Perception and automatic detection of wind-induced microphone noise

Iain R. Jackson,[a] Paul Kendrick, Trevor J. Cox, Bruno M. Fazenda, and Francis F. Li
Acoustics Research Centre, University of Salford, Salford, M5 4WT, United Kingdom

(Received 29 January 2014; revised 14 July 2014; accepted 21 July 2014)

Wind can induce noise on microphones, causing problems for users of hearing aids and for those making recordings outdoors. Perceptual tests in the laboratory and via the Internet were carried out to understand what features of wind noise are important to the perceived audio quality of speech recordings. The average *A*-weighted sound pressure level of the wind noise was found to dominate the perceived degradation of quality, while gustiness was mostly unimportant. Large degradations in quality were observed when the signal to noise ratio was lower than about 15 dB. A model to allow an estimation of wind noise level was developed using an ensemble of decision trees. The model was designed to work with a single microphone in the presence of a variety of foreground sounds. The model outputted four classes of wind noise: none, low, medium, and high. Wind free examples were accurately identified in 79% of cases. For the three classes with noise present, on average 93% of samples were correctly assigned. A second ensemble of decision trees was used to estimate the signal to noise ratio and thereby infer the perceived degradation caused by wind noise.
© *2014 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4892772]

PACS number(s): 43.60.Np, 43.38.Kb [SAF] Pages: 1176–1186

I. INTRODUCTION

Noise created by air flow over a microphone can cause problems when making sound recordings outdoors. It also degrades the sound heard via hearing aids.[1] Our study first examined how wind-induced microphone noise (henceforth referred to as wind noise) is perceived on recordings. Then it explored how the perceived quality of the degraded audio could be estimated from a microphone signal using machine learning algorithms. The study is concerned with sound recordings made by both amateurs and professionals. The prevalence of portable consumer devices, such as mobile phones, has led to a large increase in user-generated content. While inexpensive technologies have liberated amateurs to make recordings, many are made in challenging conditions. Coupled with a lack of awareness of recording techniques, challenging recording conditions can cause audio quality to be poor. Consequently, the study considered a wide range of audio qualities, along with a wide range of recording devices from smart phones to separate microphones plugged into digital recorders.

Before developing algorithms to detect the quality of recordings in the presence of wind noise, it was necessary to understand how the presence of the noise degrades perceived audio quality. The effect of wind noise on the perceived quality of particular products or environments has been explored previously, such as the effect of wind noise in moving vehicles.[2] However, little was known about how the presence of wind noise in audio recordings affects the perceived quality. Therefore, a key research issue was to determine what features of the noise make it noticeable and affect perceptions of quality. For example, how do the gusts of wind that make the noise time-variant affect quality? To develop this understanding, a set of perceptual tests were carried out as detailed in Sec. III.

A wind noise meter was then developed that estimated the perceived quality of recordings contaminated with wind noise. There are a number of published methods for detecting wind noise, although none of them were suitable for the broad range of recording applications our study considered. Hearing aids can use two microphones to detect wind noise. As wind has a much lower velocity than sound waves, low correlation between the two microphone signals indicates the presence of wind noise.[3] Our interest was in user generated content, much of which is gathered from a single mono microphone, and this meant that methods using two microphones were not suitable.

Single channel techniques to detect wind noise in hearing aids often compare long term averages of the signal and wind noise spectra, exploiting the distinctive low frequency spectrum typical of wind noise.[4] Unfortunately, the low frequency sensitivity of some consumer recording devices precludes such a method. Indeed, some consumer devices automatically apply high pass filters as a method to stop the low frequencies in wind noise overloading the preamplifier. Therefore, our detector could not simply rely on the examination of low frequencies.

A number of detection methods have been proposed to allow subsequent wind noise reduction. Nemer[5] used a decision tree to identify wind noise frames in the presence of speech using a combination of features from a linear prediction analysis and harmonic analysis. This method will fail with sounds other than speech and relies on the existence of a resonance in the wind noise, which is only present for a subset of wind-noise cases.[6] Xiaoqiang *et al.*[7] and Schmidt *et al.*[8] built up a dictionary of wind noise signatures using examples where only wind noise was present in the audio.

[a]Author to whom correspondence should be addressed. Electronic mail: jacksoniain@hotmail.com

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

As the system was trained using noise free cases, the detection is likely to be less effective in the presence of the foreground sound that is the target of the recording.

A model to estimate the perceived degradation caused by wind noise has to overcome a considerable number of factors: There are a number of different generating mechanisms which can alter the character of the wind noise;[9] recording devices have different frequency responses, and there are a vast number of possible foreground sounds that recordists might be trying to capture. This makes a machine learning approach, where an algorithm is trained to predict wind noise from salient features, a promising choice for this problem. The development of the wind noise meter, and the results achieved with it, are given in Sec. IV.

To allow the perceptual measurements and to develop the wind noise meter, a dataset of audio examples was required. Therefore, a wind noise simulator was developed to be capable of rendering an audio stream based on real meteorological data, as described in Sec. II. In addition, field measurements were taken to allow robust validation of the machine learning algorithm used to model the perceived degradation caused by wind noise.

II. DATABASE GENERATION

An algorithm was developed to allow simulated wind noise to be generated. It permitted the sound pressure level of the wind noise to be known for every sample, because the noise was generated in isolation from other sounds. The wind noise could then be readily combined with other foreground sounds to simulate a diverse range of recording scenarios. The algorithm used airflow data from anemometer measurements as the generating function. This allowed databases of pre-existing wind velocity histories to be used to simulate a wide range of wind conditions.

To produce an algorithm to simulate wind noise, it is necessary to understand some of the key characteristics of the noise. The natural atmosphere contains turbulent fluctuations in temperature, velocity, density, and humidity. Wind noise is created when the wind advects turbulent fluctuations over a microphone. There are two dominant components to wind noise in the audible range.[6,9] The first component comes from naturally occurring turbulent fluctuations in the atmosphere that are recorded as pressure variations at the microphone. This component is predominantly caused by velocity turbulence inducing stagnation pressure fluctuations at the microphone. The spectra of the atmospheric pressure and velocity fluctuations both exhibit a relationship where the power is proportional to $k^{-5/3}$, where k is the wave number. This component often dominates the overall wind noise.[9] The second component causing audible noise arises from interactions between the wind and the microphone. It includes resonant behavior such as vortex shedding and boundary layer turbulence around the object. The eddy interaction exhibits a $k^{-7/3}$ power law, while interaction of the eddies with the vertical wind shear has a $k^{-11/3}$ power relationship in the inertial region.

Two simulators have been developed to generate wind noise, accounting for microphones with and without windshields. One model was based on real wind noise recordings from an unshielded microphone in a wind tunnel, the other used a stochastic simulation of a shielded microphone based on a model by Van den Berg.[10] These two models produce noise samples at a known, static wind speed. To simulate time-varying wind conditions over a range of metrological conditions, these were combined with anemometer time histories as described at the end of this section.

Wind noise for an unshielded microphone needs to include resonant behavior due to vortex shedding. The database of unshielded microphone wind noise was based on audio recorded at known constant wind speeds at the exit of a silenced wind tunnel. The audio was recorded on an unshielded, calibrated $\frac{1}{4}$ in. measurement microphone. Simultaneously, a sonic anemometer (Metek USA-1) measured the wind speed. The fan speed was systematically adjusted and 1 min of audio was recorded for 43 different airflow speeds ranging from 0.5 to 15 ms^{-1}. A second reference microphone was placed just outside the tunnel but not in the air flow. The purpose of this was to capture the fan noise and so enable the removal of this from the wind noise recording using a Wiener filter.[11] The power spectra from these measurements indicated a mixture of the power-law relationships; an example is shown in Fig. 1.

A stochastic simulation was used to generate wind noise for a shielded microphone. A simulation was preferred over measurements because this allowed the windshield diameter to be easily varied. Van den Berg showed both theoretically and experimentally that the sound power level L_w for the one-third octave centered at frequency f, for a shielded microphone, is given by[10]

$$\mathrm{L_w(f)} = 40\log_{10}(\mathrm{v}) - 6.67\log_{10}(\mathrm{fD/V}) - 10\log_{10}(1+(3dfD/V)^2) + 42, \quad (1)$$

where V is wind speed and D is wind screen diameter.

The one-third octave power spectrum generated by Eq. (1) was converted to a linear frequency scale by linear interpolation of the sound power level spectrum over the log of the frequency. The 0 Hz value was set to 0. A time domain signal was generated by assigning a random phase to each

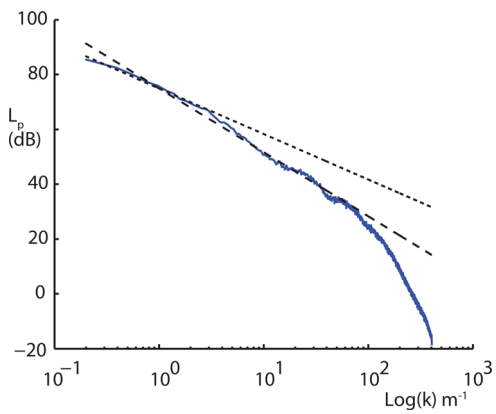

FIG. 1. (Color online) Example of a sound pressure level measurement using wind tunnel versus wave number (solid line). Also shown are the two power laws: $\mathrm{k}^{-5/3}$ (short dashes); $\mathrm{k}^{-7/3}$ (long dashes).

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

bin in the spectrum and applying an inverse Fourier transform.

Time-varying wind noise was generated by combining the outputs from the above models with wind speed time histories taken from anemometer data. This enabled wind noise samples to be generated with arbitrary wind speed time histories. The noise samples were processed in 100 ms windows with 50% overlap. In each 100 ms window, a noise sample from one of the above models was chosen that has the desired wind speed for that time in the anemometer time history. A Hanning window was applied to each of the 100 ms noise samples before they were mixed together into the final wind noise simulation.

The anemometer time histories were taken from the CASES-99 database.[12] This has measurements from eight sonic anemometers on five towers, three at 5 m and five at 2 m. Eight days of diurnal data were used, representing a wide range of wind conditions. The input parameters to the shielded model were; a wind speed time history and a wind shield diameter, the diameter for the wind meter example database was randomly varied between 2 cm and 20 cm, for the perceptual testing it was fixed at 5 cm. For the unshielded model the only input was a wind speed time history. Then, 12 000 examples, each 10 s long were generated using both models, i.e., with and without a windshield. Random sampling was used to select each time history from the CASES-99 wind speed database.

The resulting database of wind noise examples was used for both the perceptual tests and the development of the wind noise meter. Evidence of the validity of the models is demonstrated by testing the trained wind noise meter using examples of real field recordings of wind noise which were not used to train the algorithm.

III. PERCEPTUAL TESTS

To explore the relationship between wind noise and quality, a speech-in-noise task was used. Speech was selected as the foreground signal because of the prevalence of speech in audio recordings. Naive participants rated the quality of audio clips with controlled wind noise degradation. The experiments were carried out both in laboratory conditions and also across the Internet. The experiment was run in this way to allow comparison of the results from controlled laboratory conditions with data from more ecologically valid conditions where listeners auditioned sounds in everyday environments using a wide variety of consumer audio systems.

A. Laboratory test

Two key characteristics of wind noise were identified from the recordings: The level and the temporal variability or gustiness. Level is analyzed using the mean *A*-weighted sound pressure level, L_{Aeq}, which scales with wind speed. Research into wind noise inside vehicles have produced a measure for gustiness based on the ratio of the level of identified transients verses the background level.[13] In our work, however, it was felt that having a process to identify transients was an unnecessary complication. In metrology, the gust factor is defined as the ratio of the maximum velocity over a short window (for example, between 1 s and 5 s) to the hourly average.[14] As our interest is in the perception of gustiness, the gust factor definition has been refined to work with sound level rather than wind speed. The temporal variability was computed as the mean absolute difference between the L_{Aeq} over the whole 5 s sample and the L_{Aeq} in a moving 1 s window.

This measure of temporal variability was then converted into three classes representing low, medium, and high gustiness. Examples from the wind noise database with the same L_{Aeq} were grouped together, and the temporal variability parameter evaluated for all these samples. Samples with the lowest, the mean and the highest temporal variability values were selected to represent low, medium, and high levels of gustiness. (The standard deviation may seem like an obvious initial choice to quantify gustiness; however, standard deviation takes no account of the rate of change of the sound. Neither slow nor sudden changes in the average level would be perceived as gusts.)

The *A*-weighted wind noise levels were divided into eight equally sized adjacent groups, from 30 to 82 dB (at 6.5 dB intervals). The *A*-weighted speech level was set relative to the wind noise at 57 dB as this is the level of normal speech at 1 m.[15] For the gustiness, each of the eight wind level groups was evaluated separately and the range of temporal variability determined. Three examples closest to the minimum, mean, and maximum temporal variability were selected as the low, medium, and high levels of gustiness.

Audio samples for the psychoacoustic experiments were created from factorial combinations of wind level and gustiness, thus the test set consisted of 24 permutations of wind noise (three levels of gustiness and eight levels of wind noise), plus one additional sample with no wind noise present. The speech level for each sample was set to have an L_{Aeq} of 57 dB and then the wind noise added. Additionally, to prevent the possibility of participants recognizing particular signature patterns of gustiness across different samples, three variations were created for each level differing only in the temporal pattern. Which variant was heard by participants on each trial was randomized.

Each wind noise sample was paired with one of 25 spoken nonsense sentences from a subset of the corpus used by Picheny, Durlach, and Braida.[16] These sentences are grammatically correct but free of any meaningful semantic content. For example, "His quick world must pass in a flag." Each sentence contained four target words, unknown to the subjects. A measure of correctly identified words was obtained using the method from Picheny *et al*. Words were marked as incorrect if a single phoneme was omitted or misidentified. However, the incorrect addition or omission of suffixes "s," "ed," and "d" were not considered sufficient to count as an incorrectly identified word. Typos and misspellings were accepted as correctly identified words if the attempt was clear and unambiguous. Homophones of the target word (e.g., there, their, they are) were also accepted as correctly identified. Scoring of participants' submitted sentences was blind to the test condition they came from.

Thirty participants completed the test (mean = 28.3 yr, SD = 7.2 yr). None reported any known hearing impairment.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

The experiment was conducted in a large anechoic chamber. It was run on a laptop using a GUI written specifically for the task. An M-Audio MobilePre external soundcard connected the output of the computer to two powered loudspeakers, a Genelec 1029A loudspeaker paired with a Genelec 1091A subwoofer. The loudspeakers were positioned directly in front of the seated participants, 1.5 m away, with the 1029A positioned at head height and the subwoofer directly below on the floor. All samples were presented in mono, the playback level was set to 57 dBA at the listener position using a clean speech sentence.

Participants were informed that they were to be presented with spoken nonsense sentences but were given no information about the presence of wind noise on the samples. All responses were provided by participants on the test laptop using the keyboard and mouse. Playback of the samples and the rate of progression through the test were determined by participants. They were instructed to listen to each clip once and type the sentence they heard. They then had to rate both the difficulty of the task and the overall quality of each clip. For this rating task they could replay the sample if they desired. Difficulty and quality ratings were taken via user-controlled sliders which output values ranging from 0 to 100 for analysis.

The presentation order of sentences and order of wind noise permutations was fully randomized, as were the pairings of sentence and wind noise permutation in each sample. Before the test began, participants were presented with two practice trials. Participants were informed that the audio sample in one of these practice trials represented an example of the best quality of audio they would hear in the test (sample contained no wind noise) and that the other was an example of the worst quality they would hear (an example with the highest level of wind noise). The whole experiment took approximately 15 min to complete. Participants were paid for the time spent completing the experiment.

B. Web test

The web version followed the same format as the laboratory test. Participants were initially presented with an example of a spoken sentence to check their audio setup and instructing them to set their own comfortable level for playback. The interface for the web test had minor visual differences to that used in the lab, but the overall layout, function, and instructions were the same. An incentive for participants to complete the experiment was provided in the form of a prize draw for £10 vouchers. Responses gathered on the web were screened prior to analyses for non-serious participation. Any participants who progressed through three or more trials without entering text and/or moving the rating sliders for quality and difficulty were removed from the final sample. The final web test sample consisted of 5104 trials (mean = 4.26 trials per participant).

C. Results

Figures 2(a) and 2(b) present ratings for quality as a function of wind level in the laboratory and on the web, respectively. Curves represent each level of the gustiness variable. Figure 3 shows the difficulty ratings and the number of words typed by the participants that were correct versus wind level. Effect sizes for all significant main effects and interactions are summarized in Table I. The effect size is a measure of the magnitude of an effect, reflecting the proportion of variance explained by it. Higher values indicate stronger effects (typically, for this measure, effect sizes >0.01 are considered small, >0.06 medium, and >0.14 large).[17]

1. Laboratory

An 8 (wind noise level) $\times$ 3 (gustiness) repeated-measures analysis of variance (ANOVA) was performed on ratings of audio quality from the laboratory test. Mauchly's test indicated that the assumption of sphericity had been violated (i.e., the data do not support the assumption that variances are equal across conditions) for the effect of wind level [$\chi^2(27) = 46.62$, $p = 0.01$] and for the interaction of

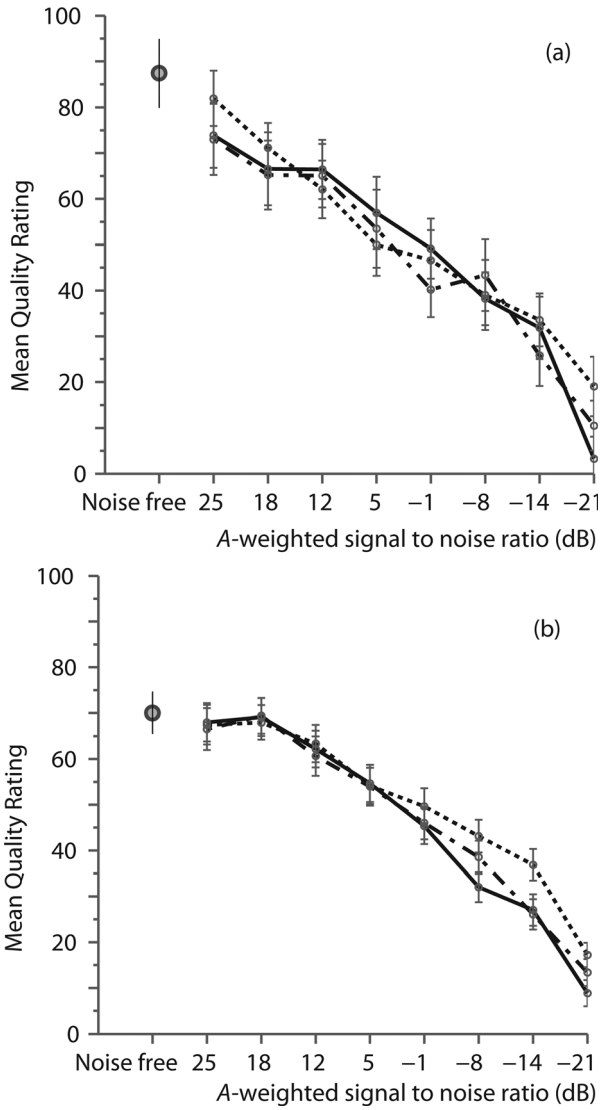

FIG. 2. Participants' ratings of audio quality by level of gustiness for (a) laboratory and (b) web. Values given for signal to wind noise ratio represent the mid-points of each of the eight windows the samples were drawn from. Lines represent: low gustiness (solid line); medium gustiness (dotted dashes), and high gustiness (short dashes). Error bars represent 95% confidence intervals.

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

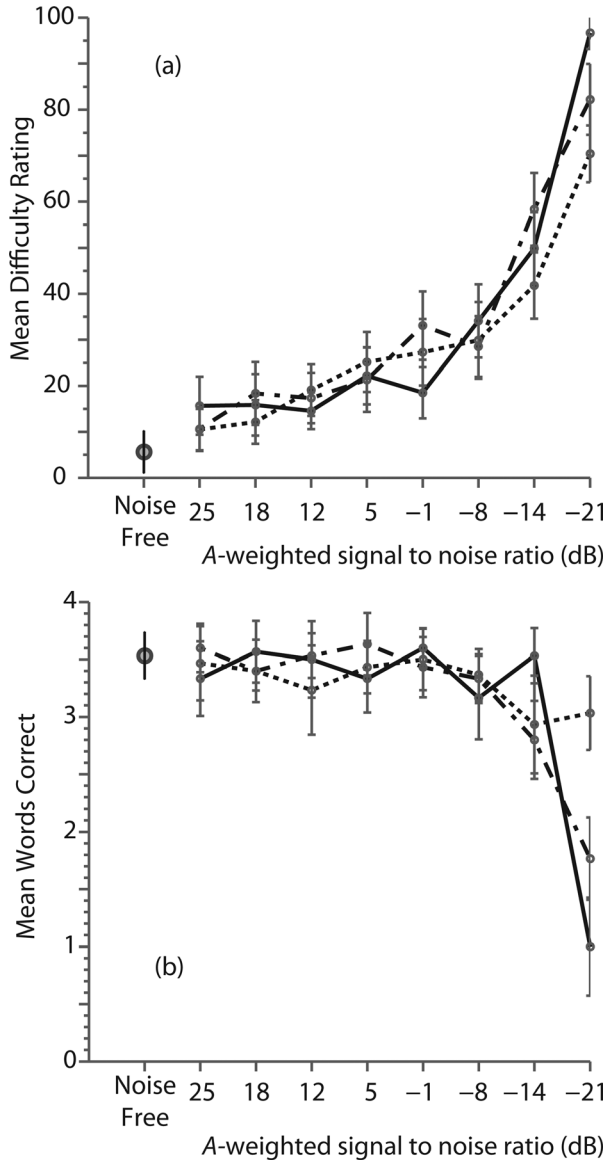

FIG. 3. Laboratory results for (a) listener's perception of task difficulty and (b) number of words correctly entered vs signal to wind noise ratio. Lines represent: low gustiness (solid line); medium gustiness (dotted dashes); and high gustiness (dashes). Error bars represent 95% confidence intervals.

TABLE I. Summary of effect sizes (partial η^2 squared) for the ratings of audio quality, difficulty of task and number of words correctly typed. Table shows main effects and interactions of the two experimental variables, wind level and level of gustiness. "NS" indicates the absence of a significant effect.

	Independent Variable		
Dependent Variable	Wind Level	Gustiness Level	Interaction (Wind Level × Gustiness)
Quality			
Lab	0.77	NS	NS
Web	0.30	<0.01	<0.01
Difficulty			
Lab	0.83	0.16	0.11
Web	0.35	<0.01	0.02
Word Scores			
Lab	0.61	NS	0.19

wind level and gustiness [$\chi^2(104) = 173.55$, $p < 0.001$]. Subsequently, Greenhouse-Geisser correction estimates for degrees of freedom were used for these analyses.[18] A very strong significant linear trend is observed in the data for wind level ($p = <0.001$, partial $\eta^2 = 0.91$), reflecting a consistent relationship between impairment of quality with each increase in level of wind noise.

A significant main effect of wind level was observed ($p < 0.001$, partial $\eta^2 = 0.77$). Gustiness was not found to have a significant effect on quality ratings ($p = 0.10$, partial $\eta^2 = 0.07$) and no interaction between wind level and gustiness was found ($p = 0.13$, partial $\eta^2 = 0.05$).

As Fig. 2(a) shows, overall, increases in wind level were associated with significant decreases in quality ratings independent of levels of gustiness. The significant main effect of the wind level variable was broken down by repeated contrasts for each successive level of the variable. Each successive increase in level of wind noise was associated with a significant decrease in quality ratings (all $ps < 0.01$, all partial $\eta^2 > 0.25$), with the exception of ratings at level 2 and 3, and levels 5 and 6, which were not found to significantly differ ($ps > 0.14$). This finding may be of relevance to future research as it implies that the just-noticeable difference for change in wind noise (or at least its effect on quality) will be below 6.5 dB.

A paired samples t-test showed that the condition with the lowest level of wind noise (the highest quality rating of any of the wind noise conditions, $M = 76.28$, $SE = 2.92$) had a significantly lower quality rating than the noise-free condition [$M = 87.40$, $SE = 4.00$, $t(29) = 2.76$, $p = 0.01$], demonstrating the sensitivity of perceptions of quality to the presence of wind noise. Indeed, this finding suggests the threshold at which wind noise begins to affect perceptions of quality is above a signal to noise ratio of 25 dB.

Figure 3(a) shows participants' mean ratings for the difficulty of the task of identifying the words versus wind noise level. The difficulty results are very similar to those for quality. Ratings were analyzed with the same procedure as the quality ratings. An 8 (wind noise level) × 3 (gustiness) repeated-measures ANOVA was performed, with difficulty ratings as the dependent variable. Mauchly's test indicated that the assumption of sphericity had been violated for the effect of wind level [$\chi^2(27) = 66.41$, $p < 0.001$] and for the interaction of wind level and gustiness [$\chi^2(104) = 202.78$, $p < 0.001$]. Greenhouse-Geisser correction estimates for degrees of freedom were used as necessary.

Significant main effects were observed for both the gustiness variable ($p = 0.01$, partial $\eta^2 = 0.16$) and for the wind level variable ($p < 0.001$, partial $\eta^2 = 0.83$), and also for the interaction of these two variables ($p < 0.01$, partial $\eta^2 = 0.11$). The relative effect sizes of these factors however suggest wind level is by far the most important influence on task difficulty.

Breaking down the interaction with repeated contrasts of wind level, by levels of gustiness, suggests the interaction occurs due to differences in perceived difficulty which emerge at higher levels of wind noise. For signal to noise ratios equal to or better than −1 dB, we do not observe any significant change in difficulty ratings across successive levels of wind noise (all $ps > 0.06$, all partial $\eta^2 < 0.12$). The

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

emerging interaction between gustiness and wind level is most easily understood with reference to Fig. 3(a), where it is clear that the rate of change in the curves for gustiness differ as wind level increases above a signal to noise ratio of −1 dB. Most notable in this respect, is the difference in ratings at the highest level of wind noise, where the task difficulty is perceived to increase considerably as gustiness moves from high ($M = 70.41$, $SE = 3.29$), to medium ($M = 82.25$, $SE = 3.91$), to low ($M = 96.66$, $SE = 1.19$).

Mean word scores for the number of words correctly typed are presented in Fig. 3(b). Each curve represents a different level of gustiness. Participants' word scores for each condition were analyzed in an 8 (wind noise level) × 3 (gustiness) repeated-measures ANOVA. Greenhouse-Geisser correction estimates were used where the assumption of sphericity had been violated for the effect of wind level [$\chi^2(27) = 43.15$, $p = 0.027$] and the interaction of wind level and gustiness [$\chi^2(104) = 183.88$, $p = <0.001$].

Significant effects were observed for wind level ($p < 0.001$, partial $\eta^2 = 0.61$) and for the interaction of wind level and gustiness ($p < 0.001$, partial $\eta^2 = 0.19$).

The word scores results are different from both the perceived quality and difficulty scales, with performance only decreasing rapidly for low signal to noise ratios. Contrasts of successive levels of wind noise show that participants' performance was not affected as wind level increased, apart from a significant decrease in word scores between −14 dB ($M = 3.09$, $SE = 0.14$) and −21 dB signal to noise ratio ($M = 1.93$, $SE = 0.12$, $p < 0.001$, partial $\eta^2 = 0.70$). The significant interaction between wind level and gustiness on word scores arises from differences in the relative impact of the highest wind levels at different levels of gustiness. Wind noise which is more consistent has a significantly greater impact on performance than wind which is gusty. This effect is most clearly seen in differences in word scores at the highest level of wind, −21 dB signal to noise ratio, for low ($M = 1.00$, $SE = 0.22$), medium ($M = 1.77$, $SE = 0.18$), and high ($M = 3.03$, $SE = 0.21$) levels of gustiness.

2. Web

A similar analysis was carried out for the web results for quality. An 8 (wind noise level) × 3 (gustiness) ANOVA was performed on the web test data, with quality rating as the dependent variable. Significant main effects were found for both variables and for their interaction, every $F > 2.21$, every $p < 0.01$. While gustiness was found to have a statistically significant effect, the size of this effect (see Table I) is considered trivial relative to that for wind noise.

Post hoc comparisons (Bonferroni corrected) of the levels in each variable participants' ratings indicate no reduction in quality is perceived until the sample at 12 dB signal to noise ratio, suggesting the threshold for quality degradation due to wind noise is between 18 and 12 dB SNR. Successive increases in wind noise beyond this threshold were each associated with significantly worsening ratings of quality. Overall, quality was found to significantly decrease as wind level increased, but marginally less so in gustier samples.

Prior to starting the web test, participants were asked about themselves and the environment within which they were completing the test. The four questions asked about: The sound reproduction equipment; how noisy the place where the experiment was being carried out was; the participant's age, and whether the participant considered themselves to be an audio expert. While significant differences were observed for many of the categories of participant information[19] it is notable that effect sizes were small across the board (all effect sizes < 0.017, partial η^2), relative to the effect size of the experimental wind level variable across the group (partial $\eta^2 = 0.30$).

D. Discussion

Overall, different levels of gustiness were not found to influence perceptions of audio quality (other than very marginally at very high levels of wind noise). Increases in wind level, however, were found to have a very large negative effect on perceptions of quality once above the detection threshold. Consequently, the development of the wind noise meter focused on the wind noise level as a measure of quality, and did not consider gustiness.

The signal to noise ratio below which wind noise levels significantly decreased the number of words correctly identified, is much lower than the signal to noise ratio below which quality and difficulty of task were perceived to be different from the noise-free case. The self-reported difficulty of the task is different from the actual measured performance. Consequently, speech intelligibility has some independence from quality, which appears more closely related to subjective perceptions of difficulty. The differences in the number of words correctly identified by gustiness at the highest levels of wind noise, also reveals some insight as to why there are only differences in quality ratings between gustiness at the worst signal to noise ratios. The word score results suggest it is slightly easier to hear words when there is gusting than when the wind is more constant.

This study allows some insight into the viability of internet-based tests for the collection of subjective ratings of quality. Despite the sacrifice of experimental control and non-optimal playback conditions via the web, the results from the web-based test generally mirrored those obtained in the lab. Many studies of web versus lab experimentation have found results replicate (for example, see review in Ref. 20) but these studies have tended to be investigations of universal cognitive processes, such as short-term memory or reaction times, for instance. Few existing studies have compared web and laboratory data in subjective judgments of quality. Our findings suggest that the efficiency gains of testing on the web (in terms of mass participation) ensure that similar results to laboratory testing can be achieved despite the non-optimal, non-controlled test conditions. The smaller effect sizes found for the effect of wind level on the web are a predictable consequence of the diversity of sources of variability and error compared to the laboratory (environment, background noise, playback equipment, playback mode, etc). This finding suggests that Web-based experiments would likely be less successful for quality judgments where

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

degradations are small or differences between stimuli are marginal (such as codec comparisons, for example). However, for experiments such as the current study, where the worst degradations were very large, the diversity of testing conditions experienced by Web participants arguably also serves to increase the ecological validity of tests, as a wider, more representative pool of participants can be reached and, importantly, participants are likely to complete the quality assessments on the same equipment and in the same environments within which they typically listen to audio.

IV. WIND NOISE METER

Figure 4 shows a schematic of the model used to estimate the perceived quality of audio degraded by wind noise. It follows a common approach in audio signal processing by first extracting Mel Frequency Cepstrum Coefficients (MFCCs) in short time frames, then classifying each frame according to the level on wind noise using a machine learning approach, before aggregating the results over a longer window. This results in a simple algorithm. First, the algorithm works on short 23 ms frames (1024 samples to allow a fast Fourier Transform to be used) overlapping by 50%. It attempts to classify the wind noise level into four categories denoted Class_L:

(a) $\text{Class}_\text{L} = 0$: $L_{Aeq} < 30$ dB (no wind);
(b) $\text{Class}_\text{L} = 1$: $30 < L_{Aeq} \leq 50$ dB (low wind);
(c) $\text{Class}_\text{L} = 2$: $50 < L_{Aeq} \leq 70$ dB (moderate wind);
(d) $\text{Class}_\text{L} = 3$: $L_{Aeq} > 70$ dB (high wind).

The lowest class boundary was derived from the perceptual measurements. These indicated that when the wind noise level was around 20 dB lower than the speech, there was little effect on quality. As hushed speech has an *A*-weighted level of about 50 dB at 1 m,[15] this implies that 30 dB is an appropriate boundary for the "no wind" threshold. Loud speech at 1 m has an *A*-weighted level of about 70 dB,[15] therefore wind noise at or above this level was considered "high wind." The other class boundary was placed at an equal level distance between the other two.

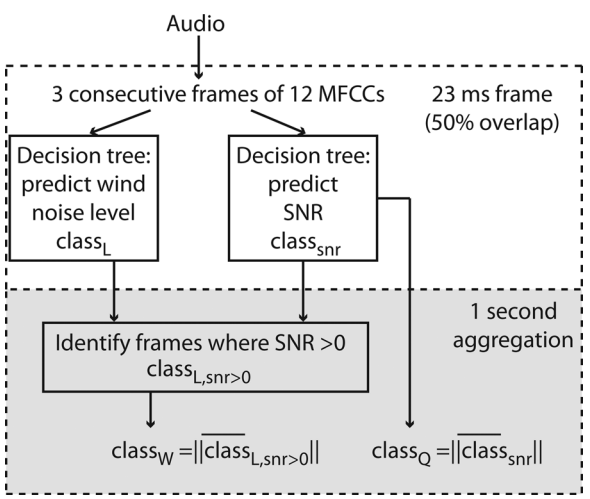

FIG. 4. Schematic of the wind noise meter.

While the absolute level of wind noise can be used to indicate the presence or lack of wind noise, the perceptual measurements showed the signal level of wind noise ratio correlates with perceived quality. Therefore, a separate decision tree was trained to classify frames according to the *A*-weighted signal to noise ratio (SNR). Six categories of signal to noise ratio are defined, denoted Class_SNR:

(a) $\text{Class}_\text{SNR} = 0$: $\text{SNR} < -20$ dB;
(b) $\text{Class}_\text{SNR} = 1$: $-20 < \text{SNR} \leq -10$ dB;
(c) $\text{Class}_\text{SNR} = 2$: $-10 < \text{SNR} \leq 0$ dB;
(d) $\text{Class}_\text{SNR} = 3$: $0 < \text{SNR} \leq 10$ dB;
(e) $\text{Class}_\text{SNR} = 4$: $10 < \text{SNR} \leq 20$ dB;
(f) $\text{Class}_\text{SNR} = 5$: $\text{SNR} > 20$ dB.

The SNR class divisions were informed by the perceptual test results. These indicated that over the SNR range from −20 to +20 dB, the quality score changed from about 90% to 10%.

The training data used calibrated models where the time history represented the pressure in Pascals induced at the microphone due to the wind. Therefore, recordings need to be scaled so that they were representative of the pressure recorded by the device. As recording devices are normally un-calibrated, the sound pressure level of any recorded audio may not be known exactly. In these cases, the signal can be calibrated by scaling the audio according to some known reference within the recording. For example, a rough calibration can be carried out by reciting a sentence at a normal speaking level into the device from one meter away in a quiet environment. The average normal speaking sound pressure level, without specifying gender, is about 57 dBA.[15]

Two ensembles of decision trees were trained to classify Class_SNR and Class_L using bagging.[21] Initial results showed that it is difficult to classify Class_L when it is masked by the foreground sound being recorded. Consequently, samples where the wind noise is quieter than the foreground sound are identified using Class_SNR, and this information is used to improve the accuracy of the meter when aggregating over a number of frames.

MFCCs are used as acoustic features to input to the decision trees as they have been shown to work well for other audio classification tasks.[22] A Hanning window is applied to each frame and 50% overlap used. The power spectrum for each frame is computed via Fourier transform, and then the Mel power spectrum computed using a triangular filter bank with 16 bands between 0 Hz and 8 kHz spaced evenly over the Mel scale,

$$f_{mel}(f) = 2595 \log_{10}(1 + f/700). \quad (2)$$

The MFCCs are the DCT (discrete cosine transform) of the log Mel power spectra; the first 12 of 16 coefficients are selected for each frame. The first MFCC is replaced by the *A*-weighted decibel value for the frame to indicate the sound pressure level. Features from the current and the next two frames are used as inputs to the decision trees. The addition of the next two frames provides the classification algorithm with information regarding the evolution of the sound over time. Class_L and Class_SNR are determined from the first

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

frame in this sequence. The same set of 36 features was used for both trees.

The MATLAB function "treebagger" was used for the ensemble supervised training of the decision trees.[23] One hundred decision trees were trained on 100 subsets of the whole training dataset selected by randomly sampling with replacement, so that the size of the subsets was the same as the original dataset. For every class decision, 6 out of 36 features were randomly selected to ensure instability in the trained trees, and the trees were not pruned. This approach is equivalent to the random forest method.[24] The resulting ensemble uses winner-takes-all voting to determine the class. Bagging has been shown to reduce the chance that a model will be overly simple. While optimization of the random forest's meta-parameters was not investigated, it is likely that this will be required for real-time implementation.

The aggregation stage outputs an estimation of the wind noise level category for a one second interval. An average of the estimated Class_{L} for the wind noise level for each frame is taken. The average is computed only using frames tagged by the second decision tree as containing wind noise at a level greater than the foreground audio. When no suitable frames are available within the 1 s interval, it is assumed the sample is free of wind noise. The average wind noise class is rounded to the nearest integer so that wind noise level over one second can be classified as none, low, moderate or high. A second aggregation is carried out to estimate the signal to noise ratio, where $\text{Class}_{\text{SNR}}$ is averaged over 1 s then rounded to the nearest integer so that the signal to noise ratio, and by association, degradation of quality can be classified.

A. Training and testing databases

Two databases were created, one for training the wind noise meter the other for testing the performance. The databases consist of foreground audio examples of speech, music, and other everyday sounds and soundscapes, some of which were corrupted with different levels of wind noise. The foreground audio in the training and testing sets were from different sources. Additionally, different sources of wind noise were used for training and testing. The algorithms were trained using only simulated wind noises, whereas the test set consisted of only real wind noises recorded on a range of devices. This was to prevent overly optimistic performance indicators being reported and ensures that reported performance is generalizable to wind noise generated from other devices and foreground audio sources.

In total, 633 samples of different foreground sounds that were the target of the simulated recording were used. This included 211 samples of male and female speech; 211 samples of music from a diverse variety of different genres; and 211 samples of everyday sounds such as animal vocalizations, traffic noise and crowd sounds. From each of the 633 sounds, a 10-s segment was selected at random to train or test the algorithms. In total there were 105 min of foreground sounds. The *A*-weighted sound pressure level of the foreground sounds were scaled so that they varied from 30 to 130 dB.

The foreground audio was corrupted with wind noise and training and test databases were segmented using tenfold cross validation. Folds were made according to the source of the foreground audio so that a particular 10-s audio sample used in training was never used in testing. For each fold, 570 examples were used in training and 63 in testing.

For the training data, unique examples of wind noise were generated using the methods described in Sec. II and added to each of the foreground sounds. Data was generated using average wind speeds over the 10-s samples ranging from 0 to 20 ms^{-1} spaced at 2 ms^{-1} intervals. The same number of samples was made for each wind speed. The simulated *A*-weighted wind noise level varied from 0 to 120 dB. A 4th order Butterworth high pass filter with a −3 dB point randomly chosen between 30 to 130 Hz was applied to the samples to simulate the different frequency responses of consumer recording devices and microphones. A sample with no wind noise was also used in the training. Hence, the training database consisted of the wind-free case, plus simulations for microphones with and without windshields, both with and without foreground audio.

As training is computationally costly, the size of the training database was reduced to 400 000 frames. This was achieved by undersampling the dataset where one-third of the data is wind noise free, one-third of the data contains only noise and one-third contains a mixture of both whose signal to noise ratios are uniformly distributed between 50 and −50 dBA and with an equal number of examples generated by each model.

A set of wind noises not present in the training was used for testing. Field measurements were made outdoors in windy conditions using a variety of audio devices. The wind noise was measured for 120 min on a small but broad range of microphone and device types: an unshielded microphone (B&K $\frac{1}{4}$ in.); the same microphone with a windshield; a portable recorder with electret condenser capsules (a Zoom H2 recorder set at its lowest input gain with automatic gain control turned off); a dynamic microphone (Shure SM58), and a mobile phone (iPhone 4). The effects of dynamic range control on the detection and perception of quality are beyond the scope of this study and will be investigated in further work.

An exposed, quiet spot, high up on the West Pennine Moors near Manchester, UK, was used. First, both measurement microphones were calibrated using a calibrator which produced 93.6 dB at 1 kHz. To calibrate the other devices, a 1 kHz tone was played over a loudspeaker at about 1 m and recorded over all devices. After applying a bandpass filter with a width of 200 Hz and center frequency of 1 kHz, the rms values were then used to calibrate the sensitivity of each device relative to one of the measurement microphones.

Table II gives a statistical description of the sound pressure levels of the wind noise. In the case of the iPhone and the shielded microphones, the distribution of levels is asymmetrical. For the iPhone this is due to the automatic gain control which is always active. For the shielded microphone, this is due to the background noise level at the site. The background noise level was less than 37 dBA (the wind noise levels were so high it was impossible to get an audio

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

TABLE II. Statistics for wind noise measurements used to test wind detector performance and models used for training.

	L_{A90}	L_{A50}	L_{A10}	L_{Aeq}
Unshielded	63.6	74.3	83.6	80.1
Shielded	36.3	42.2	50.2	47.5
Zoom	62.3	74.4	84.2	80.1
SM58	60.4	71.6	81.6	78.4
iPhone	66	81.8	90.4	85.8
Training models	30.5	80.8	113.6	107.9

sample free from wind noise even with windshields on the microphones, so the exact background noise level is unknown).

For each device, from the 120 min of recorded wind noise, 10 s of wind noise was selected at random and added to the foreground audio. Wind noise samples both with and without the foreground audio were included in the testing database. This produced a database where for each 10 s of foreground audio there were 15 variations, the wind-free case plus the following cases both with and without the foreground audio: Unshielded measurement microphone; shielded measurement microphone; Zoom H2; Shure sm58; and iPhone. The performance was evaluated for each device separately.

B. Validation tests

Performance was evaluated using the Matthews Correlation Coefficient (MCC).[25] Figure 5 shows the wind level classification performance over a range of devices. The classification performance was better for the event sounds than either music or speech. This is because the event sounds contain less low frequency content than the other sounds. Wind noise is dominated by lower frequencies; therefore, detection is more successful when the foreground audio is free of low frequencies. The shielded microphone performance was lower because of the background noise at the recording site, which meant that some cases in the test set with

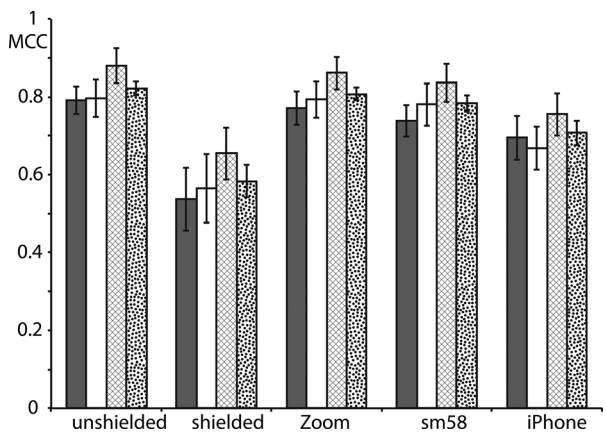

FIG. 5. Wind noise detection performance gauged by MCC for each of the five devices, error bars represent the 95% confidence limits over all ten folds. Foreground sounds: speech (gray); music (white); event sounds (cross hatch); and all (dots).

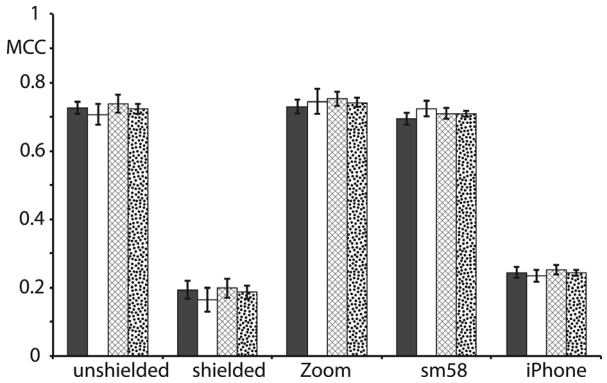

FIG. 6. Signal to noise ratio classification performance gauged by MCC for each of the five devices, error bars represent the 95% confidence limits over all ten folds. Foreground sounds: speech (gray); music (white); event sounds (cross hatch); and all (dots).

no wind noise present were mislabeled as containing low levels of wind noise.

Figure 6 shows the performance of the SNR classifier. Performance was poor for the iPhone and shielded microphone cases, whereas the correlation coefficient is about 0.7 for the other three devices. For the shielded microphone, this is because the signal to noise ratio value used to classify the frames is inaccurate, due to the presence of sound at the recording site at a comparable level to the lowest levels of wind noise. For the iPhone, as the wind noise levels were much higher than the background noise, an alternative explanation is required. Investigations indicated that is not the limited frequency response of the device that is causing this, and therefore, there is some aspect of the iPhone wind noise that is not captured by the model. This is either the presence of an automatic gain control system, not present on any other device or in the training data, or some unique feature of the wind noise, for example, a particular vortex shedding resonance not captured by the model.

Table III shows the confusion matrix for the aggregated wind noise level classifier over all devices averaged over all folds. The sensitivity is the percentage of correctly identified wind noise cases. This was significantly lower for the iPhone than the other devices, being 92% for the iPhone; 77% for the shielded case; 98% for the others; and 93% overall. The specificity is the percentage of the correctly identified wind-free cases. This was found to be the similar for all devices at 79%. As the shielded case's inaccuracy is due to the faulty assumption of no background noise at the measurement site, the detector identifies at least 92% of the true wind cases. Identification of wind noise free cases is lower at 79% and is

TABLE III. Confusion matrix, wind noise level classification for all devices.

	None detected	Low detected	Medium detected	High detected
No wind noise in sample	3413	159	217	443
Low wind noise in sample	85	683	23	1
Medium wind noise in sample	74	26	986	33
High wind noise in sample	300	2	276	2779

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

TABLE IV. Confusion matrix, signal to noise ratio classification, iPhone and the shielded microphone data separated from the rest.

		Predicted SNR class (excluding iPhone and shielded microphone)							Predicted SNR class (just iPhone and shielded microphone data)						
		SNR < −20	−20 < SNR < −10	−10 < SNR < 0	0 < SNR < 10	10 < SNR < 20	20 < SNR		SNR < −20	−20 < SNR < −10	−10 < SNR < 0	0 < SNR < 10	10 < SNR < 20	20 < SNR
Actual SNR Class	SNR < −20	784	50	11	3	1	2		10	65	175	210	154	169
	−20 < SNR < −10	7	29	30	12	4	1		0	1	6	13	27	29
	−10 < SNR < 0	2	3	19	31	14	3		0	1	2	6	15	53
	0 < SNR < 10	0	0	2	12	38	19		0	0	0	1	9	57
	10 < SNR < 20	0	0	0	1	13	62		0	0	1	1	7	70
	20 < SNR	1	2	3	7	31	703		1	2	4	8	35	776

device independent. This error is due to how well the training database of foreground sounds represents the test database. To decrease the false negative rate, the database of sounds would need to be expanded further.

Table IV shows two confusion matrices for the SNR classification, one excluding the iPhone and shielded data and one showing just the iPhone and shielded data. This allows us to examine the poor performance of the iPhone and shielded microphone results; there is a bias toward the overestimation of the signal to noise ratio. To account for this error, the final wind noise meter will combine the training and test sets to improve the generalizability of the resulting algorithm.

V. CONCLUSIONS

Perceptual tests were carried out to examine how wind noise affects the perceived quality of recorded speech. Tests were carried out both in controlled laboratory conditions and also across the Internet. The pattern of results for both experiments was similar. The trade-off between lack of experimental control and access to a very large, more representative sample via the Internet is reflected in a smaller effect size for the effect of wind level on quality ratings alongside additional significant effects (albeit of very small magnitude) not observed in the laboratory.

Increases in wind level were found to have a large negative effect on perceptions of audio quality below an *A*-weighted signal to noise ratio of approximately 15 dB. Changes in the level of gustiness, in contrast, were not found to influence quality perceptions (other than very marginally at very high levels of wind noise). Consequently, wind noise level can be considered sufficient to predict degradations in audio quality. Participants were also asked to type the words they heard during the test. The number of words typed correctly significantly decreased when the *A*-weighted signal to noise ratio was −18 dB. For many signal to noise ratios, the wind noise has a greater effect on perceived quality than it does on the ability of the subjects to correctly identify the words being spoken. Additionally, it was observed that participants' ratings of the difficulty of the task more closely reflected quality ratings than actual task performance. This finding may have implications for similar work where task performance is commonly used to assess or predict audio quality.

A meter to predict the perceived quality of recordings in the presence of wind noise was developed using a machine learning algorithm that had MFCCs as input features for bagged decision trees. The algorithm was designed to work with a single microphone and also to detect wind noise when there is limited low frequency information as some common consumer devices automatically filter out the prominent low frequencies present in wind noise. The algorithm was designed to work with a wide variety of foreground sounds: music, speech and quotidian sounds.

The algorithm worked in short 23 ms frames, with these results then being aggregated over 1-s intervals. The algorithm was designed to produce an estimation of wind noise in four classes: none, low, medium, and high. The decision trees were trained using two models that simulate devices with and without windshields. The performance of the wind

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25

noise meter was tested using a set of field measurements on five different devices. The wind noise detector accurately identified wind free examples in 79% of cases. For the three classes with noise present, on average 93% were correctly assigned to the appropriate category. A second decision tree was trained to estimate the signal to noise ratio, from which the perceived degradation to quality can be inferred. This achieved a Matthew's correlation coefficient of 0.7 for three of the devices. Poor performance for the shielded microphone was due to background noise at the recording site, while the poor performance with the iPhone is probably due to some aspect of the wind noise not captured in the model.

A version of the wind noise meter trained with both the training and test data sets is available as an open source C++ program at http://www.goodrecording.net/wind-noise-detection-open-source-program/.

ACKNOWLEDGMENTS

The authors would like to thank EPSRC for funding the project (EP/J013013/1) and our project partners BBC R&D and the British Library.

[1]S. Kochkin, "MarkeTrak VIII: Customer satisfaction with hearing aids is slowly increasing," Hear. J. **63**, 11–19 (2010).
[2]N. Otto and B. Feng, "Wind noise sound quality," SAE Tech. Pap. 951369 (1995).
[3]F. Luo and A. Nehorai, "Recent developments in signal processing for digital hearing aids," IEEE Signal Process. Mag. **23**, 103–106 (2006).
[4]Intricon Corp., "Scenic: Wind noise reduction," (2010), available at http://www.intricon.com/assets/documents/uploads/Info_Sheet_-_Wind_Noise_Reduction.pdf (Last accessed 26 January 2014).
[5]E. Nemer, "Single-microphone wind noise reduction by adaptive post-filtering," in *2009 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (2009), pp. 177–180.
[6]S. Bradley, T. Wu, and S. von Hunerbein, "The mechanisms creating wind noise in microphones," in *114th Audio Engineering Society Convention* (2003), Vol. 3, pp. 1–9.
[7]L. Xiaoqiang, L. Shuangtian, and Y. Jie, "Convolutive sparse non-negative matrix factorization for windy speech," in *2010 IEEE 10th Int. Conference on Signal Processing (ICSP)* (2010), pp. 494–497.
[8]M. Schmidt, J. Larsen, and F.-T. Hsiao, "Wind noise reduction using non-negative sparse coding," in *2007 IEEE International Workshop on Machine Learning for Signal Processing* (2007), pp. 431–436.
[9]R. Raspet, J. Webster, and K. Dillion, "Framework for wind noise studies," J. Acoust. Soc. Am. **119**, 834–843 (2006).
[10]G. P. van den Berg, "Wind-induced noise in a screened microphone," J. Acoust. Soc. Am. **119**, 824–833 (2006).
[11]N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* (Wiley, New York, 1949), 176 pp.
[12]D. Fritts, G. Poulos, and B. Blumen, "Cooperative atmosphere-surface exchange study - 1999," available at http://www.eol.ucar.edu/projects/cases99/ (Last accessed 26 January 2014).
[13]M. Blommer, S. Amman, S. Abhyankar, and B. Dedecker, "Sound quality metric development for wind buffeting and gusting noise," SAE Technical Paper, 2003-01-1509 (2003).
[14]F. K. Davis and H. Newstein, "The variation of gust factors with mean wind speech and with height," J. Appl. Meteorol. **7**, 372–378 (1968).
[15]I. R. Cushing, F. F. Li, T. J. Cox, K. Worrall, and T. Jackson, "Vocal effort levels in anechoic conditions," Appl. Acoust. **72**, 695–701 (2011).
[16]M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech," J. Speech Hear. Res. **28**, 96–103 (1985).
[17]J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Erlbaum, Hillsdale, NJ, 1988), 590 pp.
[18]S. W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," Psychometrika **24**, 95–112 (1959).
[19]I. R. Jackson, P. Kendrick, T. J. Cox, B. Fazenda, and F. Li, "Perceived quality of speech degraded by wind noise: An assessment of sources of variability in a Web experiment," in *Proceedings of the 4th International Workshop on Perceptual Quality of Systems* (2013), pp. 46–49.
[20]M. H. Birnbaum, "Human research and data collection via the Internet," Annu. Rev. Psychol. **55**, 803–832 (2004).
[21]L. Breiman, "Bagging predictors," Mach. Learn. **24**, 123–140 (1996).
[22]B. Milner, "Robust acoustic speech feature prediction from noisy Mel-Frequency Cepstral Coefficients," IEEE Audio, Speech, Language. Process. **19**, 338–347 (2011).
[23]MATLAB version 8.2.0 (The MathWorks Inc., Natick, MA, 2013).
[24]L. Breiman, "Random forests," Mach. Learn. **45**, 1–33 (2001).
[25]G. Jurman, S. Riccadonna, and C. Furlanello, "A comparison of MCC and CEN error measures in multi-class prediction," PloS one **7**, e41882 (2012).

Redistribution subject to ASA license or copyright; see http://acousticalsociety.org/content/terms. Download to IP: 146.87.136.26 On: Mon, 20 Oct 2014 12:41:25