Audio Engineering Society

Conference Paper 5

Presented at the AES 8th International Conference on Audio Forensics
2024 June 27-29, Denver, Colorado, USA

Congruence plots and their applications in forensic voice comparison

Finnian Kelly[1], Thomas Coy[1], Anil Alexander[1], and Michael Jessen[2]

[1]Oxford Wave Research Ltd., Oxford, United Kingdom
[2]Department of Text, Speech and Audio, Bundeskriminalamt, Germany

Correspondence should be addressed to Finnian Kelly (finnian@oxfordwaveresearch.com)

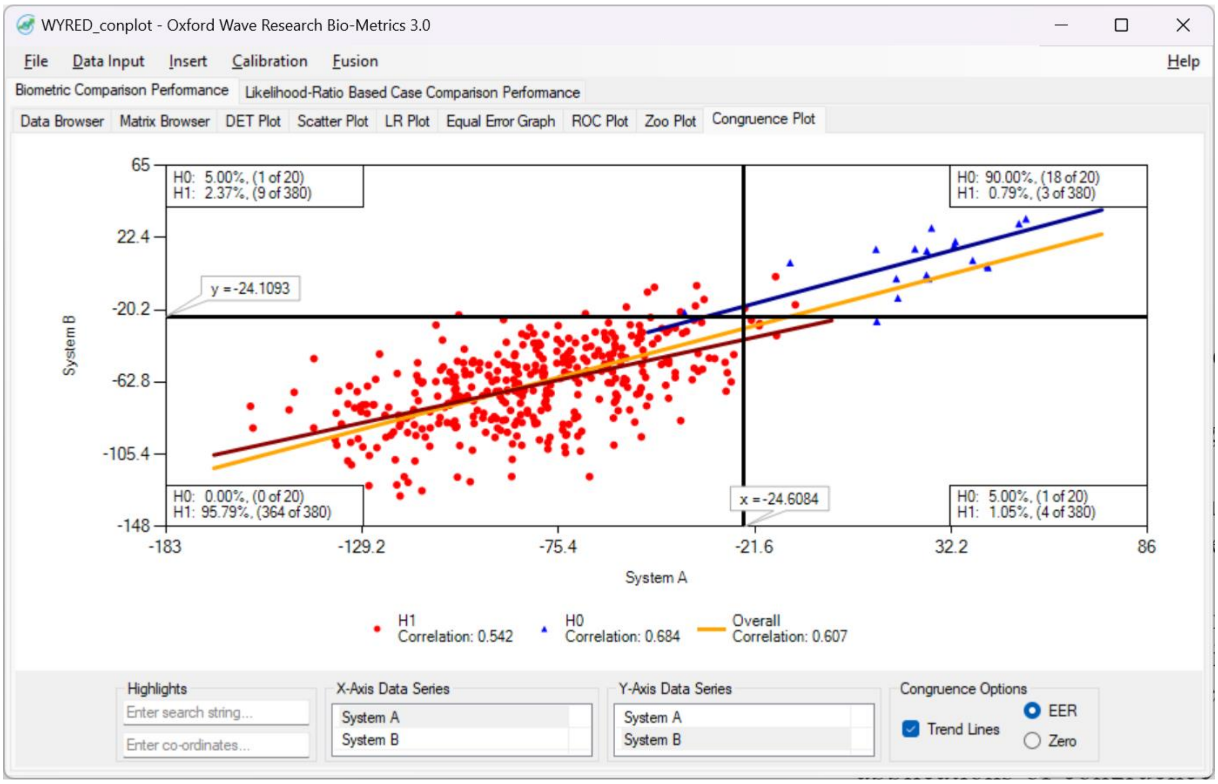
ABSTRACT

In forensic voice comparison, it is common for a practitioner to compare the same set of recordings using multiple speaker recognition systems or methods. The usual approach for comparing the results obtained from different systems or methods is to use performance metrics such as the Equal Error Rate (EER) or log-likelihood ratio cost (Cllr), and graphics like the Detection Error Tradeoff (DET) curve and Tippett Plot. While these metrics and graphics are very useful indicators of overall system discrimination and calibration, they do not indicate the level of agreement, or congruence, between the systems at the level of individual comparisons; for example, they do not inform as to whether System A and System B provide the same categorical output for an individual comparison. The congruence plot provides a visual representation of system agreement on a comparison-by-comparison basis by plotting the scores obtained for the same set of comparisons from two different systems against one another. This paper will introduce the congruence plot and its features within the Bio-Metrics software tool. The applications of congruence plots for voice comparison research and casework will then be illustrated through three forensically-relevant case studies.

1 Introduction

Traditionally, when a forensic practitioner compares the results from two speaker recognition systems as part of a research or validation exercise, or in casework, the focus is on overall performance metrics such as the Equal Error Rate (EER) or log-likelihood ratio cost (Cllr), and graphical representations like the Detection Error Tradeoff (DET) curve and Tippett Plot [1]. While such measures provide very useful indicators of overall performance, they generalise across individual speakers and sub-groups within the test data. In casework scenarios however, considering system performance at the level of individual speakers is important.

An existing approach that shifts the focus of speaker recognition performance assessment towards individual speakers is the zoo plot [2, 3]. In a zoo plot, the average same-speaker comparison score is plotted against the average different-speaker comparison score for each individual in the test set, thus showing the performance characteristics of each individual across multiple recordings within the same system. While a zoo plot successfully interrogates the individual, it is limited to the assessment of performance within a single system, which does not aid the practitioner in comparing the relative performance of individuals across multiple systems. In order to broaden the within-system system focus of the zoo plot to between-system comparison, while preserving the granularity of individual scores, we against one another. In this paper we propose the congruence plot: a graphical representation that plots the scores obtained for the present a detailed overview of the congruence plot, which was first introduced in [4], and demonstrate its features within



☑ Trend Lines ● EER ○ Zero

Figure 1: A congruence plot within Bio-Metrics software showing the scores output by two speaker recognition systems: System A (x-axis) and System B (y-axis). Same-speaker (H0) scores are indicated by blue triangles, and different-speaker (H1) scores by red circles. In this example, the axes are divided based on the EER score threshold. The H0 and H1 occupancy is indicated in each quadrant.

the Bio-Metrics[1] software tool. Some of the potential applications of congruence plots are then illustrated through several example case studies.

2 The Congruence Plot

A congruence plot (Fig. 1) displays the scores output by two speaker recognition systems[2] for the same set of comparisons, with same-speaker scores (H0) and different-speaker scores (H1) clearly differentiated. The plot is divided into four quadrants by two axis dividers. These axis dividers are positioned either at the value of zero on each axis (suitable for calibrated data such as Likelihood Ratios - LRs), or at the approximate value of the EER (Equal Error Rate) score threshold (suitable for uncalibrated data, like that in Fig. 1). The relative occupancy of each quadrant then provides a framework for assessing the degree of congruence, or agreement, between the two systems.

2.1 Congruence

Any data points appearing in the top-right or bottom-left quadrants indicate the comparisons for which the systems are *congruent*, i.e., they both support a same-speaker assessment (top-right) or a different-speaker assessment (bottom-left). For two perfectly calibrated and error-free systems, we would expect all same-speaker scores (blue triangles) to appear in the top-right quadrant, as they should be greater than zero or the EER score threshold for both systems, and all H1 scores (red circles) to appear in the bottom-left quadrant, as they should be less than zero or the EER score threshold for both systems. This behaviour would constitute 100% congruence between the systems.

[1] https://oxfordwaveresearch.com/products/bio-metrics/

[2] Congruence plots are also applicable to other biometric modalities, e.g. face recognition.

In the example in Fig. 1, it can be seen that System A and System B are largely congruent, with the majority of H0 scores appearing in the top-right quadrant and the majority of H1 scores appearing in the bottom-left quadrant. The occupancy of each quadrant, in terms of H0 and H1 scores, is indicated in the plot, providing an objective measure of congruence between the systems.

A nuance of the congruence plot is that if both systems are perfectly bad, i.e. where all same-speaker scores are in the bottom-left quadrant and all different-speaker scores are in the top-right quadrant, the two systems would still be considered 100% congruent since they agree, although incorrectly, in their categorisation of all comparisons. While this extreme scenario is unlikely to be observed, it can be expected that several congruently incorrect scores occur in situations where there is speaker- or recording condition-related variability. In Fig. 1, for example, there are 3 congruently incorrect H1 scores, i.e. red circles in the top-right quadrant. However, there are no congruently incorrect H0 scores, i.e. blue triangles in the bottom-left quadrant.

Any data points appearing in the bottom-right or top-left quadrants indicate comparisons for which the systems are *incongruent*, i.e., for which one system supports a same-speaker assessment and the other supports a different-speaker assessment. If one system is much more discriminative than the other, a relatively large number of incongruent data points can be expected.

2.2 Correlation

In addition to congruence, the correlation between the systems is also indicated on the plot. Specifically, the Spearman rank correlation coefficient is calculated for same-speaker and different-speaker scores separately, and in combination. The accompanying linear trend line is also optionally shown on the plot. In the case where the rank order of the comparisons output by both systems is very similar, the correlation coefficient will be high (as in Fig. 1), and in the case where the rank order is dissimilar, the correlation coefficient will be low.

2.3 Other features

An additional feature of the congruence plot in Bio-Metrics is the ability to highlight data points originating from an individual speaker, or group of speakers, according to a search string, in order to support a speaker-specific assessment. There is also the ability to add new data points to the plot, which can be used, for example, to show the result of a case comparison relative to those of a validation test.

3 Case Studies

Congruence plots have applications in speaker recognition research, where they can highlight individual comparisons for which systems or methods disagree, helping to diagnose speaker or condition related variables and informing of the potential for system fusion. They also have applications in forensic casework, where congruence between more-explainable and more-discriminative systems could add explainability to a case result (e.g. a likelihood ratio), in addition to informing method validation in the conditions of the case. In this section we present some of the possible applications of congruence plots through a series of speaker recognition 'case studies' involving forensically-relevant data.

3.1 Spectral vs Phonetic

In this example, we consider a scenario encountered in forensic voice comparison casework where the practitioner has two systems: System A is high in speaker discrimination but limited in explainability, and System B is lower in speaker discrimination but higher in explainability (based on phonetic theory). Specifically, an automatic speaker recognition system using spectral MFCC (Mel-frequency cepstral coefficients) features within a DNN (Deep Neural Network) x-vector framework was used as System A, and a semiautomatic system using LTF (long-term formant) features with a GMM-UBM (Gaussian Mixture Model - Universal Background Model) was used as System B. Both systems were implemented within VOCALISE software [5].

System A and System B were applied to a test set called GFS 2.0 (German Forensic Speech; [6]), which is a collection of anonymised real casework data consisting of two telephone-interception-based recordings from each of 23 male adult speakers of German. Single recordings of the same type from an additional 25 speakers were used for reference normalisation (symmetric score normalisation) with the x-vector system, and for the UBM with the GMM-UBM system.

The resulting congruence plot is presented in Fig. 2. As shown, the level of correlation between the two systems is relatively high (0.518 overall). This is expected, because both the MFCC features of the x-vector system and the formant frequencies of the LTF system are strongly influenced by vocal tract shape.

The x-vector system performs much better in terms of speaker discrimination than the LTF system. For example, if a value of 0 is used as a decision threshold (which is appropriate in this example, as the outputs of both systems are log LRs), there are many more false acceptances for the LTF system (red circles above line y=0) than the x-vector system (red circles right of line x=0). The overall performance metrics are: x-vector EER 3.49%, Cllr 0.14; LTF EER 17.53%, Cllr 0.67; fused EER 3.33%, Cllr 0.14.

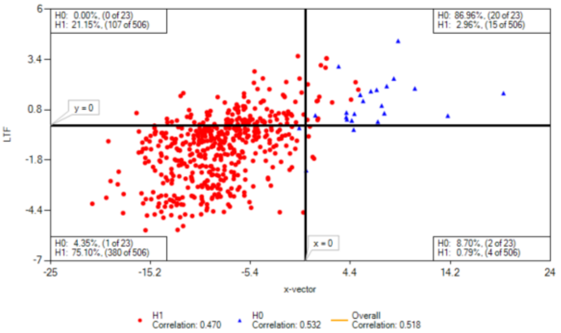


Figure 2: Congruence plot for the comparison of an x-vector system (x-axis) with an LTF system (y-axis) on the GFS 2.0 collection.

Interestingly, despite the performance difference between the systems in terms of EER and Cllr, there is a high level of congruence among the same-speaker (H0) comparisons: in only 2 of 23 comparisons is there disagreement (blue triangles in bottom-right quadrant), and in 20 comparisons both systems correctly support speaker identity. In one comparison, they both incorrectly support non-identity, but barely (blue triangle in bottom-left quadrant). Here, the observed congruence motivates the use of congruence plots to add explainability in forensic validation and casework.

3.1.1 Application of congruence plots to forensic casework

The capability of congruence plots to visualise the results from two speaker recognition systems (or methods) simultaneously has important casework implications. When validations of speaker recognition systems are performed, a common procedure is to visualise the results of the validation in the form of a Tippett plot (or Equal Error plot), along with performance indices such as EER or Cllr [1,7]. In order to evaluate the results from a case, it is useful to plot a case comparison score (an LR) into the Tippett plot (see [1: Chapter 6]) to assess where the case result occurs relative to the validation results.

With regards to validation involving multiple systems, Tippett plots are one-dimensional, allowing access to the results of only one system at a time. Congruence plots, in contrast, are two-dimensional, allowing them to be used in situations where it is meaningful to assess the results of two systems simultaneously, both in terms of validation results and case comparisons. An example of such a scenario, illustrated earlier in this Section, is the use of a maximally speaker-discriminating system (x-vector) along with a highly explanatory system (long-term formants). If the validation data and the case data are compared using both systems, the resulting LRs can be displayed on the same congruence plot, and it can be examined where the case result occurs relative to the validation.

The results in Fig. 2 show that same-speaker comparisons are mostly congruent and correct (blue triangles in top-right quadrant, where log LR > 0 for both systems). Therefore if the case result falls into the top right quadrant it not only provides evidence consistent with speaker identity from the automatic system, which in terms of discrimination is the most reliable, but there is additional confirmation pointing in the same direction from the formant system, which provides an explanatory benefit. Given the empirical results, this type of congruence among H0 comparisons is expected to occur quite often.

Considering a different scenario, if a case comparison shows a positive score for the automatic system but the formant system provides a negative result (bottom-right quadrant) to the extent that the data point from the case occurs outside of or at the margin of the cloud of data points that are typical of same-speaker comparisons, such a result would call for further scrutiny to ascertain the possible source of this incongruence.

Generally therefore, the use of congruence plots in casework can lead to endorsement (and potentially explainability) of a result when congruence occurs, and can also offer warning signs and call for further scrutiny when an incongruent situation occurs, e.g. one in which the case result occurs outside of or at the very margin of what the validation expects. What has been illustrated here with respect to x-vector and LTF systems could of course be applied to any other pair of systems for which such simultaneous analysis is meaningful.

3.2: Automatic versus Human

In this second example, we consider the scenario in which System A is a highly discriminative automatic speaker recognition system, and System B is based on human speaker recognition. Specifically, a VOCALISE MFCC x-vector system (the same system used in Section 3.1) was used as System A, and a group of lay human listeners was used as System B.

Scores for System B were drawn from an experiment presented in [8], whereby 50 human listeners were presented with pairs of short speech samples and asked to indicate whether the samples came from the same or different speakers. The aim of [8] was to explore the concept of 'voice twins', i.e., extremely similar-sounding, unrelated speakers, by presenting listeners with candidate voice twin pairings, along with 'normal' same- and different-speaker comparisons. In the present study, we use only the results from the 'normal' comparisons, which consisted of 30 same-speaker and 30 different-speaker comparisons drawn from three different datasets: WYRED [9], GBR-ENG [10], and VoxCeleb [11]. For every comparison, each individual listener confidence score was multiplied by +1 for a same-speaker judgement and -1 for a different-speaker judgement. The mean confidence score over all listeners was then taken as the output of System B.

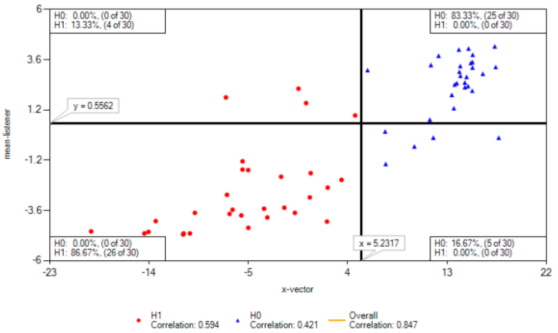


Figure 3: Congruence plot for the comparison of an x-vector system (x-axis) with a group of human listeners (y-axis) on a selection of data from [5].

The resulting congruence plot is presented in Fig. 3. The overall correlation between the systems is very high (0.847), although the H0 and H1 correlations individually are smaller (0.418 and 0.597 respectively). Such a high level of correlation is somewhat surprising, given the completely independent systems in this case (human vs machine). We note that the data in this experiment is relatively high-quality and both same- and different-speaker comparisons were randomly selected, and therefore not expected to be particularly challenging for either system.

It is evident that the discrimination of the x-vector System A exceeds that of the human listener System B: for example, if the EER threshold is used as a decision threshold (as is the case in Fig. 3), there are 5 false rejections for the human listeners (blue triangles in the bottom-right quadrant) and none for the x-vector system. There are additionally 4 false acceptances for the human listeners (red circles in the top-left quadrant), and none for the x-vector system. These results suggest that high overall correlation does not necessarily correspond with a high level of congruence. The overall performance metrics are: x-vector EER 0.00%, listener EER 10.00%, fused EER 1.67%.

The incongruent results in Fig. 3 bring to attention those comparisons for which the systems disagree; assessing these comparisons is revealing of what is most confusable for human listeners. For example, the red data points in the top-left quadrant correspond to comparisons in which the two samples are from perceptually similar-sounding speakers, and blue data points in the bottom-right quadrant correspond to comparisons in which the two samples are from the same speaker, but there is some variability present. In this case, there are no errors by the x-vector system, but if they were to occur, the corresponding human listener result may be informative as to the source of the error.

Overall, the fact that the two systems are largely congruent (83% for H0, 87% for H1) indicates that the x-vector system is largely in agreement with the perceptual judgements of human listeners.

3.3 x-vector vs ECAPA-TDNN

In this final example we consider the scenario in which both System A and System B are highly discriminative automatic speaker recognition systems with different DNN architectures. Specifically, a VOCALISE MFCC x-vector system (the same system as used in Sections 3.1 and 3.2, with additional condition adaptation, discussed below) was used as System A, and an ECAPA-TDNN system [12] (trained with the same data recipe as the VOCALISE x-vector system) was used as System B.

For this example we considered a forensically-relevant data collection, namely *forensic_eval_01*

[13], which contains 432 recordings from 166 speakers. There are two distinct conditions represented within the collection, a simulated police interview and a telephone call. The collection is divided into a training set of 105 speakers and a test set of 61 speakers. Here, we applied VOCALISE condition-adaptation [5] to the x-vector system with the training set, and then tested both systems with the test set, following the protocol specified in [13].

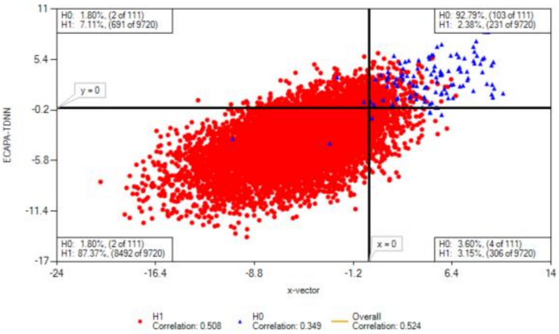


Figure 4: Congruence plot for the comparison of an x-vector system (X-axis) with an ECAPA-TDNN system (Y-axis) on the *forensic_eval_01* test dataset (official protocol).

In Figure 4 the resulting congruence plot is presented. There is a relatively high overall correlation of 0.524; this could be expected, given that the acoustic features and modelling approaches have shared characteristics across the systems, in addition to common model training data. There is a difference in discrimination however; if we take 0 as a decision threshold (as in Fig. 4) then it can be seen that the x-vector system makes slightly fewer H0 errors than the ECAPA-TDNN system: 4 H0 points to the left of the line x=0, compared with 6 H0 points below the line y=0 (note the H0 occupancy counts in each quadrant; some of the points are difficult to see in the plot). A similar pattern can be observed for the H1 points. The overall performance metrics are: x-vector EER 4.56%, Cllr 0.25; ECAPA-TDNN EER 7.46%, Cllr 0.32; fused EER 4.10%, Cllr 0.22. We note that both the individual x-vector result and the fused result exceed the performance of a previous generation VOCALISE x-vector model [14].

Referring to the performance metrics, it is evident that there is an improvement in both discrimination and calibration (decreased EER and Cllr) after fusing the two systems (via linear logistic regression applied with a leave-one-out cross-validation procedure). Referring to the congruence plot in Fig. 4, it can be seen that most of the H0 errors exist in the two quadrants of incongruence: there are a total of 6 H0 errors across top-left and bottom-right quadrants, and only 2 H0 errors in the bottom-left quadrant (again, note the H0 occupancy counts in each quadrant). The same pattern exists for the H1 errors. We suggest that this incongruence, along with the good individual discrimination performance of each individual system, demonstrates that the systems are complementary. This observation is borne out by the improvement in performance observed with fusion.

The example in Fig. 4 also highlights the use of congruence plots as a diagnostic tool: considering the two H0 points in the bottom-left quadrant, upon analysis of the recordings involved in these comparisons, it is evident that they contain poor quality audio with interfering speakers (rightmost sample), and very strong speaker-related variability (leftmost sample).

Finally, we demonstrate the capability of congruence plots to assess individual speaker performance. In Fig. 5, the comparison scores involving a specific example speaker are highlighted in the plot. The H0 and H1 occupancy statistics and correlation coefficients are updated accordingly. This provides a means to assess the performance of this individual speaker relative to that of the larger set; it is evident that for the example speaker in Fig. 5, their H0 and H1 scores fall within the typical expected range for each system.

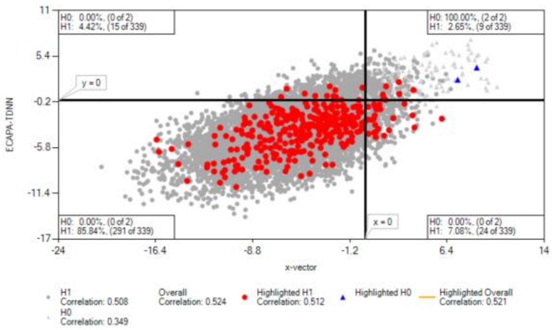


Figure 5: Congruence plot as shown in Fig. 4, with the additional highlighting of the comparisons involving a specific speaker.

4 Discussion and Conclusion

This paper has presented the congruence plot as a means to compare the output of two speaker recognition systems at the level of individual comparisons, which provides valuable insight for forensic voice comparison research and casework.

Using Bio-Metrics software, some of the applications of congruence plots were explored via three case studies involving forensically-relevant data, which demonstrated the potential of congruence plots to contribute to validation and explainability in casework, and to serve as a research tool that can inform of speaker or condition related variability and of the fusion potential between two systems.

References

[1] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, T. Niemi, “Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition”, *Frankfurt: Verlag für Polizeiwissenschaft* (2015).

[2] A. Alexander, O. Forth, J. Nash, N. Yager, “Zooplots for Speaker Recognition with Tall and Fat Animals”, *International Association for Forensic Phonetics and Acoustics (IAFPA) conference 2014*, Zürich, Switzerland (2014).

[3] G. T. Dunstone, N. Yager, “Biometric System and Data Analysis”, *New York: Springer* (2009).

[4] M. Jessen, A. Alexander, T. Coy, O. Forth, F. Kelly, “CON(gruence)-plots for assessing agreement between voice comparison systems”, *International Association for Forensic Phonetics and Acoustics (IAFPA) conference 2023*, pp. 37–38, Zürich, Switzerland (2023).

[5] F. Kelly, O. Forth, S. Kent, L. Gerlach, A. Alexander, “Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors”, *Audio Engineering Society (AES) Forensics Conference 2019*, Porto, Portugal (2019).

[6] Y. A. Solewicz, M. Jessen, D. van der Vloed, “Null-Hypothesis LLR: A proposal for forensic automatic speaker recognition”, *INTERSPEECH 2017*, Stockholm, pp. 2849–2853 (2017).

[7] G. S. Morrison et al., “Consensus on validation of forensic voice comparison”, *Science and Justice,* vol. 61, pp. 299–309 (2021).

[8] L. Gerlach, K. McDougall, F. Kelly, A. Alexander, “Voice twins: discovering extremely similar-sounding, unrelated speakers”, *INTERSPEECH 2023*, Dublin, Ireland (2023).

[9] E. Gold, S. Ross, K. Earnshaw, “The West Yorkshire Regional English Database: investigations into the generalizability of reference populations for forensic speaker comparison casework”, *INTERSPEECH 2018*, Hyderabad, India (2018).

[10] GBR-ENG database, “A telephonic speech database collected for the UK government for evaluating speech technologies”. Further details on application (2019).

[11] A. Nagrani, J. S. Chung, W. Xie, A. Zisserman, “Voxceleb: Large-scale speaker verification in the wild”, *Computer Speech & Language,* vol. 60 (2020).

[12] B. Desplanques, J. Thienpondt, K. Demuynck, “ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification”, *INTERSPEECH 2020*, Shanghai, China (2020).

[13] G. Morrison, E. Enzinger, “Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – Introduction”, *Speech Communication,* vol. 85, pp. 119–126 (2016).

[14] F. Kelly, A. Fröhlich, V. Dellwo, O. Forth, S. Kent, 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).