INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Automatic Speaker Recognition performance with matched and mismatched female bilingual speech data

Bryony Nuttall[1], *Philip Harrison*[2] *Vincent Hughes*[3]

[1]The Forensic Voice Centre, York, UK
[123]University of York, UK
bryony.nuttall@forensicvoicecentre.com
{philip.harrison|vincent.hughes}@york.ac.uk

Abstract

Validation of forensic voice comparison methods requires testing using speech samples that are representative of forensic casework conditions. Increasingly, around the world, forensic voice comparison casework is being undertaken using automatic speaker recognition (ASR) systems. However, multilingualism remains a key issue in applying automatic systems to forensic casework. This research aims to consider the effect of language on ASR performance, testing developers' claims of 'language independency'. Specifically, we examine the extent to which language mismatch either between the known and questioned samples, or between the evidential samples and the calibration data, affects overall system performance and the resulting strength of evidence (i.e., likelihood ratios for individual comparisons). Results indicate that mixed language trials produce more errors than single language trials which makes drawing evidential conclusions based on bilingual data challenging.

Index Terms: automatic speaker recognition, forensic speaker comparison, bilingualism, language mismatch

1. Introduction

In current forensic voice comparison practice, the acoustic-phonetic-linguistic method is most widely used [1], but increasingly practitioners are using automatic speaker recognition (ASR) technology, due to its speed, objectivity and replicability which are attractive to criminal justice procedures and military intelligence services across the world [2]. ASR systems utilise biometric technology to model speakers based on the biological characteristics of their voice, and to evaluate statistically both the degree of similarity between speakers and the relative typicality amongst a wider population [3]. In the context of forensic voice comparison, the output of an automatic system is a likelihood ratio (LR) which captures the relative strength of the voice evidence under the competing propositions of the prosecution and defence.

A highly attractive feature of ASR systems is the industry claim of 'language independency' [4], i.e., that the system works with all languages with equal proficiency. However, others postulate that because systems have been trained on predominantly English data [2], there is a reduction in ASR performance using non-English languages known as the "language gap" [5]. However, further research is needed to test the robustness of these claims especially under forensically realistic conditions [6]. This is especially true in the context of language-mismatch between evidential samples. With half of the world's population reported as being bilingual [7] and around 5% of UK casework enquiries containing languages other than English [8], further empirical work is required to test the performance of automatic systems with bilingual data. In this research we examine the extent to which language mismatch either between evidential comparison samples (i.e., the known and questioned samples), or between the evidential samples and the calibration data, affects overall system performance and the resulting strength of evidence (i.e., LRs for individual comparisons).

2. Methods

2.1. Data

Speech samples from 88 Canadian (English-French) bilinguals from the Royal Canadian Mounted Police, Audio and Video Analysis Unit, Speech Research Database (AVAU_UO_data) were used. The participants performed two speaking tasks in both French (Fr) and English (En): Task 1 (T1) is a read passage and Task 2 (T2) is a series of read sentences. The recordings are matched in terms of style (read speech) and channel (studio-quality recordings). Whilst these samples do not represent forensically realistic recordings, our aim was to maximise the potential performance of the system by holding confounding variables constant. Only female speakers were used from the corpus because there were insufficient male speakers who completed both tasks in both languages. Furthermore, female voices have been overwhelmingly understudied in ASR investigations and the few studies that have examined female voices (e.g., [2], [5]) found that female voices pose a greater challenge to ASR systems than male voices, thus exemplifying the importance of examining female speaker performance in greater depth.

2.2. ASR system

Testing was conducted using the state-of-the-art Phonexia Voice Inspector (v.4.0.0) x-vector system. The system generates x-vector speaker models for the 'evidential' comparison samples based on MFCC input. Scores for each comparison are then calculated using PLDA [9].

2.3. Trial configuration

The 88 speakers were divided into two equal groups; the first 44 speakers formed the 'evidential' samples (i.e., 44 pairs of evidential known-speaker (KS) and questioned-speaker (QS) samples), and the second 44 speakers formed the calibration data used to train the calibration model. To ensure stylistic consistency, T1 recordings formed the KS samples and T2 formed the QS samples. The 'evidential' group was arranged

10.21437/Interspeech.2023-680

into four language-matched or -mismatched sets (A-D); Sets A and B were same-language comparisons (English KS – English QS and French KS – French QS, respectively) and Sets C and D were different language comparisons (English KS - French QS and French KS - English QS, respectively). Each of the four sets (A-D) were compared against the four different language configurations in the calibration data to generate 16 trials (see Table 1). In each set, there was a 'fully matched' trial meaning the language configuration of the KS and QS samples in both the evidential and calibration data were an exact match.

Table 1: The overall system performance across all trials. Languages are English (En) and French (Fr). The cells highlighted in green are 'fully matched' trials.

Trial	Set	'Evidential' data languages		Calibration data languages		'Evidential' and calibration data languages Match	Single or mixed language RP match	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}
		KS	QS	KS	QS					
1	A	En	En	En	En	Match	Match	0.0016	0	0.0016
2				Fr	Fr	Mismatch	Match	0.0016	0	0.0016
3				En	Fr	Partial match	Mismatch	0.0540	0	0.0540
4				Fr	En	Partial match	Mismatch	0.1152	0	0.1152
5	B	Fr	Fr	En	En	Mismatch	Match	0.0074	0	0.0074
6				Fr	Fr	Match	Match	0.0071	0	0.0071
7				En	Fr	Partial match	Mismatch	0.2206	0	0.2206
8				Fr	En	Partial match	Mismatch	0.4066	0	0.4066
9	C	En	Fr	En	En	Partial match	Mismatch	6.28E-04	0	6.28E-04
10				Fr	Fr	Partial match	Mismatch	0.0023	0	0.0023
11				En	Fr	Match	Match	0.0738	0	0.0738
12				Fr	En	Partial match	Match	0.1487	0	0.1487
13	D	Fr	En	En	En	Partial match	Mismatch	2.2557	0.034	2.2217
14				Fr	Fr	Partial match	Mismatch	1.2312	0.034	1.1973
15				En	Fr	Partial match	Match	0.1103	0.034	0.0764
16				Fr	En	Match	Match	0.0731	0.034	0.0392

We tested three conditions that represented three different forensic scenarios. 'Single language' refers to the KS and QS samples being in the same language. 'Mixed language' refers to the KS and QS samples being in different languages.

Condition 1 – *Trials 1, 2, 5 & 6* - Single language calibration data were compared with matched and mismatched single language sets to test the effect of (mis)matched calibration data in forensic casework. This was to test the claim of 'language independency' and the 'language gap' and to test the effects of calibration data mismatch when a matched language reference database may be unavailable.

Condition 2 – *Sets C and D* - Mixed language sets were compared with single and mixed language calibration data to assess ASR performance with bilingual material.

Condition 3 - *Trials 3, 4, 7, 8 and Sets C & D* - Mixed language calibration data were compared with single and mixed language evidential data to assess the effects of a (mis)matched calibration data to determine which combinations of language yield the lowest and least severe errors. These results form a basis for drawing evidential conclusions on appropriate calibration sets in bilingual casework.

We recognise here that some of the tests we conducted would not be appropriate for validation in a forensic case, as there are intentional mismatches which would not constitute data reflective of casework conditions (e.g., using French calibration data for English evidential samples). However, our aim here is to test the magnitude of the effects of language (mis)match at various stages of ASR processing, which in turn will inform procedures for validating systems in bilingual cases.

2.4. Evaluation

Scores for comparisons in the 'evidential' group generated by the automatic system were calibrated using logistic regression based on scores generated from the calibration data. This produced sets of calibrated log LRs which were then used to evaluate system performance based on language (mis)match in both the test and calibration sets. Overall, 44 same-speaker scores and 1,892 different-speaker scores were produced in each trial.

System performance was evaluated using the log LR cost function (C_{llr}) [10] as well as its two constituent parts; C_{llr}^{min} - discrimination loss and C_{llr}^{cal} - calibration loss (see Table 1). C_{llr} is a cost function whereby a penalty is applied which is proportional to the magnitude of contrary-to-fact log LRs [3]. A system that consistently outputs LRs of 1 for all same- and different-speaker comparisons will produce a C_{llr} of 1. A C_{llr} of less than 1 indicates that a system is capturing useful speaker-discriminatory information. Discrimination loss (C_{llr}^{min}) refers to how well the system can discriminate between same-speaker and different-speaker pairs. The lower the C_{llr}^{min} value, the better the system's potential discrimination power. Calibration loss (C_{llr}^{cal}) refers to how suitable the calibration data is for calibrating the system; poor calibration indicates a problem caused by mismatch between the test and calibration data. It is calculated by subtracting the C_{llr}^{min} from the C_{llr} value. The closer the C_{llr}^{cal} value is to zero, the better calibrated the system.

3. Results

3.1. Same language comparisons (Set A and B)

The conditions for set A represent the most typical scenario in the UK whereby the KS and QS are both in English. Set B reflects the most typical scenario in predominantly French-speaking areas, such as France, as the KS and QS are both in French.

3.1.1. *Calibration*

Overall, the results are in line with a high-performing and well-calibrated system with low C_{llr} values, which are expected with analysis of single-language speech and high-quality audio. Single-language calibration data (Trials 1 & 2 in Set A and 5 & 6 in Set B) outperform mixed-language calibration data (Trials 3 & 4 in Set A and Trials 7 & 8 in Set B), with the 'fully matched' trials (language-matched and single-language calibration data-matched) producing the smallest calibration error. This is in line with our expectations. However, both single- and mixed- language calibration data result in well-calibrated systems with low C_{llr} values.

The Tippett plots in Figures 1 and 2 show sets of same-speaker (SS) and different-speaker (DS) LLRs from Sets A and B. The extent of miscalibration using the mixed-language calibration data is predominantly driven by contrary-to-fact DS LLRs using the mixed-language calibration data. These constitute fairly low magnitude errors but have caused a small "right shift" [11] of the mixed-language curves to the right of the zero line.

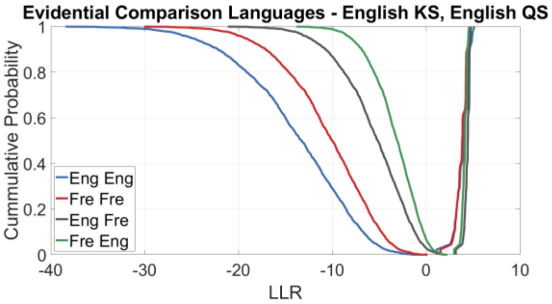

Figure 1: *Tippett plot of LLRs for Set A where the known samples and questioned samples are in English*

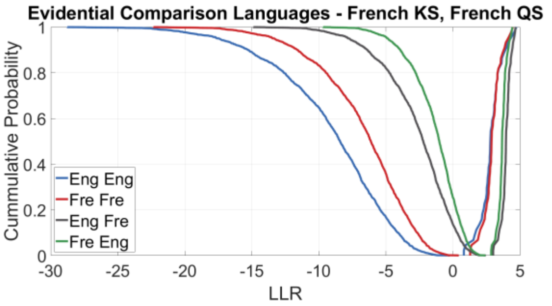

Figure 2: *Tippett plot of LLRs for Set B where the known samples and questioned samples are in French*

3.1.2. *Discrimination*

There is no discrimination error across each set indicated by a C_{llr}^{min} value of zero. This is a floor effect due to the use of high-quality, channel- and style-matched recordings.

3.2. Different language comparisons: Set C

Set C involves mixed-language comparisons; the KS is in English and the QS is in French. This would be a more uncommon scenario in UK casework, but common in bilingual areas of Canada, for example. Calibration was again performed using matched, mismatched and mixed language data.

3.2.1. *Calibration*

In Set C, the 'fully matched' trial where the configuration of the calibration data matches the configuration of the 'evidential' data (English KS - French QS) is Trial 11. We would expect this configuration to produce less severe errors because the calibration recordings are language matched to the 'evidential' recordings. However, this set produces a generally similar pattern of results to sets A and B where the single-language calibration data (i.e., Trials 9 and 10 that have English-only and French-only calibration data, respectively) produce the lowest C_{llr}^{cal} values of the set. This result is unexpected because the mismatch of mixed-language 'evidential' data and single-language calibration data is considered less well-matched and therefore less well calibrated than the 'fully matched' trial (Trial 11) or other bilingual calibration data (Trial 12). In fact, the single-language English calibration data (Trial 9) produces the lowest calibration error of both the set and of our entire experiment with an exceptionally small value of $6.28x10^{-4}$. Single-language French calibration data also produces a low calibration error that outperforms the 'fully matched' trial in Set B - Trial 6. Again, this is not as we expect as Trial 11 is considered more mismatched than Trial 6 because it uses single-language calibration data for bilingual 'evidential' data.

The Tippett plot in Figure 3 reveals the same "right shift" pattern for the mixed-language DS curves. Unlike Sets A and B, the SS curves are very closely aligned between the cumulative probabilities of 0.1 and 0.7.

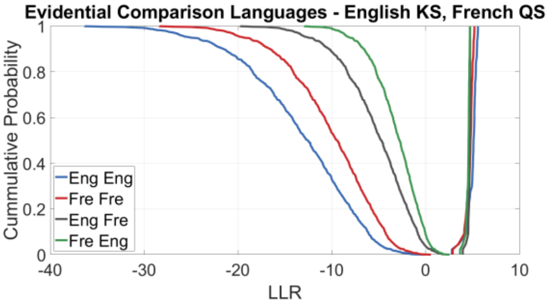

Figure 3: *Tippett plot of LLRs for Set C where the known samples are in English and the questioned samples are in French*

3.2.2. *Discrimination*

There is no discrimination error in this set, similar to Set A and B, with a C_{llr}^{min} value of 0.

3.3. Different language comparisons: Set D

Set D reflects the conditions of a case where the KS is in French, and the QS is in English; it is the same configuration as Set C but with the 'evidential' languages switched. The set produces the largest range of performance but aligns with our prediction that the bilingual calibration data performs better than single-language calibration data and the 'fully matched' trial performs the best.

3.3.1. *Calibration*

Calibration effects are highly mixed in Set D with some very low calibration error and some unusually high errors. As predicted, the best performing test in terms of calibration is the 'fully matched' trial 16 that has French KS – English QS in both the 'evidential' and the calibration data. Miscalibration gets subsequently worse as the calibration data become 'less similar' to the test languages which patterns with the descending order of the trial numbers. The single-language calibration data trials (Trial 13 – English and Trial 14 – French) are significantly less well calibrated, with English calibration data performing considerably worse than all the other tests in both the set and the experiment. These C_{llr} values are extremely high, with the C_{llr}^{cal} values reflecting highly uncalibrated systems.

The Tippett Plot in Figure 4 reveals that calibration loss for Set D is primarily driven by SS errors, with only one instance of a low magnitude DS error (Trial 16). Contrary to the patterns in the previous sets, the high magnitude SS errors causes a significant "left-shift" pattern whereby the curve crossings are positioned exclusively in the negative LLR values.

3.3.2. *Discrimination*

Set D poses the greatest problem to the ASR system in terms of discrimination. With the highest $C_{\mathrm{llr}}^{\mathrm{min}}$ value of the experiment (0.034), the mixed French-English configuration creates more discrimination error than both the single-language sets and the mixed English-French configuration (Set C).

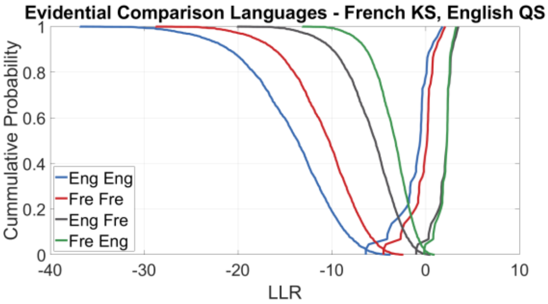

Figure 4: *Tippett plot of LLRs for Set D where the known samples are in French and the questioned samples are in English*

4. Discussion

Overall, the results are indicative of a well performing system that is well-calibrated for the majority of the trials in the sets. This is expected with good-quality, non-forensically realistic data that is channel- and style-matched. We found that, as may be expected, most of the 'fully matched' trials (i.e., ones that are matched in terms of language configuration across the 'evidential' and the calibration data) produce the lowest calibration errors. Generally speaking, the trials that are not matched in terms of single- or mixed-language pairs nor language configuration produce the highest calibration errors due to them being the least well matched.

Results indicate that mixed language 'evidential' comparison sets (C and D) pose a greater challenge to ASR systems than single language test sets (A and B), showing that the system's suitability for bilingual data still requires attention. More severe miscalibration was found in mixed-language 'evidential' and calibration data (C and D) which makes drawing evidential conclusions based on bilingual data of this kind challenging. Nonetheless, there are predictable patterns of directional shifts in log LRs, visible by off-centre zero-line curves in Tippet plots that are consistent with previous research ([2]).

A strong correlation between calibration error and the mixed-language mismatch is present with those tests producing the highest $C_{\mathrm{llr}}^{\mathrm{cal}}$ values within their sets. This demonstrates that language is an important variable when collecting data to build a well-matched calibration model for bilingual casework (see [11]). However, it is noteworthy that this is not the case for all sets and trials; in Set C, single-language English calibration data paired with mixed-language test data produced the lowest calibration error by a substantial margin, both within its set and across the entire experiment. This is an unusual result given that the use of mismatched calibration data typically causes miscalibration. This could indicate the possibility to use single-language calibration data in a case of bilingual evidential samples where a matched bilingual language calibration database is unavailable. However, further testing is needed to assess the robustness of this claim because the same did not apply for Set D. This supports the notion that mismatched data can be unpredictable within ASR systems as it produces more variability in system output due to the mismatch between comparison recordings. Interestingly, the French-only evidential data (Set B) produced slightly higher calibration error than the English-only data (Set A) which shows support for the previous findings [5], [13] of the 'language gap'. However, this finding should not be conflated given that the differences were small and other studies (e.g., [12]) found no preference for English data over non-English data. Further work is warranted to test the conditions on more forensically realistic data using different types of recordings of degraded technical quality.

5. Conclusion

The use of ASR systems when evaluating forensic voice evidence requires caution, particularly when using mismatched-language or bilingual evidence. In this study, it has been shown that bilingual mismatch between evidential data and calibration data is detrimental to system performance and matching reference data to the conditions of the case remains critical for calibration [14]. However, any data set used for calibration will be different from the evidential recordings in some known or unknown ways. Therefore, it is crucial to establish which variables have the greatest effect on system output, in order to concentrate data collection efforts around those variables. For language, the answer to this question remains complex due to the promisingly low calibration error for matched data yet anomalously high calibration error for other matched bilingual data. Whilst mismatched data in both the 'evidential' and the calibration data can produce similarly good results as matched data in certain circumstances, language-mismatched data should not be used uncritically [11]. However, predictable directional 'shifts' in output may help us to better understand expected calibration patterns. We, therefore, concur with those who suggest that dismissing automatic methods entirely when no language-matching calibration data is available would be excessive [11] because when combined with further empirical research, these shifts could provide a foundation upon which to base expected calibration error in real casework. Further work is warranted to test more mismatched language pairs and to measure the extent of the shifts, as well as the effect of technical degradation. More attention should also focus on the performance of individual speakers within these systems to glean what speaker-specific factors lead to better or worse performance, particularly for female speakers who are underrepresented in ASR studies.

6. References

[1] E. Gold and J. P. French, “International practices in forensic speaker comparison: second survey,” in *The International Journal of Speech, Language and the Law*, 2019, vol. 26, no. 1, pp. *1-20*.

[2] H. Künzel, “Automatic Speaker Recognition with cross-language material,” in *The International Journal of Speech, Language and the Law*, 2013, vol. 20, no. 1, pp. *21-44*. DOI: 10.1558/ijsll.v20i1.21.

[3] D. Watt *et al*., “Assessing the effects of accent-mismatched calibration databases on the performance of an automatic speaker recognition system,” in *The International Journal of Speech, Language and the Law:* Equinox, 2020, vol. 27, no. 1, pp. *1-34*, https://doi.org/10.1558/ijsll.41466.

[4] Phonexia. https://bit.ly/3trlEjT (retrieved 2023).

[5] L. Lu, Y. Dong, X. Zhao, J. Liu and H. Wang, “The effect of language factors for robust speaker recognition,” in *Acoustics, Speech, and Signal Processing*, 1988 and in *ICASSP*, 1988, and presented at the Conf. Proceedings of the IEEE ICASSP, Taipei, Taiwan, April 19-24, 2009.

[6] P. Rose, *Forensic speaker identification,* London, England: Taylor & Francis, 2002.

[7] F. Grosjean, *Life with Two Languages: An Introduction to Bilingualism*, Cambridge, Massachusetts: Harvard University Press, 1982.

[8] p.c. J. P. French Associates. https://www.jpfrench.com (retrieved 2023).

[9] M. Jessen, J. Bortlík, P. Schwarz, Y. A. Solewicz, “Evaluation of Phonexia automatic speaker recognition software under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01)” in *Speech Communication*, 2019, vol. 111, pp. *22-28*, ISSN 0167-6393, https://doi.org/10.1016/j.specom.2019.05.002.

[10] N. Brümmer and J. A. du Preez, “Application independent evaluation of speaker detection,” in *Computer Speech and Language*, 2006, vol. 20.

[11] D. Van der Vloed, M. Jessen, and S. Gfroerer, “Experiments with two forensic automatic speaker comparison systems using calibration datas that (mis)match the test language,” presented at the Conference on Audio Forensics, Arlington, VA, USA, June 15-17, 2017.

[12] M. A. Przybocki, A. F. Martin, and A. N. Le, “NIST speaker recognition evaluations utilizing the Mixer corpora 2004, 2005, 2006,” in *IEEE Transactions on Audio, Speech and Language Processing*, 2007, vol. 15, no. 7, pp. *1951–1959*. http://dx.doi.org/10.1109/TASL.2007.902489AFPA.

[13] J. H. Lo. “Issues of bilingualism in likelihood ratio-based forensic voice comparison,” PhD dissertation, Dept. Lang. and Ling. Science, Univ. York, 2021. [Online]. Available: https://etheses.whiterose.ac.uk/30007/.

[14] G. Morrison *et al*., “Consensus on validation of forensic voice comparison,” in *Science & Justice:* Elsevier, 2021, vol. 61, issue 3, pp. *299-309*, https://doi.org/10.1016/j.scijus.2021.02.002.