et al. (2001)
RDC-ne	Eqs. (6), (7), (10), (14) and (15)	Gustafsson et al. (2001)
MB	Eqs. (4)–(7)	Kamath and Loizou (2002)
WT	Eqs. (11) and (25)	Hu and Loizou (2004)
Wiener-as	Eqs. (3)–(7)	Scalart and Filho (1996)
AudSup	Eqs. (26) and (38), $v_b(i) = 1, 2$ iterations	Tsoukalas et al. (1997)

SPU = speech presence uncertainty, ne = noise estimation.

introduced back to the enhanced spectrum as follows: $|\widehat{S_i}(k)|^2 = |\widehat{S}_i(k)|^2 + 0.05 \cdot |\widehat{Y}_i(k)|^2$, where $|\widehat{S_i}(k)|^2$ is the newly enhanced power spectrum. A total of eight linearly spaced bands were used in Eq. (1), and δ_i was empirically set to

$$\delta_i = \begin{cases} 1, & i = 1 \\ 2.5, & 1 < i < 8 \\ 1.5, & i = 8 \end{cases} \tag{2}$$

The band-specific subtraction factor α_i is a piecewise linear function of the segmental SNR of band i and is calculated as follows (Berouti et al., 1979):

$$\alpha_i = \begin{cases} 5, & \mathrm{SNR}_i < -5 \\ 4 - \frac{3}{20}(\mathrm{SNR}_i), & -5 \leqslant \mathrm{SNR}_i \leqslant 20 \\ 1, & \mathrm{SNR}_i > 20 \end{cases} \tag{3}$$

where

$$\mathrm{SNR}_i\ (\mathrm{dB}) = 10\log_{10}\left(\frac{\sum_{k=b_i}^{e_i}|\widehat{Y}_i(k)|^2}{\sum_{k=b_i}^{e_i}|\widehat{D}_i(k)|^2}\right) \tag{4}$$

Prior to the subtraction operation in Eq. (1), the noisy speech spectrum $|Y_j(k)|$ is pre-processed to reduce the variance of the spectrum estimate using the following weighted spectral average:

$$|\widehat{Y}_j(k)| = \sum_{i=-M}^{M} W_i |Y_{j-i}(k)| \tag{5}$$

where j is the frame index, $|\widehat{Y}_j(k)|$ is the pre-processed noisy speech magnitude spectrum, $|Y_j(k)|$ is the noisy speech magnitude spectrum, $M = 2$ and the filter weights W_i were empirically set to $W = [0.09 \quad 0.25 \quad 0.32 \quad 0.25 \quad 0.09]$. A 20-ms Hamming window with 50% overlap between frames was used in the MB algorithm and in all the other Fourier-transform based algorithms tested.

A voice activity detector (VAD) was used in most of the speech enhancement methods to update the noise spectrum. More precisely, a statistical-model based voice activity detector (VAD) (Sohn et al., 1999) was used to update the noise spectrum during speech-absent periods. The following VAD decision rule was used:

$$\frac{1}{L}\sum_{k=1}^{L-1}\log \Lambda_k \underset{H_0}{\overset{H_1}{\gtrless}} \eta \tag{6}$$

where

$$\Lambda_k = \frac{1}{1+\xi_k}\exp\left\{\frac{\gamma_k \xi_k}{1+\xi_k}\right\} \tag{7}$$

where ξ_k and γ_k are defined as in (Ephraim and Malah, 1984), and ξ_k is estimated using the decision directed approach ($\alpha = 0.98$). L is the size of the FFT, H_1 denotes the hypothesis of speech presence, H_0 denotes the hypothesis of speech absence, and η is a preset threshold. In our implementation, $\eta = 0.15$ for all conditions. During the speech-absent periods, i.e., when the left side of Eq. (6) was smaller than η, the noise power spectrum was updated according to

$$N_j(k) = (1-\beta)|Y_j(k)|^2 + \beta N_{j-1}(k) \tag{8}$$

where $N_j(k)$ is the estimate of the noise power spectrum at frame j for frequency bin k, $\beta = 0.98$ is a preset smoothing factor, and $|Y_j(k)|$ is the noisy speech magnitude spectrum. The initial estimate of $N_j(k)$ was obtained from the first (speech-absent) 120-ms segment of each sentence.

The subspace methods used the VAD method proposed in (Mittal and Phamdo, 2000) with the threshold value set to 1.2. The frame windowing scheme proposed in (Jabloun and Champagne, 2003) was adopted in both VAD methods. More specifically, the signal was divided into 32-ms frames with 50% overlap between frames. The samples in the 32-ms frame were used to construct a 32×32 Toeplitz covariance matrix. The 32-ms frames were further subdivided into 4-ms frames with 50% overlap. The noisy data in each 4-ms frame were enhanced using the same eigenvector matrix derived from the 32×32 Toeplitz covariance matrix.

Table 3 lists all the algorithms evaluated along with the associated parameters and equations. MATLAB implementations of all the algorithms tested are available in Loizou (2007).

4. Subjective evaluation

To reduce the length and cost of the subjective evaluations, only a subset of the NOIZEUS corpus was processed by the 13 algorithms and submitted to Dynastat, Inc., for formal subjective evaluation. A total of 16 sentences (see Tables 1 and 2) corrupted in four background noise environments (car, street, babble and train) at two levels of SNR (5 dB and 10 dB) were processed. These sentences

were produced by two male speakers and two female speakers.

4.1. Test methodology

The subjective tests were designed according to ITU-T recommendation P.835. The P.835 methodology was designed to reduce the listener's uncertainty in a subjective test as to which component(s) of a noisy speech signal, i.e., the speech signal, the background noise, or both, should form the basis of their ratings of overall quality. This method instructs the listener to successively attend to and rate the enhanced speech signal on:

(1) the speech signal alone using a five-point scale of signal distortion (SIG) (Table 4),
(2) the background noise alone using a five-point scale of background intrusiveness (BAK) (Table 5),
(3) the overall effect using the scale of the Mean Opinion Score (OVRL) – [1 = bad, 2 = poor, 3 = fair, 4 = good, 5 = excellent].

The process of rating the signal and background of noisy speech was designed to lead the listener to integrate the effects of both the signal and the background in making their ratings of overall quality. Each trial in a P.835 test involved a triad of speech samples, where each sample consisted of a single sentence recorded in background noise. For each sample within the triad, listeners successively used one of the three five-point rating scales (SIG, BAK, and OVRL) to register their judgments of the quality of the test condition. In addition to the experimental conditions, each experiment included a number of reference conditions designed to independently vary the listener's SIG, BAK, and OVRL ratings over the entire five-point range of the rating scales.

4.2. Preparation of test sequences

The single-sentence sample files were concatenated into four triads for each of the talkers and for each condition. The P.835 standard permits the use of triads made up of either three different samples or the same sample repeated three times. For this experiment, the same sample was used three times in each triad. As per the P.835 standard, for half of the trials in the experiment the rating scale order was SIG, BAK, and OVRL, and for the other half of the trials the order was BAK, SIG, and OVRL.

The total number of test conditions was too large to present in a single P.835 test. Therefore, the test conditions were partitioned into two sub-sets which were evaluated in two separate P.835 tests. The conditions were assigned to the two tests such that the primary factors involved in the experiment (algorithm, noise type, SNR) would be confounded at the highest order interaction. Each of the two tests involved the various test conditions and the 12 standard P.835 reference conditions. Each of the test and reference conditions was represented by files from four talkers arranged in four triads. The files within each test were allocated to four presentation sets under a partially balanced/randomized-blocks experimental design. In each presentation set, the samples were ordered in a pseudo-randomized balanced-block presentation sequence to control for the effects of time and order of presentation.

Table 4
Scale of signal distortion (SIG)

5 – Very natural, no degradation
4 – Fairly natural, little degradation
3 – Somewhat natural, somewhat degraded
2 – Fairly unnatural, fairly degraded
1 – Very unnatural, very degraded

Table 5
Scale of background intrusiveness (BAK)

5 – Not noticeable
4 – Somewhat noticeable
3 – Noticeable but not intrusive
2 – Fairly conspicuous, somewhat intrusive
1 – Very conspicuous, very intrusive

4.3. Listening panels

A total of 32 listeners were recruited for the listening tests. For each of the two P.835 tests, each of the four presentation sequences was presented to a separate panel of eight naive listeners. Listeners were recruited from Dynastat's database of native speakers of North American English. Listeners were between the ages of 18 and 50 years of age. No listener had participated in a listening test in the previous three months. The listening panels in the two experiments were independent, i.e., no listener participated in more than one experiment.

4.4. Audio presentation

The processed speech material were presented to listeners seated at separate, visually screened listening stations in a soundproof room. Speech materials were presented monaurally via Sennheiser HD-25 supra-aural headphones. Subjects were instructed to use the headphone on their preferred listening ear. The other ear was open and a constant ambient noise floor was maintained at 30 dBA using Hoth noise (ITU-T P.835, 2003). The headphones were driven by a distribution amplifier set to deliver active speech at a level of 79 dB Sound Pressure Level (SPL) at the ear reference plane. Headphones were calibrated with a B&K 4153 Artificial Ear with supra-aural headphone adapter, a 4134 microphone element and a 2609 measurement amplifier. The processed speech files were channelled through a Townshend Computer Tools DAT-Link+ and recorded on Digital Audio Tape (DAT) for presentation to the listening panels. In each listening station the rating scales were presented on a PC monitor and ratings were registered with a PC keyboard.

4.5. Test sessions

The tests lasted approximately 1.25 h. Listeners took short breaks (10 min) between sessions. At the beginning of session 1, the listeners were presented with a practice block of 12 trials to familiarize them with the task and the timing in the trial presentation. The practice blocks were also designed to present the listeners with the range of conditions that would be involved in the tests on both the signal and the background scales. For each test, half the panels were presented with trials in which the rating scale order was SIG–BAK–OVRL for the first two sessions and BAK–SIG–OVRL for sessions 3 and 4. To train the listeners for the change in scale order, listeners were presented with the practice block again at the beginning of session 3. For the other half of the panels, the sessions and scale order was counter-balanced.

4.6. Evaluation results

Figs. 3–6 show the mean scores for the SIG, BAK, and OVRL scales for speech processed by 13 different speech enhancement algorithms evaluated in four types of background noise and at two SNR levels (5 dB and 10 dB). The mean scores for the noisy speech (unprocessed) files are also shown for reference.

5. Statistical analysis and discussion

We present comparative analysis at three levels. At the first level, we compare the performance of the algorithms within each of the four classes (subspace, statistical-model, subtractive, and Wiener-type). This comparison was meant to examine whether there were significant differences

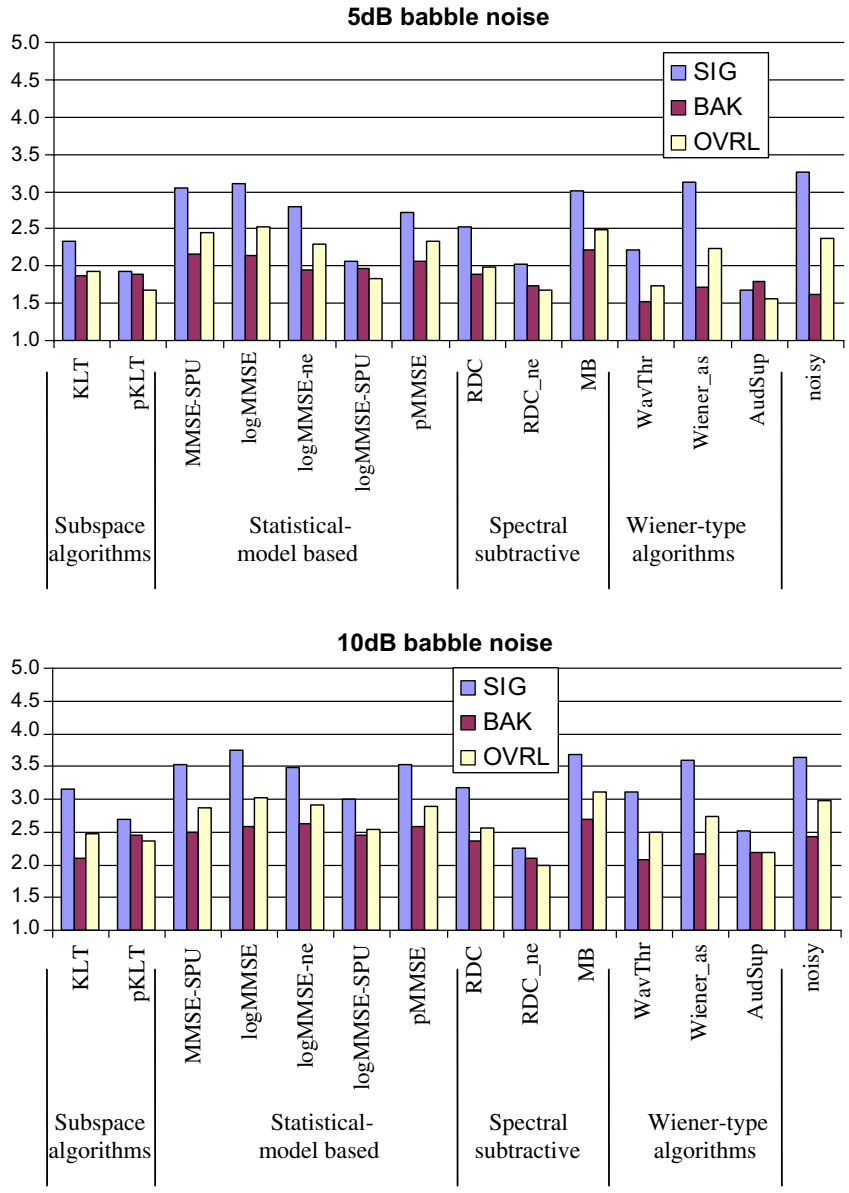

Fig. 3. The mean scores for SIG, BAK, and OVRL scales for the 13 methods evaluated in babble noise background and for SNR levels of 5 dB and 10 dB.

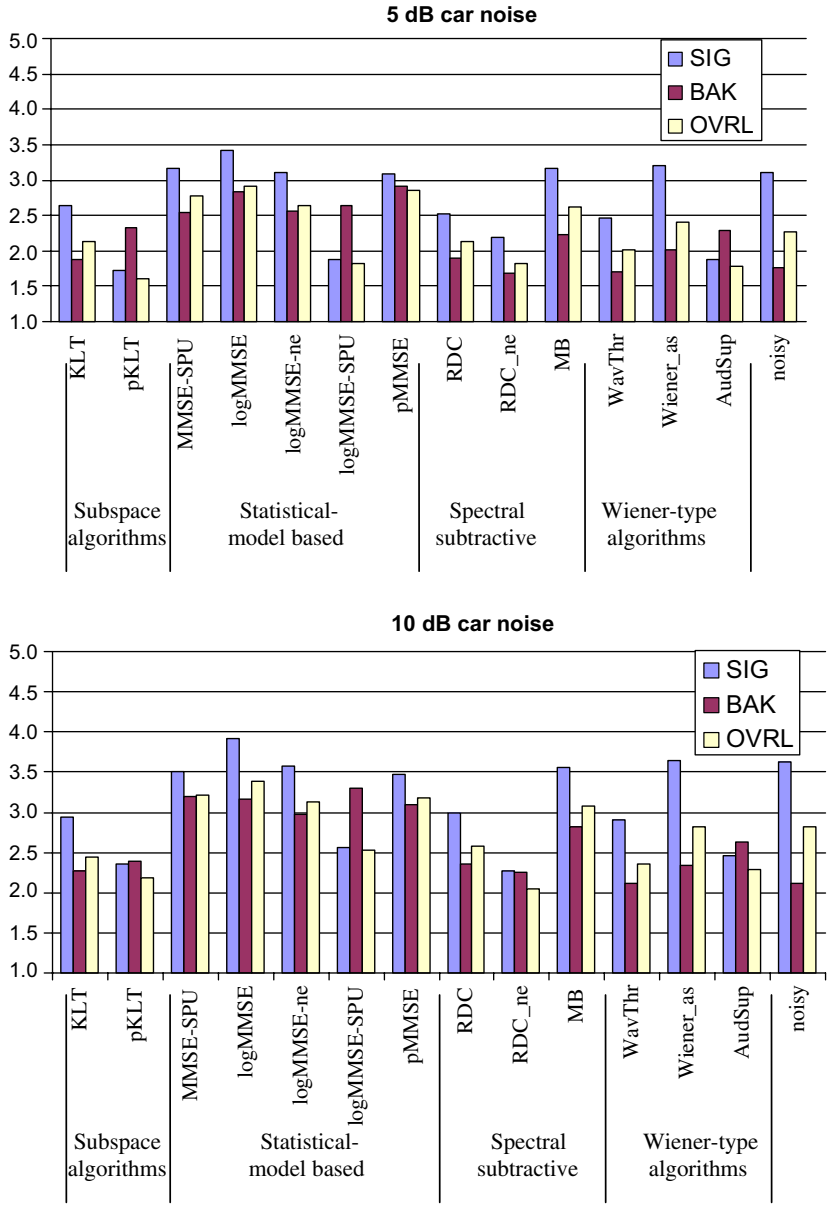

Fig. 4. The mean scores for SIG, BAK, and OVRL scales for the 13 methods evaluated in car noise background and for SNR levels of 5 dB and 10 dB.

between algorithms within each class. At the second 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. Lastly, at the third 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 algorithm(s) improved significantly the quality of noisy speech.

In order to assess significant differences between the ratings obtained with each algorithm, we subjected the ratings of the 32 listeners to statistical analysis. Analysis of variance (ANOVA) indicated a highly significant effect ($F(13,403) = 20.17$, $p < 0.0005$) of speech enhancement algorithms on the ratings of signal, noise and overall quality (a highly significant effect was also found in all SNR conditions and types of noise). Following the ANOVA, we conducted multiple comparison statistical tests according to Tukey's HSD 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.

5.1. Within-class algorithm comparisons

In terms of overall quality, the two subspace algorithms performed equally well for most SNR conditions and four types of noise, except at 5 dB car noise. The generalized subspace approach (Hu and Loizou, 2003) performed significantly ($p = 0.006$) better than the pKLT approach (Jabloun and Champagne, 2003) in 5 dB car noise. Lower noise distortion (i.e., higher BAK scores) was observed with the

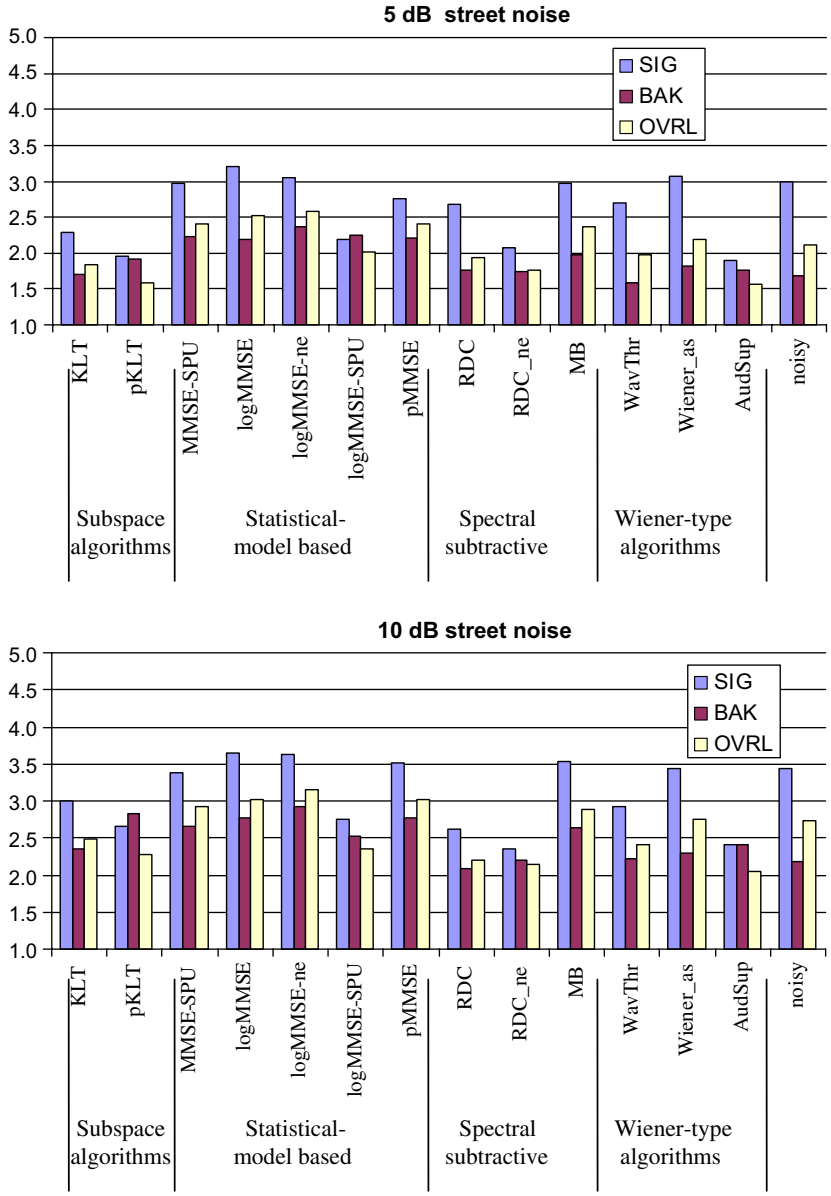

Fig. 5. The mean scores for SIG, BAK, and OVRL scales for the 13 methods evaluated in street noise background and for SNR levels of 5 dB and 10 dB.

*p*KLT method in most conditions, however, the difference in scores was not found to be statistically significant. Significantly ($p = 0.017$) lower noise distortion (i.e., higher BAK scores) was observed with the *p*KLT method only in 5 dB train noise. Lower signal distortion was generally observed with the generalized subspace method in most conditions with significant differences in 5 dB train noise and in both car noise conditions (5 dB and 10 dB). In brief, of the two subspace methods, the generalized subspace approach performed slightly better in terms of overall quality and lower signal distortion. The *p*KLT approach was more successful in suppressing background noise, however at the expense of introducing signal distortion.

The majority of the statistical-model based algorithms examined performed equally well in terms of overall quality. There was no statistically significant difference in overall quality between the MMSE-SPU, the log-MMSE, the log-MMSE with noise estimation (logMMSE-ne) and the pMMSE algorithms. The logMMSE algorithm that incorporated signal-presence uncertainty (logMMSE-SPU) (Cohen, 2002) performed significantly worse than the other algorithms in overall quality. This was surprising at first, but close analysis indicated that the logMMSE-SPU algorithm was sensitive to the noise spectrum estimate, which in our case was obtained with a VAD algorithm. Furthermore, the parameters given in (Cohen, 2002) were appropriate for a sampling rate of 16 kHz, while the present performance evaluation involved a sampling rate of 8 kHz. Hence, the experimental results do not necessarily represent the best performance obtainable with the logMMSE-SPU algorithm. Indeed, subsequent listening tests (conducted after Dynastat's subjective evaluation) confirmed that the

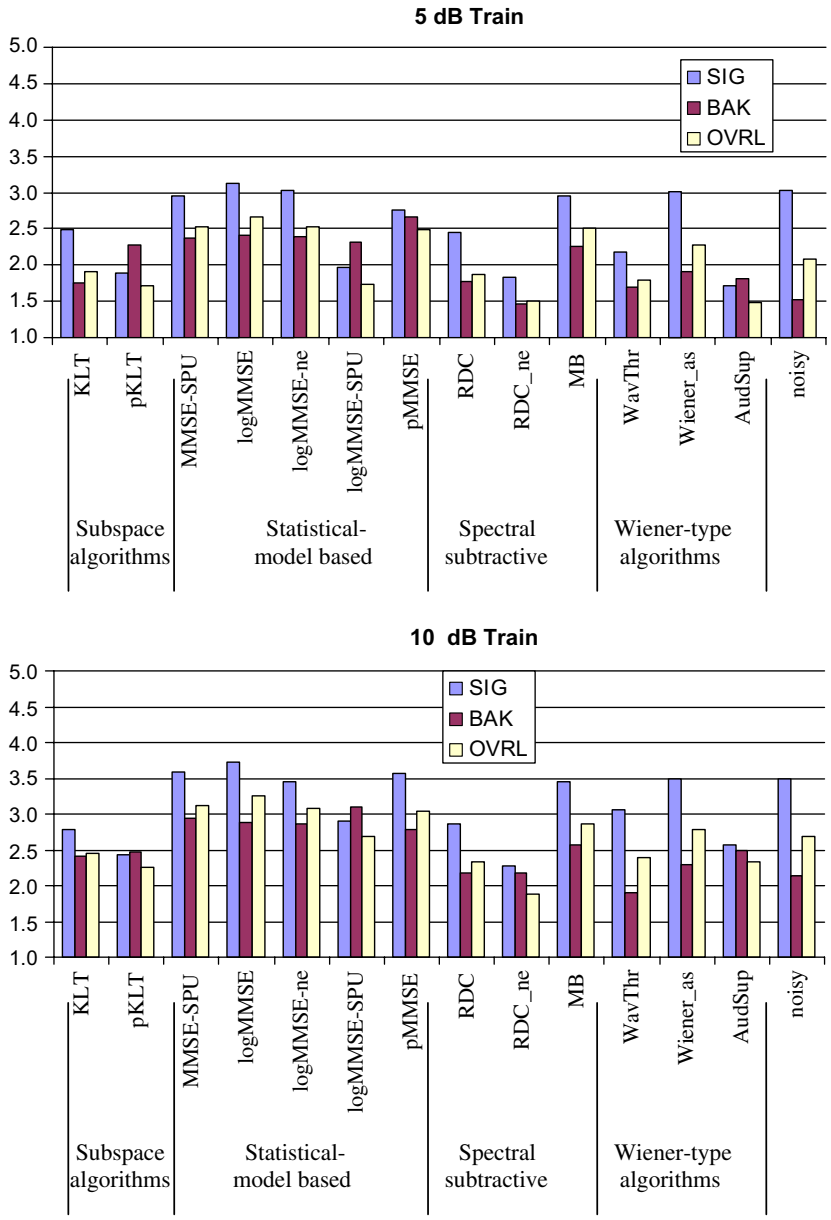

Fig. 6. The mean scores for SIG, BAK, and OVRL scales for the 13 methods evaluated in train noise background and for SNR levels of 5 dB and 10 dB.

logMMSE-SPU algorithm[2] performed better than the logMMSE algorithm when a noise-estimation algorithm (Cohen, 2003) was used to update the noise spectrum.

In terms of noise distortion, all algorithms, including the logMMSE-SPU algorithm, performed equally well. Lower noise distortion (i.e., higher BAK scores) was obtained with the pMMSE method (compared to the MMSE-SPU method) in some conditions (5 dB train, 5 dB car, 10 dB street), however the difference was not statistically significant ($p > 0.05$). In terms of speech distortion, nearly all algorithms (MMSE-SPU, log-MMSE, logMMSE-ne and pMMSE algorithms) performed equally well. Incorporating a noise estimation algorithm in the logMMSE method did not produce significant improvements in performance. One explanation for that is that the duration of the sentences was too short to observe the real benefit of noise-estimation algorithms. In brief, with the exception of the logMMSE-SPU algorithm, the MMSE algorithms performed equally well in overall quality, signal and noise distortion.

Of the three spectral-subtractive algorithms tested, the multi-band spectral subtraction algorithm (Kamath and Loizou, 2002) performed consistently the best across all conditions, in terms of overall quality. In terms of noise distortion, the MB and RDC algorithms performed equally well except in 5 dB train and 10 dB street conditions, in which the multi-band algorithm performed significantly

[2] We would like to thank Dr. Cohen for providing his code with the implementation of the logMMSE-SPU algorithm with noise estimation.

better (i.e., lower noise distortion). Performance (noise distortion) of the RDC algorithm that included noise estimation (RDC-ne) was significantly lower than the MB algorithm in all conditions. There was no real benefit, in terms of overall quality, of including noise estimation in the RDC method. In terms of speech distortion, the MB and RDC algorithms performed equally well in most conditions except in 5 dB car noise and in 10 dB street noise, in which the MB algorithm performed significantly better (i.e., lower speech distortion). In brief, the MB algorithm generally performed better than the RDC algorithm in overall quality, signal and noise distortion. It should be pointed out, however, that the MB algorithm has an unfair advantage over the RDC algorithm in that it uses non-causal filtering (Eq. (5)) to smooth out the noisy speech spectra.

Finally, of the three Wiener-filtering type algorithms examined, the Wiener-as and WT algorithms performed the best. In terms of overall quality, the Wiener-as method performed better than the WT method in three conditions: 5 dB train, 10 dB car and 5 dB babble noise. In the remaining five conditions, the Wiener-as method performed as well as the WT method (Hu and Loizou, 2004). The Wiener-as method also produced consistently lower signal distortion for most conditions, except in 10 dB train, 10 dB babble and street conditions, in which it performed equally well with the WT method. All three Wiener-type algorithms produced the same level of noise distortion in all conditions. In brief, the Wiener-as method performed, for the most part, better than the other Wiener algorithms in terms of overall quality and signal distortion.

5.2. Across-class algorithm comparisons

The above comparisons assessed differences between algorithms within each of the four classes of speech enhancement methods, but did not provide the answer as to which algorithm(s) performed the best overall across all noise conditions. Such comparisons are reported in this section.

Multiple paired comparisons (Tukey's HSD) were conducted between the algorithm with the highest score against all other algorithms. Tables 6–8 report the results for the overall quality, signal distortion and noise distortion comparisons respectively.

Table 6 shows the results obtained from the statistical analysis for overall quality. 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 6 performed equally well. It is clear from Table 6, that there is no single best algorithm, but rather that several algorithms performed equally well across most conditions. In terms of overall quality, the following algorithms performed equally well across all conditions: MMSE-SPU, logMMSE, logMMSE-ne, pMMSE and MB. The Wiener-as method also performed well in five of the eight conditions (see Table 6).

Table 7 shows the results obtained from the statistical analysis for signal distortion. The following algorithms performed the best, in terms of yielding the lowest speech distortion, across all conditions: MMSE-SPU, logMMSE, logMMSE-ne, pMMSE, MB and Wiener-as. The KLT, RDC and WT algorithms also performed well in a few isolated conditions (see Table 7).

Finally, Table 8 shows the results obtained from the statistical analysis for noise distortion. The following algorithms performed the best, in terms of yielding the lowest noise distortion across nearly all conditions: MMSE-SPU, logMMSE, logMMSE-ne, pMMSE and MB. The *p*KLT method also performed well in five of the eight conditions. The KLT, RDC, RDC-ne, Wiener-as and AudSup algorithms performed well in a few isolated conditions (see Table 8).

Comparing the results in Tables 6–8, we observe that the algorithms that yielded the lowest noise distortion (i.e., lowest noise residual) were not necessarily the algorithms that yielded the highest overall quality. The *p*KLT algorithm, for instance, performed well in terms of noise distortion, but performed poorly in terms of overall quality and speech distortion. In contrast, the algorithms that performed the best in terms of speech distortion were also the algorithms with the highest overall quality. This suggests that listeners are influenced more by the distortion imparted on the

Table 6
Results obtained from comparative statistical analysis of overall quality (OVRL) scores

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB			*	*	*		*			*			
	10 dB			*	*	*		*			*			
Babble	5 dB			*	*	*		*			*		*	
	10 dB			*	*	*		*			*		*	
Street	5 dB			*	*	*		*			*		*	
	10 dB			*	*	*		*			*		*	
Train	5 dB			*	*	*		*			*		*	
	10 dB			*	*	*		*			*			

Algorithms indicated by asterisks performed equally well. Algorithms with no asterisks performed poorly.

Table 7
Results obtained from comparative statistical analysis of speech distortion (SIG) scores

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB			*	*	*		*			*		*	
	10 dB			*	*	*		*			*		*	
Babble	5 dB			*	*	*		*			*		*	
	10 dB	*		*	*	*		*	*		*		*	
Street	5 dB			*	*	*		*	*		*	*	*	
	10 dB			*	*	*		*			*		*	
Train	5 dB			*	*	*		*			*		*	
	10 dB			*	*	*		*			*		*	

Algorithms indicated by asterisks performed equally well. Algorithms with no asterisks performed poorly.

Table 8
Results obtained from comparative statistical analysis of noise distortion (BAK) scores

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB			*	*	*	*	*						
	10 dB			*	*	*	*	*			*			
Babble	5 dB	*	*	*	*	*	*	*	*	*	*		*	*
	10 dB		*	*	*	*	*	*	*		*			*
Street	5 dB		*	*	*	*	*	*			*			
	10 dB		*	*	*	*	*	*			*			
Train	5 dB		*	*	*	*	*	*			*			
	10 dB			*	*	*	*	*						

Algorithms indicated by asterisks performed equally well. Algorithms with no asterisks performed poorly.

speech signal than on the background noise when making judgments of overall quality (more on this in Section 5.4). That is, listeners seem to place more emphasis on speech distortion rather on noise distortion when judging the quality of speech enhanced by a noise suppression algorithm.

5.3. Comparisons in reference to noisy speech

Lastly, we report on the comparisons between the enhanced speech and the noisy (unprocessed) speech. Such comparisons are important as they tell us about the possible benefits (or lack thereof) of using speech enhancement algorithms.

Multiple paired comparisons (Tukey's HSD) were conducted between the ratings obtained with noisy speech (unprocessed) samples and the ratings obtained with speech enhanced by the various algorithms. The results are reported in Tables 9–11 for overall quality, signal distortion and noise distortion comparisons respectively. In these tables, asterisks indicate significant differences (i.e., significant benefit) between the ratings of noisy speech and enhanced speech. Table entries indicated as 'ns' denote non-significant differences between the ratings of noisy speech and enhanced speech, i.e., noisy and enhanced speech were rated equally. Blank entries in the Tables indicate inferior ratings (i.e., significantly poorer ratings) for the enhanced speech compared to the ratings of noisy speech samples.

Table 9 shows the comparisons of the ratings of overall quality of noisy speech and enhanced speech. The striking finding is that only a subset of the algorithms tested provided significant benefit to overall quality and only in a few conditions (car, street and train). The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, and pMMSE improved significantly the overall speech quality but only in a few isolated conditions. The majority of the algorithms (indicated with 'ns' in Table 9) did not provide significant improvement in overall quality when compared to the noisy (unprocessed) speech.

Table 10 shows the comparisons of the ratings of signal distortion of noisy speech and enhanced speech. For this comparison, we do not expect to see any asterisks in the Table. Good performance is now indicated with 'ns', suggesting that the enhanced speech did not contain any notable speech distortion. The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, pMMSE, MB and Wiener-as performed the best (i.e., no notable speech distortion was introduced) in all conditions. The algorithms WT, RDC and KLT also performed well in a few isolated conditions.

Table 11 shows the comparisons of the ratings of noise distortion of noisy speech and enhanced speech. The algorithms MMSE-SPU, log-MMSE, logMMSE-ne, log-MMSE-SPU and pMMSE lowered significantly noise

Table 9
Statistical comparisons between the ratings of overall quality of noisy (unprocessed) speech and enhanced speech

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB	ns		*	*	ns		*	ns		ns	ns	ns	
	10 dB	ns		ns	*	ns	ns	ns	ns		ns		ns	
Babble	5 dB	ns		ns	ns	ns		ns	ns		ns		ns	
	10 dB			ns	ns	ns	ns	ns	ns		ns		ns	
Street	5 dB	ns		ns	ns	*	ns	ns	ns	ns	ns	ns	ns	
	10 dB	ns	ns	ns	ns	ns	ns	ns			ns	ns	ns	
Train	5 dB	ns	ns	*	*		ns	ns	ns		ns	ns	ns	
	10 dB	ns	ns	ns	*	ns	ns	ns	ns		ns	ns	ns	ns

Algorithms denoted with asterisks improved significantly the overall quality of noisy speech. That is, the quality of the enhanced speech was judged to be significantly better than that of noisy speech. In contrast, algorithms denoted with 'ns' did not improve the overall quality of noisy speech.

Table 10
Statistical comparisons between the ratings of speech distortion (SIG) of noisy (unprocessed) speech and enhanced speech

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB	ns		ns	ns	ns		ns	ns		ns		ns	
	10 dB			ns	ns	ns		ns			ns		ns	
Babble	5 dB			ns	ns	ns		ns			ns		ns	
	10 dB	ns		ns	ns	ns		ns	ns		ns	ns	ns	
Street	5 dB			ns	ns	ns		ns	ns		ns	ns	ns	
	10 dB	ns		ns	ns	ns		ns			ns	ns	ns	
Train	5 dB	ns		ns	ns	ns		ns			ns		ns	
	10 dB			ns	ns	ns	ns	ns			ns		ns	ns

Algorithms denoted with 'ns' did not introduce notable speech distortion. In contrast, algorithms with blank table entries introduced notable speech distortion when compared to the noisy (unprocessed) speech.

Table 11
Statistical comparisons between the ratings of noise distortion (BAK) of noisy (unprocessed) speech and enhanced speech

		KLT	*p*KLT	MMSE-SPU	log-MMSE	logMMSE-ne	logMMSE-SPU	pMMSE	RDC	RDC-ne	MB	WT	Wiener-as	AudSup
Car	5 dB	ns	*	*	*	*	*	*	ns	ns	ns	ns	ns	*
	10 dB	ns	ns	*	*	*	*	*	ns	ns	*	ns	ns	*
Babble	5 dB	ns	ns	*	*	ns	ns	ns	ns	ns	*	ns	ns	ns
	10 dB	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns	ns
Street	5 dB	ns	ns	*	*	*	*	*	ns	ns	ns	ns	ns	ns
	10 dB	ns	*	ns	*	*	ns	*	ns	ns	ns	ns	ns	ns
Train	5 dB	ns	*	*	*	*	*	*	ns	ns	*	ns	ns	ns
	10 dB	ns	ns	*	*	*	*	*	ns	ns	ns	ns	ns	ns

Algorithms denoted with asterisks significantly lowered noise distortion compared to that of un-processed noisy speech. In contrast, algorithms denoted with 'ns' did not lower noise distortion.

distortion for most conditions. The MB, *p*KLT and AudSup also lowered noise distortion in a few (2–3) conditions. The remaining algorithms (indicated with 'ns' in Table 11) did not produce significantly lower noise distortion compared to the noisy (unprocessed) speech. That is, the background noise level was not perceived to be significantly lower in the enhanced speech than the noisy speech.

5.4. Contribution of speech and noise distortion to judgment of overall quality

As mentioned earlier, the P.835 process of rating the signal and background of noisy speech was designed to lead the listener to integrate the effects of both the signal and the background in making their ratings of overall quality.

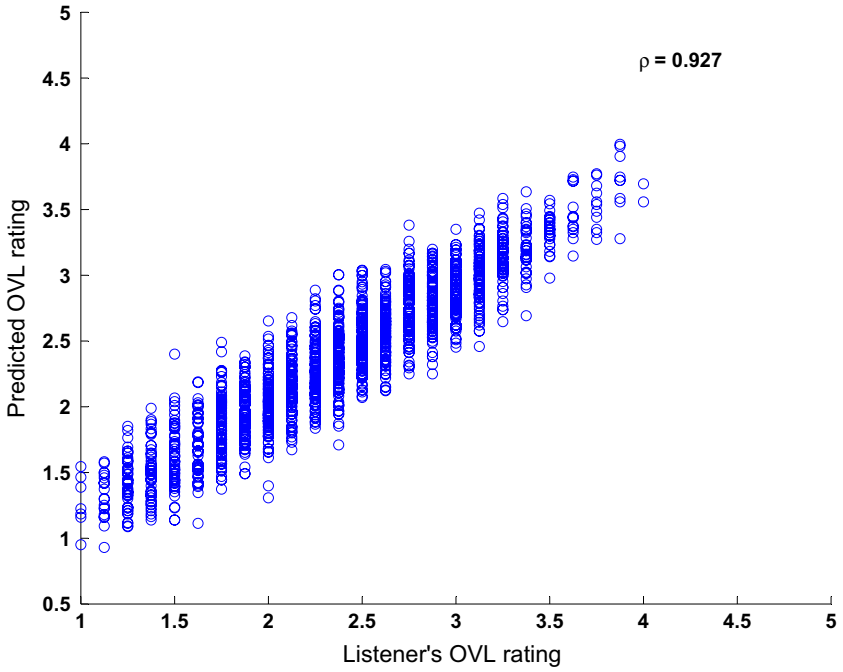

Fig. 7. Regression analysis of listener's OVRL ratings, based on SIG and BAK ratings.

Of great interest is finding out the individual contribution of speech and noise distortion to judgment of overall quality. Our previous data (Tables 6 and 7) led us to believe that listeners were influenced more by speech distortion when making quality judgments. To further substantiate this, we performed multiple linear regression analysis on the ratings obtained for overall quality, speech and noise distortion. We treated the overall quality score as the dependent variable and the speech and noise scores as the independent variables. Regression analysis revealed the following relationship between the three rating scales:

$$R_{\mathrm{OVL}} = -0.0783 + 0.571 \cdot R_{\mathrm{SIG}} + 0.366 \cdot R_{\mathrm{BAK}} \quad (9)$$

where R_{OVL} is the predicted overall (OVRL) rating score, R_{SIG} is the SIG rating and R_{BAK} is the BAK rating. The resulting correlation coefficient was $\rho = 0.927$ and the standard error of the estimate was 0.22. Fig. 7 shows the scatter plot of the listener's overall quality ratings against the predicted ratings obtained with Eq. (9). The above equation confirms that listeners were indeed integrating the effects of both signal and background distortion when making their ratings. Different emphasis was placed, however, on the two types of distortion. Consistent with our previous observation, listeners seem to place more emphasis on the distortion imparted on the speech signal itself rather than on the background noise, when making judgments of overall quality.

6. Conclusions

The present study reported on the subjective evaluation of 13 different speech enhancement algorithms using the ITU-T P.835 methodology designed to evaluate the speech quality along three dimensions: signal distortion, noise distortion and overall quality. A total of 32 listeners participated in the listening tests. Based on the statistical analysis of the listener's ratings of the enhanced speech, in terms of overall quality, speech and noise distortion, we can draw the following conclusions:

(1) In terms of overall quality and speech distortion, the following algorithms performed the best: MMSE-SPU, logMMSE, logMMSE-ne, pMMSE and MB. The Wiener-as method also performed well in some conditions. The subspace algorithms performed poorly.
(2) The algorithms that performed the best in terms of yielding low speech distortion were also the algorithms yielding the highest overall quality. This suggests that listeners were influenced for the most part by the distortion imparted on the speech signal than on the background noise when making judgments of overall quality. This was also confirmed by regression analysis (Eq. (9)).
(3) Incorporating noise estimation algorithms in place of VAD algorithms for updating the noise spectrum did not produce significant improvements in performance. One explanation for that is that the duration of the sentences was too short to observe the real benefit of noise-estimation algorithms.
(4) Comparisons of ratings of the overall quality of noisy (unprocessed) speech against that of enhanced (processed) speech revealed that only a subset of the algorithms tested provided significant benefit to overall quality and only in a few conditions (car, street and

train). No algorithm produced significant quality improvement in multi-talker babble, i.e., in highly nonstationary environments.

(5) In terms of low computational complexity and good performance, the two winners were the Wiener-as and multi-band spectral subtraction algorithms. Unlike the Wiener-as method which relies on the decision-directed approach to estimate the a priori SNR, the multi-band spectral subtraction algorithm does not make use of a priori SNR information. Yet, the multi-band spectral subtraction algorithm performed as well as the statistical-model based algorithms in nearly all conditions (Tables 6–8).

Acknowledgements

The authors would like to thank Dr. Alan Sharpley of Dynastat, Inc., for all his help and advice throughout the duration of this project. Research was supported in part by Grant No. R01 DC07527 from NIDCD/NIH.

References

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp. 208–211.

Cohen, I., 2002. Optimal speech enhancement under signal presence uncertainty using log-spectral amplitude estimator. IEEE Signal Process. Lett. 9, 113–116.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Trans. Speech Audio Proc., 466–475.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. ASSP-32, 1109–1121.

Ephraim, Y., 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., Claesson, I., 2001. Spectral subtraction using reduced delay convolution and adaptive averaging. IEEE Trans. Speech Audio Proc., 799–807.

Hirsch, H., Pearce, D., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ISCA ITRW ASR2000, Paris, France.

Hu, Y., Loizou, P.C., 2003. A generalized subspace approach for enhancing speech corrupted by colored noise. IEEE Trans. Speech Audio Proc., 334–341.

Hu, Y., Loizou, P.C., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Proc., 59–67.

IEEE Subcommittee, 1969. IEEE recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust., 225–246.

ITU-T P.56, 1993. Objective measurement of active speech level. ITU-T Recommendation P.56.

ITU-T P.835, 2003. Subjective test methodology for evaluating speech communication systems that include noise suppression algorithm. ITU-T Recommendation P.835.

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

Jabloun, F., Champagne, B., 2003. Incorporating the human hearing properties in the signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Proc. 11, 700–708.

Kamath, S., 2001. A multi-band spectral subtraction method for speech enhancement. Masters thesis, University of Texas at Dallas.

Kamath, S., Loizou, P.C., 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In: Student Research Abstracts of Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing. [Online: http://www.utdallas.edu/~loizou/speech/].

Loizou, P.C., 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. IEEE Trans. Speech Audio Proc., 857–869.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press, Boca Raton, FL.

Mittal, U., Phamdo, N., 2000. Signal/noise KLT based approach for enhancing speech degraded by colored noise. IEEE Trans. Speech Audio Proc. 8, 159–167.

Rangachari, S., Loizou, P.C., 2006. A noise estimation algorithm for highly non-stationary environments. Speech Commun., 220–231.

Scalart, P., Filho, J., 1996. Speech enhancement based on a priori signal to noise estimation. In: Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., pp. 629–632.

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

Tsoukalas, D.E., Mourjopoulos, J.N., Kokkinakis, G., 1997. Speech enhancement based on audible noise suppression. IEEE Trans. Speech Audio Proc. 5, 479–514.