1

Crowdsourcing in Speech Perception

Martin Cooke (1,2), Jon Barker (3) & Maria Luisa Garcia Lecumberri (2)

(1) Ikerbasque (Basque Foundation for Science), Bilbao, Spain
(2) Language and Speech Laboratory, University of the Basque Country, Vitoria, Spain
(3) Department of Computer Science, University of Sheffield, Sheffield, UK

1.1 Introduction

Our understanding of human speech perception is still at a primitive stage, and the best theoretical or computational models lack the kind of detail required to predict listeners' responses to spoken stimuli. It is natural, therefore, for researchers to seek novel methods to gain insights into one of the most complex aspects of human behaviour. Web-based experiments offer the prospect of detailed response distributions gleaned from large listener samples, and thereby provide a means to ask new types of questions. Instead of instructing listeners to classify speech stimuli into one of a small number of categories chosen by the experimenter, large sample experiments allow the luxury of meaningful analysis of what is effectively an open set of responses. This freedom from experimenter bias is more likely to lead to unexpected outcomes than a traditional formal test which, of necessity, usually involves far fewer participants. Web-based experimentation involving auditory and linguistic judgements for speech stimuli is in its infancy, but early efforts over the last decade have produced some useful data. Some of these early crowdsourcing experiences are related in section 1.2.

However, the promise of web-based speech perception experiments must be tempered by the realisation that the combination of audio, linguistic judgement and the web is not a natural one. Notwithstanding browser and other portability issues covered elsewhere in this volume, it is relatively straightforward to guarantee a consistent presentation of textual elements to web-based participants, but the same cannot be said currently for audio stimuli, and speech signals in particular. Similarly, while it may be possible using pre-tests to assess

This is a Book Title Name of the Author/Editor

the linguistic ability of a web user whose native language differs from that of the target material in a text-based web experiment, it is far more difficult to do so for auditory stimuli. Here, performance alone is not a reliable indicator of nativeness, since it can be confounded with hearing impairment or equipment problems. Section 1.3 examines these issues in depth.

Nevertheless, we will argue that with careful design and post-processing, useful speech perception data can be collected from web respondents. Technological advances are making it easier to ensure that stimuli reach a listener's ears in a pristine state, and that the listener's audio pathway is known. New methodological techniques permit objective confirmation of respondent-provided data. Ingenious task selection can lead to the collection of useful data even if absolute levels of performance fall short of those obtainable in the laboratory.

In the latter part of this chapter we present a comprehensive case study which illustrates one approach which seems particularly well-suited to web-based experimentation in its current evolutionary state, viz. the *crowd-as-filter* model. This technique uses crowdsourcing solely as a screening process prior to the selection of exemplars which are pursued further in formal tests. As we will see in this application, tokens which have the potential to say something valuable about speech perception are rare, and the great benefit of crowdsourcing is to increase the rate at which interesting tokens are discovered.

1.2 Previous use of crowdsourcing in speech and hearing

As early as 10 years ago psychologists were realising the potential of the Internet as an alternative to laboratory-based experimentation. In an early comparison of web-based and laboratory-based experimentation, Reips (2000) identifies a list of 18 advantages of the former. These include a range of obvious factors such as the availability of a large number of subjects and the ability to reach out to demographically and culturally diverse populations, as well as cost savings in laboratory space, equipment and subject payments. However, Reips argues that there are also subtler advantages that may be no less important. For example, participants of Internet-based experiments are 'highly' voluntary, meaning that there may be less motivation to produce deceptive responses. Likewise, results may have high external validity and generalise to a larger number of settings (e.g., Laugwitz 2001), and findings are likely to be more applicable to the general population (Horswill and Coster 2001).

Although the web-based methodology has been discussed amongst psychologists for over 10 years, it is only very recently that it has been seriously considered by hearing researchers. This is no doubt largely due to technical difficulties in the reliable delivery of audio stimuli to web-users who may be using software that is several product cycles out-of-date and who could only recently be expected to have Internet connections with adequate bandwidth. Nevertheless, the increased ease and precision with which audio-based experiments can be conducted is evidenced by the rapidly growing interest among musicologists (Honing 2006; Honing and Ladinig 2008; Kendall 2008; Lacherez 2008), audiologists (Bexelius et al. 2008; Choi et al. 2007; Seren 2009; Swanepoel et al. 2010) and, of particular relevance for the current chapter, the speech technology community (Blin et al. 2008; Kunath and Weinberger 2010; Mayo et al. 2012; Wolters et al. 2010).

The first major web-based psychoacoustic experiment, published in 2008, studied the impact of visual stimuli on unpleasant sounds (Cox 2008). In this study participants were asked to listen to a sound while observing an image and were then asked to rate the ‘horribleness’ of a sound on a 6-point scale. The experiment was run from a simple Flash-enabled web site and was accessible to anyone with a web-browser and a computer with audio output capabilities (e.g., loudspeakers or headphones). Given that attitudes to sound are highly-individual, the study required a large and demographically broad subject base in order to produce meaningful conclusions. The study could not have been conducted using a traditional lab-based methodology. However, as Cox notes, simply placing an experiment on-line does not guarantee a large number of participants, regardless of how well-designed the web interface is. In order to recruit participants some form of media campaign is required. By making use of the local, national and international press, Cox was about to collect over 1.5 million votes. Clearly, success here rests on being able to capture the public imagination and being fortunate in conducting a study to which people could easily relate. Participants were not paid for their time: once engaged with the experiment the main motivating factor was that in return for each response, the response of the general population is revealed, allowing participants to compare their view with that of others.

As the existence of the current volume testifies, crowdsourcing has also recently been recognised as a useful methodology within the speech research community. In contrast to Cox’s voluntary crowd, speech researchers have largely employed crowds that have been financially compensated, usually through the use of Amazon’s Mechanical Turk (MTurk). Annotation of speech corpora was the first problem to be addressed in this way. The key concern, clearly articulated by Snow et al. (2008), is, ‘Cheap and fast – but is it good?’. Their conclusion, echoed in the title of a similar study ‘Cheap, fast and *good enough*’ (Novotney and Callison-Burch 2010) is that if crowdsourcing output is suitably handled, many large labelling tasks can be completed for a fraction of the cost of using highly-paid experts and, crucially, with no significant loss in quality. Crowdsourcing has also been used for read-speech corpus collection. McGraw et al. (2009) employ an on-line educational game to generate a ‘self-annotating’ corpus. Similarly, the VoxForge project (Voxforge 2012) asks visitors to their site to read prompted sentences with the aim of collecting a quantity of transcribed speech sufficient for training robust acoustic models for use with free and open-source speech recognition engines.

The success of crowdsourcing in the context of corpus collection, transcription and annotation is encouraging, but does not by itself demonstrate the suitability of the methodology for the study of *speech perception*. In labelling tasks, human judgement is not being recorded as a means to judge the human perceptual system, but rather as a means to generate data that will be used to either bootstrap or evaluate learning algorithms. Error in human judgements is a source of noise in the labels which may lead to suboptimal machine learning, but it will not lead directly to false experimental conclusions. Further, labelling tasks have a high degree of inter-listener agreement allowing outlying data to be filtered. Perceptual tasks, in contrast, are often concerned with the distribution of judgements or small statistical differences between conditions that are more likely to be masked in the event of a lack of experimental control.

Examples of crowdsourced speech perception studies can be found in the *speech synthesis* community. The annual Blizzard speech synthesis challenges (e.g., King and Karaiskos 2010) use human judges to rank the quality of competing synthesis systems. The judges include a mix of expert listeners and a contingent of naïve listeners recruited via email and social networking sites. Listeners typically perform tests in their offices over the Internet using headphones, though judgements are supplemented and validated by extensive lab testing. In the context of the Blizzard challenges, Wolters et al. (2010) have recently tested the validity of using a purely crowdsourcing methodology for evaluating the intelligibility of speech synthesis systems. Using listeners recruited via MTurk they find that although absolute intelligibility is much worse than in laboratory testing (a finding echoed in other crowdsourced speech perception studies, e.g., Cooke et al. 2011; Mayo et al. 2012), crucially, the MTurk listener scores reflect the *relative* intelligibility of the systems fairly well. If the task is to compare a new system against the current state-of-the-art then reliable relative judgements may be all that is required.

1.3 Challenges

Despite their many clear advantages, web-based studies are not without their problems. Reips' early review of web-based experiments identified a carefully considered list of disadvantages (Reips 2002). The extent to which these issues invalidate the web-based methodology has been much debated in the intervening years (Honing and Reips 2008; Kendall 2008; Skitka and Sargis 2006). Despite strongly polarised views, it is clear that the validity of a web-based methodology is highly-dependent on the nature of the experiments being conducted. In this section we will re-examine the key difficulties with a specific focus on the requirements of speech perception experiments.

The most commonly cited problem for the web-based methodology is the comparative lack of experimental control. In fact, most of the difficulties discussed by Reips (2002) can be seen as symptoms of this underlying problem, and the challenges of experimental control feature prominently in studies such as Wolters et al. (2010). Generally speaking, in all web-based experiments, there is a trade-off: the experimenter accepts a reduced amount of control, but hopes that this may be compensated by the opportunity to recruit a very large numbers of subjects, i.e., the added measurement noise due to nuisance variables is, it is hoped, more than countered by the increase in the number of data points. Nevertheless, a large number of data points cannot protect against systematic biases and even if subjects are plentiful it is bad practice to waste resources through poor experimental design. It is therefore worth considering how to minimise the potentially damaging consequences of the reduced experimental control inherent in web-based experimentation.

In considering a speech perception experiment, the factors that we wish to control can be broadly categorised under three headings: (i) *environmental factors* that describe external conditions which might affect a subject's responses; (ii) *participant factors* that describe how listeners are selected; and (iii) *stimulus factors* that describe how the sounds that are heard will be controlled. We consider each factor in turn.

1.3.1 Control of the environment

The loss of environmental control is perhaps the most obvious difficulty facing the web-based methodology. For speech perception studies the *acoustic* environment is clearly very important. Laboratory listening tests are typically conducted in sound-attenuating rooms with state-of-the-art equipment for audio reproduction. In contrast, web-based tests may be performed by listeners sitting at computers at home or at work, in rooms with different amounts of environmental noise and with uncalibrated headphones of unpredictable quality.

A first consideration is whether the experiment is likely to be sensitive to environmental noise. Clearly, it would be unwise, for example, to attempt to measure signal reception thresholds in a web experiment: even quiet offices typically contain significant ambient background noise as well as the possibility of intermittent audio distractions (e.g., incoming calls, visitors). However, if the experiment involves processing speech at an adverse signal-to-noise ratio (SNR) then the additional environmental noise may not be significant if its peak intensities lie below the level of the experimental masker. Alternatively, it may be possible to reduce the unpredictability of the noise background. One practical technique which may be appropriate in some tasks is to add a fixed noise floor to the stimulus to mask variation caused by differing *low* levels of external noise.

Variance caused by differing headset quality is a separate issue. Headsets may well possess significant differences in frequency response. Subjects could be asked to provide information about their headset, but this would serve little purpose as cheap consumer headsets are uncalibrated and variation may be present even between headphones of an identical make and model. However, for many experiments these sources of variation are of little real significance. Consider in particular that there is natural variation in the spectral shaping of speech caused by room acoustics, and further variation in the audiograms of even supposedly 'normal hearing' listeners. Depending on the details of the study design, it might be argued that it would be unusual for the result of a speech perception experiment to be heavily dependent on factors that the perceptual system itself works hard to minimise or ignore.

A broader factor is the degree to which the environmental context affects the attention of a participant. Consider that in a traditional experiment a subject has been brought to a lab where they are placed in an environment designed to be free of distraction. The subject will have given up time in their day to perform the experiment and can generally be expected to be focused on the task. This is in stark contrast to the web-based situation where the participant, even if highly-motivated, situated in a quiet office and wearing good quality headphones, is far less likely to be devoting their full attention to the experiment. They may have other applications running on their computer, they may be receiving email alerts or instant-messages; they are likely to be at work and generally in possession of a multi-tasking mindset. There is little that can be done to control these factors, but two points are worth noting: first, these factors are not totally excluded from traditional experiments - subjects unfamiliar with listening experiments will find the unfamiliarity of a hearing lab a distraction in itself, and differences between the mental stamina of participants will lead to varying degrees of attention throughout an experimental session. Second, it can be argued that results that are obtained in a natural environment have greater external validity, i.e., they are more likely to be representative of the type of hearing performance achieved in day-to-day life. As we noted earlier, it has been argued that web-based findings are likely to generalise better to a greater range of real-world situations (e.g., Laugwitz 2001).

1.3.2 Participants

When conducting a behavioural experiment it is normal to seek a homogenous group of participants meeting some well-defined selection criteria. Relevant criteria in listening experiments might include factors such as gender, age, language history and normality of hearing. It is the experimenter's responsibility to ensure that the selection criteria are met when recruiting participants. The opportunity for face-to-face interaction between the experimenter and the participant allows for a robust selection process. In the web-based case participant selection is more problematic. Criteria can be made explicit, allowing participants to self-select, or information can be gathered from participants using online forms which allow non-conforming participants to be filtered out subsequently. Here, two problems arise: selection depends on participants' trustworthiness, and certain selection criteria – such as possession of normal hearing – may be difficult or impossible to apply remotely, at least with current technology.

Trust

The 'trustworthiness' of participants is an oft-cited problem with web-based experiments (McGraw et al. 2000). How can it be guaranteed that the participants are providing correct information? Moreover, how can it be guaranteed that they are providing meaningful responses during the experiment itself? However, proponents of web-experimentation point out that this question applies generally to all behavioural testing whether web-based or lab-based. Further, it has been argued that participants 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 overseen by an experimenter (Honing and Ladinig 2008). Of course, this reasoning only applies to those applications of crowdsourcing where participants are not being paid and provides an argument other than cost savings for not making crowdsourced experiments financially rewarding. Further, the ease with which a web-based subject can drop out ensures that participants completing the experiment will generally be highly-motivated and providing good quality data. (Ironically, high drop-out rate is often something that web-experimenters discuss as a concern.)

If participant trustworthiness is considered to be a serious issue then precautionary checks and measures can be put in place to screen out untrustworthy data. This can often be achieved through careful experimental design. For example, if the test requires a certain level of hearing acuity which in turn necessitates good equipment and low levels of background noise, experimenters may set a threshold with tokens which must minimally be identified correctly to ascertain that the required conditions are being met. If listeners need to have a particular linguistic background (e.g., regional, native, non-native accent, L1 of origin), which will be determined by a questionnaire, experimenters can also add criterion tokens designed to filter out participants who are not being frank about their profile (see also the use of what we call 'anchor tokens' in section 1.5.6). When membership of a specific group is sought, appropriate slang vocabulary presented orally can be a useful means to determine affiliation. More generally, to increase the quality of participant data, web forms should not be pre-filled with default values (something the current authors are somewhat guilty of in the case study presented later in this chapter!) and participants should be compelled to complete the form before commencing the experiment proper. It may be advisable to avoid disclosing the selection criteria to avoid participants supplying dishonest data in order to gain access to the experiment. This is particularly true if respondents are motivated by financial reward.

Even valid participants – i.e., those meeting the experiment's selection criteria – may need to be screened out if they are not sufficiently engaged in the task and are simply providing arbitrary responses in order to complete the work with minimum effort. It may be possible to detect such participants by monitoring response timings and checking that they fall in a normal range. Another simple and commonly-employed technique is to intersperse the genuine trials with a number of dummy trials which have a highly-predictable ‘correct’ response, where ‘incorrect’ responses to these trials is then a sure indicator that the participant is not cooperating (Sawusch 1996). Finding reliably-correct dummy trials is itself something that is aided by large listener samples in a crowdsourcing study, as we show in section 1.5.6.

Hearing impairment

The impossibility of measuring participants’ audiograms is a major limitation of the web-based methodology. Many people with mild or even moderate hearing loss do not realise that they have a deficit, especially if the loss has been progressive and not associated with trauma (as is typically the case with age-related hearing loss). Despite remaining undetected, a hearing deficit can easily lead to a measurable and significant effect on speech perception, particularly in noise, and render a potential participant unsuitable for a wide variety of speech perception experiments. Robust solutions to this problem are not obvious. Wolters et al. (2010) screened participants using a standard hearing questionnaire based on the Hearing Handicap Inventory for Adults (HHIA) (Newman et al. 1990) but HHIA scores are only weakly-correlated with audiometric measures and Wolters et al.’s conclusions appear to cast doubt on the efficacy of this approach. It is perhaps more reliable to identify abnormal subjects directly from the statistics of their stimulus responses and apply a post-hoc filtering to the results. This may be made easier if ‘diagnostic’ stimuli can be inserted into the experiment. Also, as with other issues with poor subject control, the difficulties can be reduced by designing experiments that rely on within-subject rather than between-subject comparisons.

Linguistic background

Speech perception tasks in general and accent judgements in particular are usually carried out by naïve native listeners of the target language (Major 2007), although expert judgements (Bongaerts 1999) and non-native data (MacKay et al. 2006; Major 2007; Riney and Takagi 2005) have also been collected, depending on the experimental aims and listener availability. Despite their predominance, the reliability of native judges has been questioned (Major 2007; Van Els and De Bot 1987). Dialectology and L2 studies find native listeners to be far from homogeneous as a group: listeners vary in their ability to judge accents and to some extent in their perceptual performance depending on their history of exposure to different varieties and languages, as well as other individual variables such as metalinguistic awareness, age, hearing, and, for some tasks, personality and educational factors. In the case of crowdsourcing, familiarity with technology and computer interfaces can also introduce variability in the results.

One of the issues that needs to be handled carefully is the acquisition of indicators of a participant’s linguistic background. In crowdsourcing this monitoring has to be done indirectly, since the experimenter is not usually able to ascertain a participant’s linguistic competence by means of observations of their speech. In principle, notwithstanding additional technical and ethical concerns, web-based collection of a speech sample is possible, though its analysis would be expensive in time and effort and perhaps difficult to justify in the context of a speech perception experiment. Careful questionnaires are needed which clarify which languages are spoken by listeners, to which level and from what age (i.e., which are native languages, second languages or foreign languages). In this respect, care should be taken with the terms used to describe multilingual situations (e.g., ‘bilingualism’, ‘second language learners’ etc.; see Garcia Lecumberri et al. 2010, for a review). However, the thoroughness of the questionnaire and therefore its length need to be weighed against the possibility of discouraging participation. To ensure data reliability, as mentioned above, participants’ self descriptions can be correlated with performance on criterion tokens which may provide useful indicators of the trustworthiness of questionnaire responses. Kunath and Weinberger (2010), exploring the use of MTurk listeners for perception tasks, establish a baseline pre-test to determine listeners’ accuracy. They also propose for future studies a more demanding and comprehensive ‘qualification test’ which will screen listeners before selecting them as participants in the main perception task.

1.3.3 Stimuli

In all listening experiments it is clearly important that a reproducible stimulus can be delivered undistorted to the subject. In the early years of web-based experimenting, technological constraints were such that the web was a poor substitute for the laboratory if experiments required delivery of reliable, high-quality audio and/or video. Web-browsers would not always provide media support without the installation of non-standard ‘plugins’ and file sizes could require excessive download times on narrowband connections. However, with the more widespread availability of higher bandwidth connections to the Internet and with the advent of technologies such as MPEG4, HTML standardisation and client-side web scripting, these problems have largely vanished for most users.

A number of software frameworks for constructing and hosting web experiments have emerged in recent years. Systems such as WEXTOR (Reips and Neuhaus 2002), NetCloak (Wolfe and Reyna 2002) and DEWEX (Naumann et al. 2007) perform processing on the server-side to avoid client-side compatibility issues. A fully server-side approach, however, is unsuitable for speech perception experiments that require controlled delivery of stimuli on the client. In contrast, the WebExp package employs a Java applet running on the client (Keller et al. 1998, 2009) which allows sophisticated control but at the expense of requiring that Java has been installed in the client’s browser, something that cannot be guaranteed, especially on mobile devices. A potential solution is demonstrated by the Percy framework (Draxler 2011) that makes use of the latest HTML specification, HTML5, which provides multimedia tags to control the presentation of audio. This technology permits the development of web experiments which will run in any compliant browser with no need for external media players, plug-ins or additional client-side software.

Despite ongoing technological advances a few issues remain worth noting. First, although increased bandwidth means audio stimuli can be continuously streamed over the Internet, careful software design is needed to ensure that the participant receives stimuli in a predictable fashion. In some experiments even tiny uncertainties in the start time of a stimulus can impact the result. Therefore, stimuli need to be downloaded or buffered so that their onset times can be controlled with millisecond precision. Pre-downloading an entire experiment’s stimuli may take appreciable time and participants may have little patience for watching a download bar. A solution calls for good software design e.g., using buffering and asynchronous downloads that occur in dead time while the participant is reading instructions or processing the previous stimulus.

A further issue concerns the quality of client-side audio hardware. Although a digital signal can be delivered with fidelity and in a timely manner to the participant, its reproduction can be compromised by poor-quality sound cards, variability of headphone frequency responses and interference from noise in the surrounding environment as discussed earlier.

Finally, *audio-visual* speech perception experiments require precise audio-visual synchronisation. Participants can be sensitive to asynchronies of as little as 40 ms if the audio arrives in advance of the video. For television broadcasting, for example, the Advanced Television Systems Committee recommends that audio should lead video by no more than 15 ms and audio should lag video by no more than 45 ms. The commonly employed MPEG encoding can ensure close synchronisation but only if care is taken during preparation by, for example, inserting presentation time stamps into the MPEG metadata field. Even with due care, data can potentially become desynchronised if there is significant mismatch in the video monitor and audio processing circuitry after decoding. Such timing errors would, however, be unusual in modern hardware and this problem can be expected to disappear in the near future for most users.

1.4 Tasks

Choosing a task depends mainly on the aims of the data collection and on the stimuli which will be used. In principle, we can classify speech perception tasks broadly according to what they are aiming to measure: signal properties such as speech intelligibility, quality and naturalness, speaker aspects (e.g., accent evaluation), and listeners' perceptual abilities and phonological systems.

1.4.1 Speech intelligibility, quality and naturalness

Speech intelligibility measurement is the object of a great many studies in speech perception, motivated by investigation of factors such as speech style and listener characteristics as well as the effects of maskers, vocoders and synthesis procedures. Intelligibility is normally quantified objectively[1] by having listeners report what they have heard either orally or in writing. Thus, intelligibility is frequently measured with tasks that take the form of oral reports (e.g., repetition of speech, answers to questions, utterance completion) or written reports (orthographic or phonetic transcriptions) or selection from response alternatives presented programmatically. In the case of studies which use crowdsourcing, the latter option is the only feasible modality at the present time. Participants can be asked to type what they have heard (Wolters et al. 2010) or choose their response using a custom designed interface (Garcia Lecumberri et al. 2008).

Another measure which is sometimes grouped with intelligibility is 'comprehensibility' – how easy it is to understand a particular utterance or speaker. As opposed to intelligibility, comprehensibility is a subjective measure, since it depends on a listener responding based on their impression rather than on quantifiable data (e.g., number of words/segments understood). Comprehensibility, like many other subjective listener-derived judgements,

[1] Note that the term *subjective intelligibility* is also frequently used in this context to distinguish between measures derived from listeners on the one hand and predictions made by so-called *objective intelligibility models* on the other.

is typically measured by means of Likert scales. In crowdsourcing, comprehensibility of synthetic speech has been assessed with a version of the Mean Opinion Score (MOS), a 5-point Likert scale (Blin et al. 2008).

Naturalness, alongside intelligibility and comprehensibility, is the main criterion by which synthetic and other forms of generated or coded speech are routinely judged. One of the pioneering applications of crowdsourcing has been in the annual Blizzard Challenge (e.g., King and Karaiskos 2010), which also employs Likert scales similar to the ones mentioned above.

The measures outlined above are frequently employed in conditions which simulate some elements of everyday speech perception, usually by presentation in the presence of competing sound sources or under cognitive load. The latter is especially relevant for accent judgements (see section 1.4.2 below). Formal studies have corroborated the intuition that foreign accents (Munro and Derwing 1995) and unfamiliar regional accents (Floccia et al. 2006) can make special demands on the part of the listener so that the cognitive effort required from listeners is higher than when listening to a familiar accent. One direct way of measuring cognitive effort is to monitor response times or latencies. Latencies can be calculated as a global quantification of overall accent effects or at a finer level of detail, in terms of segmental or featural variables (e.g., Bissiri et al. 2011).

From a technical perspective, reaction time monitoring in a crowdsourced experiment requires careful design, particularly with respect to the balance of responsibilities for client and server components, but the use of suitable client-side software enables reliable reaction monitoring (e.g., as demonstrated in Keller et al. 2009). However, a less tractable set of issues comes from the need for a commitment on the part of participants to focus on the task to the best of their abilities during its time-span and to avoid distractions. A related issue is that it may be difficult for the remote experimenter to measure web-respondent fatigue effects which sometimes accompany tasks involving cognitive load.

1.4.2 Accent evaluation

Outside the strict communicative confines of intelligibility as measured by narrow criteria such as the number of keywords identified correctly, speech provides a wealth of other information. For instance, speech provides cues to the geographical and/or social origin of speakers, and also conveys affect. In turn, a talker's speech provokes attitudinal responses in listeners. Accent research, for both native and foreign accents, has addressed all these areas.

Accents may be analysed according to speakers' geographical or linguistic origins. Within this broad field, some researchers are concerned with straightforward accent classification. Clopper and Pisoni (2005) suggest a perceptual regional accent classification task in which listeners are asked to indicate on a map where a particular speech sample belongs to in geographical terms. In some contexts, regional accent judgements are linked to opinions of social class (Trudgill and Hannah 2008; Wells 1982). Kunath and Weinberger (2010) used crowdsourcing to classify foreign-accented English samples according to three possible first language origins as well as in the degree of foreign accent present in each sample. The magnitude or degree of accent corresponds to the extent to which it differs from a particular norm or standard. The notion of distance from some reference is flexible. In the case of synthetic or manipulated speech, an evaluation might be designed to measure the extent to which speech differs or conforms to a 'standard' set by natural speech or even a specific voice. Conformity evaluations used for synthetic speech and foreign/regional accents typically employ similar response measures as for studies of naturalness (i.e., rating scales).

Accents can cause attitudinal reactions on listeners, who may feel charmed, soothed or interested when listening to certain voices or conversely may get irritated, anxious or bored when listening to some accents that differ from their own. Negative reactions are often present in the case of foreign accents (Brennan and Brennan 1981; Fayer and Krasinski 1987), probably due to communication breakdowns or the extra effort listeners need to make in order to repair phonetic deviations (Fernandez Gonzalez 1988). Paradoxically, foreign speakers who achieve near-native accents may also provoke unusual reactions such as suspicion or envy. Again, attitudinal style judgements are usually carried out by means of Likert scales.

Speech conveys paralinguistic and extralinguistic information and as such, listeners develop constructs about speakers' personalities and other characteristics such as intelligence, education, profession, trustworthiness and socioeconomic status, which may become generalised for particular combinations of speaker/listener groups and according to certain stereotypes. There are connections between accent and physical qualities of their place of origin (Andersson and Trudgill 1990). Thus, regional accents will be judged as euphonic if they belong to a scenic location whereas accents from industrial areas tend to be considered uglier. Research on second language acquisition has shown that for foreign accents too listeners judge speech differently depending on the perceived origin of the speakers (Hosoda et al. 2007) and these judgements extend to beliefs about economic level, status, job suitability and professional competence (Boyd 2003; Dávila et al. 1993).

Since most of these accent evaluation tasks are based on speech tokens evaluated along Likert scales, they are certainly appropriate targets for crowdsourcing studies, just as Kunath and Weinberger (2010) have done for degree of foreign accent, and as in the evaluations of speech naturalness cited earlier.

1.4.3 Perceptual salience and listener acuity

Exploring the perceptual salience of speech features and listeners' perceptual abilities in, for example, detecting just-noticeable differences (JNDs) is often the object of basic research or an adjunct to other speech perception tests. Discrimination tasks are typically used here. Discrimination tasks normally give no information as to listeners' phonological or grammatical classification of the stimuli. Stimuli may be presented in one or two pairs (AX or AA - BX), or in triads (ABX) in which a listener has to decide if X is the same or differs from the other tokens in the sequence. In the latter case, to avoid short-term memory biases towards the stimulus closest to X, the triad AXB is the preferred alternative (Beddor and Gottfried 1995).

Discrimination tasks require a very high degree of stimulus control to avoid ascribing perceptual performance to a confounding variable. It seems unlikely that crowdsourcing will be suitable (or indeed necessary) for the estimation of psychoacoustic distinctions such as JNDs in pitch or duration. It is perhaps surprising then that crowdsourcing has been used in speech-based discrimination tasks. For instance, Blin et al. (2008) developed a system to support ABX discrimination tasks used to compare manipulated and natural speech. Notwithstanding the fact that these two examples involve relatively high-level categorisations and hence to some extent prior information, the basis for any decision originates in part from low-level stimulus differences. One would certainly expect to find differences in sensitivity between formal and web-based discrimination tasks.

1.4.4 Phonological systems

One of the main aims of speech perception research has been to discover something of the structure of a listener's phonological system, addressing questions such as: what are the phoneme categories in an inventory, what is the internal structure of those categories, how is the phonological system organised, and how does it relate to other phonological systems. For these purposes, sound identification tasks have been widely employed. In identification tasks the stimuli may correspond to phonemic categories or their realisations, but dialectology also uses this type of task to classify accents. Here, the task for the listener is to provide a label (e.g., phonetic symbol, spelling, word, accent name) for the stimulus.

Identification tasks may be totally open so that the listener comes up with the label. However, this can make it difficult to compare results across listeners because of a potential profusion of labels. Additionally, the task may be too difficult, particularly for speakers with low metalinguistic awareness. In that case, individual abilities introduce a great deal of response variability. To avoid these risks, experimenters frequently provide ready-made label choices. The number of choices may be restricted or open (e.g., presenting labels for all the language's phonemes or for all the dialectal areas). While limited response sets typically lead to easier tasks, they carry the danger of leaving out the label which the listener would actually choose in response to a stimulus (Beddor and Gottfried 1995).

In phoneme labelling studies particular care needs to be taken with the labels chosen since orthography may play an important and often confounding part, particularly in cross-linguistic research (Beddor and Gottfried 1995). A study which compared English consonant perception across listeners from eight different L1 backgrounds (Cooke et al. 2010) found that for naïve listeners orthography has a strong influence. The alternative of using phonemic symbols restricts the participants to populations which are familiar with them, or runs the risk of providing unwanted perceptual training during symbol familiarisation.

In order to explore the internal structure of listeners' phonological categories and how different realisations are classified as exemplars of the same phonemic category, categorical discrimination tasks may be used. These tasks resemble straightforward discrimination tasks except that in categorical discrimination all stimuli are physically different (ABC) and listeners have to classify as 'same' those belonging to the same phonemic category (e.g., A and B). This task can also be extended to accent studies in order to group different speakers into accent groupings defined by regional origin or L1, for example.

Category Goodness Rating is a metric designed to explore the internal structure of phonemic categories at a more detailed level than is possible with categorical discrimination. Listeners rate individual stimuli based on how good an exemplar it is of a particular phonological category. This task usually accompanies either an identification task or a categorical discrimination task. Ratings can use Likert-like scales (Kuhl 1991) or continuous scales (Gong et al. 2011).

The types of task outlined above may well be suitable for crowdsourcing – there have been few studies in this domain to date – but their hallmark of multiple response alternatives which convey what might be quite subtle categorical distinctions raises a broader issue for web-perception experiments: how to instruct the participant and how to determine if the instructions have been adequately understood. In formal laboratory situations, the human-human interaction between experimenter and participant is rich, flexible, and rapid, providing an immediacy of feedback both for the participant who may be unsure of what is required, and for the experimenter, who can form a judgement about whether instructions have been understood. The experience for a web-based participant is monochrome by comparison. Instructions are usually presented in textual form, perhaps with audio examples. There is generally no personalised interaction, nor an opportunity to ask questions. One positive aspect is that instructions are the same for all participants, although there is no guarantee that instructions are followed or even read!

1.5 BIGLISTEN: A case study in the use of crowdsourcing to identify words in noise

We now describe in some detail a recent crowdsourcing exercise in which listeners attempted to identify words presented in noise via an online application. This study, which we call the BIGLISTEN, is an example of the *crowd-as-filter* approach, where the numerical advantage inherent in the crowd is used to screen a potentially vast number of stimuli to find tokens which meet some criterion, which are subsequently presented to listeners under a traditional controlled laboratory regime. In part, the BIGLISTEN web application was developed to pilot ideas in crowdsourcing for speech perception and in particular to enable comparisons between formal and web test results in order to evaluate the merits of the approach.

In this section, we describe the problem which motivated the BIGLISTEN and argue that crowdsourcing is a natural solution, before explaining the design decisions taken during development of the web application. We go on to highlight some of the principal findings and discuss the lessons from the pilot approach. More details of the BIGLISTEN can be found in (Cooke 2009; Cooke et al. 2011).

1.5.1 The problem

A better understanding of how listeners manage to communicate effectively using speech in realistic environments – characterised by the presence of time-varying degradations resulting from competing sound sources, reverberation and transmission channels – will enable the development of more robust algorithms in speech and hearing technology applications. One key ingredient is a detailed computational model which describes how listeners respond to speech presented in noise. At present, we have what have been termed *macroscopic* models which make objective predictions of subjective speech intelligibility (e.g., ANSI 1997; Christiansen et al. 2010) and quality (Rix et al. 2001). By contrast, the study of *microscopic* models which try to predict what listeners hear at the level of individual noisy tokens is only just starting (see, e.g., Cooke 2006). At the heart of the microscopic modelling approach is the need to discover *consistent* responses to individual speech-in-noise tokens across a sufficient sample of listeners, and to uncover a large enough corpus of such examples to allow comparative evaluation and refinement of microscopic models.

While less-sophisticated microscopic models might be expected to respond like listeners when tokens are correctly recognised, they are less likely to make the same errors as listeners unless the model successfully captures in some detail the processes involved in human speech perception. Therefore, while consistently-reported correct responses in noise are useful in model evaluation, unexpected responses common to many listeners are particularly valuable for the microscopic modelling enterprise.

The main requirement, then, is to collect a corpus of individual noisy speech tokens, each of which induces a high degree of consistency in listener responses for both correctly heard and misheard cases. More generally, we are interested in measuring the response distribution for each noisy token. Low entropy distributions, characterised by one, or perhaps two, clear concentrations of responses, are the goal of token screening. Robust estimation of response distributions demands the availability of a large number of *different* listeners, and hence makes this an ideal application for crowdsourcing.

1.5.2 Speech and noise tokens

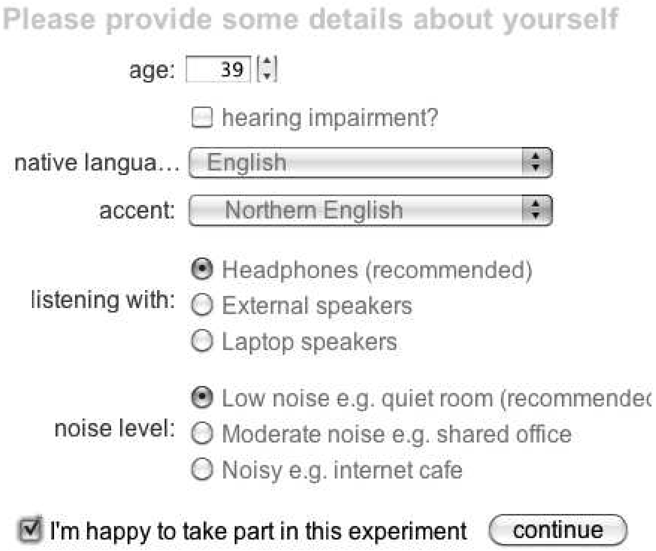
Users of the BIGLISTEN application identified one or more blocks of stimuli. Each block contained 50 monosyllabic English words mixed with one of 12 types of noise. Words came from an existing list (Cara and Goswami 2002) using selection criteria designed to encourage confusability (e.g., high spoken and written frequency and the possession of a large set of phonological neighbours) and screened to remove obscenities. Five native British English speakers, 4 males and 1 female, each recorded the subset of over 600 words which met these criteria.

A variety of noises were used to encourage different kinds of confusions, resulting, for example, from foreground-background misallocation of patches of spectro-temporal energy or masking of target speech components. Maskers included speech-shaped noise, multitalker babble for a range of talker densities (including a single competing speaker), envelope-modulated speech-shaped noise and factory noise. Each block of stimuli contained words from a single target talker and a single type of masker. The signal-to-noise ratio (SNR) was set based on pilot tests to a range low enough to create potential confusions but not so low as to lead to near-random responses. In practice, the SNR decreased within a narrow range ($SNR_{\max}$ to $SNR_{\min}$) within each block of stimuli. The first 5 stimuli in the block acted as practice tokens. Their SNRs decreased linearly from +30 dB (i.e., almost noise-free) to $SNR_{\max}$, after which the SNR decreased linearly for the remaining 45 tokens to $SNR_{\min}$. Different maskers used different SNR ranges to reflect the finding that listeners' ability to reach a criterion intelligibility level varies with noise type (Festen and Plomp 1990). The purpose of using a decreasing SNR during the block was to test a range of noise levels where consistent confusions might be expected to occur and also to provide the user with a more challenging and perhaps engaging task experience with time. Users could complete as many blocks as they wished. More details of the task and stimuli are provided in Cooke (2009).

1.5.3 The client-side experience

Visitors to the BIGLISTEN home page saw a single web page containing a small amount of motivational text, instructions and the test itself. The page also included clickable examples of words in noise which had the dual purpose of illustrating the types of stimuli in the test and allowing the volume control to be set to a comfortable level. The test interface ran via a Java applet. The applet initially displayed a form to collect a small amount of information from the respondent and to seek their consent to take part in the test (figure 1.1). Once the form was filled in and consent given, the main experimental interface – essentially a text input box –

replaced the form (figure 1.2). After completing a block, users received immediate feedback on their performance, expressed as a ranking based on their score of words correctly identified within the subset of listeners who had heard the same test block.



Please provide some details about yourself

age: 39

□ hearing impairment?

native langua… English

accent: Northern English

listening with:
● Headphones (recommended)
○ External speakers
○ Laptop speakers

noise level:
● Low noise e.g. quiet room (recommendec
○ Moderate noise e.g. shared office
○ Noisy e.g. internet cafe

☑ I'm happy to take part in this experiment

Figure 1.1 The initial page of the the test interface, showing the questionnaire filled in.

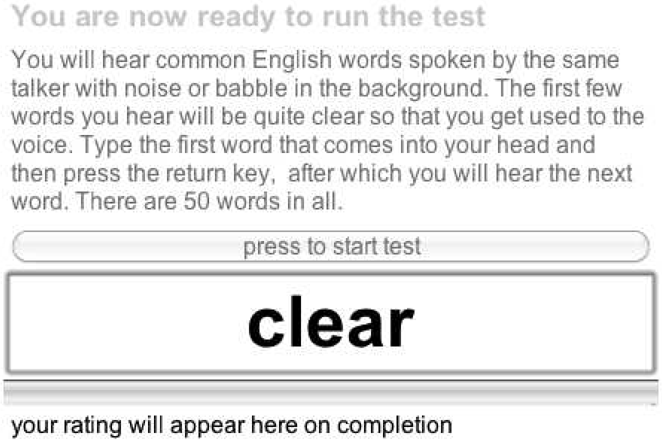
1.5.4 Technical architecture

General considerations

Previous sections highlighted those aspects of a web-based experiment that are largely outside an experimenter's control. However, the impact of many of these factors can be mitigated to a large extent through a careful consideration of software architecture and design. Key technical goals include minimising the impact of network delays (e.g., through buffering), maintaining precise control over data such as audio signals and user responses transferred between client and server, robust handling of spikes in user interest (e.g., via resource pooling), encouraging task completion through a seamless and rapid data gathering process, and by accommodating as far as possible differences in client hardware, software and location. Consequently, technical solutions are favoured which support portability, localisation, scalability and client-server load sharing in addition to a rich set of programming structures.

The BigListen architecture

The BigListen application employs Java technologies coupled with a back-end relational database. Java provides good support for audio (via the `javax.sound.sampled.*`



You are now ready to run the test

You will hear common English words spoken by the same talker with noise or babble in the background. The first few words you hear will be quite clear so that you get used to the voice. Type the first word that comes into your head and then press the return key, after which you will hear the next word. There are 50 words in all.

clear

your rating will appear here on completion

Figure 1.2 The main experiment screen.

package) and user interfaces (via `javax.swing.*`) as well as multiple threads of execution and database integration. In principle, highly-variable demand can also be accommodated using Enterprise Java technologies. These were not felt to be necessary for the initial version of the BigListen application but the ease of future migration to a scaleable web app is an attractive feature of Java.

A Java applet running on the client's browser is responsible for collecting respondent-provided information, delivering noisy speech tokens, gathering participants' responses and providing feedback at the end of each test block. A further applet supports the inclusion of buttons on the web application's introductory page to provide examples of stimuli and also to allow the user a convenient means to check the volume setting.

A Java servlet mediates all information flows from and to the applet. The servlet is responsible for all communication with the database and filestore, in addition to one-time initialisation of common resources such as connection pools. The applet-servlet design pattern permits full abstraction of implementation details (e.g., no database language code is present in the applet, nor any direct links to other back-end resources) facilitating rapid reconfiguration of the back end without affecting the user view of the application and without requiring recoding at the applet level.

Information about test blocks as well as homophone, language and accent lists is held in a relational database in the BigListen application. The database also stores participant-supplied information, word responses and timing data. For efficiency, complete blocks of 50 test stimuli are bundled into single files stored on the server. To enable delay-free presentation during the test itself, a block of stimuli is downloaded to the client applet while the user fills in the form. Intermediate buffering strategies, such as downloading the next or next-but-one stimulus while the user hears the current one, may be more appropriate than monolithic block transfer in situations where a user's results can be put to immediate use in selecting stimuli for successive users. Here, the overhead of transferring a 50-word block was not high.

Making best use of user demand

In many crowdsourcing applications the number of respondents using the system in any given time period can be difficult to predict. Too few users may mean that the required number of responses per token is not achieved, while conversely, too many users can rapidly exhaust the supply. While the former case may lead to insufficient statistical power in subsequent analyses, the latter represents a missed opportunity. A number of techniques can be used to address these issues. A low response rate can still result in valuable data if stimuli are rationed with the aim of maintaining a given number of responses per token. A lower limit on the number of tokens available at any instant might be based on the maximum expected number of tokens screened by a single individual, given that users should not in general hear the same token twice. A higher-than-expected usage can be accommodated either by dynamic generation of new stimuli to meet demand, or by overgeneration of tokens.

The BIGLISTEN application adopts a rationing approach. Blocks of stimuli progress through three states – 'unused', 'active', and 'exhausted'. At any time, a small number of blocks are active. When a block has been screened sufficiently, it is moved to the exhausted state and replaced by an unused block. Sufficiency of screening is defined in the BIGLISTEN based on reaching a criterion number (here set to 20) of 'high quality' listeners (the definition of high quality here is approximately the same as the 'subj' category described in section 1.5.6 below).

1.5.5 Respondents

Here we examine quantitative aspects of the BIGLISTEN experiment as well as the information provided by respondents themselves.

Raw response statistics

Two adverts placed 11 days apart via the University of Sheffield's internal announcement service (which has the potential to reach more than 20000 staff and students) led to 2120 respondents filling in the initial applet form within the first 20 days of the first advert. Of these, 1766 (83.4%) went on to complete the task (i.e., respond to at least one block of stimuli). Note that since respondents were not required to register to use the system, no user-tracking between page visits was possible, so what we call respondents here are actually separate page visits. Predictably, most of the activity occurred on the days of the adverts themselves, with a rapid decrease over time (see figure 1.3). Clearly, peaky demand is a consequence of the method used to garner interest in the web experiment. Ideally, publicity measures which produce a more uniform demand over time are preferred, although in this case the level of demand was not problematic for the tool.

Between them, respondents heard 157150 individual noisy tokens, corresponding to 3143 blocks, a mean of 1.78 blocks per respondent. Figure 1.4 demonstrates that while most listeners identified stimuli from a single block, a significant proportion went on to complete several blocks. The number of additional blocks screened gives an indication of how engaging the task was for listeners. An additional 0.78 blocks per listener perhaps suggests that while many respondents were curious enough to carry out the task once, most did not feel it sufficiently engaging to continue. Here, it seems likely that the relatively sparse feedback provided (essentially just a user ranking) and the lack of any reward – monetary or otherwise

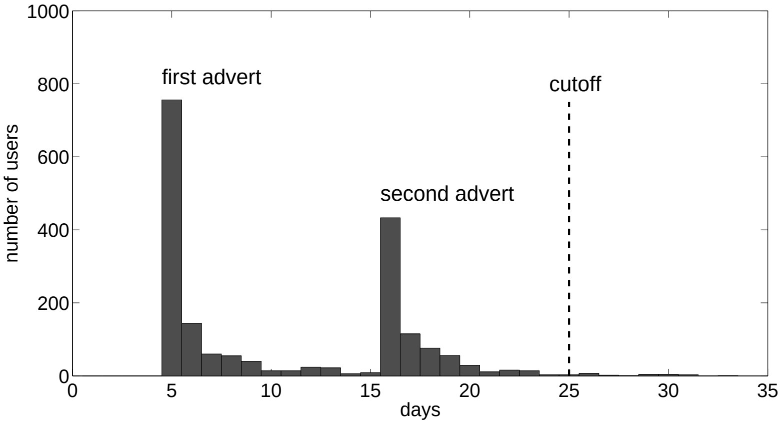


Figure 1.3 Number of responses per day.

– was responsible for the relatively low task engagement. In practice, task designers can use this kind of quantitative information to improve the web application.

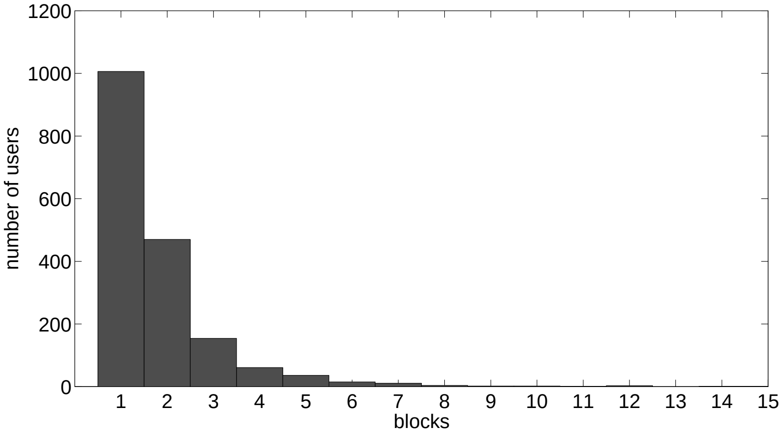


Figure 1.4 Number of stimulus blocks identified per respondent.

The mean response time per block was 155 s, i.e., just over 3 seconds per stimulus. Figure 1.5 shows the distribution of mean response times per stimulus.

Tables 1.1 to 1.4 summarise data supplied by respondents about their first language, accent, listening conditions and audio hardware respectively, while figure 1.6 plots their age distribution. In addition, 58 respondents (3.3%) reported some degree of hearing impairment. Figures are based on the 1766 respondents who completed at least one block.

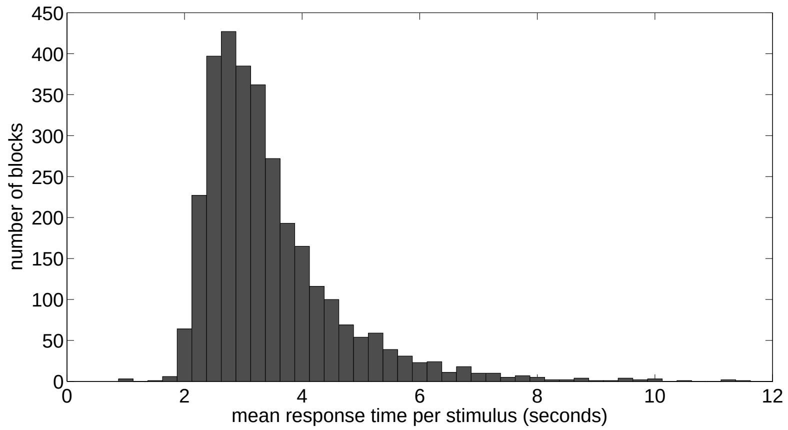


Figure 1.5 Response time.

Table 1.1 Respondents' self-reported first language

N	Percent	L1
1442	81.65	English
70	3.96	Chinese
39	2.21	German
31	1.76	Spanish
27	1.53	Bulgarian
19	1.08	Arabic
16	0.91	Hindi
11	0.62	Greek

Here and elsewhere categories with fewer than 10 respondents are omitted

First language (L1)

More than 4 out of every 5 respondents reported English as an L1, while the remaining native languages reflect the multilingual community typical of a UK university. While native English listeners were the target audience here, our experience with later versions of the BigListen application tested with large L2 populations suggests that robust confusions can also be harvested from non-native listeners, particularly from homogeneous samples such as advanced learner groups with the same L1. For L2 listeners confusions appear to be dominated by L1 influences rather than masking.

Table 1.2 Accents reported by respondents with English as L1

N	Percent	Accent
746	42.24	UK and Rep. Ireland
317	17.95	not supplied
265	15.01	Northern English
162	9.17	Southern
104	5.89	Midlands
55	3.11	Received Pronunciation
12	0.68	Scottish
10	0.57	West Country
10	0.57	Welsh
10	0.57	Northern Irish

Accent

Table 1.2 lists the dominant accents of English amongst respondents. Knowing a listener's linguistic origins within the native population can – in principle – help to make sense of their responses. One issue is the granularity at which to define accents. A detailed classification can lead to problems in finding an appropriate category for the many listeners who have moved around, producing the potential for confusion on the part of users as to the desired response. The problem is more acute for bilinguals or individuals with mixed accents. In the BIGLISTEN listeners could choose from 10 options within the UK, 7 each for Oceania and North America, and around 5 each for other English-speaking countries. A design decision was taken to also permit null responses, or one of several less-specific categories such as 'UK and Rep. Ireland', or 'General American'. The aim was to enable respondents to get through the questionnaire rapidly in order to encourage completion of the whole task. A better approach might be to forego self-classification of accent and instead to embed accent-diagnostic words within the main test, along the lines of SoundComparisons (2012). As we will see later, certain word confusions reveal something of the likely broad accent region of the listener and provide an indirect way to classify a respondent's accent.

Listening conditions

Crowdsourced listening tests will inevitably contain many responses from users listening under non-ideal acoustic conditions. This aspect of crowdsourcing is one of the most difficult to control (but see section 1.6 for some suggestions). Part of the problem stems from the robust nature of human speech perception: listeners are so capable of tracking a target source in the presence of reverberation or other sound sources that their tolerance for extraneous sound is high, and what is subjectively a quiet environment may well contain a significant level of noise. A very high proportion of respondents in the BIGLISTEN claimed to take the test in a quiet environment, a figure perhaps influenced by the availability of such spaces for a university population and not necessarily representative of a wider audience. On the other hand, the test itself demands a certain degree of quietness. We later introduce a method for selection of responses based on performance on near-universally correct stimuli, which can be expected to identify those respondents listening in reasonably quiet conditions.

Table 1.3 Respondents' listening conditions

N	Percent	Noise level
1541	87.26	low (e.g., quiet room)
207	11.72	moderate (e.g., shared office)
18	1.02	noisy (e.g., Internet cafe)

Table 1.4 Respondents' audio hardware

N	Percent	Audio delivery
815	46.15	headphones
577	32.67	external loudspeakers
374	21.18	laptop speakers

Audio hardware

Extraneous noise is attenuated by headphone listening. As for listening conditions, the fidelity of audio delivery is one area where a large amount of variability can be expected. Here, perhaps surprisingly, the majority of respondents did not use headphones but instead listened though external or laptop speakers, the latter in particular being clearly sub-optimal for speech in noise tasks.

Age

Due to factors such as the possibility of age-related hearing loss, knowing a respondent's age can be valuable for later subsetting or rejection of responses. Here, the age profile (figure 1.6) probably says more about that of the group who received the invitation to participate than it reveals of any age-related predilection for online tests. Note that the peak at age 30 stems from this being the default choice on the questionnaire, again resulting from a design decision to facilitate rapid test completion. In a large-scale test it would make more sense to force respondents to choose an age. Even so, it is interesting to observe that all but an estimated 4% of respondents did indeed go to the trouble of selecting an age rather than using the default.

1.5.6 Analysis of responses

In this section we examine the responses supplied by users of the BIGLISTEN and go on to compare them to those of a group tested using the same task and materials under traditional laboratory conditions (for details see Cooke 2009). Since not all blocks heard by the formal group were exhausted by the web group (in the sense defined in section 1.5.4 above), the following analysis is based on a subset of the exhausted web data, corresponding to material spoken by one of the male talkers in each of the 12 noise conditions.

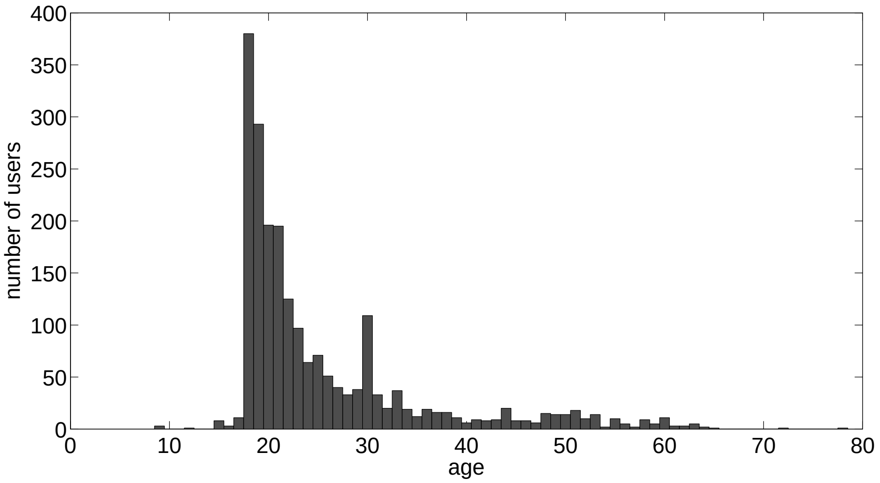


Figure 1.6 Distribution of respondents' self-reported ages.

Effect of self-reported factors on recognition rates

While the principal purpose of the BIGLISTEN web experiment is to discover interesting word confusions, most of the time in formal tests listeners reported the correct answer, so it is of interest to explore how the information supplied by respondents (e.g., first language, age) correlated with overall recognition scores. Figure 1.7 shows mean scores for each level of the factors gathered from participants.

This figure needs to be interpreted with some care. These are univariate scores i.e., computed over all other factors and thus it is important to note that control variables are not independent. For example, a correlation can be expected between those respondents who reported hearing impairment and those in the older age brackets. For a sufficiently large sample, a full conditional dependency analysis between factor levels could be carried out, but the relatively small scale of the current sample precludes this kind of analysis here. Also note that the distribution of respondents across levels for some of these factors is non-uniform. This caveat aside, we include the data to give some idea about the likely average effect of participant factors on performance.

Ambient noise in the test environment had a large effect, as did having a first language 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, although it is likely that listeners with moderate or severe HI either did not attempt the task or used a hearing aid.

While the ranking of levels within each factor is almost as expected (the exception being the poorer performance of external loudspeakers compared to internal loudspeakers), the cross-factor comparisons afforded by this type of plot are revealing. The difference between

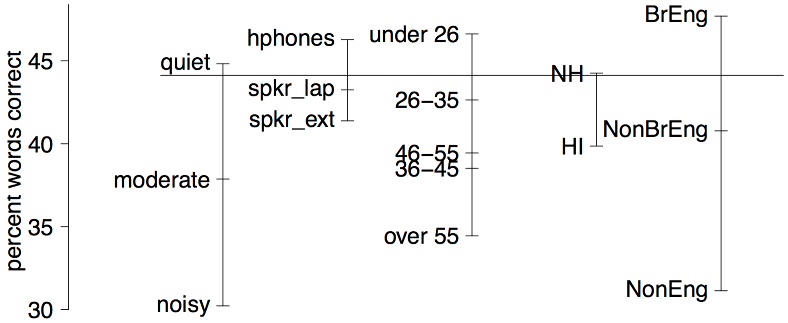


Figure 1.7 Mean word identification scores for each level of respondent-supplied factors. Figure reproduced from Cooke et al. (2011).

means for quiet and noisy conditions is of a similar size as the difference in performance between British English and non-native listeners. The benefit of headphone listening is, by comparison, not so large.

Web versus formal listening tests

The upper and lower boxplots of figure 1.8 depict word score statistics for the crowdsourced (WEB) and traditionally-tested (FORMAL) groups prior to any type of respondent-filtering. The intermediate boxplots (SUBJ, ANCHOR, SUBJ+ANCHOR) describe scores for subsets of web respondents selected on the basis of subjective and objective criteria defined below.

This figure demonstrates that mean scores obtained via unfiltered crowdsourcing are very significantly reduced – here, by well over 20 percentage points – compared to those obtained under traditional testing procedures. This outcome has been found in other web-based speech perception studies (e.g., Mayo et al. 2012; Wolters et al. 2010). For instance, in Mayo et al. (2012) MTurk listeners had an absolute performance level of around 75% of that measured in a traditionally-tested group.

Nevertheless, figure 1.8 suggests that individual web listeners are capable of high scores. Indeed, some WEB participant scores are higher than those obtained in the FORMAL group, although it should be noted that the latter employed far fewer participants (which also accounts for the wider confidence intervals for the FORMAL group).

Clearly, the WEB group includes data from respondents whose first language is not English, or who reported hearing-impairment, or might be expected to suffer from age-related hearing deficits, less-than-ideal listening conditions or audio delivery hardware. As a first post-filtering step, respondent-supplied criteria were used to select a subjectively-defined subset of web respondents (SUBJ). This subset contained only those respondents who satisfied *all* of the following criteria:

(i) listening in a quiet environment

(ii) audio delivery via headphones

(iii) British variety of English as first language

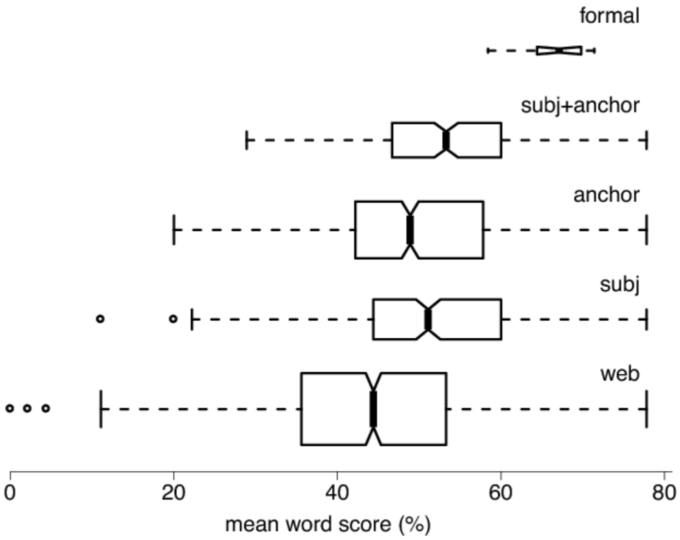


Figure 1.8 Boxplots of scores for formal and web groups. Lines extend to 1.5 times the inter-quartile range, circles indicate outliers, box thickness is proportional to the number of listeners in group and notches depict 95% confidence intervals. Figure from Cooke et al. (2011).

(iv) aged 50 or under
(v) no reported hearing problems

Around 31% of web listeners satisfied the intersection of these constraints. As anticipated, the mean score for this group (figure 1.8) is significantly higher [$p < 0.01$] than the unfiltered WEB group, although still far below the level of the FORMAL group.

Taking respondents' information at face value, subjectively-defined criteria go some way to matching conditions in traditional testing environments, where more control over the listener population can be exercised. However, they retain responses from those listeners who, for whatever reason, performed very poorly on the test compared to others in the cohort (see the outliers in figure 1.8). These listeners may have given up at some point during a block of stimuli and then entered arbitrary responses in order to receive feedback at the end of the test, for example. For this reason, it is useful to seek objective criteria to select well-motivated respondents. In the crowdsourcing scenario, one approach is to examine response consistency across listeners. In general many techniques are possible based on measuring the likelihood of a response sequence by comparing the response to each token with the distribution of responses from all other listeners who screened that token. In the BIGLISTEN we adopted an approach based on first identifying a type of criterion token (see section 1.3.2) that we call an 'anchor token' – an individual stimulus that satisfies the joint criteria of (i) having been screened by many listeners and (ii) having a very high rate of correct identification. Once anchor tokens are identified, they can be used to filter out those respondents who failed to reach a criterion score on these stimuli. Here, anchor tokens were defined as those stimuli heard by at least 30 listeners and which, as individual tokens, resulted in scores of at least 80% correct. Since not all listeners heard the same blocks of stimuli, different set of anchor tokens are used in each case. Fortunately, many anchor tokens meeting the above criteria were present in the response set.

Respondents who achieved mean scores of at least 90% on anchor tokens made up this objectively-defined ANCHOR subset. Around 63% of all web listeners met this rather strict criterion. Subjective and objective respondent filtering approaches can also be combined to produce a SUBJ+ANCHOR group. In this case, the dual criteria retained only 23% of web respondents.

The use of anchor tokens has the desired effect of removing outliers, and produces an increase in mean score, although by less than the application of subjective criteria [$p<0.05$]. Combination of the two criteria leads to higher scores, at the cost of removing more than 3 out of every 4 respondents from the analysis. However, a 13 percentage points gap still remains 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 (i.e., the set WEB−(SUBJ+ANCHOR)). For brevity, these web groups are denoted web+ and web-.

Score correlations across masker and SNR

The degree to which the different listener groups pattern in a similar way as a function of noise type and SNR is shown in figure 1.9. Each point represents responses from a single noise type in a narrow SNR range (quantised to 1 dB). The strong correlation that exists between formal and web scores suggests 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 leads to the different listener groups being affected to a very similar degree.

An even larger correlation of 0.96 in intelligibility scores across five different speech styles was reported in a comparison of MTurk and lab-tested listeners in Mayo et al. (2012), strengthening the view that even when absolute scores differ, the pattern of scores across conditions can be remarkably similar in web-based and formal speech perception tests.

Response consistency

Another way to measure similarity in responses is to look at the proportion of words where listeners reached a certain level of consensus in their decisions. Figure 1.10 shows how many words were identified correctly (upper panel) or misidentified, but in a consistent way (lower panel) as a function of the degree of agreement.

The rightmost bars in both plots depict a very strict level of agreement, with more than 90% of listeners providing the same response to a given stimulus. For the formal group, over 350 words were identified correctly on the basis of this criterion, with rather fewer for the web+ subset of crowdsourced listeners. Here, there is a clear difference between the web+ and web- groups, the latter showing far lower degrees of response consistency.

In the middle of the range, for 50% agreement upwards, we have the weaker criterion of *majority agreement*. For correct responses, the majority agreement levels are similar for

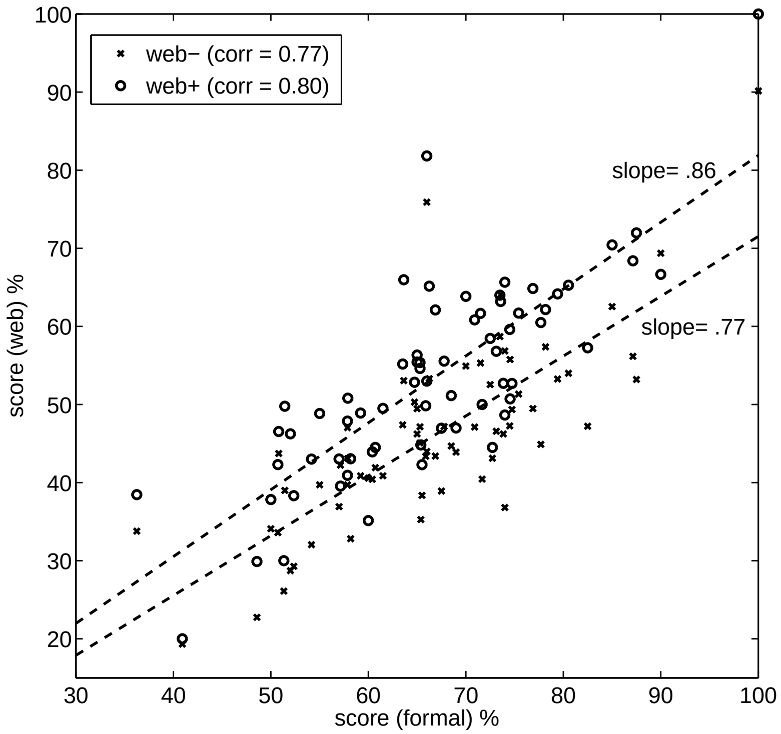


Figure 1.9 Mean scores in each masker and SNR condition for the formal and web groups. Figure from Cooke et al. (2011).

each group. However, for incorrect responses, where majority responses identify the robustly-perceived confusions that we are mainly interested in, the formal group shows a greater degree of consistency than the web+ group, while the web- group discovered relatively few consistent confusions. In fact, the formal group discovered 129 majority confusions compared to 85 and 44 respectively for the web+ and web- groups. This suggests that although the web-based procedure leads to lower overall scores, it is still effective in finding potentially-interesting word confusions in noise if both subjective and objective listener selection procedures are followed. Some of these confusions are shown in figure 1.11.

An unexpected outcome was the finding that the web+ group's majority confusions were not simply a subset of those discovered by formally-tested listeners. In fact, only 33 were common to both groups, while the remaining 96 from the formal group were not majority confusions for the web group. Intriguingly, the reverse was also the case: the web+ group crowdsourced 52 exemplars which were 'missed' by the formal group. The reasons for this finding are unclear. It is possible that the lower quality audio equipment likely to have been used by the web group led to consistent response biases. For instance, if significant high frequency attenuation was more likely to be present in the web group, confusion between certain fricatives might be more frequent, and may have led to a tendency to pattern in similar ways across the web group. This is an area which demands further investigation.

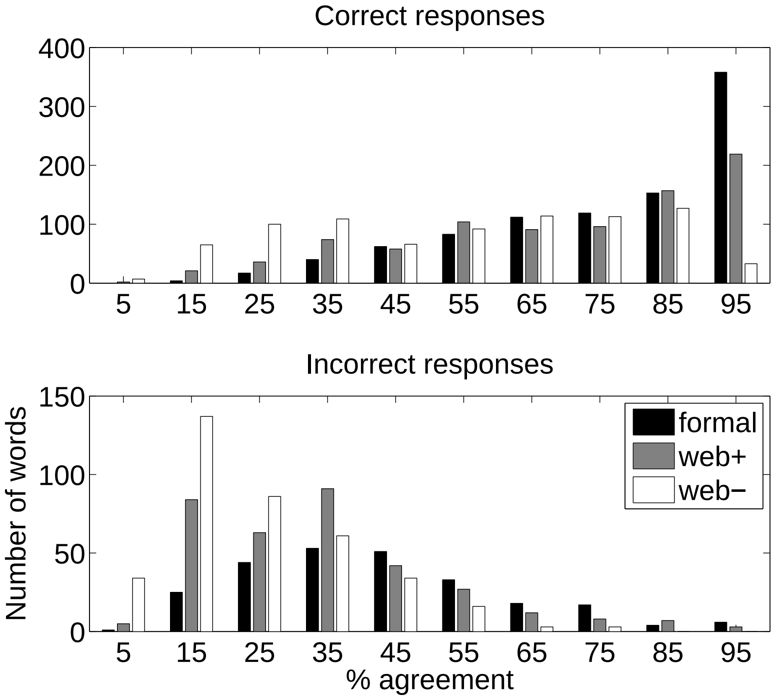


Figure 1.10 Agreement levels for correct and incorrect responses. Figure from Cooke et al. (2011).

Typical confusions

The relatively small scale of the BigListen experiment means there is insufficient data to support a comprehensive discussion of confusions. However, to date we highlight some tendencies we have observed in the data.

(i) Most confusions involve consonants rather than vowels (although the reverse was true in a subsequent unpublished study with non-native Spanish listeners). Most vowel confusions (mainly /ʌ/-/ɒ/) are likely to be caused by an accent mismatch between the speaker and listener.

(ii) Labial plosives and fricatives are often involved in onset confusions. Sometimes the confusions are inter-labial (/f/ to /p/ or /b/) involving fricative/plosive errors (Hazan and Simpson 1998), but we often observe a labial to /h/ confusion, which highlights the lack of salience of the labial gesture in acoustic/perceptual terms.

(iii) Nasals are frequently substituted or deleted, especially in coda position (Benki 2003).

(iv) Some confusions involve consonant insertion in both coda and onset position, perhaps due to incorporation of background energy fragments (e.g., ‘pea-peace’).

(v) Other confusions suggest an effect of word familiarity (e.g., ‘veil-fail’, ‘whirl-world’).

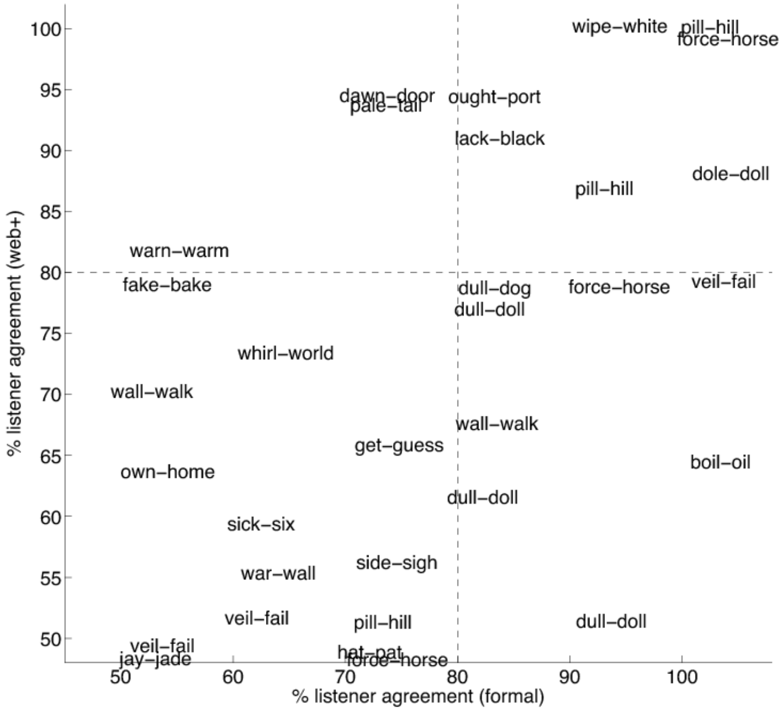


Figure 1.11 Majority confusions in common for formal and web+ listeners. Dotted lines show 80% agreement levels. Figure from Cooke et al. (2011).

Example response distributions

We end this case study with a look at some of the response distributions to individual noisy speech tokens. Each panel of figure 1.12 plots the number of times a given word was reported in response to the presented word and noise type indicated. To keep the response distributions manageable and relevant, only those responses which were reported by at least three listeners are retained. These examples have been chosen both to illustrate facets of the task and to highlight some of the issues that need to be considered when using crowdsourcing to gather responses in speech perception tasks. While we present some conjectures, the underlying mechanisms which create the response patterns are still far from understood.

(i) **‘Doll’ in 4-talker babble-modulated noise (BMN)**. This is a classic case of a very robust confusion with a high degree of listener agreement. Respondents identifying this stimulus as ‘dog’ outnumbered those reporting the correct answer by 6-to-1 here. It is possible that energetic masking of the final consonant followed by misallocation of a suitable brief noise burst from the background masker was responsible for this confusion. The vowel in this and many other examples was correctly reported. As noted above, vowels tend to be robust and survive masking at the SNRs used in BIGLISTEN.

(ii) **‘Heap’ in 2-talker babble**. Complete background words typically remain audible in two-talker maskers and in this case nearly all listeners identified the word ‘middle’

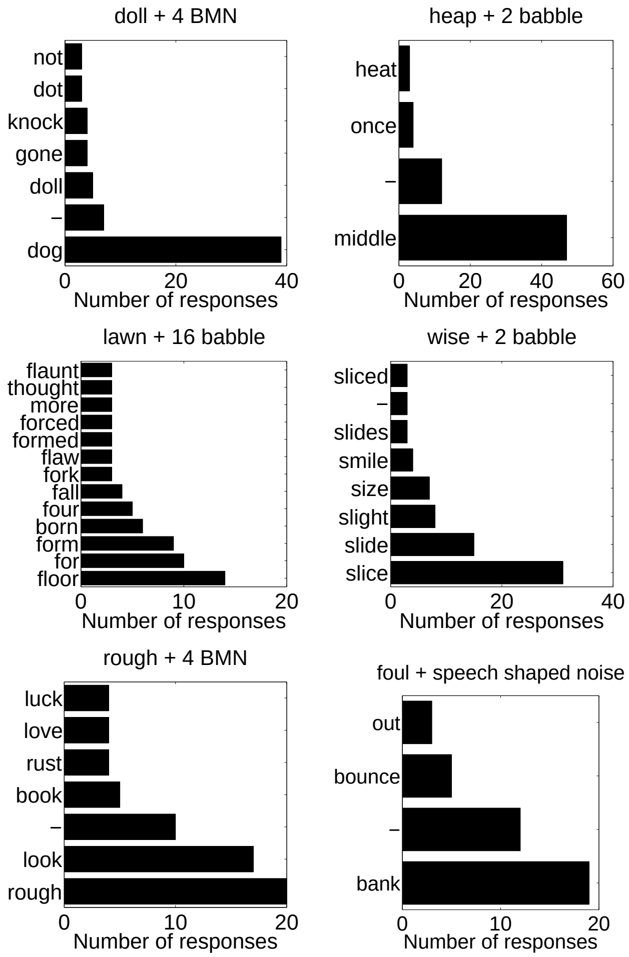


Figure 1.12 *Example response distributions. A hyphen indicates a null response. The stimulus in each case is shown as the combination of word and noise type. 2/16 babble indicates natural babble made up of 2 or 16 voices. 4 BMN is speech-shaped noise with envelope modulations from 4-talker babble.*

instead of the target ‘heap’. What is surprising about this example is that a 2-syllable word was reported in spite of listeners receiving instruction that all target words were monosyllabic. This highlights a methodological difference between crowdsourcing and traditional testing: there is no guarantee that participants bother to read the instructions, and there is less opportunity to emphasise experimental factors such as this compared to a formal testing situation where the experimenter is physically present. If necessary, immediate and automatic feedback could be provided to correct the misunderstanding.

(iii) **‘Lawn’ in 16-talker babble**. We speculate that high-entropy confusions like this are symptomatic of energetic masking, where parts of the target word are swamped by noise. How listeners fill in the inaudible segments will depend on respondent-specific 'language model' factors, leading to some variety in responses. However, this example illustrates that useful information can be obtained even in the case of a relatively high-entropy response distribution. While no respondents reported the word actually present in the noisy stimulus, nearly all agreed on the attachment of a word-initial fricative /f/, presumably recruited from the masker. Again, the vowel was reported correctly in all cases. Note also the need to handle homonyms (e.g., 'floor', 'flaw') in open-response tasks of this sort.

(iv) **'Wise' in 2-talker babble**. This is a similar example to 'lawn': a surviving target diphthong surrounded by a largely consistent initial consonant cluster and some variation in coda consonant. The different coda consonants presumably reflect both non-uniform stimulus ambiguities which favour some interpretations over others, as well as lexical constraints. It is worth noting that while mis-spellings were present (though infrequent), respondents in the main did not invent words, i.e., use nonwords to identify their response. The lesson here is that, unlike some tasks in speech perception (such as those mentioned in section 1.4.4), a task demanding words as responses is highly-appropriate for naïve listeners.

(v) **'Rough' in 4-talker BMN**. What is interesting about this example is not the fact that a small majority reported the correct response, but that some of the incorrect responses reveal something about the accent or linguistic environment of the respondents. Within the UK, regional variation in pronunciation of words is rife, with words such as 'look' and 'book' being produced with either /u:/ or /u/. In principle, these diagnostic responses might be used to corroborate respondent-supplied information on accent. This example also illustrates that homonym handling needs to be sensitive to accent e.g., 'look' and 'luck' are not homonyms for all listeners.

(vi) **'Foul' in speech-shaped noise**. Here, no listener reported the correct answer and many produced a null response, but there was enough evidence in the noisy stimulus for 19 listeners to report the word 'bank'. This is an interesting case, because the phonological transformation from 'foul' to 'bank' is not at all obvious (to say the least!) and yet the background noise type is supposedly uninformative (it does not contain speech), nor was it temporally-modulated. This concluding example demonstrates one of the primary benefits of carrying out speech perception tasks with open response sets and large numbers of listeners, viz. the emergence of intriguing and unexpected outcomes.

1.5.7 Lessons from the BIGLISTEN *crowdsourcing test*

While small-scale in nature, the BIGLISTEN experiment suggests that crowdsourcing is capable of eliciting response distributions which are of potential interest in speech perception studies. Quantitative estimates (e.g., from figure 1.10) of the rate at which even formally-tested groups make consistent mis-identifications of noisy stimuli indicate that robust confusions are rare, and motivates the use of crowdsourcing as an initial sieve prior to formal confirmation tests.

BIGLISTEN also demonstrates that both respondent-provided information and internally-generated anchor tokens can contribute to the selection of listeners who better match the levels of homogeneity and motivation which we aim for in laboratory-based tests. Nevertheless, coherent subsets of web respondents never matched scores seen in the laboratory. This finding echoes other studies of crowdsourcing with speech and/or noise stimuli (Mayo et al. 2012; Wolters et al. 2010). Clearly, crowdsourcing in speech perception is not suitable for those tasks which seek to estimate absolute performance levels of listener samples. The reasons for this discrepancy have yet to be pinpointed – in itself not an easy task – but seem likely to include differences in the overall audio delivery path from a client's computer to their auditory system: digital-to-analogue conversion, amplification, connectors, leads and headphones are all candidates for signal degeneration relative to a typical speech perception laboratory setup.

The BIGLISTEN benefitted from a surprisingly high rate of voluntary participation, estimated at around 7-10% of all those receiving one of two email invitations. While careful timing of the invitations, just following the annual influx of new students, no doubt contributed to this level of involvement, it is also possible that the promise of a rapid, hassle-free and anonymous experiment requiring no user registration appealed to many respondents. Designing a test which could generate useful data with an end-to-end time of under three minutes per user was a primary design goal, even at the expense of permitting null responses in the elicitation of user data (e.g., default values for age and accent). In hindsight, allowing default responses is not to be recommended as best practice due to its potential to invalidate sample-wide estimates of the desired factor.

One of the advantages of a large-scale listening test with a relatively unconstrained response set is the possibility of finding unforeseen yet robust responses with non-trivial explanations. For example, the BIGLISTEN has, for us, motivated a change in the way we think about the effect of noise on speech, with the notion of masking giving way to a more complex sequence of speech-noise 'reactions' which result in a given word confusion. The lesson here is that while it is possible in principle to find similar outcomes with traditional test procedures, the use of large and somewhat uncontrolled samples seems to encourage unexpected outcomes. Control of everything that can be controlled, from participants to instructions, is the official ethos in most formal tests (although it need not be), but may well be counter-productive in tasks which seek to discover 'interesting' specimens.

The finding that formal and web tests differ not only in absolute scores but also in the patterns of majority confusions suggests that additional care needs to be exercised in preparing for a web experiment. One implication is that pilots carried out in a formal setting may give a biased picture of what can be expected in a crowdsourced test.

1.6 Issues for further exploration

Further and more extensive use of crowdsourcing in speech perception seems inevitable. Some of the driving forces for greater use of non-formal testing procedures include: the increasing use of spoken language output technology which calls for large-scale comparative evaluations, for which crowdsourcing enables ranking of systems; the online delivery of simple hearing tests; and the need for more speech perception tests to better understand hearing and to develop more robust speech technology. Here, we raise some of the issues in crowdsourced speech perception that deserve further study, and highlight some technological developments which might enable better control of web experiments.

Matching traditional levels of performance

Currently, as we have seen, the absolute level of performance in web-based speech perception tests falls short of that obtained in formal settings, perhaps restricting the use of this methodology to crowd-as-filter approaches and rank ordering of conditions, assuming in addition that appropriate formal validations are carried out. What can be done to raise the performance baseline? Here are six areas to focus on:

(i) Better listener selection procedures. Pre-tests, using criterion tokens, might help to select listeners suited to the target language, for example.

(ii) Automatic determination of audio delivery hardware.

(iii) Automatic sampling of a listener's acoustic environment. While the technology already exists to make client-side estimates of, for example, background noise spectrum and level, its use raises important privacy concerns and could only be employed based on informed consent.

(iv) Improved procedures for task explanation, including mechanisms to check for correct interpretation. A short instructional video could better simulate the oral interaction typical of a lab-based experiment.

(v) Improving respondent motivation. Many opportunities exist to incorporate the collection of speech-based judgements into more entertaining applications. The provision of timely and relevant feedback is an additional facet of motivation.

(vi) Options to cope with client-side disruptions during the task. A participant with the best will in the world will find it more difficult to prevent disruptions – caused by such things as visitors or telephone calls – than an experiment under laboratory control. While response time monitoring is a passive means to identify disruption, an approach which allows participants to signal 'unreliable' trials would permit better identification of reliable data.

Decreasing variability

A key issue is how to reduce response variability, which has many of the same origins as those speculated to cause lower absolute performance – listeners, equipment, environment. Targeted advertising in special-interest communities or forums might lead to increased listener homogeneity, if this were a desirable outcome in any given web experiment, at the expense of a reduced rate of participation. More stringent respondent questionnaires are likely to produce the same tradeoff.

Equipment variability is one area which should be more easily controlled in the future. Experiments can be aimed at users of specific devices whose audio characteristics are well-understood. For more limited sample sizes (perhaps involving longer or more intensive testing), headphones could be mailed out as a gift to participants, providing an incentive to participate. Introduction of a gaming/competitive element may motivate certain types of user to undertake the test using the best equipment at their disposal.

While the BIGLISTEN made no use of IP addresses (e.g., to estimate participant location/language, or cross-session tracking), this information could be employed to increase the likelihood that different sets of responses originate from different individuals.

One source of variability in applications like the BIGLISTEN which have open-set response alternatives stems from user input errors e.g., typos and mis-spellings. If the user responds with another valid word, little can be done. Otherwise, the participant could be passively alerted to the possibility of an an input error, perhaps adopting the commonly-used method of input underlining. Handling input errors is best done on the client side and is likely to become easier to integrate into crowdsourcing in the future e.g., through the use of spellchecked forms in HTML5.

Ethics and safety concerns

Two related issue we have barely touched on are the ethical and safety dimensions of crowdsourcing in speech perception. The BIGLISTEN required explicit consent to be given before commencement of the main test, but it is not clear that this will be sufficient in all tasks or jurisdictions. Ethical and especially safety concerns involve many distinct questions, some of which have been covered in other chapters and are common to the domain of speech perception. We focus on those most relevant to the speech domain here.

First, there is the issue of possible temporary or permanent hearing damage caused by the delivery of intense stimuli. Here, we would suggest that while there are numerous examples of web-based audio delivery (e.g., online videos or music samples), and that there is very little that can be done to control the final sound intensity level at which stimuli are reproduced, deployment of crowdsourcing in speech research requires high standards to minimise user risks. Techniques include: issuing warnings about setting the output level via examples prior to reaching the main test; requiring a user to correctly-identify practice examples which are chosen to distort at high volume levels; monitoring performance in the main test and curtailing the experiment if a performance threshold on easily-recognised tokens is not reached; preventing overlong exposure to the experiment by fixing a maximum number of repeated listens from a given IP address in a fixed time period; ensuring that output levels are fixed across stimuli and tested at high volume settings on commonly-used computer hardware.

Second, detailed questionnaires, particularly those permitting complete linguistic histories which might be solicited in speech and hearing studies, should not compromise user anonymity where this has been promised. This concern applies most acutely for smaller samples that might result from targeted recruitment.

Third, feedback should be relevant, accurate and useful. In tests involving the perception of speech signals it is essential to make clear to respondents that they are not undertaking an online hearing test, and to stress in any feedback given that the results cannot be interpreted in ways which relate to their individual hearing sensitivity. The provision of useful feedback needs careful consideration in applications such as the BIGLISTEN which actively seek confusions and typically lead to low scores from listeners who are performing quite normally. Here, other feedback metrics might be required, such as the degree of listener consistency rather than raw accuracy.

1.7 Conclusions

- Crowdsourcing in speech perception can be a valuable adjunct to traditional testing methods.

- For tasks such as those which require calibration of presentation levels, or involve the reporting of fine distinctions or estimates of absolute levels of intelligibility, traditional tests remain the method of choice.
- For evaluative tasks such as accent judgements or speech synthesis quality assessment, where ranking of alternatives is the desired outcome, web-based testing is an option that merits consideration.
- In domains where the availability of a large listener sample is an essential element of experimental design, crowdsourcing may be the only practical approach.
- Further studies are required to validate the application in any new task or domain and in particular to test for the existence of consistent biases in responses from the crowd.
- Methodological innovations will be needed to enable objective confirmation of subjective wisdom.

References

Andersson L and Trudgill P 1990 *Bad Language*. Basil Blackwell, Oxford.

ANSI 1997 *S3.5-1997: American National Standard Methods for Calculation of the Speech Intelligibility Index*. American National Standards Institute, New York.

Beddor PS and Gottfried TL 1995 Methodological issues in cross-language speech perception research with adults In *Cross-language Studies of Speech Perception: Issues in Cross-language Research* (ed. Strange W) York Press.

Benki J 2003 Analysis of English nonsense syllable recognition in noise. *Phonetica* **60**, 129–157.

Bexelius C, Honeth L, Ekman A, Eriksson M, Sandin S, Bagger-Sjoback D and Litton J 2008 Evaluation of an Internet-based hearing test: comparison with established methods for detection of hearing loss. *Journal of Medical Internet Research*.

Bissiri MP, Garcia Lecumberri ML, Cooke M and Volín J 2011 The role of word-initial glottal stops in recognizing English words *Interspeech*, pp. 165–168.

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

Bongaerts T 1999 Ultimate attainment in L2 pronunciation: The case of very advanced late L2 learners In *Second Language Acquisition and the Critical Period Hypothesis* (ed. Birdsong D) Lawrence Erlbaum.

Boyd S 2003 Foreign-born teachers in the multilingual classroom in Sweden: The role of attitudes to foreign accent. *International Journal of Bilingual Education and Bilingualism* **3-4**, 283–295.

Brennan EM and Brennan JS 1981 Accent scaling and language attitudes: reactions to Mexican American English speech. *Language and Speech* **3**, 207–221.

Cara BD and Goswami U 2002 Similarity relations among spoken words: The special status of rimes in English. *Behavior Research Methods, Instruments, and Computers* **34**, 416–423.

Choi J, Lee H, Park C, Oh S and Park K 2007 PC-based tele-audiometry. *Telemedicine and e-Health* **13**, 501–508.

Christiansen C, Pedersen MS and Dau T 2010 Prediction of speech intelligibility based on an auditory preprocessing model. *Speech Communication* **52**, 678–692.

Clopper CG and Pisoni DB 2005 Perception of dialect variation In *The Handbook of Speech Perception* (ed. Pisoni DB and Remez RE) Blackwell.

Cooke M 2006 A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America* **119**, 1562–1573.

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

Cooke M, Barker J, Garcia Lecumberri M and Wasilewski K 2011 Crowdsourcing for word recognition in noise *Proc. Interspeech*, pp. 3049–3052.

Cooke M, Garcia Lecumberri M, Scharenborg O and van Dommelen W 2010 Language-independent processing in speech perception: identification of English intervocalic consonants by speakers of eight European languages. *Speech Communication* **52**, 954–967.

Cox T 2008 The effect of visual stimuli on the horribleness of awful sounds. *Applied Acoustics* **69**, 691–703.

Dávila A, Bohara A and Saenz R 1993 Accent penalties and the earnings of Mexican Americans. *Social Science Quarterly* **74**, 902–915.

Draxler C 2011 Percy: An HTML5 framework for media rich Web experiments on mobile devices *Proc. Interspeech*, pp. 3339–3340.

Fayer JM and Krasinski E 1987 Native and non-native judgments of intelligibility and irritation. *Language Learning* **37**, 313–327.

Fernandez Gonzalez J 1988 Reflections on foreign accent *Issues in Second Language Acquisition and Learning* Universitat de Valencia Valencia.

Festen J and Plomp R 1990 Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America* **88**, 1725–1736.

Floccia C, Goslin J, Girard F and Konopczynski G 2006 Does a regional accent perturb speech processing?. *Journal of Experimental Psychology: Human Perception and Performance* **5**, 1276–1293.

Garcia Lecumberri ML, Cooke M and Cutler A 2010 Non-native speech perception in adverse conditions: a review. *Speech Communication* **52**, 864–886.

Garcia Lecumberri ML, Cooke M, Cutugno F, Giurgiu M, Meyer B, Scharenborg O, van Dommelen W and Volin J 2008 The non-native consonant challenge for European languages *Interspeech*, pp. 1781–1784.

Gong J, Cooke M and Garcia Lecumberri ML 2011 Towards a quantitative model of Mandarin Chinese perception of English consonants In *Achievements and Perspectives in SLA of Speech* (ed. Wrembel M, Kul M and Dziubalska-Kowaczyk K) Peter Lang.

Hazan V and Simpson A 1998 The effect of cue-enhancement on the intelligibility of nonsense word and sentence materials presented in noise. *Speech Communication* **24**, 211–226.

Honing H 2006 Evidence for tempo-specific timing in music using a Web-based experimental setup. *Journal of Experimental Psychology* **32**, 780–786.

Honing H and Ladinig O 2008 The potential of the Internet for music perception research: A comment on lab-based versus Web-based studies. *Empirical Musicology Review* **3**, 4–7.

Honing H and Reips U 2008 Web-based versus lab-based studies: A response to Kendall (2008). *Empirical Musicology Review* **3**, 73–77.

Horswill M and Coster M 2001 User-controlled photographic animations, photograph-based questions, and questionnaires: Three Internet-based instruments for measuring drivers risk-taking behavior. *Behavior Research Methods* **33**, 46–58.

Hosoda M, Stone-Romero E and Walter J 2007 Listeners' cognitive and affective reactions to English speakers with standard American English and Asian accents. *Perceptual and Motor Skills* **1**, 307–326.

Keller F, Corley M, Corley S, Konieczny L and Todirascu A 1998 WebExp. Technical report, HCRC/TR-99, Human Communication Research Centre, University of Edinburgh.

Keller F, Gunasekharan S, Mayo N and Corley M 2009 Timing accuracy of Web experiments: A case study using the WebExp software package. *Behavior Research Methods* **41**, 1–12.

Kendall R 2008 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.

King S and Karaiskos V 2010 The Blizzard Challenge 2010 *Blizzard Challenge workshop*.

Kuhl PK 1991 Human adults and human infants show 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not. *Perception and Psychophysics* **50**, 93–107.

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

Lacherez P 2008 The internal validity of Web-based studies. *Empirical Musicology Review* **3**, 161–162.

Laugwitz B 2001 A Web-experiment on colour harmony principles applied to computer user interface design In *Dimensions of Internet Science* (ed. Reips UD and Bosnjak M) Pabst Science Publishers Lengerich, Germany pp. 131–145.

MacKay IRA, Flege JE and Imai S 2006 Evaluating the effects of chronological age and sentence duration on degree of perceived foreign accent. *Applied Psycholinguistics* **27**, 157–183.

Major RC 2007 Identifying a foreign accent in an unfamiliar language. *Studies in Second Language Acquisition* **29**, 539–556.

Mayo C, Aubanel V and Cooke M 2012 Effect of prosody changes on speech intelligibility *Proc. Interspeech*.

McGraw I, Gruenstein A and Sutherland A 2009 A self-labeling speech corpus: Collecting spoken words with an online educational game *Tenth Annual Conference of the International Speech Communication Association*.

McGraw K, Tew M and Williams J 2000 The integrity of Web-delivered experiments: Can you trust the data?. *Psychological Science* **11**, 502.

Munro M and Derwing TM 1995 Processing time, accent, and comprehensibility in the perception of native and foreign-accented speech. *Language and Speech* **38**, 289–306.

Naumann A, Brunstein A and Krems J 2007 DEWEX: A system for designing and conducting Web-based experiments. *Behavior Research Methods* **39**, 248–258.

Newman C, Weinstein B, Jacobson G and Hug G 1990 The hearing handicap inventory for adults: psychometric adequacy and audiometric correlates. *Ear and Hearing* **11**, 430.

Novotney S and Callison-Burch C 2010 Cheap, fast and good enough: Automatic speech recognition with non-expert transcription *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 207–215 Association for Computational Linguistics.

Reips U and Neuhaus C 2002 WEXTOR: A Web-based tool for generating and visualizing experimental designs and procedures. *Behavior Research Methods* **34**, 234–240.

Reips UD 2000 The Web experiment method: Advantages, disadvantages, and solutions. *Psychological Experiments on the Internet* pp. 89–114.

Reips UD 2002 Standards for Internet-based experimenting. *Experimental Psychology (formerly Zeitschrift fur Experimentelle Psychologie)* **49**, 243–256.

Riney T and Takagi, N.and Inutsuka K 2005 Phonetic parameters and perceptual judgments of accent in English by American and Japanese listeners. *TESOL Quarterly* **39**, 441–466.

Rix A, Beerends J, Hollier M and Hekstra A 2001 Perceptual evaluation of speech quality (PESQ) - a new method for speech quality assessment of telephone networks and codecs *Proc. ICASSP*, vol. 2, pp. 749–752.

Sawusch J 1996 Instrumentation and methodology for the study of speech perception In *Principles of Experimental Phonetics* (ed. Lass N) Mosby St. Louis, MO pp. 525–550.

Seren E 2009 Web-based hearing screening test. *Telemedicine and e-Health* **15**, 678–681.

Skitka L and Sargis E 2006 The Internet as psychological laboratory. *Annual Review of Psychology* **57**, 529–555.

Snow R, O'Connor B, Jurafsky D and Ng A 2008 Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 254–263 Association for Computational Linguistics.

SoundComparisons 2012 www.soundcomparisons.com.

Swanepoel W, Clark J, Koekemoer D, Hall JI, Krumm M, Ferrari D, McPherson B, Olusanya B, Mars M, Russo I and Barajas J 2010 Telehealth in audiology: The need and potential to reach underserved communities. *International Journal of Audiology* **49**, 195–202.

Trudgill P and Hannah J 2008 *International English. A Guide to the Varieties of Standard English (5th edition)*. Hodder Education, London.

Van Els T and De Bot K 1987 The role of intonation in foreign accent. *The Modern Language Journal* **71**, 147–155.

Voxforge 2012 http://www.voxforge.org/.

Wells JC 1982 *Accents of English I: An Introduction*. Cambridge University Press, Cambridge.

Wolfe C and Reyna V 2002 Using NetCloak to develop server-side Web-based experiments without writing CGI programs. *Behavior Research Methods* **34**, 204–207.

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

INDEX

This is a Book Title Name of the Author/Editor