SELECTION OF TEMPORAL WINDOWS FOR THE COMPUTATIONAL PREDICTION OF MASKING THRESHOLDS

Khan Baykaner[*], *Christopher Hummersone,* *Russell Mason*

Institue of Sound Recording
University of Surrey
Guildford, Surrey, England, GU2 7XH, UK
k.baykaner@surrey.ac.uk

Søren Bech

Bang & Olufsen
Peter Bangs Vej 15, 7600, Struer, Denmark

ABSTRACT

In the field of auditory masking threshold predictions an optimal method for buffering a continuous, ecologically valid programme combination into discrete temporal windows has yet to be determined. An investigation was carried out into the use of a variety of temporal window durations, shapes, and steps, in order to discern the resultant effect upon the accuracy of various masking threshold prediction models. Selection of inappropriate temporal windows can triple the prediction error in some cases. Overlapping windows were found to produce the lowest errors provided that the predictions were smoothed appropriately. The optimal window shape varied across the tested models. The most accurate variant of each model resulted in root mean squared errors of 2.3, 3.4, and 4.2 dB.

Index Terms— Masking, Threshold, Temporal, Prediction

1. INTRODUCTION

Ever since Fletcher's seminal work on auditory patterns [1] there has been steady progress towards the goal of predicting the occurence of the phenomenon known as auditory masking. Fletcher noticed that the human auditory system behaves like a bank of bandpass filters wherein the audibility of stimuli are constrained. Further advancements in this field included the development of gammatone filters [2] and excitation patterns [3]. Today, the most advanced auditory masking models use a chain of processes modelling the behaviour of the physical and neuronal systems (as in [4], [5]).

Such models are usually tested using simple, well defined stimuli (e.g. tone-in-noise) rather than ecologically valid stimuli (e.g. music and speech), in order to control for confounding variables when evaluating the prediction of distinct masking phenomena. When predicting the audibility of ecologically valid stimuli, the duration is often unknown. Determining the audibility of stimuli cannot, therefore, be achieved using a single batch process but must split the input into separate temporal windows. This raises the question: what kind of temporal window should be used in a masking threshold prediction model? A temporal window can be described using three parameters: duration, shape, and step (through time). The literature on human audition was consulted as a starting point, on the assumption that human listeners perform a similar function.

The human auditory system generally does not respond to signals briefer than approximately 2-20 ms [6, 7, 8, 9], which can be taken as a lower bound for the duration of the temporal window. In [10] (exponential) auditory temporal windows were suggested in which most of the energy was contained within approximately 100 ms, although it was noted that longer durations may be more appropriate for continuous signals. Results from electroencephalography studies indicate that temporal integration within the auditory cortex occurs over a period on the order of several hundred milliseconds ([11],[12]). Therefore temporal windows from 20 to a few hundred milliseconds are of most interest.

As for the shape of the temporal windows, it was demonstrated in [8] that human temporal windows are likely to have gradual skirts and in [10] the psychophysical data was mapped to an asymmetric exponential function. Hann windows have also been used in masking models ([4]).

Finally the step size of the temporal window was considered. Since auditory perception is free to respond to drastic changes in stimuli after only a few milliseconds, it appears that the temporal window 'steps' through time at a rate which implies significant overlap of windows. It may be reasonable to suggest than an optimal step size would therefore be equivalent to only a few milliseconds, yet it is not clear if this will be the case for ecologically valid stimuli. It is additionally worth investigating how detrimental to prediction accuracy the use of larger steps may be.

[*]The authors are grateful to Bang and Olufsen for funding this research.

Although some broad answers to the previous questions are known for human audition, it should not be assumed that the optimal temporal windows for masking threshold prediction models must be identical to these, since such models do not perfectly replicate the human auditory system. Thus, in order to find trends among the parameters for temporal windows an experiment is described in this paper to predict masking thresholds of ecologically valid time-varying stimuli.

2. COLLECTION OF MASKING THRESHOLD DATA

In order to evaluate the performance of masking threshold prediction models it is necessary to compare the predictions with known masking data. A masking threshold experiment was therefore conducted. Ten listeners reporting no hearing problems, aged between 21 and 38 years, and with a range of musical proficiency, participated in the experiment.

2.1. Methodology and equipment

The subjects were seated near the centre of a listening room meeting the specifications described in [13], with one target loudspeaker (Genelec 8020A) positioned 1.85 m directly in front and at a height of 0.78 m, and one interferer loudspeaker (Genelec 1032) positioned 2.2 m directly in front at a height of 1.04 m. This arrangement allowed both loudspeakers to be approximately head height without causing significant occlusion of either stimulus.

The subjects used an unmarked rotary fader to interact with a computer. The computer simultaneously replayed one audio programme (the target) via the target loudspeaker, and a different audio programme (the interferer) via the interferer loudspeaker. The subjects were instructed to adjust the level of the interferer until it was rendered 'just inaudible'.

2.2. 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, looped indefinitely. The programmes were selected to cover a range of types and genres, and included excerpts of classical music (Brahms's Hungarian Dance No.18), pop music (The Killers' On Top), and sports commentary (from a football match), for the targets and 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') for the interferers.

The target programmes were reproduced with a level of 76 dB LAeq (20 s time constant) measured at the listening position. The interferers were set to randomly selected starting levels between 70-76 dB LAeq in order to prevent listeners from simply repeating the previous fader rotation.

The experiment design was full factorial with two repetitions per trial, thus there were 18 trials per listener.

Finally, the stimuli were recorded in the listening room using a Cortex MK2 head and torso simulator, in order to produce the input stimuli for the masking threshold models.

3. MODEL STRUCTURE

Figure 1 shows an overview of the test model structure used to investigate combinations of temporal window parameters.

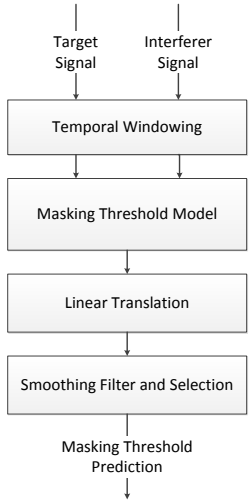

Fig. 1. Overview of the test model.

3.1. Temporal windowing

The temporal window durations tested were 100 ms, 200 ms, 300 ms, and 400 ms. The steps (distance between onset of adjacent temporal windows) tested were also 100 ms, 200 ms, 300 ms, and 400 ms, giving a maximum tested overlap of 75% (400 ms duration with 100 ms step). While it is possible that shorter steps or durations would improve prediction accuracy further, the processing time required to make such predictions quickly became impractically long. Four shapes were considered: instantaneous onset and offset (rectangular window), 50 ms raised cosine onset and identical offset (Hann window), 50 ms exponential onset with identical offset (Exp), and an exponential onset equal to 3/4 of the duration, with the remaining 1/4 an exponential offset (ExpSlope). This final, assymetrical window was selected to approximate the phenomena of forward and backward masking. The parameters were tested in all combinations except where the step exceeded the duration.

3.2. Masking threshold prediction

Three masking threshold models were implemented and tested: CASP (based on the model described in [4]), Loud_M and Loud_Z (based on the loudness model described in [5] and [14] respectively).

For a detailed description of the CASP model see [4]. In brief, the CASP model passes a known target and interferer through a series of processes which mimic the response of the human auditory system. These include a dual resonance non-linear filterbank (to model the frequency selectivity of the cochlea) and an adaptation loop (which functions similarly to temporal integration), and finally result in an 'internal representation' of the signal. Cross correlations are calculated between the internal representation of this mixture and a template mixture in which the signal is known to be audible. This is compared with a cross correlation between the internal representations of the interferer alone and the template mixture, which allows for the prediction of a probability of detecting the interferer. In order to adapt the model for the task considered in this paper, it was necessary to obtain masking thresholds, rather than a probability of detection. To do this a pre-specified probability of detection (in this case 50%) was selected at which the interferer was considered 'unmasked' and the corresponding interferer level was identified by running the model repeatedly with the interferer level adjustments prior to each run. A simple binary search algorithm was implemented to reduce the processing time required to identify the interferer level to within 0.001 dB.

Loud_M and Loud_Z, were adapted such that for each temporal window the maximum long-term loudness value (in Loud_M) and the instantaneous loudness level (in Loud_Z) were calculated for the target and interferer independently. The difference in loudness (D) between target and interferer was used to estimate the probability of detection (P) by mapping to a logistic function of the form

$$\text{P} = \frac{1}{1 + e^{-(1+0.1\text{D})}}. \quad (1)$$

The stimulus was assumed to be unmasked where P>0.5.

3.3. Linear translation

It was considered possible that an optimal calibration (i.e. producing the most accurate predictions) may exist for each masking threshold model and every combination of temporal window parameters under test. To find this array of optimal calibrations prior to testing would require an extreme processing cost, thus a post-calibration strategy was adopted. This involved using a gradient descent function to find the linear transform which produced the most accurate predictions for every model and window combination. While it might be possible to obtain more accurate masking threshold predictions by finding the optimal calibration of the masking threshold prediction model, it seems likely that the gains would be small.

3.4. Smoothing filter and selection

In order to produce a single masking threshold prediction from the set of predictions obtained (one per window) the lowest threshold was selected as the final prediction, on the assumption that listeners determine the masking threshold by attending to the moment in which the interferer is most easily audible. Additional tests were conducted in which, prior to selection, a moving average filter was applied to the set of predictions. Each filter averaged every masking threshold prediction with an equal number of adjacent predictions in both directions. Every odd filter width from 1 (no smoothing) to 23 predictions was tested. To avoid problems of end effects, the smoothing filter was set to 'wrap around' the test file such that smoothing applied to predictions at one end were averaged with predictions made at the other.

4. ANALYSIS

For each model setup a root mean squared error (RMSE_σ) between reported and predicted masking thresholds (across all listening scenarios) was calculated as a measure of accuracy. Additionally the mean error for each model, trained on every combination of 8 of the 9 listening scenarios, was taken as a cross validation measure of accuracy (RMSE_{CV}). The difference between RMSE_{CV} and RMSE_σ was used as an indicator of robustness to extrapolation (RMSE_δ), where a lower RMSE_δ indicates a more robust model. 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 for predictions which fall within the 95% confidence interval — were also calculated, which describe accuracy 'after' listener error.

4.1. Effect of duration and step (overlap)

A trend was found across all models and window shapes in which greater durations for a fixed step size (i.e. greater overlap) required wider smoothing filters to minimise prediction error (see fig. 2). This may be because as windows overlap the data is processed repeatedly, thus a greater number of windows should be averaged across in order to describe the same section of the stimuli. Beyond the optimal smoothing filter width (for a fixed overlap) the accuracy decreased as the filter widened, because the analysis tended towards an average masking threshold prediction across the whole stimuli.

The optimum width of smoothing filter was model dependent. Loud_Z had an optimum smoothing filter width at 11 predictions (for the greatest overlap tested), whereas Loud_M and CASP had optima at 3-5 predictions (depending on overlap). The Loud_Z performance may be due to its large prediction errors when the interferer was speech. In general, for speech interferers more accurate predictions can be made using the widest smoothing filters (i.e. the long-term average loudness ratio between the music and speech), and for music interferers more accurate predictions can be made using very narrow filters (i.e. the short-term average loudness ratio was a better predictor of audibility because the listener could

listen 'in the gaps' of the speech). For Loud$_Z$ relatively low average RMSE$_\sigma$s were obtained smoothing over 9-13 predictions, since this was a compromise between the narrow filters required for music interferers, and the wide filters required to reduce the error for speech interferers.

It should be noted that the most accurate prediction using Loud$_Z$ was produced with minimal smoothing (3 predictions) and no overlap (see section 4.3), with those predictions made using 100 ms step, 400 ms duration, and smoothing width 11 performing marginally worse. More generally, it was found that over the range of conditions tested both Loud$_Z$ and Loud$_M$ produced very similar RMSE$_\sigma$s in the best cases whether using no overlap and no smoothing or using significant overlap with optimal smoothing. CASP, however, produced RMSE$_\sigma$s 0.6 dB lower when using the ideal overlap and smoothing, than the most accurate prediction made without overlap (which was made using 300 ms step and 300 ms duration). This is possibly because the cross correlation approach employed by CASP may be more susceptible to the presence of transients (which larger overlap reveals) than the loudness ratio approach employed by Loud$_Z$ and Loud$_M$.

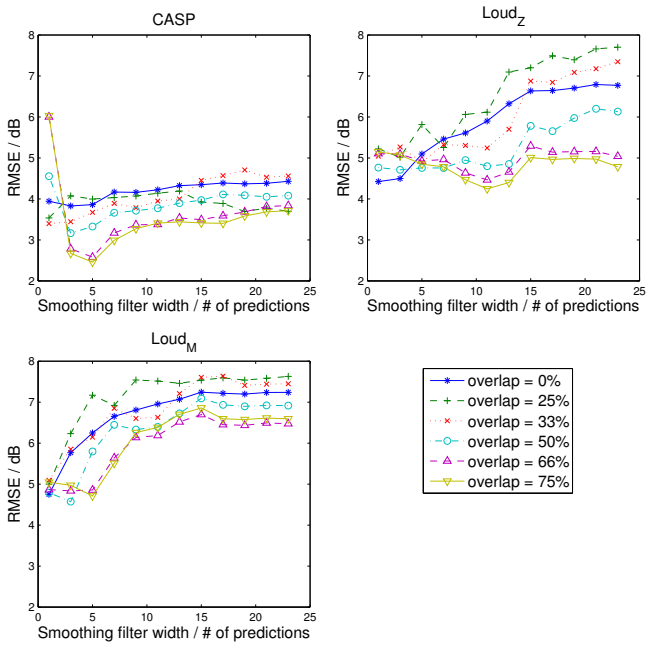

Fig. 2. RMSE$_\sigma$ calculated across target-interferer combinations based on width of smoothing filter and window overlap. Values shown are averaged across step and duration conditions with identical overlap, and across window shapes.

4.2. Effect of window shape

For all models, durations and steps, the window shape produced only small differences (usually less than 0.3 dB). It is likely that this is because over a relatively long duration (i.e. 10 second programme) a sufficient number of windows will be examined that other forces, such as selection strategy and smoothing of predictions, will have a larger effect than the weighting of data within each window.

For CASP the rectangular window usually produced the most accurate predictions. For Loud$_Z$ and Loud$_M$ there did not seem to be a consistent best shape to choose. For every combination of step and duration the maximum difference in error between window shapes was calculated for each model. The average of these maximum differences in prediction accuracy were 0.14, 0.19, and 0.43 dB for CASP, Loud$_Z$, and Loud$_M$ respectively.

4.3. Prediction accuracy

The most accurate predictions in this set of tests were made by CASP, with the best case having an RMSE$_\sigma$ of 2.3 dB and an RMSE* of 0.8 dB (using a rectangular 400 ms window stepping by 100 ms smoothed over 5 predictions). A range of similar results were produced using other window shapes, and using 300 ms duration windows. The RMSE$_\delta$ followed a similar trend to the RMSE$_\sigma$, with the most robust models also being the most accurate (RMSE$_\delta < 0.2$ dB) and the least accurate models having the greatest RMSE$_\delta$ ($0.8 - 1.4$ dB).

The most accurate predictions produced using the Loud$_Z$ model had an RMSE$_\sigma$ of 3.4 dB and an RMSE* of 2.1 dB, and were made using a 400 ms ExpSlope window stepping by 400 ms smoothed over 3 predictions. As with CASP, the RMSE$_\delta$ was lowest for those cases which were most accurate (< 0.2 dB) and the less accurate models had greatest RMSE$_\delta$s (up to 0.7 dB).

For Loud$_M$ the most accurate predictions resulted in an RMSE$_\sigma$ of 4.2 dB and an RMSE* of 2.9 dB, and were made using a 400 ms Hann window stepping by 200 ms, smoothed over 3 predictions.

5. CONCLUSION

CASP produced the most accurate predictions with a RMSE of 2.3 dB in the best case (compared to 3.4 dB and 4.2 dB for Loud$_Z$ and Loud$_M$ respectively).

Shorter steps and longer durations (i.e. greater overlap) generally produced the most accurate predictions, provided that the optimal width of smoothing filter was selected. Conversely, low overlap predictions had low error when no smoothing was applied. For CASP the use of overlap and smoothing reduced RMSE by 0.6 dB compared to the best prediction made without, whereas the effect was very small for Loud$_Z$ and Loud$_M$.

Window shape had a small effect (usually less than 0.3 dB), and the optimal selection was specific to each model.

6. REFERENCES

[1] Harvey Fletcher, "Auditory patterns," *Reviews of Modern Physics*, vol. 12, no. 1, pp. 47–65, Jan. 1940.

[2] R. D. Patterson, I. Nimmo-Smith, J. Holdsworth, and P. Rice, "Spiral Voice Operated Switch: phase 1, final report, the auditory filterbank," *Ministry of Defense*, August 1988.

[3] B. C. Moore, J. I. Alcántara, and T. Dau, "Masking patterns for sinusoidal and narrow-band noise maskers.," *The Journal of the Acoustical Society of America*, vol. 104, no. 2 Pt 1, pp. 1023–38, Aug. 1998.

[4] M. L. Jepsen, S. D. Ewert, and T. Dau, "A computational model of human auditory signal processing and perception.," *The Journal of the Acoustical Society of America*, vol. 124, no. 1, pp. 422–38, July 2008.

[5] B. R. Glasberg and B. C. Moore, "A model of loudness applicable to time-varying sounds," *The Journal of the Audio Engineering Society*, vol. 50, no. 5, pp. 331–342, 2002.

[6] R. Plomp, "The ear as a frequency analyzer," *The Journal of the Acoustical Society of America*, vol. 36, no. 9, pp. 1628, 1964.

[7] D. Ronken, "Monaural detection of a phase difference between clicks," *The Journal of the Acoustical Society of America*, vol. 47, 1970.

[8] Irwin Pollack, "Forward, backward and combined masking : Implications for an auditory integration period," *Quarterly Journal of Experimental Psychology*, pp. 37–41, February 2007.

[9] A. J. Oxenham, B. C. Moore, and D. Vickers, "Short-term temporal integration: evidence for the influence of peripheral compression," *The Journal of the Acoustical Society of America*, vol. 101, no. 6, pp. 3676–87, June 1997.

[10] B. C. J. Moore, B. R. Glasberg, C. J. Plack, and A. K. Biswas, "The shape of the ear's temporal window," *The Journal of the Acoustical Society of America*, vol. 83, 1988.

[11] Hirooki Yabe, Mari Tervaniemi, Janne Sinkkonen, and Minna Huotilainen, "Temporal window of integration of auditory information in the human brain," *Cognitive Brain Research*, pp. 615 – 619, 1998.

[12] Naoko Shinozaki, Hirooki Yabe, Yasuharu Sato, Tomiharu Hiruma, Takeyuki Sutoh, Takashi Matsuoka, and Sunao Kaneko, "Spectrotemporal window of integration of auditory information in the human brain," *Brain research. Cognitive brain research*, vol. 17, no. 3, pp. 563–71, Oct. 2003.

[13] ITU-R BS.1116, "Methods for subjective assessment of small impairments in audio systems including multichannel sound system," Tech. Rep., International Telecommunications Union - Radiocommunication, 1997.

[14] E Zwicker and H Fastl, *Psychoacoustics: Facts and Models*, Springer-Verlag Berlin Heidelberg, 1999.