RESEARCH ARTICLE

Unfamiliar voice identification: Effect of post-event information on accuracy and voice ratings

Harriet Mary Jessica Smith[*] and Thom Baguley[*]

This study addressed the effect of misleading post-event information (PEI) on voice ratings, identification accuracy, and confidence, as well as the link between verbal recall and accuracy. Participants listened to a dialogue between male and female targets, then read misleading information about voice pitch. Participants engaged in verbal recall, rated voices on a feature checklist, and made a lineup decision. Accuracy rates were low, especially on target-absent lineups. Confidence and accuracy were unrelated, but the number of facts recalled about the voice predicted later lineup accuracy. There was a main effect of misinformation on ratings of target voice pitch, but there was no effect on identification accuracy or confidence ratings. As voice lineup evidence from earwitnesses is used in courts, the findings have potential applied relevance.

The majority of research in witness identification is concerned with visual memory (Yarmey, 1995). However, in cases such as masked attacks or telephone fraud, witnesses may be required to identify culprits by their voice. Voice identification evidence has been used in at least 188 British legal cases (Read & Craik, 1995). Legal professionals have appealed for guidance on the reliability of earwitness evidence and the variables affecting accurate voice identification (Clifford, 1980; Wilding, Cook & Davis, 2000) but insufficient empirical earwitness research has been undertaken to fulfil this need (Kerstholt, Jansen, Van Amelsvoort & Broeders, 2004).

Unfamiliar voice identification

Intra-individual voice variation makes voice recognition more difficult than face recognition (Hammersley & Read, 1996; Stevenage, Howland & Tippelt, 2011). Correct identification rates for unfamiliar voices vary widely across studies (Kerstholt et al., 2004), but performance is often poor (Hammersley & Read, 1996), especially for incidental tests of memory (Clifford, 1980). Some experiments allow for intentional, and therefore more efficient, encoding. Participants are told that they will have to recognise presented voices later on (e.g. Kerstholt et al., 2004; Legge et al., 1984). Such studies may offer superficially high estimates of accuracy, which are not generalisable to real forensic contexts (Clifford, 1980).

Numerous variables, such as length of speech sample, lineup type (target-present or target-absent) and reten-

tion interval, affect unfamiliar voice identification accuracy. For example, Clifford (1980) concluded that length of speech sample was unimportant provided more than one full sentence was presented. However, Legge et al. (1984) observed a significant improvement in performance after speech duration increased from 6 to 60 seconds. Lineup type also affects accuracy. Research has uncovered high misidentification rates on target-absent lineups, even when accuracy is high on target-present lineups (Kersholt et al., 2004; Philippon, Cherryman, Bull & Vrij, 2007). Participants' use of relative judgments could prevent elimination (Philippon et al., 2007), or demand characteristics may lead to an assumption that the target is present (Van Wallendael et al., 1994). Studies have also investigated the effect of retention interval, in order to establish the point at which memory decay begins to negatively affect accuracy. In Saslove and Yarmey's (1980) between-subjects study design, there was no significant difference between performance on immediate lineup tests, and performance after a 24-hour delay. One study found no decline after 10 days (Legge et al. 1984). However, without efficient encoding, retention intervals may disrupt accuracy to a greater extent than is implied by intentional memory test results (Deffenbacher et al., 1989).

Reviews have highlighted a gap in research addressing the influence of misinformation on memory for voices (Yarmey, 1995). In a forensic context, misinformation may be introduced when witnesses discuss an incident, or after suggestive questioning. Research has demonstrated a significant negative effect of misinformation on eyewitness accuracy (Loftus, 1992). It is not clear to what extent findings apply to earwitnesses.

[*] Nottingham Trent University, United Kingdom
hmjs1@leicester.ac.uk, thomas.baguley@ntu.ac.uk

The misinformation effect in eye and earwitness research

Eyewitness research shows that witnesses are most likely to accept subtle misinformation, and that such misinformation can reduce recall accuracy by up to 30–40% (Loftus, 1992). Long retention intervals allow for greater incorporation of misinformation in memory (Frost, 2000). According to the memory impairment hypothesis, misinformation permanently overwrites original memories (Loftus, 1981). Other theorists argue that reporting false information does not indicate true memory (Zaragoza & Koshmider, 1989) because contextually rich tests show no effect of misinformation (McCloskey & Zaragoza, 1985). Reporting of misinformation may also take place because memory for the original event was not encoded (Zaragoza & Lane, 1994). According to the source monitoring hypothesis (Johnson, Hastroudi & Lindsay, 1993), memories for the original and suggested event coexist, but 'reality monitoring errors' cause the source of memories to be confused.

The majority of misinformation studies have concentrated on the disruption of recall accuracy, although misinformation also has the potential to reduce recognition accuracy. A misinformation effect has been observed in face identification studies. Loftus and Greene (1980) showed participants a face then asked them to listen to another witness' description of the face. If other witnesses mentioned a misleading feature, more than two-thirds of participants later identified a face with that feature. A more recent study has also addressed the effect of misinformation on lineup decisions. Searcy, Bartlett and Memon (2000) gave participants post-event information (PEI) with a misleading description of the target face. This biased participants to select foil faces matching the description on target-absent lineups. A foil face is an innocent distractor which has not been seen previously. However, their study did not include a target-present lineup, so the design did not detect whether hit rates would have been negatively influenced by this manipulation.

Auditory memory may be particularly suggestible. Visual memory is stronger (Posner, Nissen & Klein, 1976), and dominates auditory memory (Howard, 1982). Weaker memories are more susceptible to PEI in eyewitness research (Loftus, Miller & Burns, 1978), perhaps because inefficient encoding and storage makes them difficult to distinguish from misinformation (Johnson, Hastroudi & Lindsay, 1993).

A comparison of the effect of marked modifiers such as 'hit'/'smash' on estimates of speed in auditory and visual input modalities indicated that auditory memory was more susceptible to misinformation (MacAllister, Bregman & Lipscomb, 1988). However, use of stationary speakers prevented participants using sound localisation cues to encode a memory with realistic perceptual quality. Memories lacking perceptual quality are more prone to misinformation (Belli, 1989). Nevertheless, other studies have also found that auditory memory is relatively more distortable than visual memory (Campos & Andalonso-Quecuty, 2006).

Research has not addressed the effect of misleading PEI on memory for voices. In a police interview, witnesses give a verbal description of voices, as well as attempting a voice lineup (Yarmey, 1994). The present study aims to investigate the effect of misinformation about pitch on voice ratings and identification accuracy. Pitch is a central voice feature (Orchard & Yarmey, 1995), aiding voice identification due to its stability (Mullennix et al., 2010) and limited intra-individual variation (Hollien, 1990).

Memory for pitch and other paralinguistic voice characteristics

Paralinguistic voice characteristics include pitch, rate of speech, frequency of pauses, and enunciation. Handkins and Cross' (1985) voice profile checklist of within and between-speaker variation gives a better measure of listener perception than vague verbal description (Yarmey, 1995). Some properties of speech, including pitch, are automatically encoded with content of speech (Pisoni, 1993).

It is uncertain whether inaccurate ratings could disrupt identification accuracy. Research has not addressed whether accuracy of speech ratings predicts performance at lineup (Yarmey, 1995), or which aspects are crucial to voice recognition (Bricker & Pruzansky, 1976). However, features of pitch, expressive style, age of speaker, enunciation, and inflection were most accurately remembered at a one-week retention interval for distinctive voices (Yarmey, 1991a). Voice identification accuracy appears not to decay over the period of a week (Legge et al., 1984), so retention of these key features may predict performance (Kersholt et al., 2004).

Recent research however suggests memory for pitch may not predict accuracy on lineups. Memory for pitch is consistently inaccurate, operating according to predictable distortions. Digital manipulation of a target voice into different versions varying in pitch led to predictable errors at lineup (Mullennix et al. 2010). When the target had a high-pitched voice, foils with a higher pitch were selected. When the target was low-pitched, lower pitched foils were selected. Pitch may however operate as a cue for other paralinguistic information. Despite the difficulty of identifying unfamiliar voices, listeners are reasonably accurate when estimating a speaker's age (Linville, 1996). Cues such as lower pitch, slower rate of speech and increased pauses in adults indicate older age across genders (Linville, 1996).

Verbal recall and identification accuracy

The relationship between verbal recall and accuracy in earwitness testimony requires examination. Eyewitness studies have not found a reliable relationship between the accuracy of descriptions in free recall and accuracy of face identification (Pigott, Brigham & Bothwell, 1990). In some studies verbal descriptions disrupt recognition ability. Schooler and Engstler-Schooler (1990) found that participants with higher levels of verbal recall made more false identifications at lineup, regardless of description accuracy. Other studies have detected a verbal facilitation effect; the act of verbally describing the perpetrator can,

in some cases, make participants more likely to respond accurately at lineup (e.g. Meissner, Brigham & Kelley 2001). Application of these findings to the earwitness context has not been sufficiently investigated (Yarmey, 2007).

Self-reported confidence and identification accuracy

Self-reported confidence significantly influences mock jurors judging the reliability of witnesses' voice identifications (Van Wallendael, Surface, Parsons & Brown, 1994). However, although witness confidence may influence jurors decisions, confidence does not reliably indicate identification accuracy. The majority of studies have found confidence to have no overall value in predicting correct identification (Kerstholt et al., 2004; Yarmey, 1995; Yarmey & Matthys, 1992). Weak but significant positive correlations have occurred in some conditions, such as longer initial speech sample (Yarmey & Matthys, 1992). This supports the optimality hypothesis (Deffenbacher et al., 1989), that confidence and accuracy are only related in easy tasks. Other findings challenge this hypothesis, with studies observing positive correlations only in supposedly more difficult conditions (Orchard & Yarmey, 1995; Philippon et al. 2007). Overall, findings suggest the relationship between confidence and accuracy is insufficiently consistent to indicate performance in experimental or forensic situations. False memory studies show that the predictive value of confidence decreases even further following exposure to misleading information (Tomes & Katz, 2000). Heightened confidence often accompanies incorrect responses (Weingardt, Leonesio & Loftus, 1994).

Aims of the present study

The present study uses an incidental test of memory so that findings are generalisable to a forensic context. The study aims to initially explore misinformation effects in the under-investigated area of voice memory, addressing the effect on ratings of speech, identification accuracy, and confidence. Although some studies indicate that misinformation effect sizes may be large for auditory memory (MacAllister, Bregman & Lipscomb, 1988; Campos & Andalonso-Quecuty, 2006), other results imply that memory for pitch might be resistant to misinformation because it is central to voice recognition (Hollien, 1990; Orchard & Yarmey, 1995; Luna & Migueles, 2009; Mullennix et al., 2010). Even if memory for pitch is distorted, the research reviewed here offers conflicting evidence regarding the extent to which identification accuracy will be affected.

The present study investigates the power of various factors in predicting unfamiliar voice identification accuracy. These include verbal recall and confidence. Although previous studies have addressed the predictive power of verbal recall for face identification (e.g. Schooler & Engstler-Schooler, 1990), the effect has not been sufficiently tested for voice identification. Based on previous research, we hypothesised that there would be low overall accuracy rates, but that performance would be better on target-present lineups. We anticipated that confidence would have a weak relationship with accuracy.

Method

Design

The experiment employed a 2 × 2 mixed factorial design. The within-subjects factor was stimulus voice (male or female). The between-subjects factor was false information (about the male or female voice). The dependent variables were voice ratings (Handkins & Cross, 1985) and length of verbal recall about the target voices.

For the lineup analysis only, there was a third within-subjects factor: lineup type (target-present or target-absent). This factor only applied to the lineup analysis because the other dependent variables were measured prior to this manipulation. For this analysis the dependent variables were lineup accuracy, voice selection and self-reported confidence.

Participants

There were 33 male and 39 female adult participants (N = 72) in this study ranging in age from 18 to 64 years (M = 37.6 years, SD = 15.7 years). None of the participants reported suffering from hearing deficits.

Materials

The initial voice sample consisted of a 37-second dialogue between a white female and white male, both with relatively low-pitched voices. The male and female were aged 28 years, 11 months and 27 years, 8 months respectively. The two perpetrators disguised a discussion about drug smuggling by appearing to talk about a holiday. Neither suspect had a regional accent, idiosyncratic speaking style, or speech impediment. Both said 71 words each during the dialogue, and spoke in a normal conversational style.

Target-present and target-absent lineups were constructed for the male and female voice, and consisted of 5 voices. Voices were selected from undergraduates and graduates from Nottingham Trent University. Foil and target voices were matched for gender and age group (20–30 years). None of the foils had a pronounced regional accent, idiosyncratic speaking style, or speech impediment. Both targets and all foils were recorded saying a 4-syllable phrase from the original dialogue (males: 'I've just got back'; females: '…a red wine thanks').

The dialogue and lineups were recorded onto Audacity (2.0.0) in a silent room. Microphone settings were constant across targets and foils. All files were exported into iTunes (10.6.1 (7)). 'Stop your Sobbing' by the Pretenders, a song of duration 2:39 minutes, was used as auditory interference. Stimuli were played to participants through Sony (Model No. SRS-A202) speakers.

False information was embedded in a relative clause, in a paragraph informing participants that both suspects had been arrested. Depending on the condition, participants read that other witnesses reported either the male or female as having a high-pitched voice.

Participants completed Handkins and Cross' (1985) checklist of voice characteristics for both target voices. The checklist includes an estimate for age, as well as 10-point Likert-style rating scales for 9 voice characteris-

tics (see Appendix A). An item for pitch (1 - *low*, 10 - *high*), not included in the original checklist, was also added, so the misinformation effect could be quantified.

Procedure

Participants provided demographic information, read an information sheet and completed a consent form. To allow for an incidental test of memory, the information sheet indicated that the study was about speech perception, not that it involved false information or voice identification.

Participants were randomly and equally allocated to each of the 4 combinations of between-subject conditions using an online research randomiser (Urbaniak & Plous, 2011) (false information: about the male or female voice, and lineup type: target-present or target-absent). Participants were tested individually in a quiet room. Speaker volume was set at a constant level across all participants. Participants imagined they were eavesdropping on a conversation in a pub and tried to decide what was being discussed. When the dialogue finished, the song played. Participants then read the text containing false information. Having allowed time for careful reading, the experimenter recorded participants verbally recalling as much as possible about the target voices. Recording stopped when participants said they could not remember anything else. These descriptions were later coded by the researcher for number of descriptors, by counting each item of information relating specifically to the voice. Any reference participants made to the content of speech was not counted. Descriptions were not coded for accuracy.

Participants completed the voice characteristic checklist (Handkins & Cross, 1985), and the experimenter explained that 2 lineups of 5 voices would be played, one for each target. Participants attempted to identify the male and female suspects by selecting the corresponding number (1–5). They were told lineups would only be played once, and that the targets may or may not be present, but that all voices were different. There was an option to tick 'not present'. Following each lineup, participants rated their confidence that the target had, or had not, appeared in the lineup on 10-point Likert-style rating scales (1 - *not at all confident*, 10 - *extremely confident*). Participants who made a positive identification gave an extra rating, on the same scale, to indicate confidence in their selection. Target-present and target-absent lineups were randomly ordered, as was the order of lineup voices (Urbaniak & Plous, 2011). Participants who listened to a female target-present lineup also listened to a male target-absent lineup, and vice versa. All participants were fully debriefed after the study, and told that they had been given false information. Contact details for the researcher were provided.

Results

Ratings of voices

The descriptive statistics for ratings of target voice features (Handkins & Cross, 1985) are shown in Appendix B.

Effect of false information on ratings of pitch

A 2 (stimulus voice: male or female) × 2 (false information: male or female) fractional factorial design was employed with false information as a between-subjects factor, and stimulus voice as the within-subjects factor. This reduced the complexity of the design and reduced the interference between trials at the expense of a noisier test of the interaction effect (e.g., see Kirk, 1995, p.656). This design was analysed using the nlme package in R (Pinheiro et al., 2012; R Core Team, 2012). This approach allowed separate variances for the male and female pitch ratings, as well as the covariance between them to be estimated. There was an effect of stimulus voice, $F(1, 69) = 33.57$, $p < .001$, with the female voice being rated as higher in pitch. There was also a main effect of false information, $F(1, 69) = 4.81$, $p = .044$, with voices being rated as higher in pitch following false information. There was no significant interaction between false information and stimulus voice, $F(1, 69) = .33$, $p = .57$.

Figure 1 shows the main effect of false information on target pitch ratings, including the two-tiered 95% confidence intervals (CIs) (Baguley, 2012a) for mean pitch ratings.

Lineup performance

Accuracy was higher in the target-present lineup than in the target-absent lineup. In target-present lineups, on average, the target was correctly identified 37.5% of the time: 36.1% in male target present lineups, and 38.9% in female target-present lineups. In target-absent lineups, only 15.2% of responses correctly indicated that the target was not present. Chance performance for target-absent lineups was 16.7%, which falls within the 95% CI [5.1%, 20.1%], for participant performance on target-absent lineups. Therefore overall target-absent performance was consistent with chance. Accurate rejections constituted 11.1% of responses on male target-absent lineups, and 19.4% of responses on female target-absent lineups. The difference between the average number of accurate responses on target-present and target-absent lineups was statistically significant, as shown in the multilevel logistic regression reported below.

Predictors of accuracy

As the dependent variable (accuracy) was discrete (1 = correct response, 0 = incorrect response), a multilevel (repeated measures) logistic regression was performed on the data using the lme4 package in R (Bates, Maechler & Bolker, 2011; R Core Team, 2012). The predictor variables were lineup (1 = target-present, 0 = target-absent), false information (1 = male, 0 = female), target (1 = male, 0 = female) and number of items of voice information given in free recall. These predictors were included in a main effects only model ($\chi^2_6 = 148.7$, AIC = 160.72), and in addition, a two-way model including all two-way interaction terms. The two-way model did not improve the fit, and was not more informative ($\Delta\chi^2_6 = 5.92$, ΔAIC = 6.08). Furthermore, no two-way interaction term was statistically significant when added individually (all $p > .05$). The likelihood ratio showed that the main effects model was
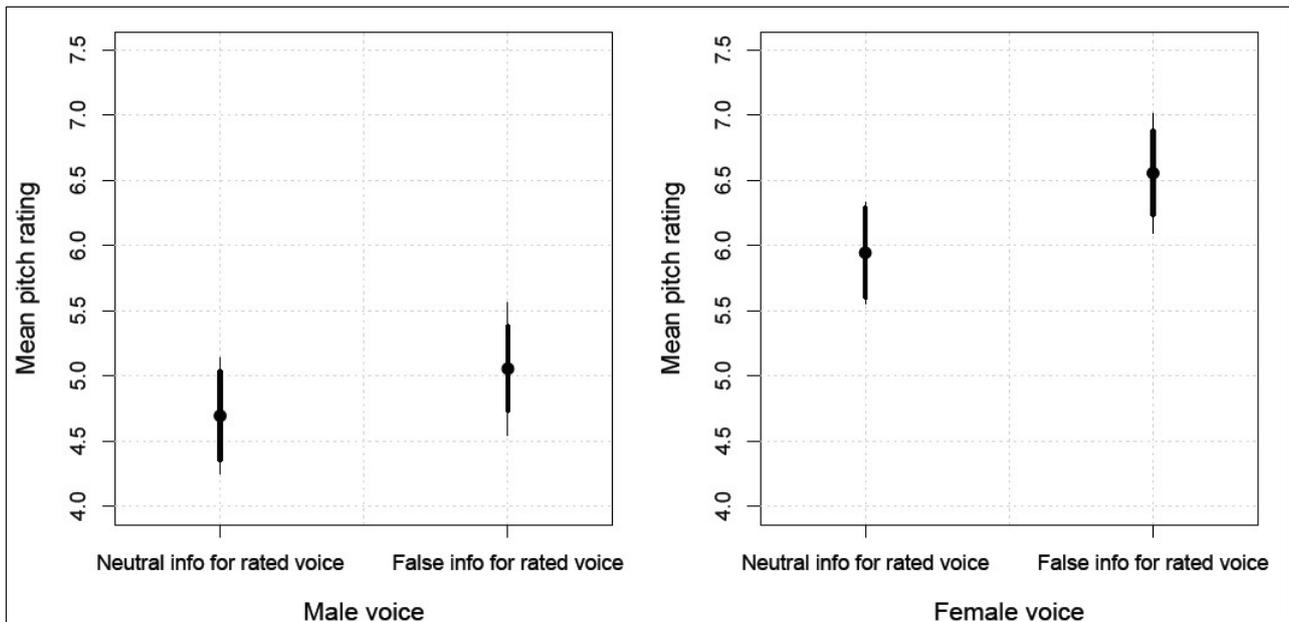
Figure 1: The effect of stimulus type (male or female voice) and false information on ratings of voice pitch, with two-tiered CI for mean pitch ratings. The outer tier (thin lines) depicts a 95% CI for the individual mean. The inner tier (thick lines) is adjusted so that means with intervals that do not overlap are different with approximately 95% confidence.

20 times more likely than the interaction model (LR_{AIC} = 20.90). For this reason we report in detail results of the main effects model only.

Each item of information increased the log odds of making a correct identification by 0.46. This is equivalent to increased odds of correct identification of approximately 1.6 for each item of voice information, $p < .05$, 95% CI [1.04, 2.42]. In the target-present condition, the log odds increased by 1.52, equivalent to increased odds of correct identification of 4.56, $p < .01$, 95% CI [1.84, 11.28]. False information ($p = .50$) and target ($p = .20$) were not significant predictors.

Accuracy and confidence ratings
A total of 8 correlations were calculated between confidence and accuracy. We report unadjusted CIs but include adjusted p values from a Holm test to control for multiple testing (Baguley, 2012b, p. 492). The degrees of freedom vary for the correlations reported here because participants who selected 'not present' at lineup only contributed ratings for confidence that the target was present or absent in the lineup. They did not contribute ratings for confidence that the correct voice had been selected. There were no reliable relationships found between overall accuracy and confidence that the target was present or absent in the lineup for the male voice, $r(70)= .13$, $p = 1.00$, 95% CI [-.11, .365] or the female voice , $r(70) = .13$, $p = 1.00$, 95% CI [-.36, .11]. For the female voice there was also no significant relationship between accuracy and confidence that the correct suspect had been identified, $r(60) = .154$, $p = 1.00$, 95% CI [-.10, .41]. However, there was a significant relationship between accuracy and confidence that the correct male voice had been identified, $r(63) = .33$, $p = .049$, 95% CI [.10, .57]. Analysis was also

conducted on lineup types. One significant relationship was found in the target-present lineup, between accuracy and confidence that the correct male suspect had been identified, $r(31) = .48$, $p = .040$, 95% CI [.155, .80]. There were no reliable relationships in target-absent lineups.

Giving participants false information about the stimulus voices did not affect confidence ratings. Lineup type did not affect ratings either. Four 2 × 2 between-subjects ANOVAs were performed on the data, with scores in each confidence condition as the dependent variable, and false information (male or female) and lineup type (target-present or absent) as the independent variables. There were no statistically significant main effects or interactions (all $p > .75$).

Discussion
This study investigated the effect of post-event information (PEI) on identification accuracy and voice ratings in an incidental test of unfamiliar voice memory. The design also captured the effect of misinformation on confidence ratings, as well as addressing the relationship that self-reported confidence and verbal recall have with accuracy. Results showed low hit rates at lineup, and particularly inaccurate target-absent lineup performance. Fullness of verbal voice description increased the odds of making a correct identification, but overall self-reported confidence was not related to accuracy. There was a main effect of false information on ratings of voice pitch. For example, compared to participants who read neutral information about the female voice, those given false information about the female voice rated it as being significantly higher in pitch. This pattern also occurred for ratings of the male pitch. Results did not show an effect of misleading information on identification accuracy or confidence ratings. Participants were accurate when rat-

ing target age. The male target was aged 28 years, 11 months, and the female target was 27 years, 8 months. This matches participants' estimates of target age; the male and female target were rated as being 26.9 and 28.1 years old respectively. The actual age of the targets were close to the centre of the confidence intervals, both of which were relatively narrow (approximately 2 years for each target) (see Appendix B).

Overall lineup performance

Hit rates in the target-present lineup were low (37.5%), but above chance, consistent with results of other incidental tests of memory (Yarmey, 1991b). However, as correct response rates vary widely across studies (Kerstholt et al., 2004), it is necessary to consider some of the possible reasons why performance was poor in the present study. Hearing two voices in the initial speech sample may have caused interference effects, preventing optimal retention of memory for either voice (Neath & Surprenant, 2003). A further possible explanation for low hit-rates is that non-distinctive voices are more difficult to identify (Orchard & Yarmey, 1995). Participants rated the targets as relatively average. 95% confidence intervals for voice feature ratings of both targets tended to cluster around the mid-point (5). All confidence intervals included a rating of 4, 5 or 6. Confidence intervals were relatively narrow. The widest confidence interval, 1.12, was observed for female rate variation. In the target-absent lineup, around 85% of responses were false identifications; performance was at chance level. The bias towards identification, and impaired performance in the target-absent condition, compared to target-present performance, is an established finding (e.g. Kerstholt et al., 2004).

Verbal recall and identification accuracy

The multilevel logistic regression showed that the odds of making a correct identification increased slightly with the fullness of description given in free verbal recall. Although there was no control condition without free verbal recall, this preliminary finding might suggest that research showing a disruptive effect of verbal description on face identification is not generalisable to voice identification (Schooler & Engstler-Schooler, 1990). This could be because verbal description, involving focus on featural aspects, is argued to selectively impair face recognition because of holistic processing (Schooler & Engstler-Schooler, 1990). Evidence suggesting that recognition of unfamiliar voices predominantly requires feature analysis (Yarmey, Yarmey & Parliament, 2001; Belin, Fecteau & Bedard, 2004), may explain why the effect was not replicated in this study. Due to the inaccuracy of unfamiliar voice identification, any measure predicting accuracy is useful to legal professionals. Results suggest that police can be more confident that a witness has correctly identified a guilty suspect at lineup if they first provided a full description of the voice. However, further research is required to ascertain whether witnesses engaging in free verbal recall are more or less accurate than those who attempt identification without having first described the voice.

Self-reported confidence and identification accuracy

In line with previous literature (e.g. Olsson et al., 1998), overall self-reported confidence was not significantly related to accuracy. The only condition in which confidence significantly correlated with accuracy was confidence that the male voice had been identified in target-present lineups. The optimality hypothesis (Deffenbacher et al., 1989), suggesting that the relationship between confidence and accuracy is stronger in easy memory tasks, does not wholly account for these results. Although target-present lineups are known to be easier, hit rates in this study were in fact higher for the female voice. Contrary to previous findings (Weingardt, Leonesio & Loftus, 1994; Tomes & Katz, 2000), confidence ratings were not inflated in misinformation conditions. This may be due to the limited effect of misinformation, discussed below. The relationship between confidence and accuracy appears to be unpredictable. Legal professionals should not use witnesses' self-reported confidence to assess the likelihood of identification accuracy.

Effect of misinformation

False information about pitch had a significant effect on ratings of target voice pitch. The influence of false information on pitch ratings was small; participants only rated the voices as 10% higher-pitched following misinformation. Other studies have observed 30–40% memory distortions when investigating misinformation and recall accuracy (Loftus, 1992). However, the present study used a 1 to 10 rating scale for pitch. This perhaps offered a more sensitive measure of the misinformation effect than some prior research. False memory studies commonly force participants to choose between items in a two-alternative forced choice recognition test (Pezdek & Lam, 2007). Such an approach polarises responses, with the potential of superficially inflating effect sizes. Other research offers alternative but complimentary explanations. It is possible that the short retention interval did not facilitate full incorporation of misinformation into memory (Frost, 2000). Results could also be explained by the fact that pitch, as a central aspect of speech (Orchard & Yarmey, 1995), is resistant to disruption by misinformation (Luna & Migueles, 2009). The misinformation effect may vary widely across voice features depending on how central or peripheral they are.

Misinformation did not affect accuracy of identification at lineup. Low hit rates in the study showed memory for unfamiliar voices is weak. Previous research suggests weak memories are more suggestible (Loftus, Miller & Burns, 1978). The failure to disrupt identification accuracy should be considered in light of other false memory research. Findings may correspond with Zaragoza and Koshmider's (1989) hypothesis that reporting false information does not necessarily reflect true memory. In contrast to the rating of pitch, identification tests could be a more accurate test of true memory. Alternatively, it is possible that whilst some source monitoring errors (Johnson, Hastroudi & Lindsay, 1993) operated when participants rated pitch in cued recall, distinguishing between the origin of memories was easier when attempting identification at lineup.

The key to understanding this result could relate to the nature of pitch and memory for pitch. Memory for pitch has been shown to be consistently inaccurate (Mullennix et al., 2010), so might not offer diagnostic information in the recognition process. Equally, owing to the large number of discrete voice features, disrupting only one may fail to distort overall voice memory to the extent that it affects selection.

Directions for future research

This exploratory study addressed several research questions not investigated by previous research. Although the study observed a limited effect of false information for a single voice feature, further research should be undertaken to clarify the mechanisms by which misinformation affects memory for voices. It used non-distinctive voices, known to secure low hit rates (Orchard & Yarmey, 1995). Future research could investigate the effect of misinformation on memory for distinctive voices. In addition, research should address whether longer retention intervals increase the extent to which misinformation is incorporated into memory.

A comparison of the effect of misinformation on central and peripheral voice features would help clarify whether the results in this study were attributable to the nature of pitch. As misinformation in a real situation may be applied to numerous voice features over a period of days, one avenue of future research could compare how identification accuracy is affected in conditions where misinformation is applied to one, two, or three voice features.

Conclusion

Results obtained in this study support warnings against the use of unfamiliar voice identification as decisive evidence in court cases. If an innocent suspect has been apprehended, misidentification is likely. The fact that fullness of verbal descriptions may predict voice identification accuracy is a useful finding from a forensic point of view. Earwitness evidence is useful in offering reliable estimates of age, which might help police distinguish between two suspects. Witnesses may however encounter misinformation in the delay between the crime and subsequent questioning. This preliminary investigation suggests that ratings of voice features have the potential to be disrupted to a greater extent than identification accuracy. The study however only addressed distortability of one voice feature. Until further studies are undertaken, it is not possible to conclude whether these findings are generalisable to other (e.g. peripheral) voice features, or to draw general conclusions about the suggestibility of auditory memory for voices. Results have highlighted various areas for future research, which may help psychologists and legal professionals better understand the malleability of voice memory in a way that is relevant and useful to forensic professionals.

Acknowledgements

The authors would like to thank Dr Andrew Dunn, whose help and expertise assisted this research.

Appendix A
Handkins and Cross' (1985) checklist of 9 voice characteristics

Rate of speech (1 - very slow, 10 - very fast)
Rate variation (1 - smooth, steady, 10 - halting, jerky)
Pitch variation (1 - monotone, 10 - highly varied)
Expressive style (1 - none, 10 - very much)
Enunciation (1 - very poor, 10 - very good)
Inflection (1 - none, flat, 10 - very much)
Tremor (1 - very steady, 10 - very shaky)
Pauses (1 - very short, 10 - very long)
Nasality (1 - none, 10 - very nasal)

Appendix B
Mean ratings and 95% CIs for target voice features (SD in brackets)

Voice feature	Male target			Female Target		
	Mean	95% Confidence Interval for Mean		Mean	95% Confidence Interval for Mean	
		Lower	Upper		Lower	Upper
Age	28.06(4.29)	27.05	29.06	26.89(4.40)	25.85	27.92
Rate of speech	5.43(1.22)	5.14	5.72	5.65(1.34)	5.34	5.97
Rate variation	5.35(2.39)	4.79	5.91	4.83(1.88)	4.39	5.28
Pitch	4.87(1.49)	4.52	5.23	6.25(1.26)	5.95	6.55
Pitch variation	3.90(2.04)	3.42	4.38	4.43(1.74)	4.02	4.84
Expressive style	3.26(2.06)	2.78	3.75	4.75(2.00)	4.28	5.22
Enunciation	5.60(1.99)	5.13	6.07	6.01(1.83)	5.58	6.44
Inflection	3.93(2.02)	3.46	4.40	4.88(1.78)	4.46	5.29
Tremor	4.13(2.23)	3.60	4.65	4.51(1.91)	4.07	4.96
Pauses	4.67(2.26)	4.13	5.20	4.32(2.06)	3.84	4.80
Nasality	4.04(2.14)	3.54	4.54	3.53(2.02)	3.05	4.00

References

Baguley, T. (2012a). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44,* 158–175. DOI: http://dx.doi.org/10.3758/s13428-011-0123-7.

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

Bates, D., Maechler, M., & **Bolker, b.** (2011). *lme4: Linear mixed-effects models using S4 classes. R package version 0.999375–42.* http://CRAN.R-project.org/package=lme4

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

Belli, R. F. (1989). Influences of misleading postevent information: Misinformation interference and acceptance. *Journal of Experimental Psychology: General, 118,* 72–85. DOI: http://dx.doi.org/10.1037/0096-3445.118.1.72.

Bricker, P., & **Pruzansky, S.** (1976). Speaker recognition. In N. Lass (Ed.), *Contemporary issues in experimental phonetics* (pp. 295–326). New York: Academic Press.

Campos, L., & **Andalonso-Quecuty, M. L.** (2006). Remembering a criminal conversation: Beyond eyewitness testimony. *Memory, 14* (1), 27–36. DOI: http://dx.doi.org/10.1080/09658210444000476.

Clifford, B. R. (1980). Voice identification by human listeners. *Law and Human Behavior, 4,* 373–394. DOI: http://dx.doi.org/10.1007/BF01040628.

Deffenbacher, K. A., Cross, J. F., Handkins, R. E., Chance, J. E., Goldstein, A. G., Hammersley, R., & **Read, J. D.** (1989). Relevance of voice identification research to criteria for evaluating reliability of an identification. *The Journal of Psychology, 123,* 109–119. DOI: http://dx.doi.org/10.1080/00223980.1989.10542967.

Frost, P. (2000). The quality of false memory over time: Is memory for misinformation "remembered" or "known"? *Psychonomic Bulletin & Review, 7,* 531–536. DOI: http://dx.doi.org/10.3758/BF03214367.

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.

Handkins, R. E., & **Cross, J. F.** (1985). Can a voice lineup be too fair? Paper presented at the annual meeting of the Midwestern Psychology Association, Chicago, IL.

Hollien, H. (1990). *The Acoustics of Crime-The New Science of Forensic Phonetics.* New York: Plenum.

Howard, I. P. (1982). *Human visual orientation.* New York: Wiley.

Johnson, M. K., Hastroudi, S., & **Lindsay, D. S.** (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28. DOI: http://dx.doi.org/10.1037/0033-2909.114.1.3.

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), 327–326. DOI: http://dx.doi.org/10.1002/acp.974.

Kirk, R. E. (1995). *Experimental Design.* (3rd Ed.) Belmont, CA: Brooks/Cole.

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

Linville, S. E. (1996). The sound of senescence. *Journal of Voice, 10,* 190–200. DOI: http://dx.doi.org/10.1016/S0892-1997(96)80046-4.

Loftus, E. F. (1981). Mentalmorphosis: Alterations in memory produced by the mental bonding of new information to old. In J. Long & A. Baddeley (Eds.), *Attention and performance IX* (pp. 417–434). Hillsdale, NJ: Erlbaum.

Loftus, E. F. (1992). When a lie becomes memory's truth. *Current Directions in Psychological Science, 1,* 121–123. DOI: http://dx.doi.org/10.1111/1467-8721.ep10769035.

Loftus, E. F., & **Greene, E.** (1980). Warning: Even memory for faces may be contagious. *Law and Human Behavior, 4,* 323–334. DOI: http://dx.doi.org/10.1007/BF01040624.

Loftus, E. F., Miller, D. G., & **Burns, H. J.** (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 19–31. DOI: http://dx.doi.org/10.1037/0278-7393.4.1.19.

Luna, K., & **Migueles, M.** (2009). Acceptance and confidence of central and peripheral misinformation. *The Spanish Journal of Psychology, 12*(2), 405–413. DOI: http://dx.doi.org/10.1017/S1138741600001797.

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: http://dx.doi.org/10.1080/00221309.1988.9711085.

McCloskey, M., & **Zaragoza, M.** (1985). Misleading postevent information and memory for events: Arguments and evidence against memory impairment hypotheses. *Journal of Experimental Psychology: General, 114,* 1–16. DOI: 10.1037/0096-3445.114.1.

Meissner, C. A., Brigham, J. C., & **Kelley, C. M.** (2001). The influence of retrieval processes in verbal overshadowing. *Memory & Cognition, 29*(1), 176–186. DOI: http://dx.doi.org/10.3758/BF03195751.

Mullennix, J. W., Stern, S. E., Grounds, B., Kalas, R., Flaherty, M., Kowalok, S., May, E., & **Tessmer, B.** (2010). Earwitness memory: distortions for voice pitch and speaking rate. *Applied Cognitive Psychology, 24*(4), 513–526. DOI: http://dx.doi.org/10.1002/acp.1566.

Neath, I., & **Surprenant, A. M.** (2003). *Human memory: An introduction to research, data, and theory* (2nd Ed.). Belmont, CA: Wadsworth.

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

Orchard, T., & A. Yarmey, A. D. (1995). The effects of whispers, voice sample duration and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 31,* 249–260. DOI: http://dx.doi.org/10.1002/acp.2350090306.

Pezdek, K., & Lam, S. (2007). What research paradigms have cognitive psychologists used to study "False memory", and what are the implications of these choices? *Consciousness and Cognition, 16*, 2–17. DOI: http://dx.doi.org/10.1016/j.concog.2005.06.006.

Philippon, A. C., Cherryman, J., Bull, R., & Vrij, A. (2007). Earwitness identification performance: The effect of language, target, deliberate strategies, and indirect measures. *Applied Cognitive Psychology, 21*(4), 417–556. DOI: http://dx.doi.org/10.1002/acp.1296.

Pigott, M. A., Brigham, J. C., & Bothwell, R. K. (1990). A field study on the relationship between quality of eyewitnesses' descriptions and identification accuracy. *Journal of Police Science and Administration, 17,* 84–88.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., & the R Development Core Team. (2012). *nlme: Linear and nonlinear mixed effects models.* R package version 3.1–104.

Pisoni, D. B. (1993). Long-term memory in speech perception: Some new findings on talker variability, speaking rate, and perceptual learning. *Speech Communication, 13,* 109–125. DOI: http://dx.doi.org/10.1016/0167-6393(93)90063-Q.

Posner, M. I., Nissen, M. J., & Klein, R. M. (1976). Visual dominance: An information-processing account of its origins and significance. *Psychological Review, 83,* 157–171. DOI: http://dx.doi.org/10.1037/0033-295X.83.2.157.

R Core Team. (2012). *R: A language and environment for statistical computing.* Vienna: R Foundation for Statistical Computing. URL http://www.R-project.org/.

Read, D., & Craik, F. I. M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied, 1,* 6–18. DOI: http://dx.doi.org/10.1037/1076-898X.1.1.6.

Saslove, H., & Yarmey, A. D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65,* 111–116. DOI: http://dx.doi.org/10.1037/0021-9010.65.1.111.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology, 22,* 36–71. DOI: http://dx.doi.org/10.1016/0010-0285(90)90003-M.

Searcy, J. H., Bartlett, J. C., & Memon, A. (2000). Influence of post-event narratives, lineup conditions and individuals differences on false identification by young and older eyewitnesses. *Legal and Criminological Psychology, 5,* 219–235. DOI: http://dx.doi.org/10.1348/135532500168100.

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

Tomes, J. L., & Katz, A. N. (2000). Confidence –accuracy relations for real and suggested events. *Memory, 8,* 273–283. DOI: http://dx.doi.org/10.1080/09658210050117708.

Urbaniak, G. C., & Plous, S. (2011). Research Randomizer (Version 3.0) [Computer software]. Retrieved on June 26, 2012, from http://www.randomizer.org/

Van Wallendael, L. R., Surface, A., Parsons, H. D., & Brown, M. (1994). "Earwitness" voice recognition: Factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology, 8,* 661–673. DOI: http://dx.doi.org/10.1002/acp.2350080705.

Weingardt, K. R., Leonesio, R. J., & Loftus, E. F. (1994). Viewing eyewitness research from a metacognitive perspective. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 157–184). Cambridge, MA: MIT Press.

Wilding, J., Cook, S., & Davis, J. (2000). Sound familiar? *Psychologist, 13,* 558–562.

Yarmey, A. D. (1991a). Description of distinctive and nondistinctive voices over time. *Journal of the Forensic Science Society, 31,* 421–428. DOI: http://dx.doi.org/10.1016/S0015-7368(91)73183-6.

Yarmey, A. D. (1991b). Voice identification over the telephone. *Journal of Applied Social Psychology, 21,* 1868–1876. DOI: http://dx.doi.org/10.1111/j.1559-1816.1991.tb00510.x.

Yarmey, A. D. (1994). Earwitness evidence: Memory for a perpetrator's voice. In D. F. Ross, J. D. Read & M. D. Toglia (Eds.), *Adult eyewitness testimony: Current trends and developments* (pp. 101–124). New York: Cambridge University Press. DOI: http://dx.doi.org/10.1017/CBO9780511759192.006.

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

Yarmey, A. D. (2007). The psychology of speaker identification and earwitness memory. In R. C. L. Lindsay, D. F. Ross, J. D. Read & M. P. Toglia (Eds.), *The handbook of eyewitness psychology, vol II: Memory for people* (pp. 101–136). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Zaragoza, M. S., & Koshmider, J. W. (1989). Misled subjects may know more than their performance implies. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 246–255. DOI: http://dx.doi.org/10.1037/0278-7393.15.2.246.

Zaragoza, M. S., & Lane, S. M. (1994). Source misattributions and the suggestibility of eyewitness memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 934–945. DOI: http://dx.doi.org/10.1037/0278-7393.20.4.934.