Hindawi
Mobile Information Systems
Volume 2017, Article ID 6986391, 6 pages
https://doi.org/10.1155/2017/6986391

Research Article

Automatic Speaker Recognition for Mobile Forensic Applications

Mohammed Algabri, Hassan Mathkour, Mohamed A. Bencherif, Mansour Alsulaiman, and Mohamed A. Mekhtiche

Center of Smart Robotics Research (CS2R), College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia

Correspondence should be addressed to Mohammed Algabri; malgabri@ksu.edu.sa

Received 2 November 2016; Accepted 10 January 2017; Published 13 March 2017

Academic Editor: Eugenijus Kurilovas

Presently, lawyers, law enforcement agencies, and judges in courts use speech and other biometric features to recognize suspects. In general, speaker recognition is used for discriminating people based on their voices. The process of determining, if a suspected speaker is the source of trace, is called forensic speaker recognition. In such applications, the voice samples are most probably noisy, the recording sessions might mismatch each other, the sessions might not contain sufficient recording for recognition purposes, and the suspect voices are recorded through mobile channel. The identification of a person through his voice within a forensic quality context is challenging. In this paper, we propose a method for forensic speaker recognition for the Arabic language; the King Saud University Arabic Speech Database is used for obtaining experimental results. The advantage of this database is that each speaker's voice is recorded in both clean and noisy environments, through a microphone and a mobile channel. This diversity facilitates its usage in forensic experimentations. Mel-Frequency Cepstral Coefficients are used for feature extraction and the Gaussian mixture model-universal background model is used for speaker modeling. Our approach has shown low equal error rates (EER), within noisy environments and with very short test samples.

1. Introduction

In a law court, science and technology are used in investigations to establish forensic evidences [1]. For many years, law enforcement agencies, lawyers, and judges have used voice forensic authentication to recognize suspects [2]. The identification of a person through speech samples with a forensic quality is challenging. In general, speaker recognition, like other bioinformatics features, is used to discriminate people through their voice. Automatic speaker recognition can be classified into two tasks: speaker verification and identification. In speaker identification, the identity of a speaker is determined by analyzing and comparing the speech of unknown speaker with that of a known speaker. Speaker verification is the process of accepting or rejecting the identity claim of a speaker. There are several applications of automatic speaker recognition that can be divided into commercial applications, such as voicemail, telephone banking, biometric authentication, and forensic applications [3]. The process of determining if a suspected speaker is the source of a questioned recording (trace) is called forensic speaker recognition (FSR) [4]. Drygajlo [5] stated that "FSR is an established term used when automatic speaker recognition methods are adapted in forensic applications." In [6] an FSR method using phonemes is described. The study used a combined method between the Gaussian mixture model (GMM) and hidden Markov model: Gaussian hidden Markov model. They used 44 s as the average duration of a recording, with 39 s as the suspect speech and 5 s as evidence. The effect of background noise on biometric parameters in FSR was proposed in [7]. In [8] an FM feature for automatic FSR was presented. The use of FM and Mel-Frequency Cepstral Coefficient (MFCC) features outperform the use of the MFCC feature alone. An enhanced FSR with colored noise was proposed in [9], where

MFCC was used to extract features and GMM to remove noise. The effect of channel variability on speaker recognition is presented in [10], and a new technique was presented to mitigate this effect. The results show that the proposed feature improves performance over a baseline system. Another study [11] proposed the channel compensation in a support vector machine to solve the problem of cross-channel degradation in speaker recognition systems. From the aforementioned literature, we can conclude that speaker recognition for forensic applications is still a challenge. Moreover, FSR for the Arabic language has not been studied and applied appropriately. In this paper, we focus on Arabic speaker recognition for forensic applications. The remainder of this paper is organized as follows: Section 2 shows the methodology of the system containing the speech corpus, feature extraction techniques, and a classifier. Section 3 presents the results obtained and discussion. Finally, Section 4 presents the conclusion and discusses future work.

2. Methodology

2.1. Speech Corpus. All the conducted experiments, through this paper, use a subset of the KSU Speech Database [12], which contains speech data acquired from 264 people (both male and female). The total recording time is approximately 159.5 h. The speakers are Arabs and non-Arabs belonging to 29 nationalities. The KSU Speech Database was recorded at three locations (office, cafeteria, and soundproof room) for three sessions with different channels: mobile medium and high quality microphones.

2.2. Feature Extraction. Feature extraction is a process of capturing the important information of a speech signal providing a useful representation. In this research, MFCCs are used to recognize speakers. MFCCs were developed by Davis and Mermelstein in 1980 and are most widely used for acoustic signals. The main concept is to divide the signal into frames and apply a hamming window for each frame. Thereafter, a cepstral feature vector is generated for each frame by applying the discrete Fourier transform for each frame. Finally, the log of amplitude spectrum must be maintained and the spectrum is smoothed using a discrete cosine transform (DCT) to obtain cepstral features for each frame. Other alternative feature representations are studied as an alternative to the MFCCs: perceptual linear prediction (PLP) and relative spectral transform-PLP (RASTA-PLP). The RASTA-PLP is dependent on the PLP acoustic model developed by Hermansky in 1989 [13], in which a band-pass filter is added to the energy in each frequency of PLP [14].

2.3. GMM-UBM Model. This state-of-the-art speaker recognition system uses a GMM with a universal background model (UBM) [15]. The first step in this modeling involves the creation of a UBM, which is a large mixture of Gaussians covering all speakers and the context of recognition. The UBM is trained using the Expectation-Maximization algorithm. Next, it is adapted using a maximum a posteriori (MAP) estimation algorithm for each speaker [16, 17].

Table 1: EER (%) of different durations of training and testing.

Training (s)	Testing (s)			
	6	4	2	1
25	0.57	0.70	1.15	2.62
20	0.89	1.15	1.15	3.52
15	0.83	1.34	1.41	2.50

2.4. Log-Likelihood Computation. To make a decision, the log-likelihood ratio is computed, and the obtained score is compared with the threshold. The log-likelihood for the test samples of feature X is calculated as follows:

$$Lk(X) = \log p\left(X \mid \lambda_{\text{hyp}}\right) - \log p\left(X \mid \lambda_{\text{ubm}}\right), \quad (1)$$

where $\log p(X \mid \lambda_{\text{hyp}})$ is the log-likelihood of testing data X for a certain target speaker mode λ_{hyp} and the $\log p(X \mid \lambda_{\text{ubm}})$ is the log-likelihood of testing data X for a UBM model.

3. Experimental Results

This section presents the identification of a person by using forensic quality samples. We studied the impact of the training and testing durations, the effect of background noise, session, and channels on the speaker recognition system. To verify the performance, we computed the equal error rate (ERR), false positive rate, and false negative rate. In addition, we show the detection error trade-off (DET) curves of each experiment.

3.1. Performance Evaluation of Different Period of Training and Testing. For many years, courts used recordings of suspects and offenders. One of the common features of such recordings is that they may not contain sufficient relevant speech material or are of a very poor quality. This complicates the determination of the speaker's identity. In this experiment, we used different durations for training sequences (10, 20, and 25 s) and testing sequences (6, 4, 2, and 1 s) for 40 speakers through clean speech recording by using the Yamaha mixer channel.

Figure 1 compares the relation between the performance of speaker recognition and the duration of the training and testing samples. As observed, DET is achieved through training sequences for durations of 25, 20, and 15 s, as shown in Figures 1(a), 1(b), and 1(c), respectively, with four testing sequence durations (6, 4, 2, and 1 s). In this case, we can conclude that the training duration greatly affects the system performance. Interestingly, we achieved a good performance of identifying a suspect speaker by using a 1 s test recording, as shown in Table 1.

3.2. Performance Evaluation of Speech and Background Noise. As previously mentioned, in forensic applications one of the common features of a recording may be very poor quality. In this experiment, we used different types of noises (babble, restaurant, and car) with different levels of signal-to-noise ratios (SNRs; 30, 20, 10, and 0 dB), as shown in

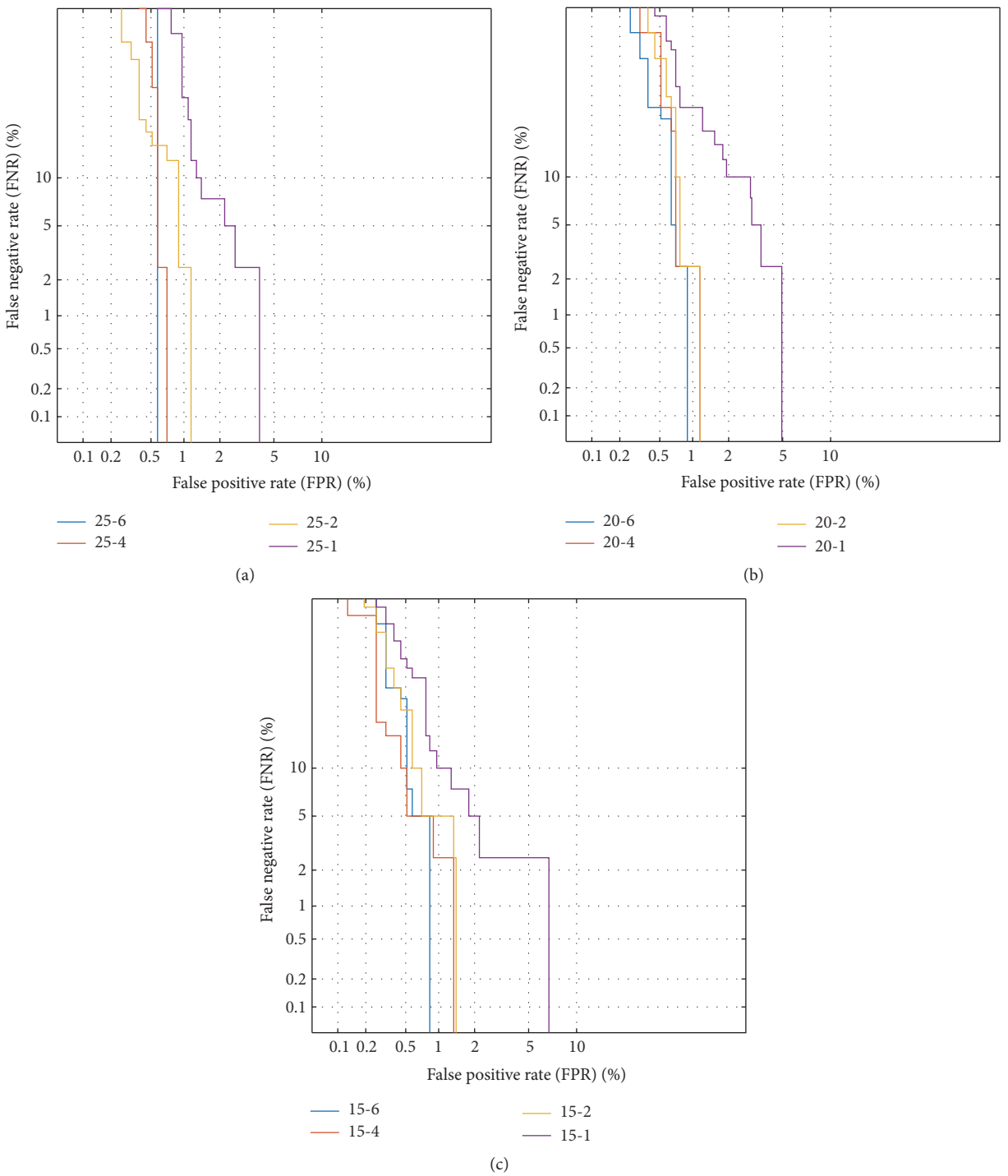

Figure 1: DET curve: training durations versus testing durations.

Table 2. We used clean and noisy speeches for training and testing, respectively. Figure 2 shows the DET curves of the experiments.

Figure 2 compares system performance of the clean training speech and degraded testing speech. The performance was computed using different types of noises for SNRs of 30, 20, 10, and 0 dB. In all cases, a decreasing performance is clearly observed when noise is added to the speech. In case

Table 2: EER (%) versus noise type.

Noise type	SNR			
	30	20	10	0
Babble	0.89	3.14	10	35.8
Restaurant	0.89	2. 75	8.1	35.0
Car	0.96	7.5	25	43.58

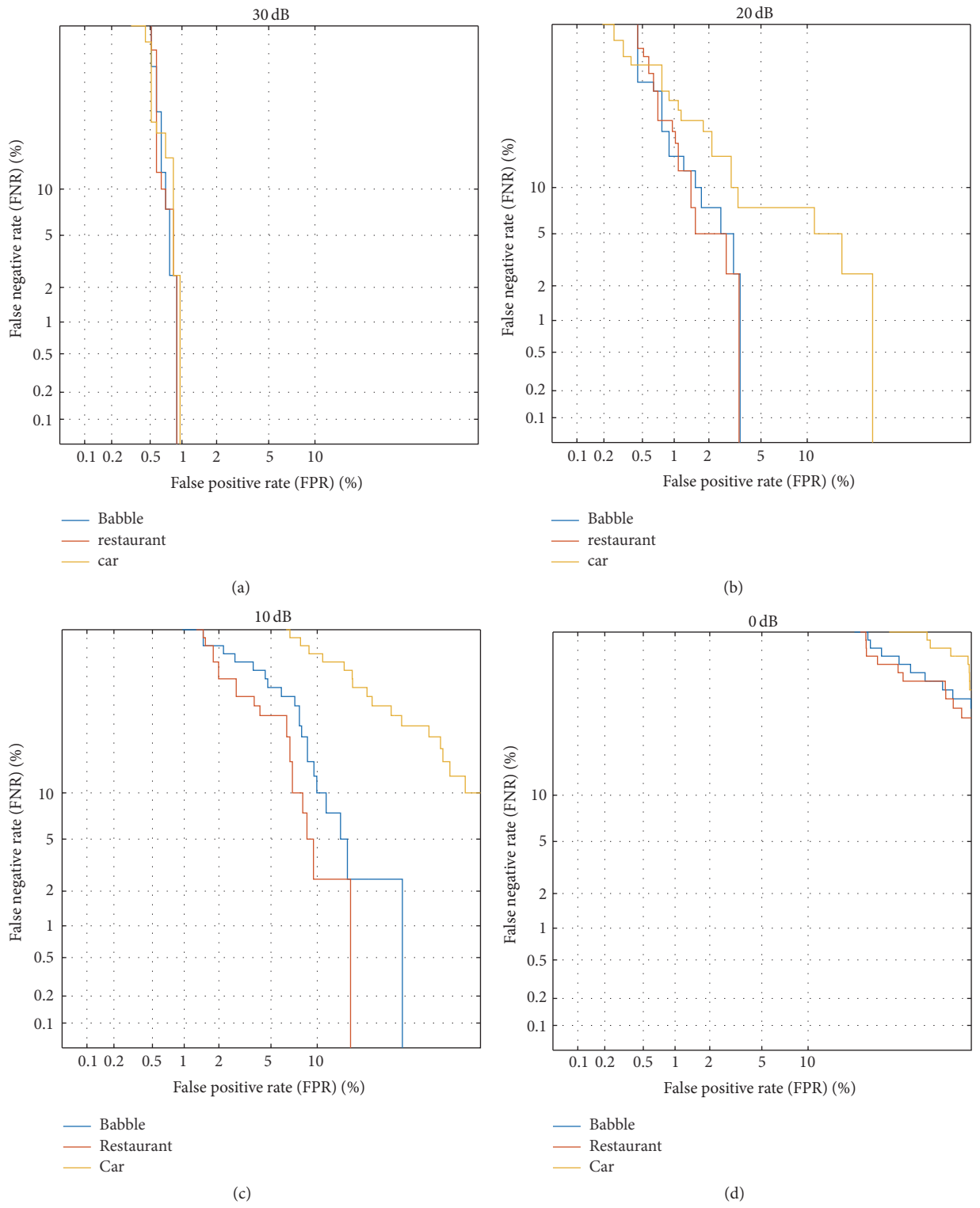

Figure 2: DET curve for different noise types and different SNR.

of low-level SNRs, that is, 10 and 0 dB, we can clearly notice a sharp decline in performance. The analysis of all DET curves shows that car noise has high EER compared with babble and restaurant noises at different levels of SNR.

3.3. Performance Evaluation of the Different Session and Source Recordings. In [2], authors state that a large effort of speaker recognition is concentrated on the mismatch or difference between the training and testing data.

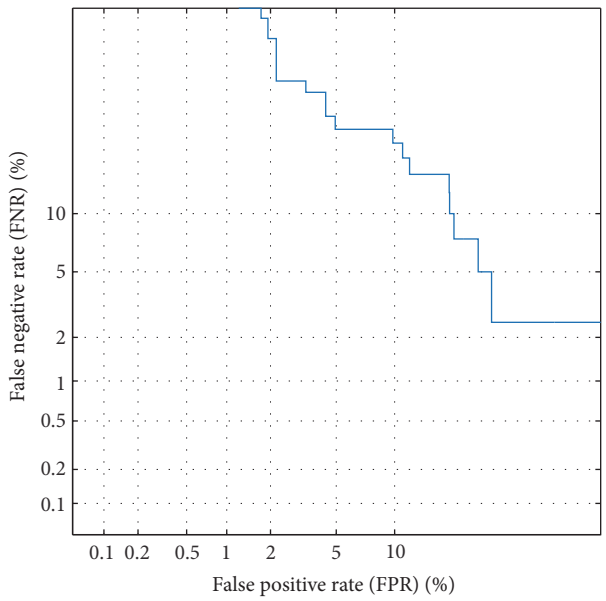

Figure 3: DET curve for different recording sessions.

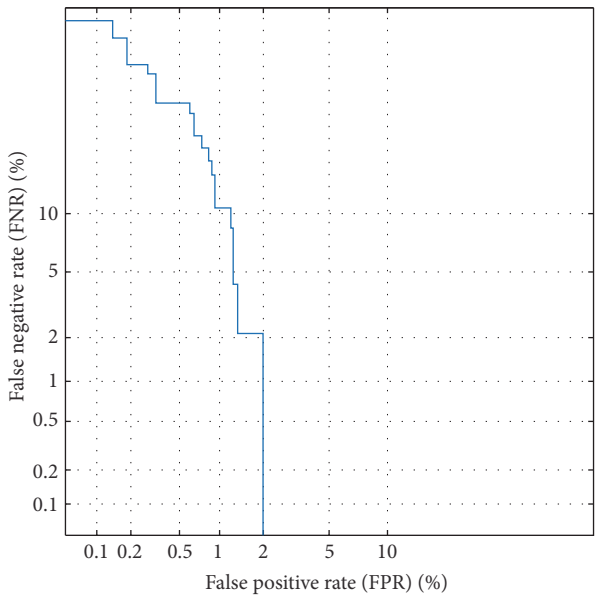

Figure 4: DET curve for mobile channel recording.

This difference is due to factors such as the different times of the recording, environment, microphone, and pathological state of speakers.

In this experiment, we used speeches of 40 speakers. The training set contains the recording of the utterances of one paragraph during an average of 30 s. The same speakers were then asked to record their voices after approximately 3 months.

We used the second recordings as testing data to examine different time durations of the recording; Figure 3 shows the DET curves. The analysis of these DET curves clearly shows that the period between recording the training and testing speeches contributes to the effect on the performance of speaker identification, despite the speech being made by the same people and in the same environment.

3.4. Performance Evaluation of the Mobile Recording. Also in forensic applications, the suspect may be talking via cell phone. In this section, we used speech of 40 speakers to validate the proposed method while the mobile channel is recording. The recording is not being good over mobile channel with low bandwidth and low quality device. The training set contains the recording of the utterances of one paragraph during an average of 30 s and 10 s for testing through clean speech recording by using the mobile channel; Figure 4 shows the DET curves of the experiments. From Figure 4, we can clearly see that the significant performance was achieved using mobile channel recording, at around 97.8% recognition rate with an EER equal to 1.98%.

In general, we can conclude that the performance of speaker recognition is sensitive to factors such as the time of recording, environment, and microphone used. Therefore, this issue requires further study to overcome this disadvantage, especially for Arabic forensic applications.

4. Conclusions

This study presents an Arabic speaker recognition system for forensic applications by using GMM-UBM, thirty-nine MFCCs were used for feature extraction, and all experiments were conducted using the KSU Speech Database.

Four situations were investigated, within suspect identification, in forensic applications. Firstly, the effect of the period of training and testing samples was investigated; this scenario occurs when judges in courts do not have sufficient recordings to identify the suspect. Secondly, when a speech suspect contains noise, this case occurs when the environment impacts the quality of the speech such as in a car or in a noisy site. Thirdly, we studied the effect of time difference or time mismatch between training and testing speeches of the same suspect; this situation occurs when the train samples are recorded months or days before the test samples. Finally, in the fourth set of experiments, we studied the case where the suspect is talking through a mobile channel, considered as the test sample, compared to a train sample that was prerecorded through a different channel, such as a microphone or another channel, also known as multimodal recording.

Our approach has shown that the test sample of a suspect can be recognized, within a noisy environment, with few seconds of speech, and at different times of training and testing. The multimodal approach did not show much improvement; this is mainly due to the big invariance between channels. In the future, we will attempt to develop a system and enhance its performance by adding new features, considering diverse noise environments, and including channel compensation.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

The authors extend their appreciation to the Deanship of Scientific Research at King Saud University for funding this work through research group no. RGP-1436-002.

References

[1] A. Drygajlo, "From speaker recognition to forensic speaker recognition," in *Biometric Authentication*, pp. 93–104, Springer, 2014.

[2] J. P. Campbell, W. Shen, W. M. Campbell, R. Schwartz, J.-F. Bonastre, and D. Matrouf, "Forensic speaker recognition," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 95–103, 2009.

[3] S. S. Alamri, *Text-independent, automatic speaker recognition system evaluation with males speaking both Arabic and English [M.S. thesis]*, University of Colorado, 2015.

[4] A. Drygajlo, "Automatic speaker recognition for forensic case assessment and interpretation," *Forensic Speaker Recognition: Law Enforcement and Counter-Terrorism*, pp. 21–39, 2012.

[5] A. Drygajlo, *Voice, Forensic Evidence of*, Springer, New York, NY, USA, 2014.

[6] P. Univaso, M. M. Soler, D. Evin, and J. Gurlekian, "An approach to forensic speaker recognition using phonemes," Technical Report, 2013.

[7] F. Beritelli, "Effect of background noise on the SNR estimation of biometric parameters in forensic speaker recognition," in *Proceedings of the 2nd International Conference on Signal Processing and Communication Systems (ICSPCS '08)*, IEEE, Gold Coast, Queensland, Australia, December 2008.

[8] T. Thiruvaran, E. Ambikairajah, and J. Epps, "FM features for automatic forensic speaker recognition," in *Proceedings of the 9th Annual Conference of the International Speech Communication Association (INTERSPEECH '08)*, pp. 1497–1500, September 2008.

[9] F. Denk, J. P. C. L. Da Costa, and M. A. Silveira, "Enhanced forensic multiple speaker recognition in the presence of coloured noise," in *Proceedings of the 8th International Conference on Signal Processing and Communication Systems (ICSPCS '14)*, pp. 1–7, IEEE, Gold Coast, Australia, December 2014.

[10] D. A. Reynolds, "Channel Robust Speaker Verification via Feature Mapping," in *Proceedings of the IEEE International Conference on Accoustics, Speech, and Signal Processing*, vol. 6, pp. 53–56, Hong Kong, China, April 2003.

[11] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for SVM speaker recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '05)*, pp. I629–I632, IEEE, Philadelphia, Pa, USA, March 2005.

[12] M. Alsulaiman, Z. Ali, G. Muhammed, M. Bencherif, and A. Mahmood, "KSU speech database: text selection, recording and verification," in *Proceedings of the UKSim-AMSS 7th European Modelling Symposium on Computer Modelling and Simulation (EMS '13)*, pp. 237–242, Manchester, United Kingdom, November 2013.

[13] L. Müller and J. V. Psutka, "Comparison of MFCC and PLP parameterizations in the speaker independent continuous speech recognition task," in *EUROSPEECH 2001 Scandinavia, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH Event, Aalborg, Denmark, September 3–7, 2001*, pp. 1813–1816, 2001.

[14] A. S. Saudi, A. A. Youssif, and A. Z. Ghalwash, "Computer Aided Recognition of Vocal Folds Disorders by Means of RASTA-PLP," *Computer and Information Science*, vol. 5, no. 2, article 39, 2012.

[15] U. Bhattacharjee and K. Sarmah, "GMM-UBM based speaker verification in multilingual environments," *IJCSI International Journal of Computer Science Issues*, vol. 9, 2012.

[16] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1, pp. 19–41, 2000.

[17] D. Povey, S. M. Chu, and B. Varadarajan, "Universal background model based speech recognition," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '08)*, pp. 4561–4564, IEEE, Las Vegas, Nev, USA, April 2008.