Eliciting and evaluating likelihood ratios for speaker recognition by human listeners under forensically realistic channel-mismatched conditions

Vincent Hughes, Carmen Llamas, Thomas Kettig

Department of Language and Linguistic Science, University of York, UK

{vincent.hughes|carmen.llamas|thomas.kettig}@york.ac.uk

Abstract

This paper describes an experiment which elicits and then evaluates LR-like scores from non-expert, human listeners in a speaker recognition task under conditions reflective of forensic casework. In doing so, it provides a framework for comparing and combining listener judgements with the output of ASR systems (or other data-driven speaker recognition approaches). Stimuli consisted of 45 same-speaker and 45 different-speaker pairs of voices from young, male speakers of Standard Southern British English, using 10 second, channel-mismatched samples. 81 listeners provided ratings of the similarity between voices and their typicality within the wider accent population, which in turn were used to calculated LR-like scores. These scores were converted to log LRs via cross-validated logistic regression calibration. Overall, the human listeners produced an EER of 26.67% and a C_{llr} of 0.773. However, considerable variation was found across individual listeners (13.3-66.7% EER). Fusion of the listener judgements with an x-vector ASR system provided very marginal improvement in performance compared with the ASR system in isolation. Importantly, the magnitude of the four errors made by the ASR system were reduced because of the listener judgements. The implications of this work for forensics will be discussed.

Index Terms: speaker recognition, forensic voice comparison, human listeners, likelihood ratio, validation

1. Introduction

1.1. Speaker recognition by humans

The ability to recognise speakers from their voice is a key facet of human perception and one that listeners rely on daily (e.g. when identifying someone in another room or over the telephone). While non-expert, human speaker recognition abilities can be relatively good with familiar, or at least predictable, voices [1], much less is known about performance with unfamiliar voices (although much related work has been conducted within the area of voice parades). Some recent work has attempted to examine this issue empirically, to generate data that can be analysed in the same way as data-driven methods, such as automatic speaker recognition (ASR) systems. [2] demonstrated that listeners are sensitive to style, producing equal error rates (EER) of as low as 6.96% with pairs of style-matched, read-speech recordings, but EERs of as high as 20.68% in style-mismatched conditions. Analysis of errors also revealed, predictably, that some speakers are more or less difficult for listeners to distinguish. [3] compared listener performance with ASR, finding marginal improvements when fusing the two. This indicates that listeners and ASR systems are sensitive to different information within the speech signal when making judgements about identity.

However, very little work has been conducted to address such issues within the domain of forensic voice comparison (FVC). There are two principal reasons why human speaker recognition performance takes on additional importance in forensics. Firstly, lay people in many jurisdictions (of which the UK is one) sit on juries and are required to make decisions about the ultimate innocence or guilt of defendants. In cases involving voice evidence, recordings are often played to juries. In this way, jurors are likely to be making judgements about the voices they hear. Whilst they may also be given expert opinion (in the form of a report or oral evidence), little is known about how listeners' own judgements override or interact with those of an expert. It is therefore important to establish how well listeners can separate same- (SS) and different-speaker (DS) pairs under conditions reflective of FVC casework. Secondly, there has been growing interest in the relative merits of different analytic methods for FVC (i.e. ASR, componential linguistic-phonetic analysis, holistic auditory analysis by experts and lay people). A key element of this has been to investigate the relative complementarity of different methods and to establish whether their combination improves overall speaker recognition performance. Much of this work has focused on comparing and combining acoustic-phonetic variables with the output of ASR systems [4] or attempts to understand error patterns produced by automatic systems using auditory analysis conducted by expert linguists [5].

Only a limited amount of research has been conducted to understand how lay listeners perform at speaker recognition within forensics. The NIST SRE 10 evaluation included a human-assisted speaker recognition element [6], involving judgements made by lay listeners in response to two sets of trials. These trials (particularly the very small HASR1 set) included SS and DS pairs that had produced contrary-to-fact results (i.e. errors) when tested with an ASR system. Listeners provided only accept-reject decisions about whether they thought pairs belonged to same- or different-speakers (i.e. posterior probability judgements). Overall performance was relatively poor, with EERs across both sets of trials between 30% and 40%, although with some variability across listeners. This also included considerable calibration error, with some listener groups producing very low false positive rates but very high false negative rates (or vice versa).

[7] extended this work to conduct more systematic testing of human listeners, treated as *systems* in the same way as any other linguistic-phonetic or ASR system. This involved testing with 45 high-quality, channel-matched pairs of SS (nine pairs) and DS (36 pairs) samples of native Swedish speakers and a panel of 52 lay listeners. Listeners provided similarity judgements on a 5-point Likert scale, which were then averaged and converted to a likelihood ratio (LR)-like score. Scores for each comparison were calibrated using logistic regression and overall performance was compared with the calibrated output

of a GMM-UBM ASR system tested on the same set of comparisons. Listeners optimally produced a C_{llr} [11] of 0.359. Much poorer performance (C_{llr} = 0.687) was reported when the same task was conducted using recordings played backwards, indicating that listeners utilise segmental information in making identity judgements. In comparison, the ASR system produced better overall performance (C_{llr} = 0.033 with the forwards samples).

1.2. Comments on previous research

With the exception of the work referenced above, few studies have tested listeners using the same methods of empirical testing as other data-driven speaker recognition system (either automatic or linguistic). Fewer still have empirically fused human judgements with ASR to assess combined performance. One key reason for this has been the considerable challenge in extracting judgements from humans that are both logically and empirically comparable with the output of data-driven systems. Some studies have simply used binary accept-reject decisions [6]. Other studies [3,7] have used scores based on similarity-only judgements. However, this does not explicitly include a judgement of typicality which is necessary to produce LR-like scores of the form:

$$\frac{p(E|H_{SS})}{p(E|H_{DS})} \quad (1)$$

and essential within the forensic domain [8,9], in order to assess the strength of the evidence. Forensic research on human speaker recognition has also typically been limited by using small numbers of samples and/or materials that are not forensically realistic. Finally, as in much speaker recognition research, the typical focus of previous work has been on overall performance (i.e. error rates), rather than on the evidential results for individual pairs of speakers.

1.3. This study

The present study contributes to this area by conducting larger-scale testing and evaluation of the speaker recognition performance of human listeners with unfamiliar voices, in a way that is analogous to testing conducted within data-driven FVC approaches (e.g. ASR). Specifically, we address the following research questions: (i) Do listener judgements about similarity and typicality produce LR-like scores that can be calibrated and evaluated like other speaker recognition systems? How do these scores compare with other types of scores (e.g. accept-reject, similarity-only)? (ii) What is the overall performance of human listeners and what is the strength of the evidence that they produce? (iii) How does listener performance compare with ASR performance?

2. Methodology

2.1. Speakers

The speech samples used as stimuli in the experiment were taken from the Dynamic Variability in Speech (DyViS) database of young (18-25), male speakers of Standard Southern British English (SSBE) [10]. Of the available 100 speakers, 45 were chosen at random. For each speaker, four 10-second edited samples were created; two taken from DyViS Task 1 (mock police interview) and two from Task 2 (exchange with a mock accomplice). The Task 1 samples were high-quality studio recordings (44.1kHz, 16-bit depth), while the Task 2 samples were far-end, landline telephone recordings (8kHz, 16-bit depth). The channel-mismatch between samples was intended to replicate common conditions within FVC casework. The files had been edited to remove as much contextual information as possible. The final set of stimuli consisted of 45 same-speaker (SS) and 45 different-speaker (DS) pairs.

2.2. Participants

81 participants aged between 18 and 74 from the United Kingdom with L1 English were recruited via Prolific to take part in the experiment (a further nine were excluded for not meeting demographic requirements). No listeners reported issues with hearing. Each participant was randomly assigned to one of three blocks. Each block contained 30 comparisons, balanced for same- (SS; 15) and different-speaker (DS; 15) pairs. Within the blocks, SS and DS pairs were presented in a random order. This produced up to 28 responses for each of the 90 comparisons.

2.3. Experimental set-up

Participants provided demographic information, as well as a judgement on a 0-100 scale to indicate how familiar they are with the SSBE accent. For each comparison, participants were first presented with a telephone sample. In line with typical FVC conditions, this was used as the nominal 'criminal' or 'unknown' sample. Participants provided a judgement on a 0-100 scale to indicate how typical they considered this voice to be relative to other speakers of the same accent (the wording of this question had been piloted to ensure that participants understood what was being asked of them). They were then presented with an interview sample (the nominal 'suspect' or 'known' sample) and asked to provide a judgement of the similarity between this and the Task 2 sample on a 0-100 scale. Finally, participants were asked to indicate on a 0-100 scale whether they thought the two voices belonged to the same speaker.

2.4. Listener score computation

Scores for each comparison were computed in a variety of ways using averages across listeners. For all types of scores, averaging was conducted using both the mean and the median, given the expectation for skew in scores across listeners. All similarity and typicality responses were firstly converted to a 0-1 scale. An offset of 0.0001 was applied to any values of 0 or 1 to avoid infinity values in the scores. In line with previous work [3,7], scores were based on (i) posterior probability (i.e. the judgement about whether the listeners thought the samples contained the same or different voices) and (ii) similarity-only judgements computed as $p/(1-p)$ where p is the average similarity rating. LR-like scores which included explicit judgements of typicality were also computed as (iii) average similarity divided by average typicality and (iv) LR-like scores computed by individual and then averaged. These individual LR-like scores were also used to assess individual listener performance. In the case of methods (ii)-(iv), natural log transforms were applied to scores prior to calibration.

2.5. ASR score computation

ASR scores were computed using the Phonexia Voice Inspector x-vector ASR system. This involves conversion of raw acoustic data to x-vector embeddings via a deep neural network. Probabilistic linear discriminant analysis is then used to compute scores which take into account both the similarity

between the two samples and their typicality (based on training data within the system).

2.6. Calibration and evaluation

Both the listener and ASR scores were converted to calibrated $\log_{10}$ LRs using logistic regression [12]. Given the relatively small number of comparisons, calibration was conducted using cross-validation whereby each score was calibrated based on coefficients derived from comparisons that did not involve any scores from the target speaker(s). In this way, calibration coefficients changed slightly for each comparison. Evaluation of performance was conducted using equal error rate (EER) and the log LR cost function (C_{llr} [11]; a validity metric based on the magnitude of contrary-to-fact LRs). In both cases, better performance is indicated by a value closer to 0. The strength of evidence produced by the different methods was also assessed with reference to the verbal scale in [13].

3. Results

3.1. Comparison of listener scores

Table 1 displays overall performance based on listener responses using the four different methods of score computation and two averaging methods. The poorest performance was found using the mean LR (EER = 64.44%, C_{llr} = 1.11). This is due to the considerable skew in the distribution of LRs for each comparison, due to a small number of extreme contrary-to-fact responses from certain listeners. With the exception of mean LR, all of the other scoring methods produced C_{llr}s of less than 1, indicating that the listener judgements are capturing some useful speaker-discriminatory information. However, on the whole, listener performance is relatively poor. C_{llr}s are also consistent across scoring methods, although there are marginal improvements when using medians rather than means (again, due to slight skew in the responses across listeners). There is some variability in EERs across the different methods, however, given the size of the data set, these differences reflect only small differences in the absolute number of comparisons that produce contrary-to-fact results.

Table 1: *EER (%) and C_{llr} values for each of the human scoring and averaging methods.*

	Score	EER	C_{llr}
Mean	(1) Posterior	24.44	0.69
	(2) Similarity-only	22.22	0.76
	(3) Similarity-typicality LR	31.11	0.77
	(4) Average LR	64.44	1.11
Median	(1) Posterior	20.00	0.74
	(2) Similarity-only	22.22	0.74
	(3) Similarity-typicality LR	26.67	0.76
	(4) Average LR	30.00	0.76

Given the similarities across methods, for the remainder of this paper we focus on the results based on median similarity and typicality judgements (3). This is because these scores are LR-like and incorporate an explicit judgement about typicality. Figure 1 displays the Tippett plot of calibrated output based on median similarity and typicality LRs. The overall strength of evidence produced is relatively weak (mean SS and DS LLRs are within the range of *limited evidence* according to the scale in [13]). However, for some DS comparisons much stronger log LRs are produced (for one pair this was equivalent to *strong evidence* in support of the DS proposition). The magnitude of contrary-to-fact SS log LRs is greater than for DS log LRs. Evaluation of the score distributions prior to calibration also reveals that generally listener judgements are miscalibrated towards making higher numbers of false negatives (i.e. SS pairs producing log scores less than 0), whereas performance with the DS pairs is relatively good overall. This suggests that, within this dataset, listeners find SS pairs somewhat more difficult than DS pairs.

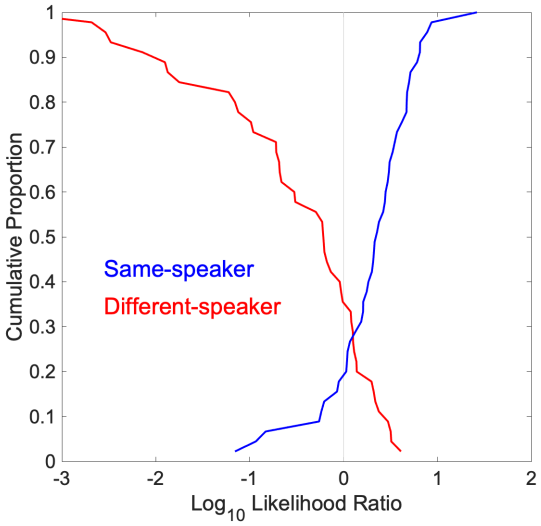

Figure 1: *Tippett plot of calibrated $\log_{10}$ SS (blue) and DS (red) LRs based on median similarity and typicality judgements across listeners (EER = 26.67%, C_{llr} =0.76).*

3.2. Performance of individual listeners

Since each listener only provided judgements for a subset of 30 comparisons (15 SS and 15 DS), there was not enough data to calibrate results by-listener. It is also likely that different listeners produce scores that are miscalibrated to different extents and in different directions, such that it isn't appropriate to use the calibration coefficients from §3.1 to apply to individual listener scores. Thus, we present here by-listener EERs as a calibration-independent measure of discrimination. Considerable variability was found across listeners in terms of their performance, with the best performing listeners achieving an EER of 13.3%, while the poorest performing listeners achieved an EER of 66.7%. We might have predicted that familiarity with the accent would aid discrimination (analogous to a well-trained data-driven system), however, no correlation was found.

3.3. Fusion of listener and ASR scores

The ASR system (EER = 4.44%, C_{llr} = 0.26) produced better overall performance compared with the human listeners. Of the 90 comparisons conducted, the ASR system produced four contrary-to-fact LRs (three false positives and one false negative). In each of those cases, the listeners produced either a consistent-with-fact LR or a much weaker contrary-to-fact LR compared with the ASR system. Fusion of the ASR scores with the median similarity-typicality scores from the listeners produced no improvement in EER (4.44%) but a marginal improvement in C_{llr}. Comparing the fused results with those of

the ASR in isolation shows that the magnitude of contrary-to-fact LRs was reduced when combined with the listener scores (hence the marginal reduction in C_{llr}). Figure 2 displays calibrated log LRs produced by the listeners (x-axis) and ASR system (y-axis). No correlation was found between the two sets of LRs for either SS or DS comparisons.

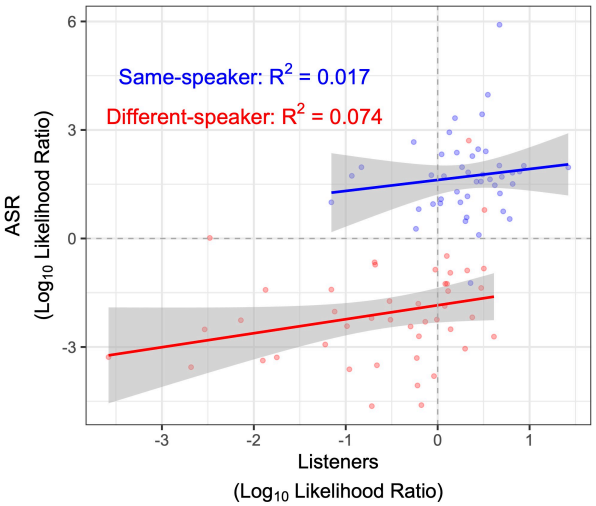

Figure 2: *Scatter plot of calibrated log_{10} SS (blue) and DS (red) LRs produced by the listeners and ASR system with linear regression line fitted.*

4. Discussion

The overall results presented here show relatively poor speaker recognition performance on the part of human listeners. However, this needs to be interpreted in light of the considerable challenges of the materials, i.e. short, forensically realistic channel-mismatched samples from a very homogeneous set of very similar sounding and demographically well-matched speakers. Reassuringly, the different methods of computing listener scores provided fairly consistent results. Given this, in line with [9], preference should be given to scores that explicitly include a judgement about typicality; in this case, the best performance was achieved by taking the median similarity and typicality judgements across listeners to produce an LR-like score. The similarity-only scores did produce marginally better performance. However, this is likely due to the use of the complement to the similarity judgement as the denominator for the score. This artificially maximises the typicality judgement which may account for stronger overall LRs and better performance. However, it is logically inconsistent with a specific alternative hypothesis (such as *the unknown voice belongs to another speaker of the same accent*) as is required in the forensic context.

As has been found in previous work, there was considerable variability across listeners in terms of their speaker recognition performance. Some listeners produced very impressive performance (13.3%), while others produced very poor performance (66.7%). In our data, however, performance was not found to be linked to a listener's familiarity with the accent (the listeners with the best and worst performance both reported their familiarity as being over 90 on the 100-point scale). Indeed, an additional set of tests using only data from participants with self-reported familiarity of over 80/100 produced worse overall performance than when including data from all listeners. This may be due to the use of a *standard* (and generally non-regional) accent in this study, with which all listeners have relatively high familiarity. The advantage of geographical proximity of listeners to the accent of speakers (see e.g. [14]) may only apply for marked regional varieties. This is something we plan to explore in future work.

The ASR system considerably outperformed the human listeners, although the conditions of the materials still made the task relatively difficult. The ASR system produced four contrary-to-fact LRs. In those cases, the listeners produced an error of a lower magnitude or were able to resolve the error. Fusion of ASR and listener scores led to only minimal improvement in overall performance (specifically C_{llr}) over the ASR in isolation. However, it is important to bear in mind that overall performance is only one relevant criterion to consider for forensics. In FVC, there is only a single pair of voices to analyse. Therefore, it is essential to know the magnitude of the evidence a system produces for those specific voices and the chance that a system would produce a contrary-to-fact result for that specific comparison. With this in mind, it is interesting that despite the poor overall performance of listeners, fusion reduces the magnitude of the small number of contrary-to-fact LRs that the ASR system produces. This has positive implications for potential interactions between listener judgements and ASR evidence presented in the courtroom, although more research is required in this area. Given the variability across listeners, it is likely that fusing the results of ASR with the better performing individual listeners would lead to greater overall improvements in performance. However, with the relatively small number of comparisons conducted per listener it was not possible to test this empirically. Finally, we found no correlation in the calibrated LRs produced by the listeners and the ASR system. Taken together with the analysis of the small number of ASR errors, the results suggest that listeners and ASR systems are sensitive to different information within the speech signal when making speaker recognition judgements. This finding is in line with the results of [3], although that study found greater improvements when fusing listener and ASR scores together.

5. Conclusions

The present study has explored ways in which LR-like speaker recognition scores can be elicited from human listeners and then compared and combined with the output of an ASR system. The marginal improvement in ASR C_{llr} after fusion potentially highlights the value of listener judgements in reducing the magnitude of contrary-to-fact LRs for individual pairs of speakers. Future work will expand this study to examine the conditions under which human listeners and ASR systems perform better or worse. We will also more directly assess the impact of listener judgements in the context of decisions made by jurors in the courtroom and explore the effects of sources of cognitive bias.

6. Acknowledgements

This research was conducted as part of the *Humans and Machines: Novel Methods for Assessing Speaker Recognition Performance* project funded by the UK Arts and Humanities Research Council (AH/T012978/1).

7. References

[1] P. Foulkes and A. Barron, "Telephone speaker recognition amongst members of a close social network," *Forensic Linguistics*, vol. 7, no. 2, pp. 180–198, 2000.

[2] A. Afshan, J, Kreiman and A. Alwan, "Speaker discrimination performance for 'easy' versus 'hard' voices in style-matched and -mismatched speech," *JASA*, vol. 151, no. 2, pp. 1393–1403, 2022.

[3] A. Afshan, J. Kreiman and A. Alwan, "Speaker discrimination in humans and machines: effects of speaking style variability," in *Proceedings of INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association,* Shanghai, China, Oct. 2020, pp. 3136–3140.

[4] V. Hughes, P. Harrison, P. Foulkes, J. P. French, C. Kavanagh and E. San Segundo, "Mapping across feature spaces in forensic voice comparison: the contribution of auditory-based voice quality to (semi-)automatic system testing," in *Proceedings of INTERSPEECH 2017 – 18th Annual Conference of the International Speech Communication Association,* Stockholm, Sweden, Aug. 2017, pp. 3892–3896.

[5] V. Hughes, P. Harrison, P. Foulkes, J. P. French and A. Gully, "Forensic voice comparison using long-term acoustic measures of voice quality," in *Proceedings of the 19th International Congress of Phonetic Sciences*, Melbourne, Australia, August 2019, pp. 1455–1459.

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

[7] J. Lindh and G. S. Morrison, "Humans versus machine: forensic voice comparison on a small database of Swedish voice recordings," in *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, August 2011, pp. 1254–1257.

[8] C. G. G. Aitken, F. Taroni and S. Bozza, *Statistics and the Evaluation of Evidence for Forensic Scientists*. Wiley, 2021.

[9] G. S. Morrison and E. Enzinger, "Score based procedures for the calculation of forensic likelihood ratios – scores should take account of both similarity and typicality," *Science and Justice*, vol. 58, no. 1, pp. 47–58, 2018.

[10] F. Nolan, K. McDougall, G. de Jong and T. Hudson, "The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research," *International Journal of Speech Language and the Law*, vol. 16, no. 1, pp. 31–57, 2009.

[11] N. Brümmer and J. du Preez, "Application-independent evaluation of speaker detection," *Computer Speech and Language*, vol. 20, pp. 230–275, 2006.

[12] G. S. Morrison, "Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio," *Australian Journal of Forensic Science*, vol. 45, pp. 173–197, 2013.

[13] C. Champod and I. W. Evett, "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,* vol. 7, no. 2, pp. 239-243, 2000.

[14] N. Atkinson, *Variable Factors Affecting Voice Identification in Forensic Contexts.* PhD Thesis, University of York, UK.