Perceived quality of speech degraded by wind noise: An assessment of sources of variability in a Web experiment

Iain R. Jackson, Paul Kendrick, Trevor J. Cox, Bruno M. Fazenda, Francis F. Li

Acoustics Research Centre, University of Salford, Manchester, UK

i.r.jackson@salford.ac.uk, p.kendrick@salford.ac.uk, t.j.cox@salford.ac.uk,
b.m.fazenda@salford.ac.uk, f.f.li@salford.ac.uk

Abstract

An online speech-in-noise task was used to assess the effect of wind noise degradation on perceptions of audio quality. Many researchers in experimental psychology are increasingly confident of the potential of web-based tasks of perception and cognition. Investigations of the perception of audio quality, however, require consideration of additional factors (e.g. different reproduction systems) which may contribute to variability. Here we report analyses of potential sources of variability, in self-reported categories by participants pre-test. These categories are participant age, audio expertise, method of audio reproduction and level of background noise in the testing environment. Some effects on quality ratings were found but, overall, effect sizes suggest these were small relative to the effect of the main experimental variable of wind noise.

Index Terms: quality perception, audio quality, quality assessment, web experiments, wind noise

1. Introduction

Over the past decade or more the use of crowdsourced data from web-based experiments has become increasingly common in psychology, particularly in studies of perception and cognition. The prospect of obtaining highly generalisable data quickly, cheaply, and on a mass scale is highly appealing to many to researchers. For experiments requiring critical audition of samples however the lack of experimental control in playback conditions can present a number of significant challenges [1]. Nevertheless, in recent years a number of web-based audio experiments have begun to emerge, in areas as diverse as psychoacoustics [2], musicology [3], and speech perception [4][5].

In a recent experiment we investigated the effect on perceived quality when audio samples of speech were degraded by the addition of wind noise [6]. Wind causes pressure fluctuations at the microphone diaphragm leading to the presence of unwanted noise in recordings. Although this is a common problem experienced by audio recordists outdoors, the exact nature of the effect on quality perceptions has not previously been investigated.

Two physical characteristics of wind noise were identified from field recordings; the level (mean A-weighted sound pressure level), and the temporal variability (or 'gustiness', defined as the mean absolute difference between the level over the whole sample and the level in a moving 1 second window). Systematic manipulation of these two properties allowed us to generate a database of 5 second audio samples of simulated wind-induced noise [7]. These samples were then paired with samples of nonsense speech (speech level was set to 58dBA for each sample). On each trial, participants were asked to type the sentence they had heard and to provide judgments of the overall quality of the audio. As speech and wind noise commonly occur together in real-world scenarios (outside broadcasts, mobile phone conversations, etc) the pairing of the two in this test was intended to provide useful data on a common audio problem while engaging participants in a familiar task.

The same experiment was carried out both in the lab and online, allowing for a comparison of the two methods. In the lab, the level of wind noise was found to account for a very large proportion of variance in quality ratings. Broadly, once above a certain threshold, participants were highly sensitive to change in level of wind noise; ratings of quality were found to significantly decrease with each successive increase on our scale of wind noise level. In contrast, the level of gustiness was not found to have a significant effect on quality.

A similar finding for the effect of wind level was found in the Web version of the experiment. Additionally however, unlike the lab data, a significant interaction of level and gustiness was also observed (at higher levels of wind noise samples with greater temporal variability were rated as being of higher quality than samples in which wind noise was more constant). Quality ratings for each version of the test, lab and Web, are presented in Figure 1 and Figure 2.

The replication on the Web of the main finding of the lab experiment (plus additional significant effects not found in the lab data) lends support to a growing number of researchers within psychology who argue that Web testing can be a viable complimentary method to lab testing (and perhaps even preferable in some circumstances) [8][9][10].

However, there are important differences between conventional psychological Web experiments, designed to assess some facet of behaviour or cognition, and the goals of many psychoacoustic-type experiments, which might aim to quantify one or more subjective or aesthetic dimensions of audio (such as quality, for instance). An experiment which measures reaction times to the onset of a stimulus, for example, is likely to be less vulnerable to variability in the reproduction of audio stimuli than one in which participants are required to make subjective ratings or decisions of preference.

In the current paper we explore some of the features which may influence judgments of quality in audio Web testing. Analyses for the speech-in-noise experiment described above were conducted across the whole sample (n = 5110 data points). Here, we attempt to improve our understanding of the relative importance of uncontrolled experimental factors in that dataset by breaking it down by additional factors reported by participants about themselves and the circumstances in which the test was completed.

10.21437/PQS.2013-8

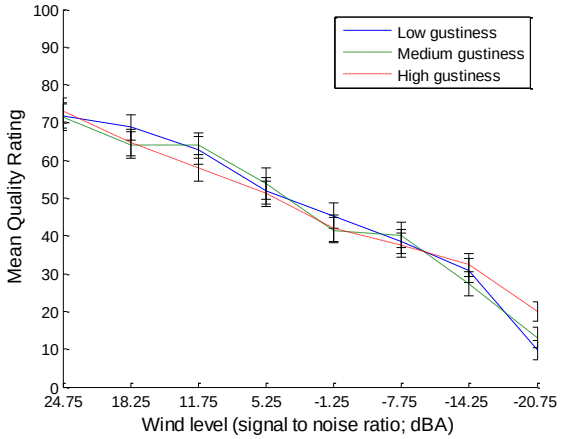

Figure 1: *Mean ratings of audio quality in the Lab test, by level of gustiness. Error bars show standard error of the mean. Units given for wind level represent the mid-points of each of the 8 consecutive windows from which the samples were drawn.*

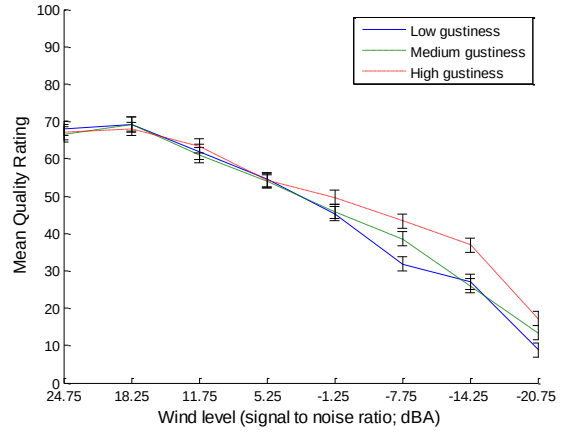

Figure 2: *Mean ratings of audio quality in the Web test, by level of gustiness. Error bars show standard error of the mean. Units given for wind level represent the mid-points of each of the 8 consecutive windows from which the samples were drawn.*

2. Method

Participants were informed that the experiment aimed to investigate how people perceived speech in noisy conditions and how this relates to audio quality. They were also informed that they would have the opportunity to enter a prize draw (for one of four £10 vouchers for Amazon) if they wished to upon completion of the task. Wind noise was not referred to anywhere in the instructions pre-test but was included in an explanation of the purpose of the experiment upon completion. After consenting to take part participants were presented with a looped example of a sample they would hear in the test (not included in test set) and advised to adjust their playback volume so they could hear the sound playing clearly. Prior to starting the test participants were required to select a response for each of 4 categories designed to gather extra information about them and the circumstances they were completing the test in. Participants were presented with the following questions and response options;

i) *I am listening to this experiment using…*
- headphones
- laptop/tablet/mobile/internal loudspeakers
- external loudspeakers
- don't know/other

ii) *The place where I am doing this experiment is…*
- very quiet
- quiet
- noisy
- very noisy

iii) *My age is…*
- (10 year windows from "0-9" to "80 or over".)

iv) *I think I am an expert in audio…*
- yes
- no
- don't know

Each response type was selected by radio button. Participants were required to submit a response for each statement before they could proceed. Once complete, participants were then presented with instructions for completing the test and given two practice trials. Samples used in the practice trials were (non-test-set) examples of the best and worst quality audio which would be heard in the test.

On each trial, up to a maximum of 25 trials, participants were asked to listen to a sample containing a unique nonsense sentence paired with one permutation of wind noise, then to type the sentence they heard and rate the overall audio quality of the sample. The sample could be played only once before submitting the typed response but as many times as desired afterwards. Quality ratings were obtained through user-controlled sliders (whose outputted values ranged from 0 to 100), labeled with "Bad" and "Excellent" at opposite ends of the scale. Playback of the samples and rate of progression through the test were determined by participants. Presentation order of samples and pairings of sentences with permutations of wind noise was fully randomized within test sets.

3. Results

All significant effects reported at $p<.05$ (with Bonferroni adjustments for multiple comparisons). Categories of participant information and their levels which met inclusion criteria for analyses are shown in Table 1. For each category only those levels with at least 100 data points were included in analyses. Each category was analysed in a three-way ANOVA with wind level and gustiness level, with quality rating as the dependent variable.

Analyses revealed:

- No effect of audio expertise on quality ratings.
- Quality ratings of those completing the test in Very Quiet locations were significantly higher than those in either Quiet or Noisy locations. Similarly, ratings in

locations considered Quiet were significantly higher than those in Noisy locations.

- Method of audio reproduction was found to significantly interact with level of wind noise; at low levels of wind noise the highest quality ratings were observed in the group listening with external speakers, and ratings from headphones users were significantly higher than devices in the remaining category. These differences rapidly disappear however as levels of wind noise increase – quality ratings converge across device types above a signal to noise ratio of around 12 dBA.
- There were no differences in quality ratings between participants in different age categories, other than for those aged 30-39 years, whose ratings were significantly higher than those in the 10-19 year category.
- To further explore the finding relating to differences in age groups an additional analysis was conducted factoring in reproduction equipment alongside age. A significant interaction was observed; participants in the 30-39 age group using external speakers gave consistently higher quality ratings than users of other devices.

Table 1. *Number of data points for each level included in analyses of categories of participant information.*

Category	Level	Data points
Reproduction of Audio	Headphones	3344
	Laptop/mobile/tablet/ internal speakers	1286
	External loudspeakers	464
Audio Expertise	Expert	496
	Non-Expert	2791
	Don't Know	1823
Environment	Very Quiet	1325
	Quiet	2947
	Noisy	767
Age	10-19 years	2081
	20-29 years	1797
	30-39 years	356

Importantly, while significant differences were observed for many of the categories of participant information it is notable that effect sizes were small across the board (all effect sizes < .017, partial η^2), relative to the effect size of the experimental wind level variable across the group (partial $\eta^2 = .30$) [11].

4. Discussion

In earlier work it was demonstrated that perceptions of quality degradation with wind noise were comparable whether the test was completed on the Web or in the lab, albeit with smaller effect sizes on the Web [6]. The analyses in that work proceeded at the level of the whole sample, ignoring potential sources of systematic variability within the sample. In this paper we have refined those analyses by breaking down the Web participants' data according to categories of additional information reported by participants about themselves and the conditions the test was completed in, and explored how these factors might influence perceptions of quality in the test.

The largest effect on quality of the four factors under investigation was the influence of noisiness of the test environment. Only 15% of participants reported completing the test in environments which were not Very Quiet or Quiet, but quality ratings were found to significantly decrease with each successive increase of self-reported noise category.

Expertise was not found to have a significant effect on ratings of quality. This is perhaps surprising when previous comparisons of experts and non-experts in the lab have found that the absolute ratings of experts tend to be lower and less variable than those of less experienced listeners (although the rank order of quality evaluations rarely differs between the two groups) [12]. It should be noted here however that nearly four times as many participants in our sample responded to the statement "I think I am an expert in audio…" with the suggestion that they did not know either way (36%) than those who did consider themselves to be experts (10%). For future studies where investigation of differences between experts and non-experts is important it would seem advisable to consider clarifying the question and/or response options with concrete examples of what would constitute an expert.

The equipment participants used to listen to the samples in the test was found to have an effect on quality ratings, but only in those samples with relatively little degradation by wind noise. In these samples the highest quality ratings were associated with participants using external loudspeakers, the lowest ratings were found in the group who completed the test using laptop, mobile, tablet, or internal speakers.

The final category to be considered was the participants' age groups. Participants in the 30-39 age group were found to give significantly higher ratings than those in the 10-19 age group. Our analyses of these categories were largely exploratory, as opposed to hypothesis driven, and it is possible that participants within the 10-19 years category are so diverse across a range of developmental and psychological traits as to preclude meaningful explanations for this particular finding. This said, the related finding that ratings associated with external speakers were consistently significantly higher in the 30-39 age group (but not in other age groups) hints towards a more complex interplay of factors. We could speculate for instance that this age group had access to higher quality equipment than younger listeners, and that this is reflected in their ratings scores.

To conclude, we observed that - for this experiment at least – the lack of experimental control over participant age, method of audio reproduction, environment and audio expertise did not present a significant hindrance to subjective testing of audio quality on the Web. Some significant differences in quality ratings were found between subsets of participants and scenarios but comparison of the magnitude of these effects relative to the overall effect size of the experimental variables suggests their influence was minor.

In other experiments however where levels of degradation in test samples are less severe, or differences between samples are more subtle, the relative contribution of the external factors on ratings of quality will become considerably more important.

Indeed, should the effect of these external factors begin to approach the magnitude of the effect of the experimental variable(s), the advantages and validity of the web-based approach would be nullified.

5. Acknowledgements

This work was funded by the Engineering and Physical Sciences Research Council (EP/J013013/1).

6. References

[1] Kendall, R., "Commentary on 'The potential of the internet for music perception research: A comment on lab-based versus Web-based studies' by Honing & Ladinig", Empirical Musicology Review, 3:8-10, 2008.

[2] Cox, T.J., "The effect of visual stimuli on the horribleness of awful sounds", Applied Acoustics, 69:691-703, 2008.

[3] Honing, H., "Evidence for tempo-specific timing in music using a web-based experimental setup", J Exp Psychol Hum Percept Perform, 32:780-6, 2006.

[4] Ribeiro, F., Florencio, D., Zhang, Cha, and Seltzer, M. "CrowdMOS: An approach for crowdsourcing mean opinion score studies", Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2011.

[5] Cooke, M., Barker, J., and Lecumberri, "Crowdsourcing in Speech Perception", in Crowdsourcing in Language and Speech, by Maxine Eskenazi (Ed.), Wiley, In Press.

[6] Jackson, I. R., Kendrick, P., Cox, T. J., Fazenda, B. M., and Li, F. F., "Perceptual evaluation of the functional and aesthetic degradation of speech by wind induced noise during recording", Proc. Meetings on Acoustics, 19:060170, [doi:10.1121/1.4799221], 2013.

[7] Kendrick, P., Cox, T. J., Li, F. F., Jackson, I. R., and Fazenda, B. M., "Wind-induced microphone noise detection – Automatically monitoring the audio quality of field recordings", IEEE Proc. Multimedia and Expo. (ICME), 2013. (accepted)

[8] Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., and Wilmer, J. B., "Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments", Psychon. Bull. Rev., 19(5):847-57, 2012.

[9] Crump, M. J. C., McDonnell, J. V., and Gureckis, T. M., "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioural research.", PLoS ONE, 18(5): e57410. doi:10.1371/journal.pone.0057410, 2013.

[10] Miller, G., "The Smartphone Psychology Manifesto", Perspect. Psychol. Sci., 7(3), 221-237, 2012.

[11] Cohen, J. "Eta-squared and partial eta-squared in fixed factor ANOVA.", Educ. Psychol. Meas., 33(1):107-112, 1973.

[12] Schinkel-Bielefeld, N., Lotze, N., and Nagel, F., "Audio quality evaluation by experienced and inexperienced listeners.", Proc. Meetings on Acoustics, 19:060016, [doi:10.1121/1.4799190], 2013.