Improving Robustness of Speaker Verification by Fusion of Prompted Text-Dependent and Text-Independent Operation Modalities

Iosif Mporas, Saeid Safavi and Reza Sotudeh

School of Engineering and Technology, University of Hertfordshire
College Lane Campus, Hatfield AL10 8PE, Hertfordshire, UK
{i.mporas,s.safavi,r.sotudeh}@herts.ac.uk

Abstract. In this paper we present a fusion methodology for combining prompted text-dependent and text-independent speaker verification operation modalities. The fusion is performed in score level extracted from GMM-UBM single mode speaker verification engines using several machine learning algorithms for classification. In order to improve the performance we apply clustering of the score-based data before the classification stage. The experimental results indicated that the fusion of the two operation modes improves the speaker verification performance both in terms of sensitivity and specificity by approximately 2% and 1.5% respectively.

Keywords: speaker verification; fusion; machine learning.

1 Introduction

Biometric technology has widely been used over the last decade to several applications, such as access control to physical places, secure login to computer systems and mobile devices, online banking and ATMs, personalized human-machine interfaces etc. One of the most widely used modalities in this area is voice-based biometrics and particularly speaker verification, due to the convenience that offers to the user as well as due to the fact that the input signal can be captured by a conventional microphone, nowadays available in most electronic devices, and does not need any specialized sensor or other hardware equipment to capture the input biometric signal.

In speaker verification the user provides a speech input, usually after a screened prompted message, and the system decides whether the user is an authorized one or not, i.e. accepts or rejects the claimed by the user identity. Based on the prompted message, speaker verification can roughly be divided into two categories, namely the text-dependent and the text-independent. In text-dependent speaker verification a text message, selected from a predefined and close-set of utterances, is prompted to the user in order for him/her to pronounce it [1-3]. In the case of text-independent speaker verification [4-7] a text generator is used to prompt to the user a text message to be pronounced, which does not belong to an apriori known to the user close-set of utterances. Thus, in the text-independent mode of operation the prompted message is ran-

dom and as a result cannot easily be reproduced by audio replay attacks from impostors. On the other hand, the use of new uttered message to the speaker verification system, which does not appear (as a whole or even partially) in the training data results in the reduction of the verification performance, thus result to a trade-off between performance and robustness against spoofing.

The concurrent technology in speaker verification is based on short-time speech signal analysis followed by machine learning based modeling. In detail, the most commonly used features for speaker recognition are the Mel frequency cepstral coefficients (MFCCs) [8, 9]. Other speech parameterization techniques, as wavelets have also successfully been applied [7]. As considers speaker modeling, the state of the art technology is dominated by the probabilistic Gaussian mixture models (GMMs) [10]. GMM technology has proved to perform well using universal background models (UBMs) trained from a large number of background speakers and maximum a-posteriori (MAP) adaptation or means-only adaptation of the UBM to speaker specific data. Except probabilistic modeling discriminative approaches, such as support vector machines (SVMs) have also successfully been used in the task of speaker verification [11]. SVMs have also been used in combination with GMMs by concatenating the means of the Gaussian components of the GMMs to super-vectors and apply discriminative classification on them [12]. Recently, subspace methods have been proposed for the speaker verification task such as the i-vectors method [13], which are based on joint factor analysis. Although in specific setups subspace methods have proved to outperform probabilistic models, the GMM-UBM approach in general offers more stable results, especially when not enough training and development data are available. For this reason, in the present evaluation we relied in this technology.

In this work, we present a methodology for fusing the speaker verification scores produced by two different modes of operation, namely the text-dependent and the text-independent. The exploitation of the advantages of each of the two modes of operation is achieved using a machine learning based scheme for fusion, in order to get a final speaker verification decision.

The rest of the article is organized as follows. In Section 2 the proposed fusion methodology for combining prompted text-dependent and text-independent speaker verification modes is presented. In Section 3 the experimental setup that was followed is described and in Section 4 the experimental results are presented. Finally, in Section 5 the conclusions of this work are given.

2 Fusion of Speaker Verification Operation Modes

In real-life voice based biometrics applications the user is asked to provide voice samples in order the system to verify whether the user is an authorized one or not. Depending on the mode of operation, the speaker verification performance as well as the vulnerability to spoofing attacks are affected. Specifically, when using text-dependent prompts the recognition accuracy is high, while when using text-independent prompts the performance significantly drops. On the other hand, prompted text-dependent operation is easy to be spoofed, for example using audio replay

attacks or synthetic speech. In contrast to this, text-independent speaker recognition mode of operation offers robustness against spoofing attacks, since due to the absence of apriori knowledge of the prompted utterance message audio replays cannot be applied, while in synthetic speech based attacks the use of phonetically rich prompted messages (which probably will not appear in the training corpus) can significantly reduce the quality of the output of a text-to-speech engine. Except this, algorithms based on phase detection can be used to identify synthetic speech.

The fusion of the prompted text-dependent (TD) and text-independent (TI) modes of operation is performed on score level. In detail, the user is asked to provide voice response to two prompted messages (usually shown on a screen), which consist of TD and TI utterance messages respectively. Each of these messages is processed by a mode specific speaker verification engine and the TD and TI verification scores are estimated. The two mode-dependent scores are concatenated to constitute a 2-dimentional feature vector which is used as input to a machine learning classification algorithm, in order to decide whether the user is an authorized or an impostor. Since the score values typically present some variation, in order to support the classification stage, we apply in advance clustering in order to separate the 2-dimentional score data to areas with less variation. After clustering the data we apply a cluster-specific classification model and get the verification decision. The block diagram of the proposed methodology is illustrated in Fig. 1.

As can be seen in Fig. 1, the user is providing to the system a prompted text-dependent and a prompted text-independent voice input. These inputs are processed by mode-specific speaker verification engines and one verification score is estimated for each mode, i.e. S^{TD} for text-dependent and S^{TI} for text-independent mode of operation. The two scores are concatenated to a score vector $V \in \mathbb{R}^2$. During the training phase a cluster algorithm separates the score vectors to C clusters. In the test phase each score vector is assigned to a cluster c, with $1 \leq c \leq C$. Based on the detected cluster for each pair of TD and TI inputs a cluster-specific classification model f will be activated and assign an acceptance or rejection decision label to the input score

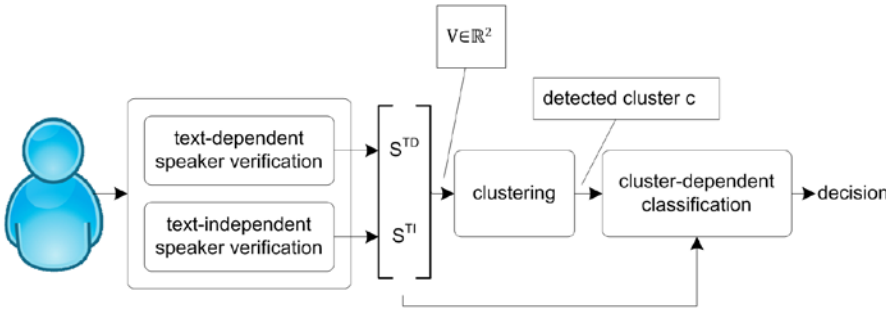

Fig. 1. Block diagram of the proposed methodology for fusion of prompted text-dependent and text-independent modes of speaker verification operation.

vector V, i.e.

$$d = f_c(V) \quad (1)$$

where f_c denotes the classification model that is dedicated to classify the inputs V which belong to cluster c.

The number of clusters is manually defined based on evaluation of performance on a bootstrap training data set. The verification in each operation mode-specific speaker verification engine is made on the basis of thresholded scores, while the speaker verification decision of the fusion methodology is made on the basis of classification results.

3 Experimental Setup

The experimental setup for the evaluation of the fusion methodology described in Section 2, is presented here. Specifically, we describe the dataset used in the evaluation, the setup of the single-mode speaker verification engines and the setup of the fusion stage.

3.1 Speech corpus

In this evaluation we relied on the RSR2015 database [3]. RSR2015 consists of recordings from 300 speakers (157 male, 143 female). For each speaker, there are 3 enrolment sessions of 73 utterances each and 6 verification sessions of 73 utterances each. In total there are 657 utterances distributed in 9 sessions per each speaker. The sampling frequency of the speech recordings is 16 kHz and the speech samples are stored with analysis equal to 16 bits.

Except RSR2015, we used TIMIT [14] for training a universal background model. TIMIT consists of recordings of 630 speakers, sampled at 16 kHz with resolution analysis equal to 16 bits per sample.

3.2 Single-mode speaker verification engines

Each of the two single-mode speaker verification engines, i.e. the text-dependent and the text-independent ones, were based on the well known GMM-UBM technique [10]. Specifically, each voice input was initially pre-processed and parameterized. During pre-processing an energy-based speech activity detector was applied to retain the speech only parts. The speech input was frame blocked using a time shifting Hamming window of 20 msec length with 10 msec overlap between successive frames. For each frame the first 19 Mel frequency cepstral coefficients (MFCCs) were estimated, which were further expanded to their first and second derivatives, thus resulting to a feature vector of length equal 57. In order to reduce the effect of handset mismatch and make the feature more robust RASTA [15] and CMVN processing were applied to the MFCC features.

For both TD and TI speaker models we used the GMM-UBM [10] approach. Specifically the universal background model (UBM) was built by a mixture of 128 Gaus-

sian distributions and was trained using all utterances from 630 speakers from TIMIT. For each of the speakers of the RSR2015 database we applied MAP adaptation (means only adaptation) on the UBM model, using the speaker-specific enrollment data and the speech utterances that corresponded to the text-dependent and the text-independent sessions for the speaker TD and TI models respectively.

3.3 Setup of the fusion of speaker verification modes

The verification scores produced by the text-dependent and the text-independent speaker verification engines described above were concatenated to 2-dimentiional feature vectors as described in Section 2. These data were clustered to groups using the k-means algorithm [16]. The number of the clusters produced by the k-means algorithm was manually defined. Based on the clustering results we trained one classification model for each cluster of data.

As considers the classification stage, each pair of TD and TI scores was processed by a cluster-specific classification model, we relied on a number of well known and widely used in the bibliography machine learning algorithms for classification. Specifically, we used the following algorithms: (i) multilayer perceptron neural networks (MLP), (ii) C4.5 decision trees (C4.5), (iii) support vector machines (SVM) using the sequential minimal optimization implementation, (iv) Bayesian networks (BN), (v) classification and regression trees (CART) and (vi) reduced error pruning tree (REP). For the implementation of these machine learning algorithms for classification we relied on the WEKA toolkit [16].

4 Experimental Results

The proposed fusion methodology for speaker verification presented in Section 2 was evaluated based on the experimental setup described in Section 3. For all evaluations we relied on a 10-fold cross validation protocol. The performance of the proposed methodology was evaluated in terms of sensitivity (i.e. the percentage of the correctly classified instances of the target speakers) and specificity (i.e. the percentage of the correctly classified instances of the impostor speakers).

As a first step we evaluated the performance of the fusion scheme without applying clustering. The experimental results for the single-mode TD and TI speaker verification engines as well as their fusion using several classification algorithms is tabulated in Table 1.

As can be seen in Table 1, the single mode speaker verification methods outperform the evaluated fusion methodologies in terms of sensitivity. However, in terms of specificity it seems that fusion of the text-dependent and text-independent modes offers an improvement of more than 2%. Specifically, the best performing fusion algorithm was the Bayesian network classifier which offered sensitivity equal to 76.33%, followed by the decision trees (C4.5, CART and REP). In terms of sensitivity, the MLP and SVM discriminative algorithms did not offered high accuracy. All evaluated fusion algorithms proved to offer specificity of more than 99.50%, while

the text-dependent and text-independent single mode methods achieved 97.46% and 91.83 respectively.

Table 1. Speaker verification, in terms of percentages of sensitivity and specificity, for different operation mode fusion methods.

Method	sensitivity	specificity
TD (single mode)	84.65	97.46
TI (single mode)	84.70	91.83
MLP	71.14	99.82
C4.5	72.89	99.79
SVM	71.12	99.82
Bayesian Network	76.33	99.66
CART	72.13	99.81
REP	73.10	99.78

In a second step we estimated the performance of the fusion methodology using different numbers of clusters. The experimental results are tabulated in Table 2. For direct comparison we replicate the results for the case where no clustering of the data was applied before the classification stage.

Table 2. Speaker verification, in terms of percentages of sensitivity (sens) and specificity (spec), for different operation mode fusion methods and different number of clusters.

	c=1		c=5		c=10		c=20	
Method	**sens**	**spec**	**sens**	**spec**	**sens**	**spec**	**sens**	**spec**
TD (single mode)	84.65	97.46	-	-	-	-	-	-
TI (single mode)	84.70	91.83	-	-	-	-	-	-
MLP	71.14	99.82	73.30	99.79	73.30	99.78	72.11	99.81
C4.5	72.89	99.79	72.97	99.79	71.92	99.81	71.53	99.82
SVM	71.12	99.82	76.07	99.70	58.89	99.97	68.43	99.88
Bayesian Network	76.33	99.66	78.70	99.57	86.55	98.88	86.47	98.92
CART	72.13	99.81	73.18	99.79	72.37	99.80	72.47	99.81
REP	73.10	99.78	73.33	99.77	72.52	99.78	72.34	99.79

As can be seen in Table 2, the application of fusion of the two modalities on clustered data results to significant improvement of speaker verification both in terms of sensitivity and specificity. In detail, the Bayesian network achieved 86.55% sensitivity for 10 clusters (i.e. for $c = 10$), which results to an improvement of 2% comparing to the text-dependent single modality. For the same setup Bayesian network achieved specificity equal to 98.88%, which corresponds to an absolute improvement of approximately 1.5% comparing to the TD single mode case. The application of cluster-based fusion of text-dependent and text-independent modes of speaker verification improved the sensitivity accuracy of all evaluated algorithms comparing to the case of fusion without clustering, i.e. for $c = 1$. This is owed to the fact that in the case of cluster-based fusion the classification algorithms are trained on data with less varying

characteristics thus can train their free parameters to be dedicated to each specific data subset's characteristics.

5 Conclusions

The score level fusion of prompted text-dependent and text-independent speaker verification modalities is a methodology that can directly be applied to real-world applications related to voice-based biometrics. The experimental evaluation using clustering of the single mode score data followed by application of classification for fusion showed an absolute improvement of more approximately 2% in terms of sensitivity and an absolute improvement of 1.5% in terms of specificity. The best performing algorithm for fusing the two modes of speaker verification operation was found to be the Bayesian network classifier. The improvement is owed to the exploitation of the underlying and complementary information between the distributions of the scores of the two modes of operation. We deem the fuse of the two modalities can lead to real-world voice biometrics based applications which will be more accurate and thus more robust to spoofing attacks.

6 Acknowledgement

This work was partially supported by the H2020 OCTAVE Project entitled “Objective Control for TAlker VErification” funded by the EC with Grand Agreement number 647850.

The authors would like to thank Dr Md Sahidullah, Dr Nicholas Evans and Dr Tomi Kinnunen for their support in this work.

References

1. Aronowitz H., Hoory R., Pelecanos J., Nahamoo D.: New Developments in Voice Biometrics for User Authentication. In Proc. Interspeech (2011)
2. Hébert M., Sondhi M., Huang Y.: Text-Dependent Speaker Recognition. Book Section, Springer Handbook of Speech Processing, pp. 743-762 (2008)
3. Larcher A., Kong Aik Lee, Bin Ma, Haizhou Li: Text-dependent speaker verification: Classifiers, databases and RSR2015. Speech Communication, Volume 60, May 2014, Pages 56-77 (2014)
4. Reynolds D.A., Quatieri T.F., Dunn R.B.: Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, Volume 10, Issues 1–3, January 2000, Pages 19-41(2000)
5. Safavi S, Hanani A., Russell M., Jancovic P. and Carey M. J.: Contrasting the Effects of Different Frequency Bands on Speaker and Accent Identification. In IEEE Signal Processing Letters, vol. 19, no. 12, pp. 829-832 (2012)
6. Safavi S., Najafian M., Hanani A., Russell M.J., Jancovic P., and Carey M.J.: Speaker Recognition for Children's Speech. In Interspeech, pp. 1836-1839 (2012)

7. Ganchev T., Siafarikas M., Mporas I. and Stoyanova T.: Wavelet basis selection for enhanced speech parameterization in speaker verification. International Journal of Speech Technology, 17(1), pp.27-36 (2014)
8. Davis S. and Mermelstein P.: Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. In IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366 (1980)
9. Furui S.: Cepstral analysis technique for automatic speaker verification. In IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 29, no. 2, pp. 254-272 (1981)
10. Reynolds D.A. and Rose R.C.: Robust text-independent speaker identification using Gaussian mixture speaker models. In IEEE Transactions on Speech and Audio Processing, vol. 3, no. 1, pp. 72-83 (1995)
11. Campbell W.M., Campbell J.P., Reynolds D.A., Jones D.A., Leek T.R.: Phonetic Speaker Recognition with Support Vector Machines. Advances in Neural Information Processing Systems 16, Neural Information Processing Systems, NIPS 2003, December 8-13, 2003, Vancouver and Whistler, British Columbia, Canada (2003)
12. Campbell W.M., Sturim D.E. and Reynolds D.A.: Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters, IEEE, 13(5), pp.308-311 (2006)
13. Kenny P., Boulianne G., Ouellet P. and Dumouchel P.: Joint factor analysis versus eigenchannels in speaker recognition. Audio, Speech, and Language Processing, IEEE Transactions on, 15(4), pp.1435-1447 (2007)
14. Campbell J.P. and Reynolds D.A.: Corpora for the evaluation of speaker recognition systems. In Proc. of ICASSP'99, vol. 2, pp. 829-832 (1999)
15. Hermansky H. and Morgan N.: RASTA processing of speech. In IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 578-589 (1994)
16. Witten I.H., Frank E., Hall M.A.: Data Mining, Practical machine learning tools and techniques, 3rd Edition. Morgan Kaufmann, San Francisco (2011).