A Sound Effect:

Exploration of the Distinctiveness Advantage in Voice Recognition.

Sarah V Stevenage[1]* Greg J Neil[2] Beth Parsons[3] & Abi Humphreys[1]

[1]Department of Psychology, University of Southampton, UK

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

[3]Department of Psychology, University of Winchester, UK

* Please address correspondence to Professor Sarah Stevenage at:

Department of Psychology

University of Southampton

Highfield

Southampton

Hampshire, UK

Email: svs1@soton.ac.uk; Tel: +44 2380 592973, Fax: +44 2380 594597

Abstract

Two experiments are presented which explore the presence of a distinctiveness advantage when recognising unfamiliar voices. In Experiment 1, distinctive voices were recognised significantly better, and with greater confidence, in a sequential same/different matching task compared to typical voices. These effects were replicated and extended in Experiment 2 as distinctive voices were recognised better even under challenging listening conditions imposed by nonsense sentences and temporal reversal. Taken together, the results aligned well with similar results when processing faces, and provided a useful point of comparison between voice and face processing.

A Sound Effect:

Exploration of the Distinctiveness Advantage in Voice Recognition.

The capacity to recognise someone from their face is relatively well-researched in terms of theoretical, behavioural and neuropsychological findings (see Schweinberger & Burton's special issue, 2011). Against this backdrop, attention has relatively recently turned to the area of voice recognition. In this regard, researchers have been keen to identify similarities in performance between voice and face processing through applying the theories and methodologies from one area to the study of the other area. The present paper draws on this approach with specific focus on the parallel effects of distinctiveness in the face recognition and voice recognition literatures.

Recognising Faces and Recognising Voices

The voice and the face are perhaps the two most readily available cues to identity (Ellis, Jones & Mosdell, 1997). Both provide rich sources of information, communicating both affective state, and linguistic speech information as well as identity. Indeed, the voice has become known as an 'auditory face' (Belin, Fecteau & Bédard, 2004; Belin, Bestelmeyer, Latinus & Watson, 2011) in recognition of the range of valuable information that it can indicate. With both faces and voices providing complementary cues about an individual, it is tempting to hold similar expectations when considering their processing. Despite this, early consideration suggested that voice recognition was substantially weaker than face recognition (see Stevenage & Neil, 2014 for a review). Notably, when recognising famous celebrities, voice recognition was significantly worse than face recognition, and voices produced significantly more 'familiar only' experiences compared to faces (Ellis, *et al.,* 1997; Hanley,

Smith & Hadfield, 1998). In fact, a series of well-designed studies suggested that voice and face recognition could only be equated when face recognition was compromised through substantial levels of blurring (Damjanovic & Hanley, 2007; Hanley & Damjanovic, 2009). Voices also served as weaker cues relative even to blurred faces when trying to retrieve both semantic details about celebrities, such as their occupation (Hanley & Damjanovic, 2009; Hanley, Smith & Hadfield, 1998), and episodic details about a time when they were previously encountered (Damjanovic & Hanley, 2007).

One possible explanation for the relative weakness of voices compared to faces, both when recognising celebrities and when retrieving information about them, is that participants may have experienced greater exposure to faces than voices given the popularity of media images. To test this account, Brédart and colleagues examined performance with personally familiar stimuli (which were likely to be heard as often as seen) and newly learned stimuli (for which face and voice exposure could be carefully controlled). As above, the results suggested that voices led to poorer retrieval of semantic details than faces, both when stimuli were personally familiar (Barsics & Brédart, 2011; Brédart, Barsics & Hanley, 2009) and when newly learned (Barsics & Brédart, 2012a). Voices also led to poorer retrieval of episodic information compared to faces when stimuli were personally familiar (Barsics & Brédart, 2011). The fact that performance was still poorer in these studies when cued with voices than faces suggested that differential exposure was unlikely to account for the previous findings (Brédart & Barsics, 2012a). Instead, the results suggested that, whilst both voices and faces could be used as cues to identity, the voice was less effective compared to the face.

This conclusion has been supported by results from several convergent methodologies. Using a priming methodology, for example, cross-modal repetition priming has been demonstrated whereby the face of a celebrity target facilitated the later recognition

of their voice, and vice versa (Ellis et al., 1997; Schweinberger, Herholz & Stief, 1997). However, the results of Stevenage, Hugill & Lewis (2012) suggested that the voice was a far weaker prime for later face recognition than the face was for later voice recognition. In an adaptation to this task, a conflicting voices paradigm was developed in which celebrity recognition was examined from the face and voice under conditions in which the face and voice either matched (both belonged to the same celebrity) or mismatched (both belonged to different celebrities). Performance indicated that face recognition remained strong and robust regardless of the identity of the accompanying voice. However, voice recognition was substantially impaired when the accompanying face belonged to a different celebrity (Stevenage, Neil & Hamlin, 2014).

Finally, the results of interference studies are relevant. In this paradigm, distractor faces were presented in between the study and test phases of a face-matching task, and distractor voices were presented in between the study and test phases of a voice-matching task (Stevenage, *et al.*, 2013). When examining performance using this interference methodology, face recognition remained strong despite the introduction of distractor faces between study and test. However, voice recognition was significantly and negatively affected by the introduction of distractor voices suggesting, once again, that voice recognition was weaker, and more susceptible to factors that affected performance, compared to face recognition.

Consideration of a Distinctiveness Advantage

Taken together, a substantial body of work now exists to suggest that the voice is measurably weaker as a cue to identity compared to the face. This said, the examination of averaged levels of performance across a voice set may mask an important factor - the distinctiveness of one voice compared to another. In this regard, evidence is emerging to

indicate a distinctiveness advantage during voice processing. When considering *familiar* voice recognition for example, Skuk and Schweinberger (2013) revealed that 12th graders were better able to recognise the voices of 20 of their classmates when those voices were distinctive rather than typical. In fact, a substantial correlation existed (r = .687) between recognition and rated distinctiveness.

In a similar vein, Foulkes and Barron (2000) asked ten friends, and two foils to record an 8-10 second scripted answerphone message. The ten friends then attempted to recognise themselves and one another from the resultant 12 voice clips. Voice recognition varied substantially. However, as above, performance was significantly better when voices were more distinctive in terms of pitch and pitch variation.

Barsics and Brédart (2012b) took a slightly different approach by examining distinctiveness effects for celebrity voices. They asked participants to make a familiarity judgement to 64 celebrity or non-celebrity voices before providing episodic details of a previous encounter, plus a name or other biographic information. In keeping with the previous results, Barsics and Brédart noted better recognition of celebrity voices, and better retrieval of semantic information, when voices were distinctive than when typical.

Distinctiveness effects have also been noted when processing *unfamiliar* voices, however here, the studies have used a broad variety of methods, and the results have not always been clear. For example, Yarmey (1991) asked participants to listen to a 36 second monologue from a single unfamiliar speaker within the context of a fictitious kidnapping scenario. Following presentation, participants provided a description of the voice either immediately, or after a delay of a day or a week. The results suggested that the descriptions of a typical voice were substantially affected by delay. However, descriptions of a distinctive

voice showed remarkable consistency even after a week, suggestive of a distinctiveness advantage with unfamiliar voices.

Mullenix, *et al.* (2009) extended this work by using a voice recognition task, again with a single unfamiliar speaker rated as either typical or distinctive. They asked participants to engage in a word classification task to spoken words before completing a surprise voice recognition test a week later for the voice of the speaker. The recognition test took the form of an old/new matching task, with 'old' clips being spoken by the target, but with 'new' clips being spoken by typical and by distinctive foil speakers. Interestingly, the results did not suggest a distinctive advantage when recognising the target speaker. However, they did indicate a significantly higher error rate to 'new' voices when the original target had been typical than when distinctive, and this was primarily due to confusions between the typical target and typical foils. As such, Mullenix *et al*. (2009) demonstrated a distinctiveness advantage with unfamiliar voices, not through better recognition of the distinctive target but through fewer false recognitions of foils.

Using a very different approach, Sauerland, Sagana and Otgaar (2013) conducted a choice blindness task in which participants were asked to listen to three pairs of voices and to choose one from each pair according to a pre-defined criterion. Following each selection, the chosen voice was then re-presented for further consideration. However, on one critical trial, the chosen voice was switched with the non-chosen voice. A failure to spot the switch was termed 'choice blindness'. Sauerland *et al*. (ibid) noted that the incidence of choice blindness was significantly reduced when the voices in the pair were less similar to one another. In other words, participants noticed the switch more readily when the foil was very different from the chosen target, possibly because they differed significantly on rated distinctiveness.

In a more recent and novel cross-modal study, Bülthoff and Newell (2015) asked participants to learn face-voice pairs, with half the faces paired with a distinctive voice (n = 12) and half with a typical voice (n = 12). In both a between-participants design (Experiment 1a) and a within-participants design (Experiment 2), the results demonstrated better subsequent recognition of the face when it had been paired with a distinctive sounding voice than with a typical sounding voice. The authors suggested that the distinctiveness of the voice made the face more distinctive and thus improved face memory. However, the results could also be interpreted in the context of multimodal person perception to which the characteristics of both the voice and the face contributed. Either way, the results suggested a unique form of distinctiveness advantage in which vocal distinctiveness facilitated subsequent person perception from the face.

In evaluating these results, it is worth noting that faces were arbitrarily paired with voices rather than being paired with their own (distinctive or typical) voices, conferring both a strength and a weakness to the design. The strength was that a common set of faces could be paired with distinctive and typical voices in a counterbalanced fashion to control item effects. However, the weakness generated by this design was that the arbitrary pairing of faces with voices could have generated mismatch effects, especially when voices were distinctive (i.e., a female voice was matched with a male face, and a Japanese speaker was matched with a Caucasian face). As such, the apparent vocal distinctiveness effect demonstrated by Bülthoff and Newell (2015) is perhaps open to interpretation.

Against this backdrop, a review of the literature suggested only one study which demonstrated a clear and direct distinctiveness advantage when recognising unfamiliar voices. This is provided by Sørensen (2012) who examined unfamiliar voice recognition by means of a delayed lineup task. Within this study, distinctiveness was operationalised based on a measure of fundamental frequency, and the results showed a distinctiveness advantage,

through superior recognition performance when the voice sounded distinctive (74%) rather than typical (56%).

Taken as a whole, distinctiveness effects when processing *unfamiliar* voices have been examined using an imaginative range of methodologies. However, the results have not always clearly indicated better recognition of distinctive versus typical targets (see Mullenix *et al*., 2009). Additionally, generalisation of the distinctiveness advantage when processing *unfamiliar* voices has, at times, been limited by the use of one, or relatively few, targets (Mullenix et al., 2009; Sørensen, 2012, Yarmey, 1991, *cf* Bülthoff & Newell, 2015). As such, the evidence for a distinctiveness advantage when recognising unfamiliar voices would benefit from replication and extension, and this is the purpose of Experiment 1.

Experiment 1 tests for a distinctiveness advantage with unfamiliar voices using a sequential same/different matching task. This method is favoured over an old/new recognition task due to the potential for interference effects in the latter task when presenting lists of voices at study and at test. However, the sequential nature of the task does impose a memory demand on the participants given that the vocal information naturally unfolds over time. Nevertheless, previous studies using a sequential same/different task have shown performance levels that avoid both floor and ceiling effects (see Stevenage *et al.*, 2013). Experiment 1 also uses a relatively large voice set to test the generalisability of previous results. Based on distinctiveness effects within the face recognition field, and the available results in the voice recognition field, it was predicted that unfamiliar voices would be recognised better when distinctive than when typical.

Experiment 1: A Distinctiveness Advantage in Unfamiliar Voice Recognition

Design

A 2 x 2 within-participants design was used in which vocal distinctiveness (distinctive, typical) and trial type ('same', 'different') were varied within a sequential matching task. The participants heard two voice clips one after the other, and were asked to decide whether the two clips came from the 'same' speaker or from 'different' speakers. Their accuracy and self-rated confidence on 'same' and 'different' trials represented the dependent variables.

Participants

A total of 72 participants (54 females) took part in return for course credit or a small monetary payment. Their ages ranged from 19 to 35 years (M = 22.53, SD = 3.51) and all participants reported normal hearing and a lack of familiarity with the speakers.

Materials

A total of 117 speaker samples were collected for the purposes of this study. For all speakers, two clips were recorded so that the clips at study and at test were not identical during a 'same' trial. In the study clip, the speaker said the phrase 'The smell of freshly ground coffee never fails to entice me into the shop' (mean duration = 5 seconds). In the test clip, they said the phrase 'The length of her skirt caused the passers-by to stare' (mean duration = 4 seconds). Both phrases were created to provide phonetic richness when exploring speaker identification, and were drawn from corpus of phrases used in the FRL2011 database (UK Home Office Centre for Applied Science and Technology).

The 117 speakers were designated as 'distinctive' or 'typical' according to the ratings of 6 independent judges. All ratings were made on a 7 point scale (1 = not at all distinctive, 7 = very distinctive) and were obtained by asking the judges to imagine that they were in a

noisy environment, such as a party, and to indicate how much each voice would stand out against the other voices. These instructions were modelled on those used to judge facial familiarity ('How much would this face stand out at a busy railway station?') (Valentine & Bruce, 1986). Based on these ratings, 32 distinctive voices (average distinctiveness for each ≥ 5) and 32 typical voices (average distinctiveness for each ≤ 3.5) were selected, with an equal number of male and female speakers in each set. In terms of the distinctiveness ratings, excellent agreement was indicated across the judges (Cronbach's Alpha = .96), and an independent samples *t*-test confirmed that the two sets of voices differed significantly in terms of their rated distinctiveness (distinctive set: $M = 6.07$, $SD = .74$; typical set: $M = 2.46$, $SD = .65$; $t_{(62)} = 20.72$, $p < .001$).

In addition to the target voices described above, the voices of 16 males and 16 females of intermediate distinctiveness (distinctiveness rating = 3.99, $SD = .99$) were selected to act as foils in the 'different' trials. Given their intermediate level of distinctiveness, the foils differed on rated distinctiveness compared to both the distinctive targets ($t_{(62)} = 9.46$, $p < .001$) and typical targets ($t_{(32)} = 7.27$, $p < .001$). As a group, these foils were matched to the set of targets on sex, and on similarity of perceived pitch, according to the judgements, by ear, of the experimenters. Subsequent analysis of F0 (as determined using Praat6039 for Windows) confirmed that the foils did not differ from either the typical or the distinctive targets in terms of F0 (typical females: $t_{(30)} = 1.22$, $p = .232$; typical males: $t_{(30)} = .48$, $p = .635$; distinct females: $t_{(30)} = .45$, $p = .659$; distinct males: $t_{(30)} = .37$, $p = .716$). Consequently, whilst the individual target voices may have varied in terms of pitch, particularly in the case of distinctive targets, the population of target voices did not stand out from the population of foil voices used.

From these stimuli, 16 'same' trials and 16 'different' trials were constructed, with 8 distinctive and 8 typical voices contributing to each set. The 'same' trials consisted of a target

speaker uttering phrase 1, and then the same speaker uttering phrase 2. The ‘different’ trials consisted of a target speaker uttering phrase 1 and a same-sex foil speaker uttering phrase 2. Finally, the assignment of target voices to ‘same’ and ‘different’ trials was counterbalanced so that each voice was heard equally often in a ‘same’ trial and a ‘different’ trial across the participant population.

The trials were presented, and data were recorded via Superlab Pro 4.5.4 (Cedrus, released 2012) via a DELL PC laptop (with an Intel i5 core and a 64-bit operating system) running Windows XP. All written instructions were presented via the 14” colour screen laptop monitor, but sound was presented via outer-ear Pro-Luxe PRO-40 Hi-Fi headphones with a frequency response of 20Hz to 20KHz. Sound volume was adjustable via the computer settings to ensure optimal listening conditions.

Procedure

All participants were tested individually within a quiet testing cubicle. After providing informed consent, a practice phase was presented in which the participants were asked to press S in response to the word ‘same’ and D in response to the word ‘different’ as it appeared on the screen. A total of 16 trials enabled the participants to map the correct key to each response, and feedback was provided.

Following this, 8 further practice trials were presented to introduce the participants to the format of the experimental trials. Instead of using voice clips, these practice trials used words. Following a ‘next trial’ prompt, a target word was presented for 500 msecs after which participants gave a rating from 1-7 for ‘pleasantness’. This ensured that the participants attended to the target. After a 5 second gap, a second word was presented and remained on screen until the participant indicated whether it was the same (S) or different (D) to the target word seen previously. Feedback was again provided.

A self-paced break followed during which the participants could ask for clarification of the task as required. After this, a randomised sequence of 32 experimental trials (16 ‘same’, 16 ‘different’) was presented, and no further feedback was available. All trials followed an identical format consisting of a ‘next trial’ prompt (250 msecs), a blank screen (100 msecs), and the presentation of the target voice clip. The participants rated this clip for vocal attractiveness using a 7 point scale, again as a way of ensuring attention to the target. An inter-stimulus interval of 16 seconds followed so that the matching task was not too easy. Finally, a second voice clip was presented, along with the on-screen question ‘same or different?’. The participants indicated their response by pressing S for ‘same’ and D for ‘different’, and the emphasis was on accuracy over speed. Finally, the participants indicated their confidence in their answer by pressing a numbered key from 1 (not at all confident) to 7 (very confident indeed).

Following completion of the task, the participants were thanked and debriefed, and the entire task lasted no more than 30 minutes.

Results and Discussion

The data from one participant were excluded through identification as an outlier reflecting poor mean performance on the easiest trials (with distinctive stimuli). The data from 71 participants remained. Given the use of a same/different task with a dichotomous response, the data were explored in line with the signal detection framework (Green & Swets, 1966). Accordingly, the accuracy scores for ‘same’ and ‘different’ trials were combined to yield primary measures of sensitivity of discrimination (*d’*) and response bias (*C*). The analysis of accuracy and confidence on ‘same’ and ‘different’ trials provided a secondary analysis.

Sensitivity of Discrimination and Response Bias

Sensitivity of discrimination (*d'*) for distinctive and for typical voices is summarised in Table 1 along with a measure of response bias (*C*). Taking sensitivity of discrimination first, a paired samples *t*-test was used to determine whether vocal distinctiveness had any effect on performance. This revealed a significant difference ($t_{(70)} = 5.80$, $p < .001$) supporting the prediction of a distinctiveness advantage. In contrast, when considering response bias (*C*), no effect of vocal distinctiveness emerged ($t_{(70)} < 1$, $p = .742$). In fact, one-sample comparisons to zero revealed no bias in responding, either overall ($t_{(70)} = 1.29$, $p = .20$) or for distinctive and typical stimuli when taken separately (both $ts_{(70)} < 1.17$, $p > .244$). These results suggested that a distinctiveness advantage was demonstrable when recognising a large set of unfamiliar voices, with this being shown through sensitivity of discrimination rather than response bias.

(Please insert Table 1 about here)

Accuracy of Performance

Accuracy of performance is summarised in Table 1 for both 'same' and 'different' trials separately. This was examined by means of a 2 x 2 repeated-measures Analysis of Variance (ANOVA) in which the effects of both vocal distinctiveness (distinctive, typical) and trial type ('same', 'different') were explored. Importantly, there was a main effect of distinctiveness ($F_{(1, 70)} = 40.09$, $p < .001$, $\eta^2_G = .09$, $MSE = .03$), with performance being better for distinctive than for typical voices. The analysis revealed no main effect of trial type ($F_{(1, 70)} = 2.95$, $p = .09$, $\eta^2_G = .01$, $MSE = .05$). Moreover, there was no interaction between distinctiveness and trial type ($F_{(1, 70)} < 1$, $p = .80$, $\eta^2_G < .01$, $MSE = .02$) indicating that the distinctiveness advantage emerged for 'same' and 'different' trials alike.

Self-Rated Confidence

Finally, analysis was conducted on self-rated confidence when recognising typical and distinctive voices. These data were calculated across all trials and are summarised in Table 1, They were analysed by means of a 2 x 2 repeated-measures ANOVA in which vocal distinctiveness (distinctive, typical), and trial type ('same', 'different') were explored. As above, this revealed a main effect of distinctiveness ($F_{(1, 70)} = 69.77$, $p < .001$, $\eta^2_G = .23$, $MSE = .49$), with confidence being greater when recognising distinctive voices than when recognising typical ones. There was, however, no main effect of trial type ($F_{(1, 70)} = 3.88$, $p = .053$, $\eta^2_G = .03$, $MSE = .80$). Again, there was no interaction between distinctiveness and trial type ($F_{(1, 70)} < 1$, $p = .81$, $\eta^2_G < .01$, $MSE = .31$) indicating that confidence was greater for distinctive than typical voices in 'same' and 'different' trials alike.

Taken together, the data from Experiment 1 were clear in supporting the prediction of a distinctiveness advantage when recognising unfamiliar voices. This advantage was revealed in sensitivity of discrimination, accuracy for 'same' and 'different' trials, and in self-rated confidence. One strength of the current study lies with the use of a large number of distinctive and typical voices, avoiding concerns that previous mixed results may have been driven by particular items within small stimulus sets. As such, this evidence sits well alongside the considerable body of work indicating a distinctiveness advantage when recognising faces (Bartlett, Hurry & Thorley, 1984; Goldstein & Chance, 1981; Light, Kayra-Stuart & Hollander, 1979; Shepherd, Gibling & Ellis, 1991; Valentine & Bruce, 1986; Winograd, 1981), as well as the findings indicating a distinctiveness advantage when recognising personally familiar or celebrity voices.

Accounting for Distinctiveness Effects using a Similarity Space Framework

In the context of face processing, the distinctiveness advantage has been elegantly accounted for by appealing to the fact that distinctive items stand out on one or more

dimensions of a face similarity space. Consequently, they suffer less confusability with near-neighbours during a recognition task compared to their typical counterparts (Valentine, 1991). Recent work has extended the concept of a similarity space to the perception of voices, with dimensions of the space reflecting the vocal characteristics that listeners use to differentiate voices (Baumann & Belin, 2010). By extension, distinctive voices again stand out on one or more dimensions that define the voice space, leading to less confusability with vocal near-neighbours compared to their typical voice counterparts. As such, Experiment 1 provides a valuable addition to the empirical evidence for a distinctiveness advantage when recognising unfamiliar voices, suggesting a robust and replicable effect which can be readily accounted for within a similarity-based voice space framework.

Experiment 2: A Distinctiveness Advantage under Challenging Conditions

The next natural question is whether vocal distinctiveness would assist the listener even when listening conditions are challenging. Such a prediction may follow by extension of the voice space framework. In this context, if listening conditions are challenging, the error when encoding a voice may be expected to increase, reducing the likelihood of a match between a voice and its previously stored representation. This additional error may be more likely to affect typical voices than distinctive voices given that typical voices are more confusable at the outset. As a result, it may be predicted that distinctive voices would retain a processing advantage even when presented under challenging listening conditions.

Two studies are of relevance to this question in as much as they suggest quite contradictory findings. The first study is provided by van Lancker, Kreiman and Emmorey (1985), who tested familiar voice recognition under three discrete conditions. First, participants listened to 2 second voice clips belonging to 45 celebrities before indicating

whether each voice was familiar (or not) from an unlimited set (task 1). Following this, participants listened to a new set of 2 second voice clips for the same celebrities, before indicating the speakers' identity by selecting one of six possible names (task 2). Participants were able to recognise nearly 27% of targets when presented in an unlimited set, and nearly 70% of targets when presented in a 6-alternate forced-choice task. Of most interest, however, was the performance in a final condition in which participants listened to 4 second clips played backwards, before again indicating speaker identity from six names (task 3). Remarkably, participants remained able to recognise over 57% of targets in the 6AFC task despite their temporal reversal. Notably, performance on these backwards voices varied substantially across the targets, with some targets being *equally recognisable* when played backwards as when played forwards. The authors considered that these unanticipated item effects may have been driven by variation in the distinctiveness of the target voices, suggesting that distinctiveness may provide an advantage when processing voices under difficult or unusual listening conditions.

In direct contrast are the findings of Orchard and Yarmey (1995). As in Yarmey's (1991) earlier work, Orchard and Yarmey asked participants to listen to either a distinctive or a typical target voice presented in the context of a fictitious kidnapping scenario. Two days later, participants were asked to identify the target from a six-person target-present or target-absent lineup. Several factors were varied including whether the target spoke normally or in a whisper, and whether the voice at lineup was of the same format (normal, whisper) to the voice at study. The results suggested that performance was significantly impaired by whispering, and by a change in speech style, in both target-present and target-absent lineups. Of more interest, however, performance was significantly affected by the distinctiveness of the target voice, but surprisingly, this indicated a trend for typical sounding voices to be better recognised – a distinctiveness disadvantage. This appeared to be mediated by several

variables including whether the speaker was whispering, and whether the listener felt confident in their recognition. As such, the evidence regarding a distinctiveness advantage under difficult listening conditions remains unclear. Experiment 2 was designed to address this issue.

Within Experiment 2, challenging listening conditions were introduced through either changing the word order within a sentence to create a nonsense clip, or through temporally reversing the voice clip. Similar manipulations have been used with faces as ways to disrupt facial processing. In such studies, the scrambling of features within an otherwise upright face, or the inversion of the face entirely, have been thought to disrupt the ability to process the critical relationships between features (see Tanaka & Farah, 1993). Whilst it cannot be assumed that nonsense speech or temporal reversal have the same disruptive effect on voices as scrambling and inversion have on faces, these manipulations have been used to good effect when making voice processing difficult (see Goggin, Thompson, Strube & Simental, 1991, Expt 4; van Lancker, Kreiman & Emmorey, 1985).

Manipulation here through the creation of a nonsense clip, or through temporal reversal, has the advantage of introducing a cognitive challenge to the listening task whilst leaving the paralinguistic properties of the stimuli unaffected. To the extent that vocal distinctiveness may be carried in these vocal properties rather than in features associated with the utterance, the distinctiveness of the voice was unchanged by the manipulation of task difficulty. Given this, if vocal distinctiveness is effective in protecting voice recognition abilities as predicted, then the recognition of distinctive voices should be superior to that of typical voices, even under these challenging listening conditions.

Design

A 3 x 2 mixed design was used in which listening condition (forwards, nonsense, backwards) was manipulated between participants, and vocal distinctiveness (distinctive, typical) was manipulated within participants. As in Experiment 1, voice recognition was examined through a sequential same/different matching task, and accuracy and self-rated confidence represented the dependant variables.

Participants

A total of 48 participants (37 females) took part in return for course credit. The participants were randomly assigned to one of three listening conditions such that they heard speech at study which was either played forwards (n = 16; 12 females), in a nonsense order (n = 16; 12 females) or backwards (n = 16; 13 females). The participants' ages ranged from 15 to 60 years (M = 23.9 years, SD = 10.2), and all participants had normal, or corrected-to-normal, hearing. The participants reported no familiarity with the stimuli, and none had taken part in the previous experiment.

Materials

Bespoke stimuli were used for this experiment, consisting of 60 speakers drawn from the same demographic population as the participants in terms of age range, and accent. All speakers were recorded uttering a study phrase under various conditions, along with a test phrase. Mirroring Experiment 1, the study phrase was 'The smell of freshly ground coffee never fails to entice me into the shop". To provide a nonsense version of this study phrase, the adjectives and nouns were repositioned within the sentence ('The shop of ground fails smell never coffee into the me freshly to entice'). The speakers practiced this nonsense phrase prior to recording until they were able to utter it with a cadence and phrasing that felt natural. To provide a temporally reversed (backwards) version of the study phrase, the 'reverse' function within Audacity 2.0.5 was used, resulting in a clip that was

incomprehensible whilst still preserving the acoustic properties of the speaker. Finally, and again mirroring Experiment 1, the speakers were recorded uttering a separate test phrase ('The length of her skirt caused the passers-by to stare'). This ensured that the study and test phrases were not identical during a 'same' trial.

Using the standard study phrase played forwards, all speakers were rated for their vocal distinctiveness by the experimenters using a 7-point scale, where 1 = not at all distinctive, and 7 = very distinctive indeed. On the basis of these ratings, 16 distinctive voices (M = 5.69, SD = .68) and 16 typical voices (M = 3.50, SD = .48) were selected as targets. In terms of the distinctiveness ratings, agreement between the raters was again good (Cronbach's alpha = .82), and the two voice sets differed significantly on vocal distinctiveness ($t_{(30)} = 10.49$, $p < .001$).

The 32 target voices were then paired to construct 16 'same' trials (8 distinctive, 8 typical), and 16 'different' trials (8 distinctive, 8 typical). The 'same' trials were constructed by pairing a study clip from one speaker with a test clip of the same speaker, whilst the 'different' trials were constructed by pairing a study clip from one speaker with a test clip from a same-sex foil speaker drawn from the remaining voices. Similarity ratings by the experimenters on a 7 point scale (1 = not at all similar, 7 = very similar indeed) confirmed that the similarity of distinctive and typical targets to their respective foils was matched (distinctive similarity: M = 5.87, SD = .64; typical similarity: M = 5.87, SD = .79; $t_{(14)} < 1$, *ns*). This ensured that the 'different' trials did not represent a trivially easy task for one or other voice set.

The voices trials were presented and data were recorded using SuperLab Pro 4.5.4 via a DELL PC laptop (with an Intel i5 core and a 64-bit operating system) running Windows 7. As in Experiment 1, all written instructions were presented via the 14" colour screen laptop

monitor, but sound was presented via outer-ear Pro-Luxe PRO-40 Hi-Fi headphones with a frequency response of 20Hz to 20KHz. Sound volume was adjustable via the computer settings to ensure optimal listening conditions.

Procedure

The participants were tested individually within a quiet experimental cubicle. Following explanation of the task, and the indication of informed consent, the participants completed a set of practice trials during which they were required to press ‘S’ or ‘D’ to the words ‘same’ or ‘different’ as they appeared on the screen. This stage enabled the participants to map the correct keyboard key to each response.

After a self-paced break, the 32 experimental trials (16 ‘same’, 16 ‘different’) were presented in a random order. All trials took the same format beginning with a ‘next trial’ prompt (250 msecs) to encourage the participants to orient towards the task. This was followed by the presentation of the study voice clip for 4 seconds. The voice was either distinctive or typical, and was heard in either the forwards, nonsense or backwards format depending on the condition to which the participant had been assigned. In contrast to Experiment 1 in which a 16 second gap was used, the present study adopted only a 4 second gap between study and test clips. This change reflected a desire to avoid floor effects associated with poor performance in the most challenging of the listening conditions. Following this 4 second gap, the test clip was played, and the participants’ task was to indicate whether it was the ‘same’ speaker or a ‘different’ speaker to the one heard at study. The participants responded by pressing ‘S’ or ‘D’ respectively. Finally, they indicated their confidence in their answer by pressing a numbered key from 1 (not at all confident) to 7 (very confident indeed).

The entire experiment lasted approximately 25 minutes, after which the participants were thanked and debriefed.

Results and Discussion

As in Experiment 1, the accuracy data for 'same' and 'different' trials were combined to provide measures of sensitivity of discrimination (*d'*) and bias (*C*). Primary analyses are reported on these measures, with secondary analyses provided using accuracy and confidence on 'same' and 'different' trials.

Sensitivity of Discrimination (d')

Sensitivity of discrimination (*d'*) and response bias (*C*) when recognising distinctive and typical sounding voices are summarised in Table 2 when voices were played in forwards, nonsense, and backwards formats at study. A 2 x 3 mixed ANOVA on sensitivity of discrimination revealed a significant main effect of distinctiveness ($F_{(1, 45)} = 15.71$, $p < .001$, $\eta^2_G = .11$, $MSE = .64$) with performance being better for distinctive than for typical voices, as predicted. In addition, there was a significant main effect of listening condition ($F_{(2, 45)} = 43.36$, $p < .001$, $\eta^2_G = .55$, $MSE = 1.11$), with repeated contrasts indicating equivalence between forwards and nonsense listening conditions ($p = .80$) but showing a substantial reduction in performance between nonsense and backwards conditions ($p < .001$). The interaction between distinctiveness and listening condition was not significant ($F_{(2, 45)} < 1$, $p = .57$, $\eta^2_G < .01$, $MSE = .64$) suggesting that distinctiveness provided an overall advantage in each listening condition.

(Please insert Table 2 about here)

Bias (C)

A 2 x 3 mixed ANOVA was also conducted on response bias (*C*). This revealed a main effect of distinctiveness ($F_{(1, 45)} = 7.35$, $p = .009$, $\eta^2_G = .06$, $MSE = .20$) in which distinctive voices attracted significantly less bias than typical ones. There was no significant effect of listening condition overall ($F_{(2, 45)} < 1$, $p = .992$, $\eta^2_G < .01$, $MSE = .29$). However, a significant interaction between distinctiveness and listening condition did emerge ($F_{(2, 45)} = 3.32$, $p = .045$, $\eta^2_G = .06$, $MSE = .20$). Tests of simple main effects revealed a significant reduction in response bias for distinctive over typical voices for nonsense speech ($F_{(1, 45)} = 13.08$, $p = .001$, $\eta^2_G = .15$, $MSE = .22$) but not when speech was played forwards ($F_{(1, 45)} < 1$, $p = .392$, $\eta^2_G = .01$, $MSE = .22$) or backwards ($F_{(1, 45)} < 1$, $p = .879$, $\eta^2_G < .01$, $MSE = .15$) when bias was minimal.

Accuracy

Accuracy of performance when recognising distinctive and typical voices is summarised in Table 2 in each of the experimental conditions. This was examined using a 3 x 2 x 2 mixed ANOVA in which listening condition (forwards, nonsense, backwards), distinctiveness (distinctive, typical), and trial type ('same', 'different') were varied. The analysis indicated no main effect of trial type ($F_{(1, 45)} = 2.83$, $p = .099$, $\eta^2_G = .02$, $MSE = .03$) and no interaction of trial type with any other variable (all Fs < 3.53, $p > .067$, $\eta^2_G < .02$, $MSE = .02$). Thus, performance was equivalent across 'same' and 'different' trials within this experiment. More importantly, the results indicated a significant main effect of distinctiveness ($F_{(1, 45)} = 30.50$, $p < .001$, $\eta^2_G = .08$, $MSE = .01$) and a significant main effect of listening condition ($F_{(2, 45)} = 40.51$, $p < .001$, $\eta^2_G = .37$, $MSE = .03$). As in the analysis of sensitivity of discrimination above, these confirmed that voice recognition was significantly better for distinctive than for typical voices, but was also significantly impaired as the message became more difficult to process.

Somewhat surprisingly, and in contrast to the analysis of sensitivity of discrimination, a significant interaction emerged between distinctiveness and listening condition ($F_{(2, 45)}$ = 4.60, p = .015, η^2_G =.03, *MSE* = .01). In line with the *a-priori* expectations, tests of simple main effects confirmed this to be due to a significant distinctiveness advantage in all conditions (forwards: $F_{(1, 45)}$ = 4.73, $p_{1\text{-tailed}}$ = .018, η^2_G = .06, *MSE* = .01; nonsense: $F_{(1, 45)}$ = 3.03, $p_{1\text{-tailed}}$ = .045, η^2_G = .04, *MSE* < .01; backwards: $F_{(1, 45)}$ = 31.95, $p_{1\text{-tailed}}$ < .001, η^2_G = .23, *MSE* = .01) but the effect was somewhat smaller in the nonsense condition.

Confidence

Self-rated confidence is summarised in Table 2 and was analysed as above using a 3 x 2 x 2 mixed ANOVA. Analysis across all trials revealed a broadly similar pattern of performance to that above. There was no main effect of trial type ($F_{(1, 45)}$ = 1.77, p = .19, η^2_G < .01, *MSE* = .24). However, there was a significant main effect of distinctiveness ($F_{(1, 45)}$ = 84.13, p < .001, η^2_G = .09, *MSE* = .22) and a significant main effect of listening condition ($F_{(2, 45)}$ = 38.82, p < .001, η^2_G = .59, *MSE* = 3.49). Two of the two-way interactions were significant (trial type x listening condition: $F_{(2, 45)}$ = 5.20, p = .009, η^2_G = .01, *MSE = .24;* trial type x distinctiveness: $F_{(1, 45)}$ = 11.38, p = .002, η^2_G = .01, *MSE* = .19), but the remaining two-way interaction between distinctiveness and listening condition was not significant ($F_{(2, 45)}$ < 1, p = .446, η^2_G < .01, *MSE* = .22). Finally, a small but significant three-way interaction emerged between all variables ($F_{(2, 45)}$ = 3.65, p = .034, η^2_G < .01, *MSE* = .19).

Further examination of this 3-way interaction through analysis of the simple main effects revealed a main effect of distinctiveness within each listening condition (forwards: $F_{(1, 45)}$ = 18.66, p < .001, η^2_G = .02, *MSE* = .22; nonsense: $F_{(1, 45)}$ = 37.35, p < .001, η^2_G = .21, *MSE* = .24; backwards: $F_{(1, 45)}$ = 29.75, p < .001, η^2_G = .05, *MSE* = .20) which interacted with trial type only in one condition (nonsense: $F_{(1, 45)}$ = 15.22, p < .001, η^2_G = .09, *MSE* = .21).

Pairwise comparisons in the nonsense condition tested the *a*-priori prediction of a distinctiveness advantage. These nevertheless revealed significantly greater confidence for distinctive than typical voices in both same trials ($t_{(15)} = 2.21$, $p = .043$) and different trials ($t_{(15)} = 5.80$, $p < .001$) suggesting that the three-way interaction may reflect noise in the data.

Taken together, these results were interesting in several regards. First, they provided support for the prediction that performance on a difficult voice recognition task would be facilitated by the distinctiveness of the voice. This was demonstrated in terms of sensitivity of discrimination, accuracy and in terms of metacognitive judgements of confidence in decision-making. This was most apparent in the temporally reversed condition when voice recognition became significantly impaired. As such, the present results complemented those of van Lancker and colleagues (1985) who suggested that distinctiveness may support voice recognition even under temporal reversal. However, the present results go one step further by providing an *a-priori* test, rather than a post-hoc explanation, of the importance of distinctiveness under difficult listening conditions.

This said, a subtlety emerged in the manipulation of task difficulty that had not been anticipated. Indeed, re-ordering the words to create nonsense speech had relatively little effect on accuracy of voice recognition performance overall. Moreover, the distinctiveness advantage in the ‘nonsense’ condition was significant but was relatively weak, and some evidence emerged of a shift in response bias in this condition. This was surprising as it has been assumed that voice recognition from nonsense speech would represent a more challenging task compared to the baseline condition.

In accounting for the results in the ‘nonsense’ condition, it is possible that the absence of a clear impact on performance in this condition reflected the relatively low power within the current design as a whole. Certainly, the current results would benefit from replication

using a greater number of participants. However, it may also be possible to explain the results in the 'nonsense' condition with reference to potential strategies that the participants could have adopted. In particular, it is possible that they concentrated on the common start of the clip ('The shop of ground…'), and disregarded the remainder of the nonsense phrase. This may have enabled the participants to perform well despite the increasing bizarreness in the nonsense phrase as it unfolded. It should also be noted that the nonsense phrase was re-presented with every speaker at study, and participants reported habituation to its bizarreness as the study wore on. As a consequence, accuracy remained relatively high in this condition (however, see Appendix for analysis of this point).

Additionally, it is possible that participants in the 'nonsense' condition were able to perform well because they disregarded the bizarre words entirely and instead utilized the melody contour (ups and downs) in the nonsense clip. Indeed, the fact that the speakers within this study had practiced the nonsense phrase meant that they could utter it with a near-natural cadence and intonation and these prosodic characteristics may have minimised the impact of the nonsense manipulation. By contrast, it is notable that when Goggin *et al*. (1991) generated nonsense clips by digitally cutting and reordering the voice clips, their clips did not retain a natural prosody, and a reduction in voice recognition performance was noted (see Goggin *et al.*, 1991).

In considering the importance of the melody contour within speech, it is conceivable that a rich melody contour may be considered an aspect of vocal distinctiveness. Some interesting work on processing the melody contour reveals that this is discernible by infants, non-musicians and musicians alike, suggesting that it may be extracted automatically (see Lee, Janata, Frost, Hanke & Granger, 2011). This said, there is some evidence to suggest that the processing of melody contours in music and in speech may differ, with the latter being far more coarse-grained than the former (see Zatorre & Baum, 2012). As such, a musical contour

explanation provides a potentially valuable interpretation of the surprisingly good performance in the ‘nonsense’ condition, but would benefit from further exploration. Moreover, given the surprising results in the ‘nonsense’ condition, there may be value in exploring performance under a different type of challenge, whilst still holding paralinguistic properties of the voice constant. Such conditions may be provided when listening to speech amidst noise (i.e., Sumby & Pollack, 1954) or when listening to speech at low volume. Further work on this issue may be of value in addressing the weaknesses of the current ‘nonsense’ condition.

General Discussion

The results presented here have provided an effective demonstration of a distinctiveness advantage when recognising unfamiliar voices. In Experiment 1, distinctive voices were recognised with greater sensitivity of discrimination, accuracy and confidence than their typical sounding counterparts and as such, the prediction of a distinctiveness advantage when recognising unfamiliar voices was supported. Moreover in Experiment 2, the distinctiveness advantage remained evident despite perceptually challenging listening conditions. In comparing performance across studies, it is notable that performance in the baseline (forwards) condition of Experiment 2 appeared better, and participants appeared more confident, than in Experiment 1. This most likely resulted from the reduction in delay between study and test in Experiment 2 (from 16s to 5s). As noted earlier, this change was important in reducing the likelihood of poor performance and thus floor effects in the disrupted listening conditions of Experiment 2. Nevertheless, it is appropriate to note this difference, and to refrain from drawing a direct comparison of absolute performance levels across the studies.

Taken together, these results confirmed the findings of a diverse set of previous studies (Bülthoff & Newell, 2015; Mullenix *et al.,* 2009; Sauerland *et al.,* 2013; Sørensen, 2012, van Lancker *et al.,* 1985; Yarmey, 1991). The benefit of the present results, however, was that the distinctiveness advantage was demonstrated here across a considerably larger voice set than has been utilised previously, and was demonstrated across a more standard voice matching task under both optimal and sub-optimal listening conditions.

This demonstration of a vocal distinctiveness advantage sits well with the face literature (Bartlett, Hurry & Thorley, 1984; Goldstein & Chance, 1981; Light, Kayra-Stuart & Hollander, 1979; Shepherd, Gibling & Ellis, 1991; Valentine & Bruce, 1986; Winograd, 1981) suggesting value in the application of methodologies and findings across the two domains. Similarly, the demonstration of a distinctiveness advantage can be readily accommodated within a similarity space explanation. This suggests that stimuli, be they faces (Valentine, 1991) or voices (Baumann & Belin, 2010), can be arranged in a similarity space on the basis of their properties along each of the dimensions that describe the space. Typical stimuli will naturally fall towards the centre of the space, and will be located in a relatively densely populated area with many near neighbours. By comparison, distinctive stimuli will, by definition, stand out on one or more of the dimension(s) that define the space and thus will fall towards the edge of the space where there are fewer near neighbours. The distinctiveness advantage has been accounted for as a natural consequence of the fact that distinctive stimuli have fewer near neighbours with which to be confused, and thus can be more easily matched to a (temporary) stored representation.

The present results may hold value when considering voice processing in an applied context, and it is useful to reflect briefly on this possibility. For instance, given the current results, it is perhaps tempting to conclude that police investigators may justifiably have greater confidence in earwitness recognition when the target voice sounds distinctive rather

than typical. Taken in a wider context, however, the present findings of better recognition memory for distinctive over typical voices should be tensioned with an indication of a greater risk of a 'false feeling of familiarity' when voices are distinctive (see Krix, Sauerland & Schreuder, 2017). Additionally, the performance of participants within a laboratory context may overestimate performance in more real-world settings for a host of reasons. As such, demonstration of a distinctiveness advantage under a range of ecologically valid conditions will require further empirical testing.

Perhaps of greater importance, however, the present paper invites a careful consideration of the concept of distinctiveness as it applies to voices. Indeed, this may represent a fruitful avenue for future work. If one adopts the statistical approach to define distinctiveness (such as that defined within models of similarity space, (Valentine, 1991)), then a distinctive voice is any voice that stands out for any reason relative to the set of voices under consideration. This is the approach that has been used within this paper[1]. Within this approach, it stands to reason that distinctiveness, by definition, is a *relative* rather than an *absolute* characteristic. Put another way, a voice that is distinctive due to an unusually low pitch (relative to some comparison set) will no longer be distinctive if all the comparison voices also have a low pitch.

Respecting this line of thought, vocal distinctiveness rests on a notable difference between a target voice and a set of comparison voices, but the type of difference is unspecified. This is the case when distinctiveness rests on a global and unspecified rating indicating that a voice 'stands out within a noisy environment', or when judging 'unusualness' or difficulty to recognise a voice in a group (Krix *et al.*, 2017). The strength of

[1] We differentiate here between distinctive of the voice (on the basis of vocal characteristics) and distinctiveness of the presentation of the voice (by scrambling for instance). In the current study, distinctiveness refers to *vocal* distinctiveness rather than unusualness created through non-standard presentation of an otherwise typical sounding voice.

such an approach, is that distinctiveness effects can be examined without constraining the basis for the distinctiveness ratings to what we may know or presume given our current understanding. The weakness of such an approach is that the basis of distinctiveness for each voice is ignored.

If, instead, one seeks to understand the particular characteristics that make a voice distinctive, it may be useful to consider those characteristics that we commonly use to distinguish one voice from another. Baumann and Belin (2010) identified pitch and formant characteristics when mapping their vocal similarity space. From this, it may be suggested that listeners judge a voice to be distinctive if it stands out on one or more of these dimensions. This definition sits well with the work of Foulkes and Barron (2000) and Sørensen (2012) who both explored voice processing when targets were distinctive in terms of pitch, or pitch variation. This said, Baumann and Belin's (2010) use of vowel sounds ('a', 'i' and 'u') as a basis for determining their voice space may have ignored other more prosodic vocal features which emerge as speech unfolds over time. Accordingly, distinctiveness has at times been operationalised through other characteristics such as accent or language (Bülthoff & Newell, 2015). Still other studies have suggested that voices may be described using rich descriptors including nasality, speed, intonation, volume, tremor, pauses (see Yarmey, 1991 for a set of descriptive ratings used). These provide an expanded set of characteristics which could serve as the basis for distinctiveness.

Adopting this line of thought would enable researchers to potentially generate distinctive versions of voices by using voices that vary naturally on some specified dimension (such as pitch or speed) or by using synthetic voices which have been manipulated to vary along a specified dimension. The effectiveness of such a manipulation will necessarily depend upon a host of factors including the extent of manipulation, the just-noticeable

differences, and the initial vocal characteristics; for a voice that is already relatively high in pitch, a further manipulation of pitch may have little impact.

One potentially promising way forward is to make use of recent developments in voice morphing software. This is capable of generating caricatures, and anti-caricatures of voices compared to some norm or reference point (Kawahara & Matsui, 2003; Schweinberger, Kawahara, Simpson, Skuk & Zäske, 2013). As such, a caricature may be considered to represent a distinctive version of a given voice, and an anti-caricature may represent a typical version of the given voice, relative to a norm. In this way, distinctiveness could be varied *within* the voice, allowing for highly controlled tests of distinctiveness effects (see Rhodes, Brennan & Carey, 1987, for a similar approach in the area of face processing). Such an approach would not enable the identification of the individual characteristics that make a voice distinctive, but it would enable the controlled manipulation of distinctiveness by exaggerating the characteristics that make *each* individual voice stand out. Future work along these lines would be valuable in providing a sophisticated test of distinctiveness effects in voices.

Summary

In summary, the present paper has reported on the results of two experiments which explored a distinctiveness advantage when recognising unfamiliar voices. The results of Experiment 1 confirmed that distinctive voices were processed with greater sensitivity of discrimination, accuracy and confidence compared to typical voices. The results of Experiment 2 extended these findings by confirming a distinctiveness advantage even under difficult listening conditions involving nonsense phrases, and backwards speech. These results sit well alongside the considerable body of research suggesting a facial distinctiveness advantage. Moreover, they can be readily explained by drawing on the concept of a similarity

space. Nevertheless, the use of generalised ratings of distinctiveness here did not address the issue of what makes a voice distinctive, making this an exciting avenue for future research.

References

Barsics, C., & Brédart, S. (2011). Recalling episodic information about personally known faces and voices. *Consciousness and Cognition, 20*, 303-308. doi:10.1016/j.concog.2010.03.008

Barsics, C., & Brédart, S. (2012a). Recalling semantic information about newly learned faces and voices. *Memory, 20 (5)*, 527-534. doi:10.1080/09658211.2012.683012

Barsics, C., & Brédart, S. (2012b). Access to semantic and episodic information from faces and voices: Does distinctiveness matter? *Journal of Cognitive Psychology, 24(7)*, 789-795. doi:10.1080/20445911.2012.692672

Bartlett, J.C., Hurry, S., & Thorley, W. (1984). Typicality and familiarity of faces. *Memory and Cognition, 4*, 373-379.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychological Research, 74*, 110-120.

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

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Science. 8 (3*), 129-135.

Brédart, S., & Barsics, C. (2012). Recalling semantic and episodic information from faces and voices : A face advantage. *Current Directions in Psychological Science, 21 (6)*, 378-381.

Brédart, S., Barsics, C., & Hanley, R. (2009). Recalling semantic information about personally known faces and voices. *European Journal of Cognitive Psychology, 21 (7),* 1013-1021. doi:10.1080/09541440802591821

Bülthoff, I., & Newell, F.N. (2015). Distinctive voices enhance the visual recognition of unfamiliar faces. *Cognition, 137*, 9-21. doi:10.1016/j.cognition.2014.12.006

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

Ellis, H.D., Jones, D.M., & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. *British Journal of Psychology, 88*, 143-156.

Foulkes, P., & Barron, A. (2000). Telephone speaker recognition amongst members of close social network. *Forensic Linguistics: The International Journal of Speech, Language and the Law*, 7, 180-198.

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.

Goldstein, A.G., & Chance, J.E. (1981). Laboratory studies of face recognition. In G.M. Davies, H.D. Ellis & J.W. Shepherd (Eds.). *Perceiving and Remembering Faces, pp. 81-104. London: Academic Press.*

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

Hanley, J.R., & Damjanovic, L. (2009). It is more difficult to retrieve a familiar person's name and occupation from their voice than from their blurred face. *Memory, 17*, 830-839.

Hanley, J.R., Smith S.T., & Hadfield, J. (1998). I recognize you but can't place you. An investigation of familiar-only experiences during tests of voice and face recognition. *Quarterly Journal of Experimental Psychology, 51A (1), 179-195.*

Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. *Proceedings of ICASSP*, 256-259.

Krix, A.C., Sauerland, M., & Schreuder, M.J. (2017). Masking the identities of celebrities and personally familiar individuals: Effects on visual and auditory recognition performance. *Perception, 46(10),* 1133-1150. doi: 10.1177/0301006617710621

Lee, Y-S., Janata, P., Frost, C., Hanke, M., & Granger, R. (2011). Investigation of melodic contour processing in the brain using multivariate pattern-based fMRI. *NeuroImage*, *57(1),* 293-300. doi:10.1016/j.neuroimage.2011.02.006

Light, L.L., Kayra-Stuart, F., & Hollander, S. (1979). Recognition memory for typical and unusual faces. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 212-228.

Mullenix, J.W., Ross, A., Smith, C., Kuykendall, D., Conard, J., & Barb, S. (2009). Typicality effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychology, 25(1),* 29-34. doi:10.1002/acp.1635

Orchard, T., & 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.

Rhodes, G., Brennan, S., & Carey, S. (1987). Identification and ratings of caricatures: Implications for mental representations of faces. *Cognitive Psychology, 19,* 473-497.

Sauerland, M., Sagana, A., & Otgaar, H. (2013). Theoretical and legal issues related to choice blindness for voices. *Legal and Criminological Psychology, 18,* 371-381. doi:10.1111/j.2044-8333.2012.02049.x

Schweinberger, S.R., & Burton, A.M. (2011). Person perception 25 years after Bruce and Young (1986): An Introduction. *British Journal of Psychology, 102,* 695-703.

Schweinberger, S.R., Herholz, A., & Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition. *Quarterly Journal of Experimental Psychology, 50A(3),* 198-517.

Schweinberger, S.R., Kawahara, H., Simpson, A.P., Skuk, V.G., & Zäske, R. (2013). Speaker perception. *WIREs Cognitive Science, 5*, 15-25. doi:10.1002.wcs.1261

Shepherd, J.W., Gibling, F., & Ellis, H.D. (1991). The effects of distinctiveness, presentation time and delay on face recognition. *European Journal of Cognitive Psychology, Special Issue: Face Recognition, 3*, 137-145.

Skuk, V.G., & Schweinberger, S.R. (2013). Gender differences in familiar voice identification. *Hearing Research,* 296, 131-140.

Sørensen, M.H. (2012). Voice line-ups: speakers' F0 values influence the reliability of voice recognitions. *International Journal of Speech Language and the Law, 19(2),* 145-158. doi:10.1558/ijsll.v19i2.145

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

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

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

Stevenage, S.V., Neil, G.J., & Hamlin, I. (2014). When the face fits: Recognition of celebrities from matching and mismatching faces and voices. *Memory, 22(3),* 284-294. doi:10.1080/09658211.2013.781654

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of Acoustic Society of America, 26*, 212-215.

Tanaka, J.W., & Farah, M.J. (1993). Parts and wholes in face recognition. *Quarterly Journal of Experimental Psychology, 46A*, 225-245.

Valentine, T. (1991). A unified account of the effects of distinctiveness, inversion and race in face recognition. *Quarterly Journal of Experimental Psychology, 43A*, 161-204. doi:http://dx.doi.org/10.1080/14640849108400966

Valentine, T., & Bruce, V. (1986). Recognising familiar faces: The role of distinctiveness and familiarity. *Canadian Journal of Psychology, 40*, 300-305.

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.

Winograd, E. (1981). Elaboration and distinctiveness in memory for faces. *Journal of Experimental Psychology: Human Learning and Memory, 7*, 181-190.

Yarmey, A.D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of the Forensic Sciences Society, 31(4)*, 421-428.

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

Table 1: Mean sensitivity of discrimination (*d'*) and response bias (*C*), together with accuracy and self-rated confidence on 'same' and 'different' trials on a same/different voice matching task with distinctive and typical voices in Experiment 1 (standard deviation in parentheses).

	Distinctive	Typical
Sensitivity of Discrimination (*d'*)	2.66 (1.24)	1.56 (1.26)
Response Bias (*C*)	-.07 (.59)	-.10 (.69)
Accuracy on 'Same' Trials	.87 (.12)	.74 (.22)
Accuracy on 'Different' Trials	.82 (.19)	.70 (.22)
Confidence on 'Same' Trials (/7)	5.28 (.82)	4.57 (1.00)
Confidence on 'Different' Trials (/7)	5.05(1.14)	4.38 (1.19)

Table 2: Mean sensitivity of discrimination (*d'*) and response bias (*C*), together with accuracy and self-rated confidence (with standard deviation) when recognising distinctive and typical voices under forwards, nonsense and backwards listening (Experiment 2).

	Distinctive	Typical
FORWARDS		
Sensitivity of Discrimination (*d'*)	3.22 (1.23)	2.70 (1.08)
Response Bias (*C*)	-.04 (.36)	-.19 (.61)
Accuracy on 'Same' Trials	.91 (.10)	.88 (.17)
Accuracy on 'Different' Trials	.91 (.10)	.83 (.13)
Confidence on 'Same' Trials (/7)	6.10 (.53)	5.80 (.63)
Confidence on 'Different' Trials (/7)	6.20 (.81)	5.49 (.45)
NONSENSE		
Sensitivity of Discrimination (*d'*)	3.29 (1.08)	2.77 (.58)
Response Bias (*C*)	.18 (.47)	-.39 (.63)
Accuracy on 'Same' Trials	.88 (.13)	.92 (.12)
Accuracy on 'Different' Trials	.94 (.09)	.80 (.13)
Confidence on 'Same' Trials (/7)	6.01 (.78)	5.72 (.67)
Confidence on 'Different' Trials (/7)	6.07 (.78)	4.92 (.83)
BACKWARDS		

Sensitivity of Discrimination (*d'*)	1.32 (.73)	.43 (.70)
Response Bias (*C*)	-.08 (.36)	-.11 (.46)
Accuracy on 'Same' Trials	.76 (.12)	.61 (.22)
Accuracy on 'Different' Trials	.70 (.21)	.54 (.19)
Confidence on 'Same' Trials (/7)	3.50 (1.50)	2.87 (1.47)
Confidence on 'Different' Trials (/7)	3.70 (1.61)	3.05 (1.53)

Appendix 1:

In examining the possibility that participants habituated to the nonsense phrase across the course of the experiment, an analysis based on the responses during the first half of the experiment *only* suggested that performance remained high when listening to nonsense messages. In fact, there was no significant difference in *d'* for distinctive voices ($t_{(15)} < 1$, $p = .675$, or for typical voices ($t_{(15)} < 1$, $p = .752$) when comparing performance across the two halves of the experiment.

Furthermore, if the data from the first half of the experiment were considered for those in the 'nonsense' group alongside all the data from those in the 'forwards' group and the 'backwards' group, the ANOVA replicated all reported findings. There was a main effect of distinctiveness ($F_{(1, 45)} = 11.45$, $p = .001$, $\eta^2_G = .11$, $MSE = 44.67$) indicating better performance with distinctive than typical voices overall. There was again, a main effect of listening condition ($F_{(2, 45)} = 54.07$, $p < .001$, $\eta^2_G = .56$, $MSE = 48.94$). Again, the interaction between distinctiveness and listening condition was not significant ($F_{(2, 45)} < 1$, $p = .761$, $\eta^2_G < .01$, $MSE = 44.67$).

Importantly for the current discussion, repeated post-hoc contrasts were used to examine the main effect of listening condition. As in the full analysis, these again revealed equivalent performance when comparing the 'forwards' condition to the 'nonsense' condition ($p = .087$), but a significant reduction in performance between the 'nonsense' condition and the 'backwards' condition ($p < .001$). Consequently, these results suggest, that a full account of the maintenance of performance from 'forwards' to 'nonsense' conditions may be more complex than a simple habituation effect.

Acknowledgements

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC) Grant (EP/J004995/1 SID: An Exploration of SuperIdentity) awarded to the first author. Colleagues on this grant are thanked for helpful contributions to the current work.