Accepted Manuscript
Paper accepted at The IEEE (53rd) International Carnahan Conference on Security Technology.
01/10/2019 – 03/10/2019, Chennai, India.

Effectiveness in the Realisation of Speaker Authentication

Heinz Hertlein, Aladdin Ariyaeeinia, Zoe Jeffrey, *Member, IEEE* and Soodamani Ramalingam
School of Engineering and Computer Science, University of Hertfordshire, Hatfield, AL10 9AB, UK
heinz.hertlein@googlemail.com, a.m.ariyaeeinia@herts.ac.uk, z.j.jeffrey@herts.ac.uk and s.ramalingam@herts.ac.uk

***Abstract*—*An important consideration for the deployment of speaker recognition in authentication applications is the approach to the formation of training and testing utterances. Whilst defining this for a specific scenario is influenced by the associated requirements and conditions, the process can be further guided through the establishment of the relative usefulness of alternative frameworks for composing the training and testing material. In this regard, the present paper provides an analysis of the effects, on the speaker recognition accuracy, of various bases for the formation of the training and testing data. The experimental investigations are conducted based on the use of digit utterances taken from the XM2VTS database. The paper presents a detailed description of the individual approaches considered and discusses the experimental results obtained in different cases.*

Keywords*—*Speaker Authentication, Biometric Applications, Speaker Recognition

I. Introduction

Speaker recognition is principally defined as that of determining the identity of an individual based on a given sample utterance [1]-[3]. The extensive progress achieved in this field over the past two decades has resulted in the technology being considered in a variety of applications ranging from remote (i.e. online & telephony) access control to identity segregation in smart environments. The two main categories of speaker recognition are defined as speaker verification and speaker identification. The former is the process of establishing whether a claimant speaker is the person he/she claims to be [4], [6]. The decision in this case is based on the match score obtained for a given test utterance against the reference model for the claimed identity. Speaker identification, on the other hand is that of determining the correct speaker from a set of registered individuals [3], [5]. For this purpose, the test utterance is compared against the reference models of the registered population. In this case, the model yielding the highest score is identified as belonging to the speaker of the given utterance. To be more specific, this process is referred to as closed-set speaker identification. If the process also includes the option of rejecting a given test utterance as not belonging to any of the speakers in the registered set, then it is termed open-set speaker identification [3]. The concern in this paper is essentially that of speaker verification. It should be noted that a detailed review of speaker recognition, covering the state of the art classification methods is presented in [2].

In practice, for any given speaker verification approach, the effectiveness of the process requires operational reliability in terms of the recognition accuracy and robustness against spoofing [6-10]. The latter has been the subject of growing investigations over the recent years [8-10]. The primary aim in those studies is to address the challenges of replay attacks and voice conversion, which are currently the two main facets of circumvention in speaker authentication.

The recognition accuracy, on the other hand, not only depends on the methods deployed in structuring the entire process (e.g. speaker modelling, speaker classification) and dealing with speech degradation (i.e. due to ambient noise and channel mismatch), but also on the paradigm defining the training and testing utterances.

The purpose of investigations presented in this paper is to facilitate the definition of frameworks for training and testing utterances for the benefit of optimising the reliability in the verification process. The remainder of the paper is structured as follows. Section 2 provides a description of the data characteristics considered in the testing phase of speaker verification in this study. Section 3 describes the investigations and discusses the experimental results. A summary of the work carried out and the overall conclusions are presented in Section 4.

II. Testing Phase Characteristics

An important requirement in the deployment of text-dependent speaker verification in authentication scenarios is that of setting up an appropriate data framework in the testing phase in relation to that in the training stage. In this regard, there are indeed a range of approaches that can be adopted. Given the characteristics of the application area considered in this study, Table 1 details the digit utterance-based data structures that are assumed in the experimental investigations.

In terms of the recognition accuracy, scenario (2) can be expected to be “in-between” scenario (1) and scenarios (3)/(4).

This scenario is not explicitly included in the experimental evaluation. In the case of Scenarios (3) and (4), it should be noted that in practice the variation of textual content from one speaker to another might cause further variation of client and impostor scores. Therefore, in such cases, particular attention should be paid to the textual richness of the speech data adopted for the implementation of score normalisations like T-norm. This is particularly the case in scenario 4, where there is no information at all about the content of the utterances, whereas in scenario (3) the system has a record of the textual content of the utterances. Additionally, scenario (3) can be designed to be based on a finite set of texts, but this is not the case in scenario (4), where clients can freely choose their own utterances.

TABLE I. DATA FRAMEWORKS CONSIDERED IN THE STUDY

Scenario	Testing/Training data conditions	Definition	Further Elaboration
(1)	Use of the same text in testing and training stages	All speakers utter the same phrase in the training and testing phases, i.e. there is no textual content variation.	All speakers say a pre-defined passphrase that is instructed by the system. For example, "My voice is my passport. Verify me." (the passphrase from the movie "Sneakers", 1992).
(2)	The text varies from test trial to test trial, but the training material is the same for all speakers	A longer recording that consists of several elements (or several recordings) is required for the enrolment. For the recognition stage, one of these elements is spoken. Therefore, the textual content can vary for individual test episodes, whilst the full set of texts is the same for all speakers (i.e. the utterance text in each test trial is a subset of the textual content of the training material).	All speakers enrol with the ten digit utterances of zero to nine. To be recognised, one or more digits will have to be uttered. This condition is suitable for a prompt-system adopted for protection against replay attacks. In this case, the system prompts the user for a certain digit combination. In practice, this prompt system requires text verification in addition to speaker recognition.
(3)	Passphrases selected by the system	Each speaker is assigned a text during the enrolment. The system decides which text is used for which speaker. The same text is to be spoken for the purpose of recognition.	In practice, words with a suitable length can be selected from a dictionary during the enrolment. The security of the system can be increased by requiring users to keep their assigned passphrase secret.
(4)	Passphrases chosen by clients	Each speaker chooses a text during the enrolment. The users can pick any text they like. The chosen phrase is to be uttered in the testing stage.	The user chooses a password at the time of enrolment. A length requirement might be enforced by the system.
(5)	Text prompt system with varying amount of training data	To maximise the acceptability of the system, users are enrolled with minimal effort using a small number of short utterances (digit utterances in this study). To increase the overall accuracy, re-training is performed with the test material following successful verification trials.	For example, only two or three digits might be sufficient to enrol. For the verification purpose, a digit prompt system is used for protection against replay attacks. To increase the recognition accuracy, test utterances with a high target probability (well above the normal threshold) are added to the training set and the respective target model is re-trained.

Depending on the application, in scenarios (3) and (4) it should be possible to require users to keep their passphrases confidential. Assuming that impostors cannot use the valid text, this improves the recognition accuracy, as verbal information verification (passphrase verification) can complement speaker verification.

III. EXPERIMENTAL EVALUATION

A. Speech data

The essence of the experimental investigations in this study is that of the evaluation of the relative speaker recognition accuracy obtainable in alternative data scenarios considered. In this respect, the corpus should have the following properties:

- For each speaker in the dataset, there should be a certain set of texts (for example, digits).
- The set of texts should be the same for all speakers.
- Each set is repeated several times by each speaker.

The XM2VTS audio corpus [11] has these properties and is therefore adopted in this study. The only issue to note is that this database provides the sequence of ten digits in a single file, rather than individual digit utterances in separate files. However,

assuming that the number of frames after silence removal is approximately the same for all ten digits, the use of single digits can be simulated by splitting the feature vector sequence of each recording into ten equal parts. It is of course acknowledged that different digit utterances have slightly different durations. However, this should not distort the outcomes as the purpose of the study is to establish the relative recognition accuracy in different cases considered.

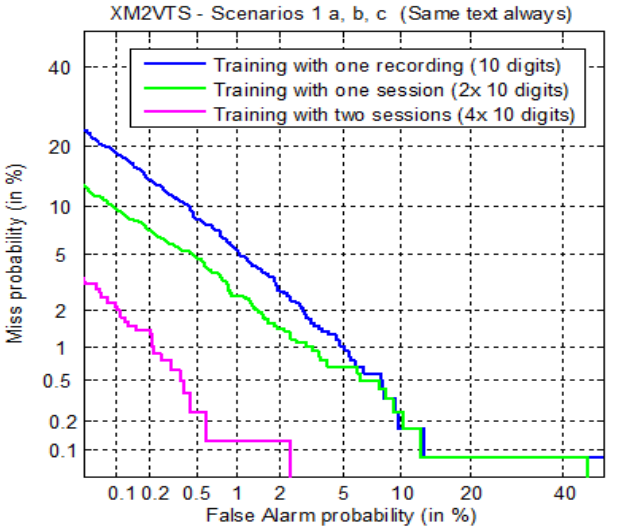

Fig. 1. GMM-UBM verification results on XM2VTS for scenarios 1 a, b and c.

B. Verification experiments and results

Table 2 illustrates the results with the full-bandwidth version of XM2VTS. It should be noted that ten speakers are removed from this corpus because of errors noted with some of the recordings. For the purpose of verification experiments, 200 speakers from this database are used as the registered speakers. Utterances from the remainder of speakers are used for the generation of the UBM.

Fig. 1 illustrates the verification results as DET plots for the experiments in scenarios 1 (a), (b) and (c). It is noted in this figure that, as expected, an increase in the training data improves the verification accuracy. The plots also confirm that when the additional training data is from a different recording session, the verification error reduces significantly. It is important to note that, as indicated in Table 2, the EER obtained for case 1 is further reduced from 0.39% to 0.25% (nearly 36% in the relative terms) when the testing data is increased in the form of passphrase repetition (scenario 1 (d)). The results for scenarios 3 (a) and 3(b) are presented as DET plots in Fig. 2. It should be noted that the training data in these user-specific passphrase cases are much shorter than those in the previous experiments. Additionally, the test trials are based on much shorter utterances (i.e. passphrases are 3 digit-long).

It is very interesting to note that 3(a) offers much higher accuracy than 1(b) where the training material is richer, and the test trials are based on longer utterances (i.e. all 10 digit utterances). This is highly attributed to the use of passphrases in the case of 3(a). However, it is worth highlighting the fact that in this case, due to data limitation, the shared acoustic content of passphrases is considerable. It is again important to note that with the use of two test utterances in these experiments, the verification error reduces significantly. This indicates that, in practice, the verification reliability can be enhanced significantly by requiring users to repeat their passphrases.

In practical applications, it will not be realistic to expect the registration of all users to be based on equal amounts of training data. Fig. 3 presents the DET plots for experiments covering this particular case. The experimental set up is as detailed in Table 2. In this respect it should be reiterated that, in here, the testing data for each verification trial is an arbitrary sequence of all digits from one complete recording (10 digits). The level of verification error noted in Fig. 3 is thought to be a consequence of using varied length of training data, which leads to considerable variation in the richness of the registered models (in some cases, the training

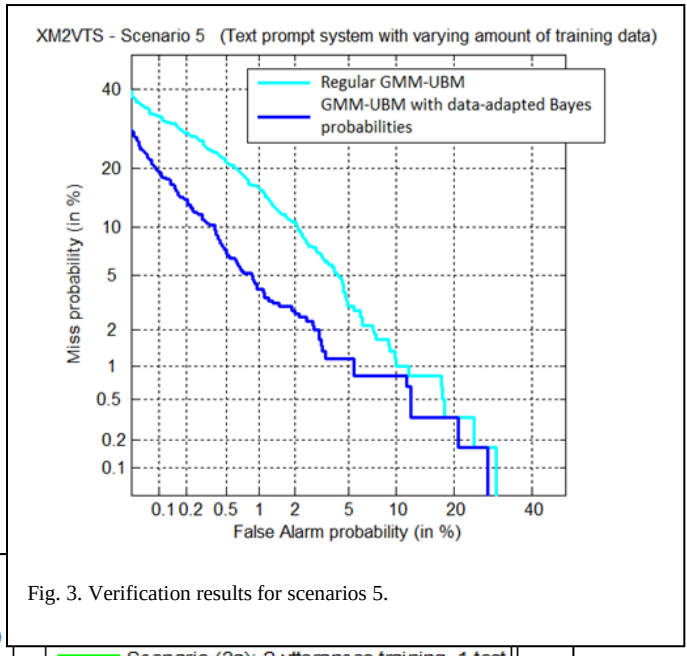

Fig. 3. Verification results for scenarios 5.

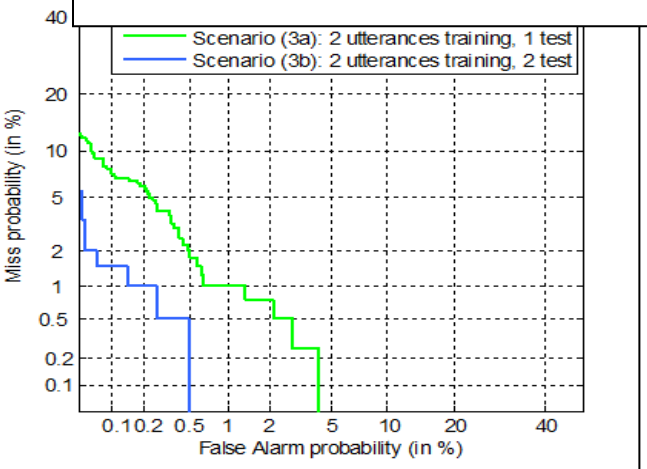

Fig. 2. GMM-UBM verification results on XM2VTS for scenarios 3 a and b.

material consist of only two-digit utterances). It is worth noting that, in practice, the testing data (prompted text) in each trial can be chosen to specifically correspond to the textual structure of the material used for training the speaker model of the claimed identity. With such an approach, the integration of verbal information verification with speaker verification is expected to considerably enhance the overall reliability of the process, as it facilitates tackling replay attacks. The complete realisation of such an approach will form part of the future work in the field.

TABLE II. EXPERIMENTAL RESULTS IN TERMS OF EER FOR THE INDIVIDUAL SCENARIOS CONSIDERED. IT SHOULD BE NOTED THAT THE ONLY SCORE NORMALISATION APPLIED HERE IS BASED ON THE UBM, WHICH IS INCORPORATED IN THE GMM-UBM [12] PROCEDURE. THIS SHOULD NOT AFFECT THE RELATIVE PERFORMANCE IN ALTERNATIVE DATA SCENARIOS, WHICH IS THE CONCERN IN THIS EXPERIMENTAL STUDY.

Scenario	**Test protocol: based on digit corpus** **Recognition accuracy: based on "clean" XM2VTS** **Speech feature: LPCC** **Classifier: GMM-UBM**	**Verification: EER**
(1) Same text always	**(1a)** Using the complete first recording of the first session (containing all 10 digits) for training and sessions 2 to 4 for testing. Each classification decision is based on one complete recording, containing all ten digits. As there are two recordings in each session, there are 6 test utterances for each speaker.	2.36 %
	(1b) Using the whole of session one for training, i.e. two recordings for each speaker. Test is the same as in (1a).	1.67 %
	(1c) Using sessions 1 and 2 for training, and sessions 3 and 4 for testing. Similar to (1a) and (1b), each verification trial is based on one recording. However, the total test set is slightly different due to the dissimilar partitioning into training and test sub corpus.	0.39 %
	(1d) The training is the same as in (1c). The test is different in the sense that each classification decision is based on one session (i.e. two recordings).	0.25 %
(2) The text varies from test trial to test trial, but the training material is the same for all speakers	The data protocol for training is identical to (1a) or (1b). However, each test trial is based on one or more digit utterances, i.e. a subset of ten digit utterances.	N/A
(3) Passphrases picked by the system	**(3a)** For each speaker, three consecutive digits are selected. This results in 2x8 distinct passphrases, due to the two different types of recordings ("0123456789" and "5069281374"). The training for each speaker is two recordings of his/her passphrase, and the testing is based on one recording. As there are 200 speakers, 12 or 13 speakers respectively share the same passphrase. In addition, there is some similarity of acoustic content of differing passphrases because only digits are used, and also because the 2x8 passphrases are not entirely unique (each partially share textual content with a number of others in the set).	1.00 %
	(3b) Training is identical to (3a). Each test decision is based on two recordings of the client's passphrase, instead of one.	0.49 %
(4) Passphrases chosen by clients	The results of scenario (3) above should give an indication for the recognition accuracy in scenario (4) as well, because there is no normalisation which takes advantage of using a fixed set of passphrases with known textual content.	

(5) Text-prompted system with varying amount of training data	**(5)** The amount of training material is varied from speaker to speaker. There are 6 groups of speakers. One group is trained with one complete session (20 digits). The other 5 groups are trained with a minimum of 2 and a maximum of 10 digits. The test material for each verification trial is an arbitrary sequence of all digits from one complete recording (10 digits). It is assumed that the prompt verification can be performed by a separate, speaker independent digit classifier, which operates independent of the speaker classifier. The results shown on the right are for the speaker classifier only.	GMM-UBM: 4.53 % GMM-UBM with data adapted Bayes probabilities: 2.46 %

IV. Conclusion

The two important facets of operational reliability in speaker authentication are the recognition accuracy and robustness against spoofing attempts. The experimental study presented in this paper has been concerned with the former and, in particular, with the way the accuracy in such a scenario is influenced by the paradigm adopted for the formation of testing and training material. Using the XM2VTS database, the study has covered a number of scenarios in the training and testing stages of speaker authentication.

As expected, the experimental results have confirmed that the recognition accuracy improves with an increase in the training data and that this improvement is more significant when the additional training data is from a different recording session. A related finding of considerable interest is that increasing the test data in the form of passphrase repetition can greatly increase the authentication accuracy.

The experiments with user-specific passphrases have shown that this scenario offers considerable improvement in the authentication accuracy, even when the data content is less rich phonetically. The outcomes of this part of the study have again confirmed that requesting users to repeat their allocated passphrases in the test phase can significantly enhance the recognition accuracy.

The investigations have also included experiments with varied duration training utterances for users. The results clearly indicate a drop in the verification accuracy because of the reduction in the textual content of the training material (i.e. reduction of the training data to as low as 2-digit utterances in some cases). In practice, the reliability against replay attacks is expected to improve considerably when the operation is in a true text-prompted mode, supported by a speaker independent speech recognition engine.

References

[1] T. Kinnunen and H. Li, 'An overview of text-independent speaker recognition: from features to supervectors', Speech Commun., 2010, 52, (1), pp. 12–40.

[2] J.H. Hansen and T. Hasan, 'Speaker recognition by machines and humans: A tutorial review', IEEE Signal Process. Mag., 2015, 32, (6), pp. 74-99.

[3] A.M. Ariyaeeinia, J. Fortuna, P. Sivakumaran and A. Malegaonkar, 'Verification effectiveness in open-set speaker identification', IEE Proceedings Vision, Image and Signal Processing 2006, 153, (5), 618-624.

[4] S. Pillay, A. Ariyaeeinia, M. Pawlewski and P. Sivakumaran, 'Speaker verification under mismatched data conditions', IET Signal Processing', Special Issue on Biometric Recognition, 2009, 3, (4), pp. 234-246.

[5] S. Pillay, A. Ariyaeeinia, P. Sivakumaran and M. Pawlewski, 'Open-set speaker identification under mismatch conditions', Proc. Interspeech 2009, pp. 2347–2350.

[6] R. Auckenthaler, M. Carey and H. Lloyd-Thomas, 'Score normalisation for text-independent speaker verification systems', Digit. Signal Process., 2000, 10, (1–3), pp. 42–54.

[7] A.M. Ariyaeeinia and P. Sivakumaran, 'Analysis and comparison of score normalisation methods for text-dependent speaker verification'. Proc. Eurospeech 1997, pp. 1379–1382.

[8] S. K. Ergünay, E. Khoury, A. Lazaridis and S. Marcel, 'On the Vulnerability of Speaker Verification to Realistic Voice Spoofing', proc. IEEE International Conference on Biometrics: Theory, Applications and Systems, Arlington, 2015.

[9] Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre and H. Li, 'Spoofing and countermeasures for speaker verification: A survey', Speech Commun. 2015, 66, pp. 130-153.

[10] T. Masuko, T. Hitotsumatsu, K. Tokuda and T. Kobayashi, 'On the Security of HMM-Based Speaker Verification Systems Against Imposture Using Synthetic Speech', Proc. Eurospeech 1999, pp. 1223-1226.

[11] http://www.ee.surrey.ac.uk/CVSSP/xm2vtsdb

[12] D. Reynolds and R. Rose, 'Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models', IEEE Transactions on Speech and Audio Processing 1995, 3, (1), pp. 72–83.