Investigating Language Variability on the Performance of Speaker Verification Systems

Amir Vaheb[1], Ali Janalizadeh Choobbasti[1], S. H. E. Mortazavi Najafabadi[1], and Saeid Safavi[2]

[1] Miras Technologies International, NO. 92, Movahed Danesh St.,
Farmanieh, Tehran, Iran
{amir, ali, hani}@miras-tech.com
http://miras-tech.com/

[2] Centre for Vision, Speech and Signal Processing, University of Surrey,
Guildford GU2 7XH, UK
s.safavi@surrey.ac.uk
https://www.surrey.ac.uk/centre-vision-speech-signal-processing

Abstract. In recent years, speaker verification technologies have received an extensive amount of attention. Designing and developing machines that could communicate with humans is believed to be one of the primary motivations behind such developments. Speaker verification technologies apply to numerous fields such as security, Biometrics, and forensics.

In this paper, the authors study the effects of different languages on the performance of the automatic speaker verification (ASV) system. The corpus used in this study is the MirasVoice speech corpus (MVSC). This corpus is a bilingual English and Farsi speech corpus. This study collects results from both an I-vector based ASV system and a GMM-UBM based ASV system. The experimental results show that a mismatch between the enrolled data used for training and verification data can lead to a significant decrease in overall system efficiency. This study shows that it is best to use an i-vector based framework with data from the English language used in the enrollment phase to improve the robustness of the ASV systems. Results collected in this study indicate that this can narrow the degradation gap caused by the language mismatch.

Keywords: speaker verification, bilingual speech corpus, Gaussian mixture model, i-vector system

1 Introduction

In the past few years, intense focus has been spent on speaker verification applications for bilingual and multilingual environments, where the training and testing materials have been taken from different languages. Speaker verification is a process in which the users claimed identity is verified based on their input voice sample, and thus involves a binary decision to see whether the test audio sample matches the voice template of the claimed speaker[21]. In the speaker verification trial, there is an identity claim asserted or provided along with the audio sample. In this case, the unknown audio sample is then analyzed and correlated with the speaker model whose label corresponds to the identity claimed by the speaker[19].

This technology plays a crucial role in Biometrics authentication and securing the computation of human-central computer interfaces[16]. Since the speaker's speech can be easily obtained over the course of cross-media human-computer interactions, this technology offers a non-intrusive means of safety and security with high utility. Although the research and development of automatic speaker verification (ASV) systems in late years have improved their accuracy and performance to the point of mass-market deployment; however, because of advances in a noise and channel compensation approaches, these systems can be susceptible to attacks such as spoofing. Voice samples from users contain information related to the environment, the spoken language, and different traits like emotional state, accent, and their vocal style [3]. This information forms the different aspects of the user's acoustic space [6]. The model learns to estimate using these aspects; thus any discrepancy in these aspects in the train and test sets would result in a degraded speaker recognition performance. The I-vector system and Gaussian mixture model-universal background model (GMM-UBM) are among the most common approaches in the design of ASV systems.

This work focuses on developing a bilingual (English and Farsi) speaker verification system. Bilingual and multilingual speaker recognition systems are considered to play a key role in the development of Biometric security systems that can function in bilingual and multilingual environments [4]. In this paper, we show how the performance of such systems degrade when the train and test sets consist of audio features from different languages. In this study, ASV systems are designed and developed using two different frameworks; I-vector and GMM-UBM. The obtained performances are reported when they are trained and tested on (i) Farsi audio data only; (ii) English audio data only; and (iii) combined Farsi and English audio data.

Speaker verification systems based on i-vector framework have grown into the state-of-the-art in the field of speaker verification over the past few years[6]. In this method, a new set of features, named i-vectors, are extracted by projecting GMM mean super-vector to a lower dimensional subspace (total variability space). The approach was originally inspired by Joint Factor Analysis (JFA) technique contained valuable speaker-discriminant information[13]. The i-vector framework was originally proposed by Dehak et al.[7] to define only one space, as opposed to the multiple spaces proposed in JFA approach. [10]

In the remainder of this article structured as follows: Section 2 talks about the related works done in this field. Section 3 describes the database used in this research while section 4 describes the approach taken towards building the ASV systems. Section 5 will present the results of different approaches and experiments. Finally, the paper summarizes in section 6.

2 Related works

While some work has been done on bilingual and multilingual speaker verification, most of the work already done in this field use monolingual databases. The work accomplished in [15] is on a bilingual (Mandarin-English) speech corpus and has shown that a language dissimilarity between the target speaker and trained model can lead to a considerable degradation in ASV performance. [5] reports that the best performance is obtained when using English audio in both training and testing phases. In [17], an automatic speaker verification system is displayed that has been trained on both Chinese and English sentences/digits that shows a smaller degrade in performance when the system tested on English versus Chinese. The results achieved in [11] also displays the effect of using different languages for training and testing phase. In [11], the corpus consisted of audio from Spanish, Tamil and Mandarin speakers. They also show that GMM-UBM models trained on a specific language tend to generalize better than GMM-UBM models trained using multiple languages. In this work, the authors present the results of both an i-vector and a GMM-UBM model trained and tested on all possible combinations.

In [18] authors have been used the SpeechDat multilingual speech database, which is composed of 21 different languages. This database has been used for the task of isolated speaker verification. The authors in [2] use a multilingual speech corpus composed of English, Hindi and a local north-eastern language in India called Arunachal Pradesh from 100 different speakers native in Arunachali languages. They show that when the global speech model is trained with more than one language, the overall performance of the speaker verification system degrades. The work done in [1] also shows how the performance of a GMM-UBM automatic speaker verification system can degrade dramatically due to a mismatch between model and target languages.

In this study, the authors have collected **33** hours of audio data from **50** individuals that are native Farsi speakers but also fluent in English. Approximately **40** minutes of audio data exist per speaker. 20 minutes of this audio is in English and the other 20 minutes is in Farsi.

3 Corpus Description

The authors in this study have collected one of the largest Farsi-English speech corpora available [1]. 33 hours of speech data from 50 individuals have been collected in this corpus making MirasVoice speech corpus (MVSC) an ideal database for our experiment. There are approximately 40 minutes of audio data available per speaker in this corpus. The audio data has been recorded using a microphone with a sample rate of 48kHz, a frequency response of 20Hz to 20kHz, a max Sound Pressure Level (SPL) of 120db and a bit rate of 16 bits. The audio data has been stored as Waveform. The speech material in this corpus consists of both read and spontaneous audio data. Recorded audio data is collected from native Iranian

[1] http://github.com/miras-tech/MirasVoice

speakers with different accents from different provinces who can speak English fluently. The database also includes meta-data such as a person weight, height, age, blood pressure, whether they smoked or not, birthplace(province), mothers birthplace(province), fathers birthplace(province), and the province they grew up in. The MVSC was initially collected for evaluation of the speaker verification systems, but it can be used in other tasks, as it contains useful labels for each speaker in the database.

4 The Automatic Speaker Verification System

In this section, we are going to explain the system components which have been used during our experiments. Figures 1 & 2 show the overall block diagram of the two speaker verification engines which have been used in this study. The following subsections introduce the details of each system individually.

4.1 Signal analysis

The feature extraction was performed as follows. An energy-based Speech Activity Detector (SAD) was used to discard periods of silence. In the next step, the speech is split into 20ms frames with a 10ms overlap using a Hamming window. The next step involves applying an FFT to obtain the short-time magnitude spectrum which is then passed along to a bank of 30 Mel-spaces triangular band-pass filters. These filters span the frequency region from 0Hz to 44kHz.

4.2 The GMM-UBM system

One of the AVS systems developed in this study is based on the Gaussian Mixture Model - Universal Background Model (GMM-UBM) method [20]. As shown in 1, a weighted sum of multiple Gaussian distributions is used to represent the feature vectors of the parameterization section. Each of the Gaussian distributions has its mean, weight, and covariances. To train the background model, the authors of this study used all the available conversational audio data stored as wave files that were available in the MVSC (except the files which were chosen for testing). Finally, by adopting the means of the universal background model with respect to the class dependent training data, speaker dependent models are trained.

4.3 The I-vector system

As mentioned earlier, i-vector-based systems have become the state-of-the-art approach for speaker recognition [10]. In this approach, the MFCC features are first extracted from the input signal to create an i-vector representation of a signal. Subsequently, the features are used to extract the Baum-Welch statistics of the signal. The i-vector is then computed using the Baum-Welch statistics.

In this study, the i-vectors are used as new low dimensional features extracted from the high dimensional mean supervectors. With the assumption that the

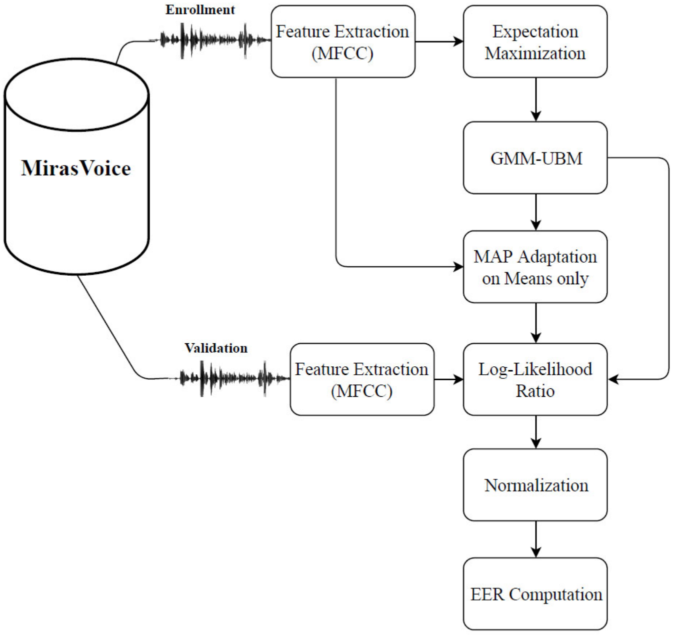

Fig. 1. Block diagram of the GMM-UBM system

GMM super-vector μ can be decomposed into $\mu = m + Tw$, where w is the i-vector sampled from a standard normal distribution, and the T matrix is used as the linear mapping from a low-dimensional total variability subspace w into a super-vector space, and m is the UBM mean super-vector trained over the enrolled audio data. To optimize the total variability subspace T and later extracting the i-vector in this study, the method described by [22] is used.

4.4 PLDA scoring

There are several scoring approaches proposed for i-vector framework, in this study the probabilistic linear discriminant analysis (PLDA) method was employed. Before applying PLDA and computing the verification scores, the i-vectors were first internally mean and length-normalized and also whitened[9]. The verification scores are computed based on the log-likelihood ratio between the different versus same speaker model hypotheses as explained in [8] and [12].

5 Experimental results

The experimental verification scheme in this study was applied similarly to the methodology used for the NIST speaker recognition evaluations. To evaluate the

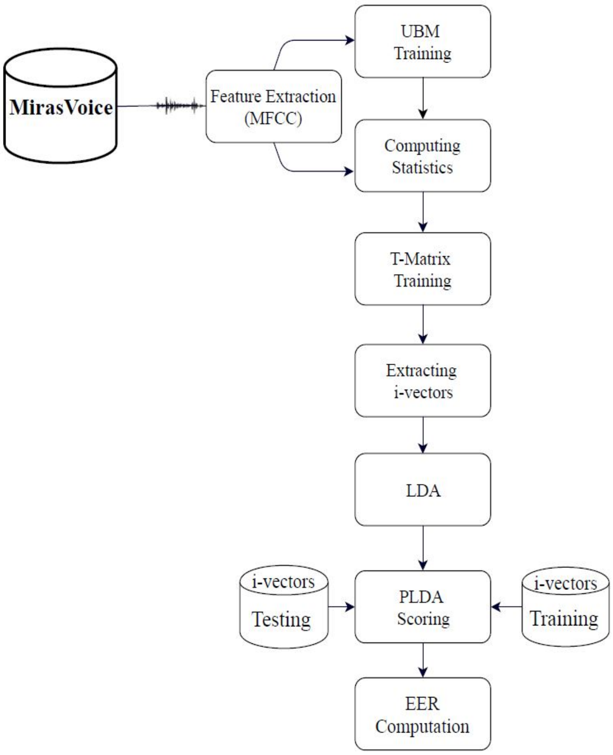

Fig. 2. Block diagram of the i-vector system

speaker models, each test sample was scored against the target speaker model and ten other imposter speaker models. The final results of this study were calculated using the standard NIST software (DetWare) and performance measure used is the percentage Equal Error Rate (EER%). Since the EER is indicative of the error rate, the smaller it shows, the better performance of the ASV system [14].

Nine different experiments were conducted in this study. Audio samples from volunteers speaking both Farsi and English were used with different setups for the enrollment step. Speaker verification models were then trained in different manners for both languages. A similar technique was used in the verification step. For each individual, **70%** of the data, equivalent to **1382** minutes of audio, was used for training and the remaining **605** minutes of audio which made up **30%** of the dataset was used for the verification step (shown in table 1). The evaluation was conducted using the following nine setups (shown in table 2):

Index	Languages	Training Data (min)	Test Data (min)
1	English	717	301
2	Farsi	665	304
3	English + Farsi	1382	605

Table 1. Amount of training and testing data

En-En Performance of the ASV system enrolled with the English audio files and verified using English audio files by the means of the English Background Model.

Fa-Fa Performance of the ASV system enrolled with the Farsi audio files and verified using Farsi audio files, utilizing the Farsi Background Model.

En-Fa Performance of the ASV system enrolled with the English audio files and verified using Farsi audio files, utilizing the English Background Model.

Fa-En Performance of the ASV system enrolled with the Farsi audio files and verified using English audio files, utilizing the Farsi Background Model.

FaEn-En Performance of the ASV system enrolled with both the Farsi and English audio files and verified using English audio files, utilizing the Farsi-English Background Model.

FaEn-Fa Performance of the ASV system enrolled with both the Farsi and English audio files and verified using Farsi audio files, utilizing the Farsi-English Background Model.

En-FaEn Performance of the ASV system enrolled with English audio files and verified using both English and Farsi audio files, utilizing the English Background Model.

Fa-FaEn Performance of the ASV system enrolled with Farsi audio files and verified using both English and Farsi audio files, utilizing the Farsi Background Model.

FaEn-FaEn Performance of the ASV system enrolled with both the Farsi and English audio files and verified using both English and Farsi audio files, utilizing the Farsi-English Background Model.

GMM-UBM and i-vector-PLDA are two ASV systems, which this study has used to conduct nine experiments. As observed from table 2 we can conclude that in each experiment, the i-vector-PLDA has resulted in better performance compared with the GMM-UBM approach. However, both algorithms have expressed similar behavioral pattern in every experiment, for example, the best performing setup is when for both training and testing the English recording were used and whenever there is a language miss-match between the training and the testing material exists we have faced with the performance degradation.

Results show that the best performance was obtained with the experiment in which the English audio samples were employed in both the training and validation stages. This demonstrates that the system is more accurate in English which may be because most of the state of the art methods are initially designed and calibrated for English recordings.

Index	Enrollment - Validation Phases	GMM-UBM EER(%)	I-Vector EER(%)
1	English - English	3.23	1.15
2	Farsi - Farsi	4.17	1.99
3	English - Farsi	6.41	4.47
4	Farsi-English	4.57	2.31
5	(English + Farsi) - English	3.68	1.78
6	(English + Farsi) - Farsi	4.27	2.12
7	English - (English + Farsi)	5.61	3.29
8	Farsi - (English + Farsi)	6.89	4.92
9	(English + Farsi) - (English + Farsi)	3.45	1.49

Table 2. Speaker verification performances in terms of EER for different experimental setup

The results also demonstrate how having a mismatch between the languages used in the enrollment and validation steps, increases the EER. Another exciting fact expressed in these experiments and shown in figure 3 is that for the miss matched conditions the model trained using data from both languages (English & Farsi), is outperformed by the model that is trained using audio samples from the English language alone. On the other hand, we have shown that for language miss-matched condition by combining speech materials from both languages in our training we can degrade the language miss-match and accommodate more information for the speaker dependent models and subsequently for the verification process.

6 Conclusion

This paper presents systematic research on language mismatch for the application of automatic speaker verification systems. The experimental results suggest that having different languages in the training and validation steps can play a significant role in the performance of the ASV systems. Based on the results, we can conclude that the system's performance is highly language-dependent. In the training stage of the ASV models, it is advised that even the non-native English speaking community use the English language. The reason for this is that the model will be more robust to their language and the user experience of the system would be higher.

In this research, two ASV systems were used, GMM-UBM and i-vector. Obtained results suggest that the i-vector-PLDA approach has better performance overall, especially in the case of language variability between the training and testing sets.

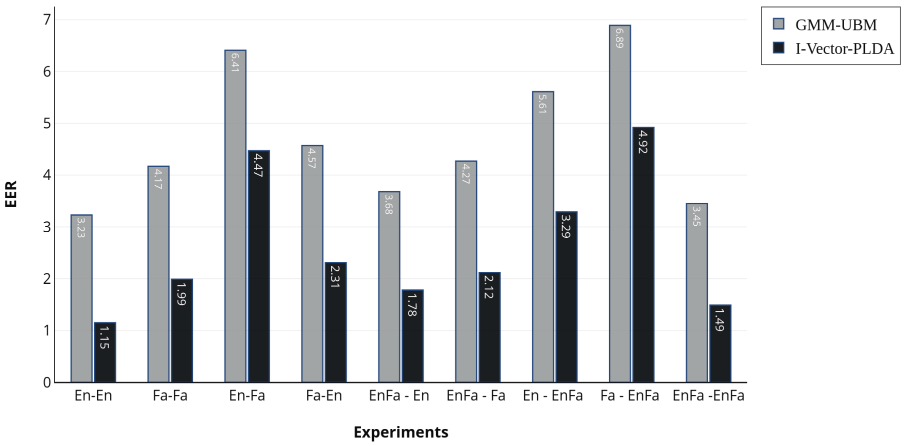

Fig. 3. Bar Chart Plot for Comparing obtained performances using GMM-UBM and I-vector Approaches for 9 different conditions

7 Acknowledgement

The authors of this study would like to thank all the volunteers and especially the people at Miras Technologies international [1] for all their efforts to help gather the MirasVoice speech corpus[5].

References

1. Bhattacharjee, U. and Sarmah, K.: GMM-UBM based speaker verification in multilingual environments. In: IJCSI International Journal of Computer Science Issues, 9(6), p.2. (2012)
2. Bhattacharjee, Utpal, and Kshirod Sarmah.: A multilingual speech database for speaker recognition. In: Signal Processing, Computing and Control (ISPCC), 2012 IEEE International Conference on, pp. 1–5. IEEE (2012)
3. Bipul Pandey, Alok Ranjan, Rajeev Kumar and Anupam Shukla: Multilingual Speaker Recognition Using ANFIS. in: 2nd International Conference on Signal Processing Systems (ICSPS) vol 3, V3–714 - V3–718 (2010)
4. Deng, L., Droppo, J., Yu, D., & Acero, A.: Learning Methods in Multilingual Speech Recognition. Speech Research Group Microsoft Research Redmond, WA, 98052.
5. A. Vaheb, A. J. Choobbasti, S.H.E. Mortazavi Najafabadi, S. Safavi: MirasVoice: A bilingual (English-Farsi) speech corpus, in: In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki (Japan) (2018)

[1] Miras Technologies International, http://miras-tech.com/

6. A. Misra and J. Hansen: Spoken language mismatch in speaker verification: an investigation with NIST-SRE and CRSS Bi-ling Corpora, in: IEEE The Spoken Language Technology Workshop (SLT) (2014)
7. N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet, and P. Dumouchel: Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In: Proceedings of Interspeech, p. 1559 1562, (2009)
8. Garcia-Romero, Daniel, and Carol Y. Espy-Wilson: Analysis of i-vector length normalization in speaker recognition systems. In: Twelfth Annual Conference of the International Speech Communication Association. (2011)
9. Saeid Safavi: Speaker characterization using adult and children's speech. In: Ph. D. dissertation, University of Birmingham. (2015)
10. N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet: Front-end factor analysis for speaker verification. In: Audio, Speech, and Language Processing, IEEE Transactions on, vol. 19, no. 4, pp. 788798, (2011)
11. Kleynhans, Neil T., and Etienne Barnard.: Language dependence in multilingual speaker verification. In: PRASA (2005)
12. Saeid Safavi, Abualsoud Hanani, Martin Russell, Peter Jancovic, and Michael Carey: Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification. In: IEEE Signal Processing Letters, Vol 19 (12), pp. 829-832. (2012)
13. Kenny, Patrick, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel: A study of interspeaker variability in speaker verification. In: IEEE Transactions on Audio, Speech, and Language Processing 16, no. 5 980–988. (2008)
14. Saeid Safavi, Martin Russell, and Peter Jancovic: Identification of Age-Group from Children's Speech by Computers and Humans. In: INTERSPEECH, 243-247. (2014)
15. R. Auckenthaler, M. J. Carey, and J. S. D. Mason: Language Dependency In Text-Independent Speaker Verification, ICASSP 2001, May (2001)
16. Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, Haizhou Lia: Spoofing and countermeasures for speaker verification: a survey. Elsevier 66, 130–153 (2015)
17. X. Qing and K. Chen.: On use of GMM for multilingual speaker verification: An empirical study. In: Proceedings of ISCSLP, pages 263-266 (2000)
18. Hge, Harald, Christoph Draxler, Henk van den Heuvel, Finn Tore Johansen, E. P. Sanders, and Herbert S. Tropf.: Speechdat multilingual speech databases for teleservices: across the finish line. (1999)
19. Stuker, S. Schultz, T. Metze, F. Waibel, A.: Multilingual articulatory features: Acoustics, Speech, and Signal Processing, Proceedings. In: (ICASSP '03). IEEE International Conference 7 (2003)
20. Saeid Safavi, Martin Russell, and Peter Jancovic: Automatic speaker, age-group and gender identification from children's speech. In: Computer Speech and Language 50. pp. 141–156 (2018)
21. Campell J.P. and Jr.: Speaker recognition: a tutorial , Proceeding of the IEEE, Vol 85, pp. 1437–1462 (1997)
22. Saeid Safavi, and Li Meng: Comparison of two scoring method within i-vector framework for speaker recognition from childrens speech. In: ICMI Workshop on Child Computer Interaction (WOCCI), Glasgow, Scotland, November (2017)