A MAP based approach to adaptive speech intelligibility measurements

Trym Holter and Svein Sørsdal

Department of Acoustics / Acoustic Research Centre, SINTEF ICT, Norway

trym.holter@sintef.no, svein.sorsdal@sintef.no

Abstract

This paper presents an adaptive procedure applied to measurements of speech intelligibility using the modified rhyme test. It is argued that the required speech-to-noise ratio (SNR) could be estimated with sufficient accuracy with as few as 25 utterances. The present procedure is based on the maximum a posteriori criterion, and it is demonstrated how the standard deviation of the SNR estimate can be improved by about 0.5 dB compared to a previously published method based on the maximum likelihood procedure.

Index Terms: speech intelligibility, listening test, modified rhyme test, maximum a posteriori.

1. Introduction

Despite the efforts put into developing robust objective measures of speech intelligibility and speech quality, subjective tests remain the gold standard for quantifying such aspects of a speech based system. Subjective testing is however time consuming compared to objective methods, and it is therefore of interest to improve the efficiency of such methods.

Speech intelligibility testing has traditionally been carried out using a constant speech-to-noise ratio (SNR) for each batch of test items, and such procedures could in this context be termed static. Adaptive procedures on the other hand offer a quicker way to estimate the SNR that corresponds to a predefined intelligibility score [1][2]. The great advantage of this is that it allows speech intelligibility to be quantified as the SNR required to achieve a given score, e.g., in a modified rhyme test (MRT) [3]. The relative performance of two different systems can then be quantified as a difference in SNR.

The most common adaptive approaches belong to the class of k-down/one-up procedures [2], where k typically equals 1, 2, or 3. The main drawback of these approaches is that they can only be used to estimate the SNR at a limited set of target intelligibility scores (P=0.5, 0.707, 0.794 for k=1, 2, 3, respectively) [4]. Also, because the up-down procedure requires a sufficient number of up-down reversals for a valid estimation, the number of responses in a trial will have to remain relatively high [5]. However, this procedure is still popular because of its simplicity. One more advantage is that it makes no assumptions about the form of the psychometric function, other than that it is monotonically increasing. This is in contrast to the methods discussed next.

An approach based on the maximum-likelihood (ML) principle has been developed for use in adaptive psychoacoustic measurements [5]. This procedure was adapted to speech intelligibility testing in [4], using the MRT. The main advantage of this scheme is that it is efficient in terms of the required number of test items. The method also in principle allows any target intelligibility score to be selected.

In this paper, a method similar to the one in [4] is presented. The new procedure is based on the maximum a posteriori (MAP) principle. The rationale behind this method is to further improve the efficiency of the ML-based method, by reducing the standard deviation of the SNR estimate for a given number of test items, while at the same time keeping the bias low. Both the ML and MAP based methods are described in detail in section 2 of this paper. Section 3 describes a range of numerical simulations done to investigate the performance of the proposed scheme. Section 4 discusses the experimental results, before the conclusions are drawn in section 5.

2. MAP-based adaptive MRT

In the MRT the test subjects are presented with rhyming words that differ only in their leading or ending consonantal phonemes. The MRT is a forced-choice task, and a list of 6 words is presented to the test subject to choose from. As discussed already, the MRT was originally meant for evaluating speech intelligibility at a fixed SNR. In the next two sections, two alternative ways to apply the MRT in an adaptive framework are discussed.

2.1. ML-based adaptive MRT

The procedure proposed in [4] relies on an ML estimation of the parameters \( \lambda \) of the psychometric function. The psychometric function in relation to an MRT quantifies the probability of a correct answer from the test subject for a given stimulus level (or correspondingly, for a given SNR). Given N responses from the test subject, \( O_N = \{o_1, \ldots, o_N\} \), the ML estimate of the psychometric function can be given as:

\[
\hat{\lambda}_N = \arg \max_{\lambda} P(O_N | \lambda)
\]

\[
= \arg \max_{\lambda} \left( \prod_{n=1}^{N} P(o_n | \lambda) \right)
\]

where it has been assumed that the observations are statistically independent. The observation \( o_n \) consists of a response value \( v_n \) (correct or incorrect) and the corresponding stimulus level \( l_n \). The conditional observation probability is given by

\[
P(o_n | \lambda) = \begin{cases} P_i(l_n) & \text{for } v_n = \text{correct} \\ 1 - P_i(l_n) & \text{for } v_n = \text{incorrect} \end{cases}
\]

where \( P_i(l) \) is the psychometric function.

The ML-based adaptive MRT (hereafter called ML-MRT) procedure works by determining the parameters of the model according to the ML criterion after every observation. The stimulus level used for the next presentation to the user is then chosen such that it corresponds to a fixed target score according to the latest estimate of the psychometric function. The assumption made is that the shape of the psychometric function associated with the task is known and that the only unknown parameter is the placement of the function along a scale defined by the level of the stimuli. Under this assumption the model parameter \( \lambda \) is a scalar. An example of a psychometric function for a 6-alternative forced-choice task.
such as the MRT is shown in Figure 1. Note that the expected word score approaches chance (i.e., 1/6) for low SNRs and certainty for high SNRs.

Figure 1: Psychometric function used in experiments.

Although little is said about this in [4], the ML-MRT needs some practical modifications in order for it to work as required. Assume for instance that the first phrase is recognized correctly by the test subject. In its raw form the ML-MRT would then set the next stimulus level to zero (or minus infinity on a log scale), which clearly is an undesirable effect. This could be seen as an extreme example of the effect well known as overfitting, which happens when a model is tailored very precisely to an insufficiently small amount of data. The heuristic approach taken here to reduce these effects is to evaluate the psychometric function only for a span of candidate offset values around the initial stimulus level. This approach was also suggested in [4]. An alternative approach would have been to collect some more initial data before the ML criterion was applied, but this idea has not been investigated here.

Even with the heuristic described above, the ML-MRT will tend to have large fluctuations in stimulus levels for the first few utterances. Figure 2 shows an example of this. The diagram illustrates how an initial incorrect response from the subject will cause the procedure to increase the levels for the first few stimuli, until it converges back towards the initial level, which in this example coincides with the true level according to the simulated psychometric function. Apart from the impact these fluctuations will have on the quality of the resulting estimates, clearly one has to ensure that the stimulus levels are not so high that they could induce hearing damages to the test subject.

Figure 2: Example showing the stimulus level relative to the initial level for each of the utterances in the MRT sequence.

2.2. The extension to MAP-based adaptive MRT

The heuristic suggested in section 2.1 could be seen as a way to introduce prior knowledge into the ML-MRT. The MAP-based adaptive MRT (hereafter called MAP-MRT) described here suggests a way to introduce such prior knowledge into an adaptive MRT framework in a more formalized way.

Assume that some prior knowledge exists about the expected placement of the psychometric function along the abscissa, and that this knowledge can be parameterized as a pdf, \( P(\lambda | \Phi) \). The parameters of the prior distribution are given by the so-called hyperparameters, \( \Phi \). Given the prior knowledge and \( N \) responses from the test subject, the MAP estimate of the psychometric function can now be written as:

\[
\hat{\lambda}_n = \arg \max_{\lambda} P(\lambda | O_n, \Phi) = \arg \max_{\lambda} \left( \prod_{n=1}^{N} P(o_n | \lambda, \Phi) \right) P(\lambda | \Phi) \tag{3}
\]

where again it has been assumed that the observations are statistically independent. The observation likelihood has not changed with the introduction of the prior, so \( P(o_n | \lambda, \Phi) = P(o_n | \lambda) \) is still given by equation (2).

2.3. Parameters of the MAP-MRT

2.3.1. The psychometric function

The psychometric function assumed in the present study of the MRT is based on a Gaussian cumulative distribution function (cdf) on a decibel SNR scale. The slope of the function corresponds to a standard deviation \( \sigma_\text{pdf} = 6.8 \text{ dB} \) of the underlying Gaussian probability density function (pdf). This resembles the approach taken in [4]. The shape of the psychometric function is shown in Figure 1. Note that the cdf has been scaled and offset so that it reflects the expected behavior of a psychometric function.

2.3.2. Prior distribution

As it is often found in Bayesian approaches, the shape of the prior distribution is not easily determined. A common approach is to assume that the prior has a Gaussian distribution, and this is the strategy chosen here. The mean of the Gaussian distribution is chosen equal to the initial stimulus level. The standard deviation of the prior, \( \sigma_\text{prior} \), will be further discussed in conjunction with the experiments in section 3.

2.3.3. Optimal stimulus selection

The preceding discussion has established a way to determine the optimal psychometric function after \( N \) trials, both in an ML and in a MAP sense. Given this knowledge, the next question that needs to be answered is how to select the stimulus value for the next trial, i.e. how to select \( s_{n+1} \). It turns out that the minimum variance of the level estimates occurs for measurements conducted at a specific point on the response curve. This is called the sweetpoint, and is thoroughly discussed in [4][5]. The calculations are repeated here because both [4] and [5] contain the same typographical error in the key result.

The aim of selecting the sweetpoint is to minimize the variance of the estimate, i.e. the variance of the placement of the psychometric function along the abscissa. Call this variance \( \sigma^2 \). The variance along the ordinate is effectively the
variance of the probability estimate $p$. This is a binomial variable, and its variance is therefore given by

$$\sigma^2_p = p(1-p)$$  \hspace{1cm} (4)

Assume now that the psychometric function can be linearised over a small region, i.e.,

$$F(x) = y = ax + b$$  \hspace{1cm} (5)

The variance along the ordinate can then be expressed as

$$\sigma^2_y = E\left\{ (y - \mu_y)^2 \right\}$$

$$= E\left\{ (ax + b - a\mu_x - b)^2 \right\}$$

$$= E\left\{ a^2 (x - \mu_x)^2 \right\}$$

$$= a^2 \sigma^2 = \left( \frac{dF(x)}{dx} \right)^2 \sigma^2_x$$

By combining equations (4) and (6) it is found that

$$\sigma^2_x = p(1-p) \left( \frac{dF(x)}{dx} \right)^2$$  \hspace{1cm} (7)

where $dF/dx$ is the derivative of the psychometric function. This expression can be minimized by a straightforward numerical analysis. For the psychometric function described in section 2.3.1 the sweetpoint is found at a target word score of 65.8%.

### 3. Experiments

The purpose of the experiments reported here is to compare the results achieved with MAP-MRT and ML-MRT in terms of the bias and standard deviation of the estimates. This was achieved through Monte Carlo simulations. In these experiments, the subject was represented by a response curve with a shape identical to that used by the adaptive MRT algorithms. At each trial, the subject’s response was simulated by a random draw from a binomial distribution with a probability defined by the subject’s response curve at the selected stimulus level. After each simulated response the optimal (according to the ML and MAP criteria, respectively) psychometric function was found, and the next stimulus level was found according to the sweetpoint of this curve.

In the experiments, the candidate psychometric functions were selected such that they were located from $3\sigma_p$ (just above 20 dB) below to $3\sigma_p$ above the default location along the abscissa, with a resolution of 0.01 dB.

The main purpose of the ML-MRT is to get reliable estimates for relatively short trials, thereby reducing the cost of performing listening tests. In [4] it is argued that the adaptive run can be done with 25 test items or fewer, and that this also is sufficient to maintain the phonetic balance in the test material to a satisfactory degree. All experiments here are therefore done with a trial length of 25 utterances. In order to reliably estimate the bias and standard deviation, the Monte Carlo simulations were done with 1000 runs.

In the first simulation, it was assumed that the user’s response curve coincided with the default location assumed by the ML-MRT and MAP-MRT. The resulting bias (estimated level – true level) and standard deviation of the estimate are shown in Figure 3 for a range of different prior distributions, characterized by $\sigma_{\text{prior}}$. The results for the ML-MRT are shown as the rightmost data point in the graph, as an indication that the MAP-MRT would be identical to the ML-MRT in the case where $\sigma_{\text{prior}}$ is infinitely large, i.e., when the prior is uniformly distributed over the range of the candidate psychometric functions. The axis for the bias estimation is shown on the left-hand side of the plot, while the axis for the variance estimation is shown on the right-hand side. The same applies to the other figures presenting bias and standard deviation estimates.

As seen in Figure 3, the bias in this case is negligible, while the standard deviation increases with $\sqrt{\sigma_{\text{prior}}}$, and reaches its maximum for the ML-MRT. Clearly, this case where there is no mismatch between the initial estimate and the true answer is artificial, and this could not be expected in real life.

Next, it was investigated what happens when the initial estimates done by the adaptive MRT algorithms are offset compared to the true response curve. Figure 4 shows the case where the initial estimate used by the algorithm lies 10 dB below that of the true response curve, i.e., the intelligibility of the system being tested is significantly overestimated.

In this experiment, it can be seen that if $\sigma_{\text{prior}}$ is selected too small, the bias of the estimate will increase (in absolute terms). If the prior distribution is selected such that the standard deviation is small, this corresponds to putting high confidence in the prior knowledge. It is therefore as expected that the bias will increase in this case if there indeed is a larger offset between the initial estimate and the true response. The same kind of effect is also seen for less serious offsets, as in Figure 5 where the initial estimate used by the algorithm lies 5 dB below that of the true response curve.
Finally, in the case where the intelligibility is initially underestimated, similar effects are also seen. In this case the bias will have a different sign, but in terms of the impact of \( \sigma_{\text{prior}} \) there is no difference. Figure 6 shows the case where the initial estimate used by the algorithm lies 5 dB above that of the true response curve.

![Image](https://example.com/image.png)

Figure 5: Bias and standard deviation of estimates when the default psychometric function used in the algorithms lies 5 dB below that of the true response curve.

![Image](https://example.com/image.png)

Figure 6: Bias and standard deviation of estimates when the default psychometric function used in the algorithms lies 5 dB above that of the true response curve.

4. Discussion

Based on the experiments presented in the previous section, a value of \( \sigma_{\text{prior}} = 6\text{dB} \) is found to be a reasonable choice. This requires that the initial estimate is not off the true value by much more than 5 dB. The initial estimate could be found by calibration towards objective measurements like the speech transmission index, or by a pilot run for example based on ML-MRT, prior to the formal testing. The accuracy of this procedure would not have to be great, as it is seen from Figure 1 that a deviation of 5 dB could mean a difference in word score of up to 20 percentage points. An overview of the bias and standard deviation for various initial offsets for ML-MRT and MAP-MRT with \( \sigma_{\text{prior}} = 6\text{dB} \) is shown in Table 1.

The experiments in the previous section show, as expected, that there is a trade-off between keeping the bias low and the standard deviation low. In all cases the standard deviation of the ML-MRT estimate is above that of the MAP-MRT, while typically, the bias of the ML-MRT will be smaller than that of the MAP-MRT. However, with a reasonable choice of \( \sigma_{\text{prior}} \), the MAP-MRT can reduce the standard deviation of the estimate by around 0.5 dB compared to ML-MRT, while the bias is kept virtually unchanged. This improvement is small, but comes at a small cost.

<table>
<thead>
<tr>
<th>Initial offset</th>
<th>Bias</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ML</td>
<td>MAP</td>
</tr>
<tr>
<td>-15.0</td>
<td>-0.1</td>
<td>2.0</td>
</tr>
<tr>
<td>-10.0</td>
<td>-0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>-5.0</td>
<td>-0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>-2.5</td>
<td>0.0</td>
<td>-0.1</td>
</tr>
<tr>
<td>0.0</td>
<td>-0.1</td>
<td>-0.1</td>
</tr>
<tr>
<td>2.5</td>
<td>-0.3</td>
<td>-0.4</td>
</tr>
<tr>
<td>5.0</td>
<td>-0.2</td>
<td>-0.6</td>
</tr>
<tr>
<td>10.0</td>
<td>-0.3</td>
<td>-1.2</td>
</tr>
<tr>
<td>15.0</td>
<td>-0.2</td>
<td>-2.0</td>
</tr>
</tbody>
</table>

The final point made here is that the standard deviation of the ML-MRT is found to be higher when the intelligibility of the system is initially underestimated, compared to when it is overestimated by the same amount. In fact, the results indicate that a reasonable strategy would be to select the minimum stimulus level as the initial estimate. The reason for this property lies in the asymmetric characteristic of the response curve.

5. Conclusions

In this paper, a novel MAP based approach to adaptive measurements of subjective speech intelligibility is presented. It is shown through computer simulations that with reasonable assumptions of the prior distribution, the standard deviation of the resulting threshold estimate can be reduced by 0.5 dB, compared to a previously published procedure utilizing the ML principle. It is suggested that a small pilot is run prior to the listening tests to establish reasonable parameters for the required prior distribution.

6. Acknowledgements

This work was funded by the Norwegian Ministry of Defence and supported by BAE Systems Operations LTD under the contract EFA-IP-NORWAY-00010506.

7. References