Automatic Speaker Recognition performance with (mis)matched bilingual speech material

Bryony Nuttall[1,2], *Phillip Harrison*[2], *and Vincent Hughes*[2]
[1] *J P French Associates, York*
bryony.nuttall@jpfrench.com
[2, 3] *Department of Language and Linguistic Science, University of York*
{philip.harrison | vincent.hughes}@york.ac.uk

The preferred presentation format would be a poster.

In order to validate any forensic voice comparison system, it is necessary to test using samples that are reflective of the conditions of the case (Morrison et al. 2021). However, the extent to which certain speaker or technical factors affect system output remains an empirical question. This research, conducted as part of a MSc dissertation, contributes towards this area by considering the effect of language in automatic speaker recognition (ASR) systems used in forensic casework. Specifically, we examine the extent to which language mismatch either between the known and unknown samples, or between the evidential samples and the reference population (RP) used for calibration, affects overall system performance and the resulting strength of evidence (i.e., likelihood ratios for individual comparisons).

Testing was conducted using the state-of-the-art Phonexia Voice Inspector (v.4.0.0) x-vector system and speech samples from 88 Canadian English-French bilinguals from the RCMP database. There were three matched and mismatched language conditions:

Condition 1 – Single language test data and different language RP data *Tests 1, 2, 5 & 6*
Single language RP data were compared with (mis)matched single language test data to test the effect of (mis)matched RP data in cases where a matched language reference database may be unavailable.

Condition 2 – Mixed language test data *Sets C and D*
Mixed language test data (where the known speaker sample is one language and questioned speaker sample is another language) were compared with single and mixed language RP data to assess ASR performance with bilingual material.

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

System performance was evaluated using the log LR cost function (C_{llr}) as well as its two constituent parts (C_{llr}^{cal} - calibration loss; C_{llr}^{min} - discrimination loss).

Results indicate that mixed language test comparison sets (C & D) pose a greater challenge to ASR systems than single language test sets (A & B), showing that the system's suitability for bilingual data still requires attention. More severe miscalibration was found in mixed language test and reference data (C & D) which makes drawing evidential conclusions based on this bilingual data challenging. Nonetheless, there are predictable patterns of directional shifts in log LRs which are consistent with previous research. When combined with further empirical research, these shifts could provide a foundation on which to base expected calibration errors in real casework.

Table 1. Overall system performance across all tests. Languages include English (En) and French (Fr).

# Test	Test set	Test language(s)		RP language		Language match?	Single or mixed language RP match?	Cllr	$Cllr^{min}$	$Cllr^{cal}$
		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

References

Morrison, G. S. et al. (2021) Consensus on validation of forensic voice comparison. *Science and Justice* 61(3): 299-309.

Royal Canadian Mounted Police speech research database, Audio and Video Analysis Unit (AVAU_UO_data). (2010-2016). Provided by the University of York.