Individual preference for the amount of noise reduction

Rolph Houben [1], *Tjeerd M.H. Dijkstra*[2,3], *Wouter A. Dreschler*[1]

[1] Clinical and Experimental Audiology, Academic Medical Center Amsterdam, The Netherlands
[2] Institute for Computing and Information Sciences, Radboud University Nijmegen, The Netherlands, [3] Dept of Electrical Engineering, Technical University Eindhoven, The Netherlands

a.c.houben@amc.uva.nl

Abstract

We measured individual preferences for pairs of audio streams differing in the amount of noise reduction (G_{min}). To describe the paired comparison data a logistic probability model was used based on a quadratic preference utility function. This model allowed for the calculation of the G_{min} that was optimal for each individual subject. For five out of ten subjects the optimized G_{min} (ranging from 5.5 to 10dB) differed significantly from the optimum obtained for the grouped data (7.1dB). The predicted preference of the individually optimized G_{min} over the group optimum was slightly higher than chance level (up to 60%).

Index Terms: Noise Reduction, Hearing Aids, Individualisation, Hearing-Aid Fitting, Quality Preference.

1. Introduction

Many noise reduction algorithms in consumer products are preset for the end user. The question is whether these presets, geared to the average end user, can be improved by individualization, especially in the case of hearing aids. Whereas there are several fitting rules for (compressive) amplification (for instance NAL-NL1 [1]) that can be used for individualizing the gain of the hearing aids to the hearing loss of the user, there are no rules for noise reduction in hearing aids. Audiologists do not have tools to fit noise reduction algorithms to their clients' hearing loss. In daily practice, the noise reduction algorithm parameters are often pre-set by the manufacturer and usually these can not be changed by the clinician. Some aids allow a choice between for instance low, medium and high amounts of noise reduction. However the difference in signal processing between these setting is not clear. The selection and setting of noise reduction is further complicated by the fact there are many noise reduction implementations and strategies [2] that usually are not revealed in the technical specifications of the hearing aids.

Previous research on optimizing noise reduction parameters is scarce since most research is focused on the comparison between distinct algorithms [3] [4] and not on optimization of a single algorithm. An exception is the work of Zakis et al. [5]. They investigated the effect of the maximum amount of gain reduction (G_{min}). After two weeks of wearing the majority of subjects (90%) preferred G_{min} to be dependent on the input sound level. No statistically significant difference was found between a fit in which G_{min} changed across frequencies and a fit in which G_{min} was the same for all frequencies. However, based on these results the authors suggested that hearing-aid users should be given the opportunity to choose between alternative configurations. In the current study we focus on the question whether G_{min} leads to different preferences in individual listeners. G_{min} strongly influences the trade-off between the amount of residual noise and unwanted distortions: a higher value of G_{min} corresponds to less residual noise, but also goes along with a higher degree of distortion. The hypothesis is that there is a perceptual optimum in G_{min} somewhere between no processing (G_{min}=0 dB, no distortion but original amount of noise) and much processing (G_{min}>> 0 dB, much distortion but little remaining noise). If this optimum differs between listeners, it might be relevant to individualize the degree of noise reduction in hearing-aids. We conducted a laboratory experiment in which the preference of naïve listeners was measured for a hearing-aid noise reduction scheme that was implemented in Matlab. The preference for different values of G_{min} for normally hearing subjects was investigated with the method of paired comparisons.

One of the intended future goals of the project is to use a similar procedure in a clinical setting (i.e. with elderly and hearing-impaired subjects). Paired comparisons are suitable in that case [6] because this method poses minimal constraints on the judge [7] [8] and allows for detection of subtle differences [7]. Therefore our assumption is that this method is better suited than for instance rating scales [9] [10]. A second intended goal is to develop a statistical model that can be used to predict a subject's answer from a reduced data set. By using such a model one could calculate during the fitting process which paired comparison will be the most informative, thus minimizing clinical measurement time.

2. Materials, methods, and design

Ten normally hearing subjects participated in this study. Their average age was 32±6 yrs, and their hearing loss was smaller than 20 dB HL for each audiometric frequency from 250 Hz to 8 kHz. Stimuli were presented with Sennheiser HDA200 headphones that were connected to the output of a sound card (Dell Latitude D630 onboard sound card) via a headphone buffer (Tucker-Davis Technologies HB-6). Speech consisted of concatenated Dutch sentences of a male speaker [11] and the time between two consecutive sentences was 500 ms. Speech was presented in a stationary speech-shaped background noise. The signal to noise level (SNR) was chosen at +5 dB. This value was chosen because Pearsons et al. [12] reported that relevant SNRs at conversation distances in daily situations for normal hearing listeners were about +5 to +14 dB [13]. For the present experiment we choose the lowest value of this range since lower SNRs are generally more challenging both for the listener and for the noise reduction algorithm. All stimuli were presented diotically at 70 dB(A). This level is well within the range for which speech intelligibility in noise is independent of the noise level (for normal hearing listeners up to about 90 dB [14]). Additionally, 70 dB(A) agrees well with reported average outdoor

10.21437/PQS.2010-17

conversation levels (66 ± 4 dB(A) [13]) and was preferred over lower levels in informal listening tests.

A noise reduction algorithm (PNR, Perceptual Noise Reduction) was implemented in Matlab [15] and was provided by GN Resound Corporation. PNR is a low-latency seventeen-channel noise reduction algorithm that is similar to those applied in the current generation of hearing aids. It uses spectral subtraction with short-term estimation of speech spectrum and long-term estimation of ambient spectrum. For the experiments we chose envelope tracking time constants of 5 and 500ms; speech probability smoothing constants of 10 and 50 ms, and noise tracking time constants of 1 and 10 s. Gain reduction was set to be independent of frequency channel and input level. The signal to noise ratio estimator was based on the identification of modulations and the applied gain depended on the estimated signal to noise ratio (below 0 dB SNR the gain reduction was equal to G_{min}, above 0 dB SNR the gain reduction increased logarithmically to zero at infinite SNR). All signal processing was run in time-frames of 50 ms [16] which allowed for (almost) real-time switching between the two stimuli of the paired comparison.

The variable under investigation was the maximum amount of noise subtraction (G_{min}). Each test consisted of a complete design (round-robin tournament), in which the stimuli were not compared to themselves. Used values for G_{min} were 0, 4, 6, 8, 10, 12, 16 dB, where 0 dB equals no noise reduction (but the signals still passed through the algorithm and received the same analysis and synthesis). A single run consisted of 21 comparisons and in total 6 runs (5 retests) were done. The design was balanced in that the odd-numbered runs contained one comparison (e.g. G_{min}=6 versus G_{min}=4) and its balanced counterpart (G_{min}=4 versus G_{min}=6) was present in the even-numbered runs. So, each subject chose the best stimulus from 21 distinct pairs, and did that 6 times. Please note that the runs were not exact retests as the speech material was different for each comparison due to the running speech.

Listener preference was measured with a paired comparison procedure (a two alternative forced choice paradigm) in which concatenated speech sentences were presented in a stationary background noise. The subject's task was to make a choice based on the question: "Imagine that you will have to listen to these signals all day. Which sound would you prefer for prolonged listening?" The question was intentionally stated in a broad context because we were interested in general preference and because we wanted to measure possible inter-individual differences in preference. A broadly formulated task is more suitable for this than specific questions such as "speech quality" or "amount of background noise" that have a large aspect of interpretation. In the broad task both questions have to be combined according to the subject's preference. Subjects were allowed to listen to the two stimuli as long as they liked and could switch between the stimuli as often as they preferred. After they decided which stimulus they preferred they could make their choice by pressing a button on a computer screen and the tests automatically moved to the next comparison. All subjects used a mouse to control the experiment.

3. Results

As a first test whether the subjects differed in their responses we applied Cochran's Q test that is often used for checking inter-subject differences. It is a nonparametric test examining change in a dichotomous variable across more than two observations and is an extension to the McNemar test for related samples [17]. All the given answers (A preferred over B, B preferred over A) for all 21 comparisons and all repetitions of a subject were compared to that of other subjects. No significant difference in subject responses were found (p=0.3, df=9, number of cases for each subject=126, the number of B responses ranged from 55 to 75). Cochran's Q test has less power than for instance parametric tests and it could be that the nonsignificant result indicates that a larger sample size was required to detect differences between subjects. Moreover, Cochran's Q compares the complete response for each subject and not specifically the most relevant (the G_{min} with highest preference). Individual differences may show up if trivial conditions (were almost all subjects agree) are excluded. And, indeed, Cochran's Q test with G_{min} = 0 dB and 16 dB excluded lead to a significant result (p=0.04).

3.1. Noise reduction versus unprocessed

A convenient way to represent the paired-comparison data is to simply count the number of times each value of G_{min} was chosen over other values of G_{min}. This results in a data reduction to one score for each of the levels of G_{min}. The win counts are plotted in Figure 1 and data for each subject is given in a separate panel. The win counts are expressed as the proportion of wins, i.e. the number of wins divided by the total number of times that G_{min} occurred in the experiment). Data was averaged over the repetitions and each point corresponds to one G_{min} value. Additionally, the data averaged over all subjects is shown in the bottom most right panel. The error bars give the 95% confidence intervals (assuming a normal distribution).
As hypothesized, the win count data for each subject plotted against G_{min} showed a "mountain shape" with an optimum for G_{min} between G_{min}=4 and G_{min}=12 dB.

3.2. Two statistical models to extract preference from the paired-comparison data

3.2.1. Model 1: Win counts

The win count data was modelled with logistic regression. This is valid because win counts are simply a summation of binomial data and thus can be approximated by a logistic linear model. Logistic regression has the advantage over other methods that it is automatically restricted to [0,1]. The data for each subject was separately fit and we applied stepwise forward selection of polynomial predictors. If the model was improved (deviance improved according to a chi-square test with p< 0.05) the polynomial term was accepted. This resulted in a second order polynomial for nine out of 10 subjects. For subject NH1 and for the pooled data the deviance of the third order significantly improved the model. However, since the effect of the third order term was small, only polynomials of the second order will be used. The resulting logistic fits are included in Figure 1. For each fit the model significance was calculated by comparing (with a chi-square test) the deviance to that of a saturated model that fits the data perfectly. The calculated p-values are shown in the top left corner of each panel in Figure 1. Models with p>0.05 were considered informative. The p-values of the models for NH1, NH2, NH3, and the pooled data set were smaller than 0.05, and can thus be considered equally informative as a saturated model.

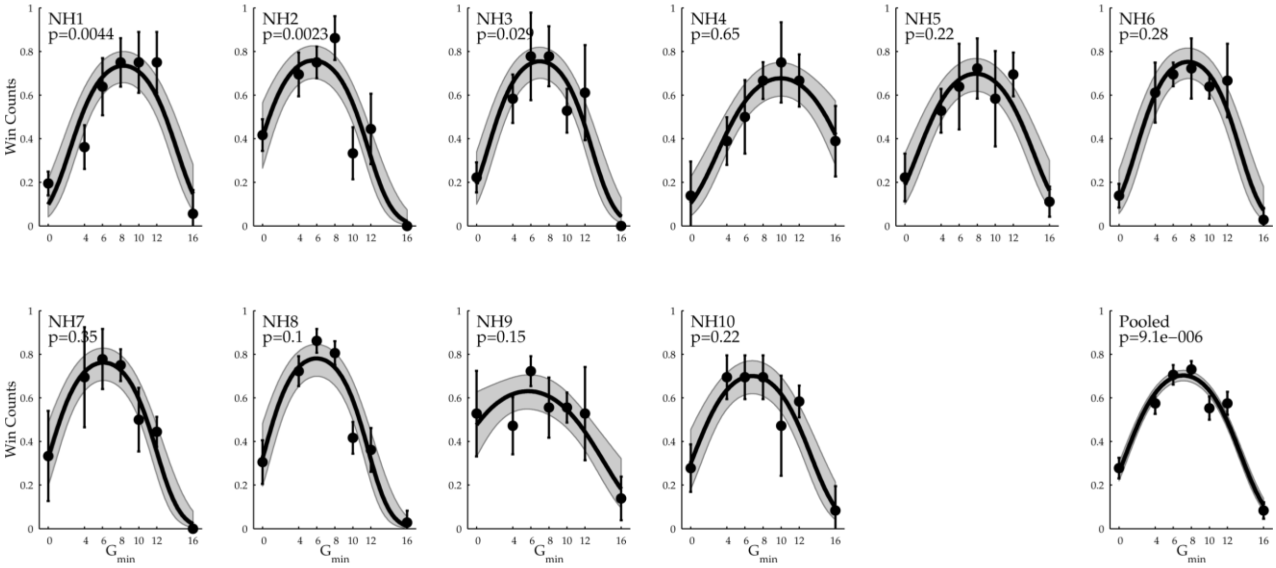

Figure 1: Win counts as function for Gmin for each individual subject. The black dots show experimental data averaged over the six repetitions. Error bars denote 95% confidence. The lines show a logistic regression fit. The shaded area denotes the 95% confidence interval of the model. The bottom right panel shows data averaged over all subjects. The top left corner of each plot shows the p-value of the model (see text).

While these fits look satisfactory, we note a shortcoming of using the win counts as basis: win counts depend on the levels chosen for the independent variable (in this case G_{min}). For example, had we chosen a more fine-grained sampling of G_{min} between 6 to 10 dB (for example, in steps of 1 dB instead of 2 dB as we did now), then the peaks in win counts would have been more shallow, making it more difficult to identify the optimal G_{min}. In contrast, the alternative model we will use below does not suffer from this deficit.

3.2.2. Model 2: Preference Utility

Instead of using aggregated win counts, we can model the data on the level of the actual paired comparisons. To do this we calculated the probability that the listener prefers one stimulus over another. By using a simple mathematical expression for the effect of G_{min}, we take into account the fact that the G_{min} levels are ordered. The principle of parsimony in modelling (i.e. use as few parameters as possible [18]) suggests that this leads to improved fitting. The new model has two basic assumptions. The first is that the subjects' dichotomous answer (0 or 1, representing the subject's choice) can be used as estimates of the underlying choice probabilities. The responses can then be described by the commonly used logit link function (log(p/(1-p)), in which p is the probability). To model the effect of G_{min} on a subject's preference we need a mathematical expression (the utility) to describe the preference in terms of G_{min}. The utility can be regarded as a measure of relative preference. This gives the second model assumption: the perceptual trade-off under investigation leads to an optimum (see Figure 1) and this can be modelled with a simple quadratic function. The quadratic utility function can be written for each subject as

$$U_i = c\,(G_{\min i} - G_{\min}^{opt})^2 \qquad (1)$$

in which $U_i = f(G_{\min i})$ is the utility and is a function of G_{min}, *i* is the stimulus (either the first *i* =1 or the second *i* =2), c is a constant (independent of G_{min}), and $G_{\min}^{opt}$ is the optimal value of G_{min} (i.e. the G_{min} at the maximum of the utility). The preference $p = f(G_{min1}, G_{min2})$ for one of the two alternatives of a paired-comparison is given by

$$\log\left(\frac{p}{1-p}\right) = U_2 - U_1 \qquad (2)$$

Insertion of (1) into (2) leads (after some simplification) to

$$\begin{aligned}\log\left(\frac{p}{1-p}\right) &= c\,(G_{\min2} - G_{\min1})\,(G_{\min2} + G_{\min1}) \\ &\quad - 2c\,G_{\min}^{opt}\,(G_{\min2} - G_{\min1})\end{aligned} \qquad (3)$$

The first term on the right hand side can be interpreted as the average G_{min} times the difference in G_{min}, and the second term as the difference in G_{min} times the optimal value of G_{min}. This model was fit to the data and Figure 2 shows the results. To show the logistic nature of the fit, the horizontal axis was chosen to be equal to the right hand side of equation (3), and the vertical axis shows the preference (p). The Figure also shows the underlying dichotomous data points and a smoothed line of the average values (dotted line). The top left corner of each plot shows the p-value of the fit. The p-values of the models for NH2, NH4, and the pooled data set were smaller than 0.05, and can thus be considered equally informative as a saturated model.

3.3. Comparison of models

The dichotomous nature of paired comparisons makes the comparison of statistical models non-trivial. Some intuitions from linear regression do not carry over directly to logistic regression: a model with no residual error constitutes as perfect fit in linear regression. The equivalent metric in logistic regression is the positive classification rate, defined as the number of answers correctly described by the model for each subject. While a positive classification rate of 100% is theoretically possible, in practice this means that the experimenter chose stimulus conditions that were too easy and hence not very informative. With statistical bootstrap simulations using the parameters fitted from the pooled model we showed that a perfect model for this data has a positive

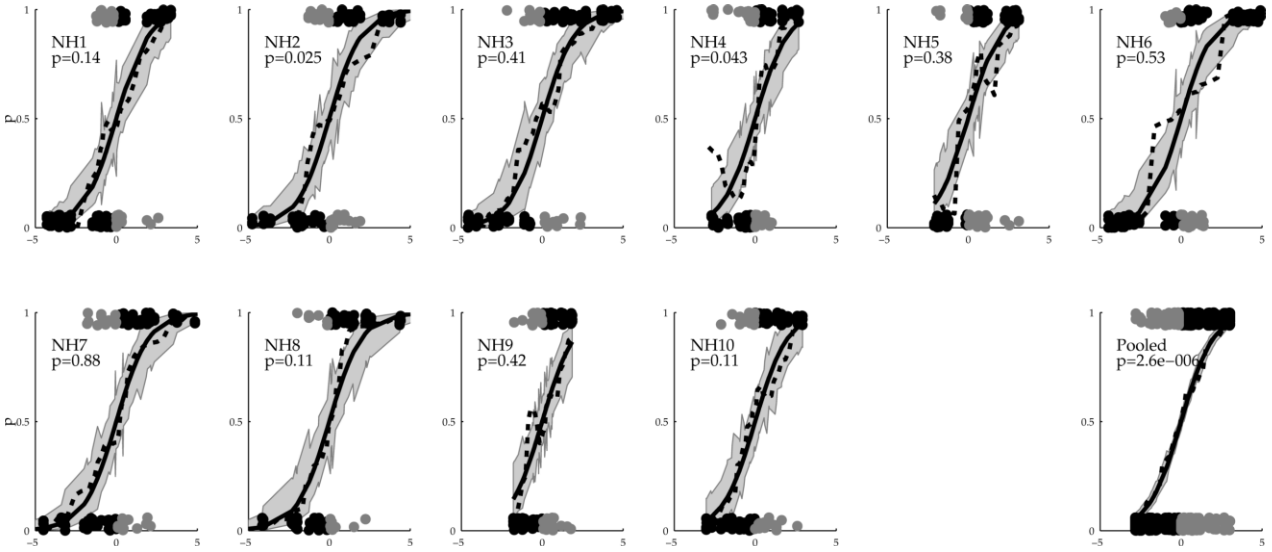

Figure 2: Paired comparison data and model fit for the model with the preference utility. The horizontal axis is described by the right hand side of equation (3). The raw dichotomous data is shown as black dots if the model had correctly classified it, otherwise in gray. The black lines show the logistic regression fits on the original 126 paired-comparison data points and the shaded area denotes the 95% confidence interval of each model. The dotted lines are a smoothed representation of the dichotomous data. In the top left corner of each plot the subject code is given as well as the p-value of the model fit (see text). The bottom right panel shows the model for the pooled data set (all subjects, 1260 paired comparisons). Note that the vertical axis represents the probability (and not the odds).

classification rate of 83%. This represents the maximal value given the observed parameters and stimulus levels used.

Table 1 gives the computed positive classification rates for the two models. The above described comparison between the two models was based on the *descriptive* accuracy of the complete data set. Another way to compare the models is to calculate the *predictive* accuracy of the models by using cross-validation. Cross-validation is a technique in which part of the data is used to fit the model to the data (training) and from which another part of the data is then predicted (testing on unseen data).

Table 1. Model comparison. Positive classification rate in % correct (higher values are better and 83% represents a perfect model) and Brier score for 3-fold cross-validation with 5000 repeats (lower values are better and a perfect model has a Brier score of 0.11)

	Positive classification rate in % correct		Brier score	
Subject	Model 1 Win Counts	Model 2 Preference Utility	Model 1 Win Counts	Model 2 Preference Utility
NH1	70	73	0.20	0.13
NH2	70	75	0.20	0.12
NH3	75	78	0.20	0.14
NH4	68	76	0.22	0.20
NH5	75	77	0.21	0.18
NH6	75	78	0.19	0.15
NH7	74	81	0.20	0.15
NH8	70	84	0.19	0.11
NH9	70	64	0.23	0.23
NH10	74	73	0.21	0.19
average NH1 through 10	72	76	0.21	0.16
pooled model	75	75	0.21	0.17

The training and testing is done repeatedly, each time with another random division of the data in training and testing set. The commonly used measure for the accuracy of a model prediction for dichotomous data is the Brier score [19]. For each answer, the prediction error is calculated by taking the square of the difference between the prediction (continuous) and the answer (dichotomous). The quadratic differences are then averaged to obtain the Brier score. Brier scores range between 0 (prediction and outcome are equal) to 1 (discordant prediction). A bootstrap calculation using the parameters fitted from the pooled data showed that the best model has a Brier score of about 0.11. A non-informative model has a Brier score of 0.25. Table 1 gives the Brier scores obtained by the models with three-fold cross-validation with 5000 repeats. Comparisons were divided at random into training (82 points for each subject) and testing set (42 points). The table also shows the Brier score for the pooled data (820 training points and 420 testing points). For all except one subject, the preference utility model had better results on both the positive classification rate and the Brier score. The exception was NH9 for which the win count model had better scores. For the pooled data set, the win count model had the same positive classification rate as the preference utility model, but the Brier score was higher.

Based on the positive classification rate and on the cross validation we conclude that the preference utility model is at least as good in describing and predicting the paired comparison data. In the following we will show results obtained with this model.

3.4. Do the individually optimized G_{min} differ from that of the group?

The statistical model can make predictions for G_{min} values that were not used in the experiment. This interpolation allows obtaining the G_{min} that corresponds to the highest preference for each individual subject. To compare the individualized values of G_{min} we need to calculate a confidence interval. The experimentally obtained data contains variance in the measured response (i.e. the preference) but the values of G_{min} were fixed in the experiment. The bootstrap method was used to obtain representative confidence intervals for the optimized G_{min}. Since we want to test for each of the ten subjects if the individually optimized G_{min} deviates from that of the group, a confidence interval is required that takes into account these multiple comparisons. For this we can use the 99.5% confidence interval. This interval corresponds to the 95% confidence interval with the inclusion of a Bonferroni correction for 10 comparisons. Figure 4 shows the best G_{min} with 99.5% confidence intervals. Five subjects (NH4,1,7,8,2) had a best G_{min} that deviated significantly ($p<0.005$) from that of the group.

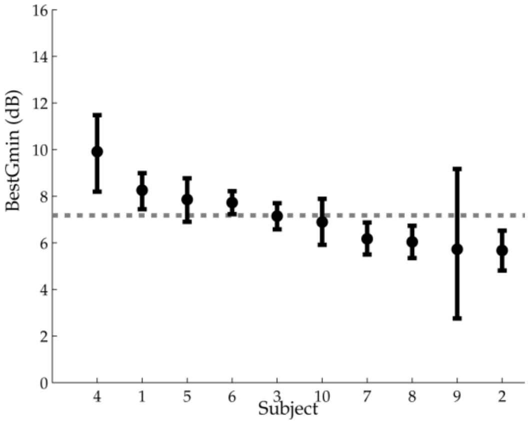

Figure 4: Individually optimized G_{min} for each subject with 99.5% confidence intervals obtained by bootstrapping (5000 times) of the preference utility model. The dotted horizontal line shows the optimal G_{min} for the group data. Subject ordering was based on the individual best G_{min}..

3.5. The predicted preference of the individually optimized G_{min} over other G_{min}

The fact that the individualized optimal value of G_{min} can differ significantly from that for the group data does not mean that this difference is relevant. To gain insight in the relevancy we can use the model to predict the preference for one setting over the other. Figure 5 shows for each subject the predicted probability of the subjects' choice for two hypothetical comparisons: 1) individually optimized G_{min} versus the group optimum, and 2) individually optimized G_{min} versus the unprocessed condition.

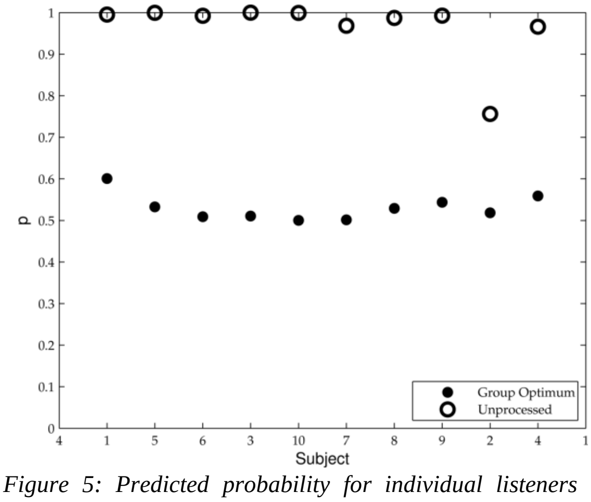

Figure 5: Predicted probability for individual listeners to choose the individually determined optimal value of G_{min} over a) the group optimum (G_{min}=7.2, full line) and b) over unprocessed (G_{min}=0, dotted line).

4. Discussion

The experimental results confirmed that most subjects preferred noise reduction over unprocessed [20] [21]. The model prediction from Figure 5 shows that the preference of the optimized G_{min} over unprocessed is quite high ($0.76<p<0.999$). Cochran's Q test gave no significant difference between the responses of the subjects. However as the other analyses showed this was most likely due to the low statistical power of the non-parametric test, and not due to small differences between subjects. For instance subject NH4 deviates from the rest in that his/her preference for G_{min}=16 is higher than that of the unprocessed condition (see Figure 1).

The win counts gave good insight in the effect of G_{min} on the response of each subject (see Figure 1). The hypothesized preference optimum for G_{min} was found at about 8 dB for the aggregated data set. For all subjects the response had a single global optimum (i.e. a G_{min} that was most preferred). Of course this optimal value of G_{min} might be different for different noises, SNRs, and noise reduction schemes, but it is likely that a quadratic model is suitable for such data.

Both statistical models could describe the experimental data. Still, the pooled models did not differ significantly from a saturated model. This is likely caused by the finding that many listeners preferred G_{min} = 8 and 12 over 10 dB. Both models assume smooth behaviour in G_{min} and are not able to track such preference changes. For instance, if for the win count model the subjects with low p-values for their individual fits (NH1, NH2, and NH3) were removed from the pooled data set, the p-value of the pooled set jumped up from $9*10^{-6}$ to $2*10^{-2}$, while removal of three random subjects changed the p-value only slightly: removal of NH8, NH9, NH10 yielded $p=9*10^{-5}$. The discontinuity at 10 dB is also the reason for the low p-values for some subjects as removal of this data point makes the models significantly different from the saturated model. The cause of the non smooth preference remains to be investigated. It could be caused by random effects or by the experimental design (e.g. interaction of the noise reduction algorithm with the speech materials) or perhaps some subjects differ from the rest in that they have additionally local maxima in their preference curve.

The preference utility model showed better results than the win count model for all except one subject, on both the positive classification rate and the Brier score. The notable exception was NH9 for which the win count model had better scores. The win count model was based on average results and might thus have been less sensitive to the relatively high variance in the data of this subject. For the pooled data set, the win count model had the same positive classification rate as the preference utility model. This indicates that for large amount of data or data that is averaged over many subjects the win count model can sufficiently describe the data. However, for individual predictions, or for smaller data sets the preference utility model showed better results. One of the reasons for this is that the preference utility model takes into account the response bias for one of the choice buttons. Wickelmeier and Choisel (2006) and Arehart et al. (2007) found a bias for the sound stimulus that was presented second. Our results showed a small (on average 56% versus 44%) bias for the first stimulus for the conditions that differed in G_{min} by 2 dB (no bias was found for the easier comparisons). The reason for preference for the first button is most likely that subjects were allowed to repeat the presentation as often as the liked.

Analyses based on win counts suffer from the fact that the scaling is influenced by the choice of stimuli that were included in the experimental design. For instance, inclusion of obvious values of G_{min} (>20 dB) would have led to flattening around the optimal value (e.g. G_{min}=8 for the aggregated data). The preference utility automatically accommodates for this with a smaller value for the fitted coefficient of the quadratic term. Therefore, the logistic prediction of each comparison makes the model more robust against inclusion of non-informative comparisons. (e.g. G_{min}>20 dB).

An additional advantage of the preference utility model is that it describes/predicts individual answers (and not the averaged win counts) with only a few fitting parameters (2 for our experiment). Such an approach might be useful to predict optimal values of $G_{min.}$ Alternatively, one could use the model to calculate during the fitting process which next paired-comparison would be the most informative.

The preference utility model showed that for 5 subjects the individually optimized G_{min} differed significantly from that of the group data. However, this is not yet enough evidence to propose routinely an individualization procedure for noise reduction because the predicted probability with which these five subjects would choose the individualized optimum over the group optimum was low. This might be different for other subjects or for subjects with hearing loss, for other noise types, for other signal-to-noise ratios, and for other types of noise reduction.

5. Conclusions

- Noise reduction as applied in hearing aids was preferred over no noise reduction by nearly all subjects.
- For this set of ten normal hearing subjects, five subjects differed significantly from the group data in their optimal value of G_{min}.
- The hypothesized preference optimum for G_{min} was found at about 8 dB for the aggregated data set. The values of G_{min} corresponding to the highest preference for individual subjects ranged from 5.5 to 10 dB.
- Although subjects differed in optimal G_{min}, the probability that a NH subject would choose their own individualized optimal G_{min} over the group best G_{min} was only small (p=0.6).
- Modelling individual responses has advantages over looking at aggregated win counts
  - automatic correction for button bias;
  - better prediction of individual answers;
  - more robust to the inclusion of trivial comparisons;
  - possibility to determine the probability that an (interpolated) setting will be chosen over another setting.

6. Acknowledgements

Approval by the Medical Ethical Committee of AMC was obtained on 24 April 2008. We gratefully acknowledge financial support from STW (Dutch Foundation for Science and Technology) for the HearClip project. We would like to thank GN Resound for providing code for the PNR noise reduction algorithm.

7. References

[1] D. Byrne, H. Dillon, T. Ching, R. Katsch, and G. Keidser, “NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures,” Journal of the American Academy of Audiology, vol. 12, J2001, pp. 37-51.

[2] A.E. Hoetink, L. Körössy, and W.A. Dreschler, “Classification of steady state gain reduction produced by amplitude modulation based noise reduction in digital hearing aids,” International Journal of Audiology, vol. 48, 2009, pp. 444–455.

[3] R. Bentler and L.K. Chiou, “Digital noise reduction: An overview,” Trends in Amplification, vol. 10, 2006, p. 67.

[4] Y. Hu and P. Loizou, “Subjective comparison of speech enhancement algorithms,” Proc. IEEE Int. Conf. Acoust., Speech, Signal Process, 2006, pp. 153–156.

[5] J. Zakis, J. Hau, and P. Blamey, “Environmental noise reduction configuration: Effects on preferences, satisfaction, and speech understanding,” International Journal of Audiology, vol. 48, 2009, pp. 853-867.

[6] S.C. Purdy and C.V. Pavlovich, “Reliability, Sensitivity and Validity of Magnitude Estimation, Category Scaling and Paired-Comparison Judgments of Speech Intelligibility by Older Listeners,” International Journal of Audiology, vol. 31, 1992, pp. 254–271.

[7] U. Böckenholt, “Hierarchical modeling of paired comparison data,” Psychological Methods, vol. 6, 2001, pp. 49–66.

[8] U. Böckenholt, “Visualizing individual differences in pairwise comparison data,” Food Quality and Preference, vol. 17, 2006, pp. 179-187.

[9] G.A. Studebaker, J.D. Bisset, D.M. Van Ort, and S. Hoffnung, “Paired comparison judgments of relative intelligibility in noise,” The Journal of the Acoustical Society of America, vol. 72, 1982, p. 80-92.

[10] L. Eisenberg, D. Dirks, and J. Gornbein, “Subjective judgments of speech clarity measured by paired comparisons and category rating,” Ear Hear., vol. 18, 1997, pp. 294-306.

[11] N.J. Versfeld, L. Daalder, J.M. Festen, and T. Houtgast, “Method for the selection of sentence materials for efficient measurement of the speech reception threshold,” The Journal of the Acoustical Society of America, vol. 107, 2000, pp. 1671-1684.

[12] K.S. Pearsons, R.L. Bennett, and S. Fidell, Speech levels in various noise environments, Washington DC: U.S. Environmental Protection Agency, 1977.

[13] W.O. Olsen, “Average Speech Levels and Spectra in Various Speaking/Listening Conditions: A Summary of the Pearson, Bennett, & Fidell (1977) Report,” American Journal of Audiology, vol. 7, 1998, pp. 21-25.

[14] K.D. Kryter, The handbook of hearing and the effects of noise, Academic Press, 1994.

[15] The Mathworks, Inc., Matlab, R2007b, Natick, MA, USA, 2007.

[16] Humphrey, R., 2008. Playrec Multi-channel Matlab Audio, Available from http://www.playrec.co.uk.

[17] K.D. Patil, “Cochran's Q Test: Exact Distribution,” Journal of the American Statistical Association, vol. 70, 1975, pp. 186-189.

[18] D.W. Hosmer and S. Lemeshow, Applied logistic regression, Wiley-Interscience, 2000.

[19] E.W. Steyerberg, Clinical Prediction Models, Springer, 2008.

[20] R. Bentler, Y. Wu, J. Kettel, and R. Hurtig, “Digital noise reduction: Outcomes from laboratory and field studies,” International Journal of Audiology, vol. 47, 2008, pp. 447-460.

[21] L. Luts, “Multicenter evaluation of signal enhancement algorithms for hearing aids,” The Journal of the Acoustical Society of America, 2010, pp. 1491-1505.