MirasVoice: A bilingual (English-Farsi) speech corpus

Amir Vaheb[1], Ali Janalizadeh Choobbasti[1], Behnam Sabeti[1],
S.H.E. Mortazavi Najafabadi[1], Saeid Safavi[2]
[1]Miras Technologies International
NO. 3, 2nd Alley, North Sheikh Bahai St., Tehran, Iran
{Amir, Ali, Behnam, Hani}@miras-tech.com,
[2] School of Engineering and Technology, University of Hertfordshire
College Lane Campus, Hatfield AL10 8PE, Hertfordshire, UK
s.safavi@herts.ac.uk

Abstract

Speech and speaker recognition is one of the most important research and development areas and has received quite a lot of attention in recent years. The desire to produce a natural form of communication between humans and machines can be considered the motivating factor behind such developments. Speech has the potential to influence numerous fields of research and development.

In this paper, MirasVoice which is a bilingual (English-Farsi) speech corpus is presented. Over 50 native Iranian speakers who were able to speak in both the Farsi and English languages have volunteered to help create this bilingual corpus. The volunteers read text documents and then had to answer questions spontaneously in both English and Farsi. The text-independent GMM-UBM speaker verification engine was designed in this study for validating and exploring the performance of this corpus. This multilingual speech corpus could be used in a variety of language dependent and independent applications. For example, it can be used to investigate the effects of different languages (Farsi and English) on the performance of speaker verification systems. The authors of this paper have also investigated speaker verification systems performances when using different train/test architectures.

Keywords: Speech Processing, Automatic Speaker Verification, Speech Corpus

1. Introduction

Automatic speaker verification (ASV) systems architecture offers a flexible and low-cost solution for biometric authentication. Although the research and development of ASV systems in recent years have improved their performance to the point of mass-market deployment; due to advances in noise and channel compensation techniques, these systems can be concerned vulnerable to spoofing (Wu et al., 2015). In recent years a number of counter spoofing techniques have been proposed. The reason behind this attention is the development of robust ASV systems for biometric authentication which has many applications in the security sector. This problem has been widely studied for English speakers, but not for Iranians or users that know multiple languages. There are several audio corpora for dominant languages like English. For instance, the Santa Barbara corpus of spoken American English (Bois et al., 2000 2005) which consists of 249,000 words spoken with the transcriptions. Another example is the Callhome American English Speech corpus developed by the Linguistic Data Consortium (LDC) (Canavan et al., 1997) which is made up of 120 unscripted 30-minute long on the phone conversations that were made in North America. There are also some purely Farsi speech corpora, but none of them are bilingual. For example, (Bijankhan et al., 1994) speech corpus containing recordings of 300 native Farsi speakers from 10 different dialect regions in Iran.

In this study, we present MirasVoice which is a bilingual audio corpus in Farsi and English. MirasVoice contains high quality recorded content from 50 speakers (27 male, 23 female), 40 of which are currently labeled. There is both read and spontaneous audio in both Farsi and English. The speakers read 3 text documents in both English and Farsi and then had to answer 17 questions in both languages. The participants were all native Iranian speakers who were educated in English. MirasVoice contains more than 33 hours of audio and can be used for a variety of audio and signal processing applications like audio speaker recognition, gender recognition and in general pattern recognition problems.

In this study, the effects of different languages (Farsi and English) on speaker verification systems was investigated. The attained results show that the best performance is obtained when the language used for both training and test phases are the same.

In the remainder of this paper, Section 2 provides more details about MirasVoice. Validation of the corpus and the experimental results are presented in Section 3. Section 4 concludes the paper.

2. Corpus Description

The MirasVoice Speech Corpus (MVSC) is one of the largest Farsi-English voice datasets currently available for general purpose studies and expert-system development. Some of the applications this dataset can be used for is for speaker recognition systems, speech recognition studies, gender recognition, cognitive science, and pattern recognition. This dataset is expected to grow larger. It currently consists of 50 individuals speaking 2 languages on 4 different texts. The convent of the read text is explained in the next section.

2.1. Speech Materials

The MVSC consists of both read speech and spontaneous speech materials. The text material read by the volunteers

Context of Text Material	Material Amount
Word	250
Sentence	63
Number	80
Question	17

Table 1: Corpus reading material information.

is available alongside the dataset. The text includes a number of words, sentences, and numbers in English. This text has been translates into Farsi by educated native speakers. This translated text has also been read by all participants. As shown in table 1, There are 250 words, 63 sentences and 80 numbers in the text. The numbers start off as easy and progress on to more complicated numbers. We also had 17 questions which we asked each participants which they answered spontaneously. We also gathered information on whether the participants smoked or not, their blood pressure, age, height, accent, birth country, mothers birthplace (province), fathers birthplace (province), time of recording and which province they grew up in. The voice files are stored in 1 minute long .wav files. there is also one 7 minute long .wav audio file in which the participant is answering the questions.

2.2. Labeling

Labeling of the dataset has been done by the authors themselves. The audio files labeled in a special manner. The generic form of the labels is LLXXXTNN. the LL stands for the Language in which the speaker is speaking and is either EN (English) or FA (Farsi/Persian). The second part XXX stands for the index the person has in the overall information .csv file. The third part represents the text the participant is reading in this file which can have 4 different character values W, S, N, Q. The characters respectively stand for words, sentences, numbers, and questions. The last part of the name, NN stands for the file number for that particular text (e.g. 03 means the third audio file). For example, the name EN002S03 means the third English audio file recorded from the second participant that was reading the sentences text. Participants name have not been shared for privacy reasons.

2.3. Recording Procedure

The recordings for the MVSC have mostly been carried out in the conference room at Miras Technologies International central office using a large microphone and stored directly onto a PC in Wave file format using a high quality microphone with a sample rate of 48kHz, a bit rate of 16 bits, a frequency response of 20Hz to 20kHz and a max SPL of 120db. Before each recording, the participants would first read out a text while the recording settings were being adjusted. Since the authors of the dataset didn't want the file lengths becoming too long, they would pause the recording when an audio file reached the threshold of 1 minute and would ask the participants to continue on a separate recording.

For the questions, the authors gave the participants a sheet containing the questions and would ask them to read the questions and answer them respectively in one recording. The length of the questions audio file is 7 minutes long. An initial plan was to collect approximately 40 minutes of recording per each speaker.

2.4. Filing System

In the dataset repository, the audio file containing male and female participants have been separated into different directories. Each directory contains a number of sub-directories indexed by the participants' index located in the .csv file at the root directory of the repository. The starting index for male participants is 001 and ends at 020, and for females, it starts from 021 and goes up to 040.

2.5. Tools

The microphone used for the recordings was a model Yeti microphone from Blue. The microphones pattern setting was set to Cardioid mode for all speakers. Cardioid mode records audio sources that are directly in front of the microphone, delivering rich, full-bodied audio. The authors also measured the speakers' blood pressure using the Beurer wrist blood pressure monitor model BC 40. We recorded the speakers systolic and diastolic blood pressure.

3. Corpus Application

MVSC is a bilingual speech corpus and could be used in speaker verification problems. This study investigates the effect of different languages on the performance of speaker verification systems. A schematic of the process of speaker verification is shown in figure 2.

3.1. Speaker Verification System

Speaker Verification is a branch of bio-metric authentication and means acceptance or refusal of speaker's claim as one of the known users of the system. During the above process, one's speech information will be compared with a speech model analogous with the claimed identification; the results would be acceptance or refusal of the speaker's claim. If the match is above a predefined threshold, the identity is accepted, otherwise, it is rejected. Both speaker identification and speaker verification tasks could be divided further into text-dependent and text-independent categories, based on a number of constraints on the contents of the test and train utterances. While in text-dependent, speakers speak the same text on both training and test phases, in the text-independent phase voice samples should be different (Mporas et al., 2016) (Singh et al., 2012).
This study uses Text-Independent Speaker Verification System (TISV) to show one of the applications of MVSC.

3.1.1. Signal Analysis

Feature extraction was performed as follows. Periods of silence were discarded using an energy-based Speech Activity Detector (SAD). The speech was then segmented into 20-ms frames (10-ms overlap) and a Hamming window was applied. The short-time magnitude spectrum, obtained by applying an FFT, is passed to a bank of 30 Mel-spaced triangular band-pass filters, spanning the frequency region from 0Hz to 44000Hz.

3.1.2. GMM-UBM System

Automatic speaker verification engine used in this study is based on the Gaussian Mixture Model - Universal Background Model (GMM-UBM) method (Reynolds et al., 2000). In this approach, the feature vectors of the parameterization section could be presented as a weighted sum of multiple Gaussian distributions, which is called the GMM. Each single Gaussian distribution has its own mean, weight, and covariance. We used all the recorded conversations wave files for building the background model. At the end class dependent speaker models are built by map adapting the means of UBM with respect to the class dependent enrollment data.

3.1.3. Verification Experiments

Verification experiments were conducted using a similar version of the methodology developed for the NIST speaker recognition evaluations. Each test utterance was scored against the ”true” (correct) speaker model and 10 other ”impostor” models. Results are presented in terms of percentage Equal Error Rate (EER), calculated using the standard NIST software (Safavi et al., 2012). In speaker verification systems, EER is one of the measure to evaluate the system performance (Wang and Cheng, 2004).

Type of speech	Recorded Audio
Total amount of labeled data (minute)	1588
Amount of data used for training models (minute)	1115
Amount of data used for testing models (minute)	473

Table 2: Database information.

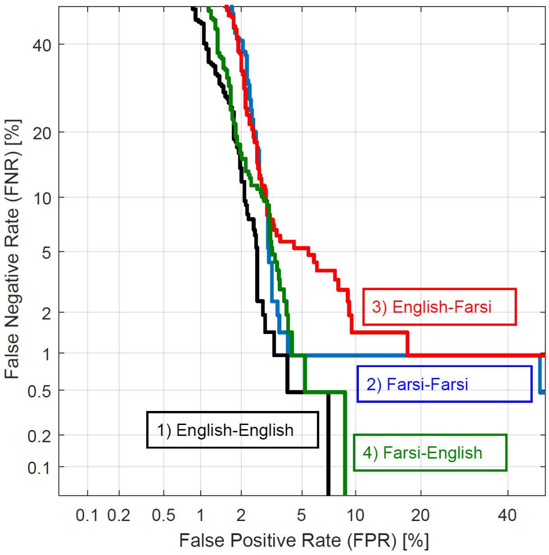

Figure 1: Comparison of 4 Different Setups

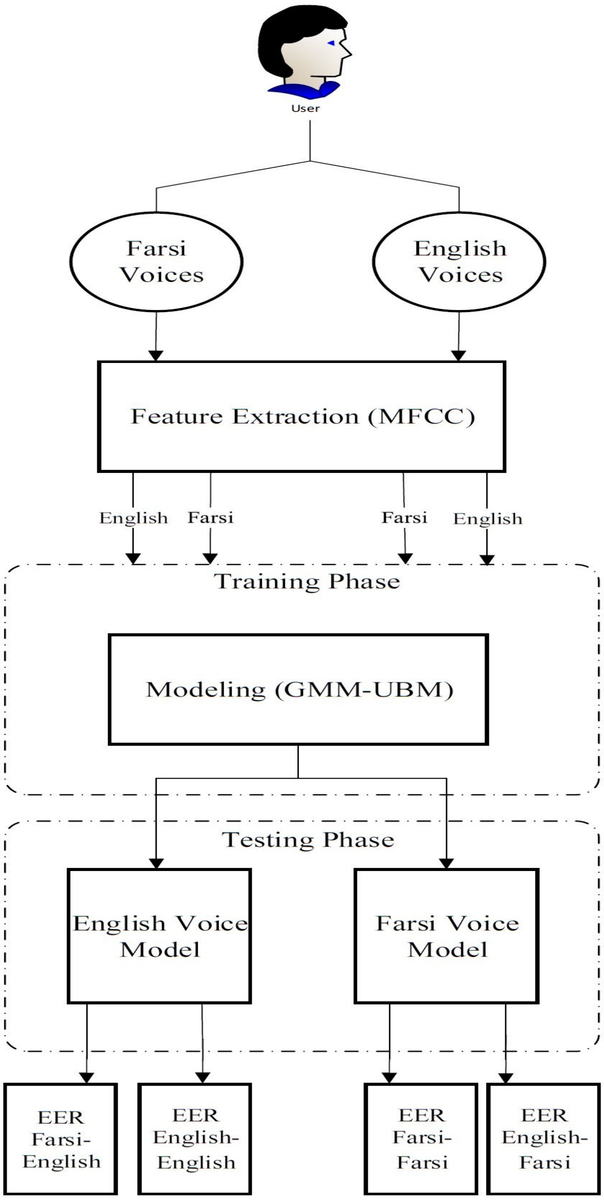

Figure 2: Block diagram of the proposed experiment system which uses text-independent speaker verification system

3.2. Experimental Results

In this study, four experiments were conducted. In the training phase, voice samples from volunteers speaking Farsi and English have been used. Speaker dependent models were then modeled for both English and Farsi. Same as training, in the testing phase we use voice samples both in English and Farsi languages. 70% of the data was used for training and the remaining 30% was used for testing. As shown in figure 1, the Evaluation was then done using 4 different setups:

1. Performance of the system when trained using English data and tested with English data as well, using the English Background Model.

2. Performance of the system when trained using Farsi data and tested with Farsi data as well, using the Farsi Background Model.

3. System performance when trained using English data and tested with Farsi data, by the means of the English Background Model.

4. Performance of the system when trained using Farsi data and tested with English data, by the means of the Farsi Background Model.
At the end of this section we are going to compare the attained results from each setups.

Index	Training-Testing Phases	EER(%)
1	English-English	2.5837
2	Farsi-Farsi	3.2381
3	English-Farsi	5.7585
4	Farsi-English	3.5885

Table 3: The Equal Error Rate (EER) for different experiments

As shown in table 3, English-English (training-testing phases) experiment has the best performance, which demonstrates that the system is more accurate in English which might be because the English language has better overall structure and a more tangible pattern for the model. The results show that having different languages in the training and testing phases increase EER. This means that the system performance is highly language-dependent.
It is worth mentioning that it's best for the non-English language community to use their native language over English in the training phase in order to use the TISV model.

4. Conclusion

In this study, MVSC which is a bilingual speech corpus in English and Farsi is presented. All the volunteers that are native Farsi speakers educated in English have been given 3 texts to read and a sheet of questions to answer in both languages. The speaker would read the texts and answer the questions in a quiet conference room with a high-quality microphone. Features like age, height, gender, etc. have also been added in the corpus repository. Currently, over 33 hours of high-quality audio from 50 speakers are available in the corpus repository but the goal is to record 50 more speakers in order to have approximately 66 hours of audio from 100 speakers in the near future.
In order to Validate MVSC, this study presents and compares the result of experiments in TISV for two different languages. The effect of using different languages on the performance of speaker verification systems is also investigated. Based on the results, it is shown that the TISV model can perform best if the data used in the training phase is the speaker's native language than English. Results revealed that the best performance is obtained when English is used for both training and testing. However, in the cases which the users are only capable of speaking their mother-tongue, it is best to use that language for both training and testing.

5. Acknowledgements

The authors of this paper would like to thank everyone at Miras Technologies International and all the volunteers especially the members of team Meissa and team Draco who put a lot of effort into building the presented dataset.

6. References

Bijankhan, M., Sheikhzadegan, J., and Roohani, M. R. (1994). Farsdat- the speech database of farsi spoken language. *Proceedings Australian Conference On Speech Science And Technology*, 2:826–830.

Bois, J. W. D., Wallace, L. C., Meyer, C., Thompson, S. A., Englebretson, R., and Martey, N. (2000 - 2005). Santa barbara corpus of spoken american english, parts 1-4. *Philadelphia: Linguistic Data Consortium.*

Canavan, A., Graff, D., and Zipperlen, G. (1997). Callhome american english speech ldc97s42. *Philadelphia: Linguistic Data Consortium.*

Mporas, I., Safavi, S., Gan, H. C., and Sotudeh, R. (2016). Evaluation of classification algorithms for text dependent and text independent speaker identification. *IEICE.*

Reynolds, D. A., Quatieri, T. F., and Dunn, R. B. (2000). Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10:19–41.

Safavi, S., Najafian, M., Hanani, A., Russell, M., Jancovic, P., and Michael, J. C. (2012). Speaker recognition for children's speech. *INTERSPEECH*, 2(1):1836–1839.

Singh, N., Khan, R. A., and Shree, R. (2012). Applications of speaker recognition. *Elsevier Ltd.*, 54:0975–8887.

Wang, H. C. and Cheng, J. M. (2004). A method of estimating the equal error rate for automatic speaker verification. *International Symposium on Chinese Spoken Language Processing (ISCSLP).*

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., and Lia, H. (2015). Spoofing and countermeasures for speaker verification: a survey. *Elsevier*, 66:130–153.