May I Speak Freely? The Difficulty in Vocal Identity Processing across Free and Scripted Speech.

Sarah V Stevenage[1*] Rebecca Tomlin[1], Greg J Neil[2] & Ashley E Symons[1]

[1]School of Psychology, University of Southampton

[2]School of Sport, Health and Social Sciences, Solent University

CORRESPONDENCE

*Please address all correspondence to Professor Sarah Stevenage at:

School of Psychology, University of Southampton, Highfield, Southampton, Hampshire, SO17 1BJ. Email: svs1@soton.ac.uk; Fax: 02380 594597; Tel: 02380 592973

Orcid ID: Sarah Stevenage: 0000-0003-4155-2939; Greg Neil: 0000-0003-1360-5490; Ashley Symons: 0000-0001-5980-6752

ACKNOWLEDGEMENTS

This work was supported by funding from the EPSRC (grant number EP/R030839/1) awarded to the first author. The authors are grateful to Dr Richard Guest and Iulia Motoc for helpful comments on the ideas described within this paper.

Abstract

In the fields of face recognition and voice recognition, a growing literature now suggests that the ability to recognise an individual despite changes from one instance to the next is a considerable challenge. The present paper reports on one experiment in the voice domain designed to determine whether a change in the mere style of speech may result in a measurable difficulty when trying to discriminate between speakers. Participants completed a speaker discrimination task to pairs of speech clips which represented either free speech or scripted speech segments. The results suggested that speaker discrimination was significantly better when the style of speech did not change compared to when it did change, and was significantly better from scripted than from free speech segments. These results support the emergent body of evidence suggesting that within-identity variability is a challenge, and the forensic implications of such a mild change in speech style are discussed.

May I Speak Freely? The Difficulty in Vocal Identity Processing across Free and Scripted Speech.

A growing literature now recognises that the task of human identification involves both the ability to *tell apart* similar instances belonging to different individuals, and the ability to *tell together* different instances belonging to the same individual (Burton, 2013). In the context of the latter task, it is noted that the ability to recognise an individual, or to match one instance to another, is made considerably more difficult when those instances reflect a natural level of variability. The result is that a perceiver may often mistake two different instances of the same individual as belonging to different people. In the face domain, this is exemplified by the difficulty in identifying an individual across changes in pose, lighting, camera angle, or even in the camera used to take the photograph (Young & Burton, 2017). This difficulty is perhaps most starkly demonstrated in a face sorting task (Jenkins, White, van Montfort & Burton, 2011) where participants are asked to sort naturally varying (or ambient) images of two identities into identity piles. If unfamiliar with the identities, participants can struggle with the sorting task, and performance will often show a failure to tell instances of the same person together (Andrews, Jenkins, Cursiter & Burton, 2015; Jenkins *et al*., 2011, Zhou & Mondloch, 2016).

The same difficulty in coping with within-identity variability also arises in the voice domain. Indeed, performance in an unfamiliar voice sorting task suggests a difficulty both when telling different instances of the same speaker together (Lavan, Burston & Garrido, 2018; Lavan Burston, Ladwa, Merriman, Knight & McGettigan, 2019; see also Lavan Merriman, Ladwa, Burston, Knight & McGettigan, 2019b), and when telling similar instances of two different speakers apart (Stevenage, Symons, Fletcher & Coen, 2019).

The vocal changes that cause difficulty for the listener are wide and varied. For instance, performance is compromised by purposeful changes such as whispering, as demonstrated in a voice recognition task (Yarmey, Yarmey, Yarmey & Parliament, 2001), a 2-day delayed lineup task (Orchard & Yarmey, 1995), and a same/different matching task (Bartle & Dellwo, 2015). Similarly, speaker discrimination in a same/different task is compromised by vocal disguises such as the adoption of an old-age voice, a hoarse voice, a hyper-nasal voice, a slow voice, or a freely selected disguise (Reich & Duke, 1979). Even when voices are familiar, a purposeful change to the voice, such as the adoption of a falsetto voice, can impair speaker recognition substantially (Wagner & Köster, 1999).

In addition to these effects of vocal disguise, a marked impairment emerges in a voice matching task when the speaker is talking in an unfamiliar language (Goggin, Thompson, Strube & Simental, 1991; Perrachioni & Wong, 2007; Winters, Levi & Pisoni, 2008). Additionally, there is difficulty when matching singing and speaking clips from the same person (Peynircioğlu, Rabinovitz & Repice, 2017, see also Bartholomeus, 1974; Zatorre & Baum, 2012). Studies have also shown an impact of a change in the type of vocalisation (vowels versus laughter) (Lavan, Scott & McGettigan, 2016), and even of a change within vocalisations such as when shifting from volitional to spontaneous laughter (Lavan *et al.,* 2016). Finally, more commonplace changes within a vocalisation which cause the listener difficulty include a change in emotional tone (Read & Craik, 1995) even when clips are presented just minutes apart (Saslove & Yarmey, 1980).

As a whole, these results suggest that different types of voice clips may vary either in the richness of the identity-relevant vocal information contained, or in the listeners' capability to perceive that information. In terms of clip richness, and in the context of the laughter work, Lavan, Short, Wilding and McGettigan (2018) suggested that spontaneous laughter may be phylogenetically older than volitional laughter, and consequently may lack the more evolved speech-based sounds that help us to distinguish between identities. In contrast, and in terms of listener capability, Belin, Fecteau and Bédard (2004) suggested that we extract three main elements from the voice: speech, affect and identity. However, Stevenage and Neil (2014) noted that these three elements may not take equal priority, with the analysis of speech and affect taking precedence, and thus distracting from, the analysis of vocal identity. Nevertheless, both a clip richness explanation and a perceiver capability explanation can account for the difficulty in processing identity across what Vernon (1952, cited in Bruce, 1994, p8) describes as 'the extent of permissible and possible permutations within a single identity'. This difficulty becomes important in a forensic context in which an earwitness, juror or lay listener may be asked to compare one speech clip to another of a different vocal style in order to determine whether the two represent a match. As such, there is value in understanding both competency and confidence that a listener has in vocal matching despite vocal change.

The purpose of the present paper is to extend the existing evidence base regarding the difficulty associated with voice processing across vocal change. However, here we adopt an everyday manipulation within speech-based clips and thus we remove the issue of non-speech-based sounds raised in Lavan *et al.*'s (2018) laughter work. Instead, we explore the impact of a change from spontaneous or conversational speech (referred to here as 'free' speech) to read speech (referred to here as 'scripted' speech).

In this regard, a number of studies detail the fact that free and scripted speech are readily distinguishable from one another, differing in segmental phonetic characteristics and prosodic characteristics (Baker & Hazan, 2010; 2011), particularly articulation rate (Dellwo, Leemann & Kolly, 2015). More specifically, scripted speech was noted to be more rapid and free from hesitation compared to free speech when story telling (Levin, Schaffer & Snow, 1982), especially in the latter part of an utterance (Remez, Rubin & Nygaard, 1986). In terms of the impact of speech style when processing vocal identity, there is value in considering the results of Smith, Baguley, Robson, Dunn and Stacey (2019) who examined speaker discrimination as speech style varied from scripted to free speech. Across two experiments, their results revealed significantly worse performance, and significantly lower confidence, when speech style changed from scripted to free speech, compared to a baseline condition in which both clips features scripted speech. Given previous demonstrations of processing difficulty resulting from vocal change, these results are interesting and important from a forensic perspective, but should perhaps not be surprising. What is missing within their study, however, is an exploration of which speech style (free or scripted) is the better style for the listener when trying process vocal identity.

The current study addressed this question by presenting a speaker discrimination task in which participants reported on whether two consecutively presented voice clips belonged to the same speaker or not. Four conditions were employed in which speech style was held constant across the two clips (scripted/scripted; free/free) and in which speech style was varied across the two clips (scripted/free; free/scripted). This design thus allowed examination of the impact of speech style *per se* alongside the impact of change in speech style. Confidence as well as accuracy was recorded in order to examine beliefs in performance as well as performance itself. Based on the above literature, it was anticipated that a change in speech style between first and second clip would impair performance in the speaker discrimination task. However, it was unclear whether scripted or free speech would produce the better performance. The present study is intended to extend the evidence base regarding the impact of vocal change. However, the results may also help to guide police and court process when considering the importance of speech style, and change in speech style, in the speaker discrimination task.

Method

Design

Participants completed a speaker discrimination task with unfamiliar voices during which speech style (free, scripted) was varied at both first and second utterance within ‘same’ and ‘different’ trials. This resulted in four experimental blocks, with two blocks representing congruent listening conditions (free speech at first and second utterance (FF); scripted speech at first and second utterance (SS)), and two blocks representing incongruent or changing listening conditions (free speech followed by scripted speech (FS); and vice versa (SF)). Accuracy of speaker discrimination was recorded together with self-rated confidence for each response.

Participants

Forty participants (30 females) took part as volunteers or in return for course credit. This number was based on an opportunity sample but exceeded the number required to obtain 80% power given a medium effect size ($\eta^2_p = .06$) and an alpha level of 0.05 (n = 23). Ages ranged from 18 to 26 years (*M* = 20.95, *SD* = 1.66), and all participants had self-reported normal hearing. Participants were unfamiliar with all speakers, as confirmed verbally at the end of the study.

Materials

The voice clips consisted of 30 unfamiliar target voices (16 females) together with 30 sex-matched foil voices. All speakers were Caucasian, and spoke English as a first language, with a southern British accent, and no audible speech impediments. Foils were paired with targets on the basis of similarity ratings. These were provided by a panel of six independent judges, who provided a similarity rating to a single scripted phrase using a 7 point scale (where 7 = very similar). All foils had a mean similarity of 4.5 or more (out of 7) and the average similarity between foil and target across all items was 5.83 out of 7 (*SD* = .77). This ensured an appropriate level of difficulty during 'different' trials.

The 30 targets were repeated across all 4 blocks of the experimental design (for comparability) necessitating 4 different free speech clips and 4 different scripted clips from each target. In contrast, foil speakers provided only 2 free speech clips and 2 scripted clips for use as the second utterance in 'different' trials. Free speech clips were obtained by asking speakers to talk on four set topics. This ensured some uniformity of content (and thus of interest) whilst not constraining vocal style or wording. Scripted clips were obtained by asking the speakers to utter four specific sentences. These were drawn from the FRL2011 database and were designed to ensure rich phonemic variability (see Appendix).

All clips were recorded within a sound-proofed studio using Audacity 3.1, with a sampling rate of 44.1 kHz, and 16 bit resolution. Voice capture was achieved using a Sennheiser EW100 wireless lapel-microphone positioned approximately 20cm from the speaker's mouth. This minimised the capture of distorted sound associated with plosive speech, but careful positioning also avoided any muffling due to clothing.

Each scripted speech clip consisted of a complete individual sentence lasting 4 seconds which did not require editing via concatenation or trimming. In contrast, free speech clips were edited within Audacity to provide 4 second segments of continuous speech extracted from a longer sample. As a result, the offset of the free speech clip did not always coincide with the end of a phrase (see also Schweinberger *et al*., 1997). Despite the standardised resultant clip length, the mean number of words (13.81, *SD* = 3.74) and vowel sounds (18.17, *SD* = 4.43) was higher for free speech clips than for scripted clips (words: *M* = 12.25, *SD* = 1.5; vowel sounds: *M* = 15.75, *SD* = 2.99). Indeed, analysis suggested that whilst the clips were matched for length, there were nevertheless more words ($t_{(119)}$ = 4.56, *p* < .001) and more vowel sounds ($t_{(119)} = 5.98$, $p < .001$) in the free speech clips than in the scripted clips, a point that is returned to in the Discussion.

From these stimuli, 15 ‘same’ trials and 15 ‘different’ trials were constructed to generate 4 blocks of 30 trials, with blocks differing only in whether free or scripted speech clips were used at first and second utterance. This resulted in four blocks consisting of free/free clips, scripted/scripted clips, free/scripted clips and scripted/free clips. ‘Same’ trials consisted of a target voice followed by a different clip from the same target voice. ‘Different’ trials consisted of a target voice followed by a different clip from the sex-matched foil voice. Individual voice clips were not repeated across each of the 4 blocks minimising the opportunity for learning. In addition, the identity of targets in ‘same’ trials was counterbalanced across blocks and across participants in order to minimise item effects.

Stimuli were presented, and data were recorded, via SuperLab 4.5.4 running on an HP laptop, with a 15.6” computer monitor set at a screen resolution of 1366 x 768 pixels. Sound was played via computer speakers, pre-set to a comfortable but adjustable level.

Procedure

Following ethical approval by the local School of Psychology ethics panel, and the provision of informed consent, participants were tested individually and completed both practice and test trials. The 20 practice trials involved presentation of the word ‘SAME’ or ‘DIFFERENT’ on screen. Participants were asked to press ‘S’ for ‘same’ and ‘D’ for ‘different’ as a way of mapping responses to response keys, and feedback was provided.

Following this, four blocks of 30 experimental trials were presented, with order of blocks counterbalanced across participants. Blocks were separated by self-paced breaks, and differed only in the style of voice clip presented in the first and second utterance. Regardless of speech style, all trials took the same format consisting of a ‘please listen…’ prompt for

250 msecs, followed by the first voice clip for 4 seconds, an inter-stimulus interval of 4 seconds, and finally, a second voice clip for 4 seconds. Participants indicated whether the speaker in the second clip was the same ('S') or different ('D') to that in the first clip and no feedback was provided. Finally, participants rated their confidence in their response on each trial using a 7 point scale (where 1 = 'not at all confident', and 7 = 'very confident indeed').

The entire experiment lasted 25-30 minutes, after which participants were thanked and debriefed.

Results

Accuracy was recorded in each condition of the speaker discrimination task, and from this, measures of sensitivity of discrimination (*d'*) and response bias (*C*) were derived (Green & Swets, 1966). These measures were used for the purposes of analysis. In addition, confidence was also explored taking 'same' and 'different' trials separately given the dissociation noted in previous work (Megreya & Burton, 2007; Ritchie & Burton, 2017). The data are summarised in Table 1 and Figure 1. Preliminary scrutiny revealed two outliers with poor discrimination in the SF condition only (with d' falling outside 1.5 x IQR). The data for these two participants were omitted from all analyses to avoid floor effects, leaving the data from 38 participants. In all analyses reported below, alpha is set to 0.05 unless otherwise stated, and exact *p* values are reported wherever possible.

(Please insert Table 1 and Figure 1 about here)

Sensitivity of Discrimination

A series of one-sample *t*-tests confirmed that performance was significantly above chance in all four conditions (all $ts_{(37)} > 11.94$, $p < .001$). Performance was evaluated using a 2 x 2 repeated-measures Analysis of Variance (ANOVA) in order to determine the impact of speech style during first utterance (free, scripted) and during second utterance (free, scripted).

The analysis revealed both a main effect of speech style in first utterance ($F_{(1, 37)} = 7.57$, $p = .009$, $\eta^2_p = .17$) and in second utterance ($F_{(1, 37)} = 7.77$, $p = .008$, $\eta^2_p = .18$), as well as a large and significant interaction ($F_{(1, 37)} = 94.38$, $p < .001$, $\eta^2_p = .72$).

Tests of simple main effects revealed this interaction to be due to a congruency advantage, with performance being better when the two utterances matched in speech style (FF vs FS: $F_{(1, 37)} = 43.85$, $p < .001$, $\eta^2_p = .54$; SS vs SF: $F_{(1, 37)} = 63.74$, $p < .001$, $\eta^2_p = .63$). Of much more interest though, the tests of simple main effects also noted significantly better performance when congruent trials involved scripted speech rather than free speech (SS > FF: $F_{(1, 37)} = 11.89$, $p = .001$, $\eta^2_p = .24$). There was no difference in performance across the two incongruent conditions (SF = FS: $F_{(1, 37)} < 1$, $p = .92$, $\eta^2_p < .01$).

Response Bias

One-sample *t*-tests revealed a significant response bias in all conditions ($ts_{(37)} < -2.73$, $p = .01$) indicating a liberal bias to say ‘same’. As a result, performance was noted as being better in ‘same’ trials than in ‘different’ trials overall. Analysis using a 2 x 2 repeated-measures ANOVA revealed neither a main effect of speech style during first utterance ($F_{(1, 37)} = .19$, $p = .67$, $\eta^2_p = .005$) nor during second utterance ($F_{(1, 37)} = .04$, $p = .85$, $\eta^2_p = .001$). However, as above, a significant interaction did emerge ($F_{(1, 37)} = 5.59$, $p = .023$, $\eta^2_p = .13$).

Tests of simple main effects again indicated this to be due to a congruency effect such that the bias to say ‘same’ showed a tendency to be stronger in congruent than incongruent trials (FF vs FS: $F_{(1, 37)} = 3.15$, $p = .084$, $\eta^2_p = .08$; SS vs SF: $F_{(1, 37)} = 4.33$, $p < .05$, $\eta^2_p = .10$). As with the discrimination data, there was no difference in the response bias across the two incongruent conditions (FS = SF: $F_{(1, 37)} = .19$, $p = .67$, $\eta^2_p < .01$). However, in contrast to the discrimination data, there was no difference in response bias between the two congruent conditions either (FF = SS: $F_{(1, 37)} = .05$, $p = .82$, $\eta^2_p < .01$). This suggested that a perceptual component rather than a decisional component was driving the previous discrimination advantage of scripted over free speech clips in the congruent listening condition.

Confidence

Average confidence ratings were obtained for ‘same’ and ‘different’ trials in each of the four experimental conditions (see Table 1 and Figure 1). These were analysed by means of a 2 x 2 x 2 repeated-measures ANOVA, examining the effects of trial type (‘same’, ‘different’), first utterance type (free, scripted) and second utterance type (free, scripted). The results indicated a main effect of trial type ($F_{(1, 37)} = 49.03$, $p < .001$, $\eta^2_p = .57$), with greater confidence on ‘same’ trials ($M = 5.84$) than ‘different’ trials ($M = 5.22$). There was no impact on confidence of either speech style at first utterance ($F_{(1, 37)} < 1$, $p = .78$, $\eta^2_p = .002$) or of speech style at second utterance ($F_{(1, 37)} < 1$, $p = .83$, $\eta^2_p = .001$). However, a small and unanticipated interaction did emerge between trial type and style at first utterance ($F_{(1, 37)} = 4.17$, $p = .048$, $\eta^2_p = .101$) which was not readily explained by post-hoc contrasts.

(Please insert Figure 2 about here)

Of more importance was the interaction between speech style at first utterance and at second utterance ($F_{(1, 37)} = 19.80$, $p < .001$, $\eta^2_p = .349$). Post-hoc examination suggested that this reflected the previously seen congruency effect, with confidence being higher when the two utterances matched in speech style (FF > FS: $F_{(1, 37)} = 11.06$, $p = .002$, $\eta^2_p = .23$; SS > SF: $F_{(1, 37)} = 11.40$, $p = .002$, $\eta^2_p = .24$). Interestingly, and in contrast to the analysis of sensitivity of discrimination, the post-hoc contrasts did not suggest greater confidence when congruent trials involved scripted speech rather than free speech (FF v SS: $F_{(1, 37)} = .00$, $p = .937$, $\eta^2_p < .001$). As before, no difference in confidence was evident in the two incongruent conditions involving a change in speech style either (FS v SF: $F_{(1, 37)} = .12$, p = .73, $\eta^2_p = .003$). No other main effects or interactions reached significance (all $Fs(1, 37) < 2.32, p > .14, \eta^2_p <$ .06).

Confidence-Accuracy Relationship

In order to examine the relationship between confidence and accuracy, the calibration method adopted by Smith *et al*. (2019) was adopted given its greater sensitivity when revealing an association compared to point-biserial correlations (Brewer & Wells, 2006). Calibration curves were generated for the population as a whole, collapsing across individual items. However, care was taken to isolate the confidence-accuracy (CA) relationship for 'same' and 'different' trials in each of the four experimental conditions given the differences noted above (see Figure 2). Within these calibration curves, perfect calibration is indicated by the dotted diagonal line and indicates that high confidence accompanies high accuracy, whilst low confidence accompanies low accuracy. Points falling below the diagonal represent over-confidence, whilst points falling above the diagonal represent under-confidence. Based on visual inspection, calibration appeared better for 'same' trials than for 'different' trials but suggested some under-confidence at lower confidence levels particularly during 'different' trials.

(Please insert Figure 2 about here)

A series of linear regressions was used to determine the statistical relationship between confidence and accuracy, using confidence as the dependent variable, and accuracy as the predictor. Alpha was adjusted to 0.05/8 given the number of regressions conducted. This revealed a significant association between accuracy and confidence for 'same' trials in all conditions (FF: $\beta = .935, p = .001$; FS: $\beta = .933, p = .002$, SF: $\beta = .930, p = .002$; SS: $\beta = .944, p = .001$) and examination of 95% confidence intervals for B indicated no discernible differences in the strength of this relationship across conditions. This suggested that participants were well-calibrated in terms of their accuracy and confidence on ‘same’ trials regardless of speech style. In contrast, there was no association between accuracy and confidence for ‘different’ trials in any of the conditions (all $\beta < .83$, $p > .021$).

Discussion

The present study was conducted in order to determine the listener’s ability to complete a speaker discrimination task despite changes in speaking style. Here, the change to speaking style was very subtle and involved a change from free speech to scripted speech or vice versa. Based on previous literature, it was anticipated that speaker discrimination would be better when there was no change in speech style compared to when there was a change. However, it was unclear whether scripted speech would be so constrained as to strip out important identity-based vocal cues, or whether it would instead be more stable and thus more beneficial when comparing one utterance to another. Consequently, it was unclear which speech style would yield the better performance on the speaker discrimination task.

The results confirmed expectations in that performance was indeed better, and confidence was higher, when there was no change in speech style between first and second utterance. Furthermore, performance was significantly better from scripted speech than from free speech. This was revealed when considering sensitivity of discrimination, and this benefit was not accounted for by response bias and was not reflected in self-reported confidence. The present analysis also revealed a strong association between confidence and accuracy overall. This was especially the case on ‘same’ trials. However, whilst there was no association between accuracy and confidence on ‘different’ trials, accuracy was around 80% when confidence was high on these trials. Consequently, whilst it is not possible in the real world to know whether a suspect is the true perpetrator, the present evidence may suggest broad trust in the confident witness. These results are perhaps surprising given the literature suggesting that confidence and accuracy are not strongly related in the earwitness domain (see Olson, Juslin & Wiman, 1998, and Smith & Baguley, 2014, p61 for reviews). Nevertheless, and as noted by Smith *et al*., (2019), the use of a calibration method rather than a point-biserial method here may have enabled the demonstration of a stronger confidence-accuracy relationship than in previous literature.

The strength of the present demonstration was that all clips were speech-based. Thus, any performance differences could not readily be attributed to one clip type containing more speech-based cues than the other (*cf* Lavan *et al*., 2018). As such, the present results confirmed previous demonstrations of a performance benefit in a speaker discrimination task when speech style was kept constant, but additionally demonstrated optimal performance when scripted rather than free speech segments were used.

Task Difficulty, Over-Estimation, Reference Stimuli and Quantity Issues

Alongside this main finding, there are several other points worthy of note. First, when considering performance overall, the task appeared to be somewhat more difficult compared to that of Smith *et al*. (2019), with an overall accuracy rate of 76.25% relative to their accuracy rate of 87%. It is possible that the greater difficulty within the current task arose due to the care when matching targets to foils within the ‘different’ trials. This was a purposeful decision as a way to avoid ceiling level performance which may have masked the effects of speech style. In support of this account, participants within the current study exhibited a response bias to say ‘same’ suggesting an overall difficulty at telling speakers apart from one another. Consequently, it was the ‘different’ trials which presented the greatest challenge both in terms of accuracy, self-rated confidence, and CA calibration.

Second, the demonstration here of better voice matching performance in the congruent listening conditions when using scripted rather than free speech clips carries implications for the interpretation of much of the voice matching literature. With most of this literature utilising scripted speech clips (in an effort to control factors such as semantic content, word count, vowel count and phonemic richness) the resultant findings may overestimate the performance level that listeners may attain when using rather more naturalistic vocal styles. This does not invalidate the findings from the existing body of work. However, it does suggest caution in generalising those results to the processing of more everyday speech styles. In this sense, whilst experimental control has been valuable in the early phase of voice identity research, the current findings remind us of the limitations that may arise when using highly controlled stimuli.

Third, it was interesting to see that performance was equivalent in the two conditions involving a change in speech style (free-scripted; scripted-free). This was the case when considering d’, response bias, confidence and the CA relationship. This implied that it did not matter which speech style was used as the first (or reference) stimulus against which the second stimulus was compared. As such, there was no evidence that either speech style would result in an ‘impoverished template’ in the same way as a noisy first stimulus may (see Smith *et al.*, 2019, Experiment 2). By extension, one may conclude that it is unlikely that the two speech styles differed fundamentally in the value of the voice identity characteristics that they contained.

Fourth, it was revealing to note that performance was better in congruent listening conditions involving scripted rather than free speech clips despite the fact that the scripted clips contained fewer words, and fewer vowel sounds than the free speech clips. Thus, despite being matched for overall length, scripted speakers spoke slower and produced fewer vocal cues for the listener to use, and yet produced the better performance. This lower word count and vowel sound count in scripted over free clips contrasts with the results reported by Levin, Schaffer and Snow (1982) who suggested that speakers demonstrated more hesitancy and thus spoke more slowly during free speech. Levin *et al.* suggested that their results reflected their use of a free speech task involving creative story-telling in which speaker hesitancy was perhaps not surprising. Their 10-second free speech segments reflected this hesitancy. In contrast, the present free speech task involved speakers talking about personal accounts, views, or preferences which may have been less demanding to generate. Our 4-second free speech segments were extracted when the speakers were mid-utterance, and thus they captured less in the way of hesitancy.

The Difference between Scripted and Free Speech

The above discussion is interesting as it starts to consider the differences that may exist between scripted and free speech. These differences appear to guide not only the ability to distinguish one speech style from another (Baker & Hazan, 2010; 2011; Dellwo, Leemann & Kolly, 2015) but the ability to discriminate between speakers as shown by Smith *et al.* (2019) and as confirmed here. At an intuitive level, it is possible that the scripted speech clips bore a greater resemblance to one another because of a greater level of consistency in delivery. In this regard, one may conclude that quality of speech is more important than quantity of speech, and that consistency of quality is cardinal in supporting accurate voice matching decisions. Equally, it is possible that the faster speech rate during free speech prevented the listener from picking up on all vocal identity-relevant cues. Further work with rather more stimuli is, however, required to separate out the consistency explanation from the speech rate explanation.

One other difference exists between the free and scripted speech clips and may be worth noting. This relates to the novelty of their semantic content. More specifically, whilst the free speech clips were constrained to some degree by the fact that they represented free answers to set questions, those answers will have varied across speakers. In contrast, the scripted speech clips were constant in their content across speakers. As a consequence, the free speech clips may have been more interesting to listen to than the scripted clips.

This variation in speech content may have had an important bearing on the listener's task because constancy of speech content during scripted clips could allow the listener to switch off from the task of speech processing and concentrate on vocal identity processing as per the task demands. In contrast, continual novelty of speech content during free speech clips may have meant that the listener remained focussed on speech processing as their *unstoppable priority* (Goggin *et al*., 1991, p457), to the detriment of vocal identity processing. This variation of speech content was perhaps unavoidable within the current study given the nature of free speech. Indeed, it could only be factored out of a future experimental design by presenting listeners with a different scripted speech clip (thus balancing the semantic novelty across free and scripted clips) on every trial. The considerable database of voice clips required to serve such an experiment may preclude its conduct. Nevertheless, this discussion does serve to highlight the various factors associated with free and scripted speech, and an explanation based on these semantic factors must sit alongside those associated with speech style as discussed above.

Conclusions and Applied Implications

The current study explored the capacity of human listeners to judge whether two voice clips came from the same speaker or from two different speakers as speech style was subtly varied. Performance was impaired when trying to match two utterances that varied in speech style. In this sense, congruency of speech style supported better matching performance. Of more interest, the results indicated that performance was better when listening to two scripted speech clips as opposed to two free speech clips.

At an applied level, the present results may hold forensic value by informing practitioners as to the optimal construction of voice matching tasks. This becomes important given the emergence of earwitness testimony as a factor in Court cases. Recent cases in England and Wales are reviewed by Smith *et al*. (2019, p272-273), but worth highlighting are the cases of R v Shannon Tamiz and others (2010) in which it is noted that translators may provide evidence to attest to a judgement that a speaker in one recording is the same as a speaker in another recording; and R v Kapikanya (2015), which depended on a jury judgement of whether the prosecution had demonstrated that a recorded voice matched the defendant. In other jurisdictions, jurors may be invited to engage in their own matching judgement when presented with two voice samples. As such, guidance may be valuable regarding best practice for forensic voice matching tasks.

Based on the current results, and all other factors being equal, it is recommended that the two utterances within a voice matching task should be consistent in speech style wherever possible. The greater consistency of speech patterns would facilitate a focus on the important vocal identity characteristics. In fact, according to the current data, two scripted style speech clips would provide the optimal conditions to enable the listener to say 'yes' to the perpetrator's voice, and to say 'no' to the voice of an innocent suspect. Of course, it may often be the case that a perpetrator's voice reveals a level of agitation or stress, limiting the generalisability of the present results within forensic settings. However, for those scenarios in which this is not the case, the current recommendations underline the utility of holding speech style constant in a matching scenario, and suggest further benefit when the voice is highly practiced or scripted in style.

At a more theoretical level, the present results extend the body of data suggesting considerable difficulty in processing identity across natural variation. Here, the focus was on vocal identity, and variation was provided by a subtle manipulation from free to scripted speech. One interesting observation is the fact that performance was affected at all by what is a relatively subtle manipulation in speech style. This suggests that the human listener is actually remarkably perceptive to changes associated with vocal style, whilst also underlining the fact that the listener does not tend to use this perceptiveness in the service of vocal identity processing.

Ethical approval:

All procedures performed in studies involving human participants were in accordance with the ethical standards of the main institution (ERGO: 8154) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A.M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68(10),* 2041-2050. doi: 10.1080/17470218.2014.1003949

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

Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavioural Research, 43:* 761-770. doi: 10.3758/s13428-011-0075-y

Bartholomeus, B. (1974). Dichotic singer and speaker recognition. *Bulletin of the Psychonomic Society, 2(4b)*, 407-408.

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

Baumann, O., & Belin, P. (2008). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research, 74*, 110-120. doi:10.1007/s00426-008-0185-z.

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

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

Bruce, V. (1994). Stability from variation: The case of face recognition. MD Vernon memorial lecture. *Quarterly Journal of Experimental Psychology, 47A*, 5-28.

Burton, A.M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology, 66(8)*, 1467-1485.

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

Green, D.M., & Swets J.A. (1966). *Signal Detection Theory and Psychophysics*. New York: Wiley. (ISBN 0-471-32420-5)

Goggin, J.P., Thompson, C.P., Strube, G., & Simental, L.R. (1991). The role of language familiarity in voice identification. *Memory and Cognition, 19*, 448-458.

Jenkins, R., White, D., van Montfort, X., & Burton, A.M. (2011). Variability in photos of the same face. *Cognition, 121(3)*, 313-323. doi: 10.1016/j.cognition.2011.08.001

Lavan, N., Burston, L., & Garrido, L. (2018). How many voices did you hear? Natural variability disrupts identity perception in unfamiliar listeners. *British Journal of Psychology, 110(3),* 576-593. doi: 10.1111/bjop.12348

Lavan, N., Burston, L.F.K., Ladwa, P., Merriman, S.E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology, 72(9), 2240-2248.* doi: 10.1177/1747021819836890

Lavan, N., Merriman, S.E., Ladwa, P., Burston, L.F.K., Knight, S., & McGettigan, C. (2019b). Please sort these voice recordings into two identities: Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology,* doi: 10.1111/bjop.12416

Lavan, N., Short, B., Wilding, A., & McGettigan, C. (2018). Impoverished encoding of speaker identity in spontaneous laughter. *Evolution and Human Behavior, 39*, 139-145. doi: 10.1016/j.evolhumbehav.2017.11.002

Lavan, N., Scott, S.K., & McGettigan, C. (2016). Impaired generalisation of speaker identity in the perception of familiar and unfamiliar voices. *Journal of Experimental Psychology: General, 145(12)*, 1604-1614. doi: 10.1037/xge0000223

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

Megreya & Burton, A.M. (2007). Hits and false positives in face matching: A familiarity-based dissociation. *Perception & Psychophysics, 69(7)*, 1175-1184.

Olson, 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

Orchard, T.L., & Yarmey, A.D. (1995). The effects of whispers, voice-sample duration, and voice distinctiveness on criminal speaker identification. *Applied Cognitive Psychology, 9(3),* 249-260. doi: 10.1002/acp.2340090306

Perrachione, T.K., & Wong, P.C.M. (2007). Learning to recognize speakers of a non-native language: Implications for the functional organization of human auditory cortex. *Neuropsychologia, 45,* 1899-1910. doi: 10.1016/j.neuropsychologia.2006.11.015

Peynircioğlu, Z.F., Rabinovitz, B.E., & Repice, J. (2017). Matching speaking to singing voices and the influence of content. *Journal of Voice, 31(2).* 256, e13-256.e17

R v Kapikanya. (2015). EWCA Crim 1507. *The Journal of Criminal Law, 2016, 80(1),* 5-16.

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

Read, D., and Craik, F.I.M. (1995). Earwitness identification: Some influences on voice recognition. *Journal of Experimental Psychology: Applied,* 1(1), pp. 6-18.

Reich & Duke, (1979). Effects of selected vocal disguises upon speaker identification by listening. *Journal of the Acoustical Society of America, 66(4).* 1023-1028.

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

Ritchie, K.L., & Burton, A.M. (2017). Learning faces from variability. *Quarterly Journal of Experimental Psychology, 70(5)*, 897-905. doi: 10.108017470218.2015.1136656

Saslove, H., & Yarmey, A.D. (1980). Long-term auditory memory: Speaker identification. *Journal of Applied Psychology, 65(1)*, 111-116.

Schweinberger, S.R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language and Hearing Research, 40(2)*, 453-463.

Smith, H.M.J., & Baguley, T.S. (2014). Unfamiliar Voice Identification: Effect of post-event information on accuracy and voice ratings. *Journal of European Psychology Students, 5(1),* 59-68. doi: 10.5334/jeps.bs

Smith, H.M.J., Baguley, T.S., Robson, J., Dunn, A.K., & Stacey, P.C. (2019). Forensic voice discrimination by lay listeners: The effect of speech type and background noise on performance. *Applied Cognitive Psychology, 33*, 272-287. doi: 10.1002/acp.3478

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

Stevenage, S., Neil, G., Parsons, B., & Humphreys, A. (2018). A sound effect: Exploration of the distinctiveness advantage in voice recognition. *Applied Cognitive Psychology*. doi: 10.1002/acp.3424

Stevenage, S.V., Symons, A.E., Fletcher, A., & Coen, C. (2019). Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task. *Quarterly Journal of Experimental Psychology.*

van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backwards voices. *Journal of Phonetics, 13*, 19-38.

Wagner, I., & Köster, O. (1999). Perceptual recognition of familiar voices using falsetto as a type of voice disguise. Proceedings of the XIVth International Congress of Phonetic Sciences, San Francisco, CA.

Williams, C.E., & Stevens, K.N. (1969). On determining the emotional state of pilots during flight: An exploratory study. *Aerospace Medicine, 40,* 1369-1372.

Winters, S.J., Levi, S.V., & Pisoni, D.B. (2008). Identification and discrimination of bilingual talkers across languages. *The Journal of the Acoustical Society of America, 12*, 4524-4538.

Yarmey, A.D., Yarmey, A.L., Yarmey, M.J., & Parliament, L. (2001). Commonsense beliefs and the identification of familiar voices. *Applied Cognitive Psychology, 15*, 283-299. doi: 10.1002/acp.702

Young, A. W., & Burton, A. M. (2017). Recognizing faces. *Current Directions in Psychological Science, 26*, 212–217. doi:10.1177/0963721416688114

Zatorre, R.J., & Baum, S.R. (2012). Musical melody and speech intonation: Singing a different tune? PLoS Biology, 10(7), e1001372.

Zhou, X., & Mondloch, C.J. (2016). Recognising ‘Bella Swan’ and ‘Hermione, Granger’: No own-race advantage in recognising photos of famous faces. *Perception, 45*, 1426-1429. doi: 10.1177.0301006616662046

Table 1:

Mean sensitivity of discrimination (d'), and response bias (C), together with proportion of accurate decisions and confidence on 'same' and 'different' trials across the four experimental conditions.

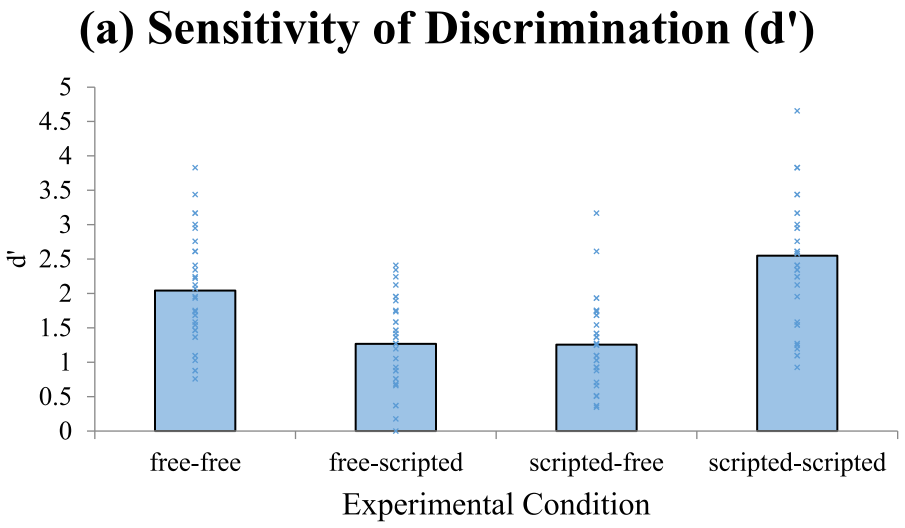
Note: confidence is rated on a 7 point scale with 7 = 'very confident indeed'. For all measures, standard deviation is provided in parentheses.

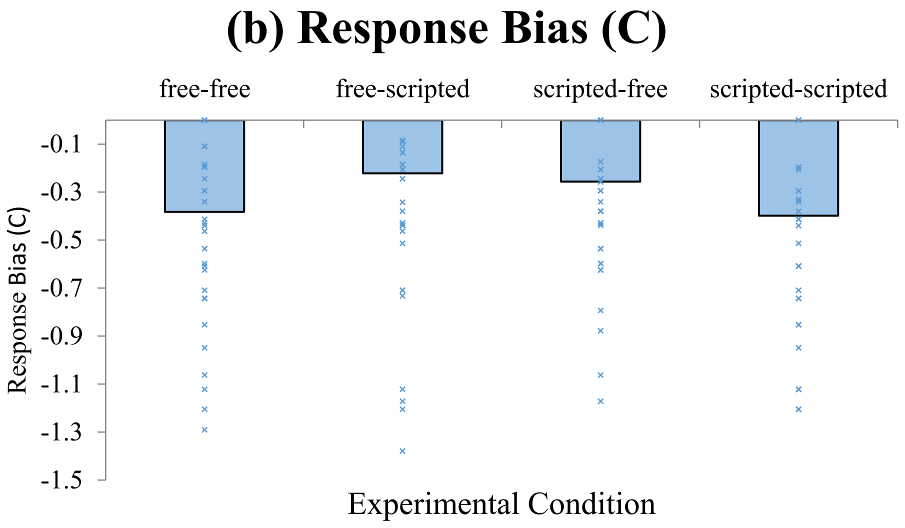
	Congruent 'free' (FF)	Incongruent 'free' (FS)	Incongruent 'scripted' (SF)	Congruent 'scripted' (SS)
Sensitivity of discrimination (d')	2.04 (.79)	1.27 (.65)	1.26 (.58)	2.55 (.93)
Prop accuracy on 'same' trials	.88 (.10)	.77 (.15)	.79 (.13)	.92 (.11)
Prop accuracy on 'different' trials	.72 (.16)	.64 (.19)	.62 (.17)	.76 (.17)
Confidence on 'same' trials	5.98 (.65)	5.60 (.77)	5.76 (.73)	6.02 (.70)
Confidence on 'different' trials	5.34 (.70)	5.17 (.90)	5.07 (.87)	5.31 (.83)
Response bias (C)	-.38 (.43)	-.22 (.50)	-.26 (.43)	-.40 (.50)

Figure 1:

Plots of (a) sensitivity of discrimination (d’), (b) response bias (C), (c) proportion accuracy, and (d) self-rated confidence across the four experimental conditions.

Note: markers show individual data points, with ties indicated by a single marker. Accuracy and confidence measures are show separately for ‘same’ and ‘different’ trials.





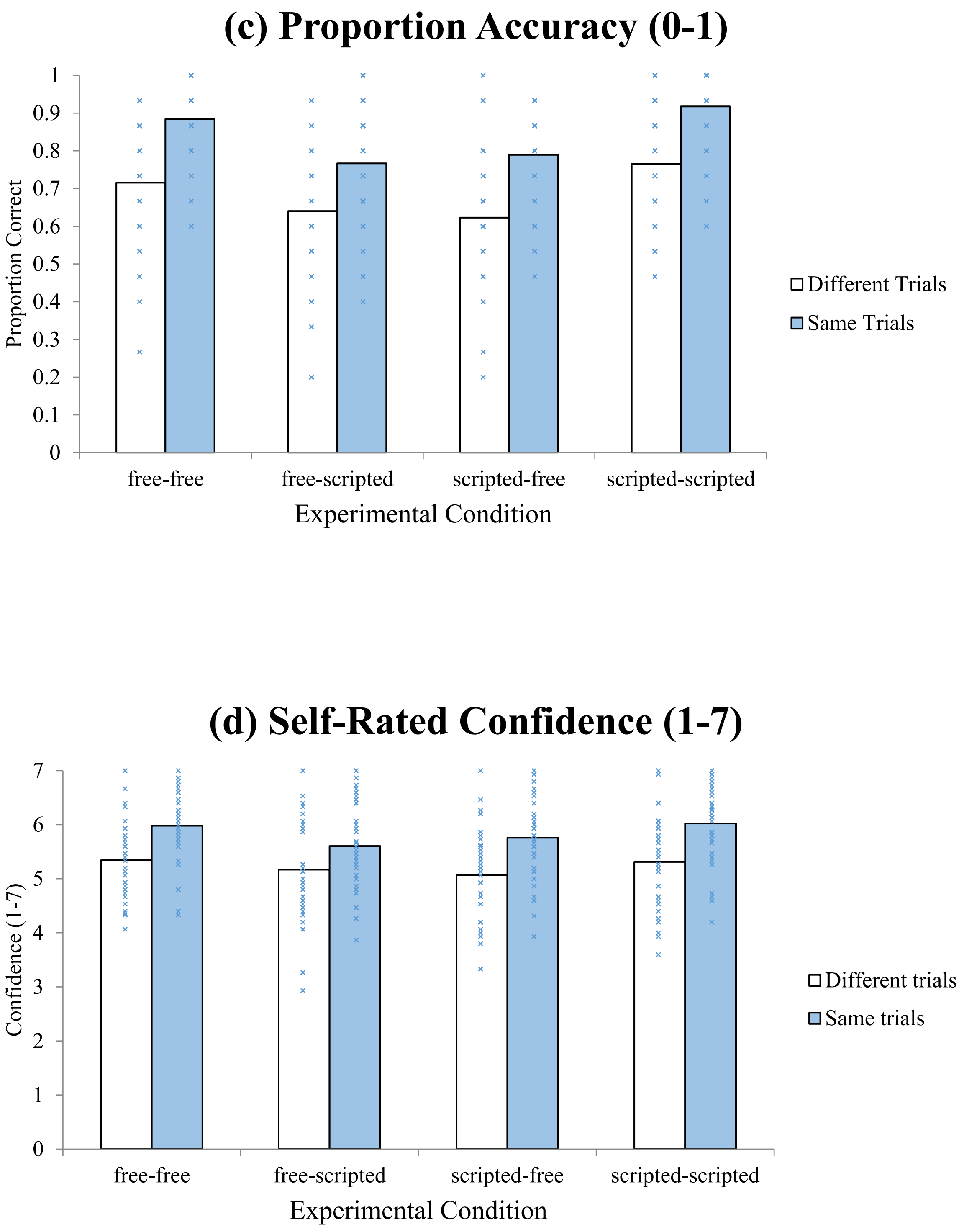
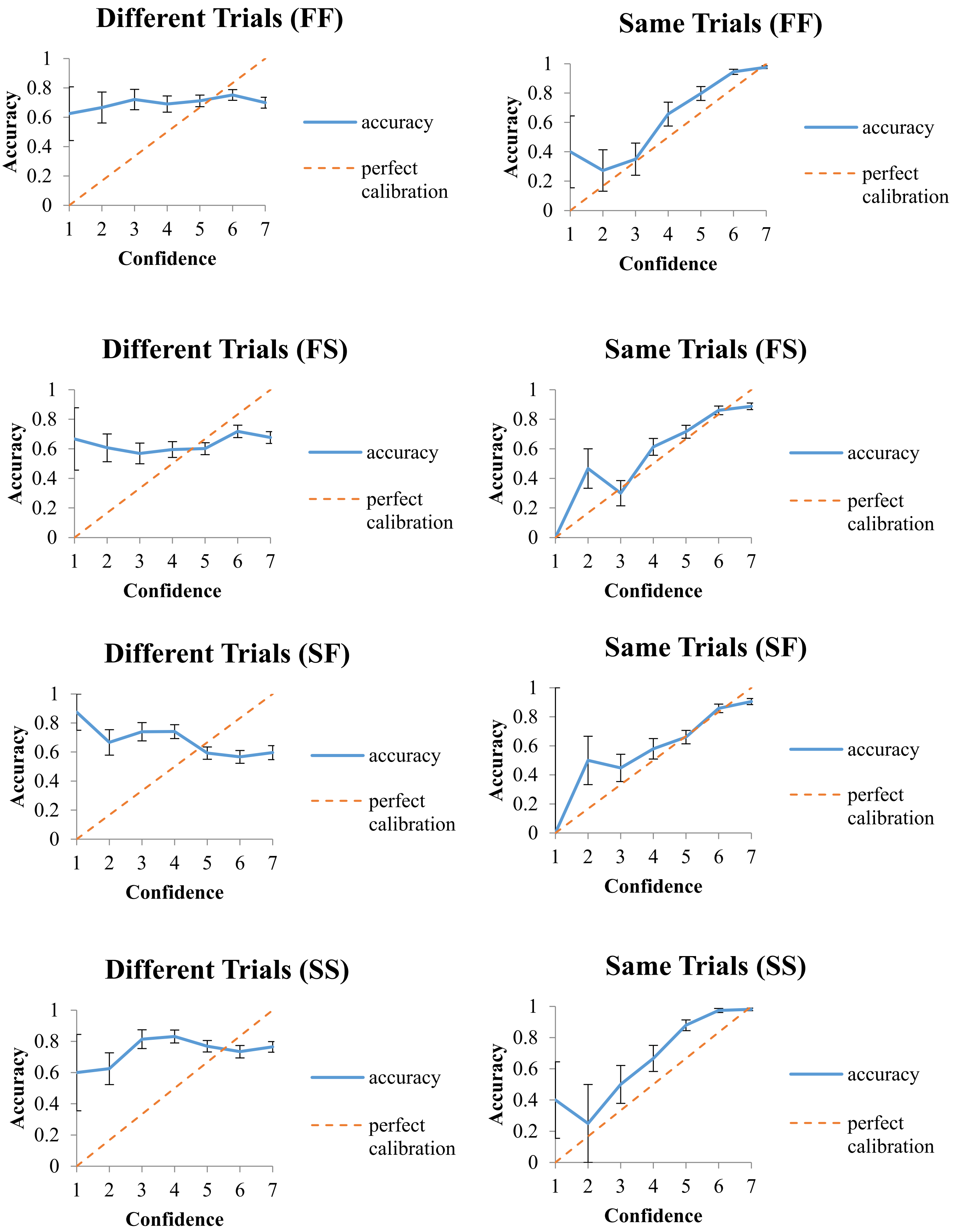

(c) Proportion Accuracy (0-1)
Proportion Correct
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
free-free
free-scripted
scripted-free
scripted-scripted
Experimental Condition
□ Different Trials
■ Same Trials
(d) Self-Rated Confidence (1-7)
Confidence (1-7)
0
1
2
3
4
5
6
7
free-free
free-scripted
scripted-free
scripted-scripted
Experimental Condition
□ Different trials
■ Same trials

Figure 2: Calibration curves for ‘same’ and ‘different’ trials in each of the four experimental conditions. Note: Error bars show standard error of the mean.



Different Trials (FF)

Same Trials (FF)

Different Trials (FS)

Same Trials (FS)

Different Trials (SF)

Same Trials (SF)

Different Trials (SS)

Same Trials (SS)

Appendix: List of free-speech topics and scripted speech sentences.

Free speech topics:

Describe a typical day in your life.

Tell us about your favourite TV show or film.

Tell us something interesting about yourself.

Tell us about your dream home.

Scripted speech sentences:

The smell of freshly ground coffee never fails to entice me into the shop.

They launched into battle with all the forces they could muster.

The length of her skirt caused the passers-by to stare.

The most important thing to remember is to keep calm and stay safe.