The Speaker and Language Recognition Workshop (Odyssey 2024)
18-21 June 2024, Quebec City, Canada

EXPLORING INDIVIDUAL SPEAKER BEHAVIOUR WITHIN A FORENSIC AUTOMATIC SPEAKER RECOGNITION SYSTEM

Vincent Hughes[1], Chenzi Xu[1], Paul Foulkes[1], Philip Harrison[1], Poppy Welch[1], Jessica Wormald[1], Finnian Kelly[2] and David van der Vloed[3]

[1] Department of Language and Linguistic Science
University of York, United Kingdom
{firstname.lastname}@york.ac.uk

[2] Oxford Wave Research
Oxford, United Kingdom
finnian@oxfordwaveresearch.com

[3] Netherlands Forensic Institute
The Hague, Netherlands
d.van.der.vloed@nfi.nl

ABSTRACT

A key issue for automatic speaker recognition (ASR), particularly for forensics, is our lack of understanding about why certain voices prove more or less of a challenge for systems. In this paper, we focus on variability in individual speaker performance within an x-vector ASR system and examine this variability as a function of the phonetic content within speech samples. The inclusion of vowels generally improved performance, but not for all speakers. Indeed, some speakers produced broadly the same C_{llr} irrespective of the phonetic content in the speech samples. Poor ASR performance was not well correlated with long-term laryngeal features (f0 and laryngeal voice quality) and these features may provide additional speaker discriminatory information for some speakers. We discuss the implications of these findings in terms of developing a speaker quality metric for flagging potentially problematic speakers prior to ASR comparison.

1. INTRODUCTION

1.1. Forensic automatic speaker recognition

Automatic speaker recognition (ASR) systems are increasingly used in forensic voice comparison (FVC) [1,2] casework around the world. ASR systems have many benefits over traditional linguistic and phonetic analysis conducted by a forensic speech scientist. Firstly, ASR is relatively quick and efficient. This reduces the amount of time needed to conduct FVC casework (where linguistic and phonetic analysis would take 10-15 hours per case for a typical 1:1 comparison). The speed of ASR also facilitates large-scale validation of systems. Secondly, ASR systems are considered, to some extent, to be more objective than linguistic and phonetic analysis. This is true in the sense that, given the same system with the same settings and input data, the output will be exactly reproducible. That is not to say, however, that ASR does not involve any subjectivity (see [3]). Further, calibrated ASR systems provide a statistically-grounded, numerical LR as output, where typicality is estimated empirically rather than based subjectively on the knowledge and experience of the analyst (as it is in linguistic and phonetic analysis).

Despite these advantages, the application of ASR to forensic materials remains challenging. In part, this is because of practical issues with the availability of large enough sets of recordings that are representative of the conditions of casework in order to train calibration models and conduct meaningful case-specific validation [4]. There are also broader issues around interpretability and explainability [5]. Development of ASR systems has typically focused on advances in modelling and algorithms in order to handle different types of technical challenges that may be encountered in real case materials. Such development has seen considerable improvement in overall ASR performance with new generations of systems [6], especially for state-of-the-art approaches that utilise deep learning to generate an embedding-based representation of the voice in a recording. The improvements in performance necessarily reflect the fact that state-of-the-art systems capture more speaker-specific information in their speaker representations than ever before. However, relatively little is yet known about why certain voices would or would not perform well within a system. Principally, this is because we do not know precisely which linguistic and phonetic properties are captured by systems - and crucially, which are missed - when generating speaker embeddings.

Such information is often overlooked because testing of ASR systems, even in forensic contexts [6], tends to focus on overall performance metrics, which average over the results of a large number of comparisons from a large number of speakers. In this way, progress within the field is measured by improvements in the overall performance metric, with less concern given to the specific causes of contrary-to-fact results (i.e. which speakers or samples are responsible for the *errors*).

1.2. Importance of the speaker factors for forensics

Relatively little work has considered the effect of speaker factors on ASR performance. There are two principal reasons why more research is needed in this area, particularly in the context of forensic evaluation. Firstly, it may be possible to

10.21437/odyssey.2024-1

leverage information about the behaviour of individuals or sub-sets of speakers within systems in order to improve overall performance; for example, if analysis reveals that speakers who produce contrary-to-fact results all share a given phonetic property, it may be that efforts should go into ensuring that systems better capture this feature (see [7], also discussed in 1.3 below). However, this relies on being able to identify particularly problematic voices or speech samples, and to utilise that information to expand data sets for training embedding extraction or for use as adaptation or calibration sets. Secondly, in the forensic domain, it is important that the analysis and conclusion provided by an analyst are transparent. This is true from the perspective of understanding whether the output of the system is consistent with what is known about the voice or the speech sample, and ensuring that different methods capture independent information (i.e. not double counting evidence). It is, therefore, essential that the analyst understands how decisions made about system architecture and choice of datasets (e.g. for calibration) affect performance for a given voice or speech sample. Transparency is also crucial from the perspective of having to explain, at some level of abstraction, how the system works and behaves to an end-user, such as a judge and/or jury [8].

1.3. Previous work

A small body of work has considered these issues. [7] examined the phonetic properties of a set of false-acceptances produced by an i-vector ASR system. Post-hoc analysis was conducted by phoneticians, who were generally able to separate the different-speaker pairs on the basis of laryngeal voice quality (amongst other features). This is consistent with other studies which suggest that MFCC-based ASR systems tend not to capture properties of laryngeal voice quality [9,10]. [11] expands this work, examining a novel 'binary-attribute-based' LR architecture [12] which is able to evaluate the contribution of different acoustic-phonetic variables (e.g. f0, vowel formants, voice quality) to speaker separation. An alternative approach is proposed in [13] which uses phonetically controlled input to explicitly test the effects of voice quality match and mismatch on system output. Very good performance was found with voice quality-matched samples, but when using voice quality-mismatched samples, a large increase in false rejections was found. This was most marked in comparisons involving supralaryngeal vocal settings which affected vocal tract length (e.g. lowered larynx and backed tongue body) as well as whisper (as in [14]). Work by the speaker recognition group at the University of Avignon has focused specifically on by-speaker variability in ASR performance. [15] presents by-speaker log likelihood ratio cost (C_{llr}) values from a test set of 30 speakers. Performance ranges from less than 0.02 to over 0.8, despite the overall system-level C_{llr} being around 0.17. The causes of the variability are different for different speakers; some of the variability is driven by same-speaker pairs (which is neatly explained in [16] as a function of the acoustic homogeneity between pairs), and some by different-speaker pairs; some by discrimination error, and some by calibration error. [15] also examines individual speaker performance as a function of the segmental phonetic make-up of samples. The removal of oral vowels generally had the biggest negative effect on individual speaker performance (i.e. performance with both same- and different-speaker pairs was better by as much as 160% with the addition of oral vowels compared with when they were removed), although for some speakers it actually led to improved performance. Similarly, the removal of fricatives improved performance for seven of the 30 speakers, but made performance worse for the remaining 23. For some speakers, it appears that performance is relatively consistent irrespective of the phonetic content of the speech samples.

1.4. This study

The present study builds on the work described in 1.3 to address key issues around the performance of ASR systems in terms of individual speakers, utilising linguistic, phonetic, and forensic knowledge to examine ASR behaviour. We do this by conducting an initial validation exercise of the kind described in [4] to assess the overall performance of an x-vector ASR system using a forensically realistic dataset. We then interrogate the results in two ways. Firstly, we assess the extent of variability in individual performance, both in terms of discrimination and calibration error. Secondly, we attempt to explain the variability across speakers as a function of the phonetic content within the speech samples, based on broad phone-level categories. This involves replicating the methodology in [15], but using a more recent x-vector architecture. We also conduct a detailed examination of the results in order to assess *why* certain speakers are more or less affected by the removal of certain phone categories. To do this, we principally focus on those speakers who perform poorest overall. The long-term aim is to use knowledge of the key phonetic variables which affect performance to develop a metric for predicting, *a priori*, which speakers are likely to be problematic within an ASR-based FVC case.

2. METHODS

2.1. Data

For this study, we utilise GBR-ENG; a dataset of forensically-realistic recordings provided by the UK Government. The dataset is ideal for forensic evaluation as it contains a very large number of both male and female speakers (1,946 total; 906 male, 1,040 female) of British English, with considerable variability in regional and social backgrounds, as well as age. Each speaker has multiple speech samples (average of 10 samples per speaker; 12,483 files in total) many of which were recorded across multiple days. Speech samples contain spontaneous conversations lasting between 181 and 373 seconds. There is also a mix of mobile and landline telephone recordings available.

2.2. Test and calibration sets

Analysis was conducted using sets of male and female speakers separately. We identified subsets of 48 male (160 files total) and 46 female (154 files total) speakers to act as test sets. Speakers were chosen based on the availability of at least three non-contemporaneous samples (defined as samples made on different days). Speakers ranged from having three to seven samples each. We also limited the analysis to good quality (determined based on over 30s of net speech, WADA SNR of over 24dB, and less than 1% clipping), mobile phone samples, in order to reduce the number of confounding variables across speakers. The calibration sets contained 50 male and 50 female speakers selected at random, with two samples recorded on

separate days that were of good quality and both made via mobile phone.

2.3. Orthographic transcription and forced alignment

In order to extract phone-level information from the speech samples, orthographic transcripts were created using the large V2 model in Whisper-timestamped [17,18], an extension to Whisper [19], with word-level timestamps. The transcripts and audio files were then processed via the Montreal Forced Aligner [20] to produce time-aligned phone boundaries.

Phones were categorised into five broad categories: (a) vowels, (b) nasals, (c) approximants, (d) fricatives, and (e) plosives. The classification is described in the MFA phone sets for UK English [21] below (affricates were excluded as there are far fewer tokens than for other categories):

(a) vowels: [i, ɪ, e, ɛ, æ, a, ɑ, ʊ, ʉ, ɐ, ɒ, ə, ɜ, ɒː, iː, ɑː, ʉː, ɜː, ɛː, əw, aw, aj, ej, ɔj]
(b) nasals: [m̩, m, mʲ, ŋ, n, n̩, ŋ, ɲ]
(c) approximants: [ɬ, l, ʎ, ɹ, j, w, ɬ̞, ʎ̞, ɻ]
(d) fricatives: [f, v, fʲ, vʲ, fʷ, vʷ, θ, ð, s, z, ʃ, ʒ, ç, h]
(e) plosives: [p, b, t̪, d̪, t, d, c, ɟ, k, g, ʔ]

The choice of these categories is based on the fact that the phones within them have similar phonetic properties. This in turn means that there is more material to assess category-level patterns, rather than focusing on individual segments, which may in total have extremely short durations. Using broad phone categories also mitigates the effect of potential errors produced during the automatic transcription and alignment stages.

2.4. Extraction of phonetic data

2.4.1. Phone-level analysis

To test the effects of the phonetic content within the speech samples, we replicated the approach taken in [15]. This involved creating versions of the original speech samples where the target category (see 2.3) had been removed (referred to in [15] as the **specific** condition). To account for the fact that there is an unbalanced amount of material for each phone category within a sample (i.e. generally more vowels than other segments), we also created **random** samples where we removed the same proportion of speech at random from the original files. The duration of these removed speech intervals was randomly drawn from a distribution where the mean and standard deviation are the same as the segments in the corresponding **specific** condition. For each phonetic class, five **random** samples were produced to account for any skew that might occur in randomly removing speech content that could affect the ASR scores.

2.4.2. Long-term phonetic analyses

From each of the original speech samples, acoustic-phonetic data were also extracted which capture long-term, laryngeal properties of the voices. This provides an additional source of phonetic information to examine individual speaker performance, which may be complementary to that captured by the phone-level analysis described in 2.4.1. Specifically, this involved extracting fundamental frequency (f0) from 25ms windows shifted by 10ms across the voiced portions of the recordings using VoiceSauce [22]. The pitch range was set to 40-300Hz for the male speakers and 75-350Hz for the female speakers. The mean and standard deviation of the f0 values were then used as summary statistics for each sample. Laryngeal voice quality acoustics were also extracted from 25ms windows shifted by 10ms across the entire file; namely the amplitude of harmonics (H1*, H2*, H4*, A1*, A2*, A3*, H2k*), spectral tilt measures (H1*-A2*, H1*-A3*, H4*-H2k*, and H2k-H5k*), energy-related measures (RMS energy), harmonics-to-noise ratios within certain frequency bands (HNR05, HNR15. HNR25, HNR35), formant frequencies and bandwidths, subharmonic-to-harmonic ratio, and cepstral peak prominence. These measures, together, broadly capture differences between modal, breathy and creaky voice qualities. Feature extraction was conducted in VoiceSauce [22] with the default settings. Given the multidimensionality of the measures, linear discriminant analysis (LDA) was used to produce a two-dimensional voice quality vector for each sample. The LDA model was trained with the speaker labels in order to maximise the ratio of between-speaker distance to within-speaker distance. This was done using the scikit-learn package in Python.

2.5. Automatic speaker recognition system

ASR testing was conducted using VOCALISE 2021 (version 3.0.0.1746, [23]), which has been widely used across the world for FVC casework [2]. VOCALISE is an x-vector-based system [24] utilising MFCCs to produce 512-dimensional speaker embeddings, which are then subjected to LDA in order to reduce dimensionality. Scoring was conducted using a pre-trained PLDA model, and no additional condition adaptation was applied. Tests were initially conducted using the whole file for each speaker. Separate tests were then conducted for the **specific** and **random** conditions for each of the five phone categories. For each set of tests we ran, same-speaker (SS) and different-speaker (DS) scores were computed for the test and calibration sets. The calibration scores were used to train a logistic regression model [25]. The coefficients from the calibration model were then applied to the test scores to produce calibrated log likelihood ratios (LLRs; males: 200 SS, 12,520 DS; females: 190 SS, 11,591 DS). These LLRs were used as the basis of the evaluation of individual speakers' results.

2.6. Evaluation

Performance was evaluated using C_{llr} [26], made up of its two constituents, C_{llr}^{min} (discrimination loss) and C_{llr}^{cal} (calibration loss; although see [27]), and equal error rate (EER). Overall system performance based on the full recordings for both male and female test sets was first evaluated, followed by by-speaker performance.

The phone-level analysis was evaluated using the ratio of the average C_{llr} for the **random** condition (across the five randomisations) and the C_{llr} for the **specific** condition expressed as a percentage. This metric is defined in [15] as C_{llr}^{R} and was chosen as a means of comparing effects across segment categories whilst not skewing results due to the amount of phonetic material included or excluded. A positive C_{llr}^{R} indicates that the **specific** condition, where a given phone category is excluded, provides better performance (i.e. lower C_{llr}) compared with the **random** condition. In such cases, the inclusion of a given phone category makes performance worse. Conversely, a negative C_{llr}^{R} means that the **random** condition, with all available segmental categories, produces better performance than the **specific** condition. In such cases, the

inclusion of a given phone category makes performance better. The C_{llr}^{R} value was calculated both overall and for each speaker individually.

3. Results

Table 1 displays overall system performance for the male and female sets based on the full samples. Performance is generally very good. For both the male and female sets, C_{llr}^{cal} is relatively low, indicating that the system produces well calibrated LLRs.

Table 1. *Overall system performance based on the male and female sets*

Set	EER (%)	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}
Male	3.03	0.13	0.10	0.03
Female	2.61	0.11	0.08	0.03

We found additional improvements in overall performance by applying S-norm [28] reference normalisation in VOCALISE (using normalisation sets of 100 files per sex) which reduces EER to 1.49% for the male speakers and to 1.53% for the female speakers. For further analysis, however, we use the baseline output of VOCALISE without normalisation as this removes an additional source of variability in individual speaker results.

3.1. By-speaker performance

Figure 1 displays by-speaker C_{llr}, C_{llr}^{min}, and C_{llr}^{cal} values for the male (above) and female (below) speaker sets. As is expected from the very good overall performance reported in Table 1, the system performs extremely well for most speakers, with C_{llr}s of less than 0.2 and in the majority of cases, no discrimination error. However, for both sets around 15% of speakers produce C_{llr}s above 0.2, with non-linear increases in C_{llr} after this point. These speakers may, therefore, be considered more problematic.

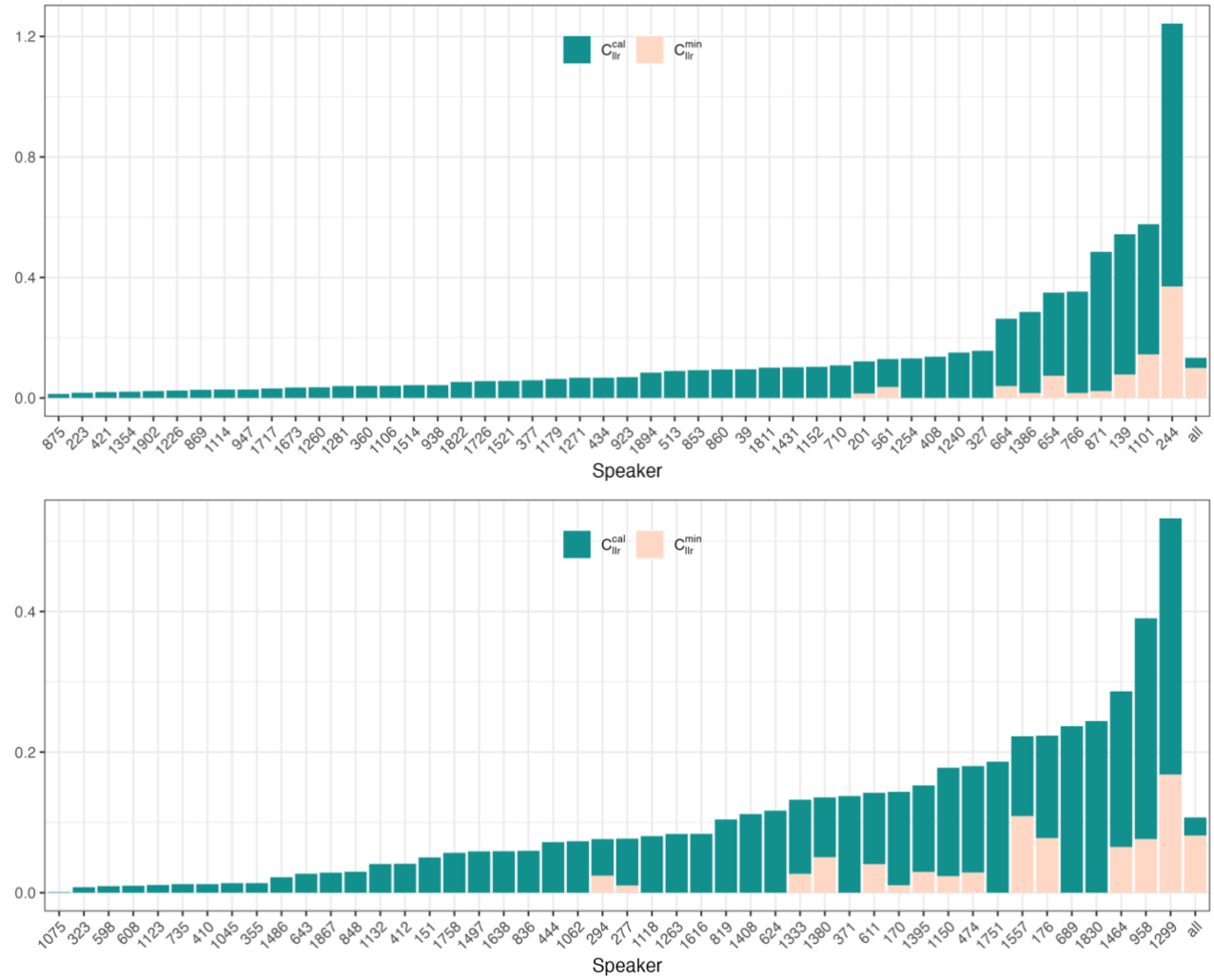

Figure 1. *By-speaker and overall C_{llr}^{cal} (green) and C_{llr}^{min} (pink) values (C_{llr} is the overall height of the bar) for males (above) and females (below) (*note differences in the scales on the y-axes).*

For a very small number of those speakers (around 3 or 4 per sex), performance is actually relatively poor, with C_{llr}s of around 0.5 or, in some cases, even higher. There appears to be one outlier speaker (#244 in the male set) whose C_{llr} is over 1.2; we return to this speaker in 3.3.

3.2. Phone-level analysis

Figure 2 displays a stacked barchart of C_{llr}^{R} values for each speaker (male, above; female, below) and each phone category. As a reminder, a negative value for C_{llr}^{R} indicates that the inclusion of a given phone category improves performance (i.e. leads to an decrease in C_{llr}). Across both datasets, and for almost all speakers, vowels contribute the most towards speaker discrimination, producing the largest negative C_{llr}^{R} of any phone category. This is perhaps unsurprising given that vowels are known to carry considerable speaker-specific information. Further, vowels make up the highest proportion of speech; the speech-active portions of samples consisted of around 39% vowel material, compared with between 8-10% for the other phone categories. This natural skew towards vowels also likely means that vowels are weighted more strongly in terms of the representation learned by the DNN in producing speaker embeddings during the initial training stage. For males, the other phone categories offered little towards improving overall performance. For females, however, nasals and approximants made some small contribution, with some improvements in C_{llr} found for most speakers with the inclusion of these categories. The inclusion of stops and fricatives makes performance worse for most speakers. There is also considerable variability across speakers in terms of the contributions of different phone categories. Some speakers display considerably higher than average contributions from the vowel material, while in some cases vowels contribute nothing to speaker discrimination. Indeed, in some extreme cases, the inclusion of vowel material actually makes a speaker's C_{llr} worse (note four male speakers towards the right of Figure 1; #139, #244, #766, #871).

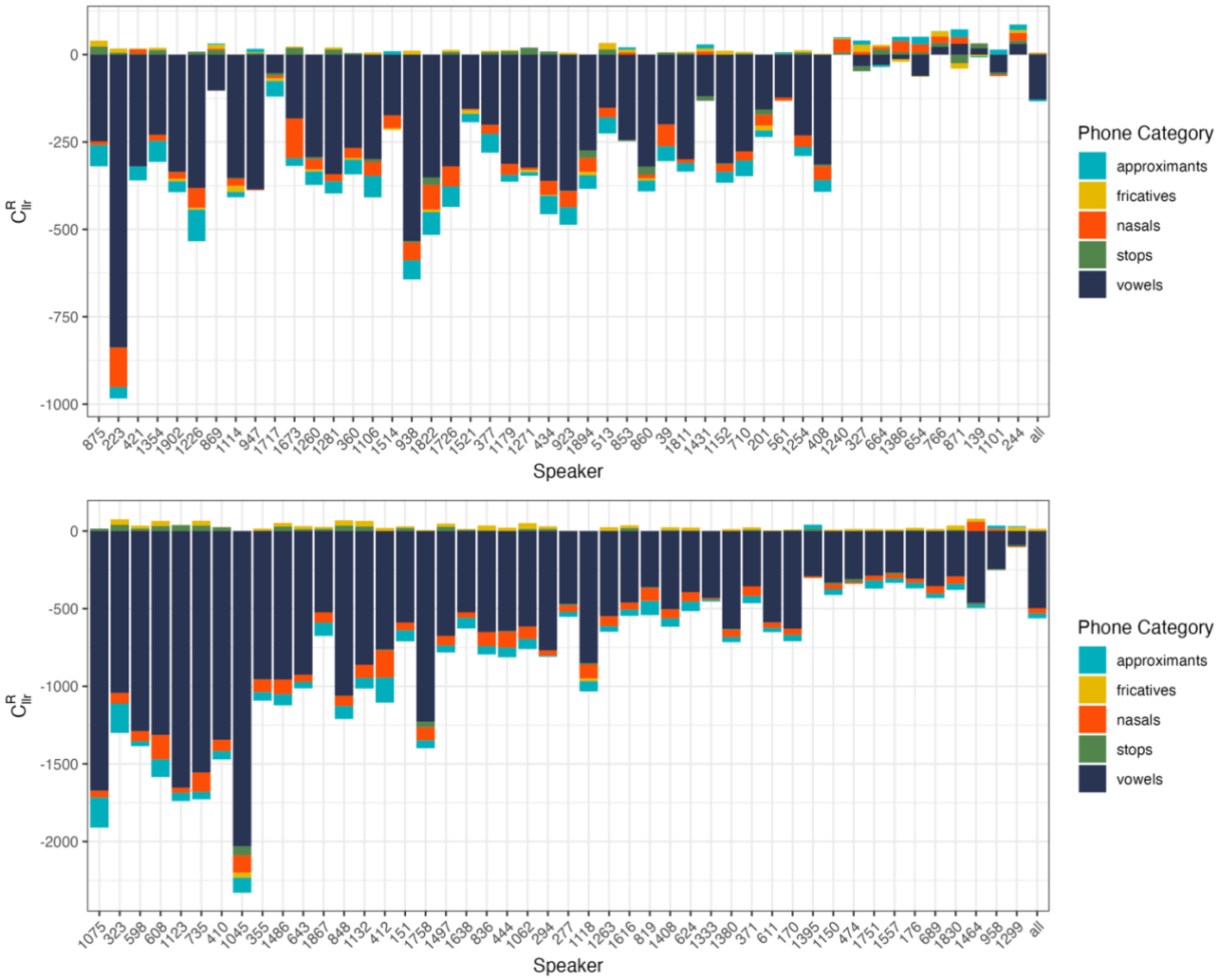

Figure 2. *Stacked barchart of by-speaker (speaker numbers on the x-axis) and overall C_{llr}^{R} values (negative C_{llr}^{R} indicates that a given category contributes to improving C_{llr} while positive C_{llr}^{R} values indicate that performance gets worse with the inclusion of a given category) for each of the five phone categories for the male (above) and female (below) sets (*note differences in the scales on the y-axes); speakers are arranged in the same order as Figure 1.*

Comparison of Figures 1 and 2 suggests a correlation between a speaker's C_{llr}^{R} and their overall C_{llr}. That is to say, those speakers who show limited improvement in performance irrespective of the phonetic content of the samples are also the speakers that generally perform worse overall. This is also reflected in relatively strong correlations between the sum of the C_{llr}^{R} and mean SS LLRs (R = -0.63) and DS LLRs (R = 0.54) for each speaker; i.e. those speakers with an overall C_{llr}^{R} closer to zero (or even positive) are more likely to produce LLRs closer to zero.

3.3. Exploring *problematic* samples and speakers

In this section, we consider *why* certain speakers perform so poorly. We initially focus on male speaker #244 given the extremely poor performance relative to any other speaker. Analysis of the individual comparisons for this speaker reveals contrary-to-fact LLRs primarily for SS comparisons involving one sample (#04). Based on listening, this sample is phonetically very different from the other two samples. More samples were available for this speaker than we used in this paper (but were lower in quality based on our heuristics). Contextual information indicates that this is not a speaker mislabelling. However, sample #04 involves a conversation with the speaker's mother, whereas in the other samples he is speaking with a peer. The phonetic differences could, therefore, be attributed to style shifting. Determining whether this could account for ASR errors, however, would require more detailed analysis of the style variability across a larger number of speakers.

On the basis of the results in 3.1 and 3.2, it appears that for the worst performing speakers, the phonetic information captured within the different phone categories is not useful for the ASR in terms of discrimination (i.e. performance remains poor irrespective of the phonetic content within a sample). This means that in order to assess *why* these speakers are performing poorly, we need to look beyond segmental information. For this purpose, we looked at long-term laryngeal features of the voice, namely voice quality (note, this does not include information about supralaryngeal vocal setting) and f0. Figure 3 displays a two-dimensional representation of the voice quality space based on LDA (described in 2.4.2), with data points coloured according to each speaker's C_{llr}. There appears to be no clear correlation between a speaker's position within the voice quality space and C_{llr}, with good performing speakers in the ASR situated both within busy clusters of speakers and at the peripheries of the voice quality space – this was also the case for f0. For some of the poorer performing speakers within the ASR system (e.g. male speakers #139, #1101; female speakers #176, #1150), good separation is revealed based on voice quality. This suggests that for some speakers (but not necessarily all), analysis of voice quality may be beneficial in addition to the baseline ASR analysis. Interestingly, the three samples for male speaker #244 are fairly close to each other within the voice quality space (the orange data points in the top plot of Figure 3). This indicates that the phonetic differences between the samples for this speaker are not related to voice quality (which is consistent with our evaluation based on listening).

4. Discussion

Examination of individual speaker behaviour has provided a rich source of information for better understanding and contextualising the overall performance of the ASR system tested in this study. Despite good overall performance, and additional improvements through reference normalisation, there remains considerable performance variability across speakers. Many speakers perform extremely well, producing C_{llr}^{min} values of 0 (i.e. no discrimination error). For the best performing speakers, there is still often some, albeit small, calibration loss which may be resolved through more careful selection of calibration data (as would be done in a forensic case; see e.g. [29]). While it is difficult to identify a hard threshold for defining whether a speaker is *problematic* (i.e. the poorer performers), we see non-linear increases in by-speaker C_{llr} by around 0.2. This accounts for around 15% of our speakers, for both the male and female datasets. In [15], similar non-linear patterns of by-speaker C_{llr} are also observed from around 0.2 (despite using an i-vector system, rather than x-vectors as in the present study), accounting for 17% of the speakers in their dataset.

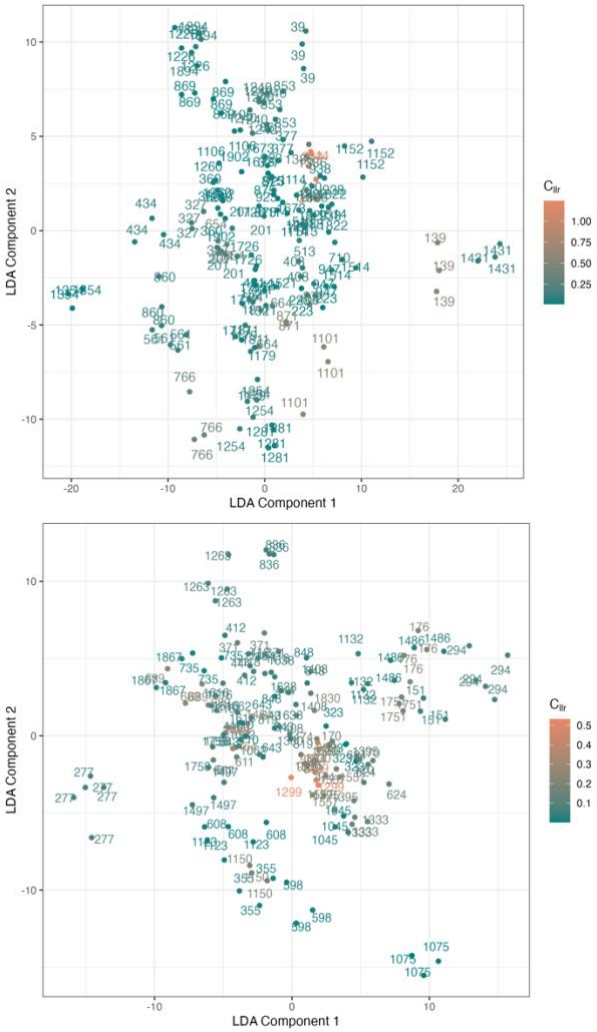

Figure 3. *Each sample (dots) and speaker (numbers) for the male (above) and female (below) sets within a two-dimensional representation of the voice quality space (based on LDA projections) coloured according to the C_{llr} calculated based on the output of the ASR system with the original files (note different scales for C_{llr} for the male and female sets).*

A crucial question, especially in the context of a single comparison in an FVC case, is whether it is possible to predict, based on linguistic and phonetic properties, whether a speaker is likely to be in the 15% of *problematic* speakers. The analysis of ASR performance as a function of phonetic content provides more fine-grained insights into speaker behaviour. For the most part, the inclusion of vowel material unsurprisingly leads to considerable improvements in performance, although there are exceptions where vowels actually make performance slightly worse. For most speakers, the inclusion of fricatives and stops makes performance worse. It may be that exclusion of these phone categories at the feature extraction stage leads to better overall system performance. In terms of trying to predict problematic speakers, our results suggest that speaker performance is related to the extent of the sensitivity to the phonetic content of a sample. That is, speakers who show big improvements in performance when certain phone categories, particularly vowels, are included in the speech samples, generally display the best overall performance. Those speakers who remain consistent in their performance irrespective of the phonetic make-up of the speech samples, are also generally those speakers who perform worse overall. It may be that this can be utilised to develop a speaker quality metric in the future.

Having identified the *problematic* speakers based on the output of the system, diagnosing the specific causes of poor performance is challenging because the variability may have multiple sources. There may, for example, be sources of technical variability that aren't revealed through overall heuristics such as net speech, SNR, and clipping. Long-term laryngeal measures of f0 and voice quality are not well correlated with the output of the ASR system and so are not good predictors of poor performance. On the contrary, as found in [7,9], laryngeal measures appear to capture complementary speaker-specific information which helps to separate some otherwise *problematic* speakers. However, methods for identifying which speakers would benefit from additional analysis of laryngeal features, especially for forensic casework, remains a challenge.

5. Conclusion

This study has provided a detailed analysis of individual speaker behaviour and attempted to understand that behaviour through the phonetic properties of each speaker. Results reveal interesting patterns related to the phonetic make-up of samples which could be used to screen *problematic* speakers prior to comparison. Future work will continue to examine the phonetic causes of poor performance within ASR systems.

6. Acknowledgements

This work was funded by an ESRC Research Grant (ES/W001241/). The Viking cluster was used during this project which is a high performance compute facility provided by the University of York. We are grateful for computational support from the University of York, IT services and the Research IT team. Thanks to Lauren Harrington for providing an independent phonetic view on the samples for speaker #244.

7. References

1. Toby Hudson, Kirsty McDougall and Vincent Hughes, "Forensic phonetics," in R.A. Knight and J. Setter, The Cambridge Handbook of Phonetics, Cambridge University Press, pp. 631–656, 2021.
2. David van der Vloed and Tina Cambier-Langeveld, "How we use automatic speaker comparison in forensic practice," International Journal of Speech, Language and the Law, vol. 29 no. 2, pp. 201–224, 2023.
3. Paul Foulkes, "Dead clade walking? On the survival prospects of the forensic phonetician," Keynote at the 18th Australasian International Conference on Speech Science and Technology, ANU, Australia, December 2022.
4. Geoffrey S. Morrison, Ewald Enzinger, Vincent Hughes, Michael Jessen, Didier Meuwly, Cedric Neumann, S. Planting, William C. Thompson, David van der Vloed, Rolf J.F. Ypma and Cuiling Zhang, "Consensus on validation of forensic voice comparison," Science and Justice, vol. 61 no. 3, pp. 299-309, 2021.
5. Yada Pruksachatkun, Matthew McAteer and Subhabrata Majumdar, Practicing trustworthy machine learning: consistent, transparent, and fair AI pipelines, O'Reilly, 2023.
6. Geoffrey S. Morrison and Ewald Enzinger, "Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – conclusion," Speech Communication, vol. 112, pp. 37-39, 2019.
7. Joaquin Gonzalez-Rodriguez, Juana Gil, Rubén Pérez and Javier Franco-Pedroso, "What are we missing with i-vectors? A perceptual analysis of i-vector-based falsely accepted trials," in Proceedings of Odyssey: The Speaker and Language Recognition Workshop, Joensuu, Finland, pp. 33–40.
8. UK Government Justice and Home Affairs Committee, Technology rules? The advent of new technologies in the justice system, 1st Report of Session 2021-22, March 2022. https://publications.parliament.uk/pa/ld5802/ldselect/ldjusthom/180/18002.htm
9. Vincent Hughes, Amanda Cardoso, Paul Foulkes, Peter French, Philip Harrison and Amelia Gully, "Forensic voice comparison using long-term acoustic measures of voice quality," in Proceedings of the International Congress of Phonetic Sciences, Melbourne, Australia, August 2019, pp. 1455-1459.
10. Vincent Hughes, Amanda Cardoso, Paul Foulkes, Peter French, Amelia Gully and Philip Harrison, "Speaker-specificity in speech production: the contribution of source and filter," Journal of Phonetics, vol. 97, 101224, 2023.
11. Imen Ben-Amor, Jean-François Bonastre, Benjamin O'Brien and Pierre-Michel Bousquet, "Describing the phonetics in the underlying speech attributes for deep and interpretable speaker recognition," in Proceedings of Interspeech, Dublin, Ireland, August 2023, pp. 3207–3211.
12. Imen Ben-Amor and Jean-François Bonastre, "Ba-lr: binary-attribute-based likelihood ratio estimation for forensic voice comparison," in Proceedings of the International Workshop on Biometrics and Forensics, Salzburg, Austria, April 2022, pp. 1–6.
13. Vincent Hughes, Jessica Wormald, Paul Foulkes, Philip Harrison, Finnian Kelly, David van der Vloed, Poppy Welch and Chenzi Xu, "Automatic speaker recognition with variation across vocal conditions: a controlled experiment with implications for forensics," in Proceedings of Interspeech, Dublin, Ireland, August 2023, pp. 591–595.

14. Finnian Kelly and John H.L. Hansen, “Analysis and calibration of Lombard Effect and whisper for speaker recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 927–942, 2021.
15. Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato and Juliette Kahn, “Phonetic content impact on forensic voice comparison,” in Proceedings of IEEE Workshop on Spoken Language Technology, San Juan, Puerto Rico, December 2016, pp. 210–217.
16. Moez Ajili, Jean-François Bonastre, Waad Ben Kheder, Solange Rossato and Juliette Kahn, “Homogeneity measure impact on target and non-target trials in forensic voice comparison,” in Proceedings of Interspeech, Stockholm, Sweden, August 2017, pp. 2844–2848.
17. Jérôme Louradour, Whisper Timestamped, GitHub Repository, 2023. https://github.com/linto-ai/whisper-timestamped
18. Tino Giorgino, “Computing and visulising dynamic time warping alignments in R: the dtw package,” Journal of Statistical Software, vol. 31 no. 7, pp. 1–24, 2009.
19. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” arXiv preprint, 2212.04356, 2022.
20. Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner and Morgan Sonderegger, “Montreal Forced Aligner: trainable text-speech alignment using Kaldi,” in Proceedings of Interspeech, Stockholm, Sweden, August 2017, pp. 498–502.
21. Michael McAuliffe and Morgan Sonderegger, English (UK) MFA Dictionary v.2.2.1, 2023, https://mfa-models.readthedocs.io/en/latest/dictionary/English/
22. Yen-Liang Shue, Patricia Keating, Chad Vicenik and Kristine Yu, “Voicesauce: a program for voice analysis,” in Proceedings of the International Congress of Phonetic Sciences, Hong Kong, August 2011, pp. 1846-1849.
23. Finnian Kelly, Oscar Forth, Samuel Kent, Linda Gerlach and Anil Alexander, “Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors,” Audio Engineering Society Forensics Conference, Porto, Portugal, June 2019.
24. David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey and Sanjeev Khudanpur, “X-vectors: robust DNN embeddings for speaker recognition,” in Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, Canada, April 2018, pp. 5329–5333.
25. Geoffrey S. Morrison, “Tutorial on logistic-regression calibration and fusion: converting a score to a likelihood ratio,” Australian Journal of Forensic Sciences, vol. 45, pp. 173–197, 2013.
26. Niko Brümmer and Johan du Preez, “Application-independent evaluation of speaker detection,” Computer Speech and Language, vol. 20 no. 2-3, pp. 230–275, 2006.
27. Geoffrey S. Morrison, “In the context of forensic casework, are there meaningful metrics of the degree of calibration?” Forensic Science International: Synergy, vol. 3, 100157, 2021.
28. Stephen Shum, Najim Dehak, Reda Dehak and James R. Glass, “Unsupervised speaker adaptation based on the cosine similarity for text-independent speaker verification,” in Proceedings of Odyssey: The Speaker and Language Recognition Workshop, Brno, Czech Republic, paper 16.
29. Luciana Ferrer, Mahesh Kumar Nandwana, Mitchell McLaren, Diego Castan and Aaron Lawson, “Toward fail-safe speaker recognition: trial-based calibration with a reject option,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, pp. 140–153, 2019.