DOI: 10.17469/O2108AISV000002

KIRSTY MCDOUGALL

Ear-catching versus eye-catching? Some developments and current challenges in earwitness identification evidence

While earwitness identification evidence collected through a voice parade can provide pivotal evidence in a criminal case, there remain many unanswered questions regarding the psychological and phonetic processes involved in this type of identification. The voice parade procedure currently used in England and Wales was developed analogously to the procedure used for eyewitness identification, yet recent research shows that, while there are some similarities between the processing of faces and voices, considerable differences exist. Research is needed to determine the optimal settings of the relevant variables in a voice parade procedure and how best to select foils for auditory comparison. Recent findings from the *IVIP* 'Improving Voice Identification Procedures' project are presented and their implications for voice parade construction discussed.

Keywords: earwitness recognition, earwitness evidence, voice parades, voice line-ups, perceived voice similarity.

1. *Introduction*

Earwitness identification evidence may be called on if a perpetrator's voice has been heard at the scene of a crime, but not recorded. If the witness received sufficient exposure to the voice, earwitness evidence may be collected through a voice parade procedure. Less well known than its visual counterpart, a voice parade is conducted using a similar format to a visual identity parade: the witness is asked whether he or she can pick out the voice of the speaker heard at the crime scene from a line-up of recorded speech samples which includes the suspect's voice and a number of 'foil' voices. Earwitness identification obtained through a voice parade can constitute crucial evidence, yet there remain many unanswered questions about the phonetic and psychological underpinnings of this type of identification and about the optimal way to collect such evidence. The present paper will highlight some of these questions, particularly with respect to the parallels often drawn between visual and auditory identification of individuals, and the importance of researching and understanding differences between the two modalities. It will commence with an outline of the current voice parade procedure in use in England and Wales which forms a backdrop for the research to be presented. A brief review of developments in psychological research showing similarities and differences between the human processing of faces and voices will be provided. This will be followed by a selection of results from two studies being undertaken in the *IVIP* 'Improving Voice

Identification Procedures' project[1] exploring these issues. The first study investigates listeners' perception of voice similarity within and between different accents. The second study is an exploration of the effect of the duration of voice parade speech samples on earwitness recognition performance. Implications for voice parade construction and directions for further research will be discussed.

2. *Current voice parade procedure in England and Wales*

In England and Wales, the recommended procedure for conducting voice parades is published in a Home Office Circular entitled 'Advice on the Use of Voice Identification Parades' (Home Office, 2003). This procedure was developed by the then Detective Sergeant John McFarlane of the Metropolitan Police, in consultation with Professor Francis Nolan of the University of Cambridge, through their work in bringing a case to the Central Criminal Court in 2002 (R v. Khan and Bains, see Nolan, 2003). McFarlane's guidelines for conducting a voice parade were devised on the basis of research available at the time such as Broeders & Rietveld (1995), Hollien (1996) and Broeders (1996), in conjunction with the existing police procedure for visual identification parades (see Code D of the Police and Criminal Evidence Act (PACE) 1984, 'Code of Practice for the Identification of Persons by Police Officers')[2]. Use of the Home Office (2003) procedure is not mandatory, but recommended.

In the United Kingdom, all suspect interviews conducted by the police are recorded. When the possibility of a voice parade arises, the forensic phonetician is usually provided with a copy of the suspect's police interview recording to assess the suitability of the individual's voice for identification via a parade. If the suspect's voice is particularly unusual in terms of accent or other idiosyncrasy, the phonetician will recommend that a voice parade is not undertaken. The phonetician also checks the quality of the speech recording and whether it provides sufficient speech material to be edited to form the suspect sample for the voice parade.

To prepare the suspect's voice parade sample the phonetician extracts short stretches of speech consisting of self-contained words or utterances, e.g. *yeah*, *I don't know*, *about half past three*, *round the corner*, etc., from the interview recording. Longer utterances and utterances revealing crime-related information are avoided. The extracts are digitally spliced into a single sound file to form a 'collage' of speech characteristic of the individual, 60 seconds in duration. The order of the extracts is jumbled so as not to allow any sense of narrative to develop.

Police interview recordings are also used for the foil speech samples in the voice parade so that all samples contain spontaneous speech in the same speaking

[1] 'Improving Voice Identification Procedures' (*IVIP*) is an interdisciplinary project on earwitness evidence funded by the UK Economic and Social Research Council (Grant Reference: ES/S015965/1) bringing together researchers in phonetics, psychology, sociolinguistics, criminology and law. https://www.phonetics.mmll.cam.ac.uk/ivip/

[2] https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/903812/pace-code-d-2017.pdf

style. Eight foil voices are required to form a parade of nine voices including the suspect's. To select the foil voices, the identification officer is asked to provide the phonetician with at least 20 recordings of police interviews from unconnected cases in which the interviewees are "persons of similar age and ethnic, regional and social background" as the suspect (Home Office, 2003: point 9) from which to select the eight foils. The guidelines state that the foil samples "must be examined to ensure that the accent, inflection, pitch, tone and speed of the speech used provides a fair example for comparison against the suspect" (point 15). The phonetician undertakes a rigorous phonetic screening of the candidate foil recordings to select a set of eight foils.

Although not described in the Home Office guidelines, when preparing a voice parade, the present author next conducts a 'perceptual distance test' (see de Jong-Lendle, Nolan, McDougall & Hudson, 2015; McDougall, 2013) to check whether the voices of the suspect and chosen foils are all roughly equally different from each other (Rietveld, Broeders, 1991). This test involves pairwise comparisons in which naïve listeners rate on a nine-point Likert scale the perceived distance between all pairings of the voices to be used in the voice parade. The ratings are interpreted using multidimensional scaling (MDS) (Schiffman, Lance Reynolds & Young, 1981). The results need to show that no particular speaker stands out as sounding markedly different from the other eight speakers and that the suspect is appropriately spaced among the foils, or else the foil selection should be reconsidered.

The Home Office guidelines recommend that a 'mock witness test' is undertaken to test that the voice parade is fair to the suspect. This involves a number of test runs of the voice parade being conducted with naïve listeners. Each listener is given a brief synopsis of the case, then listens to the voice samples. Listeners are asked to estimate how likely or unlikely they think it is that each speaker was being interviewed about the crime in question using a numerical scale, e.g. 1 to 9 where 1 = 'most likely to be about the crime' and 9 = 'least likely to be about the crime'. A mean rating is calculated for each voice sample. Providing none of the samples are given an extreme mean rating and particularly not the suspect's sample, the parade is considered to be fair on this test.

Three random orderings of the samples are chosen for the final parade materials. Three *PowerPoint* files (rather than the video cassettes of the 2003 guidelines) are prepared with the samples labelled A, B, C, D, E, F, G, H, J, one letter per slide, to accompany each voice sample. An additional slide which contains nine buttons for the nine speech samples is included to enable the officer conducting the parade to replay any of the samples at the end, if the witness requests this.

3. *Some practical and theoretical challenges*

While voice parades have been successfully implemented for a number of cases in the United Kingdom using the Home Office guidelines, the practical and resource requirements of the procedure are very time-consuming and expensive. The procedure is labour-intensive, requiring considerable technical input from an expert phonetician to prepare a tailor-made voice parade to suit the individual suspect. Finding sufficient recordings of appropriate foil voices can be difficult, both in terms of locating the number of speakers required and the amount of material needed per speaker.

Robson (2017) reports the results of a Freedom of Information investigation he conducted on the use of voice parades in England and Wales for the period 2005-2015. All 43 police forces were approached, and all but two responded. Only four forces responded that they had used the Home Office procedure. Another four forces noted that they had either considered using the procedure in particular cases which had not eventuated or that they would consider using it should such a case arise. 21 forces indicated that they had not conducted any voice parades or that they did not hold any data to suggest that a voice parade had taken place. Five further forces responded that they did not retain retrievable data on voice parades. Seven forces indicated that they did not conduct voice parades as a matter of force policy. The five forces remaining answered with 'words to the effect of "we do not undertake this process"' (2017: 46).

Thus practice in accepting and adopting the procedure is highly variable and inconsistent across England and Wales. It is not clear whether this lack of engagement with the collection of earwitness evidence is due to a lack of knowledge, or financial or other practical barriers, but it would appear likely that a combination of these is at play.

In addition to the various practical issues which may be limiting a more widespread adoption of the Home Office voice parade procedure across England and Wales, there exist gaps in fundamental research understanding of earwitness behaviour which need addressing. Many of these gaps relate to differences between the visual and auditory modalities for person perception and the fact that person identification procedures may need to be structured in different ways when dealing with auditory as opposed to visual recognition.

4. *Eyes versus ears*

There is a traditional assumption that earwitness identification of individuals may operate along similar lines to eyewitness identification. An extensive body of literature on the workings of eyewitness memory has been developing since early last century (see e.g. Lindsay, Ross, Read & Toglia, 2007 for an overview) which has shaped the practice and development of visual identification procedures. Markedly less research is available on earwitness memory, presumably due to the dominance of the visual modality, and there are many aspects of the detail of earwitness behaviour which have not yet been empirically explored. Developing an improved understanding of similarities and differences between visual and auditory processing of person-identifying information is particularly crucial. This can then inform how identification procedures designed for the visual modality should be optimally adjusted for application in the earwitness context.

Research in recent decades has shed much new light on the perception of individuals by eye and ear. A number of studies have suggested that the recognition of faces and voices involves separate, yet parallel pathways in the brain (e.g. Belin, Fecteau & Bedard, 2004; Belin, Bestelmeyer, Latinus & Watson, 2011; Ellis, Jones & Mosdell, 1997). Belin et al. (2004) propose an 'auditory face' model for processing voices, as an extension of Bruce and Young's (1986) model for the visual processing involved in recognising faces. Belin et al.'s model involves three functionally separate systems, which nonetheless do interact. These are (Box 3, p. 131):

(i) analysis of speech information,
(ii) analysis of vocal affective information,
(iii) analysis of vocal identity.

Evidence from neuroimaging can be drawn on in support of this model. For example, Imaizumi, Mori, Kiritani, Hosoi & Tonoike (1997) used positron emission tomography (PET) to observe cerebral activity during tasks requiring decisions concerning speaker familiarity, emotion identification and vowel/consonant description, and found that several cortical regions showed greatest activation when undertaking the speaker familiarity task. A functional magnetic resonance imaging (fMRI) study by von Kriegstein, Eger, Kleinschmidt & Giraud (2003) in which German-speaker listeners completed listening tasks either focussing on speaker identity or on the linguistic content of sentences showed different patterns of brain area activation for the different tasks.

Clinical research into Phonagnosia and Prosopagnosia also provides evidence for a model involving functionally separate systems. Phonagnosia is the term applied to a person with damage observed in the right anterior temporal lobe or the right superior gyrus region, who experiences an inability to recognise individuals from their voices yet is able to recognise them visually and perform name retrieval satisfactorily (Garrido, Eisner, McGettigan, Stewart, Sauter, Hanley, Schweinberger, Warren & Duchaine, 2009; Hailstone, Crutch, Vestergaard, Patterson & Warren, 2010). Prosopagnosia describes the condition in which a person is unable to recognise familiar individuals by their faces but is nevertheless able to make a recognition by voice (Barton, 2008; Neuner, Schweinberger, 2000). Thus the existence of these two conditions in which patients have selective impairments to their ability to recognise faces or voices offers further support for a model involving separate neural processing pathways for faces and voices.

The notion that an 'auditory face' architecture may operate in parallel with a visual face processing architecture has seen a growth in research interest, as is described, for example, in Brédart, Barsics (2012) and Yovel, Belin (2013). Building on this work is an increasing number of behavioural and neuropsychological findings emphasising interactions between the visual and auditory processing pathways, including in the context of person recognition (see Campanella, Belin, 2007; Stevenage, Neil, 2014). For example, Sheffert, Olson (2004) present data showing participants achieving superior identification of speakers when audio-visual information as opposed to just auditory information is provided. von Kriegstein, Kleinschmidt, Sterzer & Giraud (2005) provide functional neuroimaging evidence showing a relationship between visual face and auditory voice regions of the brain during a speaker recognition task.

Young, Frühholz & Schweinberger (2020) propose a revised model of face and voice perception, in the light of the growing body of research showing both independent processing streams for voices and faces, and contributions of multimodal regions. Their article provides a summary of the main regions of the brain which have been demonstrated to be associated with face and voice perception (see Figure 1, p. 399, and Box 2 and references therein, p. 401). It explains that distinct cortical brain areas show strong unimodal responses to voices, while others give strong unimodal responses to faces, each providing a 'basic structural analysis' of vocal and facial input (2020: 401). Areas associated with unimodal face perception have been shown to exhibit more regional functional specificity than those associated with unimodal voice perception. There are further regions of the brain which have been demonstrated to respond to both face and voice information, and different regions again in which evidence suggests that post-perceptual processing is undertaken. The model of face and voice perception proposed by Young et al. (2020) emphasises differences between the two modalities as well as commonalities. These authors argue that it is necessary to take into account the differing contextual demands required in everyday tasks involving faces and voices in order to explain the neuropsychological patterning exhibited for different activities related to face and voice perception.

In the context of person recognition by ear, it is thus important to recognise that while there is increasing evidence of functional interaction between face and voice processing pathways, the two modalities are anatomically and neurologically distinct. This means that earwitness identification procedures must be devised separately from their eyewitness counterparts and empirically tested in their own right.

5. *Study 1: Perceived voice similarity within and between different accents*

For an identity parade to be fair to the suspect, the foils must be chosen in such a way that the suspect does not stand out from the rest of the group. For visual identity parades, the PACE Code D instructions require the foils to be "at least eight other people who, so far as possible, resemble the suspect in age, general appearance and position in life" (Annex A(a) point 2). This resemblance can be achieved by choosing individuals with matching characteristics such as skin, hair and eye colours, as well as the presence/absence of facial hair, etc. In the earwitness context, however, it is not clear what profile of voice characteristics should be used to guide the determination of 'resemblance' to a suspect for a voice parade. After potential foil speakers for a voice parade have been phonetically screened by a phonetician, a perceptual distance test (see §2) can be used to verify that the set of foils chosen provides a fair comparison as judged by a group of naïve listeners. Yet this is a time-consuming, laborious process in a situation where, to minimise decay of the witness's memory, time is of the essence. Developing an understanding of the phonetic underpinnings of perceived voice similarity could lead to more efficient methods for voice parade foil selection.

Listeners' perception of voice similarity is not well understood in scientific terms and there is relatively little research into the role played by acoustic features of speech. An early study by Walden, Montgomery, Gibeily, Prosek & Schwartz (1978) found perceived voice similarity was correlated with f0 and word duration to some extent, for a single-word utterance. Remez, Fellowes & Nagel (2007) argue that formant dynamic information is relevant for perceived voice similarity, but their results are based on data from a heterogeneous group of ten speakers including males and females and American and British English dialects.

A study by Baumann & Belin (2010) elicited same/different speaker judgements on isolated Canadian French vowels and applied MDS to the data. For female speakers, the highest correlations were found between perceptual dimensions and f0 and F1. For males, highest correlations were seen between perceptual dimensions and the mean difference between F4 and F5, perhaps surprisingly given the difficulty attendant in measuring higher formants.

A study using spontaneous speech stimuli, a precursor to the present one, is described in Nolan, McDougall & Hudson (2011). Crucially, this study uses a homogeneous group of speakers, i.e. speakers of the same sex, age and accent background (male, 18-25 years, Standard Southern British English [SSBE]), so that perception of personal voice similarity can be examined with the effects of linguistic variation being relatively controlled. Listeners' ratings of voice similarity were subjected to MDS to produce pseudo-perceptual dimensions which exhibited correlations with, in order of importance, long-term f0, and F3, F2, F1 frequencies of a 'global' mean across six vowel types.

The present study extends this work to examine perceived voice similarity in multiple groups of speakers controlled for demographic background, within and across accents, and using a wider range of acoustic features. Six groups of speakers from a variety of accents are investigated to explore the consistency of patterns found within a given accent and whether these apply across different accents.

5.1 Database and speaker selection

The experiment was designed to collect listener judgements of voice similarity for three separate groups of 15 SSBE speakers, and for a group of 15 speakers from each of three accents of English spoken in Yorkshire: York, Bradford and Wakefield. Three groups of speakers of the same accent (SSBE) were chosen to enable assessment of within-accent variability, alongside the variation between SSBE and each of the Yorkshire accents. Speech data for the stimuli were extracted from three databases: *DyViS* (Nolan, McDougall, de Jong & Hudson, 2009) *YorViS* (McDougall, Duckworth & Hudson, 2015) and *WYRED* (Gold, Ross & Earnshaw, 2018). These databases all use the elicitation techniques developed for the *DyViS* database and provide spontaneous speech material for male speakers of the same age group (18-25 years for *DyViS* and *YorViS*; 18-30 years for *WYRED*). Each group of 15 speakers was chosen randomly from the appropriate database, with any speaker who impressionistically sounded particularly unusual being discarded from the selection. The speakers selected are shown in Table 1.

Table 1 - *Speaker groups used in the experiment (speaker numbers from each database)*

Group	**Database**	**Speakers**
SSBE1	*DyViS*	25, 28, 39, 53, 56, 60, 62, 65, 88, 95, 106, 111, 112, 115, 118
SSBE2	*DyViS*	1, 2, 4, 11, 21, 23, 31, 32, 35, 37, 47, 50, 76, 87, 113
SSBE3	*DyViS*	6, 19, 30, 40, 46, 54, 58, 68, 69, 75, 80, 81, 96, 99, 107
York	*YorViS*	1, 2, 4, 7, 8, 10, 11, 12, 16, 17, 18, 19, 20, 21, 22
Bradford	*WYRED*	72, 132, 135, 147, 156, 157, 167, 170, 174, 175, 176, 185, 187, 189, 191
Wakefield	*WYRED*	103, 111, 112, 127, 131, 138, 141, 143, 145, 146, 152, 158, 164, 166, 178

5.2 Stimuli and experimental set-up

Stimuli were constructed from the telephone call task (Task 2) in each of the speech databases. This is a telephone conversation between the participant in the role of 'suspect' and his 'accomplice' (a researcher), in which they discuss the detail of the police interview that the participant has just undertaken, in order to ensure that the accomplice does not provide conflicting information if he is also questioned. For each 'suspect' participant, two speech stimuli of approximately three seconds duration were created, the first (U1) containing speech in which the participant denies knowing a man called Robert Freeman, and the second (U2) involving the participant denying having visited the Yewtree Reservoir on Wednesday evening.

Within each group of 15 speakers, each speaker was paired with himself and all other speakers to form 120 pairings. A stimulus was prepared for each pairing, containing a randomly assigned U1 and U2, separated by a silence of one second. The order in which the two utterances appeared to listeners was randomly determined, as was the ordering of the stimuli. The experiment required listeners to rate the (dis) similarity of all pairings of speakers within one of the six groups, using a Likert scale.

Listener responses for the *DyViS* 1 and *YorViS* groups were collected in person using the 'ExperimentMFC' (Multiple Forced Choice) facility in *Praat* (Boersma, Weenink, 1992-2021). The experiment was conducted in a silent sound-treated room, with participants listening via headphones. The responses to the *DyViS* 1 stimuli were part of an earlier experiment in which judgements were also made on telephone-recorded utterances (see Nolanet al., 2011; Nolan, McDougall & Hudson, 2013); analysis of responses to studio-recorded stimuli only are presented here. Responses to the *YorViS* stimuli were also collected for an earlier experiment (McDougall, Hudson & Atkinson, 2014).

Responses to the *DyViS* 2 and *DyViS* 3 stimuli were collected in person using the open source software *OpenSesame* (Mathôt, Schreij & Theeuwes, 2012). Participants listened to stimuli over headphones in quiet laboratory rooms.

Due to the Covid-19 outbreak, the *WYRED* 1 and *WYRED* 2 stimuli sets were presented to participants online using the web-based platform *Gorilla* (Anwyl-Irvine, Massonnié, Flitton, Kirkham & Evershed, 2020). These participants undertook a test to check that they were using headphones or earphones (Woods, Siegel, Traer & McDermott, 2017) at the start of the experiment. Participants were allowed to take this test a second time if they did not pass it the first time.

In both the in-person and online environments, each listener was asked to judge the degree of similarity of the voices in each voice pairing, taking into account voice quality and accent, but ignoring the meaningful content of the speech. To familiarise them with the experimental set-up, listeners were given a preliminary test containing several example trials before the main test. For each stimulus pair, the screen displayed the question 'How similar are these voices?' and listeners were required to click a response on a Likert scale from 1 (very similar) to 9 (very different) before moving on to the next stimulus pair. Listeners were asked to give an immediate reaction and not agonise over comparisons, but the timing of each decision was under their control.

5.3 Listeners

120 listeners (20 per speaker group) took part in the experiment. They were recruited at the University of Cambridge, Nottingham Trent University and via the experimental participant recruitment website *Prolific* (https://www.prolific.co/). Listeners were first-language English speakers, and had been born in and lived most of their pre-18 lives in England. They were aged 18-40 years and self-reported no hearing loss or hearing difficulties. The listener group was approximately half male and half female.

5.4 Acoustic Analysis

5.4.1 Long-term fundamental frequency

Long-term fundamental frequency (f0) measures were calculated for each speaker using the 6 seconds of speech provided by the two experimental stimuli per speaker. Periods of silence and low threshold noise were removed from the files, along with any intrusive noises (coughs, furniture noises, etc.), before running the long-term pitch analysis *Praat* script to generate f0 statistics for each speaker.

5.4.2 Long-Term Formant analysis

Long-Term Formant (LTF) analysis can be used to capture a speaker's overall formant behaviour, by focussing on the long-term tendencies of each main formant rather than looking at individual vowel categories (Moos, 2010; Nolan, Grigoras, 2005). Frame-by-frame analyses through the voiced sections of a sample of speech are made using a formant tracker. In the present experiment, LTF analyses were carried out for each speaker using the Task 2 recordings. Formant-bearing speech material was manually segmented in a *Praat* TextGrid until 30 seconds net speech had been compiled for each speaker. This process involved a combined auditory and acoustic approach, inspecting the spectrogram while reviewing the material auditorily. Only speech material with a clear and visible formant structure for the first three formants was selected. Approximants were included, and laterals, nasals and speech exhibiting strong nasality were excluded. Vowels were excluded if they were produced with a very high pitch such that harmonics rather than formants were visible. Filled pauses were included if they were vocalic. The segmented material was subjected to formant analysis using the Snack Sound Toolkit (Sjölander, 1997). Four tracked formants were obtained for each frame in the material with an LPC order of 12, a maximum analysis frequency of 5000 Hz, a 20 ms frame length, a 10 ms frame advance and the remaining settings at their default values. This analysis achieved stable profiles for the first four formants for all but five speakers whose results have not been included (*DyViS* 1: 53; *DyViS* 2: 113; *DyViS* 3: 107; *WYRED* 1: 132, 156). The mean value of each of F1-F4 for each speaker was calculated.

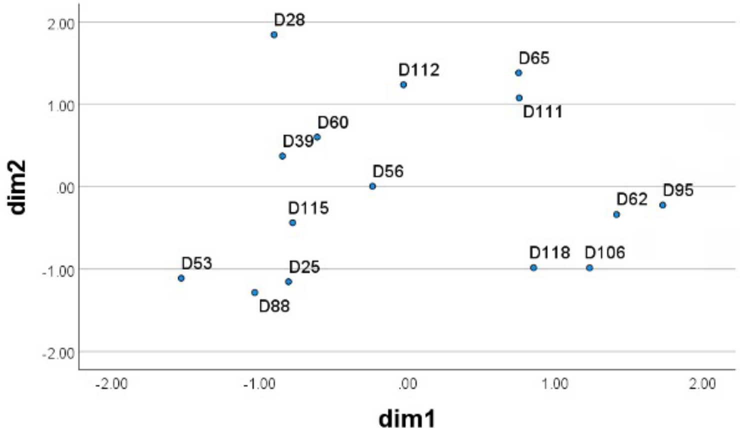
5.4.3 Articulation rate

Articulation rate (AR) was measured for each speaker using speech material from Task 2 following Jessen's (2007) procedure for calculating 'global' AR. For each speaker, 30 'memory stretches' of 5-20 syllables were analysed, with syllables determined auditorily. The syllable count for a particular utterance was defined as the actual number of syllables heard by the analyst, not the 'canonical' number that might be given for the entry for the same word(s) in a pronunciation dictionary. A 'memory stretch' was defined as a portion of fluent speech that can be retained in short-term memory of the analyst in order to count the number of phonetic syllables present (Jessen, 2007: 54). Global AR was determined by taking the mean AR across memory stretches.

5.5 Statistical Analysis

Listeners' similarity judgments on the voice pairs were subjected to MDS using INDSCAL (Individual Differences Euclidean Distance Model) in *SPSS*, with a separate analysis for each 15-speaker group. For each group, the analysis with five perceptual dimensions was chosen according to Giguère's (2006) guideline thresholds for stress (*DyViS* 1: stress = 0.174, R^2 = 0.277; *DyViS* 2: stress = 0.177, R^2 = 0.171; *DyViS* 3: stress = 0.181, R^2 = 0.145; *YorViS*: stress = 0.183, R^2 = 0.166; *WYRED* 1: stress = 0.181, R^2 = 0.210; *WYRED* 2: stress = 0.198, R^2 = 0.277). Each speaker was thus characterised along five pseudo-perceptual dimensions within a perceptual space for his group of speakers with a set of five coordinates (dim1, dim2, dim3, dim4, dim5). As an example, Figure 1 gives a plot of the first two dimensions from the analysis for the SSBE 1 group, showing the 15 speakers' locations along these dimensions. Speakers who appear relatively close on this plot were judged to be more similar-sounding (e.g. D39 and D60), while speakers who are further apart were judged less similar (e.g. D53 and D95). Each dimension accounts for successively less variance, so the lower the dimension the greater the amount of variance explained.

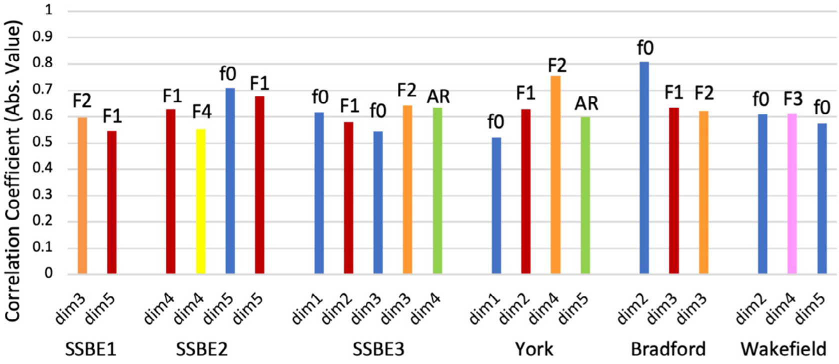
Figure 1 - *Plot of the first two dimensions (of five) produced by the MDS analysis for the SSBE 1 group, showing the relative positions among the 15 speakers*



5.6 Results

Pearson's formula was used to test the extent of correlation between the MDS dimensions from the voice similarity ratings and the set of acoustic features measured (f0, LTF1-4, AR) for each speaker group. The absolute values of correlation coefficients which were significant ($p < 0.05$) are shown in Figure 2.

Figure 2 - *Significant correlations (absolute value) between the acoustic features tested and the five pseudo-perceptual dimensions generated by the listeners' voice similarity judgements by the MDS analysis for each speaker group, dimensions 1 to 5, labelled dim1, dim2, etc.*



Long-term f0 plays a role in judgements of voice similarity in all groups except SSBE1. LTF results show significant correlations in different groups in different ways. LTF1 is significantly correlated with at least one dimension for all groups except Wakefield. LTF2 appears for two SSBE groups (1 and 3), York and Bradford. The higher formants yield a significant correlation for one group each only: in Wakefield for LTF3 and in SSBE2 for LTF4, for dim4 in both cases. AR does not play much of a role in these results, achieving significance in a higher dimension in only two of the groups: SSBE3 (dim4) and York (dim5). Within the three SSBE groups, results show some patterns, but not complete consistency. LTF1 is the only feature appearing for all three groups. The groups for the three Yorkshire accents all display a significant result for f0 in a low dimension, but differ in patterns of significance for LTF and AR features.

5.7 Discussion

Long-term f0 clearly makes an important contribution to listeners' assessment of voice similarity, showing significant correlations in all groups except SSBE1, and generally in low dimensions. It is not surprising that f0 is dominant, given that the pitch of a voice is intuitively salient. Long-term formants are also playing a role here, correlating with dimensions 2-5 in different ways in each group, notably with F1 featuring for all groups except Wakefield. These findings of f0 being the most important and formants also making a contribution are broadly consistent with the initial Nolan et al. (2011) study (whose listener ratings are also those used for SSBE1 in the present study), although the use of 'global' formant measures across vowel categories rather than LTF has obviously led to some differences. The use of the INDSCAL version of MDS in the present analysis would account for the different correlation results for f0 in SSBE1 in Nolan et al. (2011).

AR appears to make some contribution to voice similarity judgements in SSBE3 and York, but needs further investigation. Given that the stimuli were only 3s in duration, it is possible that a perceptual correlate of AR had not had sufficient opportunity to become established when listeners made their decisions. Further research looking at whether temporal-based features such as AR contribute more to the perception of voice similarity for longer stretches of speech is needed.

The data collected enable variation within and across accent groups to be evaluated. The three SSBE groups yielded similar, but not identical, patterns in the profile of acoustic features correlating with perceived voice similarity. The three Yorkshire accents produced a consistent result for f0, but differences in the patterning of LTF and AR. Although the three Yorkshire groups were labelled by their three different locations, further analysis is needed to determine how perceptually different these three accents are. Looking across the six groups, it is not clear if some of the differences observed could be attributed to speaker to speaker or sample to sample variability as much as accent grouping.

A likely source of variation in the listeners' judgements is that of their own accent background (cf. Clopper, Pisoni, 2006; Williams, Garrett & Coupland, 1999).

The listeners recruited were English speakers from England, with no more specific control on their accent background; further work should investigate regional accent background of listeners as a variable.

Developing a model of perceived voice similarity that could predict how similar two voices would sound on the basis of their phonetic properties would appear to be some way off at this stage, although it is clear that f0 and formant frequencies will have a part to play. Such a model could be central to the more efficient selection of foil voices for voice parades. A further interesting future direction for this work will be to consider possible relationships between listener-assessed and machine-assessed voice similarity using automatic speaker recognition technology (see Gerlach, McDougall, Kelly, Alexander & Nolan, 2020).

6. *Study 2: The effect of parade sample duration on voice recognition accuracy*[3]

The Home Office procedure for collecting earwitness identification evidence has been implemented effectively for a number of cases in the UK since its introduction. However, there are aspects of the procedure which were devised on the basis of the parallel eyewitness procedure from Code D of PACE that have not been subjected to rigorous experimental testing in the earwitness context. For example, the procedure calls for a line-up of nine voices, but does a nine-sample parade afford optimal earwitness recognition (cf. Bull, Clifford, 1999; Levi, 1998)? The witness is asked for a decision after listening to all voices: does this serial format lead to optimal recognition, or would having the opportunity to select or reject each voice immediately after hearing the sample yield greater accuracy of recognition (cf. Seale-Carlisle, Mickes, 2016; Smith, Bird, Roeser, Robson, Braber, Wright & Stacey, 2020)? The witness is allowed to listen to each voice sample as many times as they wish: does this provide the best opportunity for a witness to recognise a voice (cf. Pozzulo, Lindsay, 1999 regarding elimination line-ups), or could interference be at play (cf. Stevenage, Howland & Tippelt, 2011)? Further, some parameters of the procedure were chosen relatively arbitrarily and would benefit from experimental examination. For example, the procedure stipulates that the voice samples should be one minute in duration, which means that the parade will be nine minutes long: does this give the witness the optimum opportunity to recognise a voice if they have heard it previously, or is this task too long and distracting (cf. Smith et al., 2020)?

One of the key aims of the *IVIP* project is to consider the parameters of the Home Office procedure further, and in particular to determine experimentally whether there are aspects of the procedure which could be modified to optimise the performance of earwitnesses when undertaking a voice parade. In the present paper, the results of the first *IVIP* experiment investigating voice parade parameters are presented: a study of the effect of voice parade sample duration on earwitness identification accuracy.

[3] The description of this study given here is an overview of the findings of a larger study within *IVIP*, reported in greater depth in Pautz, Smith, Müller-Johnson, Nolan, Paver and McDougall (submitted).

Little previous work is available exploring whether earwitness performance is affected by the length of the samples used in a voice parade. One exception is a study by Smith et al. (2020) which investigated earwitness accuracy for parades run with 15s or 30s voice samples. The results did not show an effect of sample duration, but research is still needed to compare the currently recommended 60s with shorter sample durations, as is investigated in the present study.

6.1 Experiment design and speaker selection

Participants were exposed to a target voice for 60s and later attempted to recognise the voice of the 'perpetrator' whom they heard from a target-present or target-absent voice parade. A target-present parade simulates a situation where the guilty suspect has been apprehended, whereas a target-absent parade simulates an innocent suspect having been apprehended. The same three speech databases used in Study 1 above provided the speakers in the present experiment. Voice parades were constructed for six male target speakers of English: three speakers of SSBE (from *DyViS*) and one of each of York (*YorViS*), and Bradford and Wakefield (*WYRED*) Englishes. These target speakers were chosen as one each from the six groups of 15 speakers used in Study 1. For each target speaker, the other 14 speakers in his group were candidates for his corresponding target-present and target-absent voice parades. Using the MDS results from Study 1, the foils for each target-present parade were chosen as the eight speakers judged by the listeners as most similar-sounding to the relevant target speaker (cf. McDougall et al., 2015). The speakers chosen for each target-absent parade were the nine speakers judged most similar-sounding to the target speaker to whom the listeners had been exposed. The speakers used as targets and foil speakers are listed in Table 2.

6.2 Speech materials

The exposure material for each target speaker was taken from Task 2, the telephone call task, in each of *DyViS, YorViS,* and *WYRED.* A sample of 60s duration was selected for each target from his side of the phone call in studio quality, with only the target's speech included.

Table 2 - *Target speakers and their corresponding foil speakers (speaker numbers from each database). * indicates the additional foil used in target-absent parades*

Group	Accent	Target	Foils
SSBE1	*DyViS*	56 (D1)	25*, 28, 39, 60, 65, 95, 111, 112, 115
SSBE2	*DyViS*	23 (D2)	2, 4, 37, 31, 32, 35, 50, 76*, 87
SSBE3	*DyViS*	80 (D3)	6, 30, 40, 46, 75, 81, 96, 99*, 107
York	*YorViS*	8 (YO)	1*, 2, 4, 7, 8, 10, 16, 18, 19, 21
Bradford	*WYRED*	185 (W1)	187, 175, 170, 176*, 132, 189, 157, 156, 135
Wakefield	*WYRED*	166 (W2)	103*, 127, 138, 141, 145, 146, 152, 158, 178

The voice parades were constructed using the Home Office (2003) methodology, using the Task 1 simulated police interview task from each database as the source of speech material. Short, self-contained chunks of speech containing words, phrases or short sentences were excised from each speaker's interview. The parade samples were created by jumbling the order of each speaker's chunks to form a 'collage' of speech giving an overall impression of that speaker's voice. 15s, 30s and 60s voice samples were compiled for each of the speakers in the parades. Each listener undertook a parade containing either 15s, 30s or 60s samples. Allocation of listeners to one of the six target speakers within each of the six conditions (3 Sample Duration × 2 Target Presence) was done using a balanced randomisation procedure.

6.3 Experiment platform and format

The experiment was conducted online using the web-based platform *Gorilla* (Anwyl-Irvine et al., 2020). At the start of the experiment, participants took a test to check that they were using headphones or earphones (Woods et al., 2017). Participants were allowed to take this test a second time if they failed the first.

After listening to the 60s sample of exposure material, participants undertook a word-search task containing words for types of fruit (http://www.wordsearch-puzzles.co.uk) lasting five minutes. To prevent auditory rehearsal of the encoding, the task was accompanied by a recording of ambient noise made in a public lobby featuring unintelligible speech sounds. While a five-minute retention interval is not representative of the real-world context of a genuine voice parade case, the requirements of the task at least meant that participants' short-term memory capacity was exceeded, and long-term memory would be relied on.

Immediately before listening to the voice parade, participants were instructed that the voice that they heard in the original recording may or may not be present in the line-up (as is recommended in the Home Office guidelines); this was noted in bold lettering in the pre-parade instructions. After they had listened to the parade and prior to indicating their decision, participants were reminded once again that the voice they had heard at the beginning of the experiment might or might not have been present in the line-up. Participants listened to all nine voices before registering a decision, then they assessed how confident they were in their decision using an eleven-point scale (0-10, with 0 = 'not at all confident' and 10 = 'extremely confident').

6.4 Participants

271 participants, 136 male and 135 female, aged 18-40 years (Mean = 27.68, SD = 6.1) were recruited via the experimental participant recruitment website *Prolific* (https://www.prolific.co/). Participants had been born in England, had lived in England for most or all of their lives before turning 18 years, and spoke English as their first language. Participants self-reported having no hearing loss or hearing difficulties.

6.5 Results

The percentages of correct identifications or indications of 'not present' made in the voice parades for each of the three parade sample durations are shown in Figure 3. Chance level is at 10%, indicated by the dotted red line. [Total number of options = 9 possible 'identifications' and 1 possible 'not present'. Chance level = 100*(1/total number of options).] Each datapoint in Figure 3 represents the responses of 42-48 (Mean = 45.1) listeners.

Performance on the parades is low overall, especially in target-absent parades. Higher levels of recognition accuracy are shown for target-present than target-absent parades. A Likelihood Test for the 3 × 2 factorial design showed a significant relationship between Target Presence and accuracy of response ($G^2(1) = 19.47$, $p < 0.001$), confirming that target-present parades yield higher levels of accuracy than target-absent. Neither Sample Duration ($G^2(2) = 0.62$, $p = 0.734$) nor the interaction Target Presence × Sample Duration ($G^2(2) = 1.45$, $p = 0.484$) showed a significant relationship with accuracy of response.

In target-present parades, descriptively-speaking the 15s samples give the best performance (45% correct), followed by the 60s samples (38% correct), then 30s samples (36% correct). However, pairwise comparisons with a Hochberg correction (Hochberg, 1988) across sample durations gave negligible evidence to support the hypothesis that substantial differences were present, both between-groups (15s versus 30s versus 60s overall) and within-groups (target-present 15s versus target-present 30s, target-absent 15s versus target-absent 30s, etc.).

Listeners' accuracy of performance exhibits considerable variation by individual target speaker, as can be seen in Figure 4 which shows the percentages of accurate responses in the voice parades for each of the sample duration and target presence conditions for the six target speakers separately. Chance level is at 10%, indicated by the dotted red line. Each datapoint in Figure 4 is the result of the responses of only 5-9 (Mean = 7.53) listeners and within each sample duration, the accuracy of responses is very varied. For example in the 15s condition in target-present parades, participants produced a spread of results between 28.6% for both the Y and W1 targets and 75% for the D1 target. This suggests that some voices are much harder to remember than others.

Figure 3 - *Percentage accuracy of participants in target-present and target-absent parades for 15s, 30s and 60s conditions. Error bars show 95% confidence intervals. The dotted red line indicates the chance level of 10% for each condition*

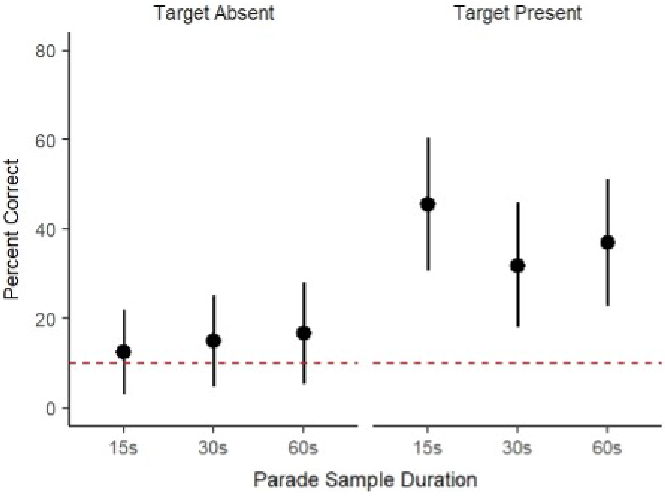
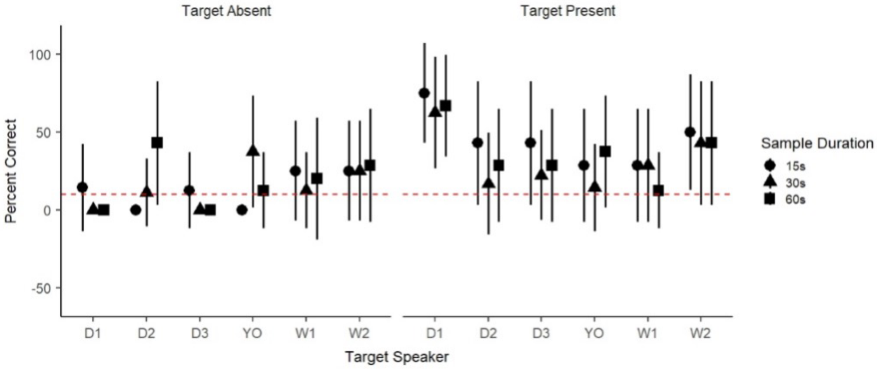


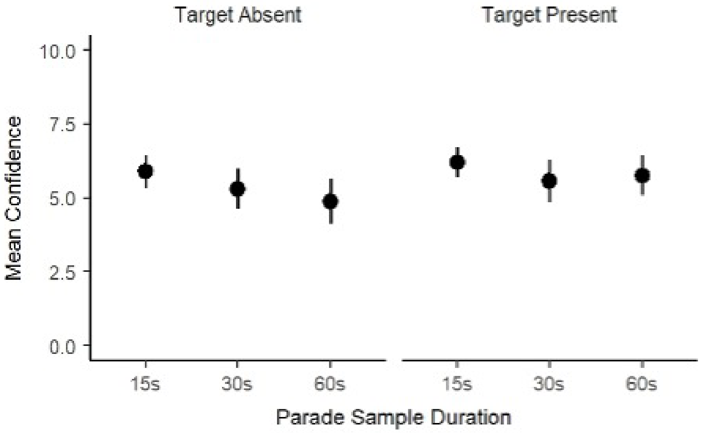
Figure 4 - *Percentage accuracy of participants in target-present (upper panel) and target-absent (lower panel) parades for 15s, 30s and 60s sample durations. Error bars show 95% confidence intervals. The dotted red line indicates the chance level of 10% for each condition*



The listeners' assessments of their confidence in selecting the correct voice from a parade or correctly choosing 'not present' are shown as mean values per sample duration and target presence condition in Figure 5.

An ordinal regression model was used to investigate the relationship between the dependent variable confidence, and the independent variables of accuracy, target presence, and sample duration. Overall, there was a statistically significant, moderately strengthened, positive association between self-related confidence and accuracy ($b = .54$, SE = 0.25, $p = 0.034$). There was no statistically significant main effect of target presence (*b* = .186, SE = 0.22, *p* = 0.402). With the 60s sample duration as the reference, no significant difference in confidence was found between the 15s (*b* = 0.48, SE = 0.26, *p* = 0.064) and 30s (*b* = 0.13, SE = 0.26, *p* = 0.621) sample durations.

Figure 5 - *Mean confidence ratings of participants in target-present and target-absent parades for 15s, 30s and 60s conditions. Error bars show 95% confidence intervals for the condition means*



6.6 Discussion

Overall recognition accuracy was low, driven in particular by false identifications in target-absent parades (Figure 3). It is hoped that this tendency could be mitigated against through using stronger warnings stressing the real-world consequences of wrongful identification (Smith, Roeser, Pautz, Davis, Robson, Wright, Braber & Stacey, submitted). Performance on the target-present parades showed higher rates of recognition accuracy, consistent with previous research. The task itself was very difficult, arguably the more so as participants were not told that their memory of the voice heard would be tested until after the retention interval had elapsed. In real-world situations, one would expect the witness's exposure to the voice to be longer than the one minute given in the present experiment, and that the witness would have expressed some preparedness to recall the voice heard at the crime scene on a future occasion. In the present study, listeners' confidence ratings tended to fall in the middle of the scale (Figure 5), possibly reflecting indecision, and suggesting that participants were aware of the difficulty of the identification task.

The results comparing parade sample durations presented here suggest that the Home Office procedure for voice parades could be modified by reducing parade sample duration to between 15 and 30 seconds satisfactorily. This would reduce the total amount of work needed from the phonetician preparing the samples and hence slightly speed up the process of constructing a parade. Further, the requirement for less speech material within each sample may open up the number of candidate foil recordings available when compiling a parade, since shorter interviews containing less material would become eligible for consideration.

The experimental results also highlight the importance of taking into account individual differences between target speakers. The six different target speakers used yielded markedly different levels of identification accuracy across the various experimental conditions. The recognisability of different target speakers varies greatly (cf. McDougall, Nolan & Hudson, 2015), yet many experimental studies investigating earwitness performance are conducted using a single target speaker. The use of different target speakers in the present experiment is likely to have contributed some of the noise in the comparison of conditions, yet it is crucial to appreciate the extent of variability of identification accuracy yielded by different target speakers. A key area for future research is what makes a voice more distinctive and/or memorable to listeners (cf. Sørensen, 2012), a topic being pursued in other studies within the *IVIP* project.

7. *Concluding remarks*

Since earwitness evidence can make a crucial contribution to a criminal case it is essential that it is collected on the basis of a comprehensive knowledge base. Given that research shows that different mechanisms are involved in processing faces and voices, earwitness identification procedures must be developed specifically in the light of auditory-oriented research. This paper has presented two studies from the *IVIP* project with implications for voice parade procedures. The first offers some initial foundations for the development of a model of perceived voice similarity which could in principle contribute to the process of selecting the foils for a voice parade. The second provides support for the notion of reducing the duration of voice samples used in the voice parade procedure in England and Wales.

Acknowledgements

Thanks are due to the *IVIP* project researchers, especially Harriet Smith, Nikolaus Pautz, Katrin Müller-Johnson, Alice Paver and Francis Nolan for their contributions to the work presented in §5 and §6. The *IVIP* project is funded by the UK Economic and Social Research Council [Grant Reference ES/S015965/1] and is hosted by the University of Cambridge, De Montfort University, Nottingham Trent University and the University of Oxford. The author is grateful to Philip Harrison for his assistance with LTF analysis. Further thanks are due to Nathan Atkinson, Toby Hudson, Christin Kirchhübel and Elaine Schmidt for their assistance with data collection and measurement in the earlier projects *VoiceSim* and *YorViS* upon which the *IVIP* project builds, and to Erica Gold for providing her AR measurements for the *DyViS* speakers.

Bibliography

Anwyl-Irvine, A.L., Massonnié, J., Flitton, A., Kirkham, N. & Evershed, J.K. (2020). Gorilla in our midst: An online behavioral experiment builder. In *Behavior Research Methods*, 31(1), 388-407. [DOI: 10.3758/s13428-019-01237-x].

Barton, J.J. (2008). Structure and function in acquired prosopagnosia: Lessons from a series of 10 patients with brain damage. In *Journal of Neuropsychology*, 2(1), 197-225. [DOI: 10.1348/174866407X214172].

Baumann, O., Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. In *Psychological Research*, 74, 110–120. [DOI: 10.1007/s00426-008-0185-z].

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

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

Boersma, P., Weenink, D. (1992-2021). PRAAT: Doing phonetics by computer. [Computer program]. http://www.praat.org/.

Brédart, S., Barsics, C. (2012). Recalling semantic and episodic information from faces and voices: A face advantage. In *Current Directions in Psychological Science*, 21(6), 378-381. [DOI: 10.1177/0963721412454876].

Broeders, A.P.A. (1996). Earwitness identification: Common ground, disputed territory and uncharted areas. In *Forensic Linguistics*, 3(1), 3-13. [DOI: 10.1558/ijsll.v3i1.3].

Broeders, A.P.A., Rietveld, A.C.M. (1995). Speaker identification by earwitnesses. In Braun, A., Köster, J.-P. (Eds.), *Studies in forensic phonetics: Beiträge zur Phonetik und Linguistik,* 64, 24-40.

Bruce, V., Young, A.W. (1986). Understanding face recognition. In *British Journal of Psychology*, 77, 305-327. [DOI: 10.1111/j.2044-8295.1986.tb02199.x].

Bull, R., Clifford, B. (1999). Earwitness testimony. In Heaton-Armstrong, A., Shepherd, E. & Wolchover, D. (Eds.), *Analysing witness testimony: A guide for legal practitioners and other professionals*. London: Blackstone, 194-206.

Campanella, S., Belin, P. (2007). Integrating face and voice in person perception. In *Trends in Cognitive Sciences*, 11(12), 535-543. [DOI: 10.1016/j.tics.2007.10.001].

Clopper, C.G., Pisoni, D.B. (2006). Effects of region of origin and geographic mobility on perceptual dialect categorization. In *Language Variation and Change*, 18(2), 193-221. [DOI: 10.1017/S0954394506060091].

de Jong-Lendle, G., Nolan, F., McDougall, K. & Hudson, T. (2015). Voice lineups: A practical guide. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 10-14 August 2015, Paper number 0598. 1-5. [https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0598.pdf].

Ellis, H.D., Jones, D.M. & Mosdell, N. (1997). Intra- and inter-modal repetition priming of familiar faces and voices. In *British Journal of Psychology*, 88(1), 143-156. [DOI: 10.1111/j.2044-8295.1997.tb02625.x].

Garrido, L., Eisner, F., McGettigan, C., Stewart, L., Sauter, D., Hanley, J.R., Schweinberger, S.R., Warren, J.D. & Duchaine, B. (2009). Developmental phonagnosia: A selective deficit of vocal identity recognition. In *Neuropsychologia*, 47(1), 123-131. [DOI: 10.1016/j.neuropsychologia.2008.08.003].

Gerlach, L., McDougall, K., Kelly, F., Alexander, A. & Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. In *Speech Communication*, 124, 85-95. [DOI: 10.1016/j.specom.2020.08.003].

Giguère, G. (2006). Collecting and analyzing data in multidimensional scaling experiments: A guide for psychologists using *SPSS*. In *Tutorial in Quantitative Methods for Psychology*, 2(1), 27-38. [DOI: 10.20982/tqmp.02.1.p026].

Gold, E., Ross, S. & Earnshaw, K. (2018). The 'West Yorkshire Regional English Database': Investigations into the generalizability of reference populations for forensic speaker comparison casework. In *Proceedings of Interspeech 2018*, Hyderabad, 2-6 September 2018, 2748-2752. [DOI: 10.21437/Interspeech.2018-65].

Hailstone, J.C., Crutch, S.J., Vestergaard, M.D., Patterson, R.D. & Warren, J.D. (2010). Progressive associative phonagnosia: A neuropsychological analysis. In *Neuropsychologia*, 48(4), 1104-1114. [DOI: 10.1016/j.neuropsychologia.2009.12.011].

Hochberg, Y. (1988). A sharper Bonferroni procedure for multiple tests of significance. In *Biometrika*, 75(4), 800-802. [DOI: 10.2307/2336325].

Hollien, H. (1996). Consideration of guidelines for earwitness lineups. In *Forensic Linguistics*, 3(1), 14-23. [DOI: 10.1558/ijsll.v3i1.14].

Home Office (2003). Advice on the use of voice identification parades. UK Home Office Circular 057/2003 from the Crime Reduction and Community Safety Group, Police Leadership and Powers Unit [http://webarchive.nationalarchives.gov.uk/20130125153221/http://www.homeoffice.gov.uk/about-us/corporate-publications-strategy/home-office-circulars/circulars-2003/057-2003/]

Imaizumi, S., Mori, K., Kiritani, S., Hosoi, H. & Tonoike, M. (1997). Task-dependent laterality for cue decoding during spoken language processing. In *NeuroReport*, 8, 899-903. [DOI: 10.1097/00001756-199803300-00025].

Jessen, M. (2007). Forensic reference data on articulation rate in German. In *Science and Justice*, 47, 50-67. [DOI: 10.1016/j.scijus.2007.03.003].

Levi, A.M. (1998). Protecting innocent defendants, nailing the guilty: A modified sequential lineup. In *Applied Cognitive Psychology*, 12(3), 265-275. [DOI: 10.1002/(SICI)1099-0720(199806)12:3<265::AID-ACP515>3.0.CO;2-O].

Lindsay, R.C.L., Ross, D.F., Read, J.D. & Toglia, M.P.(2007). *The handbook of eyewitness psychology. Volume II: Memory for people*. Mahway, NJ.: Lawrence Erlbaum.

Mathôt, S., Schreij, D. & Theeuwes, J. (2012). *OpenSesame*: An open-source, graphical experiment builder for the social sciences. In *Behavior Research Methods*, 44(2), 314-324. [DOI: 10.3758/s13428-011-0168-7].

McDougall, K. (2013). Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. In *International Journal of Speech, Language and the Law*, 20(2), 163-172. [DOI: 10.1558/ijsll.v20i2.163].

McDougall, K., Hudson, T. & Atkinson, N. (2014). Listeners' perception of voice similarity in Standard Southern British English versus York English. Paper presented at International Association for Forensic Phonetics and Acoustics Annual Conference, Zürich, August 31 – September 3, 2014.

McDougall, K., Duckworth, M. & Hudson, T. (2015). Individual and group variation in disfluency features: A cross-accent investigation. In 2015, The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, 10-14 August 2014, Paper number 0308.1-5. [http://www.icphs.info/pdfs/Papers/ICPHS0308.pdf]

McDougall, K., Nolan, F. & Hudson, T. (2015). Telephone transmission and earwitnesses: Performance on voice parades controlled for voice similarity. In *Phonetica*, 72, 257-272. [DOI: 10.1159/000439385].

Moos, A. (2010). Long-term formant distribution as a measure of speaker characteristics in read and spontaneous speech. In *The Phonetician*, 101/102, 7-24.

Neuner, F., Schweinberger, S.R. (2000). Neuropsychological impairments in the recognition of faces, voices, and personal names. In *Brain and Cognition*, 44, 342-366. [DOI: 10.1006/brcg.1999.1196].

Nolan, F. (2003). A recent voice parade. In *International Journal of Speech, Language and the Law*, 10(2), 277-291. [DOI: 10.1558/sll.2003.10.2.277].

Nolan, F., Grigoras, C. (2005). A case for formant analysis in forensic speaker identification. In *International Journal of Speech, Language and the Law*, 2(2), 143-173. [DOI: 10.1558/sll.2005.12.2.143].

Nolan, F., McDougall, K. & Hudson, T. (2011). Some acoustic correlates of perceived (dis)similarity between same-accent voices. In Lee, W.-S., Zee, E. (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 17-21 August 2011, 1506-1509. [http://www.icphs2011.hk/resources/OnlineProceedings/RegularSession/Nolan/Nolan.pdf]

Nolan, F., McDougall, K. & Hudson, T. (2013). Effects of the telephone on perceived voice similarity: Implications for voice line-ups. In *International Journal of Speech, Language and the Law*, 20(2), 229-246. [DOI: 10.1558/ijsll.v20i2.229].

Nolan, F., McDougall, K., de Jong, G. & Hudson, T. (2009). The *DyViS* database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. In *International Journal of Speech, Language and the Law*, 16(1), 31-57. [DOI: 10.1558/ijsll.v16i1.31].

Pautz, N., Smith, H.M.J., Müller-Johnson, K., Nolan, F., Paver, A. & McDougall, K., (submitted). Identifying unfamiliar voices: the influence of sample duration and parade size. Pre-print at https://psyarxiv.com/h9ynr/

Pozzulo, J.D., Lindsay, R.C.L. (1999). Elimination lineups: An improved identification procedure for child eyewitnesses. In *Journal of Applied Psychology*, 84(2), 167-176. [DOI: 10.1037/0021-9010.84.2.167].

Remez, R.E., Fellowes, J.M. & Nagel, D.S. (2007). On the perception of similarity among talkers. In *Journal of the Acoustical Society of America*, 122(6), 3688-96. [DOI: 10.1121/1.2799903].

Rietveld, A.C.M., Broeders, A.P.A. (1991). Testing the fairness of voice identity parades: The similarity criterion. In *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence, 19-24 August 1991, 46-49.

Robson, J. (2017). A fair hearing? The use of voice identification parades in criminal investigations in England and Wales. In *Criminal Law Review*(1), 36-50. [http://irep.ntu.ac.uk/id/eprint/29636/1/6791_Robson.pdf].

Schiffman, S.S., Lance Reynolds, M. & Young, F.W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.

Seale-Carlisle, T.M., Mickes, L. (2016). US line-ups outperform UK line-ups. In *Royal Society Open Science*, 3, 160-300. [DOI: 10.1098/rsos.160300].

Sheffert, S.M., Olson, E. (2004). Audiovisual speech facilitates voice learning. In *Perception and Psychophysics*, 66, 352-362. [DOI: 10.3758/bf03194884].

Sjölander, K. (1997). The Snack sound toolkit. [Computer program]. [http://www.speech.kth.se/snack/].

Smith, H.M.J., Bird, K., Roeser, J., Robson, J., Braber, N., Wright, D. & Stacey, P.C. (2020). Voice parade procedures: Optimising witness performance. In *Memory*, 28(1), 2-17. [DOI: 10.1080/09658211.2019.1673427].

Smith, H.M.J., Roeser, J., Pautz, N., Davis, J.P., Robson, J., Wright, D., Braber, N. & Stacey, P.C. (submitted). Evaluating earwitness identification procedures: Adapting pre-parade instructions and parade procedure. [Preprint https://psyarxiv.com/nxr3e/].

Sørensen, M.H. (2012). Voice line-ups: Speakers' f0 values influence the reliability of voice recognitions. In *International Journal of Speech, Language and the Law*, 19(2), 145-158. [DOI: 10.1558/ijsll.v19i2.145].

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

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

von Kriegstein, K., Eger, E., Kleinschmidt, A. & Giraud, A.L. (2003). Modulation of neural responses to speech by directing attention to voices or verbal content. In *Cognitive Brain Research*, 17(1), 48-55. [DOI: 10.1016/s0926-6410(03)00079-x].

von Kriegstein, K., Kleinschmidt, A., Sterzer, P. & Giraud, A.-L. (2005). Interaction of face and voice areas during speaker recognition. In *Journal of Cognitive Neuroscience*, 17(3), 367-376. [DOI: 10.1162/0898929053279577].

Walden, B.E., Montgomery, A.A., Gibeily, G.J., Prosek, R.A. & Schwartz, D.M. (1978). Correlates of psychological dimensions in talker similarity. In *Journal of Speech and Hearing Research*, 21, 265-275. [DOI: 10.1044/jshr.2102.265].

Williams, A., Garrett, P. & Coupland, N. (1999). Dialect recognition. In Preston, D.R. (Ed.), *Handbook of perceptual dialectology*. Philadelphia: John Benjamins, 345-358.

Woods, K.J.P., Siegel, M., Traer, J. & McDermott, J.H. (2017). Headphone screening to facilitate web-based auditory experiments. In *Attention, Perception & Psychophysics*, 79(7), 2064-2072. [DOI: 10.3758/s13414-017-1361-2].

Young, A.W., Frühholz, S. & Schweinberger, S.R. (2020). Face and voice perception: Understanding commonalities and differences. In *Trends in Cognitive Sciences*, 24(5), 398-410. [DOI: 10.1016/j.tics.2020.02.001].

Yovel, G., Belin, P. (2013). A unified coding strategy for processing faces and voices. In *Trends in Cognitive Sciences*, 17(6), 263-271. [DOI: 10.1016/j.tics.2013.04.004].