

Comparison of I-vector and GMM-UBM Approaches to Speaker Identification with TIMIT and NIST 2008 Databases in Challenging Environments

Musab T. S. Al-Kaltakchi^{1,2}, Wai L. Woo¹, Satnam S. Dlay¹, Jonathon A. Chambers¹

¹Communications, Sensors, Signal and Information Processing (ComS²IP) Group

School of Electrical and Electronic Engineering, Newcastle University, NE1 7RU, UK.

²Department of Electrical Engineering, College of Engineering, Al-Mustansiriya University, Baghdad, Iraq
{m.t.s.al-kaltakchi, Wai L. Woo, Satnam S. Dlay, Jonathon A. Chambers}@ncl.ac.uk

Abstract—In this paper, two models, the I-vector and the Gaussian Mixture Model-Universal Background Model (GMM-UBM), are compared for the speaker identification task. Four feature combinations of I-vectors with seven fusion techniques are considered: maximum, mean, weighted sum, cumulative, interleaving and concatenated for both two and four features. In addition, an Extreme Learning Machine (ELM) is exploited to identify speakers, and then Speaker Identification Accuracy (SIA) is calculated. Both systems are evaluated for 120 speakers from the TIMIT and NIST 2008 databases for clean speech. Furthermore, a comprehensive evaluation is made under Additive White Gaussian Noise (AWGN) conditions and with three types of Non Stationary Noise (NSN), both with and without handset effects for the TIMIT database. The results show that the I-vector approach is better than the GMM-UBM for both clean and AWGN conditions without a handset. However, the GMM-UBM had better accuracy for NSN types.

I. INTRODUCTION

The I-vector approach has received increasing interest for different research fields such as verification, language and emotional recognition. In [1], it was used for robust language identification and verification recognition, while [2] and [3] studied emotion and speech recognition; in [2], the I-vector results were compared with those for the GMM-UBM model. Different speaker identification challenges have been studied, for instance, with increasing numbers of speakers, channel variabilities, and the effects of noise and a handset. In [4], a GMM method was investigated for text independent speaker identification under noisy telephone channels, while elsewhere robust speaker identification in noisy environments has been studied, such as in [5], [6], [7], [8] and [9]. Moreover, [10] focused on the size of the population and the degradation produced from a noisy telephone channel and system, using the TIMIT and NTIMIT databases. However, few studies have involved a handset, AWGN, and NSN types in conjunction with fusion strategies. Handset variability effects for speaker recognition were studied in [11]. Session compensations with the I-vector approach were considered in [12] using Linear Discriminant Analysis (LDA), Nuisance Attribute Projection (NAP) and Within Class Covariance Normalization (WCCN) for text independent speaker identification. Nevertheless, this study lacked a large number of speakers, as only 50 self collected speakers were used. In [13], 1,000 speakers were

selected from YouTube to construct an I-vector speaker identification framework, but this non-standard database did not include noisy conditions.

In this paper, we establish two robust text independent closed set speaker identification systems: 1) a new fusion-based I-vector framework innovatively utilizing four feature combinations with two feature compensation methods, Capstrum Mean Variance Normalization (CMVN) and Feature Warping (FW) to the Mel Frequency Cepstral Coefficient (MFCC) and Power Normalization Cepstral Coefficient (PNCC) features. Then four combination of I-vectors are produced: FWMFCC, CMVNMFCF, FWPNC and CMVNPNC. It then exploited seven fusion types to yield multidimensional I-vectors, yielding a simple, fast, and efficient ELM classifier to identify the speakers; and 2) an extended evaluation including different NSN with/without a handset using score fusion based GMM-UBM. This exploited our previous study which evaluated clean and AWGN in [14]. In addition, late fusion techniques were employed with the GMM-UBM approach using maximum, mean and weighted sum to improve performance accuracy in clean speech, and to mitigate handset and background noise effects. However, in the current work, we make fair comparisons between the I-vector and modified fusion based on the traditional GMM-UBM approaches, using 120 speakers from each of the NIST 2008 and TIMIT databases: in total, 240 speakers with 2,400 speech utterances are employed. This work provides a clean speech evaluation for the NIST 2008 database for the stated approaches for a wide range of Gaussian Mixture Components (GMC). Also, an evaluation is presented for the TIMIT database under clean, AWGN, street traffic NSN, bus interior NSN, and crowd talking NSN with/without G.712 type handset at 16 kHz. This paper is structured as follows: Section II provides the I-vector and GMM-UBM frameworks; Section III gives the experiments and results; Section IV presents the related work; Section V includes the conclusions and the future work.

II. SPEAKER IDENTIFICATION SCHEMES USING GMM-UBM AND I-VECTOR

Fig. 1 shows two speaker identification systems using a previous study for the GMM-UBM approach [14] and the

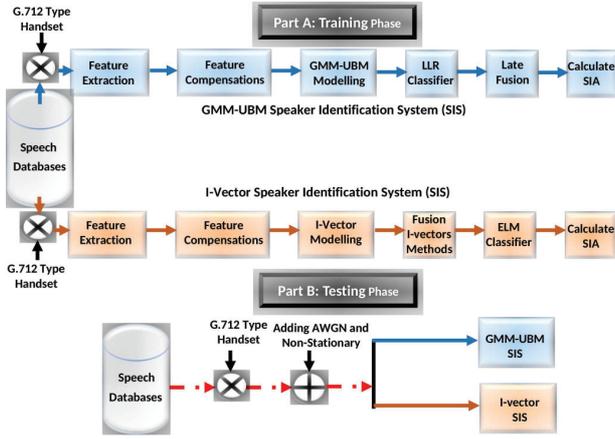


Fig. 1: Comparison of Two Speaker Identification Frameworks Using GMM-UBM and I-vector Approaches Evaluated under Different Environmental Conditions: Part A, training phase; Part B, testing phase.

proposed I-vector approach. Both systems were trained as in Part A in Fig. 1, and tested as in Part B. The full procedure for GMM-UBM evaluation can be found in [14], whereas, Table I shows that for the I-vector.

TABLE I: Summary of the algorithms for GMM-UBM and I-vector speaker identification systems, including speech pre-processing and background noise and handset effects

Steps	I-vector speaker identification system
Initial Step	All speech pre-processing and setting as well as feature extraction and compensation methods used are quite similar to the GMM-UBM approach in [14]
Step 1	Training the UBM using EM
Step 2	Compute the zero and first order Baum Welch Statistics
Step 3	Learning a Total Variability Space (TVS) from BWS
Step 4	Exploiting UBM, BW and TVS to extract the I-vector [15] [16] [17]
Step 5	We had 10 I-vectors/ speaker, six for training and the rest four I-vectors/ speaker for testing
Step 6	Fusion I-vectors using [maximum, mean, weighted sum, cumulative] for d-dimension and [interleaving, concatenation-2d] for 2d-dimension and finally [concatenation-4d] for 4d-dimension
Step 7	ELM is used for identifying speakers
Step 8	Calculate SIA using the true three tests out of four to identify each speaker
Step 9	The proposed system produces I-vector with d, 2d and 4d I-vector dimensions d=100, 200 this yields two sets [100, 200, 400] and [200, 400, 800] for TIMIT database but only [100, 200, 400] are presented in this paper. Only one set of NIST 2008 database [100, 200, 400]

A. Fusion Techniques for Combining I-vector Features

According to [15], the mathematical model for I-vector implementation is explained using:

$$S = \mu + \hat{T}i \quad (1)$$

$$i = (I + \hat{T}^t \Sigma^{-1} \hat{N}(u) \hat{T})^{-1} \cdot (\hat{T})^t \Sigma^{-1} \hat{F}(u) \quad (2)$$

where: u is the given speech utterance, $c = (1, \dots, C)$ which is the UBM mixture component, C is the number of mixture components and in this work is also denoted as $Mix = \{8, 16, 32, 64, 128, 256, 512\}$, F is the dimensionality of the acoustic feature vectors, i is the identity vector (I-vector), \hat{T} is the total variability matrix, μ is the independent speaker and channel supervector, and S is the speaker and channel dependent supervector. In addition, I is the identity matrix, Σ is a $(CF \times CF)$ diagonal covariance matrix, \hat{N} is a diagonal matrix of dimension $(CF \times CF)$, \hat{F} is the $(CF \times 1)$ dimension supervector and obtained by concatenating all first-order Baum Welch Statistics (BWS), and $(\cdot)^t$ denotes transpose. In this section, four I-vectors are employed based on the compensations for MFCC and PNCC features and denoted as i_{M1} , i_{M2} , i_{P1} and i_{P2} . The seven fusion techniques are explained below in equations (3)-(9):

$$i_{WSF} = \omega_k \times \dot{i}_j + (1 - \omega_k) \times \ddot{i}_j \quad (3)$$

where: $k = 1, 2, 3, 4$, while, $\omega_1, \omega_2, \omega_3$ and $\omega_4 = 0.7, 0.77, 0.8$ and 0.9 respectively, which have been found to yield a higher identification rate empirically.

$$i_{Maximum} = \max(\dot{i}_j, \ddot{i}_j) \quad (4)$$

$$i_{Mean} = (\dot{i}_j + \ddot{i}_j)/2 \quad (5)$$

$$i_{Cumulative} = \dot{i}_j + \ddot{i}_j \quad (6)$$

$$i_{Concatenated(2d)} = \begin{bmatrix} \dot{i}_j^t & \ddot{i}_j^t \\ \dot{i}_j^t & \ddot{i}_j^t \end{bmatrix} \quad (7)$$

$$i_{interleaving(2d)} = \begin{bmatrix} \dot{i}_{jj_{odd}}^t & \ddot{i}_{jj_{even}}^t \\ \dot{i}_{jj_{odd}}^t & \ddot{i}_{jj_{even}}^t \end{bmatrix} \quad (8)$$

$$i_{Concatenated(4d)} = [i_{M1} \quad i_{M2} \quad i_{P1} \quad i_{P2}] \quad (9)$$

and d = dimension of I-vector, where $j = 1, \dots, d$, $r = j+1, \dots, 2d$, $jj = 1, \dots, 2d$, \dot{i} is the I-vector for the normalized MFCC features with the highest SIA between (CMVNMFCF and FWMFCF) which are denoted by i_{M1} and i_{M2} , \ddot{i} is the normalized PNCC I-vector features which have the highest SIA between (FWPNCC and CMVNPNC), denoted by i_{P1} and i_{P2} . i_{WSF} , $i_{Maximum}$ and i_{Mean} are the weighted sum, maximum and mean fusion I-vectors with d -dimension I-vector. Also, $i_{Cumulative}$ is the Cumulative fusion with the d -dimension I-vector; $i_{Concatenated(2d)}$ and $i_{interleaving}$ are Concatenated and Interleaving fusion I-vectors with $2d$ -dimension I-vector; $i_{Concatenated(4d)}$ is the concatenated fusion I-vector with $4d$ -dimension for all I-vector features.

B. Identifying Speakers using Extreme Learning Machine

Recently, ELMs have been widely used in fields such as computer vision, biomedical engineering, and control and robotics, because they are simple, efficient and have impressive performance [18], [19] [20] and [21]. ELMs have single layer hidden node parameters which are randomly generated. The number of input nodes is equal to the Number I-vector Dimension (NID) and we used an almost equal Number of Hidden Neurons (NHN); however, we used different numbers when

it was necessary to achieve higher performance accuracy. In addition, the number of output neurons is equal to the number of classes, and in our work, 120 classes were used to represent 120 speakers. The ELM algorithm can be summarized as follows. We found the input weights and biases, which were randomly generated, then estimated the hidden layer of the output matrix; finally, we calculated the output weights.

III. EXPERIMENTAL RESULTS AND DISCUSSION

Two databases were exploited in this study: the TIMIT database and the 2008 NIST Speaker Recognition Evaluation Training Set Part 2. We exploited 120 speakers from TIMIT, with a total of 1,200 speech utterances, from which 480 were used for testing, and 720 for training, as in [14]. For the NIST 2008 database, we exploited 120 speakers of English using a microphone channel, and then the sampling frequency was converted from 8 to 16 kHz to mirror the TIMIT database. Only single speakers were selected by deleting the interviewers. In addition, each speech file was divided into ten equal lengths, and six out of ten were used for training (the rest for testing), with a fixed length of eight seconds. In this work, our experiments can be divided into two main parts based on NIST 2008 and TIMIT databases, in terms of the evaluations for I-vector and GMM-UBM approaches.

A. Part A: Simulations for NIST 2008

In this part, clean speech evaluations for the I-vector and GMM-UBM approaches were developed, as illustrated in Table II and Table III. The evaluations show the relationship between the SIA and the GMC, including the Number of Hidden Neurons (NHN), which is equal to Number of I-vector Dimensions (NID). According to both tables, it is clear that the highest SIA for the I-vector outperforms the GMM-UBM approach at a mixture size of 256, with 96.67% compared with 95.83%. However, the I-vector had lower results for small UBM mixture sizes.

B. Part B: The Simulations for TIMIT

The evaluations of the comparison of the TIMIT database for the I-vector and GMM-UBM techniques included various background noise types with/without a handset: clean speech, AWGN Without Handset (WOH), AWGN With Handset (WH), street traffic NSN WOH and WH, bus interior NSN WOH and WH, and finally, crowd talk NSN WOH and WH. In addition, a G.712 type handset at 16 kHz was used and each simulation was achieved by employing eleven I-vectors based on feature and fusion methods of four feature based FWMFCC, CMVNMFFC, FWPNC and CMVNMFFC with 100 I-vector dimension. There were seven other fusion methods; Weighted sum, Maximum, Mean, Cumulative I-vectors fusion with d-dimension (100), Concatenated and Interleaving fusion with 2d-I-vector dimension (200), and Concatenated fusion with 4d-dimension (400). In Fig. 2, the simulation illustrates GMM-UBM and I-vector comparisons in clean speech for TIMIT, and the best SIA for each mixture size was selected from both approaches regardless of feature or fusion

TABLE II: I-vector Speaker Identification Accuracy as a Function of UBM at Mixture Sizes $\{8, 16, 32, 64, 128, 256, 512\}$ using 1,200 speech utterances (120 Speakers) of microphone channel from NIST 2008

Simulation 1: I-vector Approach Using the NIST 2008 Database								
IV-Methods	NHN	8	16	32	64	128	256	512
i_{FW}	100	50	54.17	85.83	89.17	90.83	94.17	91.67
i_{CMVN}	100	51.67	58.33	80	86.67	95	95	91.67
i_{FW}	100	34.17	50	75	85	85.83	87.5	85.83
i_{CMVN}	100	34.17	50.83	78.33	84.17	89.17	93.33	89.17
$i_{WSF \omega_1}$	100	60	59.17	75.83	88.33	91.67	93.33	95
$i_{WSF \omega_2}$	100	50.83	59.17	75.83	92.5	89.17	92.5	94.17
$i_{WSF \omega_3}$	100	44.17	60.83	70.83	87.5	90	93.33	92.5
$i_{WSF \omega_4}$	100	49.17	55	77.5	84.17	87.5	92.5	91.67
$i_{Maximum}$	100	11.67	53.33	41.67	51.67	70.83	73.33	80
i_{Mean}	100	35	58.33	67.5	73.33	84.17	89.17	89.17
$i_{Cumulative-d}$	100	23.33	45	59.17	70	84.17	89.17	88.33
$i_{Concatenated-2d}$	200	50.83	65	81.67	89.17	95	96.67	94.17
$i_{Interleaving-2d}$	200	47.5	55	81.67	90.83	93.33	96.67	95
$i_{Concatenated-4d}$	400	25	42.5	65	80.83	87.5	88.33	87.5

TABLE III: GMM-UBM Speaker Identification Accuracy as a Function of Gaussian Mixture Components $\{8, 16, 32, 64, 128, 256, 512\}$ using 1,200 speech utterances (120 Speakers) with a microphone channel from NIST 2008 with a microphone channel

Simulation 2: GMM-UBM Approach for NIST 2008 Database							
IV-Methods	8	16	32	64	128	256	512
i_{FW}	90	89.17	92.5	95.83	93.33	92.5	94.17
i_{CMVN}	83.33	87.5	88.33	90.83	90	90.83	89.17
i_{FW}	83.33	86.67	87.5	87.5	89.17	88.33	88.33
i_{CMVN}	84.17	85	89.17	89.17	89.17	88.33	88.33
$i_{WSF \omega_1}$	89.17	90.83	94.17	95.83	95	95	95
$i_{WSF \omega_2}$	91.67	91.67	93.33	95	94.17	95	94.17
$i_{WSF \omega_3}$	90.83	91.67	93.33	94.17	94.17	95	94.17
$i_{WSF \omega_4}$	90.83	90.83	92.5	94.17	94.17	95	94.17
$i_{Maximum}$	90	86.67	93.33	93.33	92.5	92.5	91.67
i_{Mean}	88.33	90	90.83	91.67	92.5	94.17	92.5

method used. For small mixture sizes (8-64), the GMM-UBM outperformed the I-vector approach, while the I-vector showed better SIA compared with GMM-UBM at mixtures 128 and 256, but the highest SIA was with a rate 96.67% at mixture size 256, as explained in Fig. 2; thereby, the mixture size 256 was used for the evaluation for all noise conditions.

Fig. 3 and Fig. 4 explain the comparisons for both GMM-UBM and I-vector systems in AWGN, street NSN, bus NSN and crowd talking NSN with/without a handset for a wide range of SNR (0-30) dB. The continuous coloured curves with NSN square nodes for SNR levels represent the I-vector approach, while the dash-dot coloured curves with circle nodes for SNR levels depict the GMM-UBM approach. Furthermore, we used the same colour for the same noise type for both systems. The worst performance was using the AWGN because it has a constant noise spectrum, while bus NSN achieved

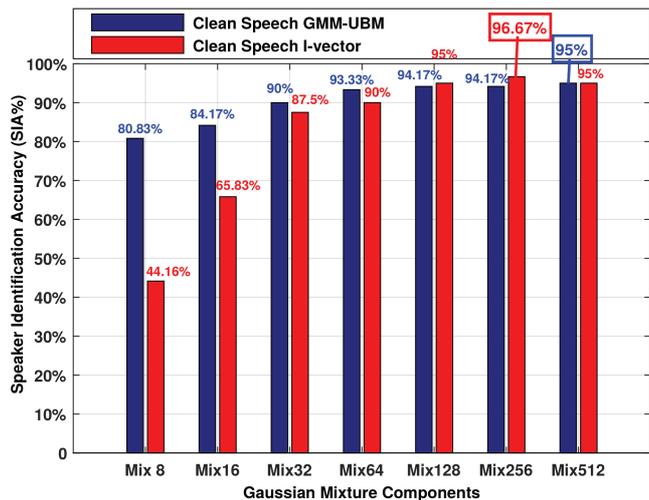


Fig. 2: Bar Chart Plot Comparisons Between SIA against Gaussian Mixture Components for GMM-UBM and I-vector Approaches in Terms of Clean Speech Using the TIMIT Database.

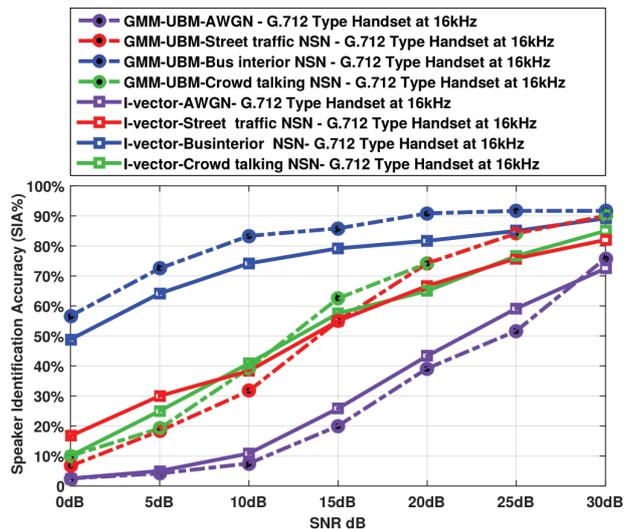


Fig. 4: Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN with G.712 Type Handset at 16 kHz at mixture size 256 Using the TIMIT Database.

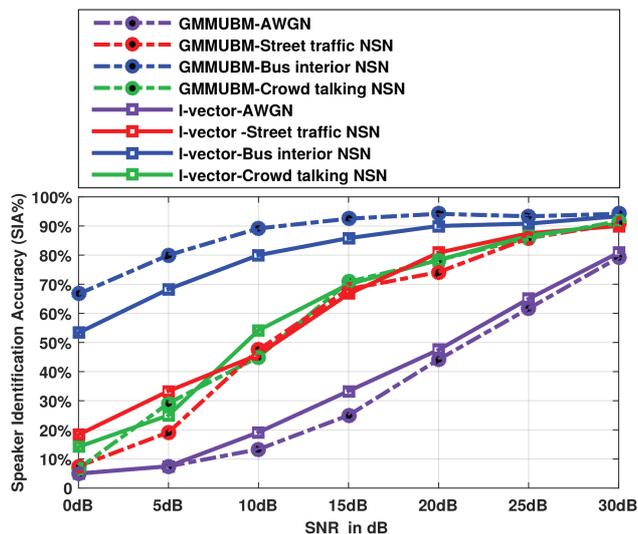


Fig. 3: Curve Plot Comparison GMM-UBM and I-vector Approaches for AWGN and NSN without handset at mixture size 256 Using the TIMIT Database.

less reduction in SIA in the presence of noise, of all other non-stationary noise types. On the other hand, both street and crowd talking NSN were accurate between AWGN and the bus NSN. The relationship between the SIA for both GMM-UBM and I-vector approaches is explained in Fig. 3 and Fig. 4 with different noise conditions with/without the handset.

IV. RELATED WORK

This section summarises the current work on I-vector and GMM-UBM approaches and other related work, alongside our

previous work and other state of the art methods [14], [22], [12], [13], [23], [24], and [5]. According to Table IV, the handset used was G.712 type at 16 kHz, and all proposed noise measurements in this table were at SNR 30 dB and mixture size 256. The best results of SIA were for clean speech, and our evaluations included various SNR levels, as explained in Fig. 3 and Fig. 4. Better SIA based on the I-vector was achieved compared with GMM-UBM under clean speech for both TIMIT and NIST 2008 databases. It also outperformed all clean speech measurements for other researchers. For TIMIT, the proposed I-vector approach achieved higher SIA under AWGN compared with the previous study on the GMM-UBM system, are compared with other work; in contrast, our previous work with GMM-UBM had better SIA than the proposed I-vector for AWGN WH, in line with other work. In addition, for non stationary background noise WH, the performance accuracy of GMM-UBM was better than the I-vector at SNR 30dB, but this reversed for some SNR levels. Finally, in [5], it seems the SIA for street noise was higher than in the proposed work, but this was achieved using a different noise database with 630 speakers.

V. CONCLUSIONS

This paper considered robust text independent speaker identification using I-vector approach for various background noises WH effects. The proposed work is compared fairly with the GMM-UBM approach and evaluated on the TIMIT and NIST 2008 databases for clean speech and also for TIMIT databases under nine different conditions, using both databases, eleven I-vectors together with feature and fusion-based methods. The system for both databases outperformed GMM-UBM techniques for clean speech, and also outperformed in TIMIT database under AWGN WOH, then it seems better for some SNR levels with street and crowd talking.

TABLE IV: Recent works Related to I-vector and GMM-UBM Techniques Speaker Identification

Approaches	The best feature / fusion based	Condition	The best SIA
(1) Proposed work using NIST 2008 with 120 Speakers (1,200 utterances)			
Proposed fusion based I-vector	Interleaving-2d	Clean	96.67%
Proposed fusion based GMM-UBM	Weighted sum	Clean	95.83%
(2) Proposed work. using TIMIT database with 120 Speakers (1,200 utterances)			
Proposed fusion based I-vector	Weighted sum	Clean	96.67%
	Concatenated-2d	AWGN	80.83% (30dB)
	Interleaving-2d	AWGN-WH	72.5% (30dB)
	FWMFCC-feature	Street traffic NSN	90% (30dB)
	Weighted sum	Street traffic NSN-WH	82% (30dB)
	Interleaving-2d	Bus interior NSN	93.33%(30dB)
	Weighted sum	Bus interior NSN-WH	89.17%(30dB)
	Concatenated-2d	Crowd talking NSN	90.83% (30dB)
	FWMFCC-feature	Crowd talking NSNWH	85% (30dB)
	Weighted sum	Street traffic NSN	90.83% (30dB)
Proposed fusion based GMM-UBM	Mean fusion	Street traffic NSN-WH	90% (30dB)
	Mean fusion	Bus interior NSN	94.17% (30dB)
	Weighted sum	Bus interior NSN-WH	91.67% (30dB)
	Weighted sum	Crowd talking NSN	91.67% (30dB)
	Weighted sum	Crowd talking NSNWH	90% (30dB)
	(3) Our previous study Using TIMIT with 120 Speakers [14][2016]		
Fusion based GMM-UBM	Weighted sum	Clean	95%
	Maximum fusion	AWGN	79.17% (30dB)
	FWMFCC-feature	AWGN-WH	75.83% (30dB)
(4) The State of The Art in [22] [2014] using NIST 2008 with 400 speakers			
I-vector Approach	Without fusion	Clean	49.5%
		White noise at 15 dB	39.3% (15dB)
GMM-UBM Approach	Without fusion	Clean	39.7%
		White noise at 15 dB	24.6% (15dB)
GMM-UBM-ZT norm	Without fusion	Clean	42.5%
		White noise at 15 dB	29.7% (15dB)
(5) In [12] [2014] using Corpus designed and MIT mobile phone with 50 speakers			
I-vector + LDA + WCCN	Without fusion	Clean	94.14% at (CDS)
		I-vector (400 dim) + LDA + WCCN	92.36% at (SVM)
(6) The state of the art in [13] [2014] using YouTube with 1,000 speakers			
I-vector retrieval	Without fusion	Clean	92% testing (10s) 96.1% testing (20s)
(7) The state of the art in [23] [2012] using TIMIT with 120 speakers			
Fusion Based GMM	Weighted sum	Clean	93.88%
(8) The state of the art in [24] [2011] using TIMIT database with 64 speakers			
GMM-UBM without fusion	Without fusion	Clean	94.5%
		AWGN-WH	74.2% (30dB)
(9) The state of the art in [5] [2007] using TIMIT database with 630 speakers			
New model with GMM at Mix 128	Without fusion	Clean	96.51%
		Street NSN (20dB)	92.86%

In contrast, for bus interior NSN, the GMM-UBM achieved less reduction in SIA compared with the I-vector approach. Additionally, fusion techniques may mitigate the reduction caused by different noise environments and the handset effect, whereas fusion weights generally seem to be the best of all feature and fusion methods used. In future work, we will also consider a new databases such as The Speakers in the Wild (SITW) Speaker Recognition Challenge database. We will also extend our evaluation of the NIST 2008 to include stationary and various NSN types with handset.

ACKNOWLEDGMENT

The first author Musab Tahseen Salahaldeen Al-Kaltakchi thanks the Ministry of Higher Education and Scientific Research, Iraq, for PhD funding.

REFERENCES

- [1] M. Li and S. Narayanan, "Simplified supervised I-vector modeling with application to robust and efficient language identification and speaker verification," *Computer Speech & Language*, vol. 28, no. 4, pp. 940–958, 2014.
- [2] L. Chen and Y. Yang, "Emotional speaker recognition based on I-vector through atom aligned sparse representation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7760–7764.
- [3] L. Mackov, J. Juh *et al.*, "Emotion recognition in I-vector space," in *2016 26th International Conference Radioelektronika (RADIOELEKTRONIKA)*. IEEE, 2016, pp. 372–375.
- [4] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *IEEE transactions on speech and audio processing*, vol. 3, no. 1, pp. 72–83, 1995.
- [5] J. Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds, "Robust speaker recognition in noisy conditions," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 5, pp. 1711–1723, 2007.
- [6] K. Kumar, Q. Wu, Y. Wang, and M. Savvides, "Noise robust speaker identification using bhattacharyya distance in adapted Gaussian models space," in *Signal Processing Conference, 2008 16th European*. IEEE, 2008, pp. 1–4.
- [7] Y. Goto, T. Akatsu, M. Katoh, T. Kosaka, and M. Kohda, "An investigation on speaker vector-based speaker identification under noisy conditions," in *Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on*. IEEE, 2008, pp. 1430–1435.
- [8] A. Khantemoori, M. Homayounpour, and M. Menhaj, "Speaker identification in noisy environments using dynamic Bayesian networks," in *Computer Conference, 2009. CSICC 2009. 14th International CSI*. IEEE, 2009, pp. 601–606.
- [9] L. Wang, K. Minami, K. Yamamoto, and S. Nakagawa, "Speaker identification by combining MFCC and phase information in noisy environments," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4502–4505.
- [10] D. A. Reynolds, M. A. Zissman, T. F. Quatieri, G. C. O'Leary, and B. A. Carlson, "The effects of telephone transmission degradations on speaker recognition performance," in *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, vol. 1. IEEE, 1995, pp. 329–332.
- [11] D. A. Reynolds, "The effects of handset variability on speaker recognition performance: Experiments on the switchboard corpus," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, vol. 1. IEEE, 1996, pp. 113–116.
- [12] T. Liu, K. Kang, and S. Guan, "I-vector based text-independent speaker identification," in *Intelligent Control and Automation (WCICA), 2014 11th World Congress on*. IEEE, 2014, pp. 5420–5425.
- [13] L. Schmidt, M. Sharifi, and I. L. Moreno, "Large-scale speaker identification," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1650–1654.
- [14] M. T. S. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Study of statistical robust closed set speaker identification with feature and score-based fusion," in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, June 2016, pp. 1–5.
- [15] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
- [16] D. Matrouf, N. Scheffer, B. G. Fauve, and J.-F. Bonastre, "A straightforward and efficient implementation of the factor analysis model for speaker verification," in *INTERSPEECH, 2007*, pp. 1242–1245.
- [17] P. Kenny, "A small footprint I-vector extractor," in *Odyssey*, 2012, pp. 1–6.
- [18] G. Huang, G.-B. Huang, S. Song, and K. You, "Trends in Extreme Learning Machines: a review," *Neural Networks*, vol. 61, pp. 32–48, 2015.
- [19] S. S. Nidhyananthan and R. S. S. Kumari, "A framework for multilingual text-independent speaker identification system," *Journal of Computer Science*, vol. 10, no. 1, p. 178, 2014.
- [20] E. Cambria, G.-B. Huang, L. L. C. Kasun, H. Zhou, C. M. Vong, J. Lin, J. Yin, Z. Cai, Q. Liu, K. Li *et al.*, "Extreme learning machines [trends & controversies]," *IEEE Intelligent Systems*, vol. 28, no. 6, pp. 30–59, 2013.
- [21] Y. Lan, Z. Hu, Y. C. Soh, and G.-B. Huang, "An extreme learning machine approach for speaker recognition," *Neural Computing and Applications*, vol. 22, no. 3-4, pp. 417–425, 2013.
- [22] R. Karadaghi, H. Hertlein, and A. Ariyaecinia, "Effectiveness in open-set speaker identification," in *2014 International Carnahan Conference on Security Technology (ICCSST)*. IEEE, 2014, pp. 1–6.
- [23] R. S. S. Kumari, S. S. Nidhyananthan *et al.*, "Fused MEL feature sets based text-independent speaker identification using Gaussian mixture model," *Procedia Engineering*, vol. 30, pp. 319–326, 2012.
- [24] R. Togneri and D. Püllella, "An overview of speaker identification: Accuracy and robustness issues," *Circuits and Systems Magazine, IEEE*, vol. 11, no. 2, pp. 23–61, 2011.