Promega

More Cases Solved

More meaningful results from your challenging samples with 8-color PowerPlex STR kits, such as PowerPlex® 35GY, and Spectrum CE. More capacity with convenient 4-plate drawer. More scheduling flexibility with continuous access to the plate drawer. The new Spectrum CE System gives you more of everything you need to power through your casework analyses.

Learn More »

www.promega.com/SpectrumCE

Received: 29 September 2022 | Revised: 8 March 2023 | Accepted: 20 March 2023

DOI: 10.1111/1556-4029.15250

JOURNAL OF FORENSIC SCIENCES AAFS

ORIGINAL PAPER

Digital & Multimedia Sciences

Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings

Dávid Sztahó PhD[1] | Attila Fejes MSc[2]

[1]Faculty of Electrical Engineering and Informatics, Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Magyar tudósok körútja 2, Budapest, 1117, Hungary

[2]Doctoral School of Law Enforcement, Hungarian National University of Public Service, H-1083 Budapest, 2 Ludovika tér, Budapest, H-1441, Hungary

Correspondence
Dávid Sztahó PhD, Budapest University of Technology and Economics, Faculty of Electrical Engineering and Informatics, Department of Telecommunication and Media Informatics, Magyar tudósok körútja 2, Budapest 1117, Hungary.
Email: sztaho.david@vik.bme.hu

Funding information
Nemzeti Kutatási Fejlesztési és Innovációs Hivatal, Grant/Award Number: FK128615

Abstract

In forensic voice comparison, deep learning has become widely popular recently. It is mainly used to learn speaker representations, called embeddings or embedding vectors. Speaker embeddings are often trained using corpora mostly containing widely spoken languages. Thus, language dependency is an important factor in automatic forensic voice comparison, especially when the target language is linguistically very different from that the model is trained on. In the case of a low-resource language, developing a corpus for forensic purposes containing enough speakers to train deep learning models is costly. This study aims to investigate whether a model pre-trained on multilingual (mostly English) corpus can be used on a target low-resource language (here, Hungarian), not represented by the model. Often multiple samples are not available from the offender (unknown speaker). Samples are therefore compared pairwise with and without speaker enrollment for suspect (known) speakers. Two corpora are used that were developed especially for forensic purposes and a third that is meant for traditional speaker verification. Speaker embedding vectors are extracted by the x-vector and ECAPA-TDNN techniques. Speaker verification was evaluated in the likelihood-ratio framework. A comparison is made between the language combinations (modeling, LR calibration, and evaluation). The results were evaluated by $\mathrm{Cllr}_{\mathrm{min}}$ and EER metrics. It was found that the model pre-trained on a different language but on a corpus with a significant number of speakers can be used on samples with language mismatch. Sample duration and speaking style also seem to affect the performance.

KEYWORDS
AusEng, ECAPA, forensic voice comparison, ForVoice120, language dependency, speaker verification, speaking style, VoxCeleb, x-vector

Highlights

- Pre-trained speaker recognition models are evaluated in forensic voice comparison scenario.
- Evaluation was done on datasets created especially for forensic purposes.
- Language mismatch between training and forensic datasets is not found using deep learning models.
- Performance increase was found using multiple samples from the known speaker.
- Sample duration and speaking style mismatch was found to affect the performance.

1 | INTRODUCTION

The fields of speaker identification (SI) and speaker verification (SV) (together: speaker recognition) have been studied for a long time, and the literature is still growing [1]. Several studies have been conducted in this field of research, resulting in a number of techniques for both topics. Over time, state-of-the-art technologies are constantly changing as new ideas emerge. Until recently, i-vectors were considered the state-of-the-art technique in speaker recognition [2], but with the advent of deep learning methods and the emergence of large speech corpora, novel classification and feature extraction methods have been developed (e.g., d-vectors [3], j-vectors [4], x-vectors [5], and ECAPA-TDNN networks [6]).

In speaker identification, the task is to identify an unknown speaker from a set of already known speakers. The closed set (or in-set) scenario is when all speakers within a given set are known. On the other hand, we talk about open set (or out-of-set) speaker identification when the set of known speakers does not contain potential test subjects [7].

In speaker verification, we verify that the speaker is who he/she says he/she is by comparing two (or more) speech samples/utterances and evaluating whether the speakers in the two samples are the same [7]. This is traditionally done, in general forensic voice comparison practice, by comparing the test sample or samples with the given speaker's sample or samples and a universal background model [8]. Another way to compare whether a pair of speakers is of the same origin is to classify the pairs as 'same' or 'different' and create a model accordingly. This is feasible for sample-by-sample comparisons, as a model can be trained on a dataset of sample pairs to predict whether speakers are identical, rather than using a UBM. From this definition, it follows that technically forensic speaker comparison is part of the speaker verification scheme, although the 'known' speaker in this case is not specifically known. A voice sample can be associated with a hypothesized speaker. The aim is to verify that the identity of this speaker (suspect) matches the identity of another unknown speaker (offender). Often the purpose of forensic voice comparison is also to verify whether the identities of two unknown speakers match. In practice, this verification is carried out using the same method.

A paradigm shift is taking place in forensic science and practice [9, 10], which allows for automatic and semi-automatic evaluation of evidence using different methods and measurement types (e.g., DNA, fingerprinting) [11, 12]. This new paradigm, the so-called likelihood ratio (LR) framework, supports a processing pipeline that can be easily computed for multiple types of evidence. Forensic voice comparison, where speaker recognition techniques are adapted to the requirements of the framework, is an area where this new paradigm can be applied. For a piece of evidence, it produces a ratio of the likelihood or probability density (at a given point) of same and different speakers [13–15].

Considering a forensic voice comparison system, we can evaluate comparisons of same and different speaker origins in two ways [16]: (1) there are multiple samples available from unknown and known speakers, and (2) samples can be compared pairwise. It naturally follows that the first scheme can achieve higher accuracy. However, there are many cases where multiple samples are not available for comparison (only a single voice recording fragment is available). Several studies have been conducted using short utterances [17–20], but these generally do not meet the requirements of forensic evaluation: the evaluation datasets do not follow a strict protocol [21] or use techniques that have already been outperformed by deep learning techniques in regular speaker recognition. This study aims to investigate this scenario in two ways: only one sample is available for the unknown speaker and (i) only one or (ii) multiple samples are available for the known speaker. An example of (i) is when one speech sample is available for both the offender and the suspect, and (ii) is when one sample is available for the offender, but multiple samples can be recorded for the suspect. Therefore, in this paper, we focus on the pairwise comparison of samples.

In the LR framework of forensic voice comparison, the likelihood of speech evidence is calculated according to two competing hypotheses, e.g., (1) "What is the possibility that the sample in question originates from the suspect?" and (2) "What is the possibility that the sample in question originates from someone else?". The ratio of these expressions expresses the strength of the evidence (Equation 1). LR is the likelihood-ratio, E is the evidence, H_{so} is the hypothesis of same-origin speakers, and H_{do} is the hypothesis of different-origin speakers.

$$LR = \frac{P(E|\,H_{so})}{P(E|\,H_{do})} \qquad (1)$$

Several large-scale corpora are available for speaker recognition [22–24], and the NIST speaker recognition challenge [25] is also often held. However, for a forensic voice comparison system, there are specific needs [21] that are not satisfied by these corpora. This paper uses a speech dataset developed for forensic expert purposes to evaluate speaker verification systems and analyze their sensitivity to sample length and speech style. The goal here is to compare the performance of state-of-the-art deep learning feature extraction models pretrained on a large dataset with models trained in a low-resource language and evaluate them in the low-resource language. This demonstrates the usefulness of pre-trained models with language mismatches (between the trained model and the test samples) for forensic voice comparisons for institutions such as public services in countries where sufficient speech data cannot be collected to adequately train a deep learning model.

Feature extraction methods based on deep learning (such as TDNN architectures) have shown better performance than previous approaches (GMM-UBM, i-vector). However, these techniques require large amounts of training data to produce suitable models. Low-resource languages do not have the data needed to train such models. Thus, the possibility of using pre-trained models for speaker recognition in smaller languages naturally arises. Kleynhans and Bernard [26] found a language-dependent tendency, but they used an outdated technology for speaker verification. However, the language dependence of deep learning feature extraction methods (such as the one used in this study) may be negligible due to the large amount of data on which they are trained. Even if the samples are from a single language, the deep learning model may be able to extract information robust enough to use an accurate speaker representation in another language. There have been studies that have investigated this cross-linguistic pattern, but none of them fit the framework of forensic voice comparison. Li et al. [27] used synthesized speech for evaluation, and the comparison was pairwise. In their study, Chojnacka and colleagues [28] conducted multilingual experiments with a multilingual training dataset, which is not the main practice we want to investigate. The language dependency of the i-vector technique has already been investigated [29, 30], but newer deep learning techniques have outperformed the older i-vector technique. Fabien and Motlicek [31] investigated the performance of x-vector models in forensic scenarios, but with acted speech, and the study did not focus on the effect of multilingualism (although the dataset was multilingual). The study by Skarnitzl and colleagues [32] is specifically related to forensic research and evaluated multilingual scenarios but uses an earlier version of the VOCALIZE [33] system based on the obsolete i-vector.

There are certain factors in a speech material that can affect the effectiveness of voice comparison. It may be important if differences in speaking style and sample duration impair performance. Few studies have focused on whether these factors actually matter [34–36], although if they do, it may bias the evaluation of the evidence. In the present study, we use the three speech styles available in the dataset developed for forensic claims and compare the results depending on sample duration.

In this study, we investigate (1) how a pre-trained deep learning speaker embedding model performs in a low-resource language that is not the same as the one in which the model was trained; (2) whether and how much performance gains are obtained when more samples are available from the known speaker (suspect); (3) how performance metrics depend on sample length; and (4) how performance metrics depend on the speech style (available in the dataset). The results may be helpful to forensic services or institutes planning forensic voice comparisons.

The structure of the paper is as follows: the methods, datasets, evaluation metrics, and scenarios used in the study are described in the next section. This is followed by a presentation of the results. Then, an overview of the resulting evaluations is given, with brief concluding reflections.

2 | EMBEDDING MODELS

In this study, two techniques were used to extract embedding vectors as features from speech samples: the x-vector and ECAPA-TDNN. These methods take a speech sample as input and output a vector that can be imagined as a vector representation of the speaker in the sample. The deep learning-based feature extraction models were applied using the SpeechBrain toolkit [37]. In this paper, we use three embedding models to evaluate cross-lingual speaker verification schemes. To compare pre-trained models on the VoxCeleb dataset (details in the Datasets section), the parameters of the custom-trained models closely followed the method of these pre-trained models. The method includes the input sound file format, extracted features, network structures, and data augmentation. These are detailed in the following subsections and in the Methods section. The pretrained models were downloaded from Huggingface [38, 39].

2.1 | The x-vector

A deep learning-based feature extraction method called x-vector was developed primarily for speaker verification [5]. It is based on a multilayer DNN architecture (with fully connected layers), with a different temporal context (which they call “frames”) in each layer. Because of the wider temporal context, the architecture is called time-delay NN (TDNN). The TDNN embedding architecture is shown in Figure 1 and Table 1.

The first five layers work on speech frames with a small temporal context centered on the current frame *t*. For example, the frame indexed as 3 sees a total of 15 frames, due to the temporal context of the previous layers. After training with the speaker identifiers used

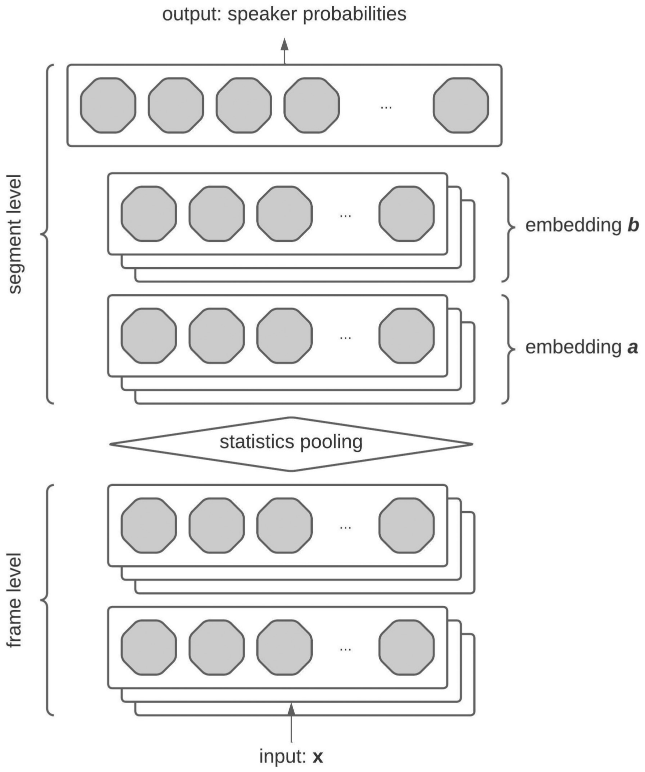

FIGURE 1 The x-vector DNN embedding architecture in [5]. The two parts: frame level (with the five frame layers) and segment level (with segment 6, segment 7, and softmax).

TABLE 1 The x-vector DNN layer architecture [5]. It contains the layers, contexts, and the input–output dimensions.

Layer	Layer context	Total context	Input × output
Frame1	[$t-2$, $t+2$]	5	120 × 512
Frame2	{$t-2$, t, $t+2$}	9	1536 × 512
Frame3	{$t-3$, t, $t+3$}	15	1536 × 512
Frame4	{t}	15	512 × 512
Frame5	{t}	15	512 × 1500
Stats pooling	[0, T}	T	1500T × 3000
Segment6	{0}	T	3000 × 512
Segment7	{0}	T	512 × 512
Softmax	{0}	T	512 × N

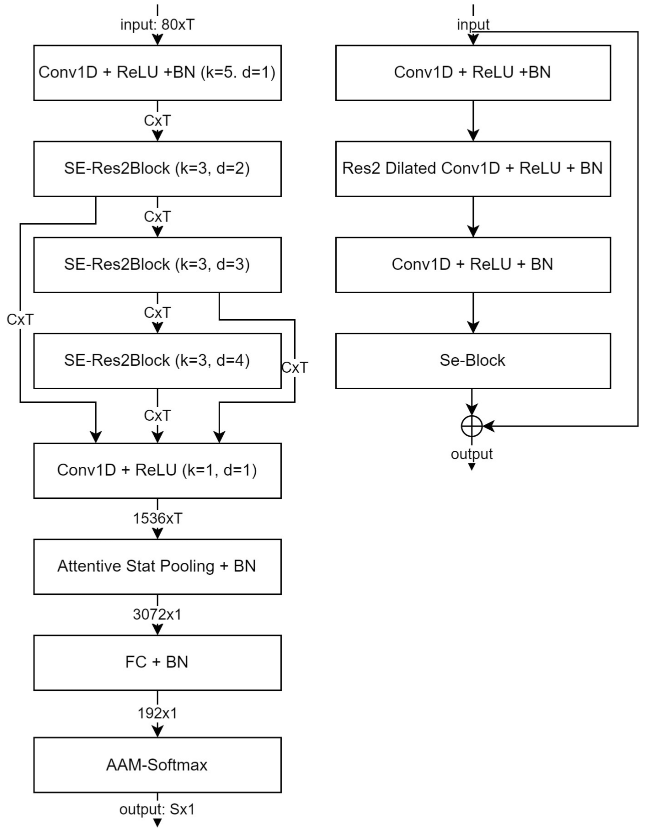

FIGURE 2 The ECAPA-TDNN layer architecture and its SE-Res2Block (taken from [6]).

as the target vector, the output of the segment6 ("x-vector") can be used as an embedding vector.

2.2 | ECAPA-TDNN

The ECAPA-TDNN model is the extension of the x-vector model architecture in three ways [6]: channel- and context-dependent statistics pooling, 1-Dimensional Squeeze-Excitation Res2Blocks (1D SE-Res2Block) and multi-layer feature aggregation and summation. The channel- and context-dependent statistics pooling enables the network to focus more on speaker characteristics that are not activated at the same or similar time instants, e.g., speaker-specific features of vowels versus speaker-specific features of consonants. Using the SE-Res2Block (taken from the field of computer vision), the limited frame context of the x-vector (15) is extended to global properties of the recording. This enables the network to see a larger context than the original x-vector architecture by applying 1D convolution layers. The multi-layer feature aggregation means that not only the activation of the selected distinguished deep layer is used as a feature map (as in x-vector), but the shallower layers (here: SE-Res2Blocks) are also concatenated, because they hold additional information that helps forming speaker vectors that may be lost in the deeper layers, so they also hold information about the speaker identity. The architecture is shown in Figure 2. For detailed information on the structure and its baseline evaluation, see [6].

3 | METHODS

3.1 | Embedding models

3.1.1 | The x-vector

The dimension of the x-vectors was set to 512, and the input was 24 mel-frequency band energies. The training was done for 35 epochs with early stopping for which the criterion was the minimum loss measured on a validation set. The training was done with Adam optimizer with a starting learning rate of 0.001. The x-vector model pre-trained on the VoxCeleb dataset was downloaded from Huggingface, and all custom-trained models followed the same input format and network structure. All samples were resampled to 16 kHz before feeding them to the network.

3.1.2 | ECAPA-TDNN

Following the method of the model pre-trained on the VoxCeleb dataset, the dimension of the extracted embedding vector in the case of custom-trained models was 192, and the input was 80 mel-frequency band energies. The training was done for 35 epochs with early stopping for which loss was measured on a validation set. The training was done with Adam optimizer with a starting learning rate of 0.001. All samples were resampled to 16 kHz before feeding them to the network.

3.2 | Cosine distance and enrollment

The cosine distance was used to evaluate the similarity of the embedding vectors extracted from the sample pairs. Cosine distance was chosen because it enables comparison of single samples, unlike the commonly used PLDA, which needs multiple enrollment samples for a speaker. The cosine distance is commonly used in speaker verification measurements. It is simply the calculation of the normalized dot product of target and test vectors (w_{target} and w_{test}), which gives a match score (Equation 2).

$$CDS(w_{\text{target}}, w_{\text{test}}) = \frac{w_{\text{target}} \cdot w_{\text{test}}}{\| w_{\text{target}} \| \cdot \| w_{\text{test}} \|} \tag{2}$$

If multiple samples are available for a known speaker, it is commonly advantageous to create an enrollment vector or model from these. In the present study, due to the use of cosine distance, the embedding vectors were averaged per speaker to get a mean embedding vector for each speaker. In the results, we compared the performance with and without using speaker enrollment. Speaker enrollment was done by averaging the embedded vectors per speaker on the session 1 samples (known speakers). The average vectors were then compared to all of the session 2 samples (unknown speakers).

3.3 | LR score calculation

To calculate LR scores, logistic regression was used (implemented with the Python sklearn package). The cosine distances were calculated for sample pairs and arranged according to the same speaker and different speaker labels, used for training logistic regression models. The output of the logistic regression model is the probability of the same speaker decision. Since H_{so} and H_{do} in Equation 1 are mutually exclusive and exhaustive events, after weighing input classes so that $P(H_{so}) = P(H_{do})$, $P(E|H_{do})$ can be alculated as $1 - P(E|H_{so})$ [40, 41]. This enables the calculation of LR in Equation 1. Figure 3 shows an example of a trained logistic regression model. Distributions of same and different origin vector pairs are shown in blue and yellow, respectively. The figure shows the $P(E|H_{so})$ probability. This implies that the LR score is 1 at the intersection of the distributions, because $P(E|H_{so}) = P(E|H_{do})$.

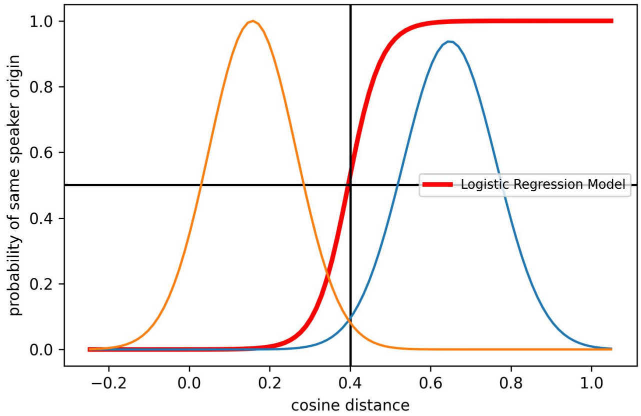

FIGURE 3 A trained logistic regression model example. Blue and yellow lines show the distributions of cosine distances of embedding vector pairs of same and different speaker origin, respectively.

3.4 | Datasets

Forensic voice comparison performance measures were evaluated on two forensic datasets (multiple speaking styles per speaker, multiple recording sessions per speaker): the Hungarian ForVoice120+ corpus and the Australian English AusEng [42] dataset. These followed the protocol specified by Morrison et al. [21]. Both datasets contain multiple speech tasks per speaker each modeling a different speaking style and (at least) two recording sessions per speaker with at least 2 weeks' delay. See Table 2 for a description of the datasets. The ForVoice120+ dataset contains 120 speakers, representing a low-resource language dataset. The AusEng corpus contains more than 500 speakers. Figure 4 shows the age distribution according to ranges defined in the AusEng dataset. As the figure shows, the ForVoice120+ used for evaluation mainly represents the 18–35 age range. However, this does not affect the statements that can be derived according to language dependency. Both corpora contain three speech tasks: free dialogue, information exchange, and monologue (simulating interrogation). The datasets were split into multiple parts. 40 speakers of the ForVoice120+ were used for LR calibration, and the remaining 80 speakers were used for evaluation. Since ForVoice120+ contains only a limited number of speakers, the Hungarian speaker embedding x-vector and ECAPA-TDNN models were trained on different samples, not explicitly made for forensic purposes: BEA [43], MRBA [44] and newly recorded samples with read text and free speech. A total of 632 speakers were used for the training; the total duration of the speech was 27.31 h. For the AusEng dataset, 395 speakers were randomly selected for embedding model training, 80 speakers for LR calibration, and 80 for evaluation.

In addition to the embedding models trained on the Hungarian dataset and the AusEng corpus, we also used pre-trained models on the VoxCeleb2 [23] corpus (available in the Huggingface repository) to extract embedding feature vectors, as the corpus contains more than 6000 speakers and represents the largest available model for speaker recognition. The two large-scale datasets represent a language that is commonly available and has many resources. Available details on the VoxCeleb dataset are also shown in Table 2. The dataset contains materials collected from YouTube. Exact information on languages included in the VoxCeleb is not available. The creators claim that it is a multilingual dataset, but only limited nationality information is made available for the dataset, not the language spoken on the original YouTube videos. It mainly represents widely spoken languages, and therefore the target low-resource language (Hungarian) is not represented. Age distribution is also not available. Hechmi et al. [45] created an enrichment for the VoxCeleb in which age information is included as well. Figure 4 shows the distribution of age by the same ranges used in AusEng based on this enrichment. The datasets used to

TABLE 2 Metadata for the For Voice120+ and AusEng datasets.

Dataset	Total number of recordings	Number of recording sessions	Number of speakers	(male/female)	Total speech length	Number of recordings per speaker
ForVoice120+	720	2	120	59/61	~32 h	6
AusEng	3899	2 or 3	555	239/316	~311 h	3–9
VoxCeleb	1,128,246	not controlled	6112	3728/2384	~2442 h	not controlled

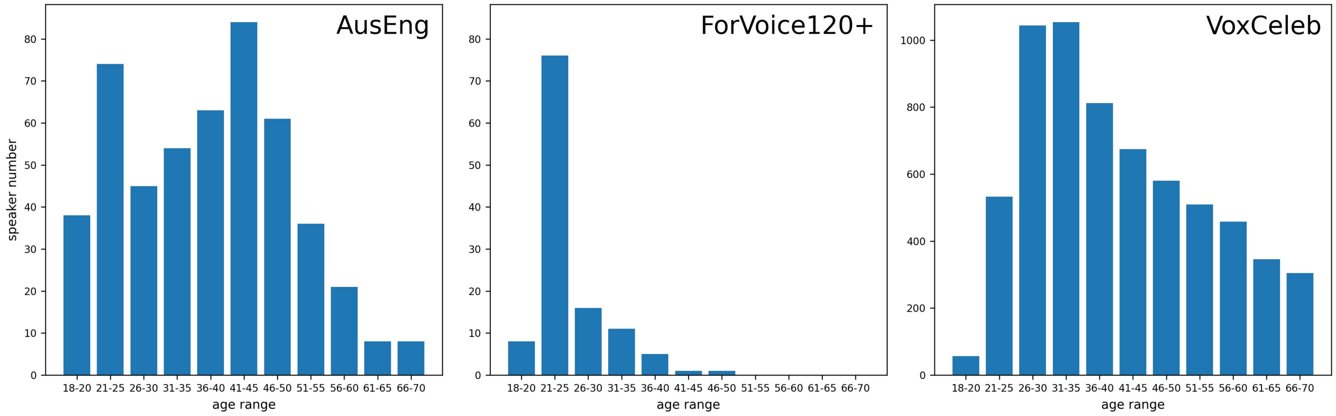

FIGURE 4 Distributions of speaker age according to ranges defined in the AusEng dataset.

train the embedding models were split into training, validation, and test sets in a ratio of 60–20-20%. The early stopping criteria were measured on the validation set and used to check for overfitting on the test set (comparing the results on the test set and the validation set).

3.4.1 | Database splitting

The recordings of the dataset used for Hungarian embedding model training, ForVoice120+ and the AusEng datasets were split into multiple parts with various lengths. The possible duration of a part was {2,3,4,5,6,7,8,9,10} s. The number of samples was (almost) evenly distributed according to the durations. First, all silence parts were removed from the recordings, and then splitting was done with 10% overlap between adjacent parts. There was no influence on the sample lengths of the VoxCeleb2 dataset because pre-trained models were used in those cases. The final distributions of sample durations are shown in Figure 5. The figure also shows the exact number of samples per durations used.

3.4.2 | Augmentation

Following the method of the pre-trained models available on Huggingface, data augmentation was applied to the Hungarian samples and the AusEng dataset during model training: the samples were augmented with every combination of time-distorted (duration was scaled with factors 0.95 and 1.05) and noise-distorted (with 15 dB white noise) variants. To compare the newly trained models with these previously trained models, the same augmentation was applied across all datasets used in the study. We do not have any control on the pre-trained models in this regard, but we have used the same method for training our custom models as described by their creators. According to the results reported on the pre-trained models, this augmentation increases the robustness of the models.

3.5 | Evaluation metrics

The outputs of the different model configurations were evaluated in terms of equal error rate (EER) of speaker verification (EER is the level at which false acceptance rate and false rejection rate are equal, commonly used in biometric security systems) and log-likelihood-ratio cost (Cllr, Equation 3) [46] defined as

$$\mathrm{Cllr} = \frac{1}{2}\left(\frac{1}{N_{so}}\sum_{i=1}^{N_{so}}\left(1+\frac{1}{LR_{so_i}}\right)+\frac{1}{N_{do}}\sum_{j=1}^{N_{do}}\left(1+LR_{do_j}\right)\right) \quad (3)$$

where N_{so} and N_{do} are the number of same-origin and different-origin comparisons and LR_{so} and LR_{do} are the likelihood ratios derived from same-origin and different-origin comparisons. Cllr is a function that measures the balance of *LR* scores of same-origin and different-origin comparisons. Ideal same-origin and different-origin comparisons have log LR >0 and log LR <0, respectively. Incorrect comparisons (which are not as ideal as the inequalities mentioned above) result in higher Cllr. The better the performance of a forensic comparison system, the more correct *LR* values it produces, the lower Cllr it achieves, supplying the evidence magnitude. In addition to Cllr, the minimum Cllr is also reported, which is the generalization of the original cost function and

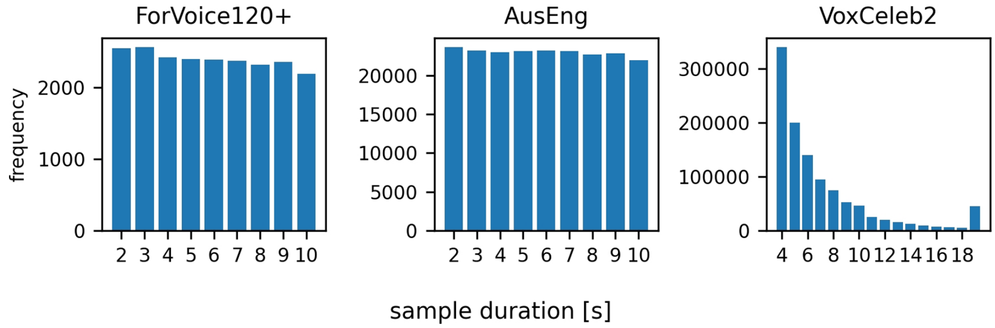

FIGURE 5 Distributions of sample durations in the datasets used for the study.

produces application-independent Cllr values by optimizing the decision threshold [47]. While Cllr is a measure of both discrimination and calibration, the calibrated Cllr has any calibration mismatch optimized away, it is a now pure measure of discrimination. The EER is a concise summary of the discrimination capability of the detector. As such, it is a very powerful indicator of the discrimination ability of the detector, across a wide range of applications. However, it does not measure calibration, the ability to set good decision thresholds [46]. Results are also displayed on Tippett plots, which show the proportion of correctly identified same and different speaker origin pairs (a visualization often used in forensic comparison).

3.6 | Evaluation scenarios

Multiple phenomena were investigated:

- language mismatch,
- speaker enrollment,
- sample duration mismatch and
- speech task mismatch.

Datasets containing different languages were used to train embedding models, calibrate LR scores, and evaluate speaker verification, all without and with enrollment. The best performing scenario was broken down into utterance durations and speech tasks to see their effect on forensic voice comparison. Training of embedding models and calibration of LR scores were performed on samples from all sessions. The evaluation compared samples from different sessions: session 1 samples were used as known speaker samples and session 2 as unknown speaker samples. Sessions were recorded with a delay of at least 2 weeks.

For the best performing dataset combination, the results are broken down by sample durations and speech tasks. The Cllr_{min} and EER values (without enrollment) are organized into matrices of the examined phenomena. For sample durations, rows and columns show results calculated by durations from 2 to 10 s. A value of a cell was calculated by filtering the sample-pair comparisons according to the known and unknown speaker sample lengths. For speech tasks, the rows and columns of the matrices contain the task numbers (1: free dialogue, 2: information exchange, 3: monologue) and the cells contain the results of the respective task pair. Using speaker enrollment, multiple sample lengths and speech tasks were not applicable for the known speakers, as the speaker vectors were averaged over all samples in session 1. The results are therefore vectors in this case. However, this does not pose a problem because it represents the real-life situation where multiple recordings can be obtained from a suspect (known speaker) and the enrollment can be performed.

4 | RESULTS

4.1 | Effect of languages used for model training and LR calibration

Table 3 shows the results without speaker enrollment using different dataset combinations. The ECAPA-TDNN models outperformed the x-vector in all cases. Significant decreases are observed in all metrics. The best performing combinations were obtained when the VoxCeleb dataset was used to train the embedding models (pre-trained models): 3.1% EER and 0.122 Cllr_{min} values for ECAPA-TDNN. This is a good value compared to the state-of-the-art results on short utterance comparisons. The LR calibration set did not make any difference. The language difference also did not reduce the performance when comparing the evaluation sets. We even found slightly higher metric values using the ForVoice120+ than using the AusEng dataset (same language although different dialect). The Tippett plot of the best performing case is shown in Figure 6.

4.2 | Effect of enrollment

The same dataset combinations were repeated using speaker enrollment. The results are shown in Table 4. The main tendencies (differences in embedding vector technique, dataset used for embedding models, evaluation datasets) are the same as before. The ECAPA-TDNN outperformed the x-vector in this case as well. The language differences did not cause performance degradation. Again, the pre-trained models available on the VoxCeleb dataset performed best. The lowest EER is 1% with a Cllr_{min} of 0.045. The Tippett plot for this case is shown in Figure 7.

TABLE 3 Speaker verification results obtained with models of different dataset combinations.

Ealuation language/ dataset	Embedding model language/dataset	LR calibration language/dataset	Model	Cllr	$\mathrm{Cllr}_{\mathrm{min}}$	$\mathrm{Cllr}_{\mathrm{cal}}$	EER
Hungarian/ForVoice120	Hungarian	Hungarian/ForVoice120	x-vector	0.632	0.601	0.031	0.189
			ECAPA-TDNN	0.405	0.401	0.005	0.116
	English/AusEng	Hungarian/ForVoice120	x-vector	0.567	0.537	0.031	0.167
			ECAPA-TDNN	0.253	0.249	0.004	0.069
	English/AusEng	English/AusEng	x-vector	0.581	0.537	0.044	0.167
			ECAPA-TDNN	0.629	0.249	0.380	0.069
	VoxCeleb	**Hungarian/ ForVoice120**	x-vector	0.365	0.349	0.016	0.102
			ECAPA-TDNN	**0.127**	**0.122**	**0.005**	**0.031**
	VoxCeleb	**English/AusEng**	x-vector	0.381	0.349	0.032	0.102
			ECAPA-TDNN	**0.168**	**0.122**	**0.046**	**0.031**
English/AusEng	English/AusEng	English/AusEng	x-vector	0.616	0.575	0.041	0.183
			ECAPA-TDNN	0.184	0.182	0.001	0.048
	VoxCeleb	English/AusEng	x-vector	0.511	0.481	0.029	0.150
			ECAPA-TDNN	0.220	0.206	0.014	0.053

The best results are highlighted in bold.

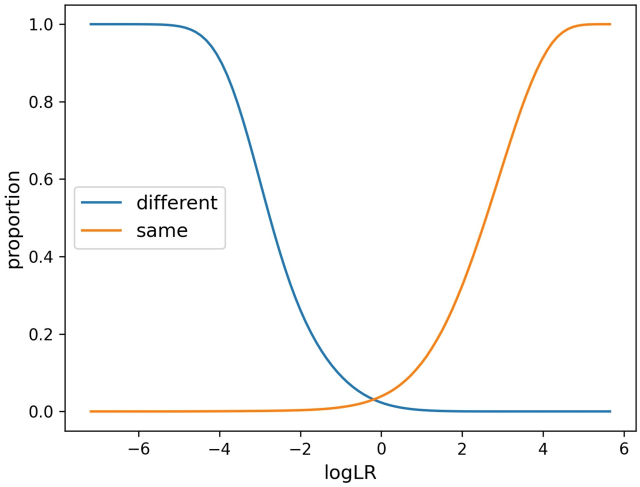

FIGURE 6 Tippett plot of dataset combination (ECAPA-TDNN pretrained on VoxCeleb and LR score calibration on ForVoice120+) with lowest $\mathrm{Cllr}_{\mathrm{min}}$.

4.3 | Effect of sample duration

Of the two best performing cases, the Hungarian was selected to examine the impact of sample duration as it is a low-resource language. Figures 8 and 9 show the $\mathrm{Cllr}_{\mathrm{min}}$ and EER values (without and with enrollment). As expected, the longer the duration of the sample, the better the results. If no enrollment is used, instead of a global 3.1% EER, an EER of 1.1% was obtained for the 10 versus 10 case. On the other hand, the shortest case (2 vs. 2) was 5%. However, if we consider that this is a comparison of the 2-s samples of the unknown and the known speakers, 5% might be an acceptable result in real life. Using speaker enrollment, samples of 10-s duration achieve an EER of 0.2%. The shortest samples go up to 1.6%.

4.4 | Effect of speaking tasks

The results of the best performing dataset combinations were also broken down into speech task combinations. The $\mathrm{Cllr}_{\mathrm{min}}$ and EER values are shown in Figures 10 and 11. Based on the results, task 3 versus task 3 (monologue, describing the events of the previous day of the speaker, EER: 1.8%) has the lowest EERs, while the highest values are obtained in the cross-task combinations (e.g., task 1 vs. task 2, EER: 3.6%). The results show the same trend for the speaker enrollment. The use of monologue gives the best results (0.8% EER) and the information exchange the worst, although only slightly higher (1.2% EER).

5 | DISCUSSION

In the present study, language mismatch effects were examined in terms of the forensic voice comparison perspective using deep speaker embeddings. The aim was to assess whether language differences matter in voice comparison and to investigate whether a model pre-trained on a large-scale dataset with language different from the target samples can be used for forensic voice comparison. The results show that it does. The lowest EER (3.1% and 1.0% without and with speaker enrollment, respectively) was obtained with the model pre-trained on the VoxCeleb dataset, evaluated on the Hungarian ForVoice120+ corpus. This was even better than evaluating the model on the English AusEng dataset. Although the VoxCeleb dataset can be considered multilingual, according to the creators, but only nationality information on speakers is available. Based on this information, widely spoken languages are represented that do not contain the target language used in this study (Hungarian). Therefore, it can be stated that language difference does not degrade the performance of the given technique based on deep learning embeddings. It should be noted that for deep learning

TABLE 4 Speaker verification results obtained with models of different dataset combinations by speaker enrollment.

Evaluation language/ dataset	Embedding model language/dataset	LR calibration language/dataset	Model	Cllr	Cllr_{min}	Cllr_{cal}	EER
Hungarian/ForVoice120	Hungarian	Hungarian/ForVoice120	x-vector	0.517	0.411	0.105	0.123
			ECAPA-TDNN	0.190	0.183	0.006	0.049
	English/AusEng	Hungarian/ForVoice120	x-vector	0.609	0.343	0.265	0.104
			ECAPA-TDNN	0.115	0.110	0.005	0.029
	English/AusEng	English/AusEng	x-vector	0.458	0.343	0.115	0.104
			ECAPA-TDNN	0.378	0.110	0.267	0.029
	VoxCeleb	**Hungarian/ ForVoice120**	x-vector	0.248	0.191	0.058	0.053
			ECAPA-TDNN	**0.050**	**0.045**	**0.006**	**0.010**
	VoxCeleb	**English/AusEng**	x-vector	0.222	0.191	0.031	0.053
			ECAPA-TDNN	**0.067**	**0.045**	**0.022**	**0.010**
English/AusEng	English/AusEng	English/AusEng	x-vector	0.472	0.358	0.114	0.104
			ECAPA-TDNN	0.064	0.061	0.004	0.016
	VoxCeleb	English/AusEng	x-vector	0.324	0.289	0.035	0.085
			ECAPA-TDNN	0.093	0.084	0.009	0.020

The best results are highlighted in bold.

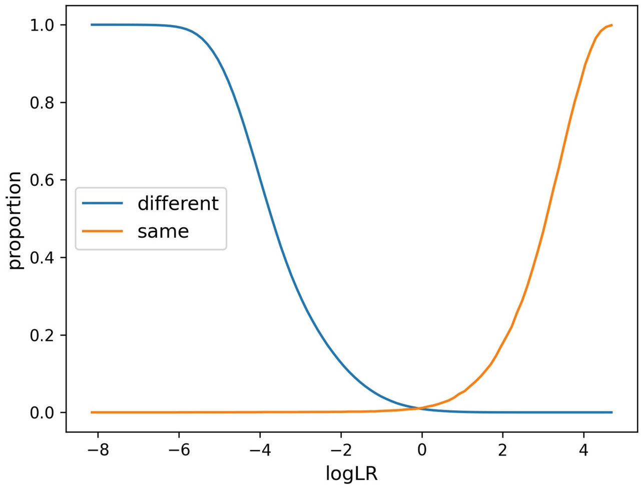

FIGURE 7 Tippett plot of dataset combination (ECAPA-TDNN pretrained on VoxCeleb and LR score calibration on ForVoice120+) with lowest Cllr_{min} for enrollment.

techniques, it is generally useful to have as many training samples as possible. Of course, if there were as many samples of the target language as in the case of VoxCeleb, the performance of the target language model would also improve. The results presented here are intended to support the fact that if there are not enough samples available (which is typical for low-resource languages), one can also use a model in a different language. The language of the corpus used for LR calibration does not seem to affect the performance. Among the two deep learning architectures used, the models structured by the ECAPA-TDNN architecture perform better than the x-vector in all corpus combinations. Although this was expected since ECAPA-TDNN inherently performs better, it was used here without the LDA/PLDA block. This should be considered when comparing results. This is in line with what Desplanques et al. reported [6].

Considering the speaker enrollments, further performance gains can be obtained. The results show that (unsurprisingly) when more samples are available from the suspect, it is better to use speaker enrollment (in this study, the average of the embedding vectors) to compare the voice sample of the offender. The best performing model achieves an absolute EER decrease of 2.1% (from 3.1% to 1.0%). Of course, this can only be achieved if more than one recording of the suspect is available, but recordings can be obtained deliberately during an investigation.

Further details can be revealed by breaking down the results by sample length and speech style (simulated with different speech tasks). The analysis of sample length shows that the longer the duration of the sample in question, the better the performance, as would naturally be expected. Comparing sample pairs of 2-s duration, an EER of 5% is achieved in the best case, while for samples of 10-s duration, this drops to 1.1%. With enrolment (comparing a single sample of the offender to the average vector of the suspect), this can be further improved: for 10 s samples, the EER is 0.5%. This means that, in practice, pre-trained models can be used on samples of other languages, but the longer the sample, the better the performance. Furthermore, it is recommended to record more samples from the suspect and to use an averaged embedding vector.

The results by speech task show that there is a slight gain in using the same speech style in the compared samples (at least the spontaneity would be the same) if enrolment is not possible. However, if multiple recordings of the suspect are available, there is no real difference if different speech styles are used, and it does not really matter.

The results show that the automatic, sample-wise forensic voice comparison technique used in this study can be used in practical, real-world scenarios. This is useful when only a single sample is

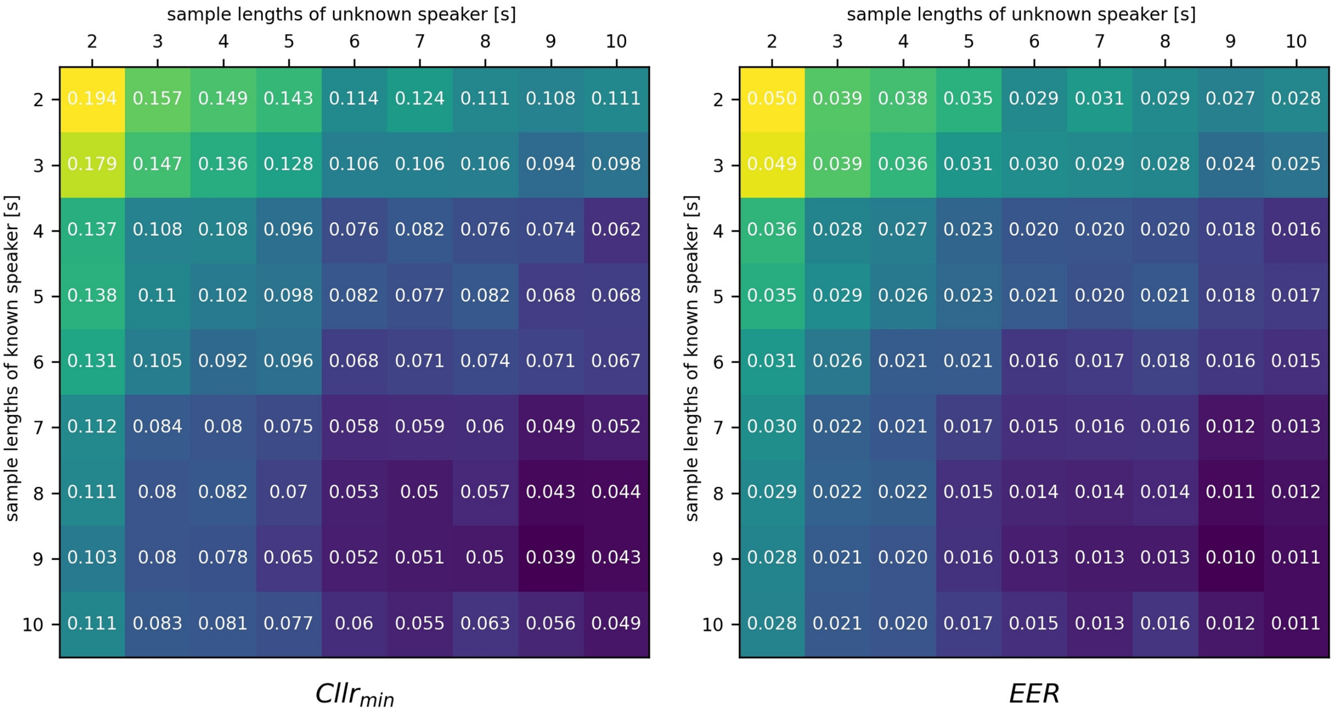

FIGURE 8 Heatmap of Cllr_{min} and EER values depending on sample duration without speaker enrollment. ECAPA-TDNN models trained on VoxCeleb, LR score calibration done on ForVoice120+.

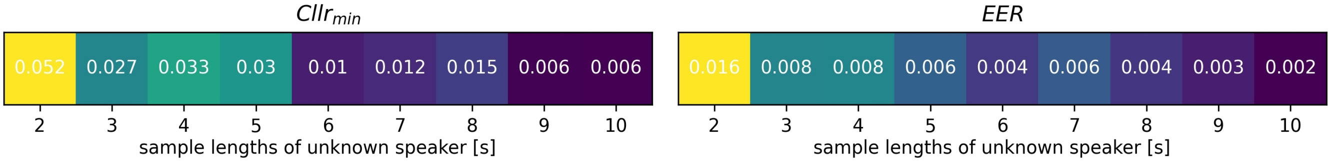

FIGURE 9 Heatmap of Cllr_{min} and EER values depending on sample duration with speaker enrollment. ECAPA-TDNN models trained on VoxCeleb, LR score calibration done on ForVoice120+.

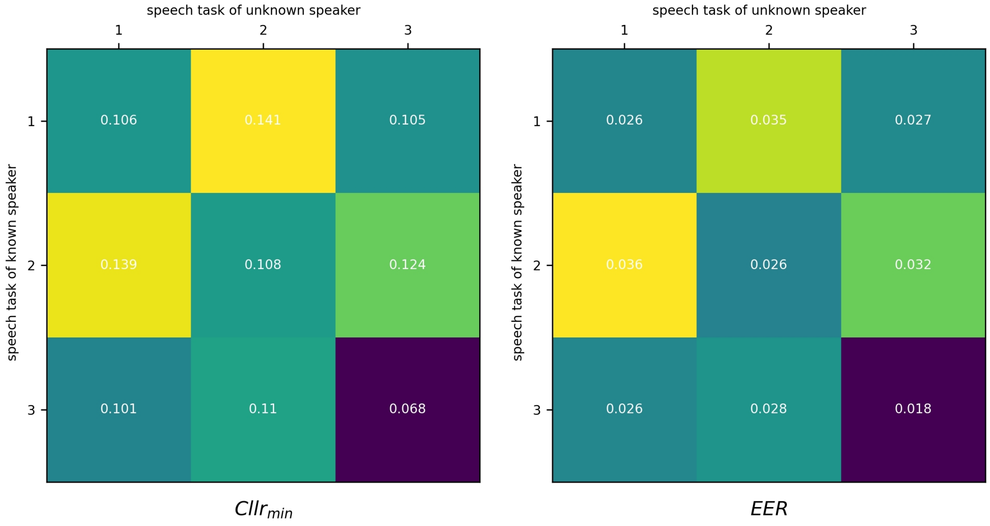

FIGURE 10 Heatmap of Cllr_{min} and EER values depending on speech task without speaker enrollment. The ECAPA-TDNN models were trained on VoxCeldeb, LR score calibration was done on ForVoice120+.

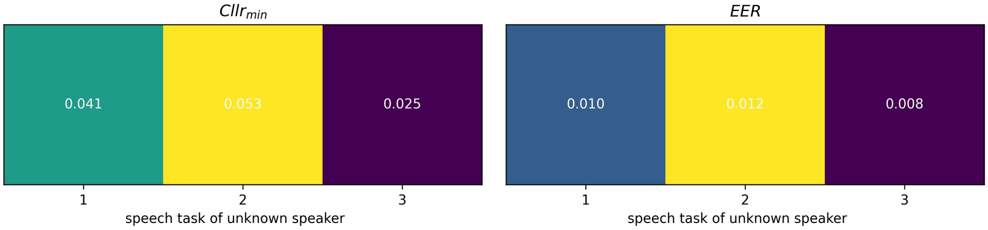

FIGURE 11 Heatmap of Cllr_{min} and EER values depending on speech task with speaker enrollment. The ECAPA-TDNN models were trained on VoxCeleb, LR score calibration was done on ForVoice120+.

available from the offender. (Multiple samples may be available from the suspect since speaker enrollment was also investigated here.) The Cllr and EER values obtained are not directly comparable with other studies (due to differences in the datasets used). The aim here was not to evaluate speaker enrolment models (these have been evaluated in other studies) but to investigate the language mismatch properties of a possible voice comparison system.

The dataset used for evaluation is in a low-resource language (Hungarian). Languages of this type often do not have sufficient speech samples to train a deep learning model and calibrate an LR framework. The results obtained here show that models pre-trained in a different language can indeed be used in the target language (which is also linguistically remote). These pre-trained models (even available online) can be universally applied to speaker verification regardless of the language of the samples in question. This can be very useful if the target language is a low-resource language where few samples are available for forensic purposes (neither proper forensic evaluation is possible nor deep learning models can be trained). The study shows that a large target language corpus is not needed to apply these models. The evaluation metrics show that even the language used to calibrate LR scores can differ from the final target language to achieve similar (good) results. A small dataset with a limited number of speakers is sufficient to evaluate the framework in the language in which the expert wants to use it.

Some comments on the legal use of pre-trained VoxCeleb models: Although VoxCeleb is a collection of YouTube videos and is licensed under the Creative Commons Attribution 4.0 International License, which means that the copyright of the original versions of the videos remains with the original owners, the models pre-trained on the dataset are licensed under Apache 2.0, which allows both research and commercial use of the models. The results presented here are not specific to the VoxCeleb, any model trained on a large-scale dataset would be sufficient.

6 | CONCLUSION

Until now, in Hungarian forensic practice, several types of audio samples are recorded during speaker verification if the suspected speaker is known. In the sampling procedure, spontaneous speech is recorded, and the suspect also reads out a text material taken from the unknown speaker's speech sample. In the acoustic–phonetic recognition methodology, the expert uses this read speech sample to compare matching sound sequences (such as words). Additionally, the spontaneous sample is also used, for example, to determine average pitch values. However, the methodology used for biometric measurements needs to be reconsidered in light of the results presented in this paper. We believe that not all voice sample types are needed to be included in the measurements for voice biometrics. Instead, it is sufficient to measure only spontaneous speech samples, and efforts should be made to have more than one voice sample from each speaker available to the expert. A limitation in this study may be that the age distribution of the dataset used for evaluation represents mainly the 18–35 age range. However, the language dependency statements derived from the results are not affected by this phenomenon because all models are evaluated on this same dataset.

This study could be useful to improve forensic speaker recognition by applying voice biometrics technology to different speech tasks and sample durations. Based on the results, future plans include investigating newly developed speaker embedding techniques, how they perform compared to the ECAPA-TDNN. Also, we plan to investigate how emotions affect forensic voice comparison and also how voice of twins may degrade the performance of the LR framework.

Audio forensics experts summarize the results of their speaker recognition measurements in an interpretation framework. Its structure and characteristics determine the final expert conclusion on the probability of speaker identity in the expert report. Using the results of this study, a new interpretation framework has been developed in the Hungarian expert field and will be published in the near future. The new framework will make the expert analysis more objective and allow for a more detailed evaluation.

FUNDING INFORMATION

The work was funded by project no. FK128615, which has been implemented with the support provided from the National Research, Development and Innovation Fund of Hungary, financed under the FK_18 funding scheme.

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

ORCID

Dávid Sztahó https://orcid.org/0000-0002-7361-4260

REFERENCES

1. Hanifa RM, Isa K, Mohamad S. A review on speaker recognition: technology and challenges. Comput Electr Eng. 2021;90:107005. https://doi.org/10.1016/j.compeleceng.2021.107005
2. Hansen JH, Hasan T. Speaker recognition by machines and humans: a tutorial review. IEEE Signal Process Mag. 2015;32(6):74–99. https://doi.org/10.1109/MSP.2015.2462851
3. Variani E, Lei X, McDermott E, Moreno IL, Gonzalez-Dominguez J. Deep neural networks for small footprint text-dependent speaker verification. Proceedings of the 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP); 2014 may 4–9; Florence, Italy. Piscataway, NJ: IEEE; 2014. p. 4052–6. https://doi.org/10.1109/ICASSP.2014.6854363
4. Chen N, Qian Y, Yu K. Multi-task learning for text-dependent speaker verification. Proceedings of the 16th annual conference of the international speech communication association (Interspeech 2015); 2015 Sept 6–10; Dresden Germany. Baixas, France: ISCA; 2015. p. 185–9. https://doi.org/10.21437/Interspeech.2015-81
5. Snyder D, Garcia-Romero D, Povey D, Khudanpur S. Deep neural network embeddings for text-independent speaker verification. Proceedings of the 18th annual conference of the international speech communication association (Interspeech 2017); 2017 Aug 20–24; Stockholm, Sweden. Baixas, France: ISCA; 2017. p. 999–1003. https://doi.org/10.21437/Interspeech.2017-620
6. Desplanques B, Thienpondt J, Demuynck K. ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification. Proceedings of the 21st annual conference of the international speech communication association (Interspeech 2020); 2020 Oct 25–29; Shanghai, China. Baixas, France: ISCA; 2020. p. 3830–4. https://doi.org/10.21437/Interspeech.2020-2650
7. Sztahó D, Szaszák G, Beke A. Deep learning methods in speaker recognition: a review. Period. Polytech Electr Eng Comput Sci. 2021;65(4):310–28. https://doi.org/10.3311/PPee.17024
8. Reynolds DA, Rose RC. Robust text-independent speaker identification using Gaussian mixture speaker models. IEEE Trans Audio Speech Lang Process. 1995;3(1):72–83. https://doi.org/10.1109/89.365379
9. Morrison GS. Measuring the validity and reliability of forensic likelihood-ratio systems. Sci Justice. 2011;51(3):91–8. https://doi.org/10.1016/j.scijus.2011.03.002
10. Saks MJ, Koehler JJ. The coming paradigm shift in forensic identification science. Science. 2005;309(5736):892–5. https://doi.org/10.1126/science.1111565
11. Bazen AM, Veldhuis RN. Likelihood-ratio-based biometric verification. IEEE Trans Circuits Syst Video Technol. 2004;14(1):86–94. https://doi.org/10.1109/TCSVT.2003.818356
12. Matz MV, Nielsen R. A likelihood ratio test for species membership based on DNA sequence data. Philos Trans R Soc Lond B Biol Sci. 2005;360(1462):1969–74. https://doi.org/10.1098/rstb.2005.1728
13. Kelly F, Forth O, Kent S, Gerlach L, Alexander A. Deep neural network based forensic automatic speaker recognition in VOCALISE using x-vectors. Proceedings of 2019 AES international conference on audio forensics; 2019 Jun 18–20; Porto, Portugal. New York, NY: AES; 2019.
14. Morrison GS. A comparison of procedures for the calculation of forensic likelihood ratios from acoustic–phonetic data: multivariate kernel density (MVKD) versus Gaussian mixture model–universal background model (GMM–UBM). Speech Commun. 2011;53(2):242–56. https://doi.org/10.1016/j.specom.2010.09.005
15. Mandasari MI, McLaren M, van Leeuwen DA. Evaluation of i-vector speaker recognition systems for forensic application. Proceedings of the 12th annual conference of the international speech communication association (Interspeech 2011); 2011 Aug 27–31; Florence, Italy. Baixas, France: ISCA; 2011. p. 21–4. https://doi.org/10.21437/Interspeech.2011-6
16. Poddar A, Sahidullah M, Saha G. Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biom. 2018;7(2):91–101. https://doi.org/10.1049/iet-bmt.2017.0065
17. Li L, Wang D, Zhang X, Zheng TF, Jin P. System combination for short utterance speaker recognition. Proceedings of the 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA); 2016 Dec 13–16; Jeju, Korea. Piscataway, NJ: IEEE; 2016. p. 1–5. https://doi.org/10.1109/APSIPA.2016.7820903
18. Min Kye S, Son Chung J, Kim H. Supervised attention for speaker recognition. Proceedings of the 2021 IEEE spoken language technology workshop (SLT); 2021 Jan 19–22; Shenzhen, China. Piscataway, NJ: IEEE; 2021. p. 286–93. https://doi.org/10.1109/SLT48900.2021.9383579
19. Rohdin J, Silnova A, Diez M, Plchot O, Matějka P, Burget L, et al. End-to-end DNN based text-independent speaker recognition for long and short utterances. Comp Speech Lang. 2020;59:22–35. https://doi.org/10.1016/j.csl.2019.06.002
20. Wang Z, Hansen JHL. Multi-source domain adaptation for text-independent forensic speaker recognition. IEEE/ACM Trans Audio Speech Lang Process. 2022;30:60–75. https://doi.org/10.1109/TASLP.2021.3130975
21. Morrison GS, Rose P, Zhang C. Protocol for the collection of databases of recordings for forensic-voice-comparison research and practice. Aust J Forensic Sci. 2012;44(2):155–67. https://doi.org/10.1080/00450618.2011.630412
22. Chanchaochai N, Cieri C, Debrah J, Ding H, Jiang Y, Liao S, et al. Global TIMIT: acoustic-phonetic datasets for the world's languages. Proceedings of the 19th annual conference of the international speech communication association (Interspeech 2018); 2018 Sep 2–6; Hyderabad, India. Baixas, France: ISCA; 2018. p. 192–6. https://doi.org/10.21437/Interspeech.2018-1185
23. Nagrani A, Chung JS, Xie W, Zisserman A. Voxceleb: large-scale speaker verification in the wild. Comput Speech Lang. 2020;60:101027. https://doi.org/10.1016/j.csl.2019.101027
24. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. Proceedings of the 40th international conference on acoustics, speech and signal processing (ICASSP 2015); 2015 Apr 19–24; Brisbane, Australia. Piscataway, NJ: IEEE; 2015. p. 5206–10. https://doi.org/10.1109/ICASSP.2015.7178964
25. Sadjadi SO, Greenberg C, Singer E, Mason L, Reynolds D. The 2021 NIST speaker recognition evaluation. arXiv preprint. arXiv:220410242 2022 https://doi.org/10.48550/arXiv.2204.10242
26. Kleynhans NT, Barnard E. Language dependence in multilingual speaker verification. In: Nicolls F, editor. Proceedings of the sixteenth annual symposium of the pattern recognition Association of South Africa; 2005 Nov 23–25; Langebaan, South Africa. Punjab, India: International Association for Pattern Recognition; 2005. p. 23–5.
27. Li L, Wang D, Rozi A, Zheng TF. Cross-lingual speaker verification with deep feature learning. arXiv preprint. arXiv170607861 2017 https://doi.org/10.48550/arXiv.1706.07861
28. Chojnacka R, Pelecanos J, Wang Q, Moreno IL. SpeakerStew: scaling to many languages with a triaged multilingual text-dependent and text-independent speaker verification System. arXiv preprint. arXiv:210402125 2021 https://doi.org/10.48550/arXiv.2104.02125
29. Misra A, Hansen JH. Spoken language mismatch in speaker verification: an investigation with nist-sre and cross bi-ling corpora. Proceedings of the 2014 IEEE spoken language technology workshop

(SLT); 2014 Dec 7–10; South Lake Tahoe, NV. Piscataway, NJ: IEEE; 2014. p. 372–7. https://doi.org/10.1109/SLT.2014.7078603

30. Vaheb A, Choobbasti AJ, Najafabadi SHE, Safavi S. Investigating language variability on the performance of speaker verification systems. In: Karpov A, Jokisch O, Potapova R, editors. Proceedings of the 20th international conference on speech and computer; 2018 Sep 18–22; Leipzig, Germany. Cham, Switzerland: Springer; 2018. p. 718–27. https://doi.org/10.1007/978-3-319-99579-3_73
31. Fabien M, Motlicek P. Open-set speaker identification pipeline in live criminal investigations. Proceedings of the 2021 ISCA symposium on security and privacy in speech communication; 2021 Nov 10–12; held virtually. Baixas, France: ISCA; 2021. p. 21–4. https://doi.org/10.21437/SPSC.2021-5
32. Skarnitzl R, Asiaee M, Nourbakhsh M. Tuning the performance of automatic speaker recognition in different conditions: effects of language and simulated voice disguise. Int J Speech Lang Law. 2019;26(2):209–29. https://doi.org/10.1558/ijsll.39778
33. Kelly F, Fröhlich A, Dellwo V, Forth O, Kent S, Alexander A. Evaluation of VOCALISE under conditions reflecting those of a real forensic voice comparison case (forensic_eval_01). Speech Commun. 2019;112:30–6. https://doi.org/10.1016/j.specom.2019.06.005
34. Afshan A, Guo J, Park SJ, Ravi V, McCree A, Alwan A. Variable frame rate-based data augmentation to handle speaking-style variability for automatic speaker verification. arXiv preprint. arXiv:200803616 2020 https://doi.org/10.48550/arXiv.2008.03616
35. Afshan A, Alwan A. Learning from human perception to improve automatic speaker verification in style-mismatched conditions. arXiv preprint. arXiv:220613684 2022 https://doi.org/10.48550/arXiv.2206.13684
36. González Hautamäki R, Hautamäki V, Kinnunen T. On the limits of automatic speaker verification: explaining degraded recognizer scores through acoustic changes resulting from voice disguise. J Acoust Soc Am. 2019;146(1):693–704. https://doi.org/10.1121/1.5119240
37. Ravanelli M, Parcollet T, Plantinga P, Rouhe A, Cornell S, Lugosch L, et al. SpeechBrain: A general-purpose speech toolkit. arXiv preprint. arXiv:210604624 2021 https://doi.org/10.48550/arXiv.2106.04624
38. Hugging Face. Pretrained ECAPA-TDNN model using SpeechBrain. https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb. Accessed 7 Feb 2022
39. Hugging Face. Pretrained TDNN (x-vector) model using SpeechBrain. https://huggingface.co/speechbrain/spkrec-xvect-voxceleb. Accessed 7 Feb 2022
40. Biosa G, Giurghita D, Alladio E, Vincenti M, Neocleous T. Evaluation of forensic data using logistic regression-based classification methods and an R shiny implementation. Front Chem. 2020;8:738. https://doi.org/10.3389/fchem.2020.00738
41. Morrison GS, Weber P, Basu N, Puch-Solis R, Randolph-Quinney PS. Calculation of likelihood ratios for inference of biological sex from human skeletal remains. Forensic Sci Int Synerg. 2021;3:100202. https://doi.org/10.1016/j.fsisyn.2021.100202
42. Morrison GS, Zhang C, Enzinger E, Ochoa F, Bleach D, Johnson M, et al. Forensic database of voice recordings of 500+ Australian English speakers. 2015. http://databases.forensic-voice-comparison.net/. Accessed 20 Feb 2022
43. Neuberger T, Gyarmathy D, Gráczi TE, Horváth V, Gósy M, Beke A. Development of a large spontaneous speech database of agglutinative Hungarian language. Proceedings of the 17th international conference on text, speech, and dialogue; 2014 Sep 8–12; Brno, Czech Republic. Cham, Switzerland: Springer; 2014. p. 424–31. https://doi.org/10.1007/978-3-319-10816-2_51
44. Vicsi K, Kocsor A, Teleki C, Tóth L. Beszédadatbázis irodai számítógép-felhasználói környezetben (A speech database for office environment). In: AlexinZ CD, editor. Proceedings of the second conference on Hungarian computational linguistics (MSZNY 2004); 2004 Dec 9–10; Szeged, Hungary. Szeged, Hungary: MSZNY; 2004. p. 315–8.
45. Hechmi K, Trong TN, Hautamäki V, Kinnunen T. Voxceleb enrichment for age and gender recognition. Proceedings of the 2021 IEEE automatic speech recognition and understanding workshop (ASRU); 2021 Dec 13–17; Cartagenda, Colombia. Piscataway, NJ: IEEE; 2021. p. 687–93. https://doi.org/10.1109/ASRU51503.2021.9688085
46. Van Leeuwen DA, Brümmer N. An introduction to application-independent evaluation of speaker recognition systems. In: Müller C, editor. Speaker classification I. Berlin/Heidelberg, Germany: Springer Berlin, Heidelberg; 2007. p. 330–53. https://doi.org/10.1007/978-3-540-74200-5_19
47. Brümmer N, Du Preez J. Application-independent evaluation of speaker detection. Comput Speech Lang. 2006;20(2–3):230–75. https://doi.org/10.1016/j.csl.2005.08.001

How to cite this article: Sztahó D, Fejes A. Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. J Forensic Sci. 2023;68:871–883. https://doi.org/10.1111/1556-4029.15250