PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Pupil dilation reveals changes in listening effort due to energetic and informational masking

James WOODCOCK; Bruno M FAZENDA; Trevor J COX; William J DAVIES[1]

University of Salford, UK

ABSTRACT

Pupil dilation has previously been shown to be a useful involuntary marker of listening effort. An inverse relationship between pupil diameter and signal to noise ratio has been shown when speech is energetically masked by noise. The work reported here aimed to investigate whether this relationship also holds for informational masking. Informational masking is a concept used in soundscape research to represent the distraction from the target sound that comes from a masking sound that is also highly salient. To investigate the effect of informational masking on listening effort, eighteen normal-hearing participants completed a speech-in-noise task in which they were asked to identify words in short sentences presented in combination with four different types of masker (competing speech, speech modulated noise, and urban and nature soundscapes) at different levels of energetic masking set using a distortion weighted glimpse proportion model. Time varying pupil dilation was measured over the course of each sentence presentation. A repeated-measures ANOVA showed a significant main effect of both the level of energetic masking and the masker type on the mean pupil dilation ($p < 0.05$). These results suggest that pupil dilation reveals changes in listening effort due to both energetic and informational masking.

Keywords: Speech, Soundscape, Masking

1. INTRODUCTION

Trying to listen to one person talking when someone else is talking is often difficult, requiring the listener to concentrate. The effect of the second talker is often modeled as exerting informational masking as well as energetic masking (1). Energetic masking happens when the masking signal contains energy in one or more of the same critical bands as the target signal and makes this part of the target signal inaudible. Energetic masking happens in the peripheral auditory system. Informational masking refers to the extra masking effect that happens in the brain when both target and masker are audible but the listener attends to (parts of) the masker instead of the target. The term informational masking has sometimes been used to refer to a distraction effect, produced when the masking signal is more salient than the target signal at some point in time (2). In this sense, informational masking could describe distraction due to speech or non-speech maskers. Informational masking has also been used to describe a situation of stimulus uncertainty where portions of the masker speech are substituted for portions of the target speech (3). The increase in listening effort that accompanies listening to speech in noise cannot be directly observed, but pupil dilation has been shown to be a useful involuntary correlate of listening effort (4).

The work reported here was performed in the context of the project S3A: Future Spatial Audio for an Immersive Listener Experience at Home, in which an object-based audio system is used to deliver broadcast audio content (5). A significant feature of object-based systems is that the scene is assembled at the listener end instead of the broadcaster transmitting a fixed mix. This means that an object-based system can make adjustments for individual listener needs, such as trying to optimize speech intelligibility by turning up the main dialogue and turning down the background music. The personalization could include monitoring intelligibility and adapting to the hearing loss profile of individual listeners. One step beyond this would be to monitor the listening effort of the listener and adjust the sound scene.

In the experiment reported here we set out to capture changes in effort as people listened to speech

[1] w.davies@salford.ac.uk

in various maskers. We attempted to discriminate between energetic and informational masking by including competing speech and speech-shaped noise as two of our masker signals.

2. METHOD

2.1 Design

The target speech was taken from the Hurricane natural speech corpus (https://doi.org/10.7488/ds/2482) (6). The recordings are of a single male native British-English talker. We used the Harvard sentences in the quiet condition. Each sentence in the corpus has an average of 8.0 words (s.d. 1.2). Four maskers were used. We attempted to represent the broad range of sounds in our audio application by selecting one sound from each of the top-level categories of a soundscape taxonomy: people (speech), nature and manmade (7). Two different speech maskers were used. For competing speech (CS) we used a single female British-English talker introducing the BBC Radio 4 Woman's Hour podcast. For speech-modulated noise (SMN), we used a modulated random noise signal with the same long-term spectrum and envelope as the competing speech masker. The SMN was thus similar to the CS but without the informational content. For a nature sound, we used a recording of a dawn chorus (https://freesound.org/people/tim.kahn/sounds/395221/). For a manmade sound (Urban) we used traffic on a street, made by combining two recordings from the Adobe Audition sound effects library (https://adobe.ly/2rDij3Z): "Ambience Busy Road 01.wav" and "Ambience Busy Traffic 01.wav".

The presentation level of the speech was set to 60 dB(A). This was done by: (i) generating a noise signal with the same long term frequency spectrum as the speech corpus; (ii) playing this noise signal at -43 dBFS; (iii) adjusting the output gain of the loudspeaker to generate $L_{Aeq,slow}$ = 60 dB(A) at the listening position. This was measured using a Type I sound level meter. Each of the speech signals were then normalised to -43 dBFS. Four levels of energetic masking were generated using Tang et al.'s *Distortion Weighted Glimpse Proportion* (DWGP) metric (8, 9). The four DWPG levels were: 0.3, 0.45, 0.6 and 0.75. These DWGP values were set by iteratively changing the gain of the masker until the DWPG was within 0.005 of the target value. The speech-noise stimuli were 10 seconds long. The speech started 4 seconds after the onset of the masker.

Experiments were conducted in the University of Salford audio booth. Light levels were kept constant across all participants. Stimuli were reproduced from a Genelec 8030A loudspeaker positioned 1.35m in front of the listener. Participants were equipped with a Pupil Labs binocular eye tracker (https://pupil-labs.com/pupil/).

2.2 Participants

Eighteen native British English speakers (mean age 24, s.d. 6) were recruited as paid participants. All participants stated that they had normal hearing at the time of the experiment.

2.3 Task

Participants were instructed to attend to the male talker in all conditions. Participants were exposed to all levels of masker (CS, SMN, Nature, Urban) and DWGP (0.3, 0.4, 0.6, 0.75) in a full factorial design. The order of conditions presented to participants was randomized. Six sentences were used for each Masker*DWGP condition meaning each participant evaluated 4*4*6 = 96 sentences. The sentences for each listener were selected randomly from the corpus of 720 sentences and there were no repeat sentences for any listener.

The stimuli were delivered using a Matlab interface on a Toshiba Portege 13.3” laptop. The interface consisted of a fixation point which the participants were instructed to look at while the stimuli were playing. After the end of each stimulus a text box appeared. Participants were instructed to type any words they heard using the laptop keyboard.

3. RESULTS AND ANALYSIS

3.1 Behavioural data

The behavioural results are given as word recognition rate, calculated for each sentence by calculating the number of correctly identified words. Figures 1 and 2 show the mean WRR with 95% confidence intervals as a function of DWGP and masker. As expected there is a monotonic relationship between DWGP and WRR (8). However, there also appear to be differences between the different masker types in Fig. 2.

A repeated-measures ANOVA using within-subject factors of DWGP and Masker shows significant main effects of the two factors along with an interaction effect ($p < 0.001$ for all effects). For DWGP the effect size, characterised by generalized eta-squared (η_G^2) was 0.75 indicating a large effect size according to Bakeman (10). For Masker $\eta_G^2 = 0.15$ indicating a small effect size. For the DWGP: Masker interaction $\eta_G^2 = 0.12$ indicating a small effect size.

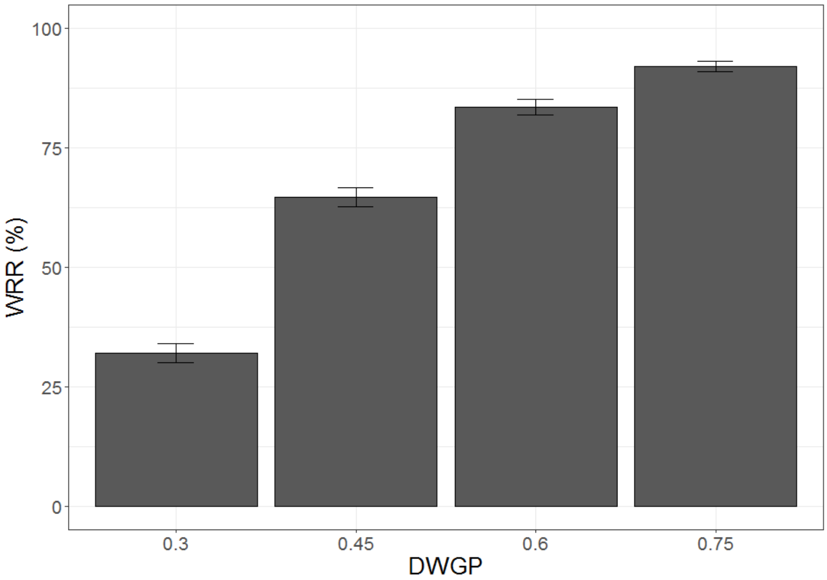

Figure 1 – Word recognition rate vs Distortion-Weighted Glimpse Proportion.

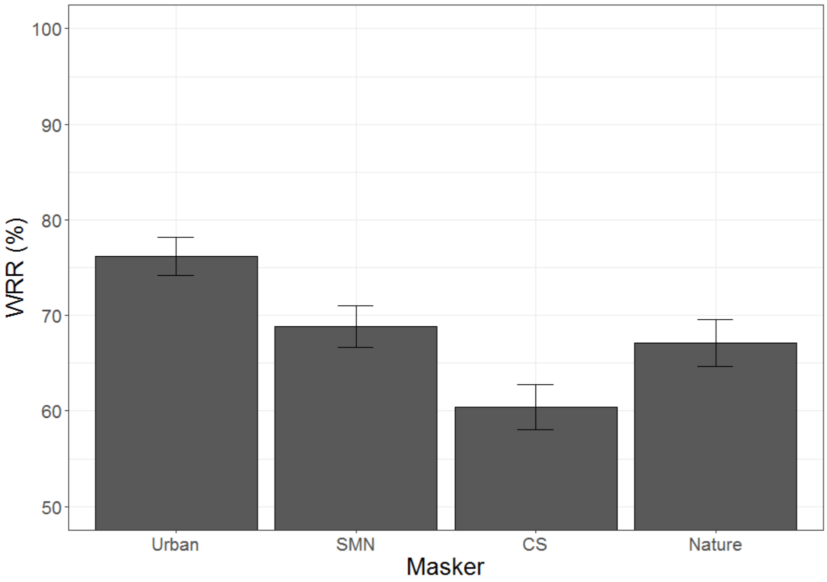

Figure 2 – Word recognition rate vs masker.

3.2 Physiological data

The physiological data were treated in a fairly conservative manner by smoothing and by treating blinks. The Pupil Labs software produces a confidence percentage that can be used to estimate where data have been corrupted by blinking or other problems. First, any trials where more than 25% of the data were less than 50% confidence were rejected. Next, the pupil responses were downsampled to 50Hz and a 0.2 second moving average filter was applied. This typically produced smooth responses corrupted by occasional blinks. The confidence data was then used to estimate where blinks occurred and a linear interpolation applied to bridge any period where confidence dropped below 50%. The response for each trial was then normalised by the average pupil diameter 0.2 seconds before the start of the trial. The mean and peak pupil dilation were then calculated for each sentence. Missing data were imputed using predictive mean matching via the R package MICE (11).

Figure 3 shows time histories of the normalised pupil response at each DWGP level. t = 0 marks the onset of the sentence. The solid lines show the mean over all participants and masker types, while the shaded areas indicate 95% confidence intervals. Pupil dilation varies considerably over the course of the sentence. Lower DWGP (higher noise level) seems to be associated with larger diameter, though the four noise levels are not cleanly separated over the whole sentence.

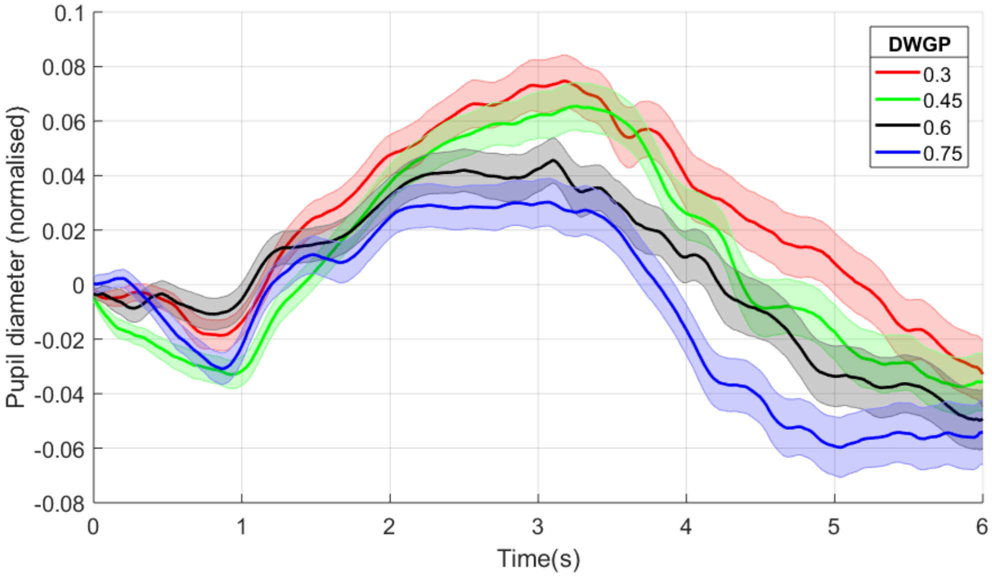

Figure 3 – Pupil diameter time history for each DWGP level.

To examine the effect of DWGP and masker type on pupil diameter, the time histories shown in Fig. 3 have been converted to a mean pupil dilation (taken over the time period in Fig. 3). Mean diameters with 95% confidence intervals are shown as a function of DWGP in Fig. 4 and as a function of Masker in Fig. 5.

A repeated-measures ANOVA using within-subject factors of DWGP and Masker shows significant main effects of DWGP and Masker ($p < 0.05$). There is no significant interaction effect ($p = 0.46$). For the effect of DWGP on mean diameter, $\eta_G^2 = 0.02$ indicating a small effect size. For the effect of Masker on mean diameter, $\eta_G^2 = 0.01$ indicating a small effect size. Post-hoc pairwise t-tests with Bonferroni corrections indicate that there are significant differences between the DWGP levels 0.3-0.6 ($p < 0.05$) and 0.3-0.75 ($p < 0.001$). Post-hoc pairwise t-tests with Bonferroni corrections indicate that there are significant differences between the Masker levels CS-Urban ($p < 0.001$) and a tendency toward a difference between CS-Nature ($p = 0.06$).

An alternative single-figure pupil variable is normalised peak diameter. When the dependent variable is taken as the peak pupil dilation, DWGP and Masker are both significant ($p < 0.05$ and $p < 0.001$ respectively). The effect size for DWGP is much smaller than that for Masker when peak pupil dilation is used as the dependent variable ($\eta_G^2 = 0.006$ and $\eta_G^2 = 0.01$ respectively).

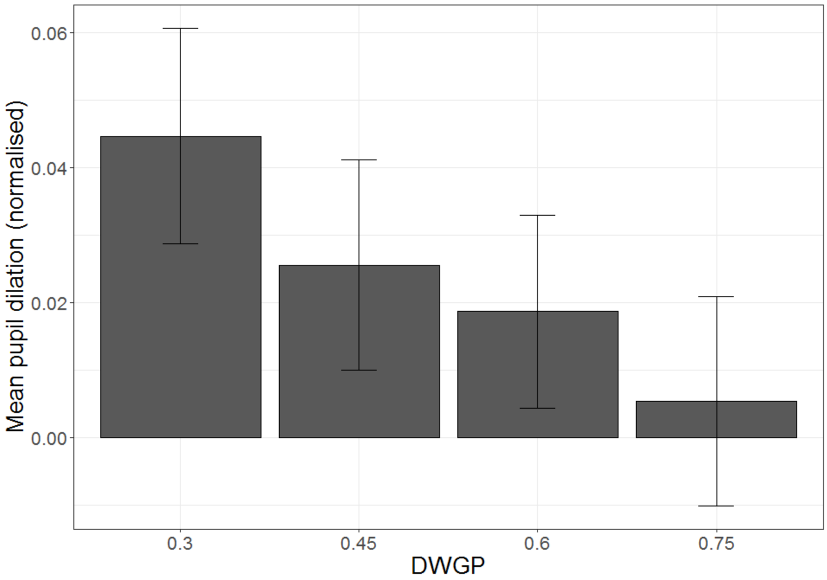

Figure 4 – Mean pupil diameter vs DWGP.

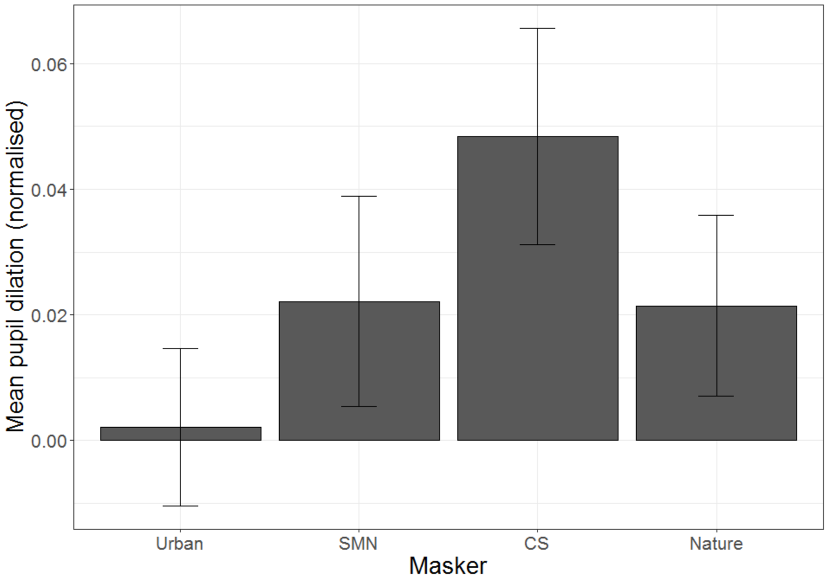

Figure 5 – Mean pupil diameter vs masker.

4. CONCLUSIONS

The results are consistent with listening effort varying with DWGP and masker type, and also over the course of each sentence. Changing DWGP demonstrates the expected energetic masking effect and this is reflected in pupil dilation as well as word recognition rate. Changing the masker seems to additionally show an informational masking effect, on both behavioural and physiological responses. The competing speech masker is associated with the lowest word recognition rate and the largest pupil diameter. Both mean and peak pupil dilation seem to reveal changes in listening effort due to both energetic and informational masking

However, while the behavioural measure (WRR) showed a large effect size for both DWGP and masker type, the effect size observed for the physiological measures was small. Further work is needed to understand the most appropriate analysis method for the pupillometry data. Possible avenues include the effect of different baseline durations and using pupil deconvolution to improve temporal resolution (12, 13).

ACKNOWLEDGEMENTS

The authors would like to thank Lara Harris for running the experiments, Yan Tang for providing valuable input to the experiment design and the code to run the DWGP model, and the experiment participants for their time. This work was supported by the EPSRC program Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and by the BBC Audio Research Partnership.

The experimental data underlying the findings are fully available without restriction. Details are available from http://dx.doi.org/10.17866/rd.salford.7931465

REFERENCES

1. Brungart DS. Informational and energetic masking effects in the perception of two simultaneous talkers. J Acoust Soc Am. 2001; 109(3):1101-9.
2. Botteldooren D, De Coensel B. Informational masking and attention focussing on environmental sound. NAG/DAGA 2009 International Conference on Acoustics; 23-26 March; Rotterdam, Netherlands; 2009.
3. Durlach NI, Mason CR, Kidd Jr G, Arbogast TL, Colburn HS, Shinn-Cunningham BG. Note on informational masking (L). J Acoust Soc Am. 2003; 113(6):2984-7.
4. Naylor G, Koelewijn T, Zekveld AA, Kramer SE. The Application of Pupillometry in Hearing Science to Assess Listening Effort. Trends in Hearing. 2018; 22:1-3.
5. Coleman P, Franck A, Francombe J, Liu Q, de Campos T, Hughes RJ, et al. An audio-visual system for object-based audio: From recording to listening. IEEE Transactions on Multimedia. 2018; 20(8):1919-31.
6. Valentini-Botinhao C, Mayo C, Cooke M. Hurricane natural speech corpus - higher quality version [dataset]. LISTA Consortium: (i) Language and Speech Laboratory UdPV, Spain and Ikerbasque, Spain; (ii) Centre for Speech Technology Research, University of Edinburgh, UK; (iii) KTH Royal Institute of Technology, Sweden; (iv) Institute of Computer Science, FORTH, Greece. ; 2019. Available from: https://doi.org/10.7488/ds/2482.
7. Bones O, Cox TJ, Davies WJ. Sound Categories: Category Formation and Evidence-Based Taxonomies. Frontiers in Psychology. 2018; 9:1277.
8. Tang Y, Cooke M, Fazenda BM, Cox TJ. A glimpse-based approach for predicting binaural intelligibility with single and multiple maskers in anechoic conditions. Sixteenth Annual Conference of the International Speech Communication Association; 6 - 10 September; Dresden, Germany; 2015.
9. Tang Y, Hughes RJ, Fazenda BM, Cox TJ. Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms. Speech Communication. 2016; 82:26-37.
10. Bakeman R. Recommended effect size statistics for repeated measures designs. Behavior research methods. 2005; 37(3):379-84.
11. Buuren Sv, Groothuis-Oudshoorn K. mice: Multivariate imputation by chained equations in R. Journal of statistical software. 2010:1-68.
12. Wierda SM, van Rijn H, Taatgen NA, Martens S. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. Proceedings of the National Academy of Sciences. 2012; 109(22):8456-60.
13. McCloy DR, Larson ED, Lau B, Lee AK. Temporal alignment of pupillary response with stimulus events via deconvolution. J Acoust Soc Am. 2016; 139(3):EL57-EL62.