Quality assessment of noise reduction for digital hearing aids: Measurements and predictions

Mark Marzinzik* and Birger Kollmeier
AG Medizinische Physik, Carl von Ossietzky Universität Oldenburg, D-26111 Oldenburg
* Email: mark@medi.physik.uni-oldenburg.de

Abstract
A novel test protocol is proposed for the evaluation of noise reduction algorithms for use in digital hearing aids. It aims at assessing the "ease of listening" (or listening effort) and quality aspects (through paired comparisons). Correlations between different "objective" measures and the subjective quality data are reported. Besides some limitations, the results suggest to use the log-area ratio (LAR) for predicting overall quality of noise reduction algorithms and Hansen's q_c (here called PMF) for predicting the amount of noise suppression.

Speech intelligibility and "ease of listening"

Most single-channel noise reduction algorithms so far have not proven to provide significant (if any) improvements in speech intelligibility in noisy environments.
A potential benefit of such algorithms, however, is increased "ease of listening" which is probably connected with less listening effort. Indeed, fatigue and increased effort when listening in noise is a common complaint of hearing-impaired subjects. Even normal-hearing subjects have this complaint after a work-day's noise exposure. However, speech recognition tests did not reflect this fatigue (Ivarsson and Arlinger, 1993). The fatigue may be related to non-auditory functions, involving concentration, attention etc.
Gatehouse (1994) suggested a "sentence verification test" in which response times are measured to assess a dimension (probably "ease of listening") that is not covered by traditional speech tests. Hoeks and Levelt (1993) use pupillary dilation as a measure of the level of mental effort needed in an attentional task. However, to obtain stable (and reliable) results these methods need averaging over numerous trials since the trial-to-trial variability is large.
For this reason we suggested a different approach (details in Marzinzik and Kollmeier, 1999; Marzinzik *et al.*, 1999): Subjects listen to unknown radio news degraded by additive noise (2.5 min.). During the listening, the subject can immediately switch between four different hearing aid settings (four different noise reduction algorithms, in this case). After the presentation, the subject has to repeat the news. The repetition is recorded with a dictaphone and motivates the subject to concentrate on the listening. This recording, however, is not evaluated. The measure of interest is a rating of each of the four algorithms on a 5-point listening effort scale (ITU, 1996) which the subject is additionally asked for. This ease-of-listening test has been used for a study on single-microphone noise reduction (Marzinzik and Kollmeier, 1999). With the ease-of-listening test, clear improvements with some single-microphone noise reduction algorithms were found compared to no noise suppression (Figure 1). However, using the same conditions it was not possible to find improvements in speech reception thresholds.
Hence, the proposed ease-of-listening test seems to be an adequate tool to assess the effect of noise reduction techniques on the required effort to listen to a target speaker in background noise over a longer period of time. However, the test has not yet been formally evaluated.

Quality aspects

Besides speech intelligibility and listening effort, it is very important to assess the subjective processing quality of the algorithms since sound quality is, in general, a major feature for the acceptance of a hearing aid.

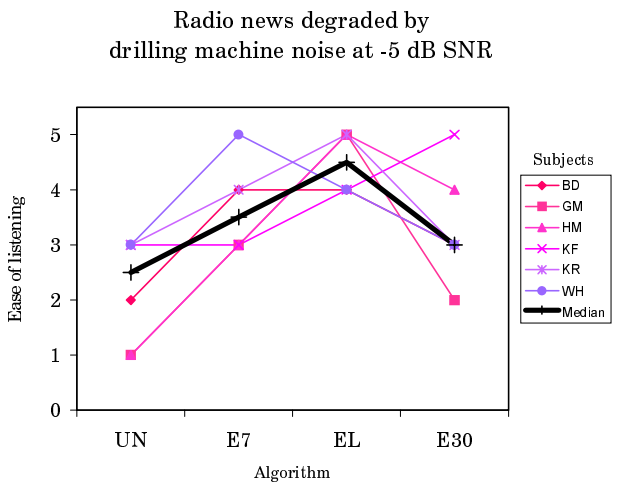

Figure 1: Results of the ease-of-listening test for radio news in drilling machine noise at -5 dB SNR. The noise reduction algorithms E7, EL, and E30 are minimum mean square error spectral amplitude estimators (Ephraim and Malah, 1984, eqs. 7 and 30; Ephraim and Malah, 1985). These algorithms were combined with a speech pause detection algorithm (for updating the noise estimate; Marzinzik, 2000). UN denotes no noise suppression. The 6 hearing-impaired subjects were fitted according to the traditional one-half gain rule. Presentation level was in the most comfortable range. The scale used for the ordinate codes as follows (ITU, 1996): 5= Complete relaxation possible; no effort required. 4= Attention necessary; no appreciable effort required. 3= Moderate effort required. 2= Considerable effort required. 1= No meaning understood with any feasible effort.

Hence, for evaluating the noise reduction algorithms we performed a complete paired comparison experiment of all algorithms (including no suppression) with regard to three different criteria: preference *with regard to the reduction of the background noise*, preference *with regard to the naturalness of the speech,* and *overall* preference. The analysis of the data was performed using the Bradley-Terry model which provides a scaling of the paired comparison data based on a difference scale. Hence, distances between algorithms are represented in a meaningful way (Bradley and Terry, 1952). Results of these measurements with the same algorithms as in Fig. 1 are given in Figure 2 (abscissas).

Predicting perceived sound quality

The following objective speech quality measures were applied to the same test signals that were used in the subjective tests: The objective speech quality measure q_c, here called PMF, introduced by M. Hansen (1998) and different quality measures proposed by J. Hansen and B. Pellom (1998): Itakura-Saito Distortion Measure (IS), Log-Likelihood Ratio Measure (LLR), Log-Area Ratio Measure (LAR), Segmental Signal-to-Noise Ratio Measure (SSNR), and Weighted Spectral Slope Measure (WSS). Correlations between objective and subjective data are reported in Table 1. A look at Figure 2, upper panel, attests the (very) high correlation between PMF and the subjective data with criterion "noise suppression".

Table 1: Pearson's correlation coefficients between objective measures and subjective data

	Subjective Criterion		
Objective measure	Noise Suppression	Speech Naturalness	Overall Preference
PMF	0.90**	0.43	0.67**
SSNR	0.71**	0.68**	0.77**
LAR	0.48	0.92**	0.87**
LLR	0.56*	0.88**	0.86**
IS	-0.50*	-0.06	-0.21
WSS	0.28	0.78**	0.68**

* *significant at a level of 0.05 (two-sided).*
** *significant at a level of 0.01 (two-sided).*

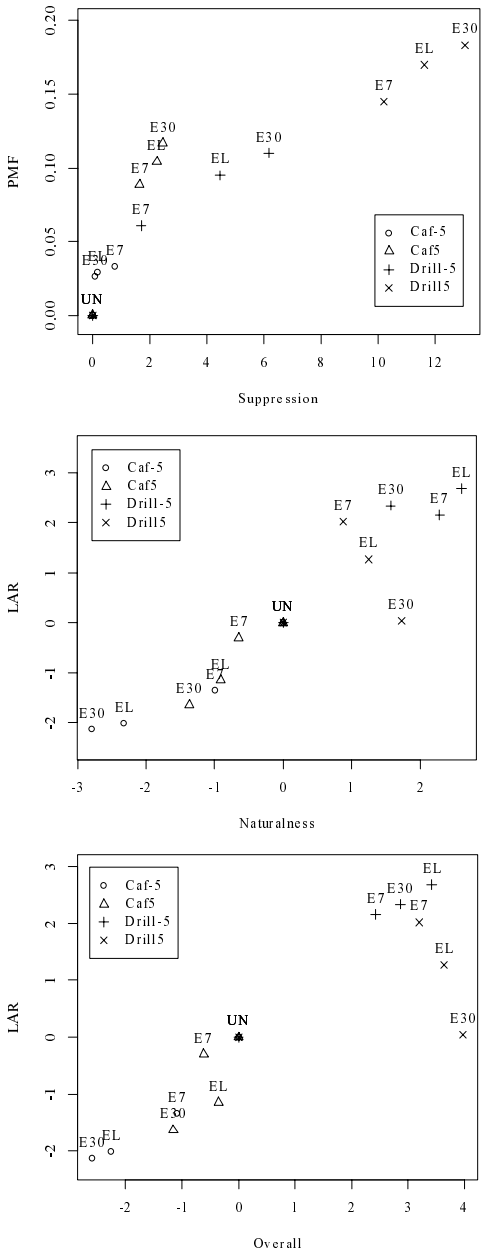

Figure 2: Scatter plots of objective measures vs. the subjective data (Bradley-Terry scale values; details of the study in Figure 1). The higher values on the abscissa indicate a higher preference value. Upper panel: *Objective measure PMF vs. subjective data for the criterion "noise suppression".* Middle panel: *LAR vs. subjective "naturalness of the speech".* Lower panel: *LAR vs. "overall preference". Noise conditions are cafeteria noise at –5 dB SNR (Caf-5) and +5 dB SNR (Caf5), drilling machine noise at -5 dB SNR (Drill-5) and +5 dB SNR (Drill5).*

Not only that the PMF measure is able to give the correct ranking of the algorithms for all different conditions, even the amount of noise suppression in the different noise conditions is very well predicted. For the criterion "speech naturalness", the correlation between the LAR measure and the subjective data is very high. While the rankings and distances are very well predicted for cafeteria noise (note that the noise suppression algorithms are judged worse than unprocessed), the LAR measure fails to give correct rankings of the algorithms for the drilling noise conditions (see Figure 2, middle panel). However, the LAR measure correctly predicts that the speech naturalness of the noise suppressed signals with drilling noise is generally judged better than unprocessed. It seems that the reversed ranking of algorithms E7, EL, and E30 for drilling noise at 5 dB SNR is due to a change of criterion by the subjects. In this noise condition they judged the speech the more natural, the more noise was suppressed. Our assumption is that above a certain signal-to-noise ratio the subjects rate the simple presence of noise as disturbing the "speech naturalness" and small distortions of the speech itself due to noise suppression processing are not weighed as much as the influence of the noise. This change in criterion, however, is not represented in the objective measures.

The LAR measure is also highest correlated with the "overall preference" judgements (Figure 2, lower panel). We see an excellent relation between LAR predictions and subjective data for cafeteria noise and drilling noise at SNRs of –5 dB. In the better SNR condition (+5 dB), the same observation can be made as discussed before: In the drilling noise at 5 dB, the subjects seem to have changed their criterion which results in a reversed rank order. Probably, this process of switching criteria is just occuring in the cafeteria noise condition at +5 dB. The judgements are better than at –5 dB (note that the noise reduction processing is still worse than unprocessed) but EL (with more noise suppression) has already overtaken E7.

With some cautions, the combination of PMF and LAR measures is recommended for objective evaluations and optimisation of parameter settings of noise reduction algorithms. However, more work has to be done with respect to subjective evaluations in order to explain some puzzling effects and discrepancies between the subjective data and the predictions.

This work was supported by the European Commission, Project SPACE (DE 3012, Telematics Applications Programme).

References

Bradley, R.A. and Terry, M.E. (1952). *Biometrika*, **39**, 324–345.

Ephraim, Y. and Malah, D. (1984). *IEEE Trans. ASSP* **32**, 1109-1121.

Ephraim, Y. and Malah, D. (1985). *IEEE Trans. ASSP* **33**, 443-445.

Gatehouse, S. (1994). *Ear and Hearing* **15**, 30-49.

Hansen, M. (1998). *Assessment and prediction of speech transmission quality with an auditory processing model*. Dissertation. BIS, Oldenburg.

Hansen, J.H.L. and Pellom, B.L. (1998). *ICSLP '98*, Sydney.

Hoeks, B. and Levelt, W.J.M. (1993). *Behavior Research Methods, Instruments, & Computers* **25** (1), 16-26.

ITU International Telecommunication Union (1996). *Recommendation P.800*.

Ivarsson, U. S. and Arlinger, S. D. (1993). *Scand. Audiol.* **23**, 159-163.

Marzinzik, M. and Kollmeier, B. (1999). In: Dau, T., Hohmann, V. and Kollmeier, B. (Eds.): *Psychophysics, physiology and models of hearing*. World Scientific Publishing, Singapore, 279–282.

Marzinzik, M., Wittkop, T. and Kollmeier, B. (1999). *J. Acoust. Soc. Am.* **105** (No. 2, Pt. 2), 1211.

Marzinzik, M. (2000). *Development and evaluation of single-microphone noise reduction algorithms for digital hearing aids*. PhD thesis, in preparation.