Forensic voice discrimination: The effect of speech type and background noise on performance

Harriet M. J. Smith[1], Thom S. Baguley[1], Jeremy Robson[2], Andrew K. Dunn[1], and Paula C. Stacey[1]

[1]Department of Psychology, Nottingham Trent University, UK

[2]Nottingham Law School, Nottingham Trent University, UK

Corresponding author:

Harriet M. J. Smith, Department of Psychology, Nottingham Trent University, 50 Shakespeare Street, Nottingham, NG1 4FQ, UK.

Email: harriet.smith02@ntu.ac.uk, Tel: +44 (0)115 848 4535

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/acp.3478

Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance

In forensic settings, lay (non-expert) listeners may be required to compare voice samples for identity. In two experiments we investigated the effect of background noise and variations in speaking style on performance. In each trial, participants heard two recordings, responded whether the voices belonged to the same person, and provided a confidence rating. In Experiment 1, the first recording featured read speech, while the second featured read or spontaneous speech. Both recordings were presented in quiet, or with background noise. Accuracy was highest when recordings featured the same speaking style. In Experiment 2, background noise either occurred in the first or second recording. Accuracy was higher when it occurred in the second. The overall results reveal that both speaking style and background noise can disrupt accuracy. Whilst there is a relationship between confidence and accuracy in all conditions, it is variable. The forensic implications of these findings are discussed.

Keywords: voice discrimination, confidence, accuracy, unfamiliar voice perception, background noise

Voices convey diagnostic identity information (Belin, Bestelmeyer, Latinus & Watson, 2011; Belin, Fecteau & Bedard, 2004; Mathias & von Kriegstein, 2014; Stevenage, Howland & Tippelt, 2011; Stevenage, Hugill & Lewis, 2012). In criminal investigations, when the perpetrator of a crime is heard but not seen, the degree of match between vocal information provided by the perpetrator and suspect constitutes 'forensic' evidence. In many cases where voice identity is disputed, a recording exists of the perpetrator. This allows direct comparisons to be drawn between voices, either by an expert phonetician, or by non-expert lay listeners, such as the police or jury members. Voice discrimination might involve either comparing different recordings to ascertain whether there is a common speaker, or comparing a single recording to the voice of a known suspect to ascertain whether they match. Despite the significant impact that a positive (or negative) match may have on the decisions made during the course of a criminal investigation, relatively few studies have addressed the ability of lay (i.e. non-technical/non-expert) listeners to make accurate comparisons. The work presented in this paper focuses on voice discrimination performance by lay listeners. Specifically, we were interested in how accuracy and confidence are affected by (1) varying the speech type, and (2) introducing background noise to recordings.

Legal application

In order to highlight the importance of this research, it is worth providing more detail about the potential role of voice discrimination decisions made by lay listeners in determining the outcome of criminal cases. Improvements in technology used to collect and preserve recordings of voices, coupled with widespread use of telephones, has increased the variety of circumstances in which correctly ascertaining the identity of a speaker is a core part of a criminal investigation. Although we primarily use the law in England and Wales to illustrate this point, voice discrimination by lay listeners is of global relevance (Edmond, Martire & San Roque, 2011; Morrison, Ochoa & Thiruvaran, 2012).

In Anglo-American legal systems, an individual can be convicted of a serious criminal offence solely on the basis of voice identification evidence (Edmond et al., 2011; McGorrery & McMahon, 2017; Robson, 2017). Examples in English case law of lay listeners being asked to decide whether the prosecution have proved that a voice recording features the disputed speaker include *R. v Shannon Tamiz and others* (2010), and *R v Kapikanya* (2015). In the leading judgment of *R. v Flynn and St John* (2008), the Court of Appeal stressed both the importance of adopting a cautious approach to the use of lay listener evidence, and the desirability of an expert witness providing evidence about whether samples feature the same individual. However, lay listener decisions still have the potential to play an important role. In cases where expert evidence is contested, the recordings can be played to the jury to demonstrate the methodology used by the expert (*R. v Suleman*, 2012). Where expert evidence of any type is admitted, it is treated as opinion evidence and juries are told it is a matter for them whether they accept or reject the conclusions of the expert (Judicial College, 2017). Courts in other jurisdictions, such as Northern Ireland (*R. v O'Doherty,* 2003) and Australia (*Bulejcik v. R*, 1995) actively endorse juries engaging in their own matching exercise without expert guidance.

The Crown Court Compendium (Judicial College, 2017) requires judges to direct juries to consider factors such as audibility, and whether or not the recording was made via a telephone, but does not assist with the extent to which the quality of a recording might have an impact on the assessment of evidence. Therefore, although juries are warned of the need to scrutinize recording quality before allowing a comparison to be made from it, the criteria by which this is assessed are not fully and consistently articulated. Further research into voice discrimination performance by lay listeners is required to ensure that the assumptions which underpin the current legal framework for decision making are accurate, and that the rules of evidence are sufficiently robust to prevent erroneous decisions being made.

Voice discrimination performance

Despite the legal weight placed on voice discrimination evidence, it is not altogether clear from existing psychological research how accurately lay listeners might be expected to perform identity verification in a court setting. Clarification is urgently required. On the one hand, it seems likely that accuracy would be relatively low. In comparison to faces, voices provide weak cues to identity **(**Hammersley & Read, 1996; Legge, Grossmann & Pieper, 1984; McAllister, Bregman & Lipscomb, 1988; Stevenage & Neil, 2014). This is likely attributable to voices primarily being encoded for content rather than identity-specific sound quality information (Fenn, Shintel & Atkins, 2011; Vitevich, 2003). Computational models suggest there is differing link strength in face compared to voice perception pathways (Damjanovic & Hanley 2007; Hanley & Turner 2000; Stevenage et al., 2012) and that mental representations of voices are more weakly encoded than faces (Stevenage et al., 2011; Stevenage, Neil, Barlow, Dyson, Eaton-Brown & Parsons, 2013). It may therefore be particularly difficult to compare identity information across two separate utterances, as is necessary in a voice discrimination task.

On the other hand, previous studies suggest that humans appear to be able to perform voice discrimination relatively accurately in certain conditions. Error rates in different conditions commonly vary between around 5% and 15% (Bartle & Dellwo, 2015; Kreiman & Papcun, 1991; Schmidt-Nielson & Crystal, 2000; Van Lancker & Kreiman, 1987; Wester, 2010, 2012; Winters, Levi & Pisoni, 2008). However, many previous voice discrimination studies testing non-expert listeners have been motivated by interest in, for example, the effect of samples featuring different languages (Wester, 2010, 2012; Winters et al., 2008) rather than potential forensic applications. Furthermore, these (Wester, 2010, 2012; Winters et al., 2008) and other studies (e.g., Kreiman & Papcun, 1991; Van Lancker & Kreiman, 1987) have tested participants on speech samples created under very similar conditions (e.g., same

speech type or no background noise), or have used vowel sounds (e.g., Lavan, Scott & McGettigan, 2016) rather than full sentences. However, what *is* clear from the previous literature is that stimulus variability negatively affects voice discrimination performance. Participants are less accurate when making judgments across different vocalisations (Lavan et al., 2016) different languages (Wester, 2012), and when linguistic content varies across to-be-compared samples (Naranyan, Mak & Bialystok, 2017).

Previous studies of voice-matching in forensic settings have failed to explore the role of different types of speech within a speaker (intra-speaker variations), even though speaking style may vary across recordings and such variations are likely to affect performance. The same person's speech can differ greatly across articulations and occasions (intra-speaker variability) (Holmberg, Hillman, Perkell & Gress, 1994; Hammersley & Read, 1996). These variations are heightened across different speaking styles. There are prosodic differences across spontaneous, read, and conversational speech (Baker & Hazan, 2010; Dellwo et al., 2012; Levin, Schaffer & Snow, 1982; Remez, Rubin & Nygaard, 1986). However, Leeman, Kolly and Dellwo (2014) showed that variations in suprasegmental temporal features are stable across changes in speaking style (spontaneous vs. read). Listeners may therefore be able to rely on stable, high-level features of speech, such as mannerisms, speaking rate, and pauses when making matching decisions (Alexander, Dessimoz, Botti & Drygajlo, 2005).

Voice recordings are likely to be made under a variety of conditions. For example, the environment might be noisy or quiet. Therefore, it is important to investigate the effect of this variable on lay listener performance. Background noise may have a disruptive effect on the perception of voice identity; it impairs speech perception (Mattys, Davis, Bradlow & Scott, 2012) and masks informative cues such as pitch (Qin & Oxenham, 2003). There is some evidence, provided by Bartle and Dellwo (2015), that the inclusion of background noise degrades human performance. In this study on voice discrimination, following ceiling level

performance in pilot testing, speech babble was added to all recordings. However, the difference between performance on clean and noisy samples was not tested statistically. Arguably, what matters more when investigating the effect of background noise is the mismatch between samples. In forensic casework, recordings are not likely to be recorded under identical environmental conditions (Alexander, Botti, Dessimoz & Drygajlo, 2005). Whilst it has been shown that human listeners perform well in comparison to computers when recording conditions are mismatched (Alexander et al., 2004, 2005), research specifically designed to address lay listener performance in isolation is required. This is an important legal issue in its own right; legal practitioners must be accurately informed about the strengths and limitations of using voice matching procedures with lay listeners.

Accuracy and confidence

A further important consideration is the relationship between participants' accuracy and confidence, as a witness or jury member who is confident that they are reporting the truth is likely to be extremely persuasive (Brewer & Burke, 2002; Cutler, Penrod & Stuve, 1988; Lindsay, Wells & Rumpel, 1981). Confidence in voice identification decisions has been investigated in earwitness contexts, showing that overall, high levels of confidence do not predict accuracy (Kerstholt, Jansen, Van Amelsvoort & Broeders, 2004; Olsson, Juslin & Winman, 1998; Yarmey, 1995; Yarmey & Matthys, 1992). However, far more research has focused on eyewitness confidence than earwitness confidence. Although eyewitness confidence tends not to be perfectly diagnostic of face identification accuracy (see Sauer & Brewer, 2015), the methods by which confidence and accuracy are analysed lead to different conclusions about the nature of the relationship. Whilst point biserial correlation points towards a weak to moderate relationship, examining the calibration between confidence and accuracy provides a richer perspective, and suggests that correlational approaches tend to underestimate the strength of the relationship (Brewer & Wells, 2006). As a result, the

literature has moved away from arguing that the confidence-accuracy relationship is weak, towards appreciating that the relationship can be strong under some conditions (Palmer, Brewer, Weber & Nagesh, 2013; Wixted & Wells, 2017). Therefore, although it has been suggested that the diagnosticity of confidence is even more limited for voice identification than it is for face identification (Olsson et al., 1998), a more detailed exploration of this relationship is needed, particularly in voice discrimination contexts. Findings relating to earwitnesses will not necessarily generalize to voice discrimination tasks, where the memory load is more limited. Previous forensic voice discrimination studies have tended not to directly address the question of whether accurate responders are also confident responders. One exception is Bartle and Dellwo (2015), whose results suggest that the overall relationship between confidence and accuracy is weaker for lay listeners than experts, but that high confidence is strongly related to accuracy. Further research is necessary to ensure that juries receive appropriate advice about the weight they should attach to witness confidence, or their own certainty, in cases which are reliant on voice discrimination evidence.

Aims

This research is overdue; assumptions about human voice discrimination performance require urgent testing. Basing legal decisions on incorrect assumptions is likely to negatively influence the course of justice. In order to learn more about the strength of forensic voice matching evidence, we tested the voice discrimination ability of lay listeners. As to-be-compared recordings are likely to vary in terms of recording conditions and context, we manipulated speech type and background noise. We were also interested in how self-rated confidence levels would vary according to different conditions, because witness confidence can be influential in the criminal investigation and trial process. We predicted that accuracy would be highest when the recordings were most similar (i.e. speaking styles matched) and the speech signal was clear (i.e. no background noise).

Experiment 1

In Experiment 1, we used a same/different task to test voice discrimination performance. Participants compared two voices (voice 1 and voice 2) for identity. We investigated whether a mismatch in speaking style would influence performance, and whether the inclusion of background noise was disruptive. We expected that participants would be most accurate when speaking styles were the same, and when there was no background noise.

Method

Design. The study employed a 2 x 2 x 2 within subject design. The factors were identity (same or different), the speech type of the second voice (read or spontaneous), and background (none or noise). The dependent variables were matching accuracy and self-rated confidence.

Participants. There were 34 participants (33 female, 1 male), with an age range of 18-36 years (M = 22.3, SD = 4.7). The participants were students, recruited from the Nottingham Trent University Psychology Department's Research Participation Scheme. They received research credits as compensation. Ethical approval for both experiments was granted by the University's Business, Law and Social Science College Research Ethics Committee.

Apparatus and materials. The stimuli were from the UCL Speaker Database (Markham & Hazan, 2002). The corpus features 35 British English speakers with either a neutral or mild South-Eastern accent, recorded performing a variety of spoken tasks. However, as not all of the speakers in the corpus are recorded performing all tasks, and we required recordings of each speaker performing a total of 3 tasks, only 24 speakers were suitable for use (13 females and 11 males). The speakers had an age range of 20-58 years (M = 30.6, SD = 9.9). In the recordings selected for this experiment, each speaker read two passages aloud from text: 'The story of Arthur the Rat', a children's story, and 'The Rainbow Passage', a simple scientific description of how rainbows are formed. In the third recording

(spontaneous speech), the speakers were recorded recalling a cartoon strip story from memory. For each of the recordings the audio quality was 705 kbits per second, 44,100 Hz, 16 bit. The recordings were edited in Adobe Audition so that there was no silence at the beginning of the recording. Each of the recordings were played from the start, for a total of 5 s, so that the speaker was heard uttering at least one full sentence. The loudness of the recordings was equalized through root-mean-square normalization in Matlab. Multi-speaker babble (Stacey, Kitterick, Morris & Sumner, 2016) was added to the recordings played in the background noise condition. For each resulting voice recording, the Signal to Noise Ratio (SNR) was 6dB (which means that the speech signal was 6dB higher in volume than the background noise).

The participants completed the experiment on an Acer Aspire laptop (screen size = 15.6 in., resolution = 1,366 × 768 pixels, Dolby Advanced Audio), with the experiment running on PsychoPy version 1.77.01 (Peirce, 2009). The voice recordings were presented binaurally through Sennheiser (HD205) headphones. The voice recordings volume was measured using a Svantek (977) sound level meter, with the headphones placed over a G. R. A. S. (RA0039) artificial ear simulator. The volume ranged between 65 – 75 dB. The sound intensity was constant across participants.

Four versions of the experiment blocks were constructed using an online research randomizer (Urbaniak & Plous, 2013) so that across versions, different combinations of voices were encountered in same identity and different identity trials. Each of the 24 speakers was heard twice in a block (once as voice 1, once as voice 2); each block consisted of 24 trials in total. There were 12 same identity trials, and 12 different identity trials. If an identity was heard as voice 1 in a same identity trial, it also featured as voice 2 in that trial, but was not heard again during the block. If an identity was heard as voice 1 in a different identity trial, that voice also featured as voice 2 in a different identity trial later in the block. On

different identity trials, both speakers were the same sex. In half of the trials both recordings featured background noise, and in the remaining half, both recordings featured no background noise. Although the order of trials was randomized between participants, each trial (within a block version) was the same.

Procedure. The procedure used in Experiment 1 is illustrated in Figure 1. The participants were allocated to two block versions (1-4) using an online research randomizer (Urbaniak & Plous, 2013). Participants completed two different block versions so that they did not encounter the same combination of stimuli twice in the experiment; the experimental trials therefore consisted of 48 pairs of voices (24 same identity, and 24 different identity) per participant. There were two practice trials before that start of each block. In the 'read vs. read' block, the 5 s recording of the Arthur the rat passage (voice 1) was compared for identity to the 5 s recording of the rainbow passage (voice 2). There was a 2 s gap between voices. While the voices played, the text 'Voice 1' or 'Voice 2' was visible in the centre of the screen. At test, the participants selected '0' if they thought the voices belonged to different people, or '1' if they thought the voices belonged to the same person. They could not respond until both recordings had finished, following which the participants responded at their own pace. No time limit was imposed. After they had registered their response, the participants were asked, '[o]n a scale of 1-10, how confident are you that you have made the correct response?' (1 – *not at all confident*, 10 – *extremely confident*). The procedure in the 'read vs. spontaneous' block was identical, apart from the second recording featuring spontaneous recall of the cartoon (voice 2). The order of the blocks was counterbalanced across participants.

Results

Accuracy. Voice discrimination accuracy was analysed using multilevel logistic regression (lme4 package in R: Bates et al., 2014) in order that both participants and stimuli could be treated as random effects. The advantages of multilevel modelling over traditional ANOVA are widely reported (Baguley, 2012; Clark, 1973; Judd, Westfall & Kenny, 2012; Smith, Dunn, Baguley & Stacey, 2016; Wells, Baguley, Sergeant & Dunn, 2013). We used the same method of analysis as Smith et al. (2016), comparing four nested models, fitted using restricted maximum likelihood. Accuracy (0 or 1) was the dependent variable. Model 1 included a single intercept, model 2 included the main effects, model 3 included the two-way interaction, and model 4 included the three-way interaction. We report likelihood ratio tests provided by lme4. They were obtained by dropping each effect in turn from the appropriate model. The chi-square statistic (G^2) and *p* value associated with dropping each effect are reported in Table 1, along with the coefficients and standard errors (on a log odds scale) for each effect in model 4. In model 4, the estimate of *SD* of the voice 1 random effect was 0.501, for the voice 2 random effect it was 0.369. The *SD* of the participant effect was 0.242.

The main effect of identity was significant, as was the main effect of voice 2 speech type. There was also a significant interaction between identity and background. Figure 2 aids the interpretation of these results. It shows the means for percentage accuracy in each condition of the factorial design. The accompanying 95% confidence intervals were obtained using the arm package in R (Gelman & Su, 2013), which simulates the posterior distributions of parameters (in this case the cell means).

Overall discrimination was 87.0% correct on average, 95% CI [82.7, 90.3]. As shown in Figure 2, accuracy was higher on different identity trials, *M* = 89.7%, 95% CI [85.5, 88.5], than same identity trials, *M* = 84.3%, 95% CI [78.9, 88.5]. In addition, participants were more

accurate when the speech types matched (read vs. read is shown in panel A, *M* = 89.23%, 95% CI [85.1, 92.3]; read vs. spontaneous is shown in panel B, *M* = 84.6%, 95% CI [79.4, 88.6]). The cell means for the 2-way interaction are near identical for three of the conditions (same identity no background noise, *M* = 84.0%, 95% CI [77.2, 89.4]; same identity background noise, *M* = 85.3%, 95% CI [79.3, 90.2]; different identity background noise, *M* = 86.18%, 95% CI [79.7, 91.3]), whereas the cell mean for different identity no background noise, *M* = 93.1%, 95% CI [89.2, 95.8], is substantially higher. This interaction has a single degree of freedom and therefore cannot be decomposed further, but it is possible to assess whether the interaction contrast equivalent to our proposed explanatory account – with weights [-1, -1, -1, 3] – 'mimics' the pattern observed cell means (Abelson & Prentice, 1997, p.321).[1] This correlation is .98 and indicates that the difference between the different identity no background noise and the other three conditions accounts for over 95% of the variance between these means (see Baguley, 2012). The interaction between identity and background is almost entirely accounted for by the observation that when no background noise is included in either recording, participants are more likely to respond that the voices belong to a different identity.

Multilevel signal detection analysis. Signal detection involves calculating sensitivity indices (*d'*) and response biases (C). The traditional approach to signal detection involves partitioning same-different data into hits (on a same identity trial, participants respond *same*), false alarms (on a different identity trial participants respond *same*), misses (on a same identity trials participants respond *different*) and correct rejections (on a different identity trial, participants respond *different*). For each participant, aggregate measures would be

[1] Extending the logic of Abelson and Prentice to a generalized linear model raises the question of whether contrast weights should be correlated with 'cell means' on a log odds scale, odds scale or probability scale. Mathematically it is arguably most reasonable to use the log odds scale (which is the underlying linear model). However, for interpretation it seems more natural to use the probability (i.e., percentage accuracy) scale which we adopt here. This also has the advantage of being more conservative; the correlation rises to .99 if the log odds are used.

calculated, and statistics performed on these values. This is problematic because it means ignoring important sources of variability in the underlying statistical model (Clark, 1973; Judd et al., 2012). The analyses reported above show that in our data there is variability at both the participant and stimulus level, underlining the importance of taking both sources of variability into account. Widely-known methods exist for flexible fitting of signal detection models with a single random factor (e.g., Wright, Horry & Skagerberg, 2009), but are at present limited to a single random factor. For our data equal variance Gaussian signal detection (EVSDT) models were estimated using a Bayesian multilevel probit model in the R package 'brms' (Bürkner 2017), making it possible to simultaneously fit models with a random effect for participants and two random effects for stimuli. Overall, the parameter estimate for C was 1.20, 95% CI [0.82, 1.60], showing that there was a bias to respond *different*. The *d'* value was 3.11, 95% CI [2.59, 3.78]. Additional modeling, not reported here, did not detect main effects of either condition or their interaction for C or *d'*.

Confidence. The means and 95% CIs (calculated from the *SE*) for each of the conditions, are shown in Figure 3. The factors were identity, voice 2 speech type, and background. In one trial, no confidence rating was recorded. This data point was removed.

The confidence data were analysed using multilevel ordered logistic regression in R using the ordinal package (Christensen, 2011). Individual effects were tested for using the same method as the matching accuracy analysis. Four models were compared, with self-rated confidence as the dependent variable. The first model included only intercepts, the second model included the predictors (identity, voice 2 speech type, and background), the third model added the two-way interaction, and the fourth model added the three-way interaction. In the three-way model, the estimate of the *SD* of the voice 1 random effect was 0.360, for voice 2 it was 0.495, and for participant random effect it was 1.031. Dropping each effect

from the null model showed that the participants were more confident when comparing read speech to read speech (b = 0.665, SE = 0.183, G^2 = 31.55, p < .001), and when comparing speech samples that did not include background noise (b = 0.099, SE = 0.210, G^2 = 4.39, p = .036). No interaction effects were detected (p > .404).

The relationship between confidence and accuracy. The likelihood of underestimating the confidence-accuracy relationship when using point biserial correlation is well-documented (Juslin, Olsson & Winman, 1996; Lindsay, Nilsen & Read, 2000). Calibration is more informative, providing information about accuracy at each level of confidence, and an indication of over/under-confidence (Juslin et al., 1996). The first step in the calibration analysis is to plot calibration curves (Brewer & Wells, 2006). Statistics are also informative (calibration (C), over/underconfidence (O/U), and the normalized resolution index (NRI)). However, both calibration curves and accompanying statistics are calculated based on aggregated data, which is problematic for this dataset (as highlighted above). Therefore, whilst we refrain from attempting to draw conclusions based on inferential statistics associated with calibration analysis, we present the calibration curves as a useful illustration of the relationship between confidence and accuracy.

Self-rated confidence was measured on a scale of 1-10. However, the majority of confidence ratings were made at the higher range. For the purposes of providing more stable estimates (following Brewer & Wells, 2006), confidence was collapsed into 3 categories (low: 1-4, medium: 5-7, high: 8-10). In Figure 4a and 4b, the overall accuracy (%) in each of the categories is plotted against the weighted mean confidence for that particular category. We collapsed across identity, and therefore deal with overall accuracy levels. The diagonal line shows where data points would fall if confidence and accuracy were perfectly calibrated. Points that fall above this line reflect underconfidence, while points that fall below the line reflect overconfidence.

Based on visual inspection, the participants' overall self-rated confidence seems to be well calibrated to their overall accuracy, especially at the higher levels of confidence. Most of the points fall below the line of perfect calibration, demonstrating that if anything, participants display a tendency towards overconfidence.

Next the relationship between confidence and accuracy in each condition was examined statistically. Separate analyses were run for each of the four conditions illustrated in Figures 4a and 4b. This was done using the ordinal package in R (Christensen, 2011) so that both participants and stimuli could be treated as random effects. Self-rated confidence was the dependent variable, and accuracy was the predictor. Two models were compared, the first including only intercepts and the second adding accuracy as a predictor. Accuracy predicted confidence in all four conditions: read vs. read, no background noise (b = 2.171, SE = 0.325, G^2 = 43.47, $p < .001$), read vs. read, background noise (b = 2.074, SE = 0.287, G^2 = 52.49, $p < .001$), read vs. spontaneous no background noise (b = 1.509, SE = 0.288, G^2 = 26.96, $p < .001$), and read vs. spontaneous, background noise (b = 1.638, SE = 0.252, G^2 = 42.91, $p < .001$).

In Figures 5a, 5b, 5c and 5d, the probability (on a log odds scale) of an incorrect match or a correct match is plotted for each level of self-rated confidence (1-10). Data from the four conditions are presented in separate figures. The plots were generated using the ordinal package in R (Christensen, 2011). A strong relationship between confidence and accuracy would be depicted by higher probability of an incorrect match at lower levels of confidence (left plots), and higher probability of a correct match at higher levels of confidence (right plots).

Figures 5a – 5d illustrate that the probability of an incorrect match tends to be high at mid-high levels of confidence (left plots), and the probability of a correct match tends to be higher at the higher levels of confidence (right plots).

Discussion

Overall mean error rates were 13%, varying between 7% and 20% across the conditions. This is on the lower side of the error rates observed in many previous studies (Bartle & Dellwo, 2015; Kreiman & Papcun, 1991; Schmidt-Nielson & Crystal, 2000; Van Lancker & Kreiman, 1987; Wester, 2010, 2012; Winters et al., 2008). Higher levels of accuracy in the different identity condition supports the conclusion that participants found it more difficult to correctly assign intra-speaker variability to a single identity than they did to correctly assign inter-speaker variability to separate identities. The degree of match between recordings appears to play a role in driving accurate performance. Participants were more accurate when the speech types were the same (read vs. read) compared to when they were different (read vs. spontaneous). There was no main effect of background noise, perhaps because when noise was included, it featured in both recordings. Alternatively, the level of noise may not have been sufficient to disrupt performance. However, there is some evidence that the presence of background noise undermines performance; accuracy was lower on different identity trials when the voices were heard with background noise. This extends the findings of previous studies which hint that background noise is associated with lower levels of accuracy (Bartle & Dellwo, 2015).

There is a relationship between confidence and accuracy in all conditions. Performance was particularly well calibrated when participants were very confident in the accuracy of their response, which is consistent with the findings of previous calibration studies (Palmer et al., 2013; Weber & Brewer, 2003). Observations based on the calibration curves are supported by the multilevel analysis, which showed that accuracy predicts

confidence across all conditions. However, it must be acknowledged that this relationship is far from perfect, and appears to be primarily driven by correct responses. The data presented in Figures 5a – 5d indicate that on balance, incorrect matches are associated with ratings in the upper half (6-10) of the confidence scale.

Experiment 2

In order to further investigate the importance of mismatched recording conditions on voice discrimination accuracy, only one of the recordings in each trial featured background noise in Experiment 2. As voices in a discrimination task are presented sequentially, we also varied whether the recording featuring background noise was presented first or second. Based on the results of Experiment 1, it seemed likely that the mismatch between background would increase task difficulty by making intra-speaker variability more salient. In the absence of previous literature, we were unsure whether an order effect would be observed.

Method

Apart from the following exceptions, the methods were identical to Experiment 1.

Design. The study employed a 2 x 2 x 2 within subject design. The factors were identity (same or different), the speech type of the second voice (reading or spontaneous), and background noise order (first or second). The dependent variables were matching accuracy and self-rated confidence.

Participants. There were 34 participants (26 female, 8 male), with an age range of 18-42 years (M = 23.3, SD = 6.7).

Procedure. In half of the trials, background noise was added to the recording of voice 1, but the voice 2 recording featured no background noise. In the other half of the trials, background noise was only added to the recording of voice 2.

Results

Accuracy. Voice discrimination accuracy was analysed in exactly the same way as in Experiment 1. The likelihood chi-square statistic (G^2) and *p* value associated with dropping each effect in turn from the appropriate model is shown in Table 2. In model 4, the estimate of the *SD* of the voice 1 random effect was 0.386, for the voice 2 random effect it was 0.473, and for the participant main effect it was 0.561.

There was a main effect of voice 2 speech type and a main effect of background noise order. Figure 6 shows that participants were more accurate overall when speech types were matched, *M* = 83.9%, 95% CI [78.5, 88.2] compared to when they were not matched, *M* = 78.9%, 95% CI [72.6, 84.1]. Overall accuracy levels were higher when the recording featuring background noise was heard second, *M* = 83.9%, 95% CI [78.2, 88.2] compared to when it was heard first, *M* = 78.7%, 95% CI [72.2, 84.2]. No other main effects or interactions approached significance. Overall accuracy was 81.3% (95% CI [75.8, 85.8]).

Multi-level signal detection analysis. The multilevel signal detection analysis revealed that overall, there was a bias to respond *different*: the parameter estimate for C was 0.96, 95%CI [0.67, 1.28]. The *d'* value was 2.14, 95% CI [1.51, 2.84]. Additional modeling detected no main effects, and no interactions for C or *d'*.

Confidence. The means and 95% CIs (calculated from the *SE*) for each of the conditions are shown in Figure 7.

As in Experiment 1, the confidence data were analysed using multilevel ordered logistic regression. The predictors were identity, voice 2 speech type and background noise order. In total 5 data points were removed owing to no confidence rating being recorded. In the three-way model, the estimate of the *SD* of the voice 1 random effect was 0.350, for voice

2 it was 0.550, and for participant random effect it was 0.938. Dropping each effect from the full model showed that the participants were more confident comparing speech samples when the background noise featured in the second voice recording (b = 0.611, SE = 0.209, G^2 = 7.87, p = .005). There was a significant two-way interaction between identity and voice 2 speech type (b =0.858, SE = 0.252, G^2 = 5.00, p = .025), and a significant three-way interaction between identity, voice 2 speech type and background noise order (b = 0.921, SE = 0.357, G^2 = 6.677, p = .010). No other main effects or interactions approached significance (p > .219). As these interactions were unpredicted, we refrain from over-interpreting them. At most, we can conclude that some conditions may promote higher confidence.

The relationship between confidence and accuracy. The calibration curves for plotting the overall accuracy (%) in each of the confidence categories (low, medium, high) plotted against the weighted mean confidence for that particular category are shown in Figures 8a and 8b.

The relationship between confidence and accuracy was analyzed using the same method as Experiment 1. Accuracy predicted confidence in all four conditions: read vs. read, background noise first (b = 1.398, SE = 0.255, G^2 = 30.21, p < .001), read vs. read, background noise second (b = 1.573, SE = 0.269, G^2 = 34.36, p < .001), read vs. spontaneous, background noise first (b = 0.815, SE = 0.223, G^2 = 13.94, p < .001), and read vs. spontaneous, background noise second (b = 0.899, SE = 0.242, G^2 = 13.10, p < .001). In Figures 9a, 9b, 9c and 9d, the probability (on a log odds scale) of an incorrect match or a correct match is plotted for each level of self-rated confidence (1-10). Each condition is presented in a separate figure.

Figures 9a – 9d illustrate a similar pattern to that observed in Experiment 1. The probability of an incorrect match tends to be high at mid-high levels of confidence (left plots), and the probability of a correct match tends to be higher at the higher levels of confidence (right plots).

Discussion

In Experiment 1, both recordings in each trial were matched for background noise: either both or neither recording featured noise. In Experiment 2 they were always mismatched: either the first or second recording featured noise. Across conditions, mean error rates in Experiment 2 varied between 14% and 25%. Although this is broadly similar to the range observed in Experiment 1, where the overall error rate was 13%, in Experiment 2 the overall error rate approached 20%. Descriptively speaking, this suggests a possible trend towards reduced accuracy, corresponding with previous findings that accuracy is lower when recording conditions differ across to-be-compared voices (Alexander et al., 2004). The speech type results provide corroborating evidence that performance is sensitive to a mismatch between recordings. Discrimination performance was more accurate when both samples featured read speech, thus replicating the results of Experiment 1. However, the overall pattern of accuracy results was different from those observed in Experiment 1. There was no main effect of identity, which suggests that the salience of inter-speaker information was reduced by the inclusion of background noise. This explanation is consistent with explanations for the identity by background interaction in Experiment 1. The main effect of background noise order revealed that the voice discrimination task is easier when the first recording does not include background noise. This could reflect the differential role of the first and second stimulus in such tasks, and the importance of the first as a model against which comparisons can be made. The results are consistent with the conclusion that when

there is a mismatch in terms of background, discrimination decisions are easier if the template is not degraded by background noise.

Overall there was a relatively clear relationship between confidence and accuracy, and further evidence that higher levels of calibration are likely to be observed when confidence is high. However, there is some indication that the relationship degrades as task difficulty increases. The lowest effect sizes were observed in the read vs. spontaneous condition, which was also the condition in which error rates were highest. These results correspond with visual inspection of Figure 8a and 8b. In keeping with the results of Experiment 1, the relationship appears to be mostly driven by high confidence in correct responses (Figures 9a – 9d).

General Discussion

In order to address gaps in the psychological literature and inform legal professionals about the evidential strength of voice discrimination, the research reported here investigated lay listener performance on voice discrimination tasks. In two experiments, we investigated whether accurate performance was influenced by a mismatch in speaking styles, and the inclusion of background noise. Overall the results show that performance is sensitive to the degree of match between recordings. Not only is accuracy reduced when speaking styles do not match, but the inclusion of background noise is disruptive. A relationship between confidence and accuracy was observed across all conditions.

Accuracy

Consistent with previous results, voice discrimination performance was not perfect, even in optimal conditions when speaking styles were matched, and neither recording featured background noise (Experiment 1) (e.g. Kreiman & Papcun, 1991; Van Lancker & Kreiman, 1987; Wester, 2010). The overall results presented here highlight the problems associated with admitting voice identification evidence based on decisions made by lay listeners. These decisions are error-prone, and subject to disruption.

Overall performance is comparable with that observed on face matching tasks (e.g. Burton et al., 2010). In light of previous research showing that voices provide weak cues to identity in comparison to faces (see Stevenage & Neil, 2014), we might have expected voice discrimination performance to be on the low side. To fully test this hypothesis, a direct comparison between face and voice matching would need to be undertaken, in which both sets of stimuli capture similar levels of inter- and intra-speaker variability. It should be noted that here, in both Experiments 1 and 2, the voice discrimination task was made relatively easy by the fact that stimuli were randomly allocated to trials rather than being matched for similarity.

In Experiment 1, there was evidence that lay listeners exhibited a bias to respond *different*, and accuracy was higher in the different identity condition. In a voice discrimination task, the listener must decide whether the voices in each trial differ because of inter-speaker variability or intra-speaker variability. The bias to respond *different* suggests that participants were more likely to incorrectly assign intra-speaker variability to different individuals than they were to incorrectly assign inter-speaker variability to the same individual. This result should be considered alongside previous results suggesting that lay listeners are more likely to respond *same* in voice discrimination tasks when they are unsure or when acoustic conditions are sub-optimal (Bartle & Dellwo, 2015). This apparent inconsistency between the Experiment 1 results and the previous literature may be attributable to the relative ease of the task in Experiment 1 when there was no background noise. Task difficulty is likely to be dictated in part by the levels of intra-speaker variability between voices, and in part by the quality of the encoded voice. Task difficulty may increase the likelihood that lay listeners will attribute the variability to the same identity because they lack expert knowledge about how individual voices can vary across instances, and are unable to isolate high level features of speech that are stable across utterances (Alexander et al.,

2005; Leeman et al., 2014). From a perceptual point of view, the presence of intra-speaker variability is problematic for lay listeners trying to discriminate between identities (Lavan et al., 2016; Naranyan et al., 2017; Wester et al., 2012). This is supported by the pattern of results observed in these experiments. The mismatch between speech types increases intra-speaker variability because of prosodic differences (Baker & Hazan, 2010; Dellwo et al., 2012; Levin et al.,1982; Remez et al., 1986). Accordingly, accuracy was lower in the read vs. spontaneous condition in both Experiments 1 and 2.

Consistent with the above discussion of bias and task difficulty, higher accuracy on different identity trials did not occur when background noise featured. The masking effects of background noise (Brungart, Simpson, Ericson & Scott 2001) are likely to compromise the quality of representations for the encoded voices. In addition, it is possible that the voice and the background noise are encoded holistically, making it difficult for the listener to isolate and attend to only the sound of the voice. Based on the results of Experiment 1, it would seem that background noise particularly masks inter-speaker variability, making different identity trials more difficult.

In Experiment 2, background noise either featured in the first or second recording. Accuracy was higher when background noise only featured in the second recording. This might indicate that the role of background noise in influencing performance is related to cognitive capacity. In a voice discrimination task, voices are presented sequentially. This means that echoic memory must be relied upon when comparing the first voice to the second. It is possible that the inclusion of background noise imposes a higher load on echoic memory, making it more difficult to make an accurate comparison. These preliminary results may suggest that when faced with a voice discrimination task in the real world, lay listeners should be encouraged to listen to the least degraded stimulus first. Further research should be undertaken with different types of degraded speech, not just background noise. In addition,

this is an important question to address in the context of expert acoustic analysis in order to explore potential ways of improving accuracy.

We cannot rule out the possibility that the design used here underestimates the performance impairments that might be observed in a forensic setting. Firstly, it should be noted that in each experiment, participants completed 48 trials (24 in each block), providing them with a significant amount of practice on this task. Witnesses or jury members would likely perform only a single comparison, so would not have the advantage of practice. Unfortunately, there is insufficient data to test this hypothesis by analysing only the first trial for each participant. Furthermore, the participants heard each identity more than once in each block: either twice in a same identity trial, or twice in two separate different identity trials. Whilst it is possible that to some extent participants may have become familiarized with the voices over the course of the experiment, we do not anticipate that this would have affected overall levels of accuracy. Each time the identity was encountered the voice said different things, and as there were 24 identities, interference would very likely undermine the participants' ability to accurately remember the voices (Stevenage et al., 2011). Secondly, the sample in both experiments was made up of students, whose mean age was around 23. Police, lawyers, and jury members would be drawn from a sample with a higher mean age. Age is likely to affect voice discrimination accuracy, as auditory acuity starts to degrade from the age of approximately 40 years (Hoffman, Dobie, Losonczy, Themann, & Flamme, 2017). In particular, older adults are more likely to struggle when attempting to extract a speech signal from background noise (Vermeire et al., 2016). On the other hand, factors may come into play that mitigate such effects and make the task easier. In court, lay listeners might have access to relatively long speech samples when making a decision (although the length of samples will probably vary widely from case to case). Here we chose short 5 second excerpts. These spanned at least one full sentence so that unlike in some previous psychological voice

discrimination studies, prosodic information could be used to make a matching decision. As has already been noted, previous studies have tended to use much shorter samples of speech (e.g., Lavan, Scott & McGettigan, 2016).

Confidence

Overall, participants exhibited high levels of confidence. Indeed, visual inspection of the calibration graphs imply a tendency towards over-confidence. Furthermore, confidence and accuracy were particularly well calibrated when participants were highly confident. We detected a relationship between confidence and accuracy across all conditions; a result that sits in stark contrast to the lack of a relationship when speaker verification is reliant on memory (Kerstholt et al., 2004; Olsson et al., 1998; Yarmey, 1995; Yarmey & Matthys, 1992). However, the results should not be taken to suggest that juries should unquestioningly rely on evidence provided by confident lay listeners, or even that they should automatically trust their own feelings of high confidence. There is at least some indication that the relationship between confidence and accuracy may degrade in line with increasing task difficulty (Experiment 2). This is in keeping with the finding that the probability of making an incorrect response remains high when participants are relatively confident that they have responded correctly (Experiments 1 and 2). It is very important to avoid assuming that a confident witness will be an accurate witness; while witnesses are generally confident when they are correct, they are not reliably underconfident when they are incorrect. It may be appropriate for judges to issue warnings to jury members about the weight they should (or should not) attach to lay listener confidence following voice discrimination decisions.

Further research

Rather than comparing lay listener performance to computers or phonetic experts as many previous studies have done (e.g. Alexander et al., 2004, 2005; Schmidt-Nielson et al., 2000; Shen, Campbell, Straub & Schwartz, 2011; Lindh & Morrison, 2011), the experiments

presented here investigated lay listener performance in isolation in order that the procedural design would not be constrained by the need for human/computer or lay listener/expert performance to be analogous. It was important that the tasks simulated as far as possible processes that are likely to underpin decision-making by acoustically untrained listeners such as the police, lawyers and jury members; a simple comparison and a yes-no response. In contrast, during casework, forensic experts would usually rate similarity on a Likert-style rating scale (Alexander et al., 2004). We believe it is important that future work is undertaken to extend our findings, using similar methods to those outlined above so that gaps in the literature can be filled.

The preliminary results presented here suggest a number of avenues for future research. Little is currently known about the information on which people base voice discrimination decisions (Pradham & Prasanna, 2011; for an exception see Alexander et al., 2004). Further work should be undertaken to identify not only what kind of perceptual cues are relied on, but, more importantly from a forensic point of view, which perceptual cues support accurate performance. Voice similarity, for example in terms of fundamental frequency, is likely to play an important role in the likelihood of a discrimination decision being accurate (Cleary, Pisoni & Kirk, 2005; McClelland, 2008, cited in Bartle & Dellwo, 2015). The trials presented in Experiments 1 and 2 featured samples that were not systematically matched according to voice features, they were only matched according to sex. This may have made different identity trials particularly easy, so to fully investigate the limits of lay listener accuracy/performance, future work could investigate the relationship between different levels of voice similarity and accuracy across matched and mismatched recordings. The stimuli used in these experiments featured read and spontaneous speech. In terms of ecological validity, the inclusion of read speech in every trial could be considered a limitation. However, although the vast majority of recorded material for comparison is

spontaneous, recorded read speech does play a role in some cases (Leeman et al., 2014), and besides, the inclusion of read speech was important in demonstrating the difference in performance when speech types were mismatched. The speaker database (Markham & Hazan, 2002) used in these experiments featured only one sample of spontaneous speech for each speaker, but more than one sample of read speech. As such, the first voice in each experiment was always an extract of read speech, and the second voice was either an extract of read speech or spontaneous speech. Our results clearly show that speaking style has the potential to affect voice discrimination accuracy. However, future research should fully explore the effects of speech type using a fully crossed design. It may be the case that samples of spontaneous speech are more difficult to match than samples of read speech. Finally, our results highlight the potential value in exploring the relationship between confidence and accuracy in voice discrimination tasks in far more detail. It is important to develop a better understanding of the conditions supporting a strong link between confidence and accuracy.

Conclusion

Despite expert forensic phoneticians commonly being called to verify that two recordings feature the same person, it is important to note the role that lay listener misidentifications can play in undermining the course of justice. At various stages of the legal process it may be necessary for police, lawyers or the jury to make decisions about whether recordings feature the same person. However, as the results show, the voice discrimination ability of lay witnesses is unlikely to be perfect. Although performance was reasonably accurate, error rates varied between 7% and 25%. We have shown that accuracy is subject to disruption by background noise and differences in speaking style, both of which may play a role in cases involving voice discrimination. Our results suggest that the way in which voice discrimination exercises are presented (i.e. the order in which recordings are heard) may

impact upon the decision made. Furthermore, confidence does not necessarily indicate accuracy. Although there is a relationship between confidence and accuracy, we have presented evidence that it is likely to be variable. Further work will help determine how and when voice comparisons should be presented to jurors, how judges should tailor directions to the jury on how the task should be approached, and under what conditions accuracy best predicts confidence. This will require collaboration between the disciplines of psychology and law to ensure that a best practice is developed to facilitate accurate decision-making within the criminal justice system.

References

Abelson, R. P., & Prentice, D. A. (1997). Contrast tests of interaction hypothesis. *Psychological Methods*, 2(4), 315-328. doi: 10.1037/1082-989X.2.4.315

Alexander, A., Botti, F., Dessimoz, D., & Drygajlo, A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. *Forensic Science International*, *146*, S95-S99. doi: 10.1016/j.forsciint.2004.09.078

Alexander, A., Dessimoz, D., Botti, F., & Drygajlo, A. (2005). Aural and automatic forensic speaker recognition in mismatched conditions. *International Journal of Speech Language and the Law*, *12*(2), 214-234. doi: 10.1558/sll.2005.12.2.214

Baguley, T. (2012). *Serious stats: A guide to advanced statistics for the behavioral sciences*. Basingstoke: Palgrave.

Baker, R., & Hazan, V. (2010, September). LUCID: a corpus of spontaneous and read clear speech in British English. In *DiSS-LPSS* (pp. 3-6).

Bartle, A., & Dellwo, V. (2015). Auditory speaker discrimination by forensic phoneticians and naive listeners in voiced and whispered speech. *International Journal of Speech, Language & the Law*, *22*(2), 229-248. doi: 10.1558/ijsll.v22i2.23101

Bates, D, Maechler, M., Bolker, B., & Walker, S. (2014) lme4: Linear mixed-effects models using Eigen and S4. R package version 1.0-6. Available at http://CRAN.R-project.org/package=lme4

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711-725. doi: 10.1111/j.2044-8295.2011.02041.x

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences, 8*(3), 129-135. doi: 10.1016/j.tics.2004.01.008

Brewer, N., & Burke, A. (2002). Effects of testimonial inconsistencies and eyewitness confidence on mock-juror judgments. *Law and Human Behavior, 26*(3), 353-364. doi:10.1023%2FA%3A1015380522722

Brewer, N., & Wells, G. L. (2006). The confidence-accuracy relationship in eyewitness identification: effects of lineup instructions, foil similarity, and target-absent base rates. *Journal of Experimental Psychology: Applied, 12*(1), 11-30. doi: 10.1037/1076-898X.12.1.11

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America, 110*(5), 2527-2538. doi: 10.1121/1.1408946

Bulecjik v R. (1995). HCA 54.

Bürkner, P. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80* (1), 1–28, doi: 10.18637/jss.v080.i01

Christensen, R. H. B. (2011). Analysis of ordinal data with cumulative link models—estimation with the R-package 'ordinal'. Available at http://cran.r-project.org/web/packages/ordinal/vignettes/clm_intro.pdf.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*(4), 335-359. doi: 10.1016/S0022-5371(73)80014-3

Cleary, M., Pisoni, D. B., & Kirk, K. I. (2005). Influence of voice similarity on talker discrimination in children with normal hearing and children with cochlear

implants. *Journal of Speech, Language, and Hearing Research*, *48*(1), 204-223. doi: 10.1044/1092-4388(2005/015)

Cutler, B. L., Penrod, S. D., & Stuve, T. E. (1988). Juror decision making in eyewitness identification cases. *Law and Human Behavior*, *12*(1), 41-55. doi: 10.1007/BF01064273

Damjanovic, L., & Hanley, J. R. (2007). Recalling episodic and semantic information about famous faces and voices. *Memory & Cognition*, *35*(6), 1205-1210. doi: 10.3758/BF03193594

Dellwo, V., Leemann, A., & Kolly, M. J. (2015). The recognition of read and spontaneous speech in local vernacular: The case of Zurich German. *Journal of Phonetics*, *48*, 13-28. doi: 10.1016/j.wocn.2014.10.011

Edmond, G., Martire, K., & Roque, M. S. (2011). Mere Guesswork: Cross-Lingual Voice Comparisons and the Jury. *Sydney Law Review*, *33*, 395-425.

Fenn, K. M., Shintel, H., Atkins, A. S., Skipper, J. I., Bond, V. C., & Nusbaum, H. C. (2011). When less is heard than meets the ear: Change deafness in a telephone conversation. *The Quarterly Journal of Experimental Psychology*, *64*(7), 1442-1456. doi: 10.1080/17470218.2011.570353

Gelman, A. E., & Su, Y. S. (2013). arm: Data analysis using regression and multilevel/hierarchical models. R package version 1.6-05. Available at http://CRAN.R-project.org/package=arm

Hammersley, R. & Read, J. D. (1996). Voice identification by humans and computers. In S. L. Sporer, R. S. Malpass & G. Koehnken (Eds.), *Psychological issues in eyewitness identification* (pp. 117-152). Hillsdale, NJ: Lawrence Erlbaum.

Hanley, J. R., & Turner, J. M. (2000). Why are familiar-only experiences more frequent for voices than for faces? *The Quarterly Journal of Experimental Psychology: Section A, 53*(4), 1105-1116. doi: 10.1080/713755942

Hoffman, H.J., Dobie, R.A., Losonczy, K.G., Themann, C.L., Flamme, G.A.. (2017). Declining prevalence of hearing loss in us adults aged 20 to 69 years. *JAMA Otolaryngology Head and Neck Surgery*,*143*(3), 274–285. doi:10.1001/jamaoto.2016.3527

Holmberg, E. B., Hillman, R. E., Perkell, J. S., & Gress, C. (1994). Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in SPL across repeated recordings. *Journal of Speech, Language, and Hearing Research, 37*(3), 484-495. doi: 10.1044/jshr.3703.484

Judicial College (2017). *The Crown Court Compendium Part I: Jury and Trial Management and Summing Up, February 2017.*

Judd, C. M., Westfall, J., & Kenny, D. A. (2012). Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*, *103,* 54–69. doi:10. 1037/ a0028347

Juslin, P., Olsson, N., & Winman, A. (1996). Calibration and diagnosticity of confidence in eyewitness identification: Comments on what can be inferred from the low confidence–accuracy correlation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1304-1316. doi: 10.1037/0278-7393.22.5.1304

Kerstholt, J. H., Jansen, N. J., Van Amelsvoort, A. G., & Broeders, A. P. A. (2004). Earwitnesses: Effects of speech duration, retention interval and acoustic environment. *Applied Cognitive Psychology*, *18*(3), 327-336. doi: 10.1002/acp.974

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar

voices. *Speech Communication*, *10*(3), 265-275. doi: 10.1016/0167-6393(91)90016-M

Lavan, N., Scott, S. K., & McGettigan, C. (2016). Impaired generalization of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General*, *145*(2), 1604-1614. doi: 10.1037/xge0000223

Leemann, A., Kolly, M. J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59-67. doi: 10.1016/j.forsciint.2014.02.019

Legge, G. E., Grossmann, C., & Pieper, C. M. (1984). Learning unfamiliar voices. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 298-303. doi: 10.1037/0278-7393.10.2.298

Levin, H., Schaffer, C. A., & Snow, C. (1982). The prosodic and paralinguistic features of reading and telling stories. *Language and Speech*, *25*(1), 43-54. doi: 10.1016/j.forsciint.2014.02.019

Lindh, J., & Morrison, G. S. (2011). Humans versus machine: forensic voice comparison on a small database of Swedish voice recordings. In *Proceedings of ICPhS* (Vol. 17, p. 4).

Lindsay, R. C., Wells, G. L., & Rumpel, C. M. (1981). Can people detect eyewitness-identification accuracy within and across situations? *Journal of Applied Psychology*, *66*(1), 79-89. doi: 10.1037/0021-9010.66.1.79

Lindsay, D. S., Nilsen, E., & Read, J. D. (2000). Witnessing-condition heterogeneity and witnesses' versus investigators' confidence in the accuracy of witnesses' identification decisions. *Law and Human Behavior, 24*, 685–697. doi: 10.1023/A:1005504320565

Markham, D., & Hazan, V. (2002). The UCL Speaker Database. *Speech, Hearing and Language: Work in progress*, *14*, 1-17.

Mathias, S. R., & von Kriegstein, K. (2014). How do we recognise who is speaking? *Frontiers in Bioscience, 6,* 92-109. doi: 10.2741/S417

Mattys, S.L., Davis, M.H., Bradlow, A.R., & Scott, S.K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*, 953-978. doi: 10.1080/01690965.2012.705006

McAllister, H. A., Bregman, N. J., & Lipscomb, T. J. (1988). Speed estimates by eyewitnesses and earwitnesses: How vulnerable to postevent information? *The Journal of General Psychology*, *115*(1), 25-35. doi: 10.1080/00221309.1988.9711085

McGorrery, P. G., & McMahon, M. (2017). A fair 'hearing' Earwitness identifications and voice identification parades. *The International Journal of Evidence & Proof*, 1365712717690753.

Morrison, G. S., Ochoa, F., & Thiruvaran, T. (2012). Database selection for forensic voice comparison. In *Odyssey* (pp. 62-77).Nolan, F. (2009). *The phonetic bases of speaker recognition.* Cambridge: Cambridge University Press

Narayan, C. R., Mak, L., & Bialystok, E. (2017). Words get in the way: Linguistic effects on talker discrimination. *Cognitive Science*, *41*(5), 1361-1376. doi: 10.1111/cogs.12396

Olsson, N., Juslin, P., & Winman, A. (1998). Realism of confidence in earwitness versus eyewitness identification. *Journal of Experimental Psychology: Applied, 4*(2), 101-118. doi: 10.1037/1076-898X.4.2.101

Palmer, M. A., Brewer, N., Weber, N., & Nagesh, A. (2013). The confidence-accuracy relationship for eyewitness identification decisions: Effects of exposure duration, retention interval, and divided attention. *Journal of Experimental Psychology: Applied*, *19*(1), 55-71. doi: 10.1037/a0031602

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics, 2*(10), 1-8. doi: 10.3389/neuro.11.010.2008

Pradhan, G., & Prasanna, S. R. M. (2011). Speaker verification under degraded condition: a perceptual study. *International Journal of Speech Technology, 14*(4), 405-417. doi: 10.1007/s10772-011-9120-6

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America, 114*(1), 446-454. doi: 10.1121/1.1579009

R. v Flynn & St John. (2008). EWCA Crim 970.

R v Kapikanya. (2015). EWCA Crim 1507.

R v O'Doherty. (2003). NICA 1.

R. v Shannon Tamiz and others. (2010). EWCA Crim 2638.

R. v Suleman. (2012). EWCA Crim 1569.

Remez, R. E., Rubin, P. E., & Nygaard, L. C. (1986). On spontaneous speech and fluently spoken text: Production differences and perceptual distinctions. *The Journal of the Acoustical Society of America, 79,* S26. doi: 10.1121/1.2023137

Robson, J. (2017). A fair hearing? The use of voice identification parades in criminal investigations in England and Wales. *Criminal Law Review*, (1), 36-50.

Sauer, J. D., & Brewer, N. (2015). Confidence and accuracy of eyewitness identification. In T. Valentine & J. P. Davis (Eds.), *Forensic Facial Identification: Theory and Practice of Identification from Eyewitnesses, Composites and CCTV* (pp. 185-208). Wiley-Blackwell: London.

Schmidt-Nielsen, A., & Crystal, T. H. (2000). Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. *Digital Signal Processing*, *10*(1-3), 249-266.

Shen, W., Campbell, J., Straub, D., & Schwartz, R. (2011, May). Assessing the speaker recognition performance of naive listeners using Mechanical Turk. In *Acoustics,*

Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on (pp. 5916-5919). IEEE.

Smith, H. M. J., Dunn, A. K., Baguley, T., & Stacey, P.C. (2016). Matching novel face and voice identity using static and dynamic facial images. *Attention, Perception, & Psychophysics*, 1-12. doi: 10.3758/s13414-015-1045-8

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing research*, *336*, 17-28. doi: 10.1016/j.heares.2016.04.002

Stevenage, S. V., Howland, A., & Tippelt, A. (2011). Interference in eyewitness and earwitness recognition. *Applied Cognitive Psychology*, *25*(1), 112-118. doi: 10.1002/acp.1649

Stevenage, S. V., Hugill, A. R., & Lewis, H. G. (2012). Integrating voice recognition into models of person perception. *Journal of Cognitive Psychology*, *24*(4), 409-419. doi: 10.1080/20445911.2011.642859

Stevenage, S. V., Neil, G. J., Barlow, J., Dyson, A., Eaton-Brown, C., & Parsons, B. (2013). The effect of distraction on face and voice recognition. *Psychological Research*, *77*(2), 167-175. doi: 10.1007/s00426-012-0450-z

Stevenage, S., & Neil, G. (2014). Hearing faces and seeing voices: The integration and interaction of face and voice processing. *Psychologica Belgica*, *54*(3), 266-281. doi: 10.5334/pb.ar

Urbaniak, G. C., & Plous, S. (2013). Research Randomizer (Version 4.0) [Computer software]

Van Lancker, D., & Kreiman, J. (1987). Voice discrimination and recognition are separate abilities. *Neuropsychologia*, *25*(5), 829-834. doi: 10.1016/0028-3932(87)90120-5

Vermeire, K., Knoop, A., Boel, C., Auwers, S., Schenus, L., Talaveron-Rodriguez, M., De Boom, C., De Sloovere, M. (2016). Speech recognition in noise by younger and older adults: effects of age, hearing loss, and temporal resolution. *Annals of Otology, Rhinology & Laryngology, 125 (4)*, 297-302. doi: 10.1177/0003489415611424

Vitevitch, M. S. (2003). Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology-Human Perception and Performance, 29*(2), 333-342. doi: 10.1037/0096-1523.29.2.333

Weber, N., & Brewer, N. (2003). The effect of judgment type and confidence scale on confidence-accuracy calibration in face recognition. *Journal of Applied Psychology, 88*(3), 490-499. doi: 10.1037/0021-9010.88.3.490

Wells, T., Baguley, T., Sergeant, M., & Dunn, A. (2013). Perceptions of human attractiveness comprising face and voice cues. *Archives of Sexual Behavior, 42*, 805–811. doi:10. 1007/ s10508-012-0054-0

Wester, M. (2010). Cross-lingual talker discrimination. In *Proc. Interspeech*, pp. 1253–1256, Makuhari, Japan

Wester, M. (2012). Talker discrimination across languages. *Speech Communication, 54*(6), 781-790. doi: 10.1016/j.specom.2012.01.006

Winters, S. J., Levi, S. V., & Pisoni, D. B. (2008). Identification and discrimination of bilingual talkers across languagesa). *The Journal of the Acoustical Society of America, 123*(6), 4524-4538. doi: 10.1121/1.2913046

Wixted, J. T., & Wells, G. L. (2017). The relationship between eyewitness confidence and identification accuracy: A new synthesis. *Psychological Science in the Public Interest, 18*(1), 10-65. doi: 10.1177/1529100616686966

Vuorre, M. (2017, October 12). What is human systems integration? [Blog post]. Retrieved from https://vuorre.netlify.com/post/2017/bayesian-estimation-of-signal-detection-theory-models-part-2/

Wright, D. B., Horry, R., & Skagerberg, E. M. (2009). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods, 41*(2), 257-267. doi: 10.3758/BRM.41.2.257

Yarmey, A. D. (1995). Earwitness speaker identification. *Psychology, Public Policy, and Law, 1*, 792-816. doi: 10.1037/1076-8971.1.4.792

Yarmey, A. D., & Matthys, E. (1992). Voice identification of an abductor. *Applied Cognitive Psychology, 6*(5), 367-377. doi: 10.1002/acp.2350060502

Table 1

Parameter estimates (b) and likelihood tests for the 2x2x2 factorial analysis, Experiment 1

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	2.486	0.297	.	.
Identity	1	1.004	0.335	7.58	.006
Voice 2 speech type	1	0.148	0.367	8.30	.004
Background	1	0.902	0.343	2.67	.102
Identity x Voice 2 speech type	1	0.147	0.450	0.53	.465
Identity x Background	1	0.813	0.448	5.57	.018
Voice 2 speech type x Background	1	0.260	0.456	1.27	.260
Identity x Voice 2 speech type x Background	1	0.117	0.590	0.04	.845

Table 2

Parameter estimates (b) and likelihood tests for the 2 x 2 x 2 factorial analysis, Experiment 2

Source	*df*	*b*	*SE*	G^2	*p*
Intercept	1	1.524	0.254	.	.
Identity	1	0.011	0.277	0.05	.822
Voice 2 speech type	1	0.233	0.262	6.52	.011
Background noise order	1	0.449	0.273	5.17	.023
Identity x Voice 2 speech type	1	0.057	0.371	0.86	.355
Identity x Background noise order	1	0.105	0.388	0.22	.639
Voice 2 speech type x Background noise order	1	0.415	0.365	0.25	.615
Identity x Voice 2 speech type x Background noise order	1	0.557	0.510	1.17	.280

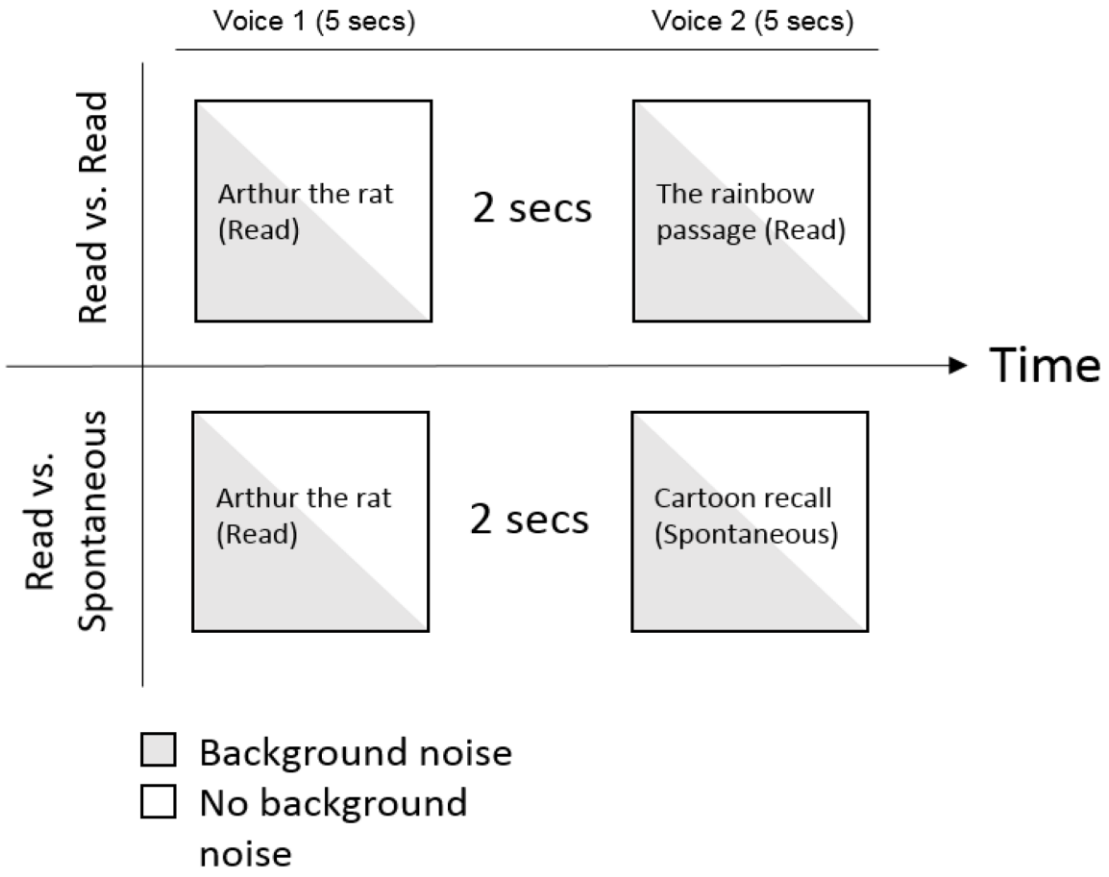

Figure 1. The procedure used in Experiment 1. In half of the trials both recordings featured background noise, and in the other half both recordings featured no background noise

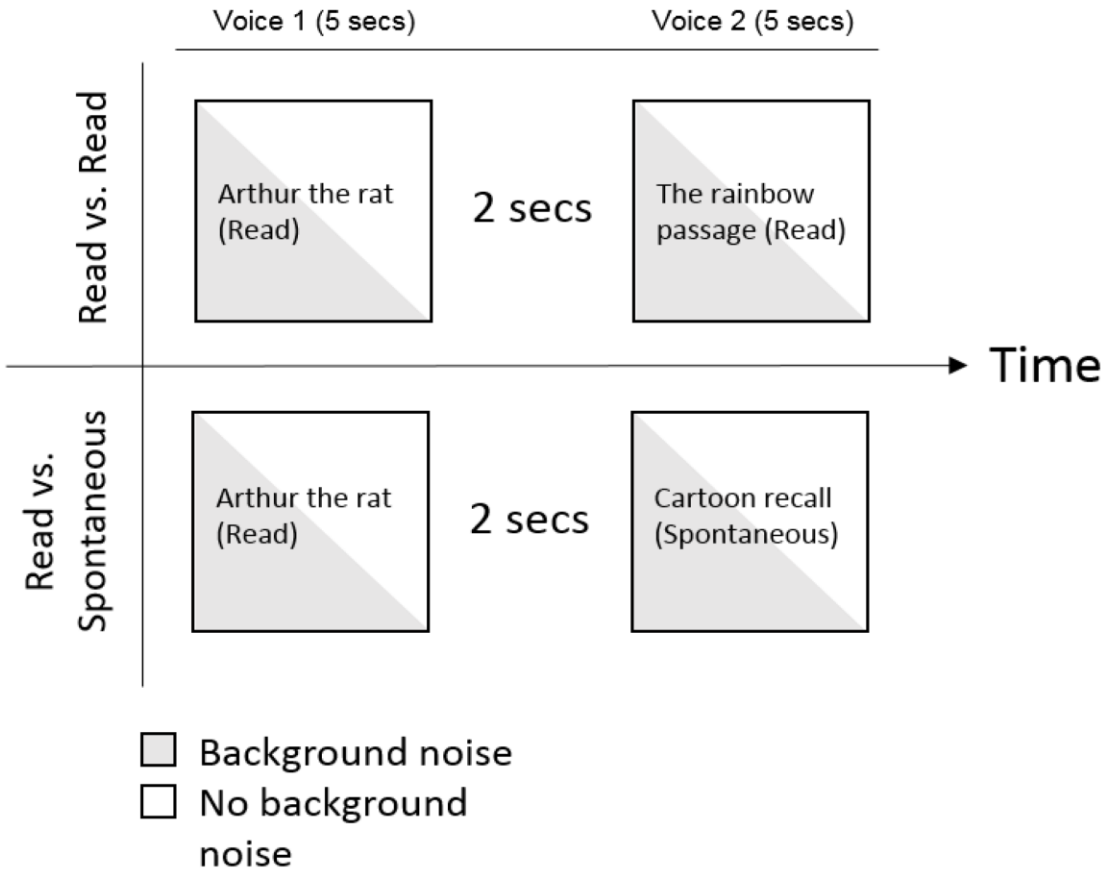

Figure 2. Voice discrimination accuracy on Read vs. Read (panel A) and Read vs. Spontaneous (panel B) trials for Experiment 1. Error bars show 95% CI for the condition means

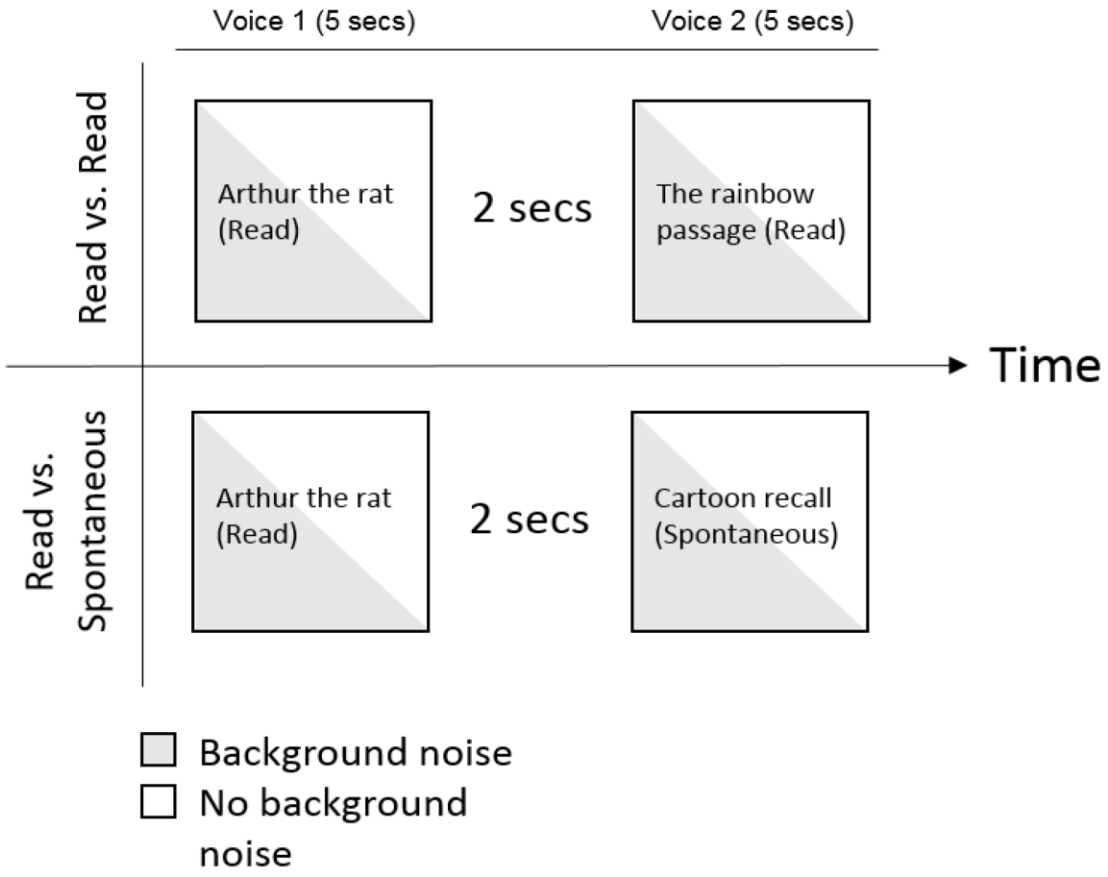

Figure 3. Self-rated confidence following voice discrimination decisions, Experiment 1

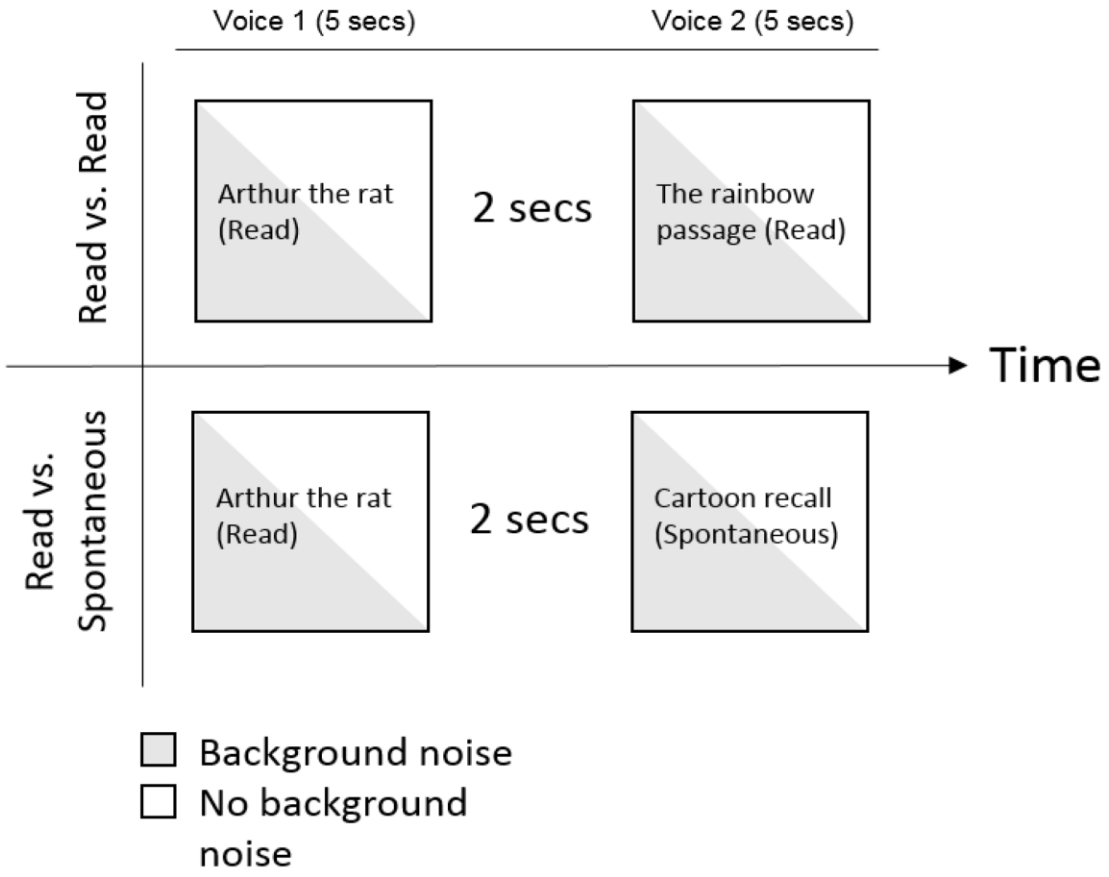

Figure 4a. Confidence-accuracy calibration for read vs. read trials, Experiment 1. Error bars are *SE*. Diagonal line shows perfect calibration

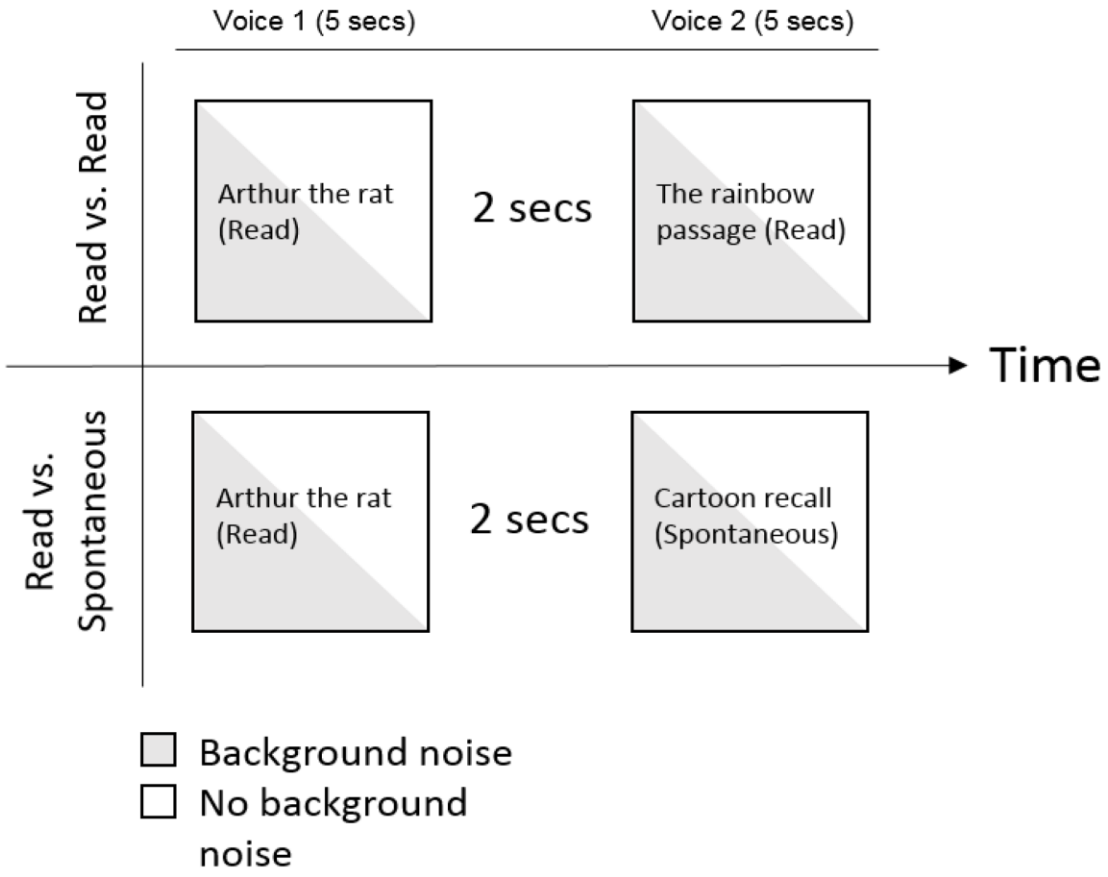

Figure 4b. Confidence-accuracy calibration for read vs. spontaneous trials, Experiment 1. Error bars are *SE*. Diagonal line shows perfect calibration

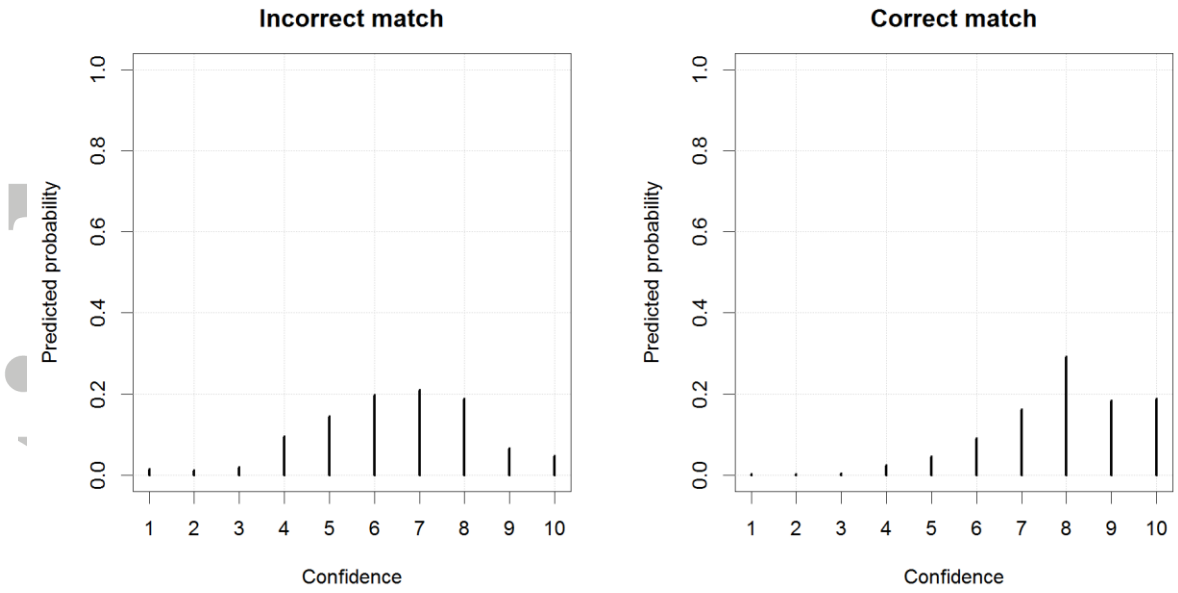

Figure 5a. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. spontaneous, no background noise condition

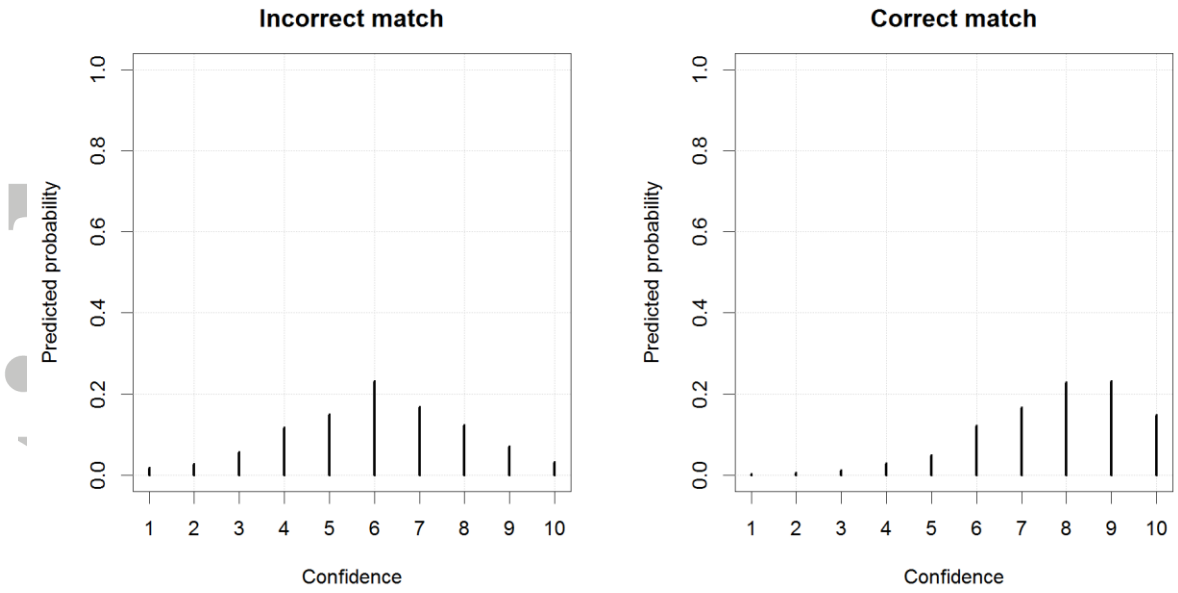

Figure 5b. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs spontaneous, background noise condition

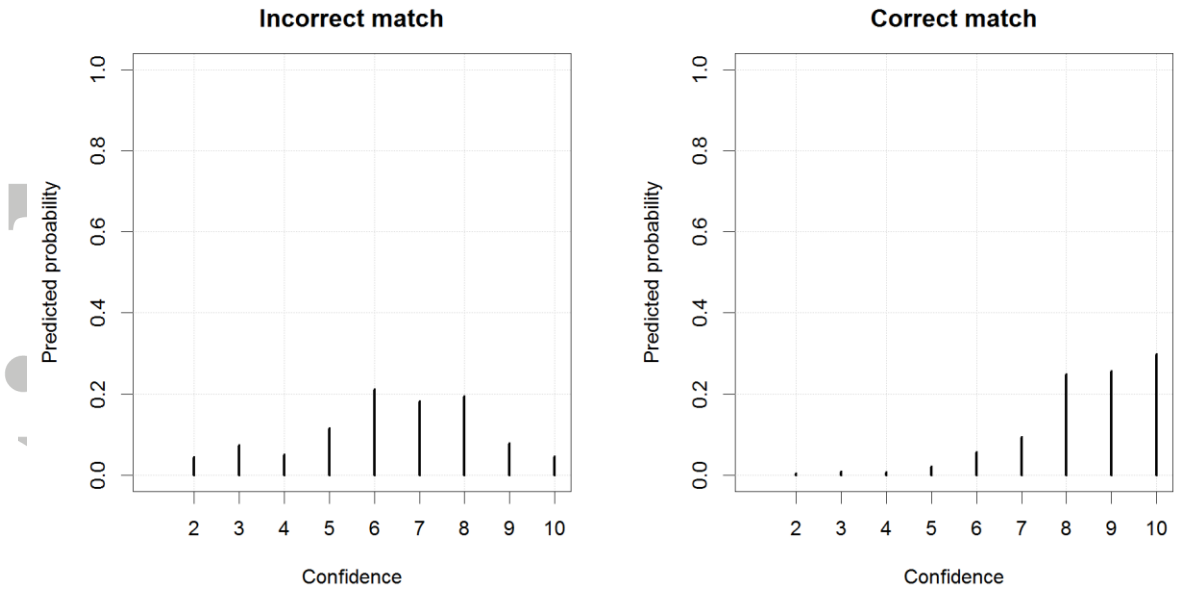

Figure 5c. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. read, no background noise condition. (N.B. In this condition, there were no confidence ratings of '1')

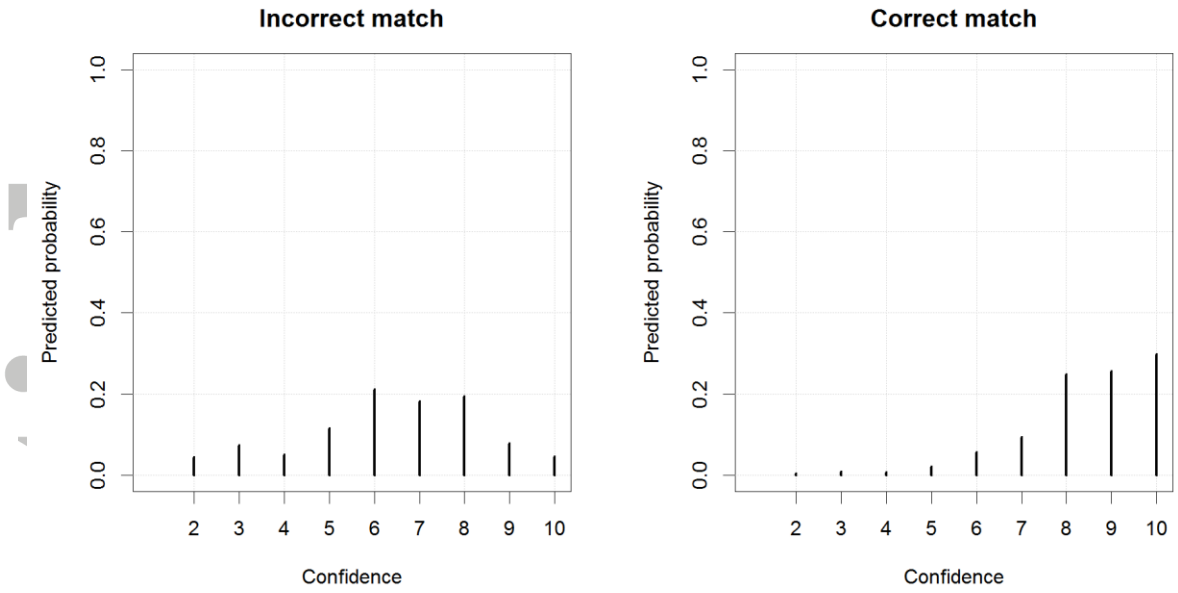

Figure 5d. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. read, background noise condition

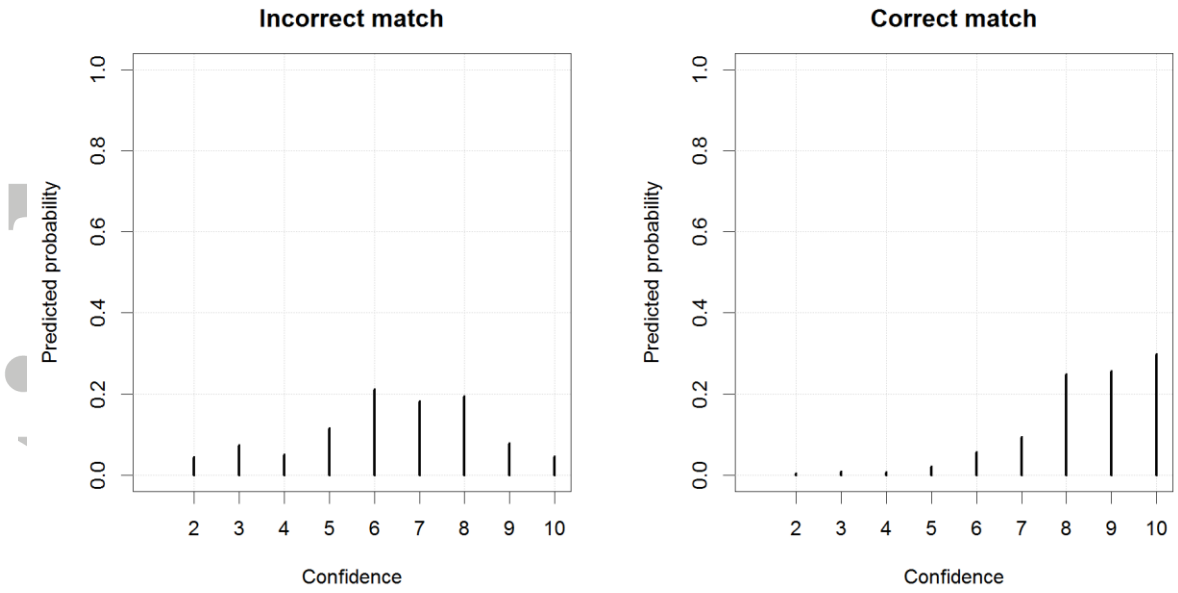

Figure 6. Voice discrimination accuracy for read vs. read (panel A) and read vs. spontaneous (panel B) trials for Experiment 2. Error bars show 95% CI for the condition means

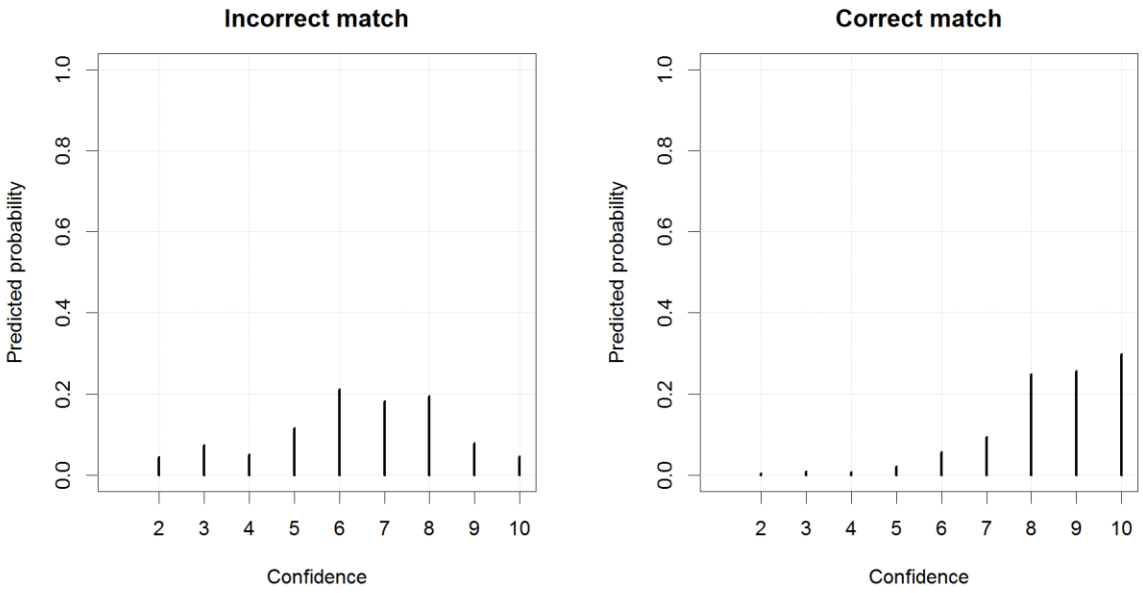

Figure 7. Self-rated confidence following voice discrimination decisions, Experiment 2

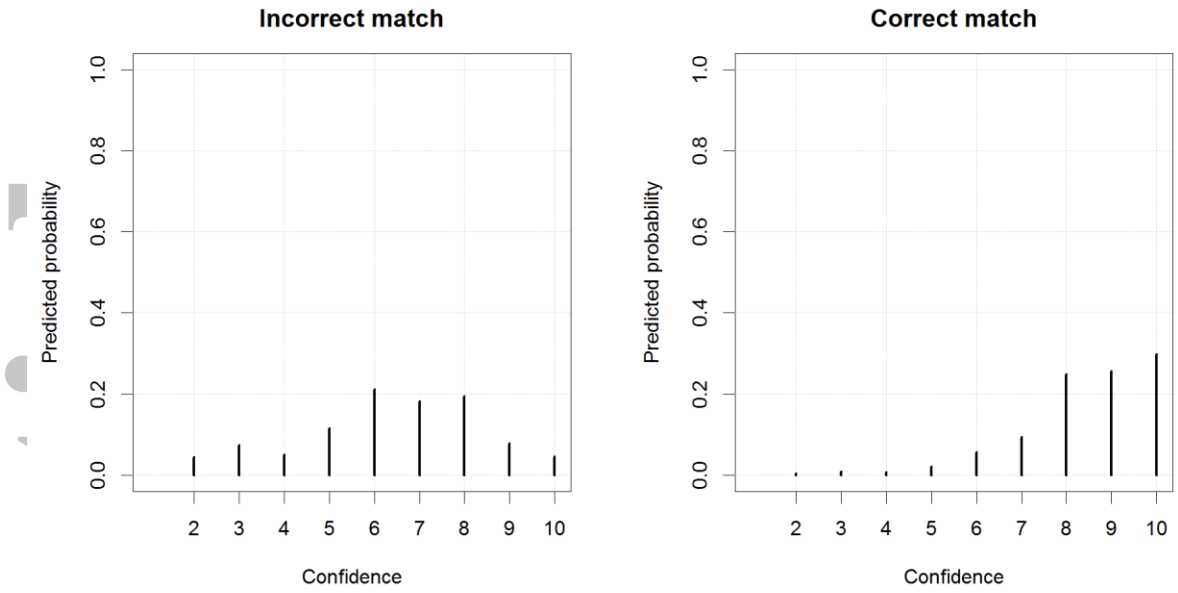

Figure 8a. Confidence-accuracy calibration in read vs. read trials, Experiment 2. Error bars are *SE*. Diagonal line shows perfect calibration

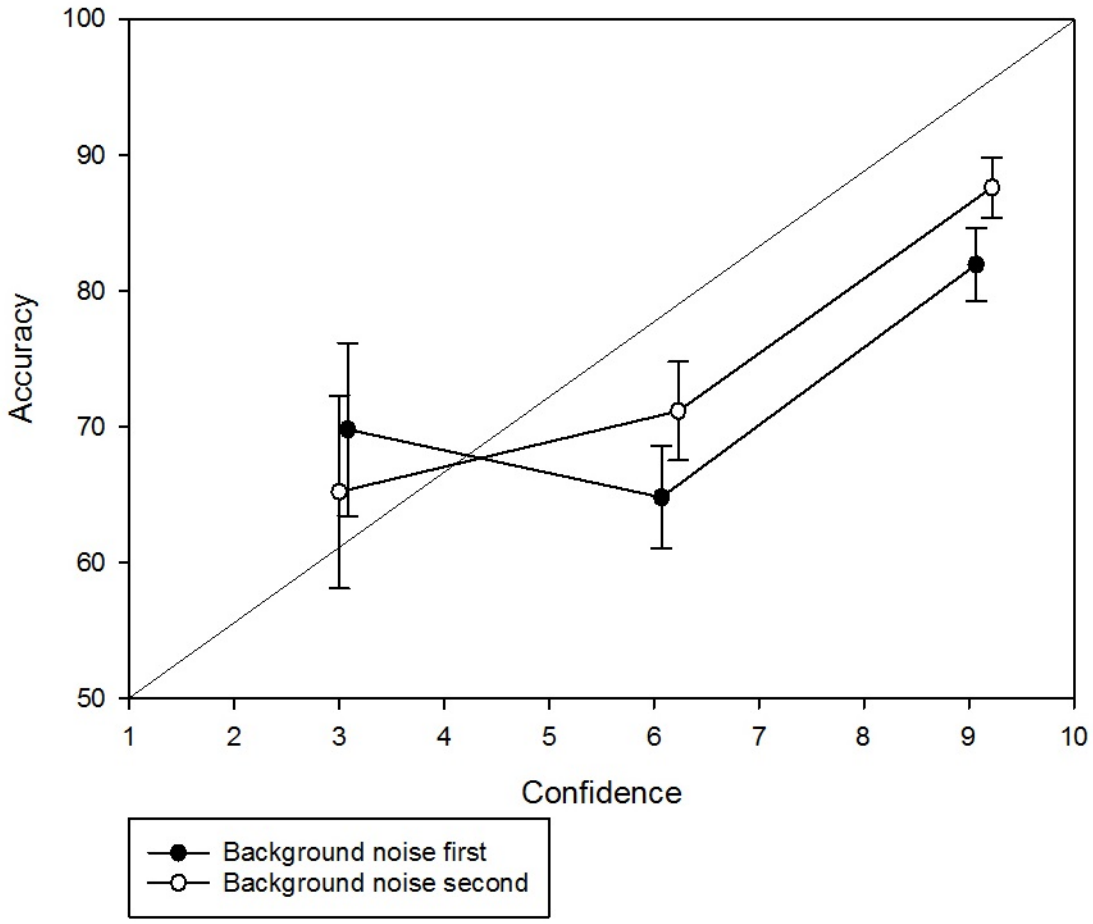

Figure 8b. Confidence-accuracy calibration for read vs. spontaneous trials, Experiment 2. Error bars are SE. Diagonal line shows perfect calibration

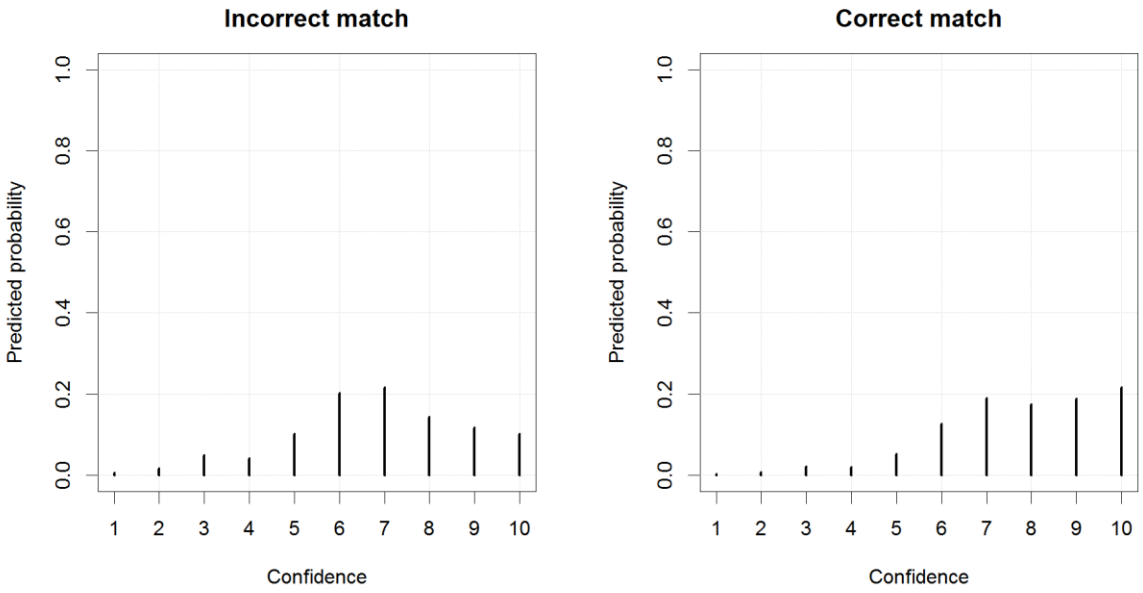

Figure 9a. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. spontaneous, background noise second condition

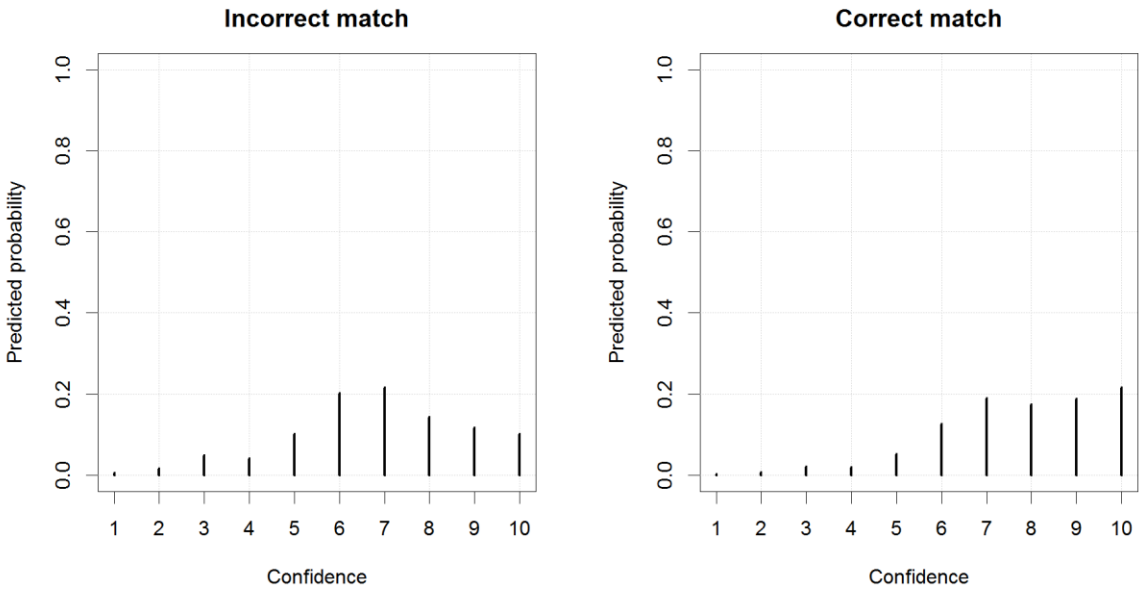

Figure 9b. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. spontaneous, background noise first condition

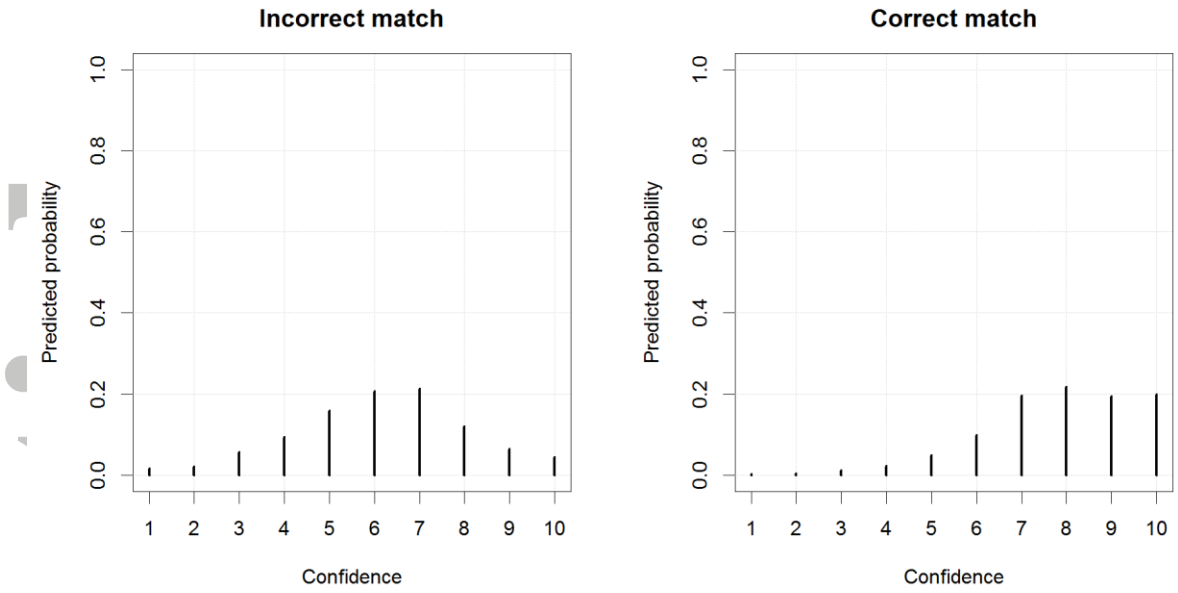

Figure 9c. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. read, background noise second condition

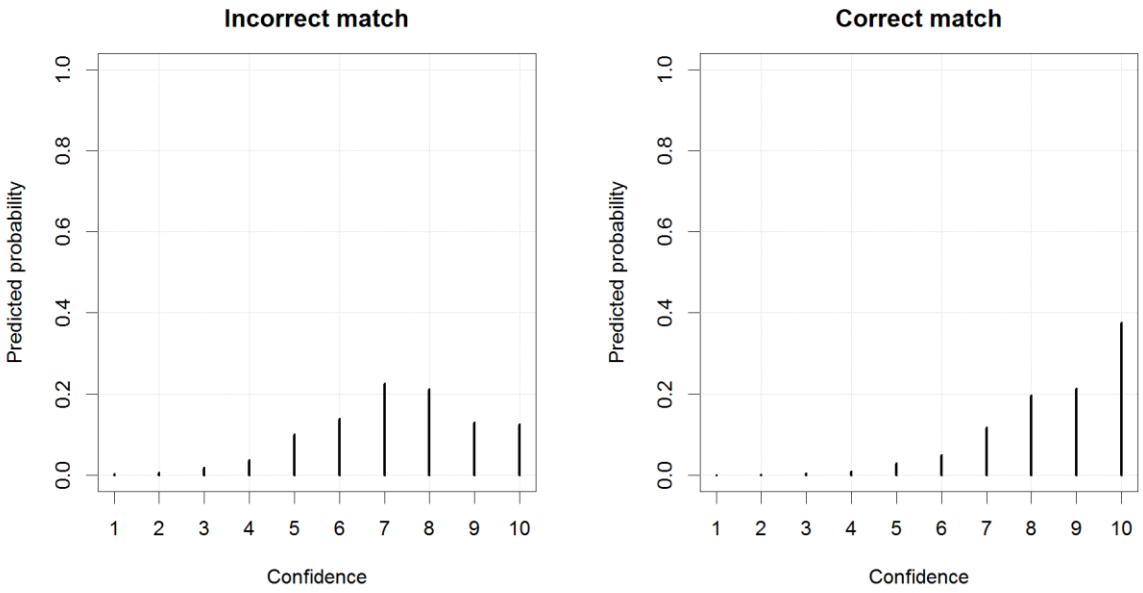

Figure 9d. Probability of an incorrect match (left) and correct match (right) at each level of confidence in the read vs. read, background noise first condition