ICA 2013 Montreal
Montreal, Canada
2 - 7 June 2013

Psychological and Physiological Acoustics
Session 3aPP: Auditory Physiology and Modeling (Poster Session)

3aPP14. The computational prediction of masking thresholds for ecologically valid interference scenarios
Khan Baykaner*, Christopher Hummersone, Russell Mason and Søren Bech

*Corresponding author's address: Institute of Sound Recording, Guildford, GU2 7XH, Surrey, United Kingdom, k.baykaner@surrey.ac.uk

Auditory interference scenarios, where a listener wishes to attend to some target audio while being presented with interfering audio, are prevalent in daily life. The goal of developing an accurate computational model which can predict masking thresholds for such scenarios is still incomplete. While some sophisticated, physiologically inspired, masking prediction models exist, they are rarely tested with ecologically valid programmes (such as music and speech). In order to test the accuracy of model predictions human listener data were required. To that end a masking threshold experiment was conducted for a variety of target and interferer programmes. The results were analysed alongside predictions made by the computational auditory signal processing and prediction model of (Jepsen et al. 2008). Masking thresholds were predicted to within 3.6 dB root mean squared error with the greatest prediction inaccuracies occurring in the presence of speech. These results are comparable to those in (Glasberg and Moore 2005) for predicting the audibility of time-varying sounds in the presence of background sounds, which otherwise represent the most accurate predictions of this type in the literature.

Published by the Acoustical Society of America through the American Institute of Physics
INTRODUCTION

When multiple listeners attend to different audio programmes in the same room the resulting cross-talk may be problematic. Such ‘auditory interference scenarios’ as these are of interest in situations where acoustically isolating solutions (such as headphones) are not desirable. An example of this type of scenario in the automotive domain would be if a passenger listens to music while the driver listens to instructions provided by a satellite navigation device. Another example, in the domestic domain, would be if multiple listeners occupy a living room in which some occupants want to watch a film while others want to listen to music. In both examples a sound zoning system, designed to control the auditory interference within spatial zones, could be utilised to minimise the effect of each unwanted programmes upon each listener, but the performance of such a system might depend heavily on the programmes involved and their relative levels. It would be useful if the performance of such systems could be objectively evaluated, without the need for listening tests, one aspect of which would be the prediction of the audibility of the interfering programme in each spatial zone. A significant amount of research has been undertaken to predict the masking effects of simple stimuli such as tones and noise, and some work has been undertaken regarding self-masking of more externally valid signals (e.g. for the perceptual models within low bit-rate audio codecs). Effective prediction of the auditory masking of one item of audio programme material by another, however, remains elusive.

The primary goal of this paper is to investigate the prediction of auditory masking for the purposes of evaluating the effectiveness of a system designed to control auditory interference, specifically within domestic and automotive environments. To that end, this paper investigates the adaptation and testing of one masking model, the Computational Auditory Signal processing and Perception (CASP) model of Jepsen et al. (2008), for use as a predictor of the auditory masking of ecologically valid programme material.

MASKING THRESHOLD EXPERIMENT

A masking threshold experiment was conducted to gather masking data for a range of ecologically valid programmes. The experimental method, stimulus details, and simulation parameters are described below.

Methodology And Equipment

The listening position was near the centre of a room meeting the specifications of ITU-R BS.1116 (1997) with one Genelec 8020A loudspeaker and one Genelec 1032 loudspeaker positioned directly in front at distances of 1.85 m and 2.2 m respectively. A further six Genelec 8020A loudspeakers were arranged in a regular hexagonal formation around the listening position at a distance of 2.0 m. All Genelec 8020As were positioned at a height of 78 cm and the Genelec 1032 was positioned at a height of 104 cm (approximately head height for a seated listener).

One audio programme (the target) was replayed via the frontal 8020A, and a different audio programme (the interferer) was replayed via the 1032. The hexagonal array was used to reproduce road noise on half of all trials. This loudspeaker arrangement was selected as a simple way to approximate the envelopment experienced when in an automobile. The subjects were given the following instructions: “You will be presented with two audio programmes; you can control the level of one of the programmes. The controllable programme will start at a level where it is audible. Using the rotary folder, please adjust the level of the controllable programme to the point where it is just inaudible”.

This methodology, known as the ‘method of adjustment’, is sometimes considered to be less accurate than other psychophysical test methods, such as alternative forced choice (AFC) style
procedures. Hesse (1986) tested the effect of a range of psychophysical procedures on masking thresholds for tone masked by noise. Thresholds fell into two groups: AFC style procedures; and non-AFC procedures including method of adjustment, adaptive control and yes/no procedures. Thresholds obtained using AFC procedures were about 2 dB lower than those found using the non-AFC procedures. Similar results were found in Watson and Nichols (1976). In an ecologically valid interference scenario, however, a listener is likely to be concerned only with whether the interfering programme is audible, which is similar to a yes/no paradigm. Since yes/no masking thresholds were very close to those using a method of adjustment, and since the method of adjustment task is both fast and intuitive, this procedure was considered appropriate for use.

Ten listeners who reported no hearing impairments, aged between 21 and 38 years, participated in the experiment. Four subjects had training in critical listening and experience conducting and participating in psychoacoustic experiments, three subjects had no such experience but were musicians, and three subjects had no experience in any of these domains.

**Stimuli**

Three items of target programme material and three items of interferer programme material were used in this experiment. All stimuli were 10 second excerpts (which was considered sufficiently short to allow for a reasonable number of trials to be conducted, yet sufficiently long to include realistic programme variability) which looped indefinitely. The targets and interferers were selected to cover a range of programme types and genres. The targets were excerpts of: classical music (Brahms’s Hungarian Dance No.18), pop music (The Killers’ On Top), and football commentary. The interferers were excerpts of: classical music (Mahler’s Symphony No. 5 Mov. 4), pop music (The Bravery’s Give in), and male speech (from the BBC Radio 4 show ‘Points of View’).

Any system designed to control interference between signals may have some effect on the magnitude spectrum of the interferer signal. In order to consider this a further six interferers, filtered replicates of the first three, were also used. Three were low pass filtered (LPF) at 200 Hz with a 9 dB/oct roll-off, based upon the results of Akeroyd et al. (2007), and three were high pass filtered (HPF) at 1 kHz with a 16 dB/oct roll-off, based upon the results of Jacobsen et al. (2011).

A single channel recording of road noise was decorrelated according to the method described in Pulkki (2007) and replayed over the 6 channel hexagonal loudspeaker array.

Benjamin and Crockett (2005) identified preferred listening levels for music in the automotive environment at between 70 and 76 dBA for a range of vehicle speeds including stationary (engine off), thus the target programmes were reproduced at a level of 76 dB LAeq measured at the listening position with a time constant of 20 seconds (i.e. programme replayed twice). The road noise was adjusted to 60 dB LAeq which was found to be a good approximation for road noise levels inside automobiles travelling at 30 mph in the above mentioned study. The interferers were set to a starting level which was randomly selected between 70 and 76 dB LAeq in order to minimise the opportunity for listeners to select the masking threshold by recalling the number of rotations of the rotary fader used on a previous trial.

**Masking Experiment Results**

Shapiro–Wilk tests of sample size n=20 showed that when the data were separated by target, interferer, road noise, and filtering, 11 of the 54 groups were not normally distributed with 95% confidence. Observations of the histograms provided little evidence to support or refute this due to the relatively low sample size per group, so all data were analysed using both parametric and non-parametric tests. No discrepancies between results were found so only results of the parametric tests are reported here.
An ANOVA revealed that all factors and interactions between target programme, interferer programme, interferer filtering, and road noise were significant, with the exception of the interactions between interferer programme and road noise, and higher level interactions featuring this combination (see table 1). Some of the factors had a smaller effect size than their higher level interactions. The partial $\eta^2$ of the interferer programme (0.341), and that of the target-interferer interaction (0.397), for example, were less than that of the three way interaction between target, interferer, and filtering (0.443). This implies that the role of the interferer was notable only in its interaction with its filtering or with the target programme. Similarly, the significant three way interaction between target, noise, and filtering had approximately the same effect size (0.469) as the two way interaction between target programme and noise (0.468).

Figures 1 and 2 show the significant three way interactions between target, interferer and filtering, and the interactions between target, road noise, and filtering, respectively. Of particular interest in figure 1 is the effect of LPF on the interferer to target ratio (ITR) of various target-interferer combinations, which was negligible in some cases and greater than 7 dB in others. Of special interest in figure 2 is the effect of LPF on ITR when the sports commentary target was presented with and without road noise.

**MODIFICATIONS TO MASKING THRESHOLD MODEL**

The CASP model makes masking threshold predictions by passing a known target and interferer programme through a series of processes which mimic the response of the human auditory system. When the signal and interferer programmes have been transformed into an ‘internal representation’ (IR) of the mixture, this IR is correlated with a template IR which is based on the same combination of programmes but with the target programme presented at a very high level (and thus assumed to be audible). The process is repeated for the interferer presented in isolation, and the difference between these two correlations is used to calculate a probability of detection via an optimal detector (described in Green and Swets (1996)). For a detailed description of the CASP model see Jepsen et al. (2008). A number of modifications were necessary to this process in order to adapt the model for the task considered in this paper.

Firstly, the desired output of the model was a prediction of the highest level at which an interferer may be replayed while remaining inaudible (i.e. the masking threshold), rather than a prediction of the probability of audibility. The distribution of masking thresholds for a large

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean sq.</th>
<th>F</th>
<th>Sig.</th>
<th>Partial $\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>918860.258</td>
<td>1</td>
<td>918860.258</td>
<td>932.068</td>
<td>.000</td>
<td>.990</td>
</tr>
<tr>
<td>Target</td>
<td>33705.706</td>
<td>2</td>
<td>16852.853</td>
<td>61.139</td>
<td>.000</td>
<td>.872</td>
</tr>
<tr>
<td>Interferer</td>
<td>1105.783</td>
<td>2</td>
<td>552.892</td>
<td>4.649</td>
<td>.024</td>
<td>.341</td>
</tr>
<tr>
<td>Noise</td>
<td>778.253</td>
<td>1</td>
<td>778.253</td>
<td>22.457</td>
<td>.001</td>
<td>.714</td>
</tr>
<tr>
<td>Filter</td>
<td>3924.350</td>
<td>2</td>
<td>1962.175</td>
<td>41.614</td>
<td>.000</td>
<td>.717</td>
</tr>
<tr>
<td>Tar*Int</td>
<td>1805.065</td>
<td>4</td>
<td>451.266</td>
<td>5.930</td>
<td>.001</td>
<td>.397</td>
</tr>
<tr>
<td>Tar*Noi</td>
<td>644.074</td>
<td>2</td>
<td>322.037</td>
<td>7.934</td>
<td>.003</td>
<td>.468</td>
</tr>
<tr>
<td>Tar*Fil</td>
<td>2024.378</td>
<td>4</td>
<td>506.095</td>
<td>21.323</td>
<td>.000</td>
<td>.703</td>
</tr>
<tr>
<td>Int*Fil</td>
<td>2932.227</td>
<td>4</td>
<td>733.057</td>
<td>34.757</td>
<td>.000</td>
<td>.794</td>
</tr>
<tr>
<td>Noi*Fil</td>
<td>678.453</td>
<td>2</td>
<td>339.227</td>
<td>22.826</td>
<td>.000</td>
<td>.717</td>
</tr>
<tr>
<td>Tar<em>Int</em>Fil</td>
<td>1222.464</td>
<td>8</td>
<td>152.808</td>
<td>7.175</td>
<td>.000</td>
<td>.443</td>
</tr>
<tr>
<td>Tar<em>Noi</em>Fil</td>
<td>648.349</td>
<td>4</td>
<td>162.087</td>
<td>7.957</td>
<td>.000</td>
<td>.469</td>
</tr>
</tbody>
</table>

**TABLE 1:** ANOVA for the effect of target programme, interferer programme, road noise, filtering, and all significant interactions between these factors on the observed masking thresholds.
**FIGURE 1:** These plots show the ITRs separated by target, interferer and filtering. The left plot shows ITRs for the classical interferer, the middle plot shows ITRs for the pop interferer, and the right plot shows ITRs for the male speech interferer. The solid lines represent the classical target, the dotted lines represent the pop target, and the dashed lines represent the sports commentary target.

**FIGURE 2:** These plots show the ITRs separated by target, road noise, and filtering. The left plot shows ITRs without road noise, and the right plot shows ITRs with road noise. The solid lines represent the classical target, the dotted lines represent the pop target, and the dashed lines represent the sports commentary target.
population will likely be asymptotic, and thus there is no way to predict the threshold at which all listeners will agree that a signal is masked (i.e. the 100% masked point). Thus the 50% masked point (the mean ITR) can be taken as a useful indicator of the masking threshold. The desired output is therefore the interferer level at which the CASP model outputs a probability of detection at 50%; this was obtained by repeating the CASP model processing with the interferer level adjusted prior to each repetition according to a binary search algorithm.

Secondly, the excerpts under test in this experiment (and those encountered within the automotive and domestic scenarios considered) were significantly longer in duration (10 seconds) than those the CASP model was tested on in Jepsen et al. (2008) (from 50 ms to 1 second). Since programmes may be of arbitrary duration within the listening scenarios considered, it is necessary to perform processing on relatively brief temporal windows of the data. In this instance temporal windows were selected to be 200 ms in duration (within the range previously tested), and did not overlap, although a variety of parameter selections may be appropriate here. Once the levels which correspond to a probability of detection of 50% had been calculated for each temporal window, the temporal window with the lowest ITR was selected (see figure 3). This process is based on the assumption that listeners selected masking thresholds by attending to the entirety of the 10 second programme, isolating the section wherein the interferer was most easily detectable, and adjusting the level until the interferer was no longer audible in that section. Other possibilities for interpreting the set of masking predictions exist, such as averaging across temporal windows, however these were not investigated in this work.

Finally, in Jepsen et al. (2008) the model was calibrated using a tone intensity discrimination task. This calibration was found to produce overly sensitive predictions (i.e. signals were predicted to be much too easily detected than the results from the masking experiment indicated), so a post calibration was implemented. This involved using a gradient descent function to find the linear transform which optimised the predictions for the given training data.

**Evaluation Of Masking Threshold Predictions**

The accuracy of the CASP model was measured by calculating the root mean squared error (RMSE) of predictions. It is also important to consider the accuracy of the model when extrapolating to new data, however, so the 54 programme combinations were split into training (T) and cross validation (CV) data sets. In each case the data was split into 38 T items (around 70% of cases) with 16 (around 30%) CV items. Ideally every possible combination of T and CV programmes would be analysed however, since there are $2.1 \times 10^{13}$ possible combinations, this would be very computationally expensive. Instead 5000 random permutations were analysed with the assumption that this would provide a sufficiently representative sample of these combinations.

The average accuracy, across the 5000 permutations, of the predictions for the T and the CV data are both presented. The difference between the two can be considered a measure of the model's ability to extrapolate to new data. Epsilon insensitive RMSEs (RMSE*), where errors are the difference between the prediction and the closest edge of the 95% confidence interval of the mean of the human masking data or 0 where predictions which fall within the 95% confidence interval, were also calculated, which describe accuracy ‘after’ listener error.

A Kolmogorov–Smirnov test showed that the 5000 CV RMSEs were not normally distributed so median, rather than mean, scores are considered. The median CV RMSE was 3.58 dB, while the median RMSEs for the T data was 3.37 dB. The difference between the RMSE of the T and CV data was therefore 0.21 dB, indicating that the model extrapolates to new cases with little additional error. It may be argued, however, that the CV and T data are not truly independent, since many programmes are filtered replicates, and thus this test of extrapolating to new data is optimistic. A more cautious estimate could be made using the case with the greatest disparity between T and CV data, which had T and CV RMSEs of 2.86 dB and 5.12 dB respectively, giving
a much greater difference of 2.26 dB.

The RMSE's were 1.78 and 1.87 dB for the T and CV data respectively, indicating that the average additional error (beyond the edge of the human listener's 95% confidence interval) was less than 2 dB.

Figure 4 shows a scatter plot of the observed ITRs against the median predicted ITRs. A strong positive correlation was found between the predicted and the observed ITRs of $R = 0.87$ ($p<0.001$), indicating a relatively linear relationship.

Further analysis revealed that prediction errors were fairly evenly distributed across different levels of the factors: target programme, interferer programme, road noise and filtering, with a few exceptions. Two cases were predicted noticeably worse than the others: those where the target was pop music and the interferer was HPF male speech with and without road noise. The predictions were that the speech would need to be reduced by 8.0 and 7.3 dB, respectively, more than was observed as necessary in order to be masked.

It was also found that the median absolute difference between predictions and observations for classical target programmes and for sports commentary target programmes were 1.99 dB and 2.38 dB respectively, whereas the median absolute difference between predicted and observed ITRs for the pop target programmes was 3.99 dB. The majority of these cases (13 of the 18) were predicted as requiring greater reduction in interferer level than was necessary, with 10 of those cases having an error exceeding 3.5 dB. Notable exceptions were for pop targets and classical interferers for all levels of filtering without road noise, where the model underestimated the reduction in the interferer level which would be required by 5.86, 4.28, and 3.72 dB for unfiltered, LPF, and HPF respectively.
FIGURE 4: Correlation between the median ITRs and the observed ITRs.

It seems, therefore, that the model tended to overestimate the reduction in interfer level which was necessary when the target was pop music (most severely when the interferer was speech), except where the interferer was classical music, where the model underestimated the necessary reduction.

CONCLUSION

Masking thresholds were collected for a range of ecologically valid programmes in order to test the CASP model in automotive and domestic environments. The CASP model was modified to produce ITRs, and the predicted and observed ITRs were compared. The results indicated that this implementation of the CASP masking model was a reasonable predictor of the auditory masking of one ecologically valid programme on another. The median CV RMSE was 3.6 dB with a median CV RMSE* of 1.9 dB, and the least accurate prediction occurred with a speech interferer and differed from observation by 8.0 dB. The results are comparable to those of Glasberg and Moore (2005), who reported a RMSE of 3 dB across test conditions with greatest prediction error of 8.2 dB for ‘pub’ background noise.

ACKNOWLEDGEMENTS

We would like to acknowledge the financial support provided by Bang & Olufsen on this work. We would also like to thank Jon Francombe for his help with preparations for the listening tests, and the subjects that took part in the test for their time and hard work.
REFERENCES


