Cross-lingual Speaker Verification: Evaluation On X-Vector Method

Hareesh Mandalapu[1], Thomas Møller Elbo[1,2], Raghavendra Ramachandra[1], and Christoph Busch[1]

[1] Norwegian University of Science and Technology, Gjøvik, Norway
{hareesh.mandalapu,raghavendra.ramachandra,christoph.busch}@ntnu.no
[2] Technical University of Denmark, Lyngby, Denmark
s144825@student.dtu.dk

Abstract. Automatic Speaker Verification (ASV) systems accuracy is based on the spoken language used in training and enrolling speakers. Language dependency makes voice-based security systems less robust and generalizable to a wide range of applications. In this work, a study on language dependency of a speaker verification system and experiments are performed to benchmark the robustness of the x-vector based techniques to language dependency. Experiments are carried out on a smartphone multi-lingual dataset with 50 subjects containing utterances in four different languages captured in five sessions. We have used two world training datasets, one with only one language and one with multiple languages. Results show that performance is degraded when there is a language mismatch in enrolling and testing. Further, our experimental results indicate that the performance degradation depends on the language present in the word training data.

Keywords: Speaker recognition · Biometrics · Cross-lingual.

1 Introduction

Biometrics characteristics are used to recognize or verify the identity of a person and to provide access to the security sensitive applications. The biometric characteristics are of two different kinds: physical and behavioral. Face, fingerprint, iris are popular physical characteristics that have been in research for many years. Behavioral biometrics are based on the way humans perform certain tasks like speaking and walking. Speaking characteristics of humans are a well-known biometric modality used to perform accurate recognition. Automatic Speaker Verification has been a famous topic in biometric applications for many years now.

The advancement of computational abilities in the recent decades encouraged applications to use biometric algorithms in many fields. Due to he wide variety of users, devices, and applications, many kinds of vulnerabilities and dependencies are evolved in operational biometric systems. The popular vulnerabilities are anomalies in the samples and presentation attacks on the biometric devices.

The dependencies are caused due to data capturing methods, change in devices, aging of the subject, and many more. There are more dependencies on behavioral biometric modalities because the behavior of the subject changes often. In speaker recognition, apart from the capturing conditions like microphone and transmission channel, background noise, the biometric algorithms also depend on the text, language, and emotion which impact the voice sample [8].

Text-dependent speaker recognition has been in use for many years [4]. In these kinds of approaches, the set of words used in testing is a subset of the words used in enrolment. Further, text-independent speaker recognition methods using Gaussian mixture models are introduced [14], and more algorithms were proposed to exclude the dependency caused by the text [6]. Language dependency is another challenging problem that emerged due to multilingual subjects and wide usage of the same biometric algorithm across the world. Language-independent approaches have been proposed on top of text-independent speaker recognition methods [1] by including multiple languages in enrolment. The National Institute of Standards and Technology Speaker Recognition Evaluation (SRE) series has been including multiple languages in their evaluation protocols over the years [3].

In this work, cross-lingual speaker verification is evaluated on a smartphone based dataset with different languages. The objective is to benchmark the performance of the state-of-the-art algorithms when different languages are mismatched in training, enrolling, and testing phases of automatic speaker verification. Thus, the following are the main contributions of this paper:

- Experiments on state-of-the-art methods that use advanced deep neural networks, like x-vector method, to check the language dependency.
- Experiments on multiple languages and multiple session datasets are included in this work.
- The dependency of trained languages used in world training data is evaluated.
- Results and discussions are presented using ISO/IEC standardized metrics for biometric performance [5].

The rest of the paper is organised as follows: Section 1.1 discusses the previous works on cross-lingual speaker recognition approaches and challenges. Section 2 describes the state-of-the-art approaches chosen for our experiments. In Section 3, the multilingual dataset is described, and Section 4, the cross-lingual experiments are presented with results and discussed. Finally, Section 5 concludes the work with the presentation of future work.

1.1 Related Work

The Automatic speaker verification as a biometric modality has emerged into many applications. The initial problems in speaker recognition have leaned over the text-dependency of the speeches in different speaker verification modules.

[3] https://www.nist.gov/itl/iad/mig/speaker-recognition

Later, the language dependency has emerged into a challenging problem in text-independent speaker verification [1]. The early works on language mismatch evaluation are performed by comparing speaker verification with world models trained on only one language and multiple languages. One could observe that when provided with all languages and enough data for world model training, there is no degradation of performance [1]. It is important to note that the enrolled and tested speaker's language are the same in these experiments. Further, the authors have also pointed out the need for new databases from different languages.

Subsequently, the research work focused on bilingual speakers and performed cross-lingual speaker verification. In the investigation of combining the residual phase cepstral coefficients (RPCC) with Mel-frequency cepstral coefficients (MFCC) work from [10], it is observed that RPCC has improved the performance of traditional speaker verification methods. The residual phase characterizes the glottal closure instants better than the linear prediction models like MFCC. The glottal closure instants are known to contain speaker-specific information [3] [12]. Considering the advantages of residual phase and glottal flow, Wang *et al.* [17] proposed a bilingual speaker identification with RPCC and glottal flow cepstral coefficients (GLFCC) as features. The experiments on NIST SRE 2004 corpus, RPCC features show the highest accuracy when compared to MFCC features.

In [9], Mishra *et al.* examined the language mismatch in speaker verification over i-vector system. When all the parameters are kept consistent, and by changing the language, there is performance degradation in EER by 135%. Also, including a phoneme histogram normalization method using a GMM-UBM system improves the EER by 16%. Li et al. [7] have proposed a deep feature learning for cross-lingual speaker verification in comparison with i-vector based method. Two deep neural networks (DNN) based approaches are proposed with the knowledge of phonemes, which is considered as a linguistic factor. The DNN feature with linguistic factor and PLDA scoring shows better performance than i-vector based method and DNN without linguistic factor.

2 X-vector based Speaker Verification system

The X-vector based speaker verification, which is a Deep Neural Network-based approach, proposed by Snyder *et al.* in [15] has the improved performance from data augmentation as suggested in [16]. The model is a feed-forward Deep Neural Network (DNN) which works on cepstral features that are 24-dimensional filter banks and has a frame length of 25 ms with mean-normalization over a sliding window of up to 3 seconds. The model consists of eight layers. The first five layers work on the speech frames, with an added temporal context that is gradually built on through the layers until the last of the five layers. A statistics pooling layer aggregates the outputs and calculates the mean and standard deviation for each input segment. The mean and standard deviation are concatenated and propagated through two segment-level layers and through the last layer, a

softmax output layer. The block diagram of x-vector based automatic verification system is show in Figure 1

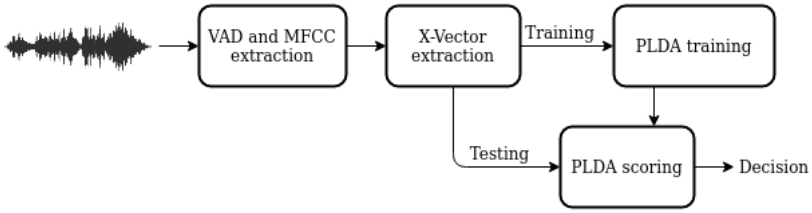


Fig. 1. Block diagram of X-vector based automatic speaker verification system

The x-vector method is used with two pre-trained variants, one trained on the combined dataset of five Switchboard datasets, SRE datasets from 2004 to 2010, and the Mixer 6 dataset and the second one is trained on the VoxCeleb 1 and VoxCeleb 2 datasets. The two models are different in multiple directions including the data capturing mechanism, languages spoken in data and variance in acquisition channels. The pre-trained models have been obtained from the Kaldi webpages namely the SRE16 model from `http://kaldi-asr.org/models/m3`, and the VoxCeleb model from `http://kaldi-asr.org/models/m7`.

2.1 NIST-SRE16 trained model

The NIST-SRE16 pre-trained model uses a total of 15 different datasets, containing a total of 36 different languages. The combined amount of speakers from the Switchboard, SRE, and Mixer datasets totals 91k recordings from over 7k speakers. Data augmentation is done, adding noise and reverberation to the dataset, and combining two augmented copies to the original clean training set. The augmentation of the recording was chosen randomly between four possible types, either augmenting with babble, music, noise, or reverb. Augmenting with babble was done by appending three to seven speakers from the MUSAN speech to the original signal, augmenting with music was selecting a music file randomly from MUSAN, trimmed or repeated to match the duration of the original signal. Noise augmentation was done by adding one-second intervals to the original signal, taken from the MUSAN noises set. Reverb augmentation was done by artificially reverberating via convolution with simulated RIRs.

The SRE16 x-vector model training is employed with with two PLDAs. The first PLDA is trained on the same datasets as the x-vector model trained, but not fitted to the evaluation dataset. As the PLDA is only trained on out-of-domain data, this PLDA is called out-of-domain (OOD) PLDA. The second PLDA (ADT) is fitted to the same datasets and has been adapted to SRE16 data by using the SRE16 major dataset, containing utterances in Cantonese

and Tagalog. Therefore, this PLDA is in-domain adapted (ADT) PLDA. The evaluation set of SRE16 is used to test the trained model. The performance of the x-vector method is observed as equal error rate (EER) of 11.73% with OOD PLDA and 8.57% with ADT PLDA.

2.2 VoxCeleb trained model

The VoxCeleb model used has been trained on the datasets VoxCeleb 1 and VoxCeleb 2 created by Chung *et al.* in [11] and [2], respectively. The development set of VoxCeleb 1 contains over 140k utterances for 1211 speakers, while the VoxCeleb 2 contains over a million utterances for 6112 speakers. All utterances in VoxCeleb1 are in English but VoxCeleb2 contains multiple languages and have been extracted from videos uploaded to YouTube. The training set size has been increased by using Data Augmentation by adding noise and reverberation to the datasets. In the same fashion as done in Section 2.1. The test set of VoxCeleb1 with 40 speakers is used to evaluate the training process and the performance is observed as EER of 3.128%.

3 Smartphone Multilingual Dataset

The SWAN (Secured access over Wide Area Network) dataset [13] is part of the SWAN project funded by The Research Council of Norway. The data has been gathered using an Apple iPhone6S and has been captured at five different sites. Each site has enlisted 50 subjects in six sessions, where eight individual recordings have been recorded. Depending on the capture site, four of the utterances are in either Norwegian, Hindi, or French, while the remaining four are in English. The utterances spoken are predetermined with alphanumerical speeches. The speakers have pronounced the first utterances in English and then in a national language depending on the site.

The six sessions of data capture are present at each site with a time interval of 1 week to 3 weeks between each session. Session 1 and 2 are captured in a controlled environment with no noise. Session 1 is primarily used to create presentation attack instruments. Therefore, we did not use session 1 data in our experiments. Session 3,4 and 6 are captured in a natural noise environment, and session five is captured in a crowded noise environment. In our experiments, we have enrolled session 2 data in all languages, and other sessions data are used for testing. This way, we can understand the session variance and the impact of noise on ASV methods. A sample of single utterance (sentence 2 in English with duration 14 seconds) is presented in Figure 2 indicting the intra-subject variation between different sessions. The Figure 2 shows the utterances of the sentence "My account number is *fake account number*" by the same subject in all sessions.

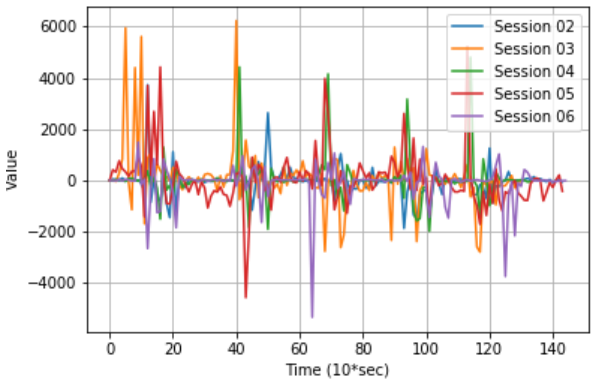


Fig. 2. A sample signal from SWAN dataset from each session.

4 Experiments and Results

We have four different sets of languages in our dataset, where English is the common language in all the sets. Experiments on four sets of different language combinations are performed. Also, we have five sessions of data capturing in each of the sets. We have followed the same protocol among all the sets by enrolling session two samples and using the rest of the sessions data for testing. To study the cross-lingual speaker recognition results, we have enrolled each language separately and tested the other languages present in that set.

The results are presented using the ISO/IEC standardized metrics for biometric performance [5]. Equal error rate (EER) is the error rate at which the false match rate (FMR) and false non-match rate (FNMR) are equal. We have plotted detection error trade-off (DET) curves, which represent the performance of the recognition of the biometric system in terms of FNMR over FMR.

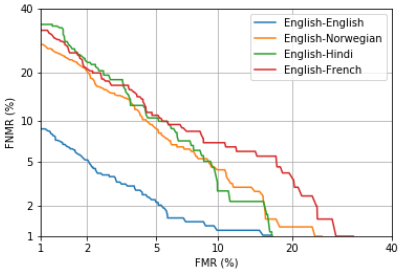
4.1 Experiment 1

The first experiment is carried out on NIST-SRE16 trained model for x-vector extraction and PLDA scoring. This experiment includes two types of PLDA scoring approaches. The first type (OOD PLDA) is an out-of-domain model trained on combined data that contains the Switchboard database, all SREs prior to 2016, and Mixer 6. The second type of PLDA (ADT PLDA) is an in-domain PLDA that is adapted to the SRE16 major partition.

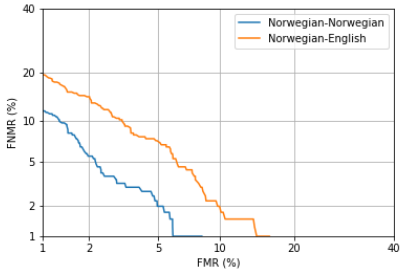
Table 1 represents the cross-lingual speaker recognition with English as the enrolment language in all four sessions. The highest error is highlighted among the block of same enrolled language in each PLDA method. It can be clearly seen that the EER values are lower when the enroll language and test languages are the same compared to different languages in test data. Similar results

Table 1. Results from SRE16-trained X-vector Model with two types of PLDAs and different sessions.

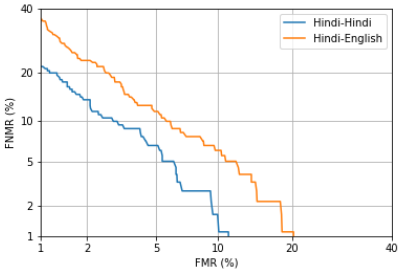
Enrolment language	Test language	S3		S4		S5		S6	
		OOD	ADT	OOD	ADT	OOD	ADT	OOD	ADT
English	English	3.21	3.20	1.65	1.76	4.05	4.15	1.78	1.83
English	Norwegian	6.45	6.65	5.89	5.61	**8.60**	**8.32**	6.16	6.11
English	Hindi	6.83	6.37	5.68	4.96	7.48	7.27	6.33	6.13
English	French	7.76	7.21	5.65	5.08	5.13	4.96	6.13	5.73
Norwegian	Norwegian	3.12	3.21	1.28	1.44	4.98	4.42	1.70	1.77
Norwegian	English	5.56	5.17	3.62	3.42	**8.46**	**7.34**	3.76	2.95
Hindi	Hindi	5.26	4.39	5.01	4.23	4.35	4.46	4.77	4.58
Hindi	English	**7.50**	**7.51**	6.18	5.73	5.45	5.49	5.23	4.72
French	French	5.33	4.32	2.45	2.40	2.62	2.35	1.88	2.06
French	English	6.13	**6.10**	3.41	3.18	**6.44**	5.22	4.63	4.64



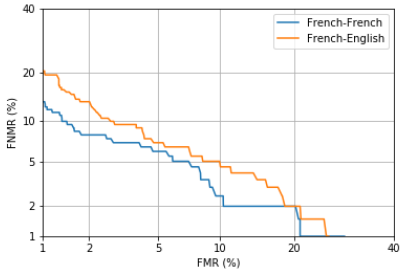
(a) Enrolment: English



(b) Enrolment: Norwegian



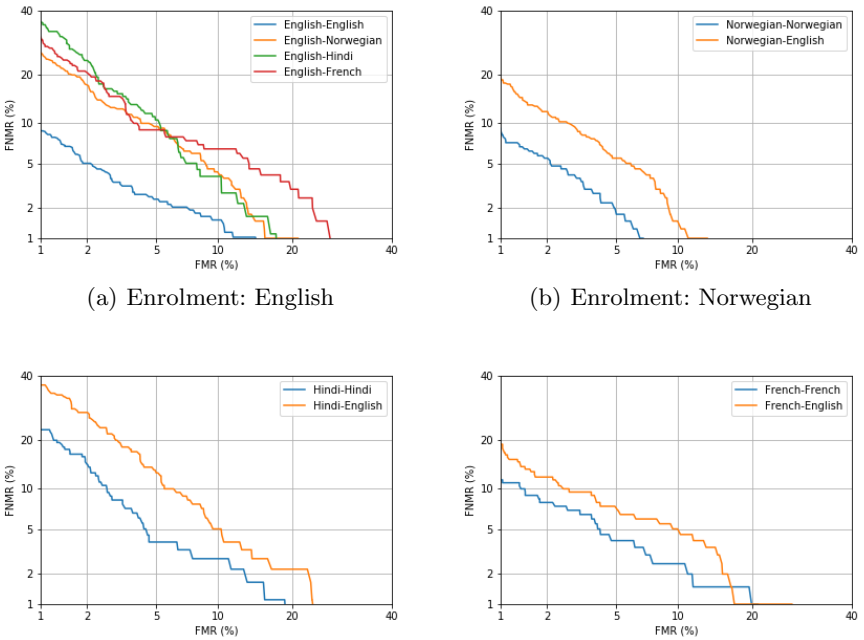
(c) Enrolment: Hindi



(d) Enrolment: French

Fig. 3. DET curves showing the performances of Session 3 with trained model on NIST-SRE16 and out-of-domain adapted PLDA (OOD).

are obtained with Norwegian, Hindi, and French. The highest difference can



(a) Enrolment: English (b) Enrolment: Norwegian

(c) Enrolment: Hindi (d) Enrolment: French

Fig. 4. DET curves showing the performances of Session 3 data and trained on NIST-SRE16 with in-domain adapted PLDA (ADT).

be observed in the case of English-French combination with a degradation in performance of more than 350% on Session 6 data.

Session 5 has displayed the least accuracy in recognizing speakers among all language combinations. The main reason for this problem could be due to the crowded environment of the data captured. The Figures 3 and 4 show the plots of DET curves from different languages used in enrolment and testing from Session 3. The error rates can be clearly seen increasing when cross-lingual speaker recognition is performed.

PLDA adaptation The adaptation of PLDA training does not show a regular trend among different languages. Although the out-of-domain PLDA adaption (OOD) displays higher error rates in many cases, in-domain adapted PLDA (ADT) does not improve the performance for some same-language and cross-language evaluations. In the future works, more experiments on different models of OOD and ADT will be studied along with multiple languages included in the data.

4.2 Experiment 2

VoxCeleb trained model is used in the second experiment. The PLDA used in this model is trained on VoxCeleb1, and Voxceleb2 combined. A similar protocol from Experiment 1 is followed here also but with only one type of PLDA model. Table 2 shows the EER values among different language combination with highest EER value highlighted. The equal error rate is increased in all cases when there is a language mismatch between enrolment and testing. However, it is interesting to observe that the difference in the drop of EER is higher than for Experiment 1.

Figure 5 shows the comparison of DET curves between the same language and cross-language speaker recognition from Session 3 of the dataset. It can be clearly seen that the performance of the system has decreased when language mismatch has happened. The difference between the same language and cross-language is much higher in the VoxCeleb model than that of the NIST-SRE16 trained model.

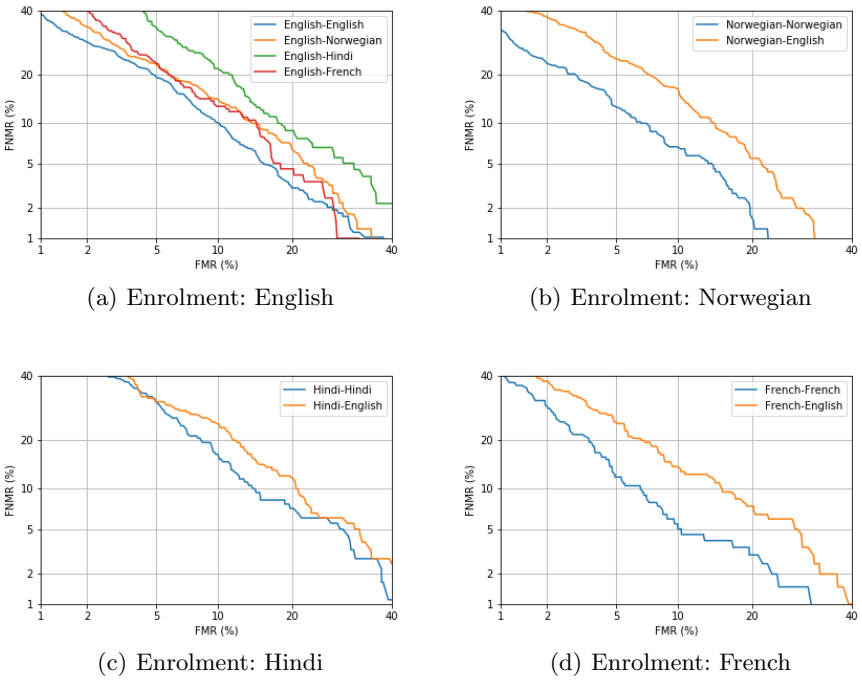
Table 2. Results from VoxCeleb X-vector Model from different sessions.

Enrolment language	Test language	S3	S4	S5	S6
English	English	9.90	7.69	10.01	7.99
English	Norwegian	11.83	10.31	**15.01**	10.48
English	Hindi	13.84	13.12	12.75	12.05
English	French	11.21	9.06	11.28	9.46
Norwegian	Norwegian	8.04	6.44	10.91	6.74
Norwegian	English	11.92	9.32	**13.71**	9.55
Hindi	Hindi	12.16	10.68	11.88	10.66
Hindi	English	**14.77**	11.70	13.11	12.72
French	French	7.64	6.58	8.29	6.94
French	English	**11.83**	9.71	8.57	9.41

The speaker recognition accuracy is consistently lower than for the NIST-SRE16 trained model in all the cases. The reason for this could be that the world training dataset in the NIST-SRE16 model contains multiple languages which attributes for cross-lingual speaker recognition robustness. On the other hand, the VoxCeleb2 dataset contains multiple languages, there is a huge variance in data and bias in the number samples per subject which could be reason that limits the ability of the system to recognize different languages in enrolling and testing.

5 Conclusion

Behavioral biometric recognition methods have multiple dependencies due to high intra-class variation caused by environmental factors and the human fac-



(a) Enrolment: English (b) Enrolment: Norwegian

(c) Enrolment: Hindi (d) Enrolment: French

Fig. 5. DET curves showing the performances of Session 3 data and trained on VoxCeleb data.

tors impacting the capture process. In the speaker recognition community, dependencies of samples like the text used in the speech and language in which speech is delivered needs to be investigated. The dependency due to language has been a problem when there is a mismatch between enrolment and tested language. In this work, we have focused on evaluating the problem of language mismatch on the state-of-the-art speaker recognition method, namely the x-vector method, which uses a deep neural network-based approach. We have chosen a multilingual dataset with four different languages and four different sessions. For the world training dataset, we included two popular publicly available datasets NIST-SRE16 and VoxCeleb.

The experiments on cross-lingual speaker recognition displayed the performance degradation when there is a mismatch in languages in enrolment and testing. Further, the dependency on the languages included in the world training dataset is observed. If there are multiple languages used in the world training dataset, which is the case of NIST-SRE16, performance degradation is less compared to the one language model VoxCeleb. In future works, a speaker recognition approach is implemented to overcome the problem of language dependency.

References

1. Auckenthaler, R., Carey, M.J., Mason, J.S.: Language dependency in text-independent speaker verification. In: 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221). vol. 1, pp. 441–444. IEEE (2001)
2. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. In: INTERSPEECH (2018)
3. Gupta, C.S.: Significance of source features for speaker recognition. Master's thesis, Indian Institute of Technology Madras, Dept. of Computer Science and Engg., Chennai, India (2003)
4. Hébert, M.: Text-dependent speaker recognition. In: Springer handbook of speech processing, pp. 743–762. Springer (2008)
5. ISO/IEC JTC1 SC37 Biometrics: ISO/IEC 19795-1:2006. Information Technology – Biometric Performance Testing and Reporting – Part 1: Principles and Framework. International Organization for Standardization and International Electrotechnical Committee (March 2006)
6. Kinnunen, T., Li, H.: An overview of text-independent speaker recognition: From features to supervectors. Speech communication **52**(1), 12–40 (2010)
7. Li, L., Wang, D., Rozi, A., Zheng, T.F.: Cross-lingual speaker verification with deep feature learning. In: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). pp. 1040–1044. IEEE (2017)
8. Lu, X., Dang, J.: An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. Speech communication **50**(4), 312–322 (2008)
9. Misra, A., Hansen, J.H.L.: Spoken language mismatch in speaker verification: An investigation with nist-sre and crss bi-ling corpora. In: 2014 IEEE Spoken Language Technology Workshop (SLT). pp. 372–377 (2014)
10. Murty, K.S.R., Yegnanarayana, B.: Combining evidence from residual phase and mfcc features for speaker recognition. IEEE Signal Processing Letters **13**(1), 52–55 (2006)
11. Nagrani, A., Chung, J.S., Zisserman, A.: Voxceleb: A large-scale speaker identification dataset. In: Lacerda, F. (ed.) Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017. pp. 2616–2620. ISCA (2017). https://doi.org/10.21437/Interspeech.2017, `http://www.isca-speech.org/archive/Interspeech_2017/abstracts/0950.html`
12. Plumpe, M.D., Quatieri, T.F., Reynolds, D.A.: Modeling of the glottal flow derivative waveform with application to speaker identification. IEEE Transactions on Speech and Audio Processing **7**(5), 569–586 (1999)
13. Raghavendra, R., Stokkenes, M., Mohammadi, A., Venkatesh, S., Raja, K., Wasnik, P., Poiret, E., Marcel, S., Busch, C.: Smartphone multi-modal biometric authentication: Database and evaluation. arXiv preprint arXiv:1912.02487 (2019)
14. Reynolds, D.A., Rose, R.C.: Robust text-independent speaker identification using gaussian mixture speaker models. IEEE transactions on speech and audio processing **3**(1), 72–83 (1995)
15. Snyder, D., Garcia-Romero, D., Povey, D., Khudanpur, S.: Deep neural network embeddings for text-independent speaker verification. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. vol. 2017-August, pp. 999–1003 (2017), `www.scopus.com`, cited By :202

16. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings. vol. 2018-April, pp. 5329–5333 (2018), www.scopus.com, cited By :270
17. Wang, J., Johnson, M.T.: Vocal source features for bilingual speaker identification. In: 2013 IEEE China Summit and International Conference on Signal and Information Processing. pp. 170–173 (2013)