The following article appeared in **J. Acoust. Soc. Am**. *150, 1390–1401 (2021)* and may be found at https://doi.org/10.1121/10.0005880

How reliable are online speech intelligibility studies with known listener cohorts?

Martin Cooke[1,a)] and Maria Luisa Garcıa Lecumberri[2]

[1]*Ikerbasque (Basque Science Foundation), Maria Diaz de Haro 3, 6, Bilbao, 48013, Spain*

[2]*University of the Basque Country, Spain*

ABSTRACT:

Although the use of nontraditional settings for speech perception experiments is growing, there have been few controlled comparisons of online and laboratory modalities in the context of speech intelligibility. The current study compares outcomes from three web-based replications of recent laboratory studies involving distorted, masked, filtered, and enhanced speech, amounting to 40 separate conditions. Rather than relying on unrestricted crowdsourcing, this study made use of participants from the population that would normally volunteer to take part physically in laboratory experiments. In sentence transcription tasks, the web cohort produced intelligibility scores 3–6 percentage points lower than their laboratory counterparts, and test modality interacted with experimental condition. These disparities and interactions largely disappeared after the exclusion of those web listeners who self-reported the use of low quality headphones, and the remaining listener cohort was also able to replicate key outcomes of each of the three laboratory studies. The laboratory and web modalities produced similar measures of experimental efficiency based on listener variability, response errors, and outlier counts. These findings suggest that the combination of known listener cohorts and moderate headphone quality provides a feasible alternative to traditional laboratory intel- ligibility studies. V

I. INTRODUCTION

Although web experiments involving auditory judgments in general, and speech in particular, have been in use for some time, the ongoing COVID-19 pandemic has added urgency to the use of nontraditional settings for the gathering of experimental data. Online delivery of audio for experiments involving speech signals that have been degraded by masking or distortion presents challenges not seen in other applications of online testing, and a key outstanding issue concerns the reliability of information gleaned from speech perception experiments performed via the web, where reliability here encompasses both the ability to match absolute scores and replicate key outcomes in labo-ratory studies. This article describes findings from web- based replications of three laboratory studies that measured speech intelligibility in a diverse range of processing conditions.

Many studies have collected auditory-based responses outside the laboratory. Early web audio experiments are reviewed by Cooke *et al.* (2013) and variously involved making holistic judgments about the unpleasantness of sounds (Cox, 2008), rating song likeability (Beasley and Chuang, 2008), assessing the quality and intelligibility of speech synthesis (Blin *et al.*, 2008; Wolters *et al.*, 2010), and voice quality in synthetic speech (Parson *et al.*, 2013). Web-based audio has also been used in audiometric applications (e.g., Bexelius *et al.*, 2008; Choi *et al.*, 2007; Seren, 2009) and the sourcing of prosodic annotations (Evanini and Zechner, 2011). More recently, the web modality has been applied in such areas as consonant identification (Schwartz and Aperliński, 2014), rating disordered child speech (McAllister Byun *et al.*, 2015), rating and transcribing dysarthric speech (Jiao *et al.*, 2019) and dysrhythmic speech (Borrie, 2018), prosodic annotation (Cole *et al.*, 2017), eliciting subjective judgments of intelligibility (Yoho *et al.*, 2019), and assessing speech quality in noise (Naderi and Möller, 2020; Zequeira Jiménez *et al.*, 2018).

The focus of this study is on speech perception experiments in which listeners are asked to transcribe words in sentences. Intelligibility studies typically involve speech whose clarity has been compromised by distortion, masking noise, or filtering, among other forms of degradation. Consequently, online measurement of intelligibility bringsits own specific technical challenges such as preservation of signal fidelity and consideration of the listening environment (cf. Jiménez *et al.*, 2020). Several recent studies have used web listening in designs which involved measurement of intelligibility, including speech-in-noise tasks. Burgos *et al.* (2015) used a web platform to obtain transcriptions of Dutch vowels spoken by Spanish learners, while Vaughn (2019) asked online listeners to transcribe Spanish-accented English in noise. Melguy and Johnson (2021) studied adaptation to Chinese-accentd English speech by asking listen- ers to transcribe seences presented in multi-talker babble.

Adaptation to distorted speech was the focus of a crowdsourced study by Van Hedger *et al.* (2019), while Yoho and Borrie (2018) asked web participants to measure the intelligibility of control and disordered speech mixed with stationary noise.

Fewer studies have compared the laboratory and web modalities on the same intelligibility-based task. In the first investigation of its kind, Wolters *et al.* (2010) found that web participants transcribing semantically unpredictable synthesised sentences in quiet had word error rates of 20% compared to 13% for those undertaking the task under laboratory conditions. Cooke *et al.* (2011) asked listeners to identify monosyllabic English words in 12 different types of masking noise at a range of signal-to-noise ratios (SNRs). Compared to their laboratory counterparts, word identification rates for web listeners were 13 percentage points (pp) lower even after the application of stringent selection criteria to the web cohort. Mayo *et al.* (2012) compared laboratory and web performances for sentences produced in plain, infant-directed, computer-directed, foreigner-directed, and shouted speech styles, finding significantly higher word error rates in all five styles for the web cohort, with dispar- ities ranging from 15 to 27 pp. Slote and Strand (2016)reported a 14 pp advantage for laboratory listeners in a spo- ken word identification task. The common thread that links these replications is the existence of an absolute penalty for intelligibility tasks performed outside of the speech perception laboratory.

Table I identifies some of the factors that have the potential to influence laboratory and web experiments. One of the principal aspects differentiating the modalities stems from their disparate participant cohorts. Many web-based studies make use of crowdsourcing platforms to recruit participants, citing rapid access to a large sample as their main motivation. An alternative—one that is particularly relevant when access to laboratory facilities is restricted—is to make use of a sample of "known" listeners drawn from the same population as those who typically volunteer to take part in laboratory experiments.[1] This approach is not as common as crowdsourcing but has been employed in previous web studies, for example, via university participants pools (e.g., Cooke *et al.*, 2011; Gould *et al.*, 2015). The principal advan-tage of a known listener sample is the ability to more closelymatch the overall listener profile of a laboratory cohort. As a consequence, a replication involving known listeners can be expected to highlight differences that are mainly due to the sorts of non-participant related aspects of web experiments listed in Table I. This is the approach taken in the current study.

TABLE I. Some of the factors that might lead to outcome differences in laboratory and web speech perception experiments.

Participants	Technology	Environment	Procedure
Age	Headphones	Noise	Instructions
Hearing	Soundcard	Distractions	Questions
Motivation	Sound encoding	Interruptions	Familiarity
Education	Connectivity		Monitoring
Native language	Browser		
Other languages			

Two questions motivated this investigation. The first concerns the scale of any disparity between a known web listener cohort and a traditional laboratory group on tasks involving the transcription of speech that has been degraded by masking, distortion, or filtering. Although the main focus of this question is on intelligibility, it is also of interest to compare the amount of within-cohort variability to assess whether web experiments have a similar statistical power to those performed in the laboratory. As a supplementary question, we were also interested in comparing the quality of responses (e.g., in terms of the frequency of non-lexicalitems in user responses) and the number of outliers from each modality.

The second issue addressed here concerns how wellweb studies match the pattern of results observed in their laboratory counterparts. Even if web cohorts fail to match scores obtained in the laboratory, online experiments maybe able to replicate critical outcomes. For example, whereas the online listeners in Slote and Strand (2016) fell short of their laboratory counterparts in terms of absolute levels of accuracy, word identification scores were highly correlated across the two modalities (e.g., words that were difficult in the laboratory also tended to be difficult to identify online). Other studies have observed an interaction between test modality and experimental conditions. Wolters *et al.* (2010) found that speech generated using diphone synthesis led to proportionally more speech transcription errors than speech generated using Hidden Markov Model synthesis, for crowdsourced listeners compared to those in the laboratory. Jiménez *et al.* (2020) used a traditional laboratory cohort alongside simulated crowdsourcing to investigate the effects of two types of common environmental noises (street noise and TV show noise) on the rating of speech quality in 15 speech degradation conditions, finding that ratings depended on the noise type, and the impact of the masker varied across the degradation conditions. Knowing how a certain type of stimulus behaves in an experiment performed via the web is critical in interpreting the outcomes of web studies involving degraded speech. Here, these issues were examined both by looking at whether performance disparities interact with experimental conditions and determining whether key outcomes of each laboratory study are also observed in the corresponding web replication.

The current study involved the replication of three speech perception experiments that had previously been performed under traditinal conditions in our laboratory (Table II). We chosethese experiments because (i) they involve different fors of speech degradation applied to the same type of sentenes; (ii) the raw responses from the labo- ratory participants re available, allowing the application of completely unform post-processing steps for the two modalities (Sec. II F); and (iii they used largely nonoverlapping subsets of stimuli (Sc. II C). Collectively, the three experiments contain 40 separate blocked conditions

TABLE II. The characteristics of the three experiments replicated in the current study.

Experiment		Laboratory study	Number of conditions	Characteristics		
				Masking	Distortion	Filtering
1	Adaptation to distorted speech	In preparation	8	No	Yes	Yes
2	Speech generated from time-frequency masks	Cooke and Garc´ ıa Lecumberri (2020)	8	Yes	Yes	No
3	Spectrally enhanced speech in noise	Experiment 1 of Tang and Cooke (2018)	24	Yes	No	Yes

involving speech that has been masked, distorted, or filtered (see Table II for a breakdown), including a number of commonly used experimental processes applied to speech, with noise-vocoding, rate compression, and time-frequency mask-based enhancement, as well as a range of maskers, including white noise, speech-shaped noise, speech- modulated noise, and competing speech.

II. METHODS

Since most methodological aspects of the experiments are similar—for instance, they all involve responding to Spanish sentences and use the same techniques and interface for stimulus delivery and response elicitation—this section describes features that are common across the experiments. Experiment-specific details are provided at the point where findings are first presented in Sec. III.

A. Participants

Web participants had a similar age, gender, and educational profile to those who had taken part in the laboratory experiments. Both cohorts were composed of students at the Alava Campus of the University of the Basque Country at the time of testing. The mean ages in the laboratory experi- ments were 21.4, 20.8, and 23.8 years old for expts. 1–3, respectively, whereas the web cohorts had slightly lower mean ages (19.3, 19.4, and 19.5 years old; see Table III), reflecting the fact that web participants were recruited from a single academic year group, since that group was familiar with the web delivery platform used for the experiments, having used it for online laboratory classes in the previous semester. The proportion of females was 0.88, 0.85, and 0.68 in the laboratory experiments compared to 0.77, 0.79, and 0.80 for the web experiments (Table III).

Before giving their consent to participate, listeners were informed that their responses would be stored in anony-mised form, made aware of the ethics permission underwhich the experiment was performed (UPV/EHU TI0146), and reminded of their rights regarding data protection. As in the earlier laboratory experiments, all of the listeners were paid for their participation.

TABLE III. Summary of self-reported participant details for the web cohorts.

	Participants			Headphones			Environment		
Experiment	*N*	Age	Female	High	Mid	Low	Good	OK	Poor
1	53	19.3	41	5	22	26	10	41	2
2	47	19.4	37	5	18	24	10	35	2
3	40	19.5	32	5	15	20	9	30	1

Prior to starting the series of experiments, participants completed a web form with details of their age, gender, and native language, and were asked if their hearing was normal to the best of their knowledge. Participants were asked to perform the experiment using headphones in a quiet space. They also provided an estimate of the quality of their headphones by choosing one of three categories with illustrative pricing as a guide: *low* (under 15 euros), *mid* (15–100 euros), or *high* (more than 100 euros). In addition, partici- pants selected one of three descriptions that best fitted their listening environment: *poor* (noise present, e.g., public transport), *OK* (mainly quiet, e.g., shared quiet space), or *good* (very quiet, e.g., private enclosed space).

Table III summarises the web cohort details after excluding two listeners, one whose native language was other than Spanish or Basque and another who reported a hearing impairment. We observe that the headphone quality was almost always *mid* or *low*, whereas the listening environments were largely in the mainly quiet (*OK*) category. Only two listeners could be considered to take the experiment under conditions comparable to the laboratory, i.e., using *high* quality headphones and a *good* listening environment. At the same time, two listeners took part with *low* quality headphones in a *poor* listening context.

B. Experiment sequence

Listeners were able to take part in as many of the three experiments[2] as they wished, but had to participate in a fixed order, namely, expt. 1–expt. 3. This approach was takenbecause the sample of potential participants meeting a simi-lar profile to the laboratory studies ws limited in number, and we were unsure at the outset how many listeners would choose to take part, leading to the risk of ending up with underpowered replications. In th event, 53 participants took part in the first experiment, f which 40 went on to complete all three of the experiments We placed the experi- ment involving adaptation to speech fst in the sequence as this experiment requires listeners to be completely unfamiliar with distorted speech forms. The order of the remaining experiments was not felt to be critical.

C. Sentence materials

All speech material came from the Sharvard Corpus (Aubanel *et al.*, 2014), which consists of everyday Spanish sentences, typically containing 6–8 words, of which five are keywords used for scoring. Throughout the three web experiments, we used exactly the same stimuli that had been used in the laboratory studies (apart from transformation to a web-compatible audio format; see Sec. II D). Consequently, thesame sentence subsets from the Sharvard Corpus that were used in the equivalent laboratory studies were employed here. As a result of the fact that the three studies were performed at different times and the current replication required 720 sentences while the Sharvard Corpus contains only 700 sentences, there was some limited but unavoidable overlap in the sentence subsets used. Specifically, expt. 1 used sentences numbered 241–480, expt. 2 used sentences numbered 1–240, and expt. 3 involved numbers 401–640. All stimuli made use of the male talker from the Sharvard Corpus and were sampled at 16 kHz as in the original laboratory studies.

D. Web platform

Participants took part via a custom-built client-server application, which handled the entire experimental process. The platform was built using the Python-based Flask web development framework (Flask, 2021) on the server, coupled with an HTML5/Javascript/AJAX client layer running on the participant's web browser. To avoid within-block interruptions or delays resulting from network problems and provide a smooth experience for the listeners, all stimuli in a block of sentences were downloaded from the server as a single unit at the outset. Similarly, all user responses were stored and uploaded to the server at the end of each block. The Javascript Howler library (Howler, 2021) was used for audio playback. All of the stimuli were encoded as mpeg-4 format audio files at a sampling rate of 16 kHz usingFFmpeg (FFmpeg, 2021) with the quality set to five (maxi- mum). The mpeg format was chosen as it is supported natively by most modern browsers. Listeners provided their responses by typing in a text entry box. A progress bar and textual indication of the percentage of stimuli heard thus far in the current block was also provided.

E. Procedure

Unlike the laboratory experiments, which were completed in a single session lasting under 1 h in all cases, during which participants were encouraged to take a short break between blocks, in the web experiments the minimal unit for completion at a single sitting was the block. This approach was taken to provide more flexibility for participants to allow for suspension and resumption following potential distractions and mitigate the effects of network connection problems that might occur during the experiment.

At the outset of each experiment, a participant was assigned to one of N orderings, where N is the number of conditions in the experiment. The condition orderings followed a balanced Latin square design. Participants were allocated to each condition ordering sequentially to ensure similar numbers in each permutation of conditions.

Listeners were first presented with a short instruction screen, which introduced the type of stimuli that they would hear and identified the number of blocks in the experiment and number of sentences in each block. During the experiment, a further brief set of instructions remained on thescreen. These instructions reminded listeners to type as many words as they could hear even if they were not sure, noted that they did not have to use vowel stress marks, and asked them to choose a comfortable volume level. For expt. 3, because some of the conditions involved a (female) com- peting talker, listeners were additionally reminded to respond to the male talker in the relevant blocks.

All of the experiments started with a practice block consisting of a number of sentences (15, 5, and 6 for expts. 1–3, respectively). Listeners were able to start typing their responses as soon as each audio stimulus started. No feedback on responses was provided in any of the experiments apart from a visual indication of the progress.

F. Postprocessing

To ensure identical post-processing for the two modalities, the responses from each web study were combined with the raw responses from its laboratory counterpart, and the combined dataset was subjected to the same processing stages and outlier analysis as described below. Note that as a consequence of this reanalysis, very minor numerical differences are possible between the laboratory results reported here and those in the published studies.

An analysis of mean scores per participant in each modality and experiment indicated that two participants had scores that were more than 1.5 times the inter-quartile range below the first quartile boundary. One outlier came from the laboratory study of expt. 1; the other was an outlier in bothof the web replications in expts. 2 and 3. These two partici- pants were excluded from further analysis.

Participant responses were processed by automatic application of the following steps: (i) removal of vowelstress marks because these were regarded as optional; (ii) removal of all non-alphanumeric characters; (iii) conversion of numbers represented as digits to full orthographic forms; and (iv) identification of any words not present in a Spanish dictionary. Subsequently, any errors that occurre more than once were manually corrected in any cases in which the intended word could be identified unambiguously. Such errors are mainly "typos" (e.g., auqnue/aunque), missing tilde symbols (e.g., ninos/niños), homophones [e.g., the company name (Glovo) to its lexical homophone (globo)] or spelling errors (e.g., haros/aros). A total of 84 distinct errors across the experiments and cohorts were corrected in this way. Although effort could have been expended to inspectall non-dictionary responses, we the decision to only inspect and attempt to correct errors that occurred more thanonce as this procedure can be regarded as a reasonable strat- egy when handling large response sets and, as such, wewere interested in whether the modality affected the number of corrected errors under a typical error-correction protocol.

TABLE IV. Counts and percentages of corrected and uncorrected errors in each experiment and modality.

	Laboratory		Web	
Experiment	Corrected	Uncorrected	Corrected	Uncorrected
1	165 (0.17%)	432 (0.44%)	118 (0.17%)	462 (0.66%)
2	55 (0.13%)	212 (0.51%)	97 (0.14%)	401 (0.58%)
3	40 (0.15%)	228 (0.83%)	44 (0.11%)	241 (0.6%)

Although analysis of such errors is not a main focus of the current study, it is informative to compare the two modalities. Table IV lists counts and proportions of errors after performing the above steps. The proportions of corrected and uncorrected errors are similar in the two modalities. On average across the three experiments, 1 word in 889 was a correctable error in the laboratory modality compared to 1/952 in the web modality. For the uncorrected errors, the numbers are 1/225 and 1/217, respectively. These statistics also suggest that manual correction of errors will have an insignificant impact on reported outcomes, and of greater relevance for the current study, there is no evidence in this data that online participants made more errors than those taking part in the laboratory. This outcome is commensurate with Slote and Strand (2016), who found a similar proportion of correctable errors (in their case, amounting to approximately 1% of the responses) for the laboratory and web modalities.

III. RESULTS 1. ABSOLUTE SCORES AND INTERACTIONS WITH CONDITION

This section outlines the three laboratory experiments and describes the outcomes of the web replications in terms of how well they match absolute intelligibility scores and whether any laboratory-web disparity varies with experimental condition. The variability across participants in the two modalities is examined in Sec. III E. We start by introducing the common statistical models used in the analysis of the three experiments.

A. Statistical models

In this section, our principal interest i the effect of test modality (Modality, with levels Lab and) and the existence of any interaction with experimental condition (Condition) rather than any main effect of Condition itself. We employed generalised linear mixed-effects models to predict the proportion of keywords correct in each trial, with Modality and Condition as fixed effects, by-subject random intercepts and per-condition slopes, and by-sentence random intercepts. Model estimation was via the glmer function of the `lme4` package (Bates *et al.*, 2015) in *R* (*R* Core Team, 2021) using the nloptwrap optimizer setting. The importance of retaining factors in interactions and main effects was determined by model comparison using the anova function; we report χ^2 statistics and p-values resulting from this model comparison procedure.

Similar model comparison procedures were used to examine the influence of self-reported headphone quality (Headphones) and listening environment (Environment) for the Web cohort alone. Due to the low number of listeners with *high* quality headphones (Table III), the *mid* and *high* quality subsets were combined into a single level "*mid-high*," which led to similar numbers of participants in the *low* and *mid-high* subgroups. Likewise, the Environment factor was reduced to two levels because only two partici- pants reported listening in *poor* conditions. These listeners were combined with those in the *OK* group. Given the dif- ference in meaning between *poor* and *OK*, this is not a natu- ral combination, but is preferable to the alternative of discarding responses from these listeners.

Table V provides a of the generalised linear mixed-effects models for each comparison described in this

TABLE V. Statistical models and outcomes. The random effects structure is identical for each model (per-sentence intercepts, per-subject intercepts, and by-condition slopes). Interactions between the listed factor (Modality, Headphones, or Environment) and Condition, as well as the main effect of the factor, are reported. p-values greater than 0.01 are provided explicitly. Conventions. ***, $p < 0.001$; **, $p < 0.01$; *, $p < 0.05$; ·, $p < 0.1$.

Experiment	Cohort(s)	Fixed effects	Interaction with Condition		Main effect	
1	Lab, Web	Modality, Condition	$\chi^2(7) = 24.4$	***	$\chi^2(1) = 11.8$	***
2	Lab, Web	Modality, Condition	$\chi^2(7) = 15.5$	* (0.03)	$\chi^2(1) = 4.24$	* (0.04)
3	Lab, Web	Modality, Condition	$\chi^2(23) = 38.1$	* (0.03)	$\chi^2(1) = 1.02$	$p = 0.31$
1	Lab, Web+	Modality, Condition	$\chi^2(7) = 15.1$	* (0.04)	$\chi^2(1) = 3.45$	· (0.06)
2	Lab, Web+	Modality, Condition	$\chi^2(7) = 9.2$	$p = 0.24$	$\chi^2(1) = 0.16$	$p = 0.69$
3	Lab, Web+	Modality, Condition	$\chi^2(23) = 25.7$	$p = 0.32$	$\chi^2(1) = 0.78$	$p = 0.38$
1	Web	Headphones, Condition	$\chi^2(7) = 3.3$	$p = 0.85$	$\chi^2(1) = 4.3$	* (0.04)
2	Web	Headphones, Condition	$\chi^2(7) = 6.2$	$p = 0.52$	$\chi^2(1) = 9.8$	**
3	Web	Headphones, Condition	$\chi^2(23) = 12.1$	$p = 0.97$	$\chi^2(1) = 9.0$	**
1	Web	Environment, Condition	$\chi^2(7) = 3.1$	$p = 0.87$	$\chi^2(1) = 0.68$	$p = 0.41$
2	Web	Environment, Condition	$\chi^2(7) = 1.8$	$p = 0.97$	$\chi^2(1) = 0.26$	$p = 0.61$
3	Web	Environment, Condition	$\chi^2(23) = 25.4$	$p = 0.33$	$\chi^2(1) = 0.44$	$p = 0.51$

section, alongside a listing of statistical outcomes. Note that the Web group mentioned in Table V corresponds to the subset of the Web cohort that used *mid-high* quality headphones (see Sec. III F).

B. Experiment 1: Adaptation to distorted speeech

Listeners adapt to previously unheard forms of speech, demonstrating improvements in intelligibility with increasing exposure (e.g., Davis *et al.*, 2005; Dupoux and Green, 1997; Warren *et al.*, 1995). Experiment 1 measured adaptation to eight types of distorted speech: fast speech, sinewave speech, tone-vocoded and noise-vocoded speech, speech restricted to a narrow spectral band, locally time- reversed speech, speech resynthesised from a time- frequency mask, and a similar masking condition with amusical substrate. The aim of the laboratory study was to identify the shape of the detailed time course of adaptationto these forms of degraded speech.

Listeners heard 15 sentences of undistorted speech by the same talker used as the basis for generating distorted speech prior to starting the main experiment in order toremove any procedural learning effects without providing any exposure to the distorted speech conditions. They then heard the eight distortion types in blocks of 30 sentences. Laboratory and web listeners followed exactly the same pro-cedure using identical stimuli (as was also the case for expts.2 and 3). The combined dataset from both modalities con- sists of 28 560 sentence-level responses.

The leftmost column of Fig. 1 depicts the outcomes of expt. 1. The upper panel compares mean intelligibilities in each of the eight distorted speech conditions for the Lab and Web modalities. On average, participants in the Lab study outperformed Web listeners by 5.5 pp, with a significant interaction indicating greater gains for the more intelligible conditions (Table V). The intelligibility disparities ranged from 2 to 13 pp, where the upper limit corresponds to the time-compressed speech condition (i.e., fast speech, speeded up by a factor of 2.5); for the remaining seven conditions, the disparity was 6 pp or less.

The effect of headphone quality within the Web cohort is shown in the middle panel of Fig. 1. There was a small but significant positive impact of using mid or high quality equipment and no interaction with the experimental condition (Table V). The lower panel compares the outcomesfor those Web listeners reporting a *poor* or *OK* listening environment with those reporting a *good* environment. The quality of the listening environment had no effect on intelligibility in expt. 1 (Table V).

C. Experiment 2: Speech generated from time-frequency masks

Experiment 2 replicates a study into the intelligibility of speech produced by passing different kinds of driving signal through a fixed time-frequency mask (Cooke and Garc ́ ıa Lecumberri, 2020). The goal of the study was to determine the extent to which a binary time-frequency mask alone is capable of supporting intelligibility, regardless of the signal that is used for resynthesis. Eight experimental manipulations were tested, one, a baseline condition consisting of intact speech mixed with speech-shaped noise, and a further seven conditions that varied in the amount of speech information in the driving signal. The latter seven conditions were presented without masking noise. Listeners responded to 30 sentences per block. The combined Web/Lab dataset consists of 17 505 sentences. Due to a software problem, which affected 4 web listeners, a total of 15 responses (0.08% of the dataset) were not recorded, 10 of which occurred in the condition where scores were at the floor.

The outcomes for expt. 2 are shown in the middle column of Fig. 1. One of the conditions led to a chance level of identification (3.3% words correct) just as in the Lab study (2.7%); this point is not plotted or used for the best-fit lines to prevent misleading fits. In all of the conditions, the Web score was equal to or lower than the corresponding Lab score, with disparities ranging from 1 to 9 pp, and a small but significant overall mean difference of 3.6 pp. The largest disparity came in response to the baseline speech-plus-noise mixture condition. A modest interaction is in evidence with a similar pattern to that of expt. 1, i.e., the size of the Lab group advantage tended to be larger in the higher intelligibility conditions (Table V).

Regarding the impact of the headphone quality (middle panel of Fig. 1), a substantial difference in mean scores of nearly 7 pp is evident (Table V). Headphone quality did not interact significantly with Condition. As in expt. 1, the quality of the listening environment (lower panel of Fig. 1) had no effect on overall scores (Table V).

D. Experiment 3: Spectrally enhanced speech in noise

Experiment 3 replicates the first of the two experiments in Tang and Cooke (2018), a study which looked at how the intelligibility of masked speech can be improved without changing overall SNR by the application of learnt static spectral weightings. The Lab experiment tested unmodified and enhanced speech at two SNRs and mixed with six different maskers—speech-shaped noise, speech-modulated noise, white noise, competing speech, low-pass filtered noise, and high-pass filtered noise; see Fig. 1 in Tang and Cooke (2018)—leading to a total of 24 conditions. Listeners heard 20 sentences in each block, with each block consistingof 10 unmodified and 10 modified sentences in a random order. The combined dataset for this experiment consistedof 14 877 sentence responses (3 sentences were lost due to the technical issue described in expt. 2).

The rightmost column of Fig. 1 presents outcomes for the 24 conditions of expt. 3. For this analysis, CONDITION was treated as a single factor and not broken down into its constituent subfactors viz., masker, SNR, and presence of enhancement. Across conditions, the laboratory modality resulted in an overall gain of 3.2 pp, with a range of disparities from -7 pp (i.e., in favour of the WEB cohort) to +15 pp. Three of the largest disparities favouring the LAB cohort came in conditions with a competing speech masker.

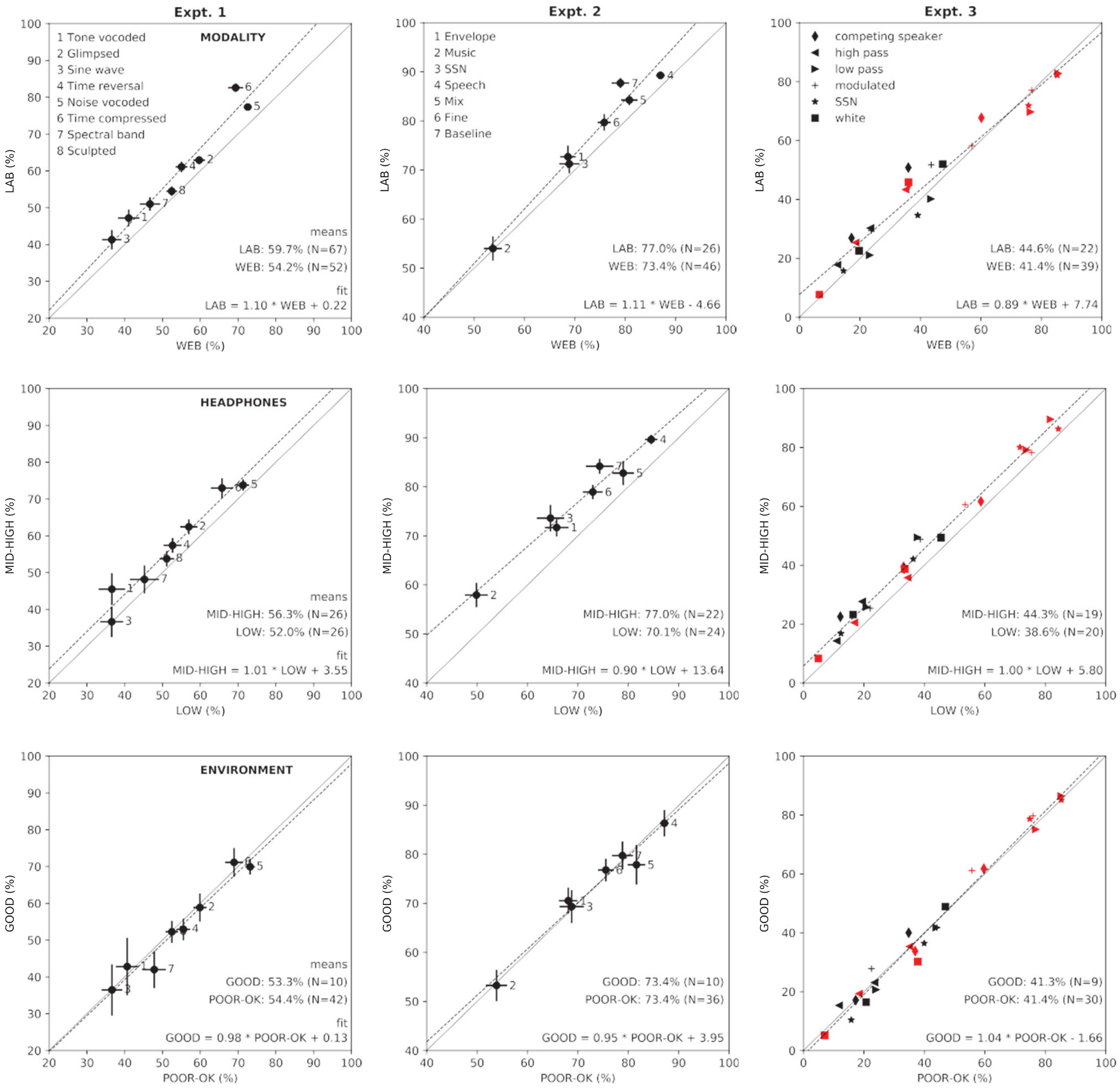

FIG. 1. (Color online) Keyword scores for the conditions of expt. 1 (left column), expt. 2 (middle column), and expt. 3 (right column). The top row com- pares Lab and Web modalities, the middle row compares the impact of "low" and "mid-high" headphone quality for the Web cohort, and the lower row com-pares Web participants who reported "poor-OK" vs "good" listening environments. Solid lines indicate equal performance, while dotted lines show the best linear-fits, whose equation is provided at the lower-right of each plot. Cohort sizes are indicated with *N*. Horizontal and vertical error bars denote ±1 standard error. For expt. 3 (right column), the red symbols denote enhanced speech conditions, whereas the black symbols signify unmodified speech. The two SNRs are not distinguished in this plot, but the higher intelligibility of each pair corresponds to the higher SNR value.

Modality showed a modest interaction with Condition, but there was no significant effect of Modality overall, i.e., Lab and Web cohorts produced statistically equivalent scores across conditions (Table V). Unlike the case for expts. 1 and 2, the laboratory advantage tended to be larger at lower intelligibilities.

Within the Web modality there was a clear benefit amounting to 5.7 pp from the use of mid-high quality headphones, and no interaction with the Condition (Table V). As in expts. 1 and 2, the Environment factor had no impact (Table V).

E. Variability among participants

Figure 2 plots the standard deviations of the per-cohort mean intelligibility scores for all of the conditions of the three experiments and suggests no clear differences between the two modalities, although there are differences in the size and spread of the standard deviations between the experi- ments, with the smallest values in expt. 2, the largest values in expt. 1, and the greatest spread in expt. 3, presumably resulting from the differing of the conditions in those experiments. Visual impressions were confirmed by an ANOVA with

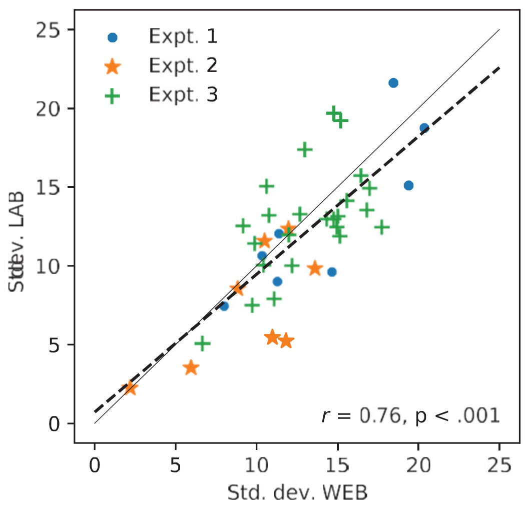

FIG. 2. (Color online) Comparison of standard deviations for the 40 conditions of expts. 1–3. Each point represents standard deviations for the Lab and Web cohorts for one condition. The solid gray line represents equal standard deviations, whereas the dotted line is the best linear-fit.

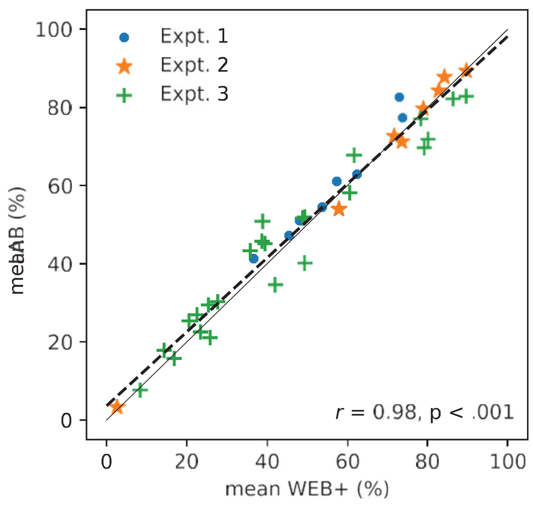

FIG. 3. (Color online) Comparison of mean scores for the 40 conditions of expts. 1–3 for the Lab and Web groups. The solid gray line represents equal performance, whereas the dotted line is the best linear-fit.

factors of Modality and Experiment, which indicated a significant effect of Experiment $[F(2, 74) = 11.1,\ p < 0.001]$ but no Modality effect $[F(1, 74) = 1.16;\ p = 0.29]$ nor interaction [F(2, 74]= 0.40; p = 0.67].

F. Interim discussion

The three web replications present a consistent pattern for identifying keywords in sentences in a range of conditions: a web cohort of known listeners fell short of scores obtained by equivalent groups in the laboratory by 5.5, 3.6, and 3.2 pp for expts. 1–3, respectively, and all but the latter are statistically significant. In all three cases, there was a mild interaction with experimental condition, although the direction of the interaction varied with the experiment, with larger Lab-Web disparities for the more intelligible conditions in expts. 1 and 2 and the reverse in expt. 3. This contrasting outcome might stem from differing experimental conditions. For example, in expts. 1 and 2, listeners mainly heard processed rather than masked speech, whereas in expt. 3 all of the conditions involved masking noise.

In all three experiments, online listeners who self-reported the use of *mid* or *high* quality headphones produced statistically higher scores than those who used *low* quality headphones, with no interaction with condition. This is a key outcome that suggests that headphone quality is of criti- cal importance in web experiments.

To further investigate the impact of headphone quality, we compared Lab performance with that of the subset of the Web cohort who used *mid* or *high* quality headphones. This latter group is denoted as the Web+ cohort in the remainder of this article. Figure 3 demonstrates a strong correlation of mean scores in the 40 condition s of expts. 1–3 for the Lab and Web+ groups. Differences in mean scores between the Lab and Web+ groups amount to 3.4, 0.1, and 0.3 pp for expts. 1–3, respectively. Statistical outcomes are shown in

rows 4–6 of Table V. Experiment 1 exhibited a marginal effect of Modality (p = 0.06), but there was no difference between the two groups in expts. 2 and 3 (Table V). Similarly, there was a (marginal) interaction of the modality and condition in expt. 1 (p = 0.04) but not in expts. 2 and 3 (Table V). These results suggest that most of the Lab-Web disparity is due to the use of low quality headphones by a subset of the latter group.

In no experiment was the self-reported listening environment an important factor in predicting intelligibility for the Web modality (Table V). Our decision to add the two participants who reported *poor* listening conditions into the group reporting *OK* conditions for the purpose of analysing potential listening environment effects ought to have promoted the appearance of any such effects. The fact that there was no differential effect of the self-reported listening environment on absolute scores suggests that the use of headphones largely obviates the need to provide an "optimal" laboratory-like listening context for web experiments.

IV. RESULTS 2. REPLICATION OF KEY OUTCOMES

While the ability to match absolute scores and the across-condition pattern of intelligibility is of relevance in assessing the value of an online modality for speech percep- tion experiments, another important element is determining whether online participants can replicate the key outcome of a given experiment. This section examines the degree to which the key outcome of each experiment was matched by the Web cohort and its Web+ subset.

A. Experiment 1

The main finding in the laboratory study of adaptation to distorted speech was that the improvemnt in intelligibil- ity with increasing exposure to sentencesas the block progressed follows a "rapid-then-gradual" attern, best fit with

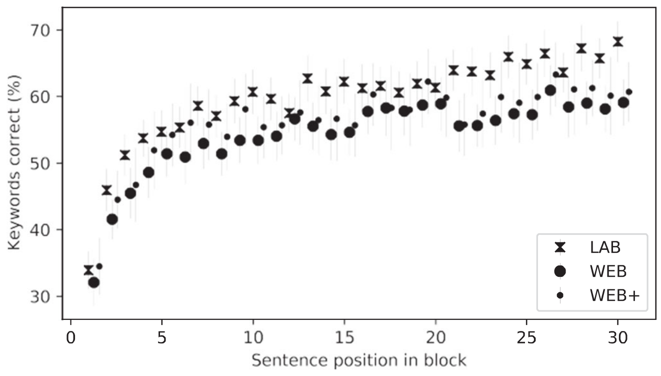

FIG. 4. Sentence scores as a function of the sentence position in the block in expt. 1 for the LAB and WEB cohorts and WEB+ subset. Error bars indi- cate 95% confidence intervals.

a logarithmic function. Figure 4 illustrates this pattern for the LAB, WEB, and WEB cohorts and suggests that online listeners exhibit a reduced degree of adaptation, with a dis- parity that grows with the position of the sentence in the block. The pattern for the WEB group is somewhat inter- mediate betweeen those of the LAB and web groups.

A generalised linear mixed-effects model with MODALITY as a fixed factor, the logarithm of sentence position in the block (POSITION) as a covariate, and the same random effects structure for the subject and sentence as described in Sec. III A was constructed to examine whether MODALITY was a significant predictor of the pattern of keyword scores across the block. For the LAB-WEB constrast, POSITION interacted with MODALITY [$\chi^2(1) = 20.8$, $p < 0.001$] with a main effect of MODALITY [$\chi^2(1) = 11.4$, $p < 0.001$] and a clear effect of POSITION [$\chi^2(1) = 2836$, $p < 0.001$] POSITION [$\chi^2(1) = 2315$, $p < 0.001$] and interaction [$\chi^2(1) = 13.2$, $p < 0.001$] effects were also present for the LAB vs WEB+ comparison but the impact of MODALITY was marginal [$\chi^2(1) = 3.37$, $p < 0.07$]. Caution should be exercised in interpreting this outcome as an unqualified replication of the key adaptation finding due to differences in the samples sizes of the LAB and WEB+ cohorts (see Sec. IV D).

B. Experiment 2

The principal finding of Cooke and Garc´ıa Lecumberri (2020) was that a time-frequency mask is not the only determinant of intelligibility, as evidenced by the wide spread of scores across conditions. Unsurprisingly, this finding is replicated by the WEB cohort. A more stringent test of replication is to ask whether two subsets of theoretically relevant conditions that were found to be statistically equivalent in the LAB study are likewise equivalent in the online study. The first subset contains the conditions labelled Baseline, Mix, and Speech in Fig. 1, corresponding to speech in noise,speech in noise passed through a mask generated by estimat- ing speech glimpses in noise, and speech alone passed through the mask. Their equivalence supports the idea that listeners are making sole use of those regions where the speech is locally dominant when processing speech in noise.The other subset involves the conditions labelled Envelope and SSN, corresponding to signals generated by passingeither the speech envelope or a noise envelope through the same mask. Their equivalence suggests that the non-binary envelope from the speech adds no information over andabove the binary envelope represented by the mask.

Cooke and Garc´ıa Lecumberri (2020) found that listeners' performance improved over the first two blocks; consequently, condition comparisons were performed using blocks 3–8, i.e., after performance had stabilised. Keyword scores from these blocks for both equivalence-subsets are shown in Fig. 5. A generalised linear mixed-effects model was constructed to predict the proportion of keywords correct in each sentence, with fixed effects of MODALITY and CONDITION and the same random effects structure for the subject and sentence as described in Sec. IV A, followed by pairwise contrasts with Tukey corrections for multiple comparisons using the `emmeans` package in *R* (Lenth, 2021). For the first equivalence, the WEB cohort displayed statisti- cally distinct outcomes (Baseline vs Mix, 4.1 pp, $z = 3.0$, $p = 0.04$; Baseline vs Speech, 7.9 pp, $z = 6.4$, $p < 0.0001$) unlike the LAB cohort (Baseline vs Mix, 2.0 pp, $z = -1.6$, $p. = 0.64$; Baseline vs Speech, 0.8 pp, $z = 0.7$, $p = 0.99$). It is apparent from Fig. 5 that this outcome stems entirely from the substantially lower scores produced by the WEB cohort in the Baseline condition. For the second equivalence, the WEB cohort successfully replicated the LAB finding in that neither distinguished the envelope and SSN conditions (LAB, 1.7pp, $z = 1.3$, $p = 0.86$; WEB, 1.4pp, $z = -1.2$, $p = 0.88$). For theWEB+ group, both equivalences were obtained: envelope-SSN (1.0 pp, $z = -0.71$, $p = 0.99$); three-way equivalence (Baseline vs Mix, 0.2pp, $z = 0.15$, $p = 0.99$; Baseline vs Speech, 3.8pp, $z = 2.4$, $p = 0.21$). This result demonstrates that the LAB and WEB+ cohorts produced identical key outcomes for expt. 2.

C. Experiment 3

The main finding of Tang and Cooke (2018; expt. 1) was that spectral enhancement leads to substantial intelligi- bility improvements in all masker conditions except white noise at both of the SNRs tested. To compare the two modalities, we examined per-participant gains in scores as a function of SNR and masker type. Gains were computed as

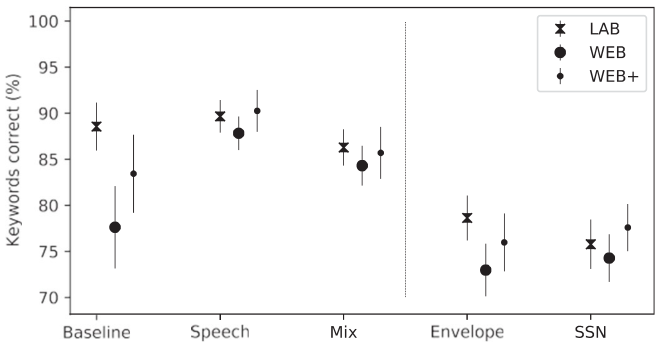

FIG. 5. Keyword scores for two sets of conditions in expt. 2 for the LAB and WEB cohorts and the WEB+ subset. Error bars indicate 95% confidence intervals.

the difference in keyword scores between the ten enhanced sentences in each block and the ten non-enhanced sentences.

Across modalities, the ranking of conditions was identical, with very similar absolute gains (Fig. 6). A linear mixed-effects model with fixed effects of the Condition, SNR, and Modality, and random by-subject intercepts, was constructed to predict gains in pp. Unlike the statistical models used earlier, the response variable (gain) was averaged across blocks because listeners did not hear the same sentences in unmodified and enhanced conditions and, consequently, the random by-subject slopes per condition were not included as there is only one data point for each (subject,condition) pair. Inclusion of Modality or any of its interactions with SNR or masker had no explanatory impact for the Lab vs Web contrast [$\chi^2(12) = 17.5$, p= 0.13] nor for the Lab vs Web+ comparison [$\chi^2(12) = 13.1$, p= 0.36]. These findings confirm that the critical outcome of expt. 3 isnot affected by modality.

D. Interim discussion

The key outcomes in each of the three laboratory experiments were matched by online listeners to differing degrees, with the weakest replication in expt. 1 and the strongest replication in expt. 3. In expts. 1 and 2, the Lab and Web groups displayed differences in key outcomes, whereas in expt. 3, both cohorts replicated the main finding. In all three experiments there was no statistically significant effect on Modality for the Lab-Web+ comparison.

In expt. 1, it would be unsafe to conclude that the Lab and Web+ groups behave similarly with respect to the key outcome, since one consequence of selecting a subset of Web listeners is to reduce the statistical power of any comparisons. This has its greatest impact in expt. 1, where the 67-strong Lab cohort is contrasted with 26 in the Web+ group. On the other hand, in expts. 2 and 3, the choice of a subset of the Web group led to a greater balance in participant numbers in the two modalities (expt. 2, 26 vs 22; expt. 3, 22 vs 19 for the Lab and Web+ groups, respectively).

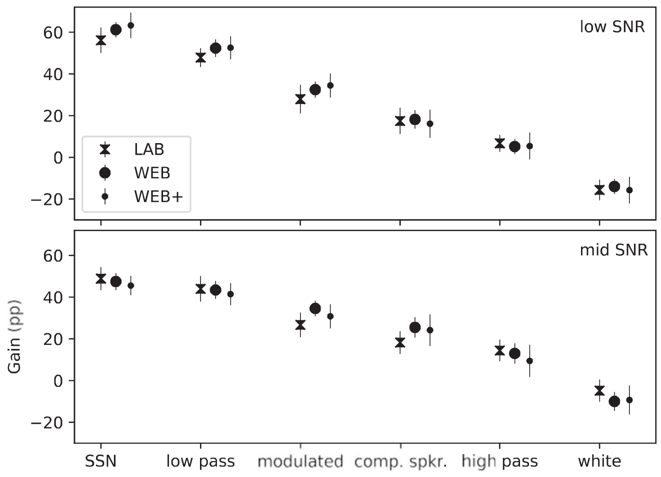

FIG. 6. Gains due to spectral enhancement in expt. 3 as a function of masker type and SNR for the Lab and Web cohorts and the Web subset. Error bars indicate 95% confidence intervals.

Although there are clear indications of a possible structural difference in intelligibility across modalities in expt. 1, the situation is less clear in expt. 2 in which the outcome differences hinge on a single condition. In the absence of within-modality replications (e.g., test-retest in the laboratory), the level of detail at which it makes sense to compare outcomes from different experimental modalities is an open question.

V. GENERAL DISCUSSION

The first question that the current study set out to address concerns the size of any laboratory advantage in absolute terms for experiments measuring speech intelligibility across a range of processing conditions. The outcomesindicate that the laboratory advantage is modest (under 5.5 pp), and minimal (under 1 pp) when data from partici- pants who used low quality headphones are excluded. These findings contrast with the substantially greater differences reported in earlier comparisons (e.g., Cooke *et al.*, 2011; Mayo *et al.*, 2012; Slote and Strand, 2016; Wolters *et al.*, 2010). The extent to which this outcome is the result of tech-nological advances over the last decade in, for example,soundcards or browser audio rendition, or the use of a known sample of listeners that are well-matched to their lab-oratory counterparts is hard to judge. Equivalent replications using crowdsourced participants are needed to answer this question.

The importance of avoiding the use of low quality headphones is evident from a comparison of absolute scores. Whereas screening techniques have been developed to detect whether participants are listening over headphones or loudspeakers (Milne *et al.*, 2020; Woods *et al.*, 2017), there remains a need for techniques to distinguish headphone characteristics. It may also be worth measuring the level of performance that can be obtained using smart speakers relative to low quality headphones in light of the finding that accurate speech audiometry is possible using such devices (Ooster *et al.*, 2019). The limited range of smart speakers may also bring a degree of consistency to listening experiments.

In some conditions, the Web cohort outperformed their laboratory counterparts in strictly numerical terms, particularly in expt. 3 (see Fig. 3). Leaving aside the issue of whether these differences lie within expected levels of variation, it is plausible that certain factors benefit web participants. For example, while laboratory participants are encouraged to take breaks between blocks, these are often minimal, whereas web participants in the current replications were able to engage and disengage between blocks at will. The web participants also presumably performed the tasks at times when they were motivated to do o, in contrast to taking part at a pre-agreed timeslot. Theremay also be an element of stress in visiting a physical labortory and undergoing a test under the vigilance of experienters. Finally, participants can be expected to be highly familiar with their own equipment, unlike the case for laboratory experiments.

The second question concerned whether the modalities lead to different outcomes in terms of both the pattern of results across conditions and in the replication of key outcomes from each experiment. Concerning the first factor, while modality interacted with the experimental conditionfor the full web cohort, these interactions were attenuated (expt. 1) or disappeared (expts. 2 and 3) for the cohort using reasonable quality headphones. Similar findings were observed for the key outcomes of each experiment. In spite of the overall success in matching the pattern and detail of laboratory findings, caution is required in extrapolating to speech processing and masking conditions in general. While the study lacked the statistical power necessary to adequately explore replications at the level of individual experimental manipulations, we observed that the time-compressed speech condition in expt. 1 led to the largest laboratory-web disparity and also that several of the largest disparities in expt. 3 came from the conditions involving a competing speech masker. Further studies may be needed to confirm the reliability of web outcomes for these and other specific families of speech modification processes or mask- ing conditions.

The value of using known listeners is highlighted by three findings that relate to the ultimate statistical efficiency of any experiment. First, inter-listener variability was essentially equivalent for the Lab and Web groups regardless of headphone quality. Consequently, similar sample sizes can be used to address the same experimental question in thetwo modalities. Second, the proportion of textual responses containing nonwords that required automatic or manual correction was very similar for the two groups. Finally, the application of identical outlier criteria resulted in similar levels of participant exclusion (here, the removal of one person from each modality). This level of inclusion contrasts with typical participant wastage in crowdsourced experiments. For example, in a review of eight studies involving the use of online crowdsourcing to assess disordered speech, Sescleifer *et al.* (2018) reports participant exclusion rates ranging from 22% to 60%.

One limitation of this study, which results from the ongoing pandemic, was the non-simultaneity of testing of the two modalities. Ideally, the interval between tests in the two modalities would be minimised to reduce the influence of factors such as technological improvements in online delivery of audio and headphone quality that could conceiv- ably affect any comparisons of the outcomes in experiments that take place several years apart.

VI. CONCLUSIONS

Online speech perception experiments performed by known participants, who avoid the use of low quality headphones, can match laboratory studies in terms of absolute scores and the pattern of results across conditions, as well as replicating key findings, while achieving a similar level of statistical efficiency. In times when access to formal laboratory facilities is problematic, the option of using known participants with heterogeneous equipment and listening environments to respond to stimuli delivered by web platforms provides a viable approach for researchers in speech perception and, in addition, enables the online pursuit of studies involving speech intelligibility in global contexts where formal speech perception laboratories are not available.

ACKNOWLEDGMENTS

This work was funded by the Basque Government Consolidados programme under Grant No. IT311-10 (Language and Speech).

[1] Although we refer to such participants as known listeners, this does not mean that their anonymity is in any way compromised.

[2] An additional experiment involving speech in noise was also performed at position two in the sequence, but due to an oversight, the sampling frequency of the web version did not match that of the original laboratory study, therefore, this experiment was not pursued further. Some participants also went on to take part in a fifth experiment involving isolated word confusions in noise, which lies outside the scope of the current study.

Aubanel, V., Garc´ ıa Lecumberri, M. L., and Cooke, M. (2014). “The Sharvard corpus: A phonemically-balanced Spanish sentence resource for audiology,” Int. J. Audiol. 53, 633–638.

Bates, D., Mâchler, M., Bolker, B., and Walker, S. (2015). “Fitting linear mixed-effects models using lme4,” J. Stat. Software 67(1), 1–48.

Beasley, R., and Chuang, Y. (2008). “Web-based music study: The effects of listening repetition, song likeability, and song understandability onEFL learning perceptions and outcomes,” TESL-EJ 12(2), 1–17.

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,” J. Med. Internet Res. 10(4), e32, available at https://www.jmir.org/2008/4/e32/.

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

Borrie, S. A. (2018). “Understanding dysrhythmic speech: When rhythm does not matter and learning does not happen,” J. Acoust. Soc. Am. 143, EL379–EL385.

Burgos, P., Sanders, E., Cucchiarini, C., van Hout, R., and Strik, H. (2015). “Auris populi: Crowdsourced native transcriptions of Dutch vowels spoken by adult Spanish learners,” in *Proc. Interspeech*, pp. 2819–2823.

Choi, J., Lee, H., Park, C., Oh, S., and Park, K. (2007). “PC-based tele-audiometry,” Telemed. e-Health 13(5), 501–508.

Cole, J., Mahrt, T., and Roy, J. (2017). “Crowd-sourced prosodic annotation,” Comput. Speech Lang. 45, 300–325.

Cooke, M., Barker, J., and Garcia Lecumberri, M. L. (2013). “Crowdsourcing in speech perception,” in *Speech Processing: Applications to Data Collection, Transcription and Assessment*, edited by M. Eskenazi, G.-A. Levow, H. Meng, G. Parent, and D. Sundermann (Wiley, New York), pp. 141–176.

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

Cooke, M., and Garc´ ıa Lecumberri, M. L. (2020). “Sculpting speech from noise, music, and other sources,” J. Acoust. Soc. Am. 148, EL20–EL26.

Cox, T. (2008). “The effect of visual stimuli on the horribleness of awful sounds,” Appl. Acoust. 69(8), 691–703.

Davis, M. H., Johnsrude, I. S., Hervias-Adelman, A., Taylor, K., and McGettigan, C. (2005). “Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences,” J. Exp. Psych. Gen. 134, 222–241.

Dupoux, E., and Green, K. (1997). “Perceptual adjustment to highly compressed speech: Effects of talker and rate changes,” J. Exp. Psych. Human Percept. Perform. 23, 914–927.

Evanini, K., and Zechner, K. (2011). “Using crowdsourcing to provide prosodic annotations for non-native speech,” in *Proc. Interspeech*, pp. 3069–3072.

FFmpeg (2021). “Ffmpeg v4.4,” available at https://www.ffmpeg.org (Last viewed 8/7/2021).

Flask (2021). “Flask v1.1.2,” available at https://palletsprojects.com/p/flask/ (Last viewed 8/7/2021).

Gould, S. J. J., Cox, A. L., Brumby, D. P., and Wiseman, S. (2015). “Home is where the lab is: A comparison of online and lab data from a time-sensitive study of interruption,” Hum. Comput. 2, 45–67.

Howler (2021). “Howler v2.2.1,” available at https://howlerjs.com (Last viewed 8/7/2021).

Jiao, Y., LaCross, A., Berisha, V., and Liss, J. (2019). “Objective intelligibility assessment by automated segmental and suprasegmental listening error analysis,” J. Speech, Lang., Hear. Res. 62(9), 3359–3366.

Jiménez, R. Z., Naderi, B., and Möller, S. (2020). “Effect of environmental noise in speech quality assessment studies using crowdsourcing,” in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.

Lenth, R. V. (2021). “emmeans: Estimated marginal means, aka least-squares means,” *R* package version 1.5.5-1, available at https://CRAN.R-project.org/package=emmeans (Last viewed 8/7/2021).

Mayo, C., Aubanel, V., and Cooke, M. (2012). “Effect of prosodic changes on speech intelligibility,” in *Proc. Interspeech*, pp. 1708–1711.

McAllister Byun, T., Halpin, P. F., and Szeredi, D. (2015). “Online crowdsourcing for efficient rating of speech: A validation study,” J. Commun. Disord. 53, 70–83.

Melguy, Y. V., and Johnson, K. (2021). “General adaptation to accented English: Speech intelligibility unaffected by perceived source of non-native accent,” J. Acoust. Soc. Am. 149, 2602–2614.

Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Billig, A. J., and Chait, M. (2020). “An online headphone screening test based on dichotic pitch,” Behav. Res. Methods (published online, 2020).

Naderi, B., and Möller, S. (2020). “Application of just-noticeable difference in quality as environment suitability test for crowdsourcing speech quality assessment task,” in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, pp. 1–6.

Ooster, J., Moreta, P. N. P., Bach, J.-H., Holube, I., and Meyer, B. T. (2019). “Computer, test my hearing”: Accurate speech audiometry with smart speakers,” in *Proc. Interspeech 2019*, pp. 4095–4099.

Parson, J., Braga, D., Tjalve, M., and Oh, J. (2013). “Evaluating voice quality and speech synthesis using crowdsourcing,” in *Text, Speech, and Dialogue*, edited by I. Habernal and V. Matoušek (Springer, Berlin), pp. 233–240.

R Core Team (2021). “R: A language and environment for statistical computing” (*R* Foundation for Statistical Computing, Vienna, Austria), available at https://www.R-project.org/ (Last viewed 8/7/2021).

Schwartz, G., and Aperliński, G. (2014). “The phonology of CV transitions,” in *Crossing Phonetics-Phonology Lines* (Cambridge Scholars, Newcastle), pp. 277–298.

Seren, E. (2009). “Web-based hearing screening test,” Telemed. e-Health 15(7), 678–681.

Sescleifer, A. M., Francoisse, C. A., and Lin, A. Y. (2018). “Systematic review: Online crowdsourcing to assess perceptual speech outcomes,” J. Surg. Res. 232, 351–364.

Slote, J., and Strand, J. F. (2016). “Conducting spoken word recognition research online: Validation and a new timing method,” Behav Res. 48, 553–566.

Tang, Y., and Cooke, M. (2018). “Learning static spectral weightings for speech intelligibility enhancement in noise,” Speech Commun. 49, 1–16.

Van Hedger, S. C., Heald, S. L. M., Nusbaum, H. C., Batterink, L. J., Davis, M. H., and Johnsrude, I. S. (2019). “Learning different forms of degraded speech as a cognitive skill,” in *Annual Meeting of the Psychonomic Society*, Montreal, Canada.

Vaughn, C. R. (2019). “Expectations about the source of a speaker’s accent affect accent adaptation,” J. Acoust. Soc. Am. 145, 3218–3232.

Warren, R. M., Riener, K. R., Bashford, J. A., and Brubaker, B. S. (1995). “Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits,” Percept. Psychophys. 57, 175–182.

Wolters, M., Isaac, K., and Renals, S. (2010). “Evaluating speech synthesis intelligibility using Amazon’s Mechanical Turk,” in *Proc. 7th Speech Synthesis Workshop (SSW7)*, pp. 136–141.

Woods, K. J. P., Siegel, M. H., Traer, J., and McDermott, J. H. (2017). “Headphone screening to facilitate web-based auditory experiments,” Atten. Percept. Psychophys. 79, 2064–2072.

Yoho, S. E., and Borrie, S. A. (2018). “Combining degradations: The effect of background noise on intelligibility of disordered speech,” J. Acoust. Soc. Am. 143, 281–286.

Yoho, S. E., Borrie, S. A., Barrett, T. S., and Whittaker, D. B. (2019). “Are there sex effects for speech intelligibility in American English? Examining the influence of talker, listener, and methodology,” Atten. Percept. Psychophys. 81, 558–570.

Zequeira Jiménez, R., Mittag, G., and Möller, S. (2018). “Effect of number of stimuli on users perception of different speech degradations. A crowdsourcing case study,” in *2018 IEEE International Symposium on Multimedia (ISM)*, pp. 175–179.