Human and automatic voice comparison with regionally variable speech samples

Vincent Hughes, Department of Language and Linguistic Science, University of York, York, YO10 5DD, United Kingdom, vincent.hughes@york.ac.uk (corresponding author)

Carmen Llamas, Department of Language and Linguistic Science, University of York, York, YO10 5DD, United Kingdom, carmen.llamas@york.ac.uk

Thomas Kettig, Department of Languages, Literatures and Linguistics, York University, Toronto, M3J 1P3, Canada, tkettig@yorku.ca

Abstract

In this paper, we compare and combine human and automatic voice comparison results based on short, regionally variable speech samples. Likelihood ratio-like scores were extracted for 120 pairs of same- (45) and different-speaker (75) samples from a total of 896 British English listeners. The samples contained the voices of speakers from Newcastle and Middlesbrough (in North-East England), as well as speakers of Standard Southern British English (modern RP). In addition to within-accent comparisons, the test included between-accent, different-speaker comparisons for Middlesbrough and Newcastle, which are perceptually and regionally proximate accents. Scores were also computed using an x-vector PLDA automatic speaker recognition (ASR) system. The ASR system (EER=10.88%, C_{llr}=0.48) outperformed the human listeners (EER=23.55%, C_{llr}=0.75) overall and no improvement was found in the ASR output when fused with the listener scores. There was, unsurprisingly, considerable between-listener variability, with individual error rates varying from 0% to 100%. Performance was also variable according to the regional accent of the speakers. Notably, the ASR system performed worst with Newcastle samples, while humans performed best with the Newcastle samples. Human listeners were also more sensitive to high-salience between-accent comparisons, leading to almost categorical different-speaker conclusions, compared with the ASR system, whose performance with these samples was similar to within-accent comparisons.

Keywords

Voice comparison, automatic speaker recognition, human listeners, accent variation

1. Introduction

This paper examines the performance of untrained, human listeners and an automatic speaker recognition (ASR; not to be confused with automatic *speech* recognition which is also commonly referred to as ASR) system at the task of voice comparison. Voice comparison involves analysing two (or more) samples of speech with a view to assessing whether they contain the voices of the same or different speakers. While much work has been done to examine listener performance at voice comparison with familiar voices (e.g., Roebuck and Wilding, 1997; Plante-Hébert and Boucher, 2015), much less work has been conducted using unfamiliar or regionally variable voices; this is despite the identification of, and adaptation to, new voices being extremely important for speech processing. There has also been relatively little work examining the processing involved in voice comparison by human listeners relative to the empirical, algorithmic approaches of ASR systems.

Our aim here is to better understand what linguistic and phonetic information human listeners and ASR systems are sensitive to when comparing unfamiliar voices, specifically those of different regional accents, and to assess whether this information is complementary. This provides a means of explaining some of the underlying processing within both human listeners and automatic systems, which are both opaque to some extent (embedded in the brains of listeners or the deep neural networks of automatic systems).

1.1 Voice comparison by humans

Previous research has demonstrated a range of factors influencing human abilities to discriminate between pairs of same- and different-speaker samples of unfamiliar voices. With good quality but very short recordings (around two seconds), overall accuracy of 82.4% is reported by Kreiman and Papcun (1991). Legge et al. (1984) demonstrate that performance improves as a function of duration, with accuracy of only around 50% (not statistically above chance) reported using six-second samples, and accuracy of 70% reported using 60-second samples. Performance degrades as a function of audio quality (Smith et al. 2019), while style mismatch (e.g., read vs. spontaneous speech) between samples also leads to degradation in listener performance compared with style-matched samples. Further, research has shown that listeners use different strategies for recognising unfamiliar voices depending on whether the pair of voices belong to the same or different speakers (see Afshan et al. 2022), exemplifying differences in perceptual strategies for ‘telling voices together’ and ‘telling voices apart’ (Johnson et al., 2020).

Work on this topic has also been conducted in the forensic domain, specifically in the context of earwitness evidence. Such evidence is used when a witness to a crime hears the voice of the criminal, but does not see their face. In such cases, the witness needs to demonstrate the ability to recognise the voice of the criminal in the vocal equivalent of a line-up. Research has often shown considerable between-listener variation in earwitness performance, but with some systematic effects as a function of factors relating to the target voice (e.g., cases of voice disguise or face coverings), factors relating to the listener (e.g., young listeners perform better than older listeners), and case-specific factors (e.g., time between exposure and line-up) (Bull and Clifford, 1984; van Wallendael et al., 1994). Most notably for the purposes of the present study, Atkinson (2015) showed better performance among listeners more familiar with the accent of the target speaker in a voice line-up task than listeners less familiar with the accent (see similar findings in Braun et al. 2018).

1.2 Voice comparison by ASR

The last 20 years has seen considerable improvement in the performance of ASR systems. Systems now utilise deep neural networks (DNNs) trained on large amounts of data to convert sets of acoustic features extracted from the speech signal into a speaker embedding extracted from the DNN (e.g., an x-vector; Snyder et al., 2018), often followed by dimension reduction via linear discriminant analysis (LDA). Given that systems still typically utilise MFCCs as input and that features are extracted from across the entire speech-active portion of a sample, the speaker embedding is thought to capture information principally about the supralaryngeal vocal tract, both in terms of physiology and also long-term vocal setting (for a phonetic definition of long-term vocal setting see Laver, 1980). Systems then compare embeddings from two speech samples (one of a known speaker, one of an unknown speaker) to generate a *score*. This score captures the two samples' similarity (and often their typicality, if using probabilistic linear discriminant analysis (PLDA)), but generally the score cannot be interpreted directly until it is calibrated using representative scores from sets of comparisons where the ground truth is known. The output is then a likelihood ratio (LR), a measure of the strength of the evidence given the same- and different-speaker propositions.

State-of-the-art ASR performance is now extremely good, even with challenging materials. A forensic evaluation coordinated by Morrison and Enzinger (2019) reports equal error rates (EER) of up to 13.9% with older generations of systems (such as the Gaussian Mixture Model Universal Background Model (GMM-UBM) approach; see Reynolds et al., 2000), compared with 2.2% for the best performing x-vector system. Considerable work has been dedicated to developing systems that are increasingly robust to technical effects, such as telephone transmission, background noise, and duration. Yet large sources of mismatch between samples remain a challenge. These includes speaker factors, especially those related to the supralaryngeal vocal tract, as demonstrated by Hughes et al. (2023). The extent to which systems are sensitive to other linguistic information such as segmental variation or laryngeal features remains under-researched.

1.3 Comparing human and ASR approaches

Approaches to testing listeners and ASR systems in voice comparison tasks with unfamiliar voices have varied widely and reported a range of results. ASR systems have generally been shown to outperform humans (Hautamäki et al., 2010; Khan et al., 2011; Das and Prasanna, 2016). With short audio samples, results are much more mixed.

Park et al. (2018) compared the performance of 65 human listeners with an i-vector (see Dehak et al., 2011; Morrison et al., 2020) PLDA ASR system using two-second samples across two speaking styles (read speech, pet-directed speech). Overall performance was relatively poor, although humans (EER=30.58%) outperformed the automatic system (EER=36.52%) in almost all conditions tested. The exception was the case of style-mismatched pairs with 'perceptually marked' speakers (defined as "non-American dialect, overly precise articulation and/or unusual disfluencies in reading" (2018: 377), where human performance (EER=46.23%) was considerably worse than system performance (EER=37.35%). Park et al. (2018) also found only a weak correlation between human and ASR responses, suggesting that the two approaches are sensitive to different information when making judgments. Afshan et al. (2020) tested the performance of 30 human listeners (of which 24 were L1 English speakers) and a state-of-the-art x-vector PLDA system using three-second, style-matched and -mismatched audio samples. As in Park et al. (2018), humans outperformed the automatic system in style-matched conditions (EER: 6.96% vs. 14.35% for read speech, 15.12% vs. 19.87% for conversational speech), although performance was comparable in style-mismatched conditions. Fusion of human and machine responses showed improvements in overall performance. This, again, provides evidence to suggest that humans capture complementary speaker-specific information to that captured by an ASR system.

A small number of studies have assessed human and ASR voice comparison from a forensic perspective. In 2010, a human-assisted element was introduced to the US National Institute of Standards and Technology (NIST) speaker recognition evaluation (Greenberg et al., 2010) by eliciting binary judgments from lay listeners using sets of speech samples that had been misclassified by an automatic system (i.e., those comparisons that were difficult for the automatic system). Listener performance was comparable to the results reported in Section 1.3 (EERs=30-40%), although some groups performed better with same-speaker pairs and others with different-speaker pairs. Lindh and Morrison (2011) performed a small-scale study of human and ASR performance (based on a GMM-UBM system) with a focus on forensic voice comparison using 45 pairs of high-quality Swedish voice samples and responses averaged over 52 lay listeners. Performance, as measured by the log likelihood ratio cost function (C_{llr}; Brümmer and du Preez, 2006), was considerably better for the automatic system (C_{llr}=0.033) than for the human listeners (C_{llr}=0.359). Human performance was also better when samples were played forwards, rather than backwards, indicating that segmental information is useful for listeners in distinguishing same- and different-speaker pairs. No fusion of human and automatic responses was undertaken as the automatic system was already performing at ceiling in terms of discrimination.

Basu et al. (2022) compared the performance of individual listeners to that of an x-vector ASR system in order to assess the extent to which ASR can be considered a form of expert evidence; that is, the extent to which it provides information beyond what could be reasonably expected from a judge or jury member. The study used challenging, forensically realistic samples of Australian English and tested Australian

English, American English, and Spanish-speaking listeners. Performance was evaluated on a by-speaker basis using uncalibrated scores and then compared with the performance of a calibrated x-vector system (which incorporated adaptation data for tuning LDA/PLDA). The automatic system optimally produced a C_{llr} of 0.42, with $\log_{10}$ LRs ranging between -3 and +3. Considerable between-listener variability was reported, with the best listener producing a C_{llr} of 0.51, while half of the native English listeners produced C_{llr}s of over 1. Listener performance was also affected by background, such that Australian listeners outperformed American listeners, who in turn outperformed Spanish-speaking listeners.

1.4 This study

Previous work has demonstrated that a range of factors affect human and automatic voice comparison performance, although little work has considered the effect of regional accent. Regional accent, however, provides an interesting avenue for exploring the extent to which humans and ASR systems capture both group- (i.e., accent features) and individual-level (i.e., speaker-specific) information encoded both in segmental and suprasegmental features, and whether what is captured by the two approaches is complementary.

In the present study, we use forensically-realistic speech samples from three British English accents: a standard accent (Standard Southern British English), a regional accent which is familiar to most British listeners (Newcastle), and a regional accent which is similar to Newcastle, but much less familiar to most British listeners (Middlesbrough). As well as a potential difference in levels of familiarity between Newcastle and Middlesbrough, the two accents have been chosen because they share many features that can be said to be characteristic of accents of the North East region (so are liable to be mistaken for each other), but at the same time have features which are highly localised and can serve to differentiate them. This not only allows us to test the contribution that particular segmental features make to the task of voice comparison, but also presents a realistic context of potential speaker misidentification.

These sets of voice samples allow us to address the following research questions:

- Are humans or ASR systems generally better at voice comparison and can combining results improve overall performance?
- To what extent are humans and ASR systems capturing complementary speaker-specific information?
- To what extent is human and ASR performance sensitive to accent variation?

The aim is not to replicate the exact processes of voice comparison and decision-making in the courts (as in Basu et al. 2022). Rather, we examine human and automatic processing more generally whilst considering the implications for forensics.

In addressing these research questions, we also address methodological issues with previous work which limit our understanding of the complementarity of human and ASR methods. Firstly, only a small handful of studies have attempted to empirically fuse human and ASR responses to assess whether there are improvements in performance when combining approaches. Secondly, the majority of work in this area has elicited binary accept-reject decisions and evaluated performance in terms of overall error rates. Such responses are expressions of posterior probability about identity (i.e., the probability of the samples containing the voices of the same or different speakers given the evidence) and are therefore logically inconsistent with the LR framework utilised by automatic systems and widely considered appropriate for the evaluation of forensic evidence (e.g Aitken, Taroni and Bozza 2021). Some studies have asked listeners to provide judgments about the similarity of a pair of voices using Likert scales, which can be converted to a LR-like score (see Lindh and Morrison 2011). While such an approach provides a gradient, numerical response, the output is still logically inconsistent with that of modern ASR systems, where typicality is incorporated into the computation of scores (Morrison and Enzinger, 2018). The exception is Basu et al. (2022) who asked listeners to provide a numerical value to express the relative probability of the voice given the same-speaker relative to the different-speaker propositions. This value was then used as an uncalibrated, LR-like score. While this approach is semantically consistent with the LR framework, the extent to which listeners understood what they were asked to do and whether this judgment captures the similarity and typicality element of the LR is questionable.

2. Methods

2.1 Speech samples

Speech samples for this study were taken from the TUULS (The Use and Utility of Localised Speech Forms in Determining Identity; Llamas et al. 2016-2019) and DyViS (Dynamic Variability in Speech; Nolan et al. 2009) corpora; for the purposes of the present study, we used only adult male speakers. These corpora contain speakers of Newcastle English, Middlesbrough English, and Standard Southern British English (SSBE). These accents were chosen to test the effects of standard versus non-standard/regional accents and the effects of listener accent familiarity on voice comparison performance. SSBE is essentially modern Received Pronunciation, i.e the standard accent of English English, which all listeners in our study will have had considerable exposure to and therefore should have strong familiarity with. Newcastle English is a regional accent associated with the North East of England. It is very well known in the UK and well represented in the media. Middlesbrough is a town in the North East of England which is geographically proximate to Newcastle, although Middlesbrough English is much less well-known, especially to people outside of the North East. Middlesbrough and Newcastle accents share levelled Northern and general North Eastern English features (e.g., monophthongal /eɪ/ (in words such as *face*) and /əʊ/ (in words such as *goat*)) (Kerswill, 2003), but the two accents are linguistically distinct in a number of ways e.g., Newcastle has features such as an open vowel at the end of words such as *comma* and *letter*, and a monophthongal realisation of the /aʊ/ vowel (in words such as *mouth*) that Middlesbrough does not share (Llamas, 2015). Despite this, it is extremely common for Newcastle and Middlesbrough to be confused, with accents from the North East commonly labelled as ‘Newcastle’ by non-linguists and non-locals. In the context of forensic casework, it is conceivable that samples of Newcastle and Middlesbrough English would be submitted for comparison under the assumption that they are the same speaker and the same accent, despite the rather subtle accent difference.

2.2 Creation of stimuli

From the 100 speakers in the DyViS database, 45 were chosen at random as the basis for our SSBE stimuli. These speakers were all aged between 18 and 25 years. Studio samples of a mock police interview (DyViS Task 1) were always used as the nominal ‘known’ (or reference) samples. These recordings were of high quality, with a sampling rate of 44.1kHz and 16-bit depth. Landline telephone samples involving a conversation with a mock accomplice (DyViS Task 2) were used as the nominal ‘unknown’ (or disputed) samples. Far-end telephone-transmitted versions of the Task 2 recordings (bandpass filtered by the telephone) were used in our study, with a sampling frequency of 8kHz and a 16-bit depth. The conditions of the comparisons (both in terms of speech style and channel) were similar to what would be found in typical forensic voice comparison cases.

We used the TUULS database to create samples of Middlesbrough and Newcastle speakers. Samples from a total of 15 male speakers were created per locality. For each accent, these included ten young speakers, aged 18 to 25, and five older speakers, aged 40 to 65. TUULS contains recordings of a police interview task designed to replicate DyViS Task 1 both in terms of the recording setup (position and type of microphones, albeit in different studios) and speaking style (question and answer format, where the participant is forced to ‘lie’ about certain information relating to a crime). As with the DyViS corpus, these recordings were used to represent the ‘known’ sample in forensic voice comparison cases. Since TUULS did not contain an ‘accomplice’ task recorded over a phone line (like DyViS Task 2), we created telephone-quality samples from the sociolinguistic interviews for the same speakers by bandpass filtering between 300 Hz and 3400 Hz. When filtering, we applied 100Hz smoothing at the edges of the filter to avoid infinitely steep cut-off. We also added 25 dB of white noise and then equalised all samples to 60dB.[1]

From each of the two recordings for each of the 45 SSBE, 15 Middlesbrough, and 15 Newcastle speakers, edited samples of 10 to 11 seconds were created. In creating these samples, we avoided including content that was overtly incriminating or suspicious from mock police interviews and, in the SSBE case, from the accomplice telephone calls as well, in case this affected listeners’ responses. Sections taken from the TUULS sociolinguistic interviews avoided the inclusion of any identifying information about the speakers, as well as any identifiable British place names. The second author then listened to each of the Newcastle and Middlesbrough samples and noted which ones contained highly localised speech features. Samples with marked localised

[1] This was done using a Praat script from Philip Harrison and the Praat Vocal Toolkit (Corretge 2023).

features were marked as H (high-salience), while samples with non-salient or no highly localised features were marked as L (low-salience).

These samples were then arranged into pairs containing one telephone-quality sample and one high-quality sample. Blocks of eight pairs were constructed. Each block contained one same-speaker (SS) pair each from an SSBE, a Newcastle, and a Middlesbrough speaker; one different-speaker (DS) pair representing two SSBE, two Newcastle, and two Middlesbrough speakers; one DS pair containing a high salience Newcastle sample and a high salience Middlesbrough sample; and one DS pair containing a low salience Newcastle sample and a low salience Middlesbrough sample. In total, each block therefore contained three SS pairs and five DS pairs. Blocks were balanced such that, in cross-accent DS comparisons, if a high salience Newcastle sample was the first (telephone-quality) sample and a high salience Middlesbrough sample was the second (HQ) sample in a mixed-region pair, the low salience mixed-region pair had Middlesbrough first and Newcastle second.

Pairs were constructed to avoid any identical speech content or similar semantic cues appearing in both samples of a pair. Listeners were presented with three blocks of eight pairs embedded in an online game-based experiment, with each block of comparisons used as a separate level. The first level was a basic experimental survey interface, of the kind often found with online tools such as Qualtrics. The second and third levels involved additional variables to test different research questions (e.g., more graphics/ gameplay, conclusions provided by a forensic expert). For the purposes of the present study, we focus on listener responses to pairs in just the first block of eight comparisons with a basic, non-gameplay interface. Thus, for the present study we only analyse eight responses per listener. Listeners were presented with pairs in a random order and never heard the same voice more than once.

2.3 Listener testing

Listeners initially consented to the study and provided demographic information about their gender, age, and where in the UK they grew up. They also provided measures of their self-assessed familiarity (on a 0-100 scale) with SSBE, Middlesbrough, and Newcastle accents. The experimental set-up follows the procedure described in Hughes et al. (2022). For each comparison, listeners were first presented with the 8kHz, telephone-quality speech sample. While this audio was playing, they were presented with the question: *This is a {Middlesbrough/Newcastle/SSBE} speaker. How typical is this voice relative to other speakers of the same accent?* They responded on a 0-100 scale. The high-quality (44.1kHz) interview sample then appeared on the screen and participants were asked to respond on 0-100 scales to two further questions: *How similar are the two voices?* and *Do these voices belong to the same speaker?* The two audio clips could be replayed as many times as the participants wanted, and only one clip at a time could be played. After providing their responses to a pair, listeners progressed to the next pair of samples. Listeners were not supervised while participating in the experiment.

Data from 896 listeners from across the UK were collected via Prolific (12 participants were excluded from the data set for not completing the experiment or providing incorrect information in the initial screening stages), which equates to around 60 responses to each comparison pair. On average, listeners took 5.4 minutes (range = 1.5 – 50.3 minutes) to complete the eight comparisons for this study (excluding gaps between comparison pairs). The similarity and typicality responses that listeners provided were divided by 100 to produce a probability on a 0 to 1 scale. Any values of 0 were converted to 0.0001 and any values of 1 were converted to 0.9999. This removed the possibility of LR-scores of 0 or infinity, but did limit scores to between 0.0001 and 9,999. This was not considered problematic as we intended to not use these responses directly as LRs, but rather to calibrate the scores to produce interpretable LRs.

As in Hughes et al. (2022), group-level LR-like scores for each comparison pair were computed by dividing the median similarity response by the median typicality response[2] and then taking the natural logarithm. The reason for averaging over listener responses was to provide a general measure of how listeners perform at voice comparison, as a single point of direct comparison with the output of the ASR system and as a single set of scores which could be fused with the ASR. This allows us to assess whether there is improvement in ASR when combined with scores from the 'average' listener, rather than specifically comparing ASR results with individual listeners (representing judges or jurors) as in Basu et al. (2022). The use of the median as the measure of central tendency reflects the fact that similarity and typicality responses are often highly skewed. It also reduces the effect of the best and worst performing individual participants in evaluating overall group-level performance. This produced a set of 45 SS and 75 DS scores.

2.4 Automatic system testing

ASR testing was conducted using the commercial x-vector Phonexia Voice Inspector (v.4.0.2) system. The system architecture is similar to that described in Snyder et al. (2018); for a fuller description see Jessen et al. (2019) and Morrison et al. (2020). The system requires samples to have an 8kHz sampling rate, as this is what it has been trained on originally. For this reason, the ASR system downsampled the 44.1kHz interview samples to 8kHz prior to analysis. This means that the ASR system analysed different data from the human listeners. We consider this unproblematic as it reflects the standard analytical practice for the ASR system, which would be applied in real world cases involving channel mismatch.

The first stage of ASR processing involves applying voice activity detection which removes all non-speech elements within the signal. Typically, Mel-frequency cepstral coefficients (MFCCs) are extracted from overlapping frames across the speech-active portion of the signal. MFCCs are produced by fitting a discrete cosine transform to the logarithm of the mel-scaled, short-term energy spectrum (see Davis and

[2] Z-score normalising participant responses prior to calculating the median produced much poorer overall performance, so the results are not reported here. Differences in how the scale is used are also removed by using the median for similarity and typicality and then calibrating the data.

Mermelstein,1980). Sequences of MFCC frames (a target frame ±7 adjacent frames) are then fed into a deep neural network (DNN) which has been pre-trained on a large set of diverse data with the aim of recognising speakers. The DNN is a time-delay neural network, which takes the two-dimensional matrices of MFCCs across 15 frames as input. Frame-level information is fed through the network and summarised with a mean and variance for the full recording. These statistics are then fed through further layers of the network and a fixed length vector (x-vector embedding) is extracted prior to the softmax layer (see Jessen et al.,2019; Morrison et al., 2020). This embedding captures speaker-specific information within each sample. Linear discriminant analysis (LDA; see Dehak et al., 2011; Morrison et al., 2020) is applied to the x-vector to remove unwanted dimensions of variability and produce a speaker representation for each sample. LDA is a dimension reduction technique which utilises speaker labels to maximise between-speaker variability and minimise within-speaker variability. A score is then computed by comparing the dimension reduced x-vectors from two samples using probabilistic linear discriminant analysis (PLDA) to produce a LR-like score, accounting for both similarity and typicality (see Prince and Elder, 2007; Kenny, 2010; Morrison et al., 2020). Note that the PLDA scoring model within the system was pre-trained, rather than using data that represented the specific accents in our experiment. In this way, the listeners' estimations of typicality were more specific to the accents in question than those within the PLDA scoring. No adaptation data or reference normalisation was applied. As in 2.3, the ASR analysis produced scores for the 45 SS and 75 DS comparisons.

2.5 Calibration and evaluation

The group-level listener scores and ASR scores were both separately calibrated using logistic regression (Pigeon et al. 2000; Morrison, 2013) with Morrison's (2009) robust implementation of the train_llr_fusion.m function from Brümmer's FoCal toolkit.[3] Calibration is a means of shifting and scaling scores to convert them to log LRs (LLRs). The LLR is a numerical value which is interpretable as a measure of the magnitude of the evidence, where 0 is the threshold between support for the SS proposition (LLRs > 0) and support for the DS proposition (LLRs < 0). Calibration is also a means of reducing the bias within a set of scores (i.e., susceptibility to more false positives or more false negatives), which could be due to a mismatch between the data used to train the system (either human or ASR) and the data used for testing. Unlike Basu et al. (2022), we apply calibration to the listener scores. While this is not reflective of what a jury member would do in a courtroom (i.e., they aren't able to shift and scale their initial response based on knowledge of responses from a set of calibration data), we use calibration in this study to maximise the comparability of listener and ASR performance.

Given the relatively small number of comparisons (principally because of the amount of data that could be provided by the human listeners), calibration was performed via leave-one/two-out cross-validation, whereby each score was calibrated based on scores that did not involve the speaker(s) in either of the samples under comparison. In

[3] https://sites.google.com/site/nikobrummer/focal

practice, this means that scores from comparisons involving all the accents in our experiment were used for the calibration models for both the listeners and ASR system. This, in turn, means that the scores are not evaluated relative to a specific, forensically meaningful pair of propositions. This was a pragmatic decision given the small number of test pairs in each accent set and the fact that general benchmarking of ASR systems often involves calibrating with scores from regionally variable sets of speakers. Cross-validated logistic regression fusion was also applied to combine the scores from the listeners and the automatic system. This applies the same calibration technique as described above, but with multivariate sets of scores to produce a single set of calibrated LLRs for evaluation (see Pigeon et al., 2000; Morrison, 2013).

Overall performance was evaluated using EER and C_{llr} (Brümmer and du Preez, 2006) with the calibrated LLRs (45 SS, 75 DS) as input. EER is a threshold-independent measure of discrimination which captures the percentage error at the point where the percentage of false positives is equal to the percentage of false negatives. C_{llr} assigns a weight to each LLR and penalises the system more heavily based on the magnitude of the contrary-to-fact LLRs (i.e., *errors*), capturing both discrimination and calibration error. Since between-accent tests were only conducted with DS pairs (i.e., no speaker spoke in more than one accent), the results were evaluated using the false positive rate and the DS half of the C_{llr} equation. As highlighted above, the reason for testing DS, between-accent pairs was because of the general confusability of the Newcastle and Middlesbrough accents; i.e., it is conceivable that a non-expert could think a pair of samples belonged to the same speaker despite one being from Newcastle and the other being from Middlesbrough. The performance of individual listeners was also assessed. However, due to the very small number of comparisons per listener (eight in total), uncalibrated scores were used to calculate a C_{llr} and EER. Given the use of uncalibrated scores, the false positive and false negative rates were also calculated separately to assess any bias in listener responses to one type of error over another. Note, however, that listeners responded to different sets of eight comparisons, so caution is needed in interpreting differences in listener performance.

The magnitude of the LRs was also interpreted using a well-established scale (Champod and Evett 2000) which converts numerical LRs into verbal equivalents which express the relative strength of the evidence in support of the same-speaker and different-speaker propositions. An adapted version of this scale is shown in Table 1.

Table 1: Adapted version of the verbal LR scale from Champod and Evett (2000)

Log_{10} LR	**Verbal expression**
0 – ±1	Limited evidence
±1 – ±2	Moderate evidence
±2 – ±3	Moderately strong evidence
±3 – ±4	Strong evidence
±4 – ±5	Very strong evidence

3. Results

3.1 Overall performance

Figure 1 displays the Tippett plot of calibrated LLRs (here we use base-10 LLRs) output by the ASR and by the human listeners. Tippett plots display the cumulative distribution of same-speaker (SS) LLRs, and in the case of Figure 1, the inverse cumulative distribution of different-speaker (DS) LLRs. Tippett plots provide information about the magnitude of the LLRs produced, overall performance (the point at which the SS and DS lines cross is the EER), and the extent of calibration. For more information, see Meuwly, 2001 and Morrison et al., 2021). Overall, the ASR system outperformed the listeners by a considerable margin (ASR: EER=10.89%, C_{llr}=0.48; Listeners: EER=23.55%, C_{llr}=0.75). The magnitude of the LLRs was also slightly greater with the automatic system both for SS (ASR mean SS LLR=0.94; Listener mean SS LLR=0.38) and DS (ASR mean DS LLR=-1.06; Listener mean SS LLR=-0.47), although both produced average LLRs broadly equivalent to *limited evidence*. More detailed analysis of the listener results suggests a weak negative correlation between the number of times an utterance was listened to and accuracy, such that more listens generally lead to slightly poorer performance. This suggests listeners' first intuitions about voices are most reliable for voice comparison. However, this is confounded by the fact that listeners were likely to listen to more difficult pairs more times than easier pairs, and the fact that most utterances were only listened to once or twice.

No improvement in performance was found when fusing the scores from the ASR system and the listeners. In fact, while EER remained the same compared with the ASR in isolation (10.89%), there was a slight increase in the C_{llr} when fused (0.52). There was also a slight decrease in the magnitude of both mean SS and DS LLRs with the fused results compared with the ASR system, although again on average LLRs were equivalent to *limited evidence*. However, analysis of the individual contrary-to-fact LLRs (i.e., *errors*) produced by the ASR system reveals some interesting effects in terms of the human responses. The ASR system produced five contrary-to-fact SS LLRs. For three of these comparisons, the listeners produced LLRs within the range of *limited evidence* for the SS proposition (i.e., in the correct direction). For the remaining two SS ASR errors, the listeners also produced errors, but these were around one order of magnitude weaker than the ASR system. The patterns for DS comparisons were more mixed. Of the three DS errors produced by the ASR system, only one correctly provided support for the SS proposition within the listener results.

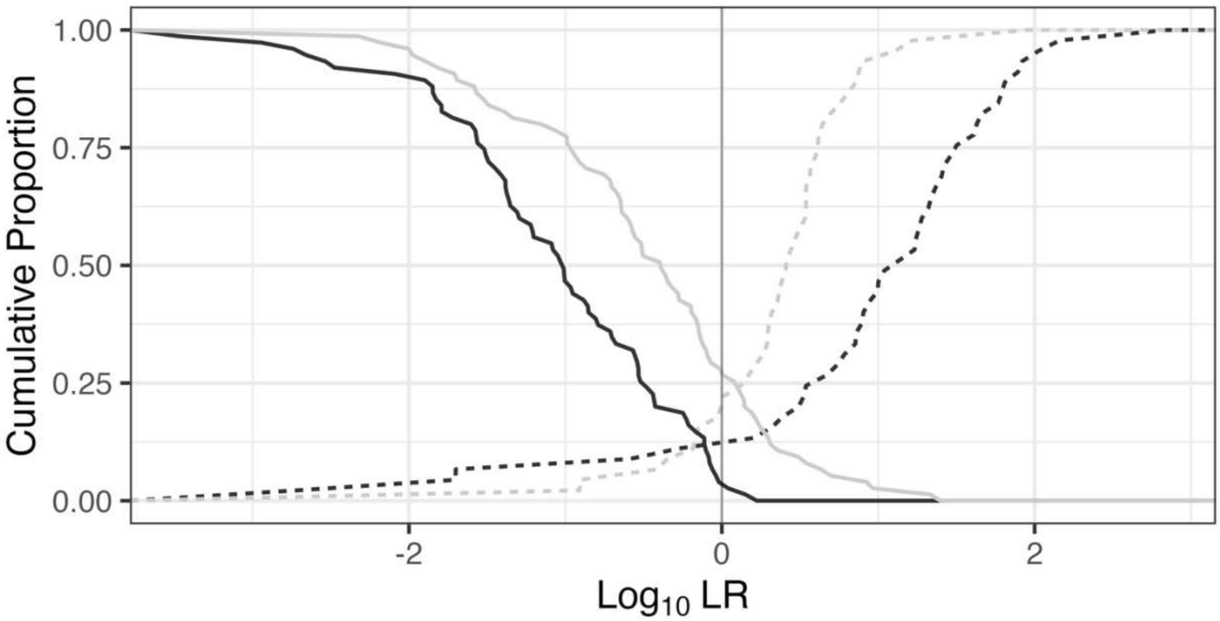

Figure 1: Tippett plot of the SS (dashed line, right) and DS (solid line, left) LLRs produced by the ASR system (black) and human listeners (grey)

3.2 Accent effects

ASR and listener performance within accents was also considered. The ASR system performed best with the Middlesbrough samples (EER=0%, C_{llr}=0.24) and markedly worse with the Newcastle samples (EER=13.33%, C_{llr}=0.711). Note here that the relatively large difference in EER in fact represents just four Newcastle comparisons (3 SS and 1 DS) producing contrary-to-fact evidence. This reflects the relatively small sample size within each accent group. Intermediate performance was found for the ASR system with the SSBE samples (EER=13.33%, C_{llr}=0.374). Less variable, but overall poorer, performance was found for the human listeners. The listeners performed best with the Newcastle (EER=20%, C_{llr}=0.70) and SSBE samples (EER=20%, C_{llr}=0.77), and worst with the Middlesbrough samples (EER=33.33%, C_{llr}=0.84). This difference could reflect a familiarity effect (see Figure 2 which displays the underlying distribution of familiarity ratings for each accent within our data set), whereby listeners are more aware of, and have greater general exposure to, Newcastle English and SSBE compared with Middlesbrough English. No improvements to the within-accent ASR performance were found when fused with the listener scores.

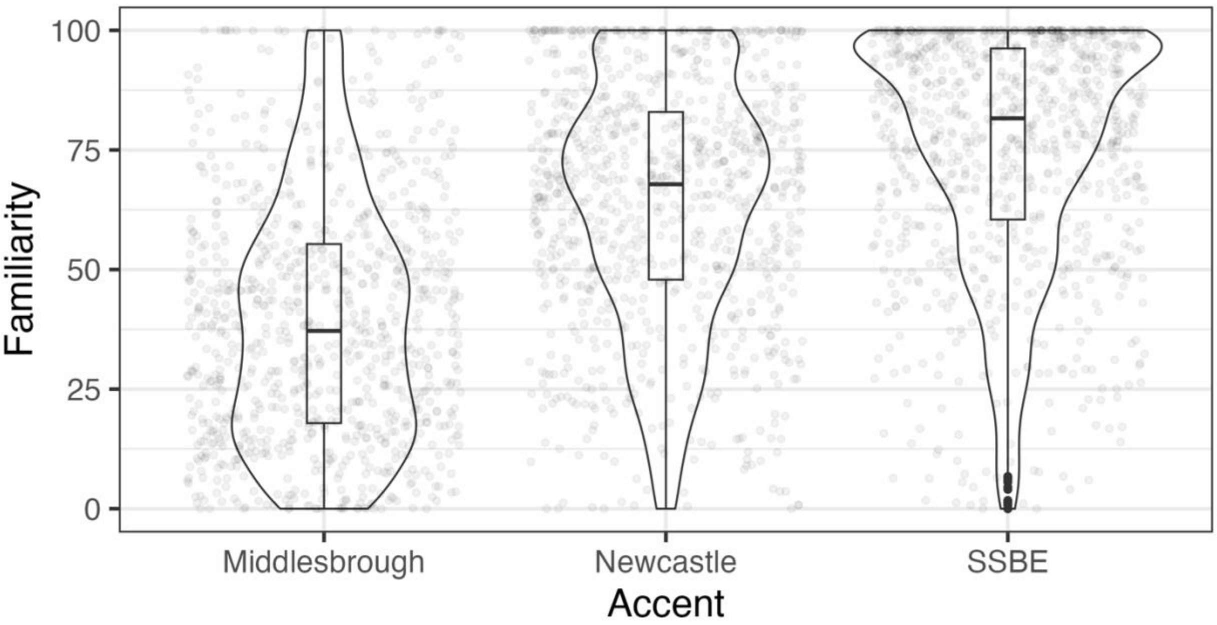

Figure 2: Distributions of self-reported familiarity responses across the 896 participants for Middlesbrough (left), Newcastle (middle) and SSBE (right)

Interesting patterns of performance were found in the between-accent DS comparisons. For the high salience between-accent comparisons, listener performance (EER=6.67%, C_{llr}=0.35) was comparable to that of the ASR system (EER=6.67%, C_{llr}=0.28). Further, this was the only condition in which the fusion of ASR and listener scores produced an improvement in performance over the ASR in isolation (EER=0%, C_{llr}=0.21), indicating that the listeners were sensitive to complementary speaker-specific information in these comparisons. For both the listeners and the ASR system, there was a marked drop-off in performance from the high salience to the low salience between-accent comparisons. For the listeners, this condition produced the poorest overall performance compared with any of the within-accent subsets (EER=33.33%, C_{llr}=1.17), with the C_{llr} suggesting poor calibration of LLRs in addition to poor discrimination. For the ASR system, performance with the low salience between-accent comparisons was actually better than with the Newcastle subset (EER=13.33%, C_{llr}=0.51), but poorer than with the Middlesbrough and SSBE subsets.

3.3 Individual listeners

The listener results presented in Sections 3.1 and 3.2 reflect averages across all 896 participants in our study. As in previous studies (e.g., Basu et al. 2022), considerable variability in performance was found across listeners' uncalibrated scores. Note, though, that results need to be interpreted with caution given that listeners responded to different sets of eight comparisons (three SS, five DS). In our study, by-listener EERs ranged from 0% to 100%, with 58 of the 896 producing EERs of 0%. Of those 58 listeners, only 15 produced 0% errors based on an LLR threshold of 0. The uncalibrated scores also showed a bias towards false negatives (across listeners, mean false negative was 61.05% compared with mean false positive rate of 21.52%). The best performing listener produced a C_{llr} of 0.52, while the worst produced a C_{llr} of 5.09. This suggests that there is considerable by-listener variability both in terms of discrimination and calibration error. All the listeners with 0% EER produced C_{llr}s below 1, indicating that they are all capturing useful speaker-specific information in a way that many of the other listeners are not. The small number of comparisons per listener mean it is not possible to break down listener performance by accent subset.

4. Discussion

As has been demonstrated previously (e.g., Basu et al. 2022), in our study the state-of-the-art automatic system outperforms human listeners at the task of speaker recognition. This is true not only in terms of overall listener performance, but also when compared on a by-listener basis with even the very best performing listeners. Overall performance of both the human listeners and the ASR system is relatively poor. This reflects the challenging nature of the materials under analysis (very short, channel-, quality-, and style-mismatched samples) and is conceivably representative of ASR and listener performance with real forensic materials. Despite this, both approaches produce C_{llr} values below 1 – and in the case of the ASR system, considerably less than 1. Both approaches are thus capturing useful speaker-specific information for discriminating between SS and DS pairs. In the following sections, we discuss our findings in terms of what they tell us about human and ASR processing during speaker recognition, and we consider the implications for forensic speech science.

4.1 Human and ASR processing

The results of this study provide insights into the different processing strategies used by human listeners compared with automatic systems when recognising unfamiliar speakers. This is nicely demonstrated through analysis of the between-accent DS comparisons. In the high salience condition, listeners produce markedly better performance compared with their overall performance, or compared with any of the within-accent conditions. This suggests that listeners are intuitively sensitive to the salient segmental accent differences between Middlesbrough and Newcastle and respond to these pairs of samples in an almost categorical way (i.e., the accent difference means that the voices must belong to different speakers). When those segmental cues are not available, such as in the case of the low salience between-accent comparisons, listener performance is markedly poorer (in fact, this is the condition where listeners perform worst). Conversely, the performance of the ASR system with the high salience between-accent comparisons is slightly poorer than with Middlesbrough-only comparisons, although better than with the Newcastle-only and SSBE-only samples. This suggests that the ASR system is much less sensitive to important segmental differences between accents that can be utilised for speaker recognition and also that the ASR system approaches variability in a continuous, rather than categorical, way. This may be because the samples used here were relatively short (c. 10s) and so the system doesn't have enough information to capture the acoustic differences between speakers of different accents.

In 3.1, we reported a lack of improvement in speaker recognition performance when fusing the ASR and listener scores, i.e., the addition of average listener scores was not capable of improving performance over and above what is achieved by the ASR system alone. While this provides evidence as to relative discrimination performance, it provides little insight into the extent to which ASR systems and human listeners are sensitive to the same or different speaker-specific information when conducting a speaker recognition task. This is because the average performance of the human listeners is so much poorer than that of the ASR system, such that the listeners are, on average, capturing much less speaker-specific information. This in turn is driven by the fact that there is so much variability in performance across the listeners, with the poorest listeners producing EERs of 100%, and so capturing little to no useful speaker-discriminatory information. To understand more about the complementarity of ASR and listener processing, in this section we focus only on the results produced by the best performing listeners. This is explicitly a *post hoc* analysis intended to examine the speaker-specific information captured by the best performing listeners, removing some of the noise from the poorest performing listeners.

Using data from only the 58 listeners who produced 0% EERs, we re-ran the tests conducted in 3.1 (note that there was no obvious correlation between self-reported familiarity and performance, with the 58 best performing listeners reporting a wide variety of familiarity with all three accents). Because of the between-subjects nature of the human listener experiment, data from only 104 (39 SS, 65 DS) of the 120 comparisons were available for analysis. With this subset of the data, the listener performance (EER=12.56%[4], C_{llr}=0.42) was much closer to that of the ASR system (EER=7.69%, C_{llr}=0.37). Fusing scores from the two approaches did not produce an improvement in EER, but did considerably reduce the C_{llr} to 0.277. This demonstrates that the best performing listeners are capturing complementary information to that of the ASR, at least in terms of improving calibration (if not discrimination). Further, Figure 3 displays the relationship between the LLRs produced by the ASR and the best performing listeners for each of the comparisons in our subset. No correlation is found between the two sets of results when analysing SS and DS pairs separated, such that ASR LLRs predict less than 1% of the variance in the listener LLRs. A stronger overall correlation was found when pooling data from SS and DS comparisons (R^2=0.254), but this captures the fact that both listeners and ASR are generally separating SS and DS pairs to some extent. The lack of correlation is exemplified by the fact that one DS comparison produces a LLR of -7.56 for the listeners (three orders of magnitude higher than the *very strong evidence* category on Champod and Evett's (2000) verbal scale), while only producing a LLR of -0.28 for the ASR (equivalent to *limited evidence*). Conversely, another DS LLR produces a LLR of -5.22 for the ASR, but 0.55 (i.e., contrary-to-fact *limited evidence*) for the listeners. Interestingly, this comparison is also a between-accent high salience pair.

[4] Note that the combination of LR-like scores produces an EER of 12.56% despite individuals producing 0%. This is because the scores are uncalibrated when they are averaged, meaning that different listeners produce scores at different points on the scale. In the averaging process, this leads to what become contrary-to-fact LLRs post-calibration.

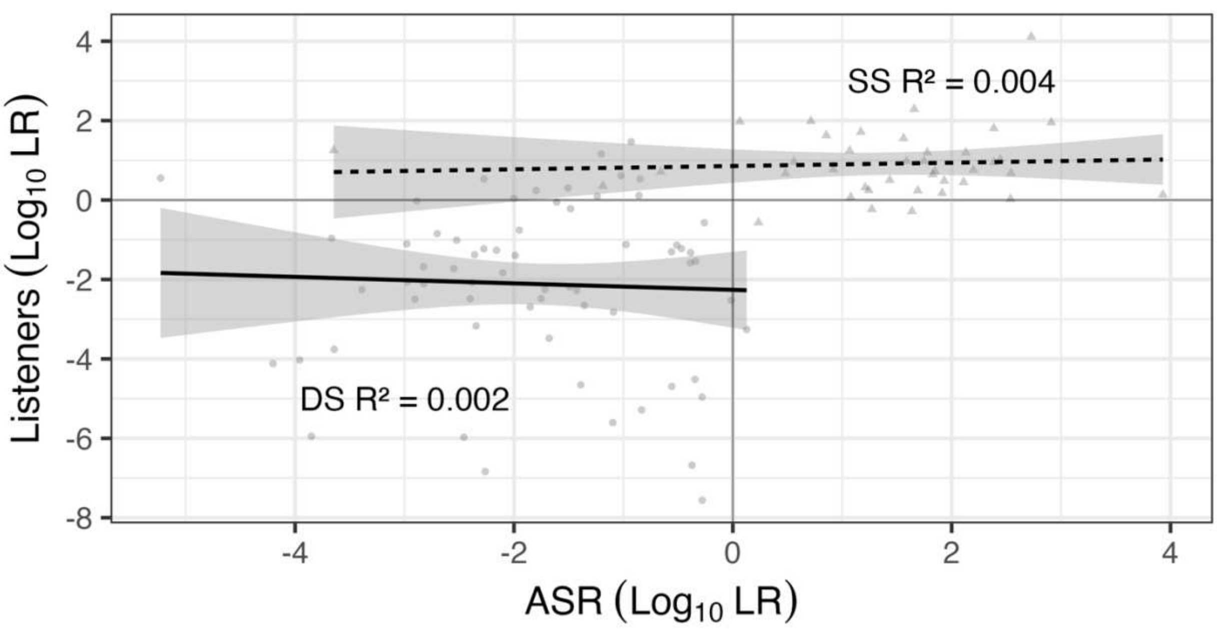

Figure 3: Scatter plot with linear regression lines (and 95% confidence intervals) for calibrated SS (triangles/ dashed line, top right) and DS (circles/ solid line, bottom left) LLRs output by the ASR system (x-axis) and the average of the best performing 58 listeners (y-axis) based on a subset of 104 comparisons (39 SS, 65 DS)

Taken together, these data indicate that listeners are sensitive to different speaker-specific information when making speaker recognition judgments compared with automatic systems. Listeners appear to respond to segmental differences and are particularly able to utilise their pre-existing linguistic knowledge of regional accents to aid with speaker recognition. The ASR system is less sensitive to such information in part potentially because of the shortness of the samples, but also because of the extraction of MFCCs and pooling of data in generating speaker embeddings.

4.2 Implications for forensic speech science

With our data, the ASR system provides better voice comparison performance compared with averaging across naive human listeners. This is true even with the additional benefit of calibrating listener responses to reduce calibration error (unlike the Basu et al. 2022 study), producing more comparable results with those from the ASR system. The ASR system also outperforms even the very best individual listeners - although it is worth reiterating that by-listener results were not calibrated due to the small amount of data. While our study did not intend to replicate the decision-making process of individual jurors or judges, the results demonstrate that ASR is providing speaker-specific information for separating SS and DS pairs beyond what could reasonably be expected of the average lay listener – a criterion in the England and Wales Crown Prosecution Service’s definition of expert evidence.

Of more interest is the fact that listener and ASR strategies of voice comparison appear to be somewhat complementary, such that the best performing listeners are capable of improving on the baseline performance of an ASR system. This raises an interesting question about the extent to which the views of potential members of a jury may be usefully integrated with ‘expert evidence’ (in this case, the output of an ASR system) in order to improve the validity of the evidence. Given how much variability there is across listeners, however, the key issue would be whether it is possible to pre-emptively predict which listeners are likely to be ‘good performers’ for speaker recognition. *Post hoc* analysis of our data reveals no demographic predictors of good listener performance. The exception here is that familiarity with an accent appears to aid performance, at the group level, as evidenced through poorer performance with samples of the much less well-known Middlesbrough accent. (No clear correlation was found between individual listener performance and familiarity, although the number of data points per listener, per accent is very small.) We also considered the performance of listeners on other levels within our game as a means of predicting performance with the stimuli in the present study. This was done by fusing responses from subsets of listeners identified as ‘good’ performers in other levels with the output of the ASR system. However, this did not prove to be a useful diagnostic, with none of these listener groups providing the same improvements in performance as the best listeners from within the data set presented in this paper.

5. Conclusion

In this paper, we have explored the relative and combined performance of an ASR system and human listeners using stimuli from a range of British English accents, including a set of between-accent comparisons. We used a novel methodology to produce comparable data from the automatic system and the listeners. Consistent with previous research, our results suggest that ASR systems and well performing listeners are sensitive to complementary speaker-specific information due to the differences in the ways they process speech. Future work will continue to explore factors explaining group- and individual-level differences in listener performance, specifically in terms of what phonetic information is available in a sample (e.g., the type and frequency of regionally-marked features, and more global features such as voice quality). We will also compare the findings reported in this paper with the results for the other levels in our game in order to assess the extent to which listeners’ judgments are affected by other information in a criminal case and the conclusion of a forensic expert.

Acknowledgements

This work was supported by a UK Arts and Humanities Research Council Early Career Grant (AH/T012978/1) entitled *Humans and Machines: Novel Methods for Assessing Speaker Recognition Performance*. Thanks also to Joe Cutting and Dan Slawson who helped us develop our game and build the infrastructure to run it as an experiment.

References

Afshan, A., Kreiman, J. and Alwan, A. (2020) Speaker discrimination in humans and machines: effects of speaking style variability. In *Proceedings of Interspeech*, October 25-29, 2020, Shanghai, China, pp. 3136–3140.

Afshan, A., Kreiman, J. and Alwan, A. (2022) Speaker discrimination performance for 'easy' versus 'hard' voices in style-matched and -mismatched speech. *Journal of the Acoustical Society of America* 151: 1393–1403.

Aitken, C. G. G., Taroni, F. and Bozza, S. (2021) *Statistics and the Evaluation of Evidence for Forensic Scientists (3rd ed).* Wiley: Hoboken, NJ.

Atkinson, N. (2015) *Variable Factors Affecting Voice Identification in Forensic Contexts*. PhD dissertation. University of York, UK.

Basu, N., Bali, A. S., Weber, P., Rosas-Aguilar, C., Edmond, G., Martire, K. A. and Morrison, G. S. (2022) Speaker identification in courtroom contexts – Part I: Individual listeners compared to forensic voice comparison based on automatic-speaker-recognition technology. *Forensic Science International* 341: 111499.

Braun, A, Llamas, C., Watt, D., French, J. P. and Robertson, D. (2018) Sub-regional 'other accent' effects on lay listeners' speaker identification abilities: a voice line-up study with speakers and listeners from the North East of England. *International Journal of Speech, Language and the Law* 25: 231–255.

Brümmer, N. and du Preez, J. (2006) Application-independent evaluation of speaker detection. *Computer Speech and Language* 20: 230–275.

Bull, R. and Clifford, B. R. (1984) Earwitness voice recognition accuracy. In G. L. Wells and E. F. Loftus (eds) *Eyewitness Testimony: Psychological Perspectives.* CUP: Cambridge. pp. 92–123.

Champod, C. and Evett, I. W. (2000) Commentary on A. P. A. Broeders (1999) 'Some observationson the use of probability scales in forensic identification' Forensic Linguistics 6(2): 228-41. *Forensic Linguistics* 7: 239–243.

Corretge, R. (2023) Praat Vocal Toolkit. https://www.praatvocaltoolkit.com/index.html (Last viewed November 10, 2023).

Das, R. K. and Prasanna, S. R. M. (2016) Exploring different attributes of source information for speaker verification with limited test data. *Journal of the Acoustical Society of America* 140: 184–190.

Davis, S. B. and Mermelstein, P. (1980) Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing* 28(4): 357–366.

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P. and Ouellet, P. (2011) Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech and Language Processing* 19(4): 788–798.

Greenberg, C., Martin, A., Brandschain, L., Campbell, J., Cieri, C., Doddington, G. and Godfrey, J. (2010) Human assisted speaker recognition in NIST SRE2010. In *Proceedings of Odyssey: The Language and Speaker Recognition Workshop*, June 28-July 1, 2010, Brno, Czech Republic, pp. 180–185.

Hautamäki, V., Kinnunen, T., Nosratighods, M., Lee, K.-A., Ma, B. and Li, H. (2010) Approaching human listener accuracy with modern speaker verification. In

Proceedings of Interspeech, September 26-30, 2010, Makuhari, Chiba, Japan, pp. 1473–1476.

Hughes, V., Llamas, C. and Kettig, T. (2022) Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions. In *Proceedings of Interspeech*, September 18-22, 2022, Incheon, Korea, pp. 5238–5242.

Hughes, V., Wormald, J., Foulkes, P., Harrison, P., Kelly, F., van der Vloed, D., Welch, P. and Xu, C. (2023) Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics. In *Proceedings of Interspeech*, August 20-24, 2023, Dublin, Ireland, pp. 591–595.

Jessen, M., Bortlík, J. Schwarz, P. and Solewicz, Y. A. (2019) Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). *Speech Communication* 111: 22–28.

Johnson, J., McGettigan, C., and Lavan, N. (2020) Comparing unfamiliar voice and face identity perception using identity sorting tasks. *Quarterly Journal of Experimental Psychology* 73: 1537–1545.

Kahn, J., Audibert, N., Rossato, S., and Bonastre, J. F. (2011) Speaker verification by inexperienced and experienced listeners vs. speaker verification system. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 22-27, Prague, Czech Republic, pp. 5912–5915.

Kenny, P. (2010) Bayesian speaker verification with heavy tailed priors. In *Proceedings of Odyssey: The Language and Speaker Recognition Workshop*, June 28-July 1, 2010, Brno, Czech Republic. [Keynote]

Kerswill, P. (2003) Dialect levelling and geographical diffusion in British English. In D. Britain and J. Cheshire (eds) *Social Dialectology*. Benjamins: Amsterdam. pp. 223–243.

Kreiman, J. and Papcun, G. (1991) Comparing discrimination and recognition of unfamiliar voices. *Speech Communication* 10: 265–275.

Laver, J. (1980) *The Phonetic Description of Voice Quality*. CUP: Cambridge.

Legge, G. E., Grosmann, C. and Pieper, C. M. (1984) Learning unfamiliar voices. *Journal of Experimental Pyschology: Learning, Memory and Cognition* 10: 298–303.

Lindh, J. and Morrison, G. S. (2011) Humans versus machine: forensic voice comparison on a small database of Swedish voice recordings. In *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, August 17-21, 2011, Hong Kong, pp. 1254–1257.

Llamas, C. (2015) Middlesbrough. In R. Hickey (ed) *Researching Northern English.* Benjamins: Amsterdam. pp. 251–270.

Llamas, C., Watt, D. and French, J. P. (2016-19) *The Use and Utility of Localised Speech Forms in Determining Identity: Forensic and Sociophonetic Perspectives.* ESRC-funded project (ES/M010783/1).

Meuwly, D. (2001) *Reconnaissance de locuteurs en sciences forensiques: l'apport d'une approche automatique*. PhD dissertation. University of Lausanne, Switzerland.

Morrison, G. S. (2009) train_llr_fusion_robust.m [Software release 2009-07-02]. http://geoff-morrison.net/#TrainFus

Morrison, G. S. (2013) Tutorial on logistic-regression calibration and fusion: Converting a score to a likelihood ratio. *Australian Journal of Forensic Sciences* 45: 173–197.

Morrison, G. S. and Enzinger, E. (2018) Score based procedures for the calculation of forensic likelihood ratios – Scores should take account of both similarity and typicality. *Science and Justice* 58: 47–58.

Morrison, G. S. and Enzinger, E. (2019) Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Conclusion. *Speech Communication* 112: 37–39.

Morrison, G.S., Enzinger, E., Ramos, D. and Gonzalez-Rodriguez, J. (2020) Statistical models in forensic voice comparison. In Banks, D.L., Kafadar, K., Kaye, D.H. and Tackett, M. (eds.) *Handbook of Forensic Statistics*. Boca Raton, FL: CRC. pp. 451-497.

Nolan, F., McDougall, K., de Jong, G. and Hudson, T. (2009) The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language and the Law* 16: 31–57.

Park, S. J., Yeung, G., Vesselinova, N., Kreiman, J., Keating, P. and Alwan, A. (2018) Towards understanding speaker discrimination abilities in humans and machines for text-independent short utterances of different speech styles. *Journal of the Acoustical Society of America* 144: 375–386.

Pigeon, S., Druyts, P. and Verlinde, P. (2000) Applying logistic regression to the fusion of the NIST'99 1-speaker submissions. *Digital Signal Processing* 10: 237–248.

Plante-Hébert, J. and Boucher, V. J. (2015) Effects of nasality and utterance length on the recognition of familiar speakers. *Proceedings of the International Congress of Phonetic Sciences (ICPhS)*, August 10-14, 2015, Glasgow, UK.

Prince, S. J. D. and Elder, J. H. (2007) Probabilistic linear discriminant analysis for inferences about identity. In *Proceedings of the IEEE 11th International Conference on Computer Vision*, pp. 1–8.

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

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018) X-vectors: Robust DNN embeddings for speaker recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 15-18, 2018, Calgary, Canada, pp. 5329–5333.

Roebuck R and Wilding J. (1997) Effects of vowel variety and sample length on identification of a speaker in a line-up. *Applied Cognitive Psychology* 7: 475–481.

Van Wallendael, L. R., Surace, A., Hall Parsons D. and Brown, M. (1994) 'Earwitness' voice recognition: factors affecting accuracy and impact on jurors. *Applied Cognitive Psychology* 8: 661–677.