ADJUSTING A COMMERCIAL SPEECH ENHANCEMENT SYSTEM TO OPTIMIZE INTELLIGIBILITY.

GASTON HILKHUYSEN AND MARK HUCKVALE

Speech, Hearing and Phonetic Sciences, University College London, London, U.K.
g.hilkhuysen@ucl.ac.uk
m.huckvale@ucl.ac.uk

To improve the quality of noisy speech recordings, sound engineers have at their disposal a variety of signal processing techniques. These techniques often have a wide range of parameters which need to be adjusted to obtain optimal processing results. This paper investigates the difficulty of finding the best parameter settings for a commercial noise-reduction system. In a first experiment, operators adjusted the settings of a particular system while attempting to maximise the intelligibility of speech corrupted with babble noise at different signal-to-noise ratios. Their preferences were then evaluated in a listening experiment - showing that their chosen settings actually reduced intelligibility compared to the original signal. In another experiment a range of parameter settings for the same system were evaluated using both listeners and an intelligibility model based on a speech envelope distortion measure. Although the measure is imperfect, it is still able to predict optimal parameter settings better than the human operators.

INTRODUCTION

To improve the quality of noisy recordings, forensic audio operators have at their disposal a wide variety of speech enhancement methods. Among these are noise-reduction algorithms that aim to attenuate the level of any interfering noise whilst preserving the speech signal content. It is clear that such noise-reduction processing does improve the physical signal-to-noise ratio in the signal, and that listeners report that such processing leads to an increase in the perceived quality of recordings [1]. However, objective measurements of the actual speech intelligibility tend to show that noise-reduction decreases rather than increases performance [2]. One of the weaknesses of these studies that show the deleterious effects of noise reduction is that the algorithms are chosen and applied independently from an analysis of the nature and level of the signal distortion in any particular case. Given that numerous parameters are available on most enhancement tools, one could hypothesize that the observed deleterious effects on intelligibility could have resulted from sub-optimal parameter settings.

In practice, parameters of noise reduction systems are typically adjusted based on the opinion of the operator. One could question whether such settings are optimal: studies in the past have shown a mismatch between opinion-based estimates of the intelligibility of a speech signal compared to actual performance-based measures [3,4,5].

How best, then, to choose optimal settings for a speech enhancement method? Although with any real forensic recording operators do not have access to the clean input speech signal, one could create "equivalent" speech materials by distorting and corrupting known signals to a similar degree as real recordings. One could then compare the enhanced noisy signal against the clean original signal using intrusive measures of intelligibility. That way intrusive intelligibility models could provide an objective and cost-effective method to find optimal parameter settings, replacing time-consuming intelligibility experiments involving numerous listeners. Unfortunately, traditional intrusive intelligibility models such as the Speech Intelligibility Index (SII) [6] or the Speech Transmission Index (STI) [7] do not account properly for the effects of non-linear processing such as noise reduction, tending to predict improvements in intelligibility in contrast to the deteriorations actually observed [8].

In an attempt to overcome this limitation of the SII and STI, we have recently developed an extension of the SII that better accounts for the deleterious effects of noise reduction [9]. In this model, it is assumed that changes in intelligibility can be predicted from the accuracy with which speech envelopes are transmitted across various audio bands. Since additive noise and noise reduction techniques with non-optimal parameter settings distort these envelopes, the level of distortion can be used to predict intelligibility. We measure the quantity of envelope distortion after enhancement processing in terms of an equivalent amount of added noise that would lead to the same degree of distortion. This provides us with an equivalent signal-to-noise ratio (SNR) for each audio band. These equivalent SNRs are then used in a SII calculation, defining a quantity labelled SII_{mod} that we have found is monotonically related to intelligibility and a better predictor for intelligibility after noise-reduction than the SII or STI. One of the goals of the present study is to determine whether SII_{mod} might provide a means to determine optimal noise reduction settings.

In this paper, we investigate the extent to which judgements of audio engineers or the predictions of an intelligibility model lead to selection of the best parameter settings for a commercial noise reduction system. In Experiment 1, human expert listeners are asked to adjust settings of a noise reduction system in order to improve the intelligibility of some noisy speech signals. Their opinion-based settings are evaluated in Experiment 2 with an objective performance-based measure of intelligibility. In Experiment 3, a wide range of different parameter settings are investigated for their effect on intelligibility. Lastly, to investigate whether an intelligibility model can be used to identify optimal settings, SII_{mod} is used to predict the effects of the parameter settings. The effectiveness of these predictions are compared to the outcomes of Experiments 2 and 3.

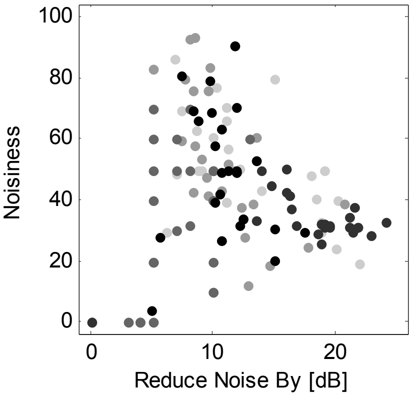

Figure 1: settings chosen by operators. Gray tints represents operators.

1 EXPERIMENT 1

1.1 Method

Five listeners, or "operators", adjusted the parameter settings for the VST(tm) 'Adaptive Noise Reduction' plug-in available in the Adobe Audition v3.0 Digital Audio Workstation (DAW). All operators were experienced listeners, familiar with different speech enhancement algorithms. Out of the seven processing controls available on the plug-in, only the ones labelled 'Reduce noise by' (Red) and 'Noisiness' (Noi) were manipulated. It was found that changing the settings of the other controls had little perceptible effect on the signal, and their settings were fixed at the levels suggested by the manufacturer in all experiments. During testing, operators listened to three concatenated IEEE sentences, which were corrupted by babble noise at five SNRs, ranging from -12 up to 0 dB SNR in 3 dB steps. The operators were asked to find the positions of the Red and Noi controls which gave "maximum intelligibility". Each operator adjusted the controls for a particular SNR five times, leading to OPERATOR(5) x SNR(5) x REPEAT(5) = 125 pairs of settings.

1.2 Results and discussion

Figure 1 shows a scatterplot of the settings chosen by the operators. High values for Red are never combined with high values of Noi, suggesting some trade-off in performance. However, low setting for Red are combined with settings across the whole range of Noi, and vice versa. Consequently, parameters settings show no significant correlation ($r = -0.16$, $p > 0.05$). Settings for each parameter were submitted to an Analysis of Variance for repeated measurements, including the factors OPERATOR(5) × SNR(5) × REPEAT(5). Both analyses showed significant effects for OPERATOR ($F(1,4) = 44.6$; $p < 0.01$| Red) ($F(1,4) =108.7$; $p < 0.05$ | Noi). None of the other factors or any of the interactions reached significance. We conclude that operators held different opinions on the settings that they thought would maximise intelligibility. There was no indication that operators changed their opinion during the course of the experiment, nor that they proposed different settings for different SNRs. All operators had considerable experience with noise reduction systems - most of them were engineers involved in the development or evaluation of speech enhancement algorithms. Nevertheless, given the same noise reduction system and equivalently noise-perturbed speech, they opted for different parameter settings, possibly leading to different intelligibilities. To evaluate their overall success at maximising intelligibility in the next experiment, we took the overall average parameter settings across all conditions and operators to be the most appropriate.

2 EXPERIMENT 2

2.1 Method

IEEE sentences [10,11], each containing five keywords, were mixed with babble noise at the SNRs used in Experiment 1. Half of these sentences were processed by the 'Adaptive Noise Reduction' plug-in available on the DAW using 13 dB and 48% as settings for the Red and Noi controls, respectively. The other half was not processed with noise reduction. Processed and non-processed sentences were randomly presented to ten normal hearing naïve listeners, who were asked to repeat each sentence verbatim. A particular sentence was only presented once to each listener. Intelligibility performance in each condition was defined by the two-based logarithm of the ratio between the number of correct and number of incorrect keywords in the listeners' responses and expressed as performance levels in Berkson (Bk) units.

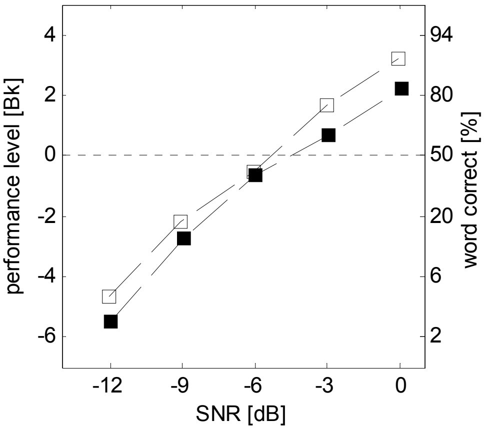

Figure 2: intelligibility before (□) and after (■) noise reduction.

2.2 Results and discussion

Figure 2 displays intelligibility of speech in babble noise with and without adaptive noise reduction. Open and closed markers indicate non-enhanced and enhanced speech, respectively. It can be observed that at each SNR, the open marker is located below the closed marker, suggesting universal deterioration in intelligibility due to noise reduction at the average operator settings. These drops in performance levels were assessed with a mixed effects logistic regression model. At -3 and 0 dB SNR, the deleterious effect of noise reduction reached statistical significance ($\chi2(1)$ = 10.3, $p < 0.01$; $\chi2(1)$ = 7.2, $p < 0.01$, respectively). At these SNRs performance dropped by about 1 Bk, meaning that for a fixed number of correct words, the number of incorrect words doubled due to noise reduction. Why did the experts choose settings that actually degraded intelligibility? One possibility is that the experts were unable to perform the task. Another possibility could be that the deteriorating effects of the noise suppressor are invariant across parameter settings, in which case each settings was as good as any other. This is perhaps reflected in the disagreements between operators over the best settings. Lastly, one might hypothesize that operators adjusted the parameters to values that were optimal in the sense that they were the least detrimental. These explanations were investigated in Experiment 3.

3 EXPERIMENT 3

3.1 Method

IEEE sentences were mixed with babble noise at -3 dB SNR. Using 16 pairs of parameter settings, equal numbers of sentences for each pair were processed by the 'Adaptive Noise Reduction' plug-in. Pairs of parameter settings were created by combining four levels of Red (0, 13, 26 and 39 dB) with four levels of Noi (0, 33, 66 and 99 %). Stimuli consisting of non-processed speech were added as a control. Intelligibility was measured using the same procedure as Experiment 2, with 10 normal hearing naïve listeners who had not been exposed to the sentences before.

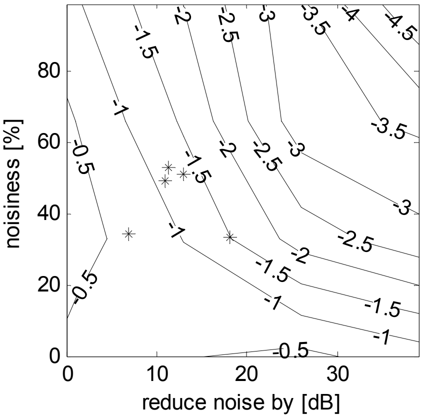

Figure 3: observed shifts in intelligibility in Berksons as a function of two noise reduction parameter settings.

3.2 Results and discussion

Figure 3 shows a contour plot based on the data obtained in Experiment 3. Curves represent combinations of Red and Noi that gave rise to equal intelligibility. Numbers on the curves specify the shifts in performance levels due to noise reduction expressed in Berksons. Negative values indicate a drop in intelligibility due to noise reduction. For settings of Red close to zero, changes in Noi have little consequence and vice versa, and at these settings noise reduction has little effect on intelligibility. When both parameters are set to higher values, intelligibility deteriorates rapidly and is poorest when Red and Noi are set at their maximum values. The five star symbols indicate the mean settings at -3 dB SNR proposed by each operator. Clearly these are not settings that lead to highest intelligibilities, but to drops in performance of about 1 Bk as previously found in Experiment 2.

4 MODELLING

4.1 Predicting Experiment 2

Figure 4 shows the relation between intelligibilities observed in Experiment 2 and predictions obtained from SII_{mod} for the five SNR levels. The model predicts intelligibility with values between 0 and 1, representing fully unintelligible and fully intelligible speech, respectively. Filled and open markers indicate intelligibility with and without noise reduction. Second order polynomials were fitted to each set of markers, giving rise to two 'performance' functions. For SII_{mod} to be valuable, the two performance functions should coincide, meaning that SII_{mod} relates to performance level, whether or not noise reduction is applied. In Figure 4 this holds for high values of SII_{mod}, but at low values the curve for noise reduced speech is higher. Calculating a SII_{mod} for noise reduced speech while predicting its intelligibility on the basis of the performance function for non-processed speech, will results in an underestimation of intelligibility of noise reduced speech at low SNRs.

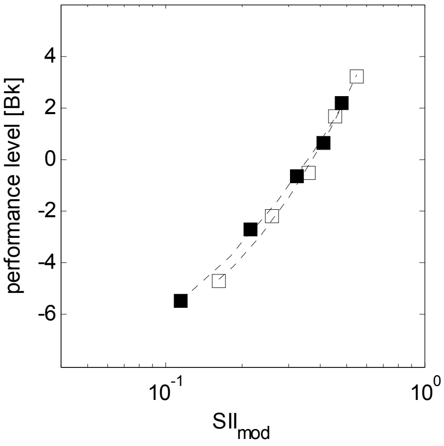

Figure 4: performance functions of SII_{mod} without (□) and with (■) noise reduction.

4.2 Predicting Experiment 3

SII_{mod} was calculated for each of the 16 pairs of parameter settings. Using the performance function for non-processed speech shown in Figure 4, the predicted effects of noise reduction on intelligibility were calculated. Figure 5 shows the contour plot for shifts in intelligibilities, analogous to Figure 3 but now based on estimates from SII_{mod}. A cross shows the parameter settings used in Experiment 2.

The density of the contours in Figure 4 is less than in Figure 3, indicating that SII_{mod} predicts smaller changes in intelligibility with parameter setting than were actually observed in Experiment 3. Closer inspection reveals that for increasing values of Noi, SII_{mod} predicts increases in its deteriorating effect on intelligibility only if Red exceeds 15 dB. The Red parameter only has an effect on SII_{mod} when restricted to values below 15 dB. Here increasing the value of Red is expected to deteriorate intelligibility. For setting of Red above 15 dB, only small changes in intelligibility are predicted, in contrast with the effects observed in Experiment 3.

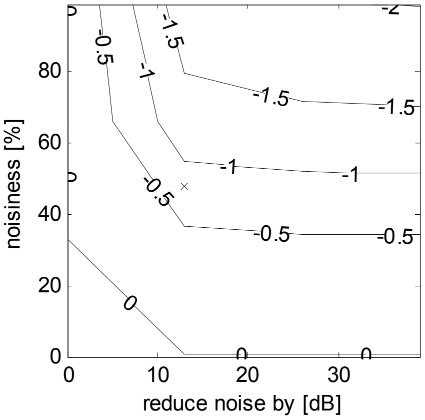

Figure 5: predicted shifts in intelligibility in Berkson as a function of two noise reduction parameter settings.

When both Red and Noi are set to zero, SII_{mod} is positive, indicating some restoration of the speech envelopes corrupted by the babble noise. Although the improvement is too small to be noticeable in the measured intelligibilities, we find it encouraging to notice that with some settings the noise reduction system is actually generating speech envelopes that are less distorted than before processing. Additionally we note that even when both parameters are set to zero, the output of the plugin differs from its input.

5 CONCLUSIONS

When asked to maximise the intelligibility of speech with a noise reduction tool, we have shown that operators set the parameters to values that in fact reduced intelligibility. Although these proposed settings differed significantly across operators, Figure 3 shows that four out of five operators suggested settings that introduced a drop of 1 Bk or more in performance. The fact that opinion-based intelligibility differs from performance-based measures has been noticed before [3,4,5], but to our knowledge never in the context of speech enhancement tools.

To overcome the limitations of settings based on opinions, we explored the use of SII_{mod} and found that although this model can partly account for the deleterious effects of noise reduction, in its current form it fails to predict the effects with some settings, and underestimates the size of the reductions in general. For the latter, this is partly due to the estimation of the performance functions by regression, which gives rise to systematic underestimation of high and low performance levels through 'regression toward the mean'. Ideas for improving $\mathrm{SII}_{\mathrm{mod}}$ have been formulated elsewhere [12]. Notwithstanding the current imperfections of $\mathrm{SII}_{\mathrm{mod}}$, the model outperformed the operators and suggested parameter settings that in a performance based intelligibility test had little deleterious effect. Of course to achieve this, $\mathrm{SII}_{\mathrm{mod}}$ needs access to both clean and noisy speech signals, while the operators only had access to the noisy speech. Although results show that the effects of noise reduction vary with the SNR, it remains to be seen whether different SNRs, speakers or noises require different parameters settings to optimize intelligibility, questions that are left for future studies. If such is the case, the value of $\mathrm{SII}_{\mathrm{mod}}$ in operational situations will depend on the quality of the simulated audio environment. If optimal settings vary little across recording conditions, $\mathrm{SII}_{\mathrm{mod}}$ could be a valuable tool for estimating settings that optimize intelligibility in the forensic audio area.

6 ACKNOWLEDGEMENTS

This work is supported by the Centre for Law-Enforcement Audio Research, funded by the U.K. Home Office

REFERENCES

[1] Hu, Y. and Loizou, P. C., "Subjective Comparison and Evaluation of Speech Enhancement Algorithms" *Speech Communication*, vol. 49, pp. 588-601, Jul, 2007.

[2] Hu, Y. and Loizou, P. C., "A comparative intelligibility study of single-microphone noise reduction algorithms" *J Acoust Soc Am*, vol. 122, pp. 1777-1786, Sep, 2007.

[3] Larsby, B. and Arlinger, S., "Speech recognition and just-follow-conversation tasks for normal-hearing and hearing-impaired listeners with different maskers" *Audiology*, vol. 33, pp. 165-176, Jun, 1994.

[4] Preminger, J. E. and Van Tasell, D. J., "Quantifying the relation between speech quality and speech intelligibility" *J Speech Hear Res*, vol. 38, pp. 714-725, Jun, 1995.

[5] Eisenberg, L. S., Dirks, D. D., Takayanagi, S., and Martinez, A. S., "Subjective judgements of clarity and intelligibility for filtered stimuli with equivalent speech intelligibility index predictions" *J Speech Lang Hear Res*, vol. 41, pp. 327-339, Apr, 1998.

[6] ANSI, "Methods for the Calculation of the Speech Intelligibility Index" American National Standards Institute, ANSI Standard S3.5-1997 (R2007), 1997.

[7] EIC, "Sound system equipment - Part 16: Objective rating of speech intelligibility by speech transmission index" EIC Standard 60268-16, 2003.

[8] Ludvigsen, C., Elberling, C., and Keidser G., "Evaluation of Noise Reduction Method: Comparison between Observed Scores and Scores Predicted from STI" Scan Audiol, vol. 38, pp. 50-55, 1993.

[9] Hilkhuysen, G. and Huckvale, M., "Understanding the intelligibility of speech after noise reduction: a comparison of predictive models", British Society of Audiology Short Papers Meeting on Experimental Studies of Hearing and Deafness, Southampton UK, 2009.

[10] Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., and Weinstock, M., "IEEE Recommended Practice for Speech Quality Measurements" *IEEE Transactions on Audio and Electroacoustics*, vol. AU17, pp. 225-246, 1969.

[11] Smith, M. W. and Faulkner, A., "Perceptual adaptation by normally hearing listeners to a simulated "hole" in hearing" *J Acoust Soc Am*, vol. 120, pp. 4019-4030, Dec, 2006.

[12] Hilkhuysen, G. and Huckvale, M., "Signal properties reducing intelligibility of speech after noise reduction", European Conference on Signal Processing (EUSIPCO), Denmark, 2010.