Variability in performance across four generations of automatic speaker recognition systems

Anonymous submission to Interspeech 2025

Abstract

The field of automatic speaker recognition (ASR) has seen a series of generational changes to speaker modelling approaches in the last 3 decades. Adoption of new approaches has mainly been driven by improvements observed in overall system-level performance metrics on common datasets. There is now considerable debate within the field around understanding why systems perform better for some speakers than others. In this study, we compare the performance of 4 generations of ASR systems with the same set of forensically-relevant test and calibration data. On a system- and individual speaker-level, we observe improvements from GMM-UBM to i-vector to x-vector but not for ECAPA-TDNN. We find that certain individuals remain difficult to recognise across all systems. Our findings show that both file- and speaker-level factors contribute to the performance of individual speakers and systems overall, which supports calls for more detailed exploration of system performance.

Index Terms: automatic speaker recognition, forensic applications, by-speaker performance

1. Introduction

1.1. Speaker modelling approaches

Over the last few decades, there has been a series of step changes in speaker modelling approaches used in automatic speaker recognition (ASR) systems. In the early 2000s, Gaussian Mixture Models (GMM) were the predominant approach; these are generative models of raw short-term acoustic features such as MFCCs, summarised with a series of means, variances, and weights. The GMM-UBM approach [1] incorporates a Universal Background Model (UBM) to increase generalisability of the model, as well as Maximum a Posteriori (MAP) adaptation, which involves adapting the UBM towards the data from a target speaker to build a target speaker model. The early 2010s welcomed i-vectors [2], an extension of GMMs whereby features are converted to a compact, fixed-length vector via projection in a total variability subspace. Then x-vectors [3] were introduced, which incorporate neural architectures to produce fixed-length speaker models from an embedding within a time-delay neural network (TDNN). The latest generation of speaker modelling is Emphasized Channel Attention, Propagation and Aggregation in TDNN (ECAPA-TDNN) [4], which shares a similar approach to x-vectors but with the addition of a ResNet neural architecture and an attention mechanism.

The aims of each new generation of speaker modelling are to maximise between-speaker variability and minimise within-speaker variability, and to reduce the effects of nuisance variables (often technical, e.g. noise, channel, duration). Improvements in overall system performance are generally reported from one generation to the next, and the community converges around the new approach without necessarily exploring why the new approach works better for some speakers than for others.

1.2. Evolution of the field of speaker recognition

Two major issues with current approaches to ASR system development were raised at a special panel session of the 2024 Odyssey workshop. First, there is a convergence of approaches. As soon as a new approach is shown to outperform its predecessor, the community jumps to the new state-of-the-art. This is driven by the second issue raised, which is the focus on benchmarking exercises such as the regular speaker recognition evaluations organised by NIST [5]. Current evaluative approaches centre around achieving the best performance on benchmarking datasets, with improvements measured at a global level, e.g. Equal Error Rate (EER). This focus on overall error metrics leads to a number of problems, e.g. it masks variability in system performance as a function of speaker or other factors, and does not consider specific use cases, such as the application of ASR systems in the forensic domain.

Increasingly, ASR systems are being used to generate forensic evidence in voice comparison cases [6]. For forensic applications, users need to know that the system works under the conditions of their specific case. It is therefore necessary to test and validate the system prior to use in a forensic case, in order to fully understand the extent of variation in performance as a function of factors commonly encountered in casework. Further, of crucial concern to the analyst is the specific voices being compared, thus understanding system performance at a speaker-specific level is a priority, i.e., *how does the system perform for the specific type of speakers in this case?* [7] and [anon] begin to explore performance variability at an individual speaker-level, investigating why certain speakers may prove more difficult to recognise than others. Both papers investigate the phonetic content of recordings and how the inclusion or exclusion of different types of speech sounds impacts ASR performance (measured via by-speaker C_{llr}).

1.3. This study

This study builds on [anon], which focused on speaker-level variability in ASR performance and started to explore *why* some speakers prove more challenging, namely by manipulating the phonetic content of the samples. In the present study, we use the same set of forensically-realistic recordings to explore speaker-level variability in performance across four generations of an ASR system based on different speaker modelling approaches. The four approaches, from oldest to most recently developed, are GMM-UBM, i-vector, x-vector and ECAPA-TDNN. We first compare performance at an overall system-level, with a focus on speaker discrimination. Then we compare performance at the speaker level, exploring how consistently the approaches perform for individual speakers. We conduct a detailed examination of the results, at both the level of the speaker and of individual comparisons, in order to assess why some speakers consistently prove challenging even to the best-performing speaker modelling approach.

2. Methods

2.1. Data

The data for this study comes from GBR-ENG, a dataset of forensically-realistic recordings collected and provided by the UK Government. The full dataset contains 1,946 speakers (906 male, 1,040 female) of British English, with considerable variability in age, and regional and social background. There are multiple samples for each speaker (mean = 10; 12,483 files in total), typically recorded over a number of days. Samples contain spontaneous conversational speech, have a duration of between 181 and 373 seconds, and are telephone recordings with a mix of landline and mobile recordings.

2.2. Test and calibration sets

For this study we used the same subset of GBR-ENG as in [anon], composed of 98 male speakers; these are divided into a test set comprising 48 speakers with between 3 and 7 files each (160 files total) and a calibration set comprising 50 speakers with 2 files each. All recordings are mobile telephone calls recorded on different days, with between 41 and 236 seconds of net speech, and are relatively good quality in terms of signal-to-noise ratio and little to no clipping.

2.3. Automatic speaker recognition system

Testing was carried out using VOCALISE 2021 (version 3.0.0.1746) [8], which has been widely used for forensic speaker comparison casework. We used three approaches to speaker modelling currently available within the software, along with another comparably-trained approach:

1. **GMM-UBM** with Maximum A Posteriori adaptation
2. **i-vector** with dimension reduction via Linear Discriminant Analysis (LDA) and scoring with a pre-trained Probabilistic Linear Discriminant Analysis (PLDA) model
3. **x-vector** with dimension reduction via LDA and scoring with a pre-trained PLDA model
4. **ECAPA-TDNN** with Cosine Distance scoring

The GMM-UBM and i-vector approaches are trained on the same data, which is a subset of a larger dataset used to train both the x-vector and ECAPA-TDNN approaches. High-level details about training data can be found in [9].

For each approach, same-speaker (SS) and different-speaker (DS) scores were computed for both the test and calibration sets. The calibration scores were used to train a logistic regression model [10] and the coefficients were applied to the test scores to produce calibrated $\log_{10}$ likelihood ratios (LLRs). There were a total of 200 SS LLRs and 12,520 DS LLRs per approach. LLRs were used as the basis for evaluating performance at a system- and speaker-level.

2.4. Evaluation

Overall performance was evaluated using Equal Error Rate (EER) and Log Likelihood Ratio Cost Function (C_{llr}) [11]. In both cases, the closer the value to 0, the better the performance. In the case of C_{llr}, a value of 1 or above means that the system is not providing any useful speaker discriminatory information. C_{llr} has two components: $C_{\text{llr}}^{\text{min}}$ (a measure of discrimination error, where 0 means perfect separation of SS and DS scores) and $C_{\text{llr}}^{\text{cal}}$ (a measure of calibration error).

We use a combination of LLRs (including mean SS and DS LLRs) and $C_{\text{llr}}^{\text{min}}$ to evaluate performance at a speaker-level. We opted for $C_{\text{llr}}^{\text{min}}$ since $C_{\text{llr}}^{\text{cal}}$ is volatile with a small number of files per speaker and given our interest in discrimination rather than calibration. A $C_{\text{llr}}^{\text{min}}$ of 0 means that perfect discrimination can be achieved for that speaker with the optimally-selected calibration data.

3. Results

3.1. Overall system performance

Table 1 shows the overall performance for each speaker modelling approach. Major improvements are observed in both EER and C_{llr} from GMM-UBM to i-vector and from i-vector to x-vector. On our dataset, ECAPA-TDNN performs better than i-vector but not as well as x-vector, with EER and C_{llr} values almost double the corresponding values for the x-vector approach.

Table 1: *Overall performance of the four speaker modelling approaches.*

	EER (%)	C_{llr}	$C_{\text{llr}}^{\text{min}}$	$C_{\text{llr}}^{\text{cal}}$
GMM-UBM	44.5	0.97	0.92	0.05
i-vector	23.5	0.67	0.58	0.09
x-vector	3.0	0.13	0.10	0.03
ECAPA-TDNN	7.0	0.27	0.21	0.06

Table 2 shows the correlations between the uncalibrated scores of the comparisons from one generation of speaker modelling approach to the next, separated for SS and DS comparisons. Strong positive correlations are observed in every case for SS comparisons, while moderate to strong correlations are found for DS comparisons.

Table 2: *Pearson's correlation coefficients and p-values for comparisons of raw scores of individual comparisons across each approach and its successor, separated for same-speaker (SS) and different-speaker (DS) comparisons.*

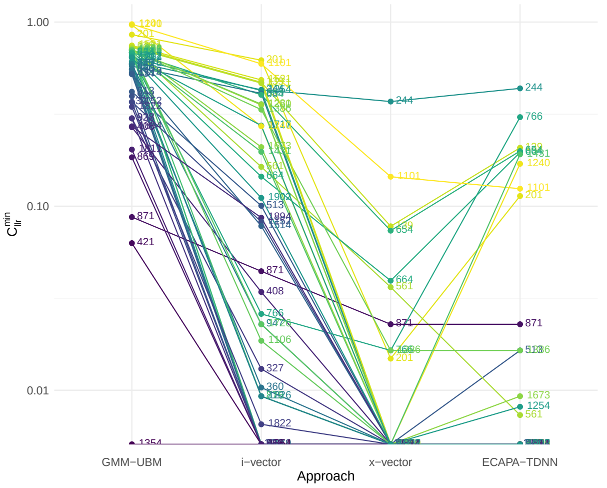
	Approach 1	**Approach 2**	r	**p**
SS	GMM-UBM	i-vector	0.745	<.0001
	i-vector	x-vector	0.742	<.0001
	x-vector	ECAPA-TDNN	0.760	<.0001
DS	GMM-UBM	i-vector	0.834	<.0001
	i-vector	x-vector	0.630	<.0001
	x-vector	ECAPA-TDNN	0.496	<.0001

3.2. By-speaker performance

3.2.1. Speaker discrimination

Figure 1 tracks the by-speaker $C_{\text{llr}}^{\text{min}}$ across the four speaker modelling approaches. First, we focus on the three less recent approaches (GMM-UBM to i-vector to x-vector). For all speakers, $C_{\text{llr}}^{\text{min}}$ decreases from one approach to the next, showing that performance improves with newer generations of modelling approaches. The range of $C_{\mathrm{llr}}{}^{\mathrm{min}}$ across the speakers also decreases from GMM-UBM (0.97) to i-vector (0.62) to x-vector (0.37), showing that overall performance also becomes more consistent within a system.

Figure 1: *By-speaker $C_{llr}{}^{min}$ across four different speaker modelling approaches. Note that the y-axis is log_{10} scaled to provide better resolution at the lower end of the scale (particularly for the x-vector and ECAPA-TDNN approaches).*



An optimal $C_{\mathrm{llr}}{}^{\mathrm{min}}$ of 0 was achieved for 1 speaker using the GMM-UBM approach and 13 speakers using the i-vector approach. The x-vector approach resulted in an EER of 0% and a $C_{\mathrm{llr}}{}^{\mathrm{min}}$ of 0 for 38 of the 48 speakers (79%). EER for the other 10 speakers ranged from 0.21% to 27.92% (0.21% to 2.87% excluding speaker 244) and $C_{\mathrm{llr}}{}^{\mathrm{min}}$ ranged from 0.015 to 0.37. We return to these speakers later.

ECAPA-TDNN did not outperform x-vector, despite being the latest generation tested. An optimal $C_{\mathrm{llr}}{}^{\mathrm{min}}$ of 0 was achieved for 33 speakers (69%) by the ECAPA-TDNN approach; thus, ECAPA-TDNN had perfect discrimination ($C_{\mathrm{llr}}{}^{\mathrm{min}} = 0$) for 5 fewer speakers than x-vector. $C_{\mathrm{llr}}{}^{\mathrm{min}}$ stayed the same for 35 speakers (73%) and increased (i.e., got worse) for 11 speakers (23%). Only 2 speakers (4%) had a lower (i.e., better) $C_{\mathrm{llr}}{}^{\mathrm{min}}$.

Table 3: *Spearman's rank correlation coefficients and p-values for comparisons of by-speaker $C_{llr}{}^{min}$ across each approach and its successor.*

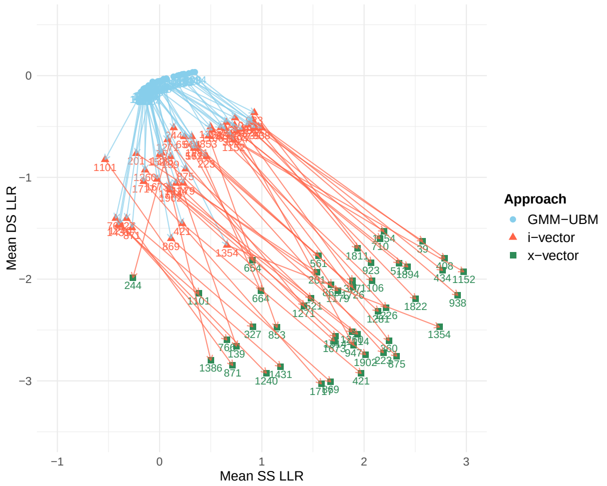
Approach 1	Approach 2	ρ	p
GMM-UBM	i-vector	0.655	<.0001
i-vector	x-vector	0.482	<.001
x-vector	ECAPA-TDNN	0.799	<.0001

The ranking of speakers within the group was highly correlated across the different approaches (see Table 3), particularly across the GMM-UBM and i-vector approaches and the x-vector and ECAPA-TDNN approaches. The slightly weaker correlation between i-vector and x-vector is likely a result of the large number of speakers for whom $C_{\mathrm{llr}}{}^{\mathrm{min}}$ dropped to 0 when using the x-vector approach. In general, then, if one generation of speaker modelling works well for an individual speaker, the next generation generally also works well for that speaker.

3.2.2. *Log Likelihood Ratios*

Figure 2 shows the mean same-speaker and different-speaker LLRs for each individual speaker across the GMM-UBM, i-vector and x-vector approaches. In general, LLRs become stronger (i.e., further from 0) with regard to both SS and DS comparisons from each approach to the next.

Figure 2: *Mean same-speaker (SS) and different-speaker (DS) log likelihood ratios (LLRs) for GMM-UBM, i-vector and x-vector approaches.*



In every case, the DS LLRs increase in magnitude (towards a larger negative value) from GMM-UBM to i-vector to x-vector, demonstrating better speaker discrimination in newer approaches. SS LLRs also tend to increase in magnitude (towards a larger positive value), with a few exceptions from GMM-UBM to i-vector at the bottom end of the distribution and one exception from i-vector to x-vector.

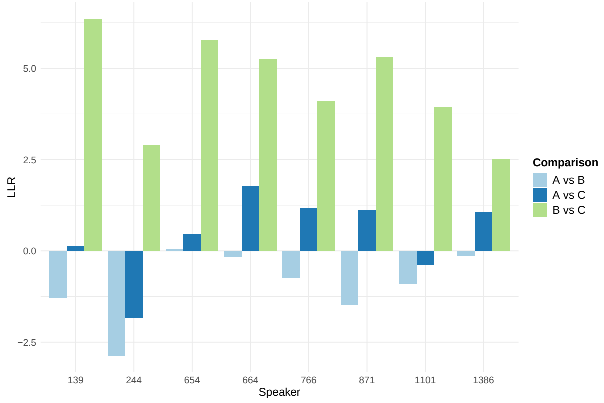
The relationship between the x-vector and ECAPA-TDNN approaches (not shown in Figure 2) was not as clear, though the LLRs generally decreased in magnitude (i.e., got closer to 0) from x-vector to ECAPA-TDNN for SS and/or DS comparisons.

3.3. By-file performance

For this part of the analysis, we focused on the results of the x-vector approach as it outperformed the other methods both overall and at the speaker-level. There were 10 speakers for whom the x-vector approach did not achieve perfect speaker discrimination, i.e., their $C_{\mathrm{llr}}{}^{\mathrm{min}}$ was greater than 0, which is the result of overlap in the LLRs for the SS and DS comparisons involving each of these speakers. It was found that these speakers also had low mean SS LLRs (ranging between -0.26 and 1.56), compared with the mean SS LLRs for the other 38 speakers (ranging between 0.91 and 2.98).

Inspection of the SS LLRs for these 10 speakers revealed that 8 of them had what could be termed a *problem* file. These 8 speakers have 3 files in the test set and therefore 3 SS comparison pairs, i.e., A vs B, A vs C and B vs C. For these speakers, the 2 comparisons involving the problem file (A) produced considerably lower LLRs than the comparison between B and C (see Figure 3). These results suggest that the poorer performance of these speakers is the consequence of the single *problem* file, rather than an inherent speaker issue. We consider the potential causes of this in the discussion.

Figure 3: *LLRs for same-speaker comparisons for 8 speakers with C_{llr}^{min} above 0 (using the x-vector approach) with 1 problem file (A) out of 3.*



No *problem* file was found for the other 2 speakers (numbers 201 and 561; not shown in Figure 3), who have 5 and 4 files in the test set respectively and mean SS LLRs around 1.5, which is higher than the mean SS LLRs of below 1 for the other 8 speakers. For these 2 speakers, examination of the DS LLRs that overlap with the SS LLRs (i.e., those which cause the non-perfect discrimination) shows that almost all of these DS comparisons are with a single speaker. For speaker 201, 3 out of 4 DS comparisons are with speaker 1811 and for speaker 561, 8 out of 9 DS comparisons are with speaker 1254. All LLRs for these comparisons are greater than 0. These results show that a large contributory factor to the poorer performance for these 2 speakers is the comparisons with a single speaker. The same finding was also apparent for some of the other 8 speakers with problem files, highlighting that poorer performance can have multiple causes.

4. Discussion

Using our set of forensically-realistic recordings, improvements in performance at both an overall system- and individual speaker-level were observed from GMM-UBM to i-vector to x-vector; these results were expected given the major developments from one generation to the next. We also found that the x-vector approach outperformed a newer generation of speaker modelling, ECAPA-TDNN. However, the ECAPA-TDNN and x-vector systems used in this study were trained with the same data: all sampled at 8kHz and containing a significant quantity of telephone speech, making it relatively matched to the test conditions. Reference implementations of ECAPA-TDNN shown to outperform x-vector, e.g. [4], have been trained with speech sampled at 16kHz. We speculate that a combination of narrow bandwidth and telephone test condition is responsible for the more effective x-vector system in this study.

The analysis of individual speaker performance showed general trends, revealed some differences in speaker behaviour and highlighted speakers with poorer performance. The individual speaker performance metrics ($C_{\mathrm{llr}}^{\mathrm{min}}$, mean SS and DS LLRs) provided useful insights into the variation found across systems for individual speakers. However, they can still mask details in the results for individual files which may allow some of the differences in performance to be explained.

We concentrated our attention on the 10 speakers who were not perfectly discriminated by the best performing x-vector system. We found that 8 of the 10 speakers had a *problem* file that caused weaker LLRs for SS comparisons involving that file. Preliminary auditory and acoustic analysis of the files revealed observable differences between the *problem* file (A) and the other two files (B and C) for all 8 speakers. The differences related to a range of technical factors (e.g. distance from microphone, background noise, attenuation of frequency bands), speaker factors (e.g. voice quality) and stylistic factors (e.g. increased vocal effort, pitch variability). These factors were present (or absent) to a greater or lesser extent and in different combinations across the *problem* files. If these factors are the cause of the lower SS LLRs, then their number, interdependencies and potential to vary within recordings leads to a complex situation. However, these findings are encouraging as they show that some of the potential causes of poorer performance can be readily observed in recordings. They are also amenable to control, allowing their impact on performance to be tested.

The remaining 2 speakers did not have a *problem* file, but their overlapping DS LLRs resulted almost exclusively from multiple comparisons with only 1 other speaker. Preliminary assessment of these files indicates some similarity in speaker-related rather than technical factors, although not all comparisons between these speakers resulted in LLRs that overlap with SS LLRs. A similar finding was also made for some of the other 8 speakers, which suggests that their poorer performance was due to factors affecting both their SS and DS comparisons.

The findings presented clearly demonstrate that system-level performance metrics mask a wealth of detail about the behaviour of speakers in ASR systems. Even considering performance at a speaker-level can hide valuable insights which are only revealed when looking at the file-level behaviour. These findings clearly support the calls made at the 2024 Odyssey workshop for more detailed investigations into the factors affecting individual performance. Future investigations should focus on disentangling and objectively measuring the factors which influence individual speaker performance.

5. Conclusion

The novel approach we have taken to evaluating ASR performance provides insights into variability both between and within different speaker modelling approaches. Performance is generally shown to improve over generations, which is reflected in both overall performance metrics and speaker-level trends (i.e. good speakers remain good and poor speakers remain poor). This work goes beyond the findings of [anon] and suggests that variability in performance is related to both speaker- and file-level factors. While speaker factors, e.g. the phonetic make-up of samples, may still contribute to a particular file proving problematic, the specific cause is likely a complex combination of technical, speaker and stylistic factors. Understanding individual speaker and file behaviour will ultimately allow us to predict what types of behaviour are more likely to influence system performance, which in turn will assist analysts using ASR systems in forensic voice comparison cases. In future work, we will begin to disentangle the complex interaction of factors contributing to certain files or speakers proving problematic to the ASR system by running a series of experiments using controlled data to explore, for example, the effects of vocal conditions combined with technical factors to understand their impact on performance.

6. References

[1] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Processing*, vol. 10, no. 1-3, pp. 19–41, 2000.

[2] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *Proc. 2018 IEEE International Conference on Acoustics, Speech and Signal Processing*, Calgary, Canada, Apr. 2018, pp. 5329–5333.

[4] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN based speaker verification,” in *Proc. INTERSPEECH 2020 – 21st Annual Conference of the International Speech Communication Association*, Shanghai, China, Oct. 2020, pp. 3830–3834.

[5] C. S. Greenberg, L. P. Mason, S. O. Sadjadi, and D. A. Reynolds, “Two decades of Speaker Recognition Evaluation at the National Institute of Standards and Technology,” *Computer Speech and Language*, vol. 60, p. 101032, 2020.

[6] D. van der Vloed and T. 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.

[7] A. Moez, B. Jean-François, B. K. Waad, R. Solange, and K. Juliette, “Phonetic content impact on forensic voice comparison,” in *Proc. 2016 IEEE Spoken Language Technology Workshop*, San Juan, Puerto Rico, Dec. 2016, pp. 210–217.

[8] F. Kelly, O. Forth, S. Kent, L. Gerlach, and A. Alexander, “Deep neural network based forensic automatic speaker recognition in vocalise using x-vectors,” in *Audio Engineering Society International Conference on Audio Forensics*, Porto, Portugal, Jun. 2019.

[9] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, and A. Alexander, “Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01),” *Speech Communication*, vol. 112, pp. 30–36, 2019.

[10] G. 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.

[11] N. Brümmer and J. D. Preez, “Application-independent evaluation of speaker detection,” *Computer Speech and Language*, vol. 20, no. 2-3, pp. 230–275, 2006.