Effects of noise suppression on intelligibility: Dependency on signal-to-noise ratios[a)]

Gaston Hilkhuysen[b)]
Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom

Nikolay Gaubitch and Mike Brookes
Electrical and Electronic Engineering Department, Imperial College, Exhibition Road, London SW7 2BT, United Kingdom

Mark Huckvale
Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, United Kingdom

(Received 2 December 2010; revised 3 November 2011; accepted 10 November 2011)

The effects on speech intelligibility of three different noise reduction algorithms (spectral subtraction, minimal mean squared error spectral estimation, and subspace analysis) were evaluated in two types of noise (car and babble) over a 12 dB range of signal-to-noise ratios (SNRs). Results from these listening experiments showed that most algorithms deteriorated intelligibility scores. Modeling of the results with a logit-shaped psychometric function showed that the degradation in intelligibility scores was largely congruent with a constant shift in SNR, although some additional degradation was observed at two SNRs, suggesting a limited interaction between the effects of noise suppression and SNR. © *2012 Acoustical Society of America*. [DOI: 10.1121/1.3665996]

PACS number(s): 43.72.Kb, 43.72.Dv [EB] Pages: 531–539

I. INTRODUCTION

Many methods for the reduction of added noise have been designed to improve noisy speech signals (for a review, see Loizou, 2007). The majority of these noise suppression methods have a similar general structure. The noisy speech is first divided into overlapping time frames and a front-end process estimates the noise power spectrum. A subsequent back-end process then applies an attenuation factor that varies in time and frequency as a function of the estimated signal-to-noise ratio (SNR). In general, listeners give noisy speech after noise reduction higher quality ratings than the nonprocessed signal (Loizou, 2007).

Despite the large number of proposed algorithms, only a few studies have addressed the consequences of noise suppression for speech intelligibility (Lim, 1978; Boll, 1979; Ludvigsen *et al.*, 1993; Tsoukalas *et al.*, 1997; Arehart *et al.*, 2003; Loizou, 2007; Hu and Loizou, 2007). In some of these studies, an occasional improvement in intelligibility was reported (Tsoukalas *et al.*, 1997; Arehart *et al.*, 2003; Hu and Loizou, 2007), but in general, noise suppression either had little or a detrimental effect on intelligibility. In terms of the range of SNR values tested, previous studies only included one (Boll, 1979), two (Ludvigsen *et al.*, 1993; Arehart *et al.*, 2003; Loizou, 2007; Hu and Loizou, 2007), or at times three (Lim, 1978; Tsoukalas *et al.*, 1997) SNR values per noise type. Loizou (2007) found that for a majority of noise reduction algorithms, percentage word-correct scores reduced more at lower SNRs, suggesting that the deteriorating effects of speech enhancement on intelligibility increases at lower SNRs.

The impact of an algorithm on intelligibility may vary considerably with SNR. The effect of an algorithm might depend on the intelligibility of the input signal, and it may not be the case that effects observed at a particular level of intelligibility can be generalized to other levels or that shifts in intelligibilities, caused by noise suppression and observed at two or more levels of intelligibility, can be interpolated or extrapolated to other levels. One might expect an interaction between SNR and processing due to inaccuracies in estimating specific parameters used in noise-suppression algorithms at different SNRs. For example, it may be easier to estimate accurately some noise-reduction parameters, such as the SNR or noise spectrum, at high SNRs compared to low SNRs. Consequently, the noise-reduction algorithm could be more effective at high SNRs than at low SNRs. In this paper, we investigate whether such SNR dependent effects of noise suppression on intelligibility exist. It will be assumed that the functions relating SNR to percentage correct scores, the so called psychometric functions (PMf), have logistic shapes. Then if the effects of processing are independent of SNR, the PMf should simply shift along the SNR axis as a consequence of noise reduction. Changes in the PMf's slope would be considered as SNR dependent effects.

A better understanding of SNR dependent effects on intelligibility could be of benefit to designers and users of noise suppression methods and could support the development of predictive models of intelligibility that might be used to design new noise suppression algorithms or to optimize the application of existing algorithms for particular signal

[a)] Part of this work was presented at the British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness, Southampton UK, 2009.

[b)] Author to whom correspondence should be addressed. Electronic mail: g.hilkhuysen@ucl.ac.uk

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

conditions. Additionally, if the effects of noise suppression are stable across SNRs, one may consider measuring these effects in the future with more time-efficient adaptive procedures, as for example proposed by Leek (2001). However, to be efficient, these adaptive procedures typically focus on SNRs that are much lower than the SNRs relevant for most human communication (Brand and Kollmeier, 2002). If effects do vary with SNR, the outcome of such tests may not generalize to the SNRs of interest.

The effects of noise suppression at low SNR are particularly of interest for applications in law-enforcement and forensic audio, where recordings are often made in adverse conditions (Manchester, 2010). Low SNR values can also be found in the mobile telecommunications area, where conversations tend to be held in noisier conditions than previously encountered in the wired network (Jellyman, 2009). In this paper, we collect a wider range of intelligibility data than before, including three different noise suppression methods for two different noise types over five SNRs.

II. METHODS

A. Participants

Sixty participants were recruited from the Psychology Subject Pool of University College London (UCL), staff and students at the UCL department of Speech, Hearing and Phonetic Sciences, and collaborators within the Centre for Law Enforcement Audio Research. All participants indicated they attended primary school in the UK and used English as their principal language during childhood, criteria to identify them as native speakers of British English. Their pure-tone air-conducted hearing thresholds at octave frequencies between 0.125 and 8 kHz included were 20 dB HL or less for both ears, interpreted as indicating normal hearing. The median age of the participants was 26 years, ranging from 18 up to 60 years. The effects of three speech enhancement algorithms were investigated incorporating 20 listeners per algorithm.

B. Materials

Stimuli were presented diotically over headphones (Sennheiser HDA-200) driven by a digital I/O system (DAC) (RME Fireface 400). The experimental setup was calibrated using an artificial ear (B&K 4153) equipped with a flat-plate adaptor and a 1/2 inch condenser microphone (B&K 4192), connected to a microphone power supply (B&K 2804); and a spectrum analyzer (OnoSokki cf-350 z). Levels observed in 1/3-octave bands with center frequencies ranging from 0.16 to 6.3 kHz indicated that the errors introduced by the equipment were less than 1.3 dB.

Signal processing and stimulus presentation were accomplished with software written in MATLAB release 2008a (Mathworks, 2008) using 64-bit floating point representations for all signals and in all signal manipulation. The DAC was controlled using the ASIO driver supplied by its manufacturer using a resolution of 24 bits and 44.1 kHz sampling rate.

The speech materials used in the experiments were the UCL recordings of the IEEE sentences (Rothauser *et al.*, 1969; Smith and Faulkner, 2006). Throughout the experiment, two types of noise were used: babble and car noise. Babble noise was taken from the NATO noises (TNO, 1990), a recording of 235 s from 100 people speaking in a canteen in which individual voices are slightly audible. The car noise recording was of a Ford Escort driven at approximately 110 km/h (70 mph) on a dry tarmac test track. No other cars were driven in the car's proximity, hence only the sound of the Escort is heard. During this 168 s recording, the omnidirectional microphone was located close to the dashboard surface in the car's cabin. Long-term average spectra of speech and noise signals are shown in Fig. 1.

C. Noise suppression

Three speech enhancement algorithms were evaluated for the intelligibility tests: spectral subtraction (SS), minimum mean square error spectral estimation (MMSE), and subspace enhancement (SSA). These algorithms form a representative set of the standard algorithms mentioned in Loizou (2007). The parameter settings used in each case are discussed in the following text, and a complete MATLAB implementation of the first two algorithms is available in VOICEBOX (Brookes, 2008). All the algorithms processed the speech signal in overlapping frames and used an overlap-add procedure on the processed frames to synthesize the enhanced speech output.

Each of the noise reduction algorithms requires an estimate of the noise power spectrum. For this purpose, we used an algorithm by Martin (2001) that eliminates the need for explicit speech activity detection by tracking the minimum power in each frequency band of the noisy signal's smoothed short-term power spectrum. The algorithm assumes that this minimum power represents the noise floor, and an estimate of the average noise power in each frequency band is then obtained by applying a bias compensation factor whose derivation is described in detail in Martin (2006). This algorithm implicitly assumes that the noise power spectrum is stationary over intervals of a few seconds.

The SS algorithm is described by Berouti *et al.* (1979) and operates in the spectral magnitude domain. The speech signal was divided into 32 ms Hamming windowed frames

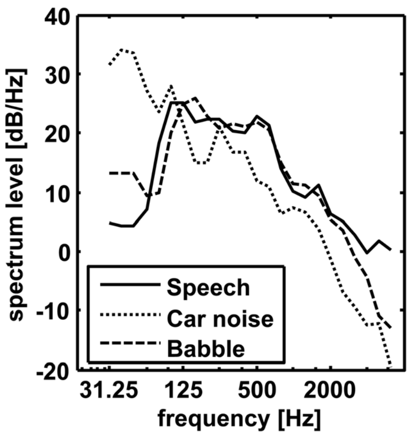

FIG. 1. One-third octave spectrums of speech and noises. All signals at broadband levels of 51 dB SPL. Continuous, dotted and dashed lines represent speech, car noise, and babble, respectively.

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

overlapping by 50%, and an estimate of the noise magnitude was subtracted in every frame and each frequency bin. To reduce the perception of musical noise, the algorithm incorporates a gain floor of –20 dB and an oversubtraction factor. Following the recommendations of Berouti *et al.* (1979) for reducing speech distortion, the oversubtraction factor had a maximum value of 3 and varied with the estimated frame SNR according to the equation:

$$\alpha_{os} = 2 + 0.04(|\mathrm{SNR} - 20| - |\mathrm{SNR} + 5|). \quad (1)$$

The minimum mean squared estimator of the log-spectrum described by Ephraim and Malah (1985) was implemented as a second algorithm. Assuming a Gaussian model for the complex spectral amplitudes of both speech and noise, this algorithm determines the optimum estimate of the log-spectrum of the clean speech signal. For estimating the *a priori* SNR within the algorithm, we used the decision-directed approach (Ephraim and Malah, 1984) with a time constant of 0.4 s ($\alpha = 0.96$), and we limited the *a posteriori* SNR to a maximum of 30 dB to limit the effect of high-level transients. The frame size and overlap factor were the same as for the SS algorithm.

The subspace approach to speech enhancement was introduced by Ephraim and Van Trees (1995) and generalized by Hu and Loizou (2003) to accommodate a colored noise spectrum. The approach takes advantage of correlations within the speech signal to distinguish it from the corrupting noise and seeks to minimize the speech distortion introduced subject to a maximum permitted noise level in the enhanced speech. The trade-off between distortion and noise is controlled by an algorithm parameter, μ, which is made adaptive so that less distortion is permitted when the SNR is high. The value of μ was restricted to the range 1-30, and all remaining algorithm parameters followed the recommendations given by Hu and Loizou (2003). The noise covariance matrix was calculated recursively from the estimated noise power spectrum using a smoothing time-constant of 3.1 ms ($\alpha = 0.98$).

D. Stimuli generation

The process for generating the stimuli can be divided into several steps. In the initial step, all speech and noise materials were downsampled from their original sampling rates and 16-bit representations to 16 kHz and 64-bit depth using the MATLAB resample function. In the second step, sentence levels were adjusted such that all had equal RMS values (Brady, 1968; ITU, 1994). These levels were measured using the activlev function available in VOICEBOX (Brookes, 2008). The first 200 sentences of the speech material were divided into sets of 20 sentences each. Within a set, sentences were assigned to one of the experimental conditions. In the following description, the sentence assigned to a particular experimental condition will be called the target sentence. To generate the stimulus for a particular trial, two sentences were randomly drawn from the same set as the target sentence with the restriction that these sentences should be different from the target sentence. The target sentence was embedded in these two additional sentences, inserting 150 ms of silence between the target and the embedding sentences, resulting in a sentence triplet. A noise fragment, the duration of which equaled that of the sentence triplet, was drawn randomly from the noise file and mixed with the triplet. The experimental conditions assigned to the target sentence controlled the noise type and SNR. In noise suppressed conditions, the noise-disturbed sentence triplet was passed through a noise suppressor; a step that was omitted in non-processed conditions. In the final step, the target sentence including the noise-filled silences was extracted from the sentence triplet, and 150 ms linear fades were applied to the onset and offset. Embedding the target sentence followed by extraction after noise suppression gave the noise suppressor the opportunity to stabilize on past and future information in the signal when processing the target sentence. The stimulus was upsampled to 44.1 kHz, passed through a 1024-point FIR filter that corrected the headphone frequency response and played back over the DAC and the headphones. Apart from the low-pass filtering introduced by down-sampling the signals, no additional band-pass filtering was applied.

The experimental setup was calibrated such that the RMS speech level in the absence of noise and noise suppression, so called original speech, was 51 dB SPL. This level was determined from Speech Intelligibility Index (SII) (ANSI, 1997) calculations for speech in quiet assuming a participant with a frequency independent hearing loss of 20 dB HL and a long-term average speech spectrum derived from the current materials. The SII proved highest for speech presented between 43 and 78 dB SPL approximately, indicating optimal intelligibility. It was judged that with speech at 51 dB SPL, the stimulus would be comfortably loud even when measuring at the lowest SNRs resulting in the highest presentation levels.

Across experimental conditions, the level of the speech was fixed, while the noise level varied. The noise levels were selected from a pilot study, in an attempt to obtain word correct scores in the range of 10% to 90% for non-processed speech. For babble, this resulted in SNRs from −12 to 0 dB in 3 dB steps. For car noise, SNRs ranging from −21 to −9 dB in 3 dB step were expected to result in similar word correct scores. To convert the dB SPL values used in this report into dB (A) values, one should add −3, −4, and −11 dB to the levels of speech, babble and car noise, respectively.

E. Experimental design

The three noise reduction algorithms were evaluated in separate experiments that will be addressed as Exp SS, Exp MMSE, and Exp SSA. Effects on intelligibility scores were assessed for two types of noise (NOISE = {car, babble}) using SNRs that targeted five performance levels (SNR = {−21, −18, −15, −12, −9} for car noise; SNR = {−12, −9, −6, −3, 0} for babble) with the noise suppressor switched on or off (SUPPRESSOR = {off, on}). Consequently, the kernel experimental design contained 20 experimental conditions. In Exp SS and Exp MMSE, the conditions were assigned to a set of 20 sentences according to a Latin square, such that across

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

participants each sentence was presented in all experimental conditions. This design allows one to assess differences in intelligibility scores between individual sentences without confounding by experimental conditions. In Exp SSA, the 20 sentences in a set were assigned randomly to the experimental conditions. These 20 sentences, forming an experimental block, were presented in random order.

To each listener, 10 experimental blocks were presented, using different sets of 20 sentences per block. Each set of sentences was taken from two consecutive lists within the IEEE corpus. Because each IEEE sentence contains five keywords, the intelligibility score in each experimental condition is based on SUBJECTS(20) × KEYWORDS(5) × BLOCKS(10) = 1000 responses.

F. Procedures

Data acquisition took place in a sound proof booth. After obtaining informed consent, the participant's pure-tone air-conducted hearing thresholds were measured (ISO, 2004). Subsequently, the participant received instructions for the intelligibility task: i.e. speaking aloud the sentences heard over the headphones. The participant was informed that some sentences would have poor intelligibility in which case the participant was encouraged to guess even if this would result in a nonsense or incomplete sentence. During the intelligibility task, the participant faced a computer screen and controlled stimuli presentation by clicking on buttons displayed on the screen. Each sentence was presented only once; verbal responses were audio recorded.

After responding to each experimental block, the participant determined the number of words correct in the verbal responses that the previous participant had given while listening to the same sentence set. The first participant scored responses obtained in a pilot study, the responses of the last participant were scored by a supplementary participant who did not perform the intelligibility task. The advantages of having subjects instead of the experimenter determine the correctness of the responses were fourfold: (1) Varying between intelligibility testing and scoring made the participants' task less monotonous. (2) The setup resulted in a test that was semi self-administered. (3) The fact that the experiment leader was a non-native speaker of English did not influence the test scores. (4) The effect of having a single judge score the sentences might possibly introduce bias, while the effect of multiple judges will only add random variation into the test scores—which can be handled in the statistical analysis.

While scoring, the recorded verbal response of the previous participant was played back over the headphones. The keywords of the corresponding sentence were displayed on the screen. The participant was instructed to tick the words mentioned in the response, disregarding differences in word order, verb tense, or noun quantities. In contrast to the intelligibility task, participants were allowed to change the presentation level and could replay a single response as many times as desired. On average, the intelligibility and scoring task took about 3 and 2 minutes per block, respectively. Typically, participants finished their 10 blocks of intelligibility testing and scoring within 70 minutes.

G. Statistical analysis

Following the approach taken by Ihlefeld *et al.* (2010), a logit transform was applied to the percentage word correct scores. But rather than a natural logarithm, we applied a base-two logarithm to give the transformed unit a simpler interpretation. The resulting outcome measure is labeled performance level, defined as

$$\mathrm{Perf} = \log_2(\mathrm{p}/(1-\mathrm{p})). \qquad (2)$$

Perf is a dimensionless quantity that we gave the unit "Berkson" (Bk). The unit is a tribute to Joseph Berkson (1899-1982), a physicist and statistician who popularized the usage and analysis of log odds. In Eq. (2), p denotes the proportion correct words. For each 1 Bk increase in performance level, the number of correct words doubles relative to a fixed number of incorrect words. Consequently, performance levels of −3; −2; −1; 0; 1; 2; and 3 Bk correspond to 11, 20, 33, 50, 67, 80, and 89% correct. If PMfs expressed on percentage scale have logit shapes, these functions become linear on a Berkson scale. Given this linear relation, a constant shift on a Berkson scale corresponds to a constant shift on a decibel scale. The inverse logit function

$$\mathrm{P} = 100 \cdot (2^{\mathrm{perf}}/(1+2^{\mathrm{perf}})) \qquad (3)$$

allows one to convert performance levels back into percentages.

Ihlefeld *et al.* (2010) analyzed log odds with analysis of variance for repeated measurements (RM-ANOVA). The current study uses multilevel logistic regression, as proposed by Goldstein (1995) and Hox (2010). The technique is also known as hierarchical generalized linear models for binary outcomes (Raudenbusch and Bryk, 2002) or generalized mixed models for binary data (Agresti, 2007). Various authors (e.g. Max and Onghena, 1999; Quéne and van der Bergh, 2004; Rellini *et al.*, 2005) have argued that multilevel analysis is superior to RM-ANOVA, since it does not assume sphericity, allows one to specify non-Gaussian error terms and gives higher statistical power.

III. RESULTS

To examine whether the scoring by different participants introduced inconsistencies in the outcomes, the 20,000 recorded responses obtained in Exp SS were scored independently by two additional judges. "Ground truth"' scores were generated by combining the scorings provided by the subjects and the two additional judges, using a majority-vote rule. Scorings of the judges differed 0.7% and 0.8% from this ground truth. The scorings of the subjects differed 0.7% from ground truth.

Figure 2 shows the percentage word correct scores observed in Exp SS, Exp MMSE, and Exp SSA as a function of SNR, i.e., PMfs. Experimental conditions with the suppressor switched off and on are represented by open and filled markers, respectively. Circles and squares indicate results for speech in car noise and babble, respectively. Vertical lines represent estimates of the standard deviation of

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

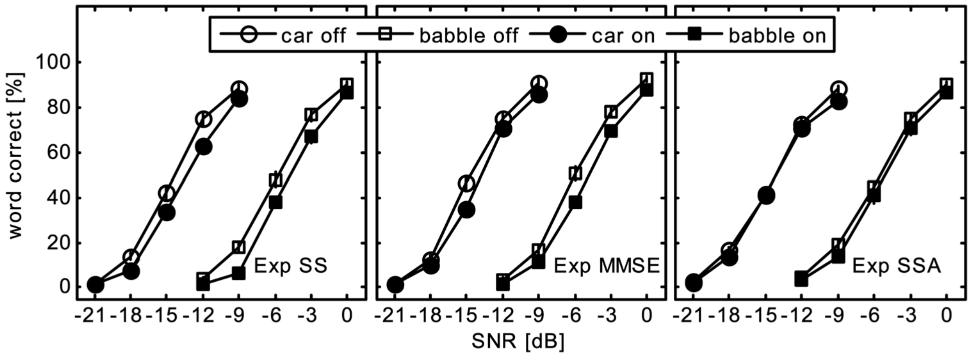

FIG. 2. Observed intelligibilities (percentages). Filled markers represent performance for conditions with the noise suppressor switched on. Circles and squares indicate results for car noise and babble, respectively. Observations from Exp SS, Exp MSSE, and Exp SSA are visualized in panels from left to right.

the sampling error; the estimates tend to be smaller than the markers. Almost all filled markers are located below the open maskers, meaning that algorithms reduced intelligibility scores. Effects appear to be different at the lowest SNRs of each noise type (SNR = −21 dB|car noise; SNR = −12|babble) for all noise suppressors, where markers for noisy speech without and with noise reduction coincide. In Exp SS and Exp MMSE, similar effects are visible at the highest SNRs (SNR = −9 dB|car noise; SNR = 0 dB|babble). Because these markers are close to the 0% and 100% limits of the percentage scale, these findings are usually treated as floor and ceiling effects, respectively.

A different view emerges from Fig. 3, where all data are plotted with the ordinate expressing intelligibility scores in performance levels. For convenience, corresponding percentages are displayed on the right hand-side axis. Whereas in Fig. 2 the PMfs exhibit sigmoid shapes, in Fig. 3, functions are almost linear and in fact would be straight lines if the PMfs had logit shapes. Expressed on a Berkson scale, the distance between the open markers and their corresponding filled markers seem to vary little across SNRs. To address the statistical significance of these differences, a multilevel loglinear regression analysis was performed for each combination of noise type and experiment separately. These analyses included the fixed factors SNR {−21, −18, −15, −12, −9|car noise} and {−12, −9, −6, −3, 0|babble}, SUPPRESSOR {off, on} and SUBJECTS as a random factor. Outcomes of the statistical analyses are displayed in Tables I and II for car noise and babble, respectively. Cell entries represent the coefficients estimated in these analyses expressed in Berksons. Stars indicate statistical significance: Effects with corresponding χ^2 values between 3.8 and 6.6 and 1 degree of freedom are denoted by an asterisk, representing p-values between 0.05 and 0.01. Coefficients with corresponding χ^2 values above 6.6 and 1 degree of freedom, hence p-values less than 0.01 are denoted by double asterisks. Effects are expressed relative to a reference condition, which was the experimental condition SNR = −15 dB for car noise; SNR = −6 dB for babble; SUPPRESSOR = off.

Table I shows the coefficients of multilevel logistic regressions on the intelligibilities obtained in car noise. In Exp SS, the intelligibility score in the reference condition is estimated at -0.4 Bk (41%), which differs significantly from 0 Bk (50%). For changes in the SNR to −9, −12, −18, and −21 dB, effects are 3.6, 2.2, −2.3, and −6.0 Bk, resulting in estimated intelligibility scores of 3.2 Bk (90%); 1.7 Bk (77%), −2.7 Bk (13%), and −6.5 Bk (1%), respectively. Applying SS in the reference condition altered intelligibility scores by −0.6 Bk (9%), leading to an estimated performance level of −1.0 Bk (33%). All these effects are statistically significant. This is not the case for the effects of SS at the other SNRs as displayed in rows 7−10. Besides the shift of −0.6 Bk (9%) due to SS at −15 dB SNR, additional effects at −9, −12, −18, and −21 dB SNR are estimated at 0.0, −0.3, −0.4, and 0.8 Bk. These coefficients do not differ significantly from zero. In other words, compared to the effect found at −15 dB SNR, there is no evidence that SS had different effects on intelligibility scores at the other SNRs. In Exp SS, performance after noise reduction is estimated at 2.6 Bk (86%), 0.8 Bk (64%), −3.7 Bk (7%), and −6.2 Bk (1%) at −9, −12, −18, and −21 dB SNR, respectively.

In Exp MMSE, outcomes are similar to Exp SS with one exception. The estimated performance in the reference condition of −0.2 Bk (47%) does not differ significantly from 0 Bk (50%). All other effects correspond to the ones observed in Exp SS. Results from Exp SSA differ in two aspects from the finding in Exp SS. First, the data show no significant

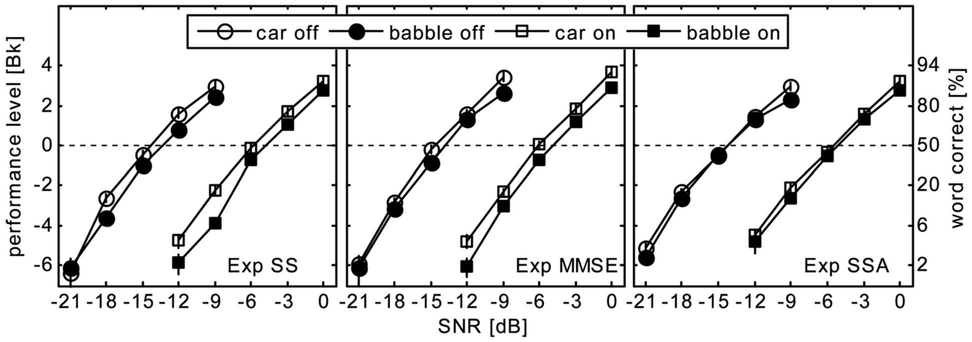

FIG. 3. Observed intelligibilities (logit transformed percentages). Filled markers represent performance for conditions with the noise suppressor switched on. Circles and squares indicate results for car noise and babble, respectively. Observations from Exp SS, Exp MMSE, and Exp SSA are visualized in panels from left to right.

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

TABLE I. Multi-level logistic regression coefficients for intelligibility scores in car noise. Effects on performance are expressed in Berksons relative to the reference condition, which is the performance at -15 dB SNR with the noise suppressor switched off. Coefficients that differ statistically significant from zero with corresponding *P* values between 0.01 and 0.05, or below 0.01 are indicated by * and **, respectively.

			Experiment		
Row No.	Effect		SS	MMSE	SSA
1	Reference	−15, off	−0.4*	−0.2	−0.5*
2	SNR	−9	3.6**	3.6**	3.6**
3		−12	2.2**	1.8**	2.0**
4		−18	−2.3**	−2.6**	−1.8**
5		−21	−6.0**	−5.7**	−4.7**
6	SUPPRESSOR	on	−0.6**	−0.7**	0.0
7	SNR	−9; on	0.0	−0.1	−0.7**
8	X	−12; on	−0.3	0.4	−0.4
9	SUPPRESSOR	−18; on	−0.4	0.3	−0.2
10		−21; on	0.8	0.5	−0.5

effect of SSA in the reference condition. Intelligibility scores at −15 dB SNR before and after noise reduction are equal at −0.5 Bk (41%). Second, at −9 dB SNR performance changed from 3.0 Bk (89%) to 2.3 Bk (84%), a significant shift of −0.7 Bk. Analogue to Exp SS, reductions of −0.4, −0.2, and −0.5 Bk at, respectively, −12, −18, and −21 dB SNR are non-significant. Estimated intelligibility scores after noise reduction are 1.3 Bk (71%), −2.7 Bk (13%), and −5.7 Bk (2%), respectively. The results obtained in car noise can be summarized as follows: (1) With SS and MMSE, the deteriorating effect of noise reduction was found to be independent from SNR. (2) SSA had no significant deteriorating effects on intelligibility scores except at the highest SNR.

Table II shows the regression coefficients for intelligibilities in babble. The reference condition is speech at −6 dB SNR without noise reduction. In Exp SS, the performance in the reference condition is estimated at −0.1 Bk (48%). Shifting the SNRs to 0, −3, −9, and −12 dB altered performance by 3.3, 1.8, −2.1, and −4.6 Bk, respectively. At these SNRs, performance is estimated at 3.2 Bk (90%), 1.7 Bk (77%), −2.0 Bk (18%), and −4.6 Bk (4%), respectively. Evidently, intelligibility scores vary across SNRs: All coefficients differ significantly from zero. Applying SS to speech in babble at -6 dB SNR affects performance by −0.6 Bk. The intelligibility score after noise reduction is estimated at −0.7 Bk (38%). Rows 7−10 display the regression coefficients for the effects of noise reduction at other SNRs. These coefficients represent effects of noise reduction additive to the one found for the reference condition. At 0, −3, and −12 dB, SNR coefficients do not differ significantly from zero. The effects of noise reduction at these SNRs are similar to the effect found in the reference condition. Performance after noise reduction at 0, −3, and −12 dB SNR is 2.7 Bk (87%), 1.0 Bk (67%), and −5.9 Bk (2%), respectively. At −9 dB SNR, the coefficient of 1.0 Bk is significant, indicating an additional reduction in the intelligibility score compared to the effect of SS at −6 dB SNR. The total shift in performance is −1.6 Bk (11%), resulting in an estimated intelligibility score of −3.8 Bk (6%) for noise reduced speech in babble at −9 dB SNR.

TABLE II. Multi-level logistic regression coefficients for intelligibility scores in babble. Details as for Table I.

			Experiment		
Row No.	Effect		SS	MMSE	SSA
1	Reference	−6, off	−0.1	0.0	−0.3
2	SNR	0	3.3**	3.6**	3.6**
3		−3	1.8**	1.8**	2.0**
4		−9	−2.1**	−2.4**	−1.8**
5		−12	−4.6**	−4.8**	−4.2**
6	SUPPRESSOR	on	−0.6**	−0.8**	−0.2
7	SNR	0; on	0.1	−0.1	−0.3
8	X	−3; on	−0.1	0.1	−0.1
9	SUPPRESSOR	−9; on	−1.0**	0.1	−0.3
10		−12; on	−0.5	−0.5	−0.1

In general, the outcomes of Exp MMSE and Exp SSA with babble are similar to Exp SS. However, MMSE did not exhibit the additional reduction in the intelligibility score at −9 dB SNR as found with SS in babble. None of the coefficients for MMSE displayed in rows 7−10 differs significantly from zero. In other words, with MMSE, there is no evidence that the effect of noise reduction differs across SNRs. Also Exp SSA showed no evidence of SNR dependent effects of noise reduction. None of coefficients that expressed effects of noise reduction was significant.

The results for noise reduction in babble can be summarized as follows: (1) SSA had no effects on the intelligibility scores at any SNR. (2) MMSE had a deteriorating effect on intelligibility scores that was independent of SNR. (3) With the exception of one SNR, SS reduced intelligibility scores equally across SNRs. At the exceptional SNR, the reduction in intelligibility scores was worse than observed at all other SNRs.

IV. DISCUSSION

In most speech intelligibility studies that use open response sets, the experimenters decide whether responses are correct (e.g., Brand and Kollmeier, 2002; Terband and Drullman, 2008). While attempting to automate data collection, various authors (e.g., Versfeld *et al.*, 2000; Hu and Loizou, 2007; Terband and Drullman, 2008) have requested their participants to type their responses. However, as Terband and Drullman (2008) point out, this may confound intelligibility with spelling and typing abilities. To overcome this weakness, these authors utilized automatic spelling checking and dynamic alignment but found that test-retest variability was higher than if an experimenter scored the verbal responses. In the current study, each participant scored the responses of the previous participant, leading to a semi self-administered intelligibility test. Outcomes suggest that these scores are as reliable as scoring by experimenters, at least for the normal hearing listeners in the age range employed here.

In our evaluation of the intelligibility of noisy speech after enhancement with a representative set of standard

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

algorithms, we found little evidence that the effects of noise reduction vary with SNR. Three algorithms were tested with two noise types, each at five SNRs. Effects of noise reduction at four SNR were compared to its effect at a reference condition corresponding to unprocessed speech at an SNR that gives an intelligibility score close to 0 Bk (50%). Noise reduction effects were significantly different from the reference condition in only 2 of 24 possible comparisons. These findings suggest that effects of speech enhancement on intelligibility are generally independent of SNR. Such a conclusion apparently differs from the observations made by Loizou (2007), who reported little effects when the intelligibilities scores prior to processing were high, but deleterious when the intelligibility scores prior to speech enhancement were around 0 Bk (50%). However, Loizou's findings can be seen to be congruent with ours as described in the following text.

Where in most studies statistical analysis was absent (Lim, 1978; Boll, 1979; Ludvigsen *et al.*, 1993; Tsoukalas *et al.*, 1997), Loizou (2007) performed RM-ANOVAs on percentages. Such analyses provides useful information about the effects in general or at a particular SNR, but they give limited information on differences in the effects of noise reduction across SNRs. Ideally a significant interaction of processing with SNR should indicate that such differences in effects are present. However, while performing RM-ANOVAs on intelligibility scores expressed as percentages, the significance of this interaction merely indicates that the processing effect is not a constant percentage across SNRs. Its significance should be interpreted as evidence against the hypothesis that processing shifts the PMf along the percentage axis. The fact that on the basis of this hypothesis one would predict intelligibility scores below 0% or above 100% illustrates that the rejection of this hypothesis is trivial. Significance of the processing by SNR interaction becomes inevitable with a broad range of SNRs or enough statistical power. It can be shown that similar limitations exist when the percentages are transformed in arc-sine units or rationalized arc-sine units prior to the RM-ANOVA.

S-shaped PMfs can be approximated by a number of functions such as the cumulative Weibull, normal, or the logit function (Klein, 2001). The last two are similar and have been used frequently in intelligibility studies (e.g. Versfeld *et al.*, 2000; Brand and Kollmeier, 2002). A logit shaped PMf was assumed here. If the effects of noise reduction are independent of SNR, we argued that the PMf should shift along the SNR axis. Table III illustrates such a shift based on our observations for MMSE with babble noise. The slope of the corresponding curve in Fig. 2 is about 0.7 Bk dB^{-1}. In Table III, Column 1 lists various SNRs, Columns 2 and 4 exhibit the predicted intelligibilities expressed on a Berkson scale with and without noise suppression, respectively, i.e. within these columns performance changes at rate of 0.7 Bk dB^{-1}. Column 7 displays the differences between Columns 2 and 4, differences that are constant (-0.8 Bk), reflecting the constant effect of MMSE in babble. The percentages in Columns 3 and 5 correspond to the performance levels expressed in Columns 2 and 4, respectively. Column 7 shows the difference between the Columns 5 and 6. This percentage shift differs with the intelligibility. Similar to the findings of Loizou (2007), the effect of noise suppression is small or large, when intelligibility is high or close to 0 Bk (50%), respectively. Findings in Table III additionally suggests that effects expressed on a percentage scale will diminish when the SNRs decreases below −6 dB.

TABLE III. Estimated effects of MMSE in babble. Intelligibilities scores with and without noise suppression and their differences at various SNRs expressed in performance levels and in percentages.

SNR (dB)	Suppressor on		Suppressor off		Suppressor effect	
	(Bk)	(%)	(Bk)	(%)	(Bk)	(%)
0.0	3.5	92	4.3	95	−0.8	3
−1.5	2.4	84	3.2	90	−0.8	6
−3.0	1.3	72	2.1	82	−0.8	10
−4.5	0.3	55	1.1	68	−0.8	13
−6.0	−0.8	36	0.0	50	−0.8	14
−7.5	−1.9	21	−1.1	32	−0.8	11
−9.0	−2.9	12	−2.1	18	−0.8	7
−10.5	−4.0	6	−3.2	10	−0.8	4
−12.0	−5.1	3	−4.3	5	−0.8	2

For speech in babble at −9 dB SNR processed with SS and for speech in car noise at −9 dB SNR processed with SSA, the effects did differ significantly from the reference condition. Of the 1000 keywords presented, 75 and 832 correct words were observed, respectively. Based on the effects found in the corresponding reference conditions, the expected numbers of correct words in these conditions are 124 and 886, respectively. We do consider these effects relevant but find it hard to present a proper explanation, more because a proper understanding of the detrimental effects of noise suppression on intelligibility is currently lacking (Loizou and Kim, 2011). Moreover, no systematic effects have been found. Hence the outcomes of the current study provide little evidence that the efficiency of the parameter estimation used during noise suppression varies with SNR.

It may seem somewhat surprising that the performance level in the reference condition for speech in car noise of Exp SS and Exp SSA differed significantly from 0 Bk, while in Exp MMSE, performance in the same reference condition did not differ significantly from 0 Bk. However, also in Exp MMSE, the observed performances were negative, and multilevel logistic regression that combined the results of the three experiments found no significant effects in the reference conditions across all experiments. Results of this analysis are not presented here because the effects particularly of interest to this study end up at high-order interactions, complicating interpretation.

In contrast to previous studies, where occasionally increasing intelligibilities due to noise suppression were observed (Tsoukalas *et al.*, 1997; Arehart *et al.*, 2003; Hu and Loizou, 2007), such improvements could not be observed in the current study. One could argue that the SNRs examined in the current study are too low, hence too difficult to be handled by the algorithms. Previous studies (Tsoukalas *et al.*, 1997; Arehart *et al.*, 2003) typically

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

examined intelligibility in the range of -5 dB up to 5 dB SNR, while here values down to -21 dB SNR are employed. However, comparing the SNRs across different studies is obstructed by the differences in bandwidth of the stimuli. In the studies by Lim (1978), Ludvigson *et al*. (1993), and Arehart *et al*. (2003), these were restricted to frequencies below 4.7, 4, and 4 kHz, respectively. Hu and Loizou (2007) limited their stimuli to 0.3$-$3.4 kHz. The car noise employed here has most of its energy below 0.125 kHz, as can be noticed in Fig. 1. These low frequencies do not contribute to intelligibility (ANSI, 1997), hence the car noise needs to be raised to substantial levels before it exceeds the energy of speech in the 0.125$-$8 kHz range containing the audio bands that contribute most to speech intelligibility. If one would only consider the noise and speech energy in the 0.3-3.4 kHz range, as done in the study by Hu and Loizou (2007), the levels of the car noise used in the current experiment would range from -15 to -2 dB SNR. Restricting speech and babble to 0.3$-$3.4 kHz has no effect on its SNR, because their spectrums are similar. Thus while most of the SNRs used in the current study are lower than the levels used before, it is likely that the highest SNRs used here do overlap with the SNRs used in previous studies (Tsoukalas *et al*., 1997; Arehart *et al*., 2003). This idea is further strengthened by the fact that intelligibilities for the non-processed speech observed here, fully covers the intelligibilities reported in all previous studies.

V. SUMMARY AND CONCLUSIONS

The effects of noise suppression on the intelligibility of speech in noise at various SNRs were studied. It showed that in 22 of 24 comparisons, the effect could be considered as a constant shift of the psychometric function along the SNR axis. Only little evidence was found that could indicate that the effects of noise suppression vary with SNR. For just two psychometric functions was the intelligibility reduction at one SNR significantly larger than observed near the midpoint. Because shifts are largely independent from SNR, we see few limitations in the use of speech reception thresholds to study the effects of noise suppression on intelligibility.

ACKNOWLEDGMENTS

We thank three anonymous reviewers for their valuable comments. This work was funded by the U.K. Home Office.

Agresti, A. (**2007**). *An Introduction to Categorical Data Analysis,* 2nd ed. (Wiley, Hoboken, NJ), Chaps. 4 and 10.

ANSI (**1997**). S3.5-1997. *American National Standards Institute Methods for Calculation of the Speech Intelligibility Index* (Acoustical Society of America, New York).

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

Berouti, M., Schwartz, R., and Makhoul, J. (**1979**). "Enhancement of speech corrupted by acoustic noise," Proc. IEEE Intl. Conf. Acoustics, Speech Signal Process. **4**, 208–211.

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

Brady, P. T. (**1968**) "Equivalent peak level—A threshold-independent speech-level measure," J. Acoust. Soc. Am. **44**, 695–699.

Brand, T., and Kollmeier, B. (**2002**). "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," J. Acoust. Soc. Am. **111**, 2801–2810.

Brookes, M. (**2008**). "Voicebox: Speech Processing Toolbox for MATLAB." http://www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html (Last viewed 11/23/11).

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

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

Ephraim, Y., and Van Trees, H. (**1995**). "A signal subspace approach for speech enhancement," IEEE Trans. Speech Audio Process. **3**, 251–266.

Goldstein, H. (**1995**). *Multilevel Statistical Models.* (Arnold, London, UK), Chap. 7.

Hox, J. (**2010**). *Multilevel Analysis: Techniques and Applications.* (Erlbaum Associates, Mahwah, NJ), Chap. 6.

Hu, Y., and Loizou, P. (**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. (**2007**). "A comparative intelligibility study of single-microphone noise reduction algorithms," J. Acoust. Soc. Am. **122**, 1777–1186.

Ihlefeld, A., Deeks, J. M., Axon, P. R., and Carlyon, R. P. (**2010**). "Simulations of cochlear-implant speech perception in modulated and unmodulated noise," J. Acoust. Soc. Am. **128**, 870–880.

ISO (**2004**). ISO 389-8:2004, *Reference Zero for the Calibration of Audiometric Equipment—Article 8: Reference Equivalent Threshold Sound Pressure Levels for Pure Tones and Circumaural Earphones*, (International Organization for Standardization, Geneva, CH).

ITU (**1994**). ITU-T P.56, *Objective Measurement of Active Speech Level* (International Telecommunication Union, Geneva, CH).

Jellyman, K. A. (**2009**). "An Assessment Of Speech Intelligibility In The Context Of Coders In High Noise," Ph.D. thesis, University of Wales, United Kingdom.

Klein, S. A. (**2001**). "Measuring, estimating, and understanding the psychometric function: a commentary," Percept. Psychophys. **63**, 1421–1455.

Leek, M. R. (**2001**). "Adaptive procedures in psychophysical research," Percept. Psychophys. **63**, 1279–1292.

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. C. (**2007**). *Speech Enhancement: Theory and Practice* (CRC Press, Boca Raton, FL), Chaps. 5-9.

Loizou, P. C., and Kim, G. (**2011**). "Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions," IEEE Trans. Audio, Speech, Lang. Process. **19**, 47–56.

Ludvigsen, C., Elberling, C., and Keidser, G. (**1993**). "Evaluation of noise reduction method: comparison between observed scores and scores predicted from STI," Scan. Audiol. **38**, 50–55.

Manchester, P. (**2010**). "Found sound: an introduction to forensic audio," Sound on Sound. **750**, 90–95.

Martin, R. (**2001**). "Noise power spectral density estimation based on optimal smoothing and minimum statistics," IEEE Trans. Speech Audio Process. **9**, 504–512 (**2001**).

Martin, R. (**2006**). "Bias compensation methods for minimum statistics noise power spectral density estimation," Signal Process. **86**, 1215–1229 (**2006**).

Max, L., and Onghena, P. (**1999**). "Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research," J. Speech Lang. Hear. Res. **42**, 261–270.

Quené, H., and van den Bergh, H. (**2004**). "On multi-level modeling of data from repeated measures designs: a tutorial," Speech Commun. **43**, 103–121.

Raudenbush, S. W., and Bryk, A. S. (**2002**). *Hierarchical Linear Models, 2nd ed.* (Sage Publications: Thousand Oaks, CA), Chap. 10.

Rellini, A. H., McCall, K. M., Randall, P. K., and Meston, C. M. (**2005**). "The relation between women's subjective and physiological sexual arousal," Psychophysiol. Res. **42**, 116–124.

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M. (**1969**). "IEEE

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp

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

Smith, M. W., and Faulkner, A. (**2006**). "Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing," J. Acoust. Soc. Am. **120**, 4019–4030.

Terband, H., and Drullman, R. (**2008**). "Study of an automated procedure for a Dutch sentence test for the measurement of the speech reception threshold in noise," J. Acoust. Soc. Am. **124**, 3225–3234.

TNO (**1990**). NATO Noises NATO: AC243/(Panel 3)/RSG-10 ESPRIT: Project No. 2589-SAM (Compact Disk).

Tsoukalas, D. E., Mourjopoulos, J. N., and Kokkinakis, G. (**1997**). "Speech enhancement based on audible noise suppression," IEEE Trans. Speech Audio Process. **5**, 497–514.

Versfeld, N. J., Daalder, L., Festen, J. M., and Houtgast, T. (**2000**). "Method for the selection of sentence materials for efficient measurement of the speech reception threshold," J. Acoust. Soc. Am. **107**, 1671–1684.

Downloaded 23 Jan 2012 to 155.198.125.228. Redistribution subject to ASA license or copyright; see http://asadl.org/journals/doc/ASALIB-home/info/terms.jsp