INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Evaluation of a forensic automatic speaker recognition system with emotional speech recordings

Robert Essery, Philip Harrison, Vincent Hughes

Department of Language and Linguistic Science, University of York, UK
robert.essery22@gmail.com, {philip.harrison|vincent.hughes}@york.ac.uk

Abstract

In forensic contexts, speakers often feel emotional, which will likely influence their speech. Emotional mismatch between samples is therefore a source of variability which could have substantial effects on the performance of a forensic automatic speaker recognition system. This paper examines the issue of emotional speech in forensic casework, both in terms of emotional match and mismatch between test samples and in terms of the data used to calibrate the system (i.e. the reference population). Specifically, we tested system performance on samples of neutral and acted angry and fearful speech data across 37 test conditions. The best system performance was achieved when the test data and reference population conditions matched exactly. However, in 16 of the 37 tests, the system produced a Cllr greater than 0.8, 10 of which also exceeded a Cllr of 1. As a result, caution should be used to interpret the results of automatic and semi-automatic forensic analysis on emotional speech data.

Index Terms: forensic voice comparison, emotion, mismatch, likelihood ratio, validation, automatic speaker recognition

1. Introduction

1.1. FASR Systems and Testing

Forensic Automatic Speaker Recognition (FASR) systems are currently being used to varying extents around the world in criminal casework by police forces and forensic analysts. However, as with all forms of forensic evidence, it is important that such systems are tested and empirically validated under conditions that are reflective of casework [1]. This testing focuses on how well a system can differentiate between pairs of same-speaker (SS) and different-speaker (DS) speech samples.

The process begins like non-forensic automatic speaker recognition (ASR) systems with an initial *feature-to-score* stage. First, at least two speech recordings are input into the system: one questioned sample (QS) and one known sample (KS). Next, vectors of acoustic features are extracted from overlapping frames from across the voice-active portions of the samples, which are then converted to speaker models, such as an x-vector [1]. These models are then compared against each other to generate a *score*. However, this score is generally not appropriate for forensic purposes as it fails to consider the typicality of the score in relation to a relevant population of similar-sounding speakers [2].

The output of a FASR system, however, is a (hopefully well-calibrated) likelihood ratio (LR), which quantifies the strength of evidence under the competing propositions of the prosecution and the defence. In the context of FASR, the LR is defined as:

$$\frac{p(s|H_{ss})}{p(s|H_{ds})} \quad (1)$$

where s is the score and H_{ss} and H_{ds} are distributions of same-speaker and different-speaker scores from the relevant population. The LR framework has been recommended by the European Network of Forensic Science Institutes (ENFSI) [3] and many expert practitioners of FASR [4].

The validity of the system can then be assessed based on these calibrated LRs using metrics such as the log-likelihood ratio cost (C_{llr}), comprised of discrimination error (C_{llr}^{min}) and calibration error (C_{llr}^{cal}). This means that C_{llr} captures the magnitude of contrary-to-fact LRs rather than just the quantity [5]. A C_{llr} over 1 is considered very poor system performance, while 0 is optimal [4].

1.2. Data Mismatch and Emotional Speech

Both the quantity and magnitude of contrary-to-fact outputs that a FASR system makes can be affected by mismatches between the QS, KS, and reference population samples. These mismatches can be categorised in terms of technical factors and linguistic factors. Much work in FASR testing has focused on technical mismatch. However, previous studies testing FASR systems have also found that linguistic mismatch, in terms of language [6] and accent [7], degrade system performance, likely due to variability in speech production. Both found that increased calibration error was the main driving factor for system degradation.

The effect of emotion on speech is highly variable for every individual. A speaker may react differently to the same emotion. Perceptually, anger could be characterised in speech by increased vocal effort ranging from harsh voice to shouting [8]. However, it is also possible for a speaker to become quieter or use more precise articulation than in neutral speech. Realisations of fear speech could include quiet strained mumblings caused by vocal tract rigidity [9] or loud screamed speech, for example. Acoustically, both anger and fear are known to influence features like phoneme duration, fundamental frequency [10], and formant frequencies, with wide variability when compared to neutral speech [11][12].

In forensic casework, QS recordings are often taken from scenarios involving highly charged emotions. This will cause an emotional mismatch if the KS only contains neutral speech from police interviews, for example. A mismatch between samples should generally be avoided but suppose the only available QS speech data contains emotional speech. To conduct robust analysis, it is essential that forensic experts understand how the mismatch may affect their results to draw more informed conclusions.

10.21437/Interspeech.2023-2214

To the best of our knowledge, no previous studies have evaluated an LR-based FASR system on emotional speech data. However, studies testing non-forensic ASR systems strongly suggest that the performance will be degraded. [13] found that matched emotion comparisons achieved a correct identification rate greater than 99%. When neutral speech was compared with emotional speech, including anger and fear, correct identification rate fell to less than 60%, and the system was considered to have failed. [8] considered shouted and increased vocal effort speech data and found that neutral and angry matched comparisons achieved identification rates greater than 99%, but this fell below 30% in mismatched comparisons. In a more recent study, [14] investigated different methods of model generation to improve performance on emotional speech. Their most successful method achieved identification rates between 60% to 65% in neutral-emotional mismatched conditions for anger and fear.

1.3. Current Study

This study investigates whether emotional speech causes the same pattern of system performance degradation in forensic systems when compared to neutral speech. The results will also evaluate the extent to which a match between the reference population and test data in terms of emotion affects system performance. The study focuses on *Anger* and *Fear* as we believe these are likely to occur in forensic casework, meaning an increased likelihood that a speaker's voice could be affected by these emotions during a criminal event, or in a reference recording.

1.3.1. Limitations

It must first be acknowledged that the speech data used for the experiment is high quality, acted speech data and thus not forensically realistic. This means that the absolute performance of the system will be overoptimistic relative to real forensic conditions. However, our interest here is in the relative effects of different emotions and emotion mismatch as a means of gauging the importance of emotion as a source of variability in FASR testing. The acted speech samples also have slightly varied realisations of the same emotion based on the actor's interpretation of fear and anger. This is problematic as the various possible emotional realisations between speakers could be considered another form of data mismatch.

2. Method

2.1. Corpus

The speech recordings used in the experiment were taken from The Crowd-sourced Emotional Multimodal Actors Dataset (CREMA D) [15]. This data set contains 7442 recordings of acted emotional speech by 91 speakers (48 male and 43 female), aged between 20 and 74 years, of various ethnicities. However, all spoke with a General American English accent. The actors spoke the same twelve sentences in six emotions: anger, disgust, fear, happiness, neutral, and sadness.

2.2. Preparation of Audio Files

Only samples from the male speakers were used for the experiment due to the higher quantity of usable speakers who sounded similar. *Fear* and *Anger* were chosen as the two non-neutral emotions due to their higher likelihood of occurrence in forensic casework. Also, the stereotypical effects of these emotions on speech differ enough from neutral speech and each other to the extent that it could be classed as a mismatched linguistic factor when compared against each other.

Based on these reasonings, after an initial auditory assessment of all the recordings for the relevant emotions, 20 speakers were eliminated, leaving 28 speakers for testing. Speakers were eliminated if their neutral speech samples sounded significantly different from the majority of other speakers in terms of linguistic features. Also, speakers were eliminated when their realisation of emotional speech significantly differed from the majority of other speakers or was too similar to their neutral speech samples that it could not be considered a linguistic factor mismatch.

Each speaker's twelve sentences were concatenated into one sound file for each emotion. These were then split in half to simulate a KS and QS for each speaker in each emotion. The average net speech for each emotion's KSs and QSs ranged from 9.1 seconds to 12.8 seconds. Greater amounts of net speech were observed in the Anger samples.

2.3. Experiment

Scores were calibrated for nine possible KS and QS emotion combinations with multiple reference populations, totalling 37 different test conditions as shown in Table 1. Each test condition simulates a possible scenario in forensic casework. For example, Tests 12 to 15 simulate a mismatched condition where the KS contains neutral speech from a police interview, and the QS contains angry speech from a covertly recorded conversation, for example.

Within each test, 28 same-speaker (SS) and 756 different-speaker (DS) comparisons were conducted. The state-of-the-art Phonexia Voice Inspector [16] FASR system was used to compute the scores. Voice Inspector is an x-vector system using deep neural networks to convert acoustic features into compact speaker representations (x-vectors). For each SS and DS pair, the x-vectors are compared to generate a score using probabilistic linear discriminant analysis. Due to the limited number of speakers available, calibration was conducted using cross-validation. For each comparison within a test, two speakers' scores from the relevant score-set were isolated. The reference population comprises the remaining speakers' SS and DS scores from the score-set that matched the emotion combination required for that test. Those scores were used to train a logistic regression model, which was then applied to the scores for the comparison speakers to generate a calibrated log LR (LLR). In this way, the calibration coefficients change slightly for each pair of samples under analysis. System performance for each test was then evaluated based on the entire set of calibrated LLRs using the equal error rate (EER), C_{llr}, C_{llr}^{min}, and C_{llr}^{cal}. This process was then repeated for all 37 tests.

3. Results

The results of the experiment are presented in full in Table 1. Test 1, with matched neutral test data paired with a matched neutral reference population, resulted in a 0% EER, C_{llr} of 0.05, C_{llr}^{min} of 0.014, and C_{llr}^{cal} of 0.034. Thus, in optimal conditions, the system performs very well. This is unsurprising, especially given the use of high-quality, channel-matched samples. When the other matched emotion reference populations were used for calibration in Tests 2 and 3, the system still performed well with

Table 1: Experiment plan detailing the emotion combinations of the test data and paired reference population samples for each of the 37 test conditions, as well as the validation test results. KS and QS refer to the 'known sample' and 'questioned sample' respectively.

Test	Test Data		Reference Population		EER	C_{llr}	C_{llr}^{min}	C_{llr}^{cal}
	KS	**QS**	**KS**	**QS**				
1	Neutral	Neutral	Neutral	Neutral	0%	0.048	0.014	0.034
2			Anger	Anger	0%	0.053	0.013	0.040
3			Fear	Fear	0%	0.071	0.013	0.058
4	Anger	Anger	Neutral	Neutral	3%	0.178	0.053	0.125
5			Anger	Anger	3%	0.104	0.062	0.042
6			Neutral	Anger	4%	0.973	0.105	0.869
7			Anger	Neutral	3%	0.870	0.058	0.812
8	Fear	Fear	Neutral	Neutral	7%	0.312	0.221	0.091
9			Fear	Fear	7%	0.294	0.235	0.058
10			Neutral	Fear	7%	0.697	0.220	0.477
11			Fear	Neutral	7%	0.504	0.216	0.288
12	Neutral	Anger	Neutral	Neutral	14%	2.042	0.411	1.630
13			Anger	Anger	14%	4.324	0.402	3.922
14			Neutral	Anger	14%	0.510	0.429	0.081
15			Anger	Neutral	14%	0.517	0.409	0.108
16	Anger	Neutral	Neutral	Neutral	11%	1.378	0.329	1.048
17			Anger	Anger	11%	2.978	0.313	2.665
18			Neutral	Anger	14%	0.440	0.322	0.117
19			Anger	Neutral	14%	0.402	0.338	0.064
20	Neutral	Fear	Neutral	Neutral	18%	2.689	0.475	2.214
21			Fear	Fear	12%	1.635	0.479	1.156
22			Neutral	Fear	18%	0.586	0.500	0.086
23			Fear	Neutral	18%	0.661	0.494	0.167
24	Fear	Neutral	Neutral	Neutral	13%	1.357	0.338	1.018
25			Fear	Fear	14%	0.841	0.344	0.497
26			Neutral	Fear	14%	0.517	0.341	0.176
27			Fear	Neutral	14%	0.502	0.353	0.148
28	Anger	Fear	Neutral	Neutral	18%	1.427	0.454	0.973
29			Anger	Anger	18%	3.058	0.464	2.593
30			Fear	Fear	18%	0.904	0.449	0.455
31			Anger	Fear	18%	0.641	0.477	0.164
32			Fear	Anger	18%	0.610	0.452	0.158
33	Fear	Anger	Neutral	Neutral	14%	0.999	0.410	0.589
34			Anger	Anger	14%	2.197	0.393	1.804
35			Fear	Fear	14%	0.632	0.405	0.227
36			Anger	Fear	14%	0.504	0.401	0.102
37			Fear	Anger	14%	0.506	0.420	0.085

C_{llr}s less than 0.07. However, the matched fear reference population resulted in slightly more calibration error. Performance was slightly worse with matched test data using the anger and fear data paired with matched reference populations (Tests 4/5 and 8/9). However, these tests did achieve C_{llr}s ranging from 0.104 to 0.312, so performance was still relatively good, with anger resulting in better performance than fear. Also, for both anger and fear, the system performed better when the matched reference population reflected the emotion of the test data, with the matched neutral reference populations leading to slightly higher C_{llr}s.

The system performed relatively poorly in tests with matched emotion test data paired with mismatched neutral-emotion/emotion-neutral reference populations (Tests 6/7 and 10/11). Although C_{llr}s for these tests did not exceed 1, they ranged relatively high between 0.70 and 0.98. However, the fear test data paired with a fear-neutral reference population resulted in a more reasonable C_{llr} of 0.50. It appears these higher values are more driven by calibration error, especially for fear, as the C_{llr}^{min} values remained rather consistent between 0.216 to 0.235 regardless of reference population match or mismatch.

The system performed well on all but two tests with mismatched neutral-emotion/emotion-neutral test data paired with a mismatched neutral-emotion/emotion-neutral reference population (Tests 14/15, 18/19, 22/23, and 26/27) with C_{llr}s less than 0.517. However, most of these tests also produced relatively high EERs of 14%, and the tests with neutral-fear test data (Tests 22 and 23) produced EERs of 18% and high C_{llr}^{min} values ranging from 0.49 to 0.5, meaning high discrimination error. This indicates poorer system performance, but the lower C_{llr}^{cal} values meant the C_{llr}s for these tests did not exceed 0.661.

Performance was very poor on tests with mismatched neutral-emotion/emotion-neutral test data paired with matched emotion reference populations (Tests 12/13, 16/17, 20/21, and

24/25). Most of these tests resulted in C_{llr}s ranging from 0.84 to 2.9, and the neutral-anger test data paired with matched anger reference population (Test 13) produced a C_{llr} of 4.3, the highest of all the tests. This high C_{llr} is primarily driven by calibration error with a high C_{llr}^{cal} of 3.92, while the C_{llr}^{min} remained similar to the other tests using the same test data ranging from 0.402 to 0.429. The neutral-fear test data tests (Tests 20 and 21) also stood out for producing high C_{llr}^{min} scores, both at 0.48, with the matched neutral reference population (Test 20) producing a relatively high 18% EER. Interestingly, the matched neutral reference population resulted in worse performance in the fear tests than the matched fear reference population. However, in the anger data tests, the matched neutral reference population performed better than the matched anger reference population.

The system performed reasonably well on the tests with mismatched anger-fear/fear-anger test data paired with the mismatched anger-fear/fear-anger reference populations (Tests 31/32 and 36/37) with C_{llr}s ranging from 0.50 to 0.64. However, these are some of the highest values of all the tests with mismatched test data paired with a mismatched reference population. Also, the EER ranges from 14% to 18% in these tests. This suggests that the lack of neutral speech in any sample results in poorer performance.

The system performed even worse on the tests with mismatched anger-fear/fear-anger test data paired with matched emotion reference populations (Tests 28/29/30 and 33/34/35). C_{llr}s were very high with three tests: 28, 29 and 34, exceeding 1, and two more tests: 30 and 33 exceeding 0.9. The highest C_{llr} values came from the tests with matched anger reference populations (29 and 34), driven by high C_{llr}^{cal} values of 2.6 and 1.8. The matched neutral reference populations resulted in the next best performance, followed by matched fear.

4. Discussion

4.1. Test Data Match/Mismatch

The results show the best performance when the emotion of the speech in the KS and QS are the same, regardless of the reference population. As expected, the matched neutral test data tests produced the lowest C_{llr} and EER of all the tests. While performance degraded slightly in the matched anger and matched fear tests, the system still performed reasonably well, depending on the paired reference population. The leading cause for this performance loss is calibration error.

Interestingly, the specific emotion appears to influence the extent of the degradation in performance in the matched test data tests. For example, anger consistently resulted in slightly higher calibration error than fear. However, fear consistently produced much higher discrimination error, resulting in the poorest system performance of the matched test data tests.

System performance is worse when the test data is mismatched. Test 19 showed the best system performance of all the mismatched test data tests with a C_{llr} of 0.402. This is higher than the C_{llr} values in seven of the eleven matched test data tests. This shows that even if a reference population with the same mismatched conditions as the test data can be used, performance will still not be better than in a matched test data with matched reference population scenario.

Performance degrades even further when there is an emotion mismatch, and neither emotion is neutral. This was somewhat expected as systems have generally been trained with neutral speech, so any type of analysis using neutral speech should result in better performance. The C_{llr} values of the best-performing anger-fear/fear-anger tests are worse than all the neutral-emotion/emotion-neutral tests. Also, EERs are higher on average, especially for the anger-fear tests.

4.2. Reference Population Match/Mismatch

Unsurprisingly, the results also show that system performance is better when the conditions of the reference population match the test data as closely as possible.

In the matched anger-anger and fear-fear tests, relatively good system performance was only achieved when the reference population was matched neutral or matched emotion. As expected, using the matched reference population in the relevant emotion resulted in better performance. However, a matched neutral reference population still resulted in good system performance for both the anger-anger and fear-fear conditions. The mismatched emotion reference populations resulted in extremely poor C_{llr} values. This was driven by increased calibration error, while discrimination error remained consistent regardless of the reference population.

In both the mismatched neutral-emotion/emotion-neutral tests and the mismatched anger-fear/fear-anger tests, the best performance was observed when paired with mismatched neutral-emotion/emotion-neutral reference populations. Like the matched tests, when the conditions of the test data and reference population matched, the C_{llr} showed good performance. Still, the tests with mismatched test data paired with a matched reference population gave the worst performance, with C_{llr} values ranging from over 1 to just over 4 in ten of the fourteen tests. Again, this is heavily driven by calibration error.

Interestingly, in the mismatched anger-fear/fear-anger tests, the matched anger reference population caused the worst performance. Given that neither testing sample contained neutral speech, it was expected that this would result in the worst performance due to a complete mismatch though this was not the case. Regardless, the C_{llr} with any matched emotion reference population was very high in each of these tests.

5. Conclusion

This study provides insights into the relative performance of a FASR with emotion matched and mismatched data of the kind found in forensic casework. We found that emotional speech degrades system performance. To achieve the best performance with emotional speech data, the test data must be matched in terms of emotion and paired with a matched reference population of the same emotion. In the event of a mismatch in the test data, a reference population should be used that has the same emotion mismatch as the test data to achieve the best performance. Using any reference population that does not match the emotion combination in the test data would result in poor performance driven by calibration error. These results suggest that caution should be exercised in using FASR with emotional speech data. Future studies should look at further testing of different FASR systems on actual emotional speech data in forensically realistic conditions to better determine how reliable these systems are. Also, future studies looking at data mismatch regarding same-emotion realisation and intensity would produce beneficial results for actual casework.

6. References

[1] G. S. Morrison, E. Enzinger, E. Ramos, J. González-Rodríguez and A. Lozano-Díez, "Statistical models in forensic voice comparison" in *Handbook of Forensic Statistics*, D. L. Banks, K. Kafadar, D. H. Kaye, and M. Tackett, Eds. Boca Raton, FL: CRC, 2020, pp. 451-497.

[2] 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 & Justice*, vol. 58, no. 1, pp. 47–58, 2018.

[3] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological guidelines for best practice in forensic semiautomatic and automatic speaker recognition*. Frankfurt: Verlag für Polizeiwissenschaft, 2015.

[4] G. S. Morrison, E. Enzinger, V. Hughes, M. Jessen, D. Meuwly, C. Neumann, S. Planting, W. C. Thompson, D. van der Vloed, R. J. F. Ypma, C. Zhang, A. Anonymous, and B. Anonymous, "Consensus on validation of Forensic Voice comparison," *Science & Justice*, vol. 61, no. 3, pp. 299–309, 2021.

[5] P. Rose and E. Winter, "Traditional forensic voice comparison with female formants: Gaussian mixture model and multivariate likelihood ratio approaches.," in *Proceedings of the 13th Australasian Conference on Speech Science and Technology. Melbourne, Australia*, pp. 42–45, 2010.

[6] D. van der Vloed, M. Jessen, and S. Gfroerer, "Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language," in *Proceedings of the Audio Engineering Society International Conference on Audio Forensics*., Papers 2-1, June 2017.

[7] D. Watt, P. Harrison, V. Hughes, P. French, C. Llamas, A. Braun, and D. Robertson, "Assessing the effects of accent-mismatched reference population databases on the performance of an automatic speaker recognition system," *International Journal of Speech Language and the Law*, vol. 27, no. 1, pp. 1–34, 2020.

[8] C. Hanilçi, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, and F. Ertaş, ''Speaker identification from shouted speech: Analysis and compensation,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., pp. 8027–8031. May 2013.

[9] J. Pittam and K. R. Scherer, "Vocal expression and communication of emotion," in *Handbook of emotions*, M. Lewis, and J. M. Haviland, Eds. The Guilford Press, 1993, pp. 185–197.

[10] M. A. Hagenaars and A. van Minnen, "The effect of fear on paralinguistic aspects of speech in patients with panic disorder with agoraphobia," *Journal of Anxiety Disorders*, vol. 19, no. 5, pp. 521–537, 2005.

[11] S. Yildirim, M. Bulut, C. Lee, A. Kazemzadeh, C. Busso, Z. Deng, S. Lee, and S. Narayanan, "An acoustic study of emotions expressed in speech," in Proc. 8th Int. Conf. Spoken Language Process., Jeju Island, Korea, pp. 2193–2196, Oct. 2004.

[12] T. Özseven, "The acoustic cue of fear: investigation of acoustic parameters of speech containing fear," *Archives of Acoustics*, vol. 43, no. 2, pp. 245–251, 2018.

[13] M. V. Ghiurcau, C. Rusu, and J. Astola, ''A study of the effect of emotional state upon text-independent speaker identification,'' in Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP), pp. 4944–4947, May 2011.

[14] A. B. Nassif, I. Shahin, A. Elnagar, D. Velayudhan, A. Alhudhaif, and K. Polat, "Emotional speaker identification using a novel capsule nets model," *Expert Systems with Applications*, vol. 193, p. 116469, 2022.

[15] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," IEEE transactions on affective computing, vol. 5, no. 4, pp. 377–390, 2014.

[16] *Phonexia Voice Inspector*. (4.0.2) [Computer Software]. Available: https://bit.ly/3trlEjT