Crowdsourcing for word recognition in noise

Martin Cooke[1,2]*, Jon Barker*[3]*, Maria Luisa Garcia Lecumberri*[2]*, Krzysztof Wasilewski*[3]

[1]Ikerbasque (Basque Science Foundation)
[2]Language and Speech Laboratory, Universidad del País Vasco, Spain
[3]Department of Computer Science, University of Sheffield, UK

m.cooke@ikerbasque.org

Abstract

Access to large samples of listeners is an appealing prospect for speech perception researchers, but lack of control over key factors such as listeners' linguistic backgrounds and quality of stimulus delivery is a formidable barrier to the application of crowdsourcing. We describe the outcome of a web-based listening experiment designed to discover consistent confusions amongst words presented in noise, alongside an identical task carried out using traditional laboratory methods. Web listeners were graded according based on information they provided as well as via their responses to tokens recognised robustly by a majority of participants. While overall word identification scores even for the best-performing web subset were well below those obtained in the laboratory, word confusions with high levels of cross-listener agreement were obtained nevertheless, suggesting that focused application of crowdsourcing in speech perception can provide useful data for scientific analysis.
Index Terms: speech perception, noise, web experiment

1. Introduction

Traditional studies in human speech perception have focused on mean response characteristics across sets of items, to the extent that objective 'macroscopic' models now exist which are capable of making good predictions of average intelligibility in different types of noise [1, 2]. However, far less is known about the detailed processing of individual tokens such as words in the auditory system, and we do not yet possess 'microscopic' models of speech perception with the ability to make robust estimates of responses to individual items. One factor impeding the development of such models is a lack of examples of robust confusions i.e. incorrect responses which have a high level of inter-listener agreement. Constructing a large corpus of robust confusions is not a straightforward task since the rate at which such confusions occur is relatively low [3] and by definition a large number of listeners is required to determine consistency. The current paper describes an appeal to 'crowdsourcing' – web-enabled citizen science – to screen potential confusions in sufficient quantities with large numbers of listeners.

Members of the public have been contributing valuable data points in science for well over a century (e.g. [4]), but the Web has created the possibility of very large scale participation of individuals in scientific projects (e.g. [5]). Speech researchers are starting to recognise the potential of outsourcing tasks such as evaluations, transcription and perception tests to a wider community of workers. However, experiments involving speech present special challenges such as variability in audio hardware and listeners' hearing thresholds, as well as lack of homogeneity of linguistic experience of participants. Prior to wider application of web-based experiments involving speech stimuli, measurements of the reliability of web-derived data are needed.

The current study aimed to compare the outcomes of traditional and web-based perception tests in which listeners identify words in noise, to relate subjective (participant-provided) information and objective measures to overall scores, and to assess the scientific value of the information obtained in web experiments in terms of the discovery of consistent patterns of listener confusions.

2. Web-based studies in audio

The growing use of the web for experiments involving audio is evidenced by a burgeoning interest among musicologists [6, 7], audiologists [8] and speech researchers [9, 10, 11]. Web-based studies, however, are not without their problems [12]. One commonly-cited issue is the comparative lack of experimental control. While laboratory listening tests are typically conducted in sound-attenuating rooms with state-of-the-art equipment for audio reproduction, web-based tests may take place in non-quiet rooms using uncalibrated headphones of unpredictable quality.

A second concern is the 'trustworthiness' of responses [13]. How can it be guaranteed that the subjects are providing meaningful responses? This question arises because participants in web-based experiments are less controlled than those in a formal laboratory experiment. However, it has been pointed out by proponents of web-experimentation that this question applies generally to all behavioural testing whether web- or lab-based and further that subjects in a web-based experiment may generally have less motivation to give deceptive responses: if they are taking time to complete the experiment it will be because they have a genuine interest (they are 'highly' voluntary) rather than because they have been drawn in by the promise of financial reward and are being coerced to complete the session while being monitored by an experimenter [14]. On the other hand, crowdsourcing technologies such as Amazon Mechanical Turk [15] do involve financial gain.

High drop-out rate is often cited as a further problem, but ironically it is the ease with which subjects *can* drop out that ensures that subjects completing the experiment will generally be well-motivated and contributing good quality data. Nevertheless, precautionary checks and measures are needed to screen out untrustworthy data (e.g. by monitoring time stamps of responses or inserting dummy trials which have predictable responses). Good web-experiment design can also motivate participants and minimise drop out rates [12, 16].

3. Web experiment

Listeners heard monosyllabic English words presented in 12 different types of noise at a range of signal-to-noise ratios. 613 words were chosen from an existing list [17] based on a set of criteria designed to maximise the potential for confusability, principally by selecting words with a high phonological neighbourhood density. Four male and one female native British English speakers reproduced the words in isolation. More details of stimuli and the formal listening test are provided in [3].

3.1. Web interface

The Web face of the application is a Java Applet which communicates with a back-end database engine whose role is to store user details and responses and to upload sets of speech-in-noise stimuli. The interface was designed to permit participants to complete the test in less than 3 minutes. Participants operate throughout from a single web page containing a description of the scientific motivation for the test and a small number of familiarisation stimuli which also provided an indirect reminder to set the volume. Listeners fill in a short questionnaire (figure 1, top). To eliminate network delays during the test itself, the applet uploads a complete block of 50 stimuli while participants are reading about the task. After providing consent, listeners start the test, entering their guess after each item presentation (figure 1, bottom). To engage the participant and familiarise them with the target voice (which is the same throughout the block of stimuli), the noise level ramps linearly from +30 dB to a level around 0 dB for the first 5 stimuli, after which the level increases more slowly. On completion, participants receive feedback in the form of their ranking amongst all listeners who have heard the same test block. Participants are then able to hear further blocks of stimuli should they wish to.

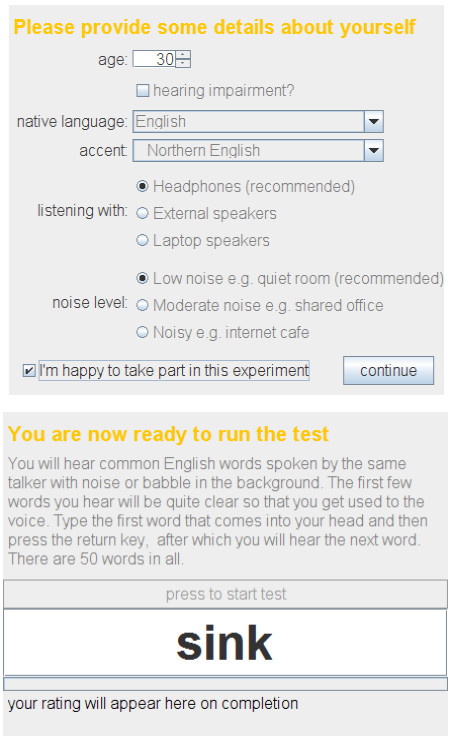

Figure 1: *Web interface.*

3.2. Respondents

Two adverts placed 11 days apart via the University of Sheffield's internal announcement service (which has the potential to reach more than 20 000 staff and students) led to 2120 respondents completing the task within 20 days of the first advert. On average, participants completed 1.48 blocks in 155 seconds per block. In total, 157 150 individual noisy tokens were presented. Here we analyse 77 400 individual responses from 903 listeners, based on their responses to tokens spoken by one of the male speakers in 12 noise conditions, since these were common to the formal listening tests described in [3].

Figure 2 shows univariate mean scores for each level of the factors gathered from participants, permitting a cross-factor comparison of effect sizes. Ambient noise in the test environment had a large effect, as did having a first language (L1) other than English. More surprisingly, the performance of listeners having as their L1 a variety of English other than British English (NonBrEng) was substantially lower than the level obtained by native British English speakers (BrEng). Predictably, older listeners fared less well than younger, and similarly users with headphones outperformed those relying on internal or external speakers. Listeners who reported hearing impairment showed relatively little degradation. However, this data should be interpreted with caution since different numbers of listeners contributed to each factor level (see caption).

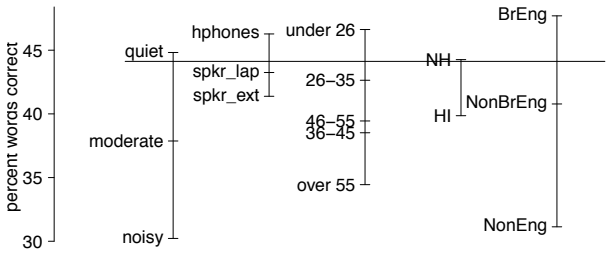

Figure 2: *Mean word identification scores as a function of participant-related factors. Response percentages (i) environment: quiet 91, moderate 8, noisy 1; (ii) hardware: headphones 50, laptop speakers 17, external speakers 34; (iii) age range: under 26 (52), 26-35 (36), 36-45 (10), 46-55 (1), over 55 (2); (iv) normal hearing 97, hearing impaired 3; (v) language/accent: BrEng 75, NonBrEng 5, NonEng 19.*

3.3. Listener group word identification scores

Here we explore the performance of selected subsets of listeners based on both subjective and objective criteria, and compare their scores to listeners tested under formal conditions [3].

The subjective approach uses information supplied by participants to define a subset well-matched to those undertaking the formal test. Here, we examine a subjectively-defined subset ('subj'), consisting of the 31% of web listeners who reported all of the following: listening in a quiet environment over headphones, aged 50 or under with no known hearing problems, and with a British variety of English as their first language.

Participants who score highly on 'anchor tokens' – items which have a very high rate of correct identification across listeners – are likely to be highly-motivated. Anchors were defined as those tokens heard by at least 30 listeners and which resulted in scores of at least 80% correct. Participants who achieved mean scores of at least 90% across anchor tokens constituted an

objectively-defined ‘anchor’ subset (63% of all web listeners).

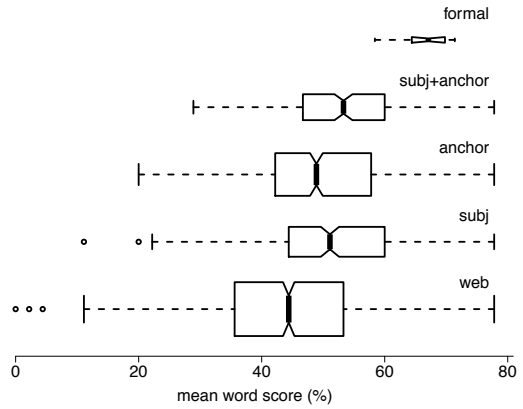

Figure 3: *Boxplots of scores for formal and web groups. Lines extend to 1.5 times the inter-quartile range, circles indicate outliers, box thickness is proportional to the number of listeners in group and notches depict 95% confidence intervals.*

Figure 3 depicts word score summaries for participants in the formal and web tests, including the 23% of listeners who met both subjective and objective criteria (‘subj+anchor’). The group tested under formal conditions far outperformed any of the web subsets. Scores in the ‘subj’ subset were higher than the web average, but in spite of their satisfying strict participant-supplied criteria, this group still contained poor-performing outliers. Application of the anchor tokens constraint had the effect of removing these outliers, while the combination of subjective and anchor criteria resulted in higher scores than either alone. Even so, a 13 percentage points gap exists between the traditionally-tested and best web subset. In subsequent analyses the responses of the formal group are compared with the best-performing web subset (subj+anchor) and its complement. These groups are denoted web+ and web- for brevity.

3.4. Response correlations

Figure 4 compares scores for the formal group with those obtained by each of the web+ and web- groups. Each point represents a single SNR level for one of the 12 maskers. Recall that tokens in block of stimuli were presented with decreasing SNRs. To obtain the points in figure 4, SNRs were rounded to the nearest integer. Each point is therefore based on a subset of words. The strong correlations which exist between formal and web scores suggest that both the varying difficulty in identifying word subsets at a given SNR as well as the challenge produced by each of the masker types affected groups to a similar degree.

3.5. Response consistency

Figure 5 shows the numbers of words as a function of listener agreement for both correct (upper panel) and incorrect (lower) responses. For both correct and incorrect responses, greater consistency (i.e. more tokens with a high level of agreement) is observed for the formal group, while the web+ subset is more consistent than the web- subset.

The formal group discovered 129 majority confusions (defined as those responses with agreement $\geq 50\%$) compared to 85 and 44 for the web+ and web- groups. This suggests that although the web-based procedure leads to lower overall scores,

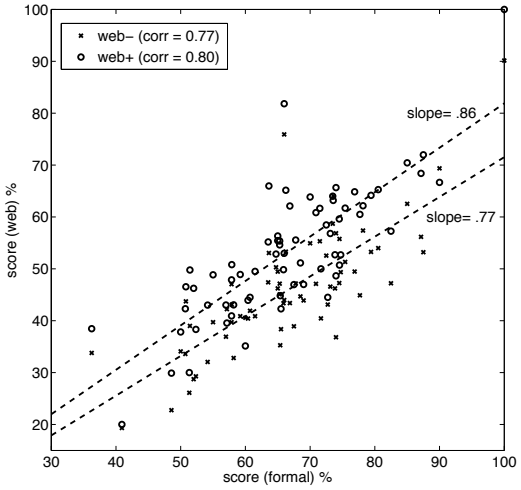

Figure 4: *Mean scores in each masker and SNR condition for the formal and web groups.*

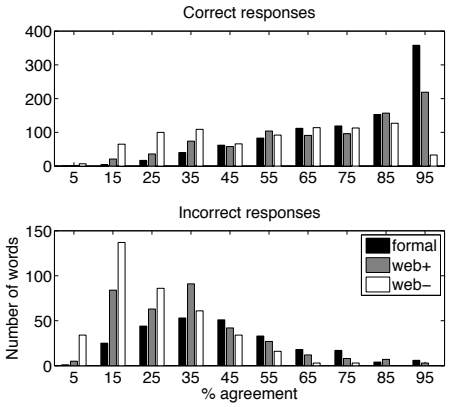

Figure 5: *Agreements for correct and incorrect responses.*

it is still effective in finding potentially-interesting word confusions in noise if both subjective and objective listener selection procedures are followed.

However, not all majority confusions are the same for web and formal listeners. Here, 33 were common to both (shown in figure 6), but 96 discovered in formal listening tests were not majority confusions for the web+ group of web listeners, while the web+ group discovered 52 exemplars which were not majority confusions for the formal listeners. A detailed analysis of confusions is beyond the scope of the current paper. However, findings to date include the following:

- Most confusions involve consonants rather than vowels, and the vowel confusions discovered (mainly /ʌ/-/ɒ/) are likely to be due to a mismatch between the speaker’s accent and the mean expectation of the listener sample.
- Within onset confusions, labials (both plosives and fricatives) are often involved. Sometimes the confusions are inter-labial (/f/ to /p/ or /b/) involving fricative/plosive errors [18], but often there is a labial to /h/ confusion, which highlights the weakness of the labial gesture in acoustic/perceptual terms.
- Nasals are frequently substituted or deleted, especially

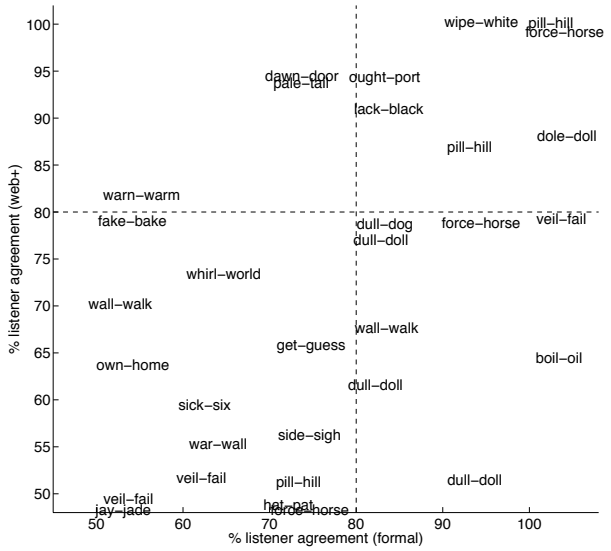

Figure 6: *Majority confusions in common for formal and web+ listeners. Dotted lines show 80% agreement levels.*

in coda position [19].

- Some confusions involve consonant insertion in both coda and onset position, perhaps due to incorporation of background energy fragments (e.g. 'pea-peace').
- Other confusions suggest an effect of word familiarity (e.g. 'veil-fail', 'whirl-world').

4. Discussion

The current study suggests that crowdsourcing has the potential to elicit robust word confusions in noise. A subset of web-based participants selected on the basis of both self-supplied criteria and performance on anchor tokens discovered 7.9% majority confusions among the tokens screened compared to 11.9% found by listeners under conventional laboratory conditions. A subset of web listeners who failed to meet either one or both subjective or objective criteria was far less successful, uncovering 4.1% robust confusions.

Web-based listening appears well-suited to initial token screening, with 'interesting' examples followed up in formal tests. However, only a minority of robust confusions were jointly discovered by the formal and web groups, with significant numbers of confusions unique to each group. While the more homogeneous formal group might be expected to reach a high level of agreement on a larger number of individual tokens, it is surprising that the web cohort made consistent decisions on 52 tokens which formal listeners either did not find confusing or were unable to agree upon. In fact, the formally-tested group had majority correct decisions on 35% of such tokens, while of the majority confusions found by the formal group, web-listeners showed majority agreement on the correct answer in 18% of cases. This outcome raises the possibility that some web-based confusions arise from sources such as low fidelity audio delivery, a notion supported by the finding that absolute levels of performance for web-based participants, even after stringent selection criteria, were far lower than those achieved in the traditional approach (cf. [11]). Further work is needed to explore which processes might cause consistent web-only confusions (see [3] for an approach to confusion diagnosis).

The blocked design of stimulus delivery meant that the web-based test was not optimised for confusion elicitation. A more effective approach is to remove confusion candidates adaptively on the basis of information from earlier listeners, since the likelihood of a candidate reaching consistent confusion status drops rapidly with the number of correct responses.

Acknowledgements. This work was supported by the Sound to Sense (S2S) Marie Curie RTN. The authors thank Stuart Wrigley for assistance with web-hosting.

5. References

[1] K. S. Rhebergen and N. J. Versfeld, "Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," *J. Acoust. Soc. Am.*, vol. 117, no. 4, pp. 2181–2192, 2005.

[2] C. Christiansen, M. S. Pedersen, and T. Dau, "Prediction of speech intelligibility based on an auditory preprocessing model," *Speech Comm.*, vol. 52, no. 7-8, pp. 678–692, 2010.

[3] M. Cooke, "Discovering consistent word confusions in noise," in *Proc. Interspeech*, Brighton, UK, 2009, pp. 1887–1890.

[4] `http://www.britastro.org/vss`.

[5] `http://www.galaxyzoo.org`.

[6] H. Honing, "Evidence for tempo-specific timing in music using a web-based experimental setup," *Journal of Experimental Psychology*, vol. 32, no. 3, pp. 780–786, 2006.

[7] R. Kendall, "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*, vol. 3, pp. 8–10, 2008.

[8] J. Choi, H. Lee, C. Park, S. Oh, and K. Park, "PC-based tele-audiometry," *Telemedicine and e-Health*, vol. 13, no. 5, pp. 501–508, 2007.

[9] L. Blin, O. Boeffard, and V. Barreaud, "Web-based listening test system for speech synthesis and speech conversion evaluation," in *International Conference on Language Resources and Evaluation*, 2008, pp. 2270–2274.

[10] S. Kunath and S. Weinberger, "The wisdom of the crowd's ear: Speech accent rating and annotation with amazon mechanical turk," in *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*. Association for Computational Linguistics, 2010, pp. 168–171.

[11] M. Wolters, K. Isaac, and S. Renals, "Evaluating speech synthesis intelligibility using Amazon's Mechanical Turk," in *Proc. 7th Speech Synthesis Workshop (SSW7)*, 2010.

[12] U.-D. Reips, "Standards for internet-based experimenting," *Experimental Psychology (formerly Zeitschrift fur Experimentelle Psychologie)*, vol. 49, no. 4, pp. 243–256, 2002.

[13] K. McGraw, M. Tew, and J. Williams, "The integrity of web-delivered experiments: Can you trust the data?" *Psychological Science*, vol. 11, no. 6, p. 502, 2000.

[14] H. Honing and O. Ladinig, "The potential of the internet for music perception research: A comment on lab-based versus web-based studies," *Empirical Musicology Review*, vol. 3, pp. 4–7, 2008.

[15] `https://www.mturk.com`.

[16] L. Skitka and E. Sargis, "The internet as psychological laboratory," *Annu. Rev. Psychol.*, vol. 57, pp. 529–555, 2006.

[17] B. D. Cara and U. Goswami, "Similarity relations among spoken words: The special status of rimes in English," *Behavior Research Methods, Instruments, and Computers*, vol. 34, pp. 416–423, 2002.

[18] V. Hazan and A. Simpson, "The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise," *Speech Comm.*, vol. 24, pp. 211–226, 1998.

[19] J. Benki, "Analysis of English nonsense syllable recognition in noise," *Phonetica*, vol. 60, pp. 129–157, 2003.