Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

# On the influence of language similarity in non-target speaker verification trials

*Paul M. Reuter*[1], *Michael Jessen*[2]

[1]Fraunhofer Institute for Digital Media Technology, Division Hearing, Speech and Audio Technology, Oldenburg, Germany
[2]Department of Text, Speech and Audio, Bundeskriminalamt, Wiesbaden, Germany
paul.maria.reuter@idmt.fraunhofer.de, michael.jessen@bka.bund.de

## Abstract

In this paper, we investigate the influence of language similarity in cross-lingual non-target speaker verification trials using a state-of-the-art speaker verification system, ECAPA-TDNN, trained on multilingual and monolingual variants of the VoxCeleb dataset. Our analysis of the score distribution patterns on multilingual Globalphone and LDC CTS reveals a clustering effect in speaker comparisons involving a training language, whereby the choice of comparison language only minimally impacts scores. Conversely, we observe a language similarity effect in trials involving languages not included in the training set of the speaker verification system, with scores correlating with language similarity measured by a language classification system, especially when using multilingual training data.
**Index Terms**: speaker recognition, speaker verification, cross-lingual, forensics

## 1. Introduction

There is continued interest over the years in how and to what extent automatic speaker recognition is affected by language. A common research paradigm to address this question has been to study mismatch between the language spoken by the questioned speaker and known speaker [1, 2, 3, 4]. Some researchers approached the language effect by studying language mismatch not between questioned and known speaker but between the language spoken by both of them (test language) and the data used for embedding-level training [5, 6, 7], for score normalisation [5, 8], or calibration [7, 9]. Most of these studies focus on overall performance characteristics such as EER or Cllr and some of them show how the performance loss due to language mismatch relative to language match could be mitigated. Forensic automatic speaker recognition, the main target of this work, has an interest not only in overall speaker discrimination performance, but also in how the results of a system test (validation) are distributed in Tippett plots or other distribution graphs [10, 11]. The current study focuses entirely on different-speaker comparisons and on the score distributions that these comparisons generate. Some of these comparisons involve the same language (language match) others different languages (language mismatch) from a set of languages available in the corpora to be used here. We hypothesise that different-speaker comparisons involving the same language yield higher scores than those involving different languages (H1). We also hypothesise that among the different-language (aka cross-lingual) comparisons (aka trials), those that involve languages that are in some way similar to each other yield higher scores than those that involve languages that are more distant to each other (H2). What counts as similar or distant will be operationalized by using a language classification system. Moreover, it will be examined to what extent the composition of the deep-level training data influences the results, by using either multi-lingual or mono-lingual training sets. The novelty of this research lies in its focus on the language-induced score shift patterns and the range of factors considered for their explanation. It has implications for forensics in particular, but also more generally, including the current discussion on potential bias in AI processing. The score distribution of different-speaker trials is important for score normalisation methods such as t-norm or s-norm ([12] for a summary of forensically-oriented studies using these normalisation methods; see also [6]). It also has relevance for calibration. When no data are available to form a normalization or calibration set that is equilingual to the recordings in the forensic comparison, it is important to know to what extent the normalisation or calibration procedure would be biased when using a non-matching language set and whether the effect can be predicted by language distance or other factors.

## 2. Methods

### 2.1. Language classification

For language classification, a ResNet34 [13] was trained on VoxLingua107 [14] using cross-entropy loss. The model was used to analyze the distribution of languages spoken in the VoxCeleb1 and 2 datasets [15, 16] which served as training data for the speaker verification models (see Fig. 1).

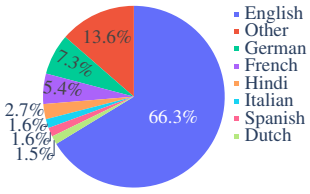


Figure 1: *Language distribution in VoxCeleb1 and 2*

### 2.2. Speaker verification

ECAPA-TDNN [17] was adopted as the speaker verification model architecture as it is commonly used in state-of-the-art speaker verification [18, 19, 20]. It is built upon the original x-vector architecture [21] and employs several enhancements regarding channel attention, propagation and aggregation. In our experiments, we used two models of the same ECAPA-TDNN architecture trained on different datasets. The first model is the pretrained ECAPA-TDNN model from Speechbrain [22] trained on VoxCeleb1 and 2 (also referred to as system Vox1+2). The second model was trained exclusively on English utterances


10.21437/Interspeech.2025-526

from VoxCeleb2 (also referred to as system Vox2-en). Both models were trained with the AAM-Softmax loss [23] for 15 epochs, following the Speechbrain recipe, resulting in an EER of 0.90 % and 1.49 % on the VoxCeleb1-O test set, respectively.

### 2.3. Datasets and preparation

We conducted our experiments on two multilingual datasets.

#### *2.3.1. Globalphone*

Globalphone [24] is a multilingual database consisting of high-quality recordings of read sentences. It is uniform across languages with respect to the recording conditions and the collection scenario. We performed voice activity detection on every file to remove non-speech segments and cut the remaining speech content to 5 seconds length. From every available language, 40 male and/or female speakers were randomly selected with two recordings per speaker. Since English is the main language in VoxCeleb but not contained in Globalphone, we included 40 random male and female speakers from Librispeech [25], which is very similar in acoustic conditions. This left us with a total of 10 languages for male speakers and 13 languages for female speakers.

#### *2.3.2. LDC Multilingual CTS 2011*

LDC Multilingual CTS 2011 [26] is a collection of telephone speech datasets in various languages. We performed voice activity detection and cut all recordings to 20 seconds length. Per language, 20 male and/or female speakers were randomly selected with two recordings per speaker. We ended up with a total of 16 languages for male speakers and 18 for female speakers.

### 2.4. Comparison procedure

For both datasets, we performed pair-wise comparisons between all possible different-speaker combinations of files within and across languages. Speaker verification was performed using cosine similarity of the speaker embeddings. No post-processing (like score normalization) was applied. To measure the language similarity of two utterances, we use the cosine similarity between the outputs of the last hidden layer (embeddings) of the language classifier. This way, for every two-file comparison, a speaker similarity and a language similarity score were obtained.

## 3. Results

### 3.1. Qualitative analysis

Fig. 2 shows the score distributions of different-speaker trials in Globalphone involving German and Russian, respectively, for the system trained on VoxCeleb1+2. For both reference languages, speaker similarity scores are highest on average for the same-language case (German-German, Russian-Russian) (H1 fulfilled). However, there are noticeable differences in the distributions of different-language scores. While scores of comparisons involving German are largely unaffected by the language being compared (H2 not fulfilled), there is a shift in scores of Russian comparisons depending on the compared language with generally higher scores for related (Slavic) languages such as Bulgarian, Croatian and Czech (H2 fulfilled). Examining the average female different-speaker similarity scores of all language comparisons (see Table 1) reveals a pattern similar to German for English and French as well as a pattern similar to

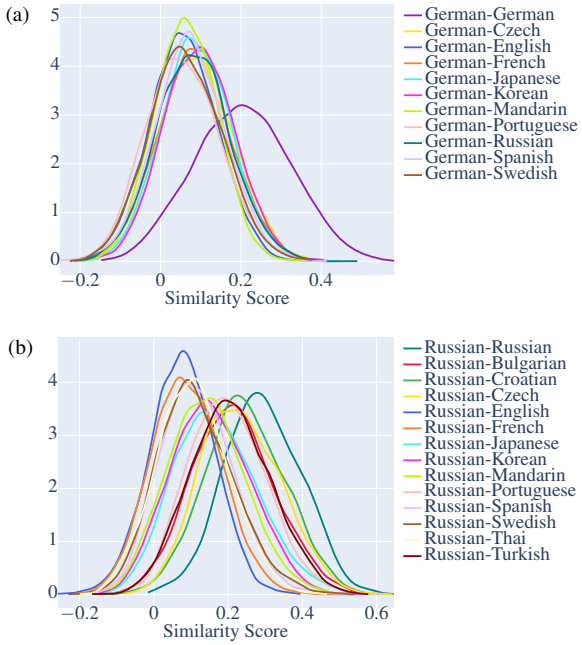


Figure 2: *Score distributions of different-speaker trials involving a) German (males) and b) Russian (females) in Globalphone set with VoxCeleb1+2 training*

Russian for other Slavic languages in the Globalphone set. We also observe higher speaker similarity scores among Mandarin, Japanese, Thai and Korean. In the LDC set, again Slavic-Slavic comparisons resulted in higher average scores than Slavic-Non-Slavic comparisons. Increased speaker similarity scores were also found among Bengali, Punjabi, Urdu and Pashto (Indo-Iranian languages) and among Thai and Lao (Tai languages). Due to space constraints LDC results are not presented here; however, the relationship between speaker similarity and language similarity in LDC will be explored later on.

The results suggest that in cross-lingual different-speaker trials, the choice of comparison language has a small effect on speaker similarity scores if the reference language is well represented in the training dataset (such as English, German, French, see Fig. 1). As part of this pattern, the score distributions of all the compared languages are tightly clustered. For reference languages not included or barely present in the training dataset (such as Slavic languages), speaker similarity scores are higher for related than for non-related comparison languages and no clustering of score distributions is observed. Cross-lingual comparisons involving a training language quite robustly produce the lowest scores (see Table 1), presumably because the system interprets known vs known/unknown language pairs as a strong contrast, whereas unknown vs unknown language pairs are interpreted as less contrasting.

Score distributions of the system trained on English-only VoxCeleb2 from German and Russian comparisons are depicted in Fig. 3. Once again, same-language comparisons achieve the highest average score (H1 fulfilled) and the system outputs lowest average scores in comparisons involving the training language (English) on Globalphone and LDC. (These two are robust patterns that re-emerge throughout this study; they will not be further mentioned beyond here.) Compared to system Vox1+2, neither is a clustering effect shown in the German-based comparisons of Fig. 3a nor is a clear language similar-

Table 1: *Average female different-speaker similarity scores of language comparisons in Globalphone set with VoxCeleb1+2 training*

| | English | Swedish | French | Spanish | Portuguese | Czech | Bulgarian | Croatian | Russian | Turkish | Mandarin | Japanese | Thai | Korean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| English | 0.121 | 0.083 | 0.086 | 0.086 | 0.073 | 0.076 | 0.072 | 0.074 | 0.077 | 0.068 | 0.087 | 0.085 | 0.087 | 0.079 |
| Swedish | | 0.303 | 0.080 | 0.112 | 0.111 | 0.115 | 0.095 | 0.126 | 0.114 | 0.146 | 0.131 | 0.127 | 0.148 | 0.115 |
| French | | | 0.243 | 0.075 | 0.056 | 0.098 | 0.066 | 0.063 | 0.089 | 0.079 | 0.094 | 0.094 | 0.091 | 0.102 |
| Spanish | | | | 0.274 | 0.145 | 0.080 | 0.085 | 0.117 | 0.117 | 0.118 | 0.163 | 0.152 | 0.184 | 0.153 |
| Portuguese | | | | | 0.232 | 0.167 | 0.160 | 0.193 | 0.198 | 0.170 | 0.157 | 0.141 | 0.123 | 0.132 |
| Czech | | | | | | 0.297 | 0.195 | 0.205 | 0.234 | 0.154 | 0.112 | 0.146 | 0.140 | 0.125 |
| Bulgarian | | | | | | | 0.214 | 0.194 | 0.214 | 0.150 | 0.09 | 0.123 | 0.079 | 0.103 |
| Croatian | | | | | | | | 0.261 | 0.245 | 0.203 | 0.137 | 0.155 | 0.102 | 0.138 |
| Russian | | | | | | | | | 0.293 | 0.207 | 0.142 | 0.168 | 0.106 | 0.154 |
| Turkish | | | | | | | | | | 0.242 | 0.202 | 0.186 | 0.127 | 0.202 |
| Mandarin | | | | | | | | | | | 0.392 | 0.284 | 0.274 | 0.298 |
| Japanese | | | | | | | | | | | | 0.287 | 0.230 | 0.275 |
| Thai | | | | | | | | | | | | | 0.398 | 0.204 |
| Korean | | | | | | | | | | | | | | 0.338 |

ity effect shown in Russian-based comparisons of Fig. 3b (H2 not fulfilled). Reduction of the clustering effect may have to do with the fact that German no longer is a training language. Reduction of the language similarity effect is probably due to the fact that the system is not trained for any language other than English.

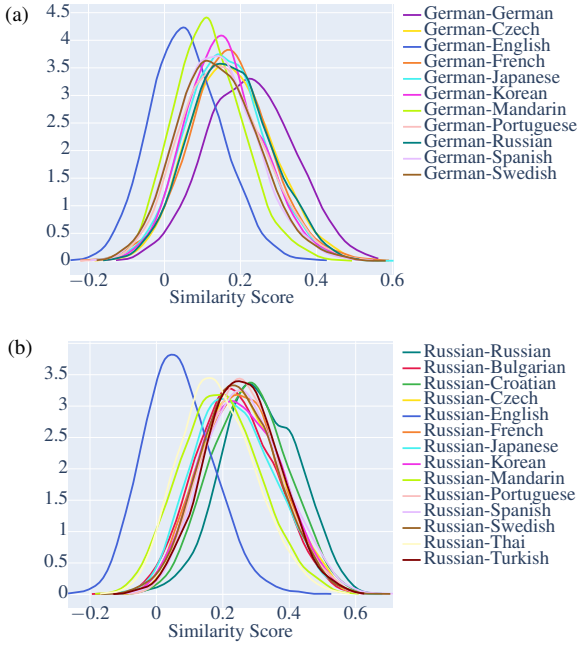


Figure 3: *Score distributions of different-speaker trials involving a) German (males) and b) Russian (females) in Globalphone set with English-only VoxCeleb2 training*

To further understand the two different score distribution patterns in Fig. 2 and 3, Globalphone embeddings were averaged per speaker and visualized using t-SNE [27] (see Fig. 4). We make two observations: Firstly, the t-SNE map is divided into male and female speakers. This is expected as biological sex affects vocal characteristics. Secondly, for the system trained on VoxCeleb1+2, speakers of languages well represented in the training dataset (English, German, French) or related languages (Swedish) form clusters in t-SNE map that are clearly separated from clusters of speakers of other languages. We assume that in the training of a system, speakers of the main training languages are arranged in the embedding space such that their distance to speakers of different languages is largely unaffected by the language spoken (what was seen in Fig. 2a). As a result, speakers of training languages have no preferred association with other languages and isolate themselves in the t-SNE map unlike speakers of untrained languages. When trained on English exclusively, only English speakers form their own cluster in the t-SNE map.

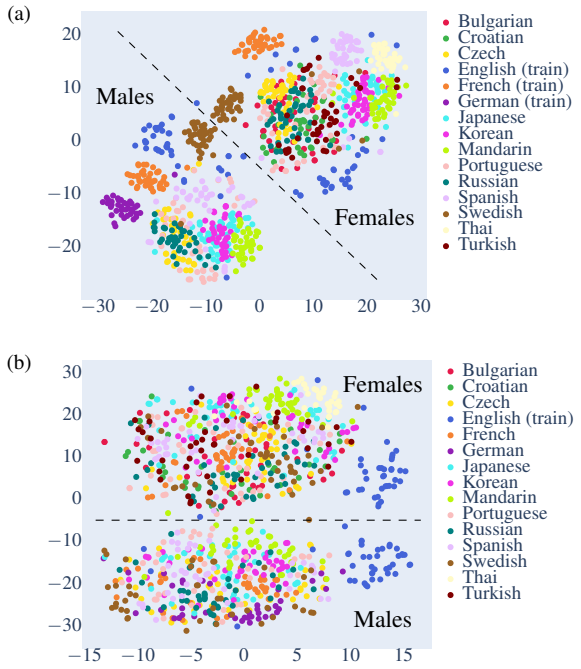


Figure 4: *t-SNE visualization of averaged speaker embeddings from Globalphone set with a) VoxCeleb1+2 training and b) English-only VoxCeleb2 training. The perplexity parameter was set to 50. Male and female speakers are fairly accurately separated by the drawn dashed line.*

### 3.2. Quantitative analysis

So far the results have shown that the effect of language similarity on non-target speaker verification scores is more complex than originally assumed when H2 was expressed. A language similarity effect tends to occur only when a system is trained on several languages and the languages that are compared are not among those the system is trained for. When instead the trained languages constitute the basis of the comparisons, the score distributions of the languages that are compared against

the trained ones tend to form a cluster. To separate the effect of training in the following analysis, we split languages into the category trained (English, German and French for system Vox1+2 and English for Vox2-en) or untrained (all other languages) and consider only trained vs untrained and untrained vs untrained comparisons. In order to quantify the trends that were shown in the previous section, two procedures were used.

Table 2: *Score standard deviations from trained vs untrained (highlighted) and untrained vs untrained comparisons in Globalphone set for systems Vox1+2 and Vox2-en*

| Language | Vox1+2 | Vox2-en |
|---|---|---|
| English | 0.086 | 0.097 |
| German | 0.090 | 0.106 |
| French | 0.097 | 0.113 |
| Swedish | 0.099 | 0.111 |
| Spanish | 0.107 | 0.112 |
| Portuguese | 0.109 | 0.112 |
| Czech | 0.112 | 0.112 |
| Bulgarian | 0.116 | 0.119 |
| Croatian | 0.113 | 0.116 |
| Russian | 0.113 | 0.115 |
| Turkish | 0.113 | 0.115 |
| Mandarin | 0.123 | 0.118 |
| Japanese | 0.122 | 0.116 |
| Thai | 0.121 | 0.120 |
| Korean | 0.125 | 0.117 |

Firstly, we measured the standard deviation of cross-lingual comparison scores. The average standard deviations of male and female language sets (if available) are displayed in Table 2. It can be seen that for system Vox1+2 the standard deviation of scores from trained vs untrained comparisons is lower than that from untrained vs untrained comparisons. For system Vox2-en, English produces the lowest standard deviation of scores. The results show that comparisons involving a training language are less affected by the comparison language and hence have a smaller score variation which confirms the hypothesis of a clustering effect.

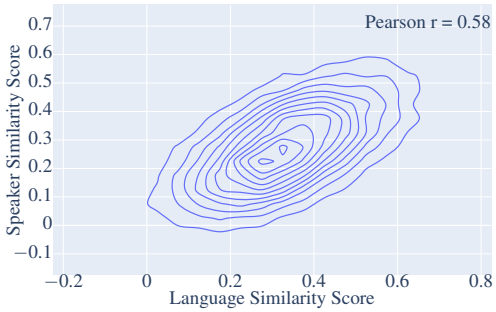


Figure 5: *Language and speaker similarity score distribution contours of male different-speaker cross-lingual trials involving Punjabi in LDC set with VoxCeleb1+2 training*

Secondly, we performed a correlation analysis between speaker similarity and language similarity. For every language considered, the Pearson correlation coefficient between speaker and language similarity scores obtained from different-speaker cross-lingual comparisons (trained vs untrained and untrained vs untrained) was calculated. In Fig. 5, the speaker similarity scores of the system trained on VoxCeleb1+2 are plotted against the language similarity scores for Punjabi. There is a moderate positive correlation between speaker and language similarity. Table 3 displays the average of the calculated correlation coefficient from the male and female set of each language (if available).

Table 3: *Pearson correlation coefficient between speaker and language similarity scores obtained from trained vs untrained (highlighted) and untrained vs untrained comparisons in Globalphone and LDC sets for systems Vox1+2 and Vox-en*

| | Globalphone | | LDC | |
|---|---|---|---|---|
| Language | Vox1+2 | Vox2-en | Vox1+2 | Vox2-en |
| English | 0.076 | 0.104 | 0.313 | 0.349 |
| German | 0.047 | 0.079 | / | / |
| French | 0.104 | 0.175 | / | / |
| Swedish | 0.196 | 0.155 | / | / |
| Spanish | 0.126 | 0.175 | 0.458 | 0.433 |
| Portuguese | 0.207 | 0.138 | / | / |
| Czech | 0.295 | 0.123 | 0.511 | 0.358 |
| Polish | / | / | 0.622 | 0.481 |
| Slovak | / | / | 0.578 | 0.431 |
| Bulgarian | 0.404 | 0.235 | / | / |
| Croatian | 0.334 | 0.198 | / | / |
| Ukrainian | / | / | 0.574 | 0.428 |
| Russian | 0.359 | 0.185 | 0.549 | 0.404 |
| Turkish | 0.243 | 0.149 | 0.482 | 0.417 |
| Mandarin | 0.530 | 0.250 | 0.459 | 0.332 |
| Japanese | 0.402 | 0.272 | / | / |
| Thai | 0.454 | 0.338 | 0.581 | 0.427 |
| Korean | 0.476 | 0.264 | / | / |
| Lao | / | / | 0.535 | 0.435 |
| Arabic | / | / | 0.422 | 0.348 |
| Farsi | / | / | 0.428 | 0.320 |
| Bengali | / | | 0.518 | 0.502 |
| Punjabi | / | / | 0.539 | 0.479 |
| Urdu | / | / | 0.490 | 0.431 |
| Tamil | / | / | 0.505 | 0.459 |

On both datasets system Vox1+2 shows higher correlation coefficients for languages not included in the training dataset than for languages well represented (English, German, French). This confirms the hypothesis of a language similarity effect for untrained languages. Speaker similarity scores of system Vox2-en correlate less with language similarity on average than those of system Vox1+2, presumably, because the system extracts less language information as it was only trained for English. LDC yields higher correlations than Globalphone, possibly because with LDC there are more untrained languages to compare between and a wider range of both similar and dissimilar languages.

## 4. Conclusion

In this paper, we focussed particularly on the influence of language similarity in non-target cross-lingual speaker verification trials for a state-of-the-art speaker verification system. Results show that in comparisons involving a training language the choice of comparison language has only a small effect on the generated scores (clustering effect). In comparisons among languages the system was not trained on, speaker similarity scores correlate with language similarity (language similarity effect). This effect is more pronounced for a system with multilingual training. Insights like these are important for forensic speaker recognition, which is characterised by sparseness of case-relevant data. Knowing what factors lie behind language-related score shifts can improve score normalisation and also calibration.

## 5. Acknowledgements

This work was partially funded by the Federal Ministry of Education and Research of Germany (BMBF) in the VIKING project (13N16239).

## 6. References

[1] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. A. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the stbu submission for the nist speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072–2084, 2007.

[2] A. Misra and J. H. Hansen, "Modelling and compensation for language mismatch in speaker verification," *Speech Communication*, vol. 96, pp. 58–66, 2018.

[3] T. Nechanský, T. Bořil, A. Houzar, and R. Skarnitzl, "The impact of mismatched recordings on an automatic-speaker-recognition system and human listeners," *Acta Universitatis Carolinae Philologica 1 / Phonetica Pragensia*, pp. 11–22, 2022.

[4] J. Thienpondt, B. Desplanques, and K. Demuynck, "Tackling the score shift in cross-lingual speaker verification by exploiting language information," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7187–7191.

[5] P. Matějka, O. Novotný, O. Plchot, L. Burget, M. D. Sánchez, and J. Černocký, "Analysis of score normalization in multilingual speaker recognition," in *Interspeech*, 2017, pp. 1567–1571.

[6] F. Bahmaninezhad, C. Zhang, and J. H. Hansen, "An investigation of domain adaptation in speaker embedding space for speaker recognition," *Speech Communication*, vol. 129, pp. 7–16, 2021.

[7] D. Sztahó and A. Fejes, "Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings," *Journal of Forensic Sciences*, vol. 68, no. 3, pp. 871–883, 2023.

[8] R. Skarnitzl, M. Asiaee, and M. Nourbakhsh, "Tuning the performance of automatic speaker recognition in different conditions: effects of language and simulated voice disguise," *International Journal of Speech, Language and the Law*, vol. 26, no. 2, p. 209–229, 2020.

[9] D. van der Vloed, M. Jessen, and S. Gfroerer, "Experiments with two forensic automatic speaker comparison systems using reference populations that (mis)match the test language," in *AES International Conference on Audio Forensics*, 2017.

[10] A. Drygajlo, M. Jessen, S. Gfroerer, I. Wagner, J. Vermeulen, and T. Niemi, *Methodological Guidelines for Best Practice in Forensic Semiautomatic and Automatic Speaker Recognition*. Frankfurt: Verlag für Polizeiwissenschaft, 2015. [Online]. Available: https://enfsi.eu/wp-content/uploads/2016/09/guidelines%20_fasr_and_fsasr_0.pdf

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

[12] G. S. Morrison and E. Enzinger, "Multi-laboratory evaluation of forensic voice comparison systems under conditions reflecting those of a real forensic case (forensic_eval_01) – conclusion," *Speech Communication*, vol. 112, pp. 37–39, 2019.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[14] J. Valk and T. Alumäe, "Voxlingua107: A dataset for spoken language recognition," in *IEEE Spoken Language Technology Workshop (SLT)*, 2021, pp. 652–658.

[15] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, 2017, pp. 2616–2620.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[17] B. Desplanques, J. Thienpondt, and K. Demuynck, "Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification," in *Interspeech*, 2020, pp. 3830–3834.

[18] J. Thienpondt, B. Desplanques, and K. Demuynck, "Integrating frequency translational invariance in tdnns and frequency positional information in 2d resnets to enhance speaker verification," in *Interspeech*, 2021, pp. 2302–2306.

[19] Z. Zhao, Z. Li, W. Wang, and P. Zhang, "Pcf: Ecapa-tdnn with progressive channel fusion for speaker verification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[20] Z. Chen, B. Han, X. Xiang, H. Huang, B. Liu, and Y. Qian, "Build a sre challenge system: Lessons from voxsrc 2022 and cnsrc 2022," in *Interspeech*, 2023, pp. 3202–3206.

[21] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5329–5333.

[22] M. Ravanelli, T. Parcollet, P. Plantinga, A. Rouhe, S. Cornell, L. Lugosch, C. Subakan, N. Dawalatabad, A. Heba, J. Zhong, J.-C. Chou, S.-L. Yeh, S.-W. Fu, C.-F. Liao, E. Rastorgueva, F. Grondin, W. Aris, H. Na, Y. Gao, R. D. Mori, and Y. Bengio, "Speechbrain: A general-purpose speech toolkit," 2021.

[23] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[24] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *7th International Conference on Spoken Language Processing (ICSLP 2002)*, 2002, pp. 345–348.

[25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[26] S. Strassel, K. Walker, K. Jones, D. Graff, and C. Cieri, "New resources for recognition of confusable linguistic varieties: the LRE11 corpus," in *The Speaker and Language Recognition Workshop (Odyssey 2012)*, 2012, pp. 202–208.

[27] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.