Speaker Verification Using Cosine Distance Scoring with i-vector Approach

Musab T. S. Al-Kaltakchi
Department of Electrical Engineering, College of Engineering, Mustansiriyah University
Baghdad, Iraq
Email: musab.tahseen@gmail.com

Raid Rafi Omar Al-Nima
Technical Engineering College of Mosul Northern Technical University
Mosul, Iraq
Email: raidrafi2@gmail.com

Mahmood Alfathe
Department of Computer and Information Engineering, Ninevah University
Mosul, Iraq
malfathe2012@my.fit.edu

Mohammed A. M. Abdullah
Dept. of Computer and Information Engineering, Ninevah University
Mosul, Iraq
mabdul@ieee.org

***Abstract*— In this paper, a robust yet simple speaker verification system is implemented. The speaker verification system is investigated employing the i-vector approach with the Cosine Distance Scoring (CDS) for system classification. In addition, to measure the system performance, Equal Error Rate (EER), Detection Error Trade-off (DET) Curve, Receiver Operating Characteristic (ROC) curve as well as Detection Cost Function (DCF) were utilized. Experimental results are conducted on the TMIT database using 64 randomly selected speakers. The proposed system utilizes the Mel Frequency Cepstral Coefficients (MFCC) and Power Normalized Cepstral Coefficients (PNCC) for feature extraction. In addition, features normalization methods such as Feature Warping (FW) and Cepstral Mean-Variance Normalization (CMVN) are used in order to mitigate channel effect noise. The speakers are modeled with the i-vector while CDS is used for classification. Experimental results demonstrate that the proposed system achieved promising results while being computationally efficient.**

Keywords*— Speaker verification, i-vector, Cosine Distance Scoring (CDS), TMIT database*

I. INTRODUCTION

Speaker recognition is one of the hot topics in biometrics and forensics where speakers are identified based on their unique speech characteristics. For speaker recognition, the i-vector can be employed in order to form an efficient yet simple speaker recognition system [1]. Speaker verification using i-vector is firstly proposed by Dehak [1]. Following this, several works in the literature utilized the i-vector for this purpose. In [2], a distance measure is employed in order to calculate the similarity between the target and test i-vectors for context-independent speaker verification task. The researchers in [2] demonstrate that the similarity distance attained better results compared with traditional methods such as the GMM-UBM and the Probabilistic Linear Discriminant Analysis (PLDA). In [3], in the GMMs of i-vector, and the Baum-Welch statistics are calculated through the weighting method. Then weighting parameters are produced for training and testing frames in order to redefine the optimization problem. In addition, a new updating rule is used based on combining weights for both the covariance and mean vectors matrices to the posterior probabilities and the system tested with the speaker in The Wild (SITW) database. Recently, in [4] an overview paper is presented for speaker verification task by employing Deep Neural Networks (DNN). The work in [5] presented a new technique based on short length utterance i-vector for speaker recognition. In order to compensate for the session variation, source-normalized LDA (SN-LDA), Linear Discriminant Analysis (LDA), and within-class covariance normalization (WCCN) are used. Furthermore, PLDA is also used for short utterance variation. The work in [6], gave an overview of the i-vector in terms of extracting the i-vector, where this application is used in different applications of speech processing such as the speaker, speech, language, and accent recognition. Also, in [7] the i-vector is evaluated based on forensic applications. Moreover, the Cosine Distance Scoring (CDS) without a score normalization method is employed in [8]. Similarly, in [9] robust speaker identification system is studied using the i-vector for different databases under stationary and non-stationary background noise and handset effect. Similarly, the i-vector for speaker identification task is presented in previous work in [10], [11] and [12]. Although the previous work proposed robust speaker recognition systems, however, the complexity of these methods is rather high. Therefore, in this paper we propose a robust yet simple method for speaker verification.

In this work, our contributions can be summarized as follows. Firstly, the CDS is tested for speaker verification due to its simplicity compared to the sophisticated neural network methods employed in the aforementioned works. Secondly, the system performance is tested with different modalities such as the Equal Error Rate (EER), Detection Error Trade-off (DET) Curve, Receiver Operating Characteristic (ROC) curve as well as Detection Cost Function (DCF). Experimental results on the TIMIT demonstrated that the proposed system achieved promising results while being computationally efficient.

978-1-7281-5249-3/20/$31.00 ©2020 IEEE

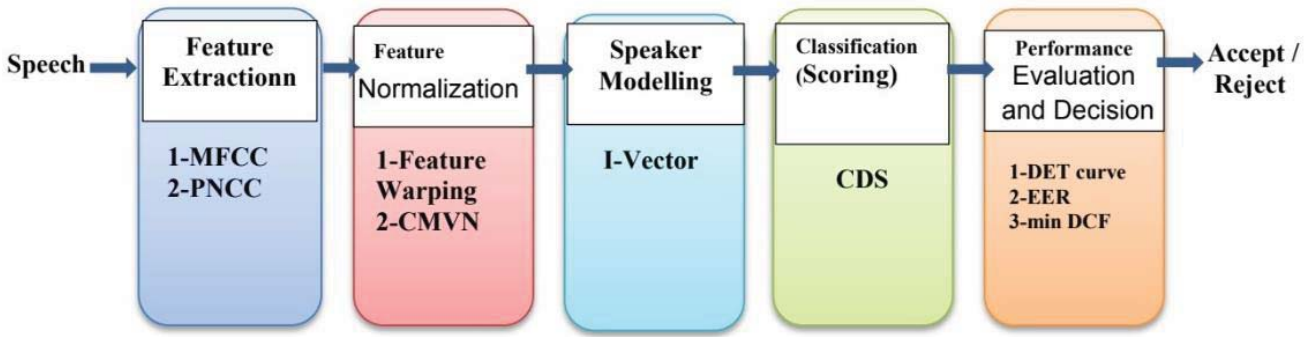

Fig. 1. The speaker verification block diagram using CDS to the I-vector approach

This paper is organized as follows. Section II presents the proposed method. In Section III the experimental results are given while section IV concludes this paper.

II. The Proposed Method

Figure 1 shows the main block diagram of the proposed speaker verification system. Two feature extraction methods are employed in the proposed system namely: MFCC and PNCC. Feature warping and CMVN are applied for feature normalization in order to compensate for channel effect noise. The CDS is used for classification and the performance of the system is measured with EER, DET and ROC curves.

A. Feature Extraction and Normalization Methods

Two types of feature extraction approaches are employed in our system: the Mel Frequency Cepstral Coefficients (MFCC) and the Power Normalized Cepstral Coefficients (PNCC). The main stages to implement the MFCC are: pre-emphasis, frame blocking and windowing, fast Fourier transform, Mel-scaled filter bank, and finally Cepstrum [9]. The PNCC features can be divided into three stages: initial processing, environmental processing and final processing. In the PNCC, the Gammatone filter is used instead of the Mel filter bank. In addition, the asymmetric noise filter is used to remove the noise level. Both implementations for MFCC and PNCC are explained with all details in [9]. Moreover, feature normalization methods are used such as Cepstral Mean and Variance Normalization (CMVN) and the feature warping methods in order to mitigate the linear channel effects [13]. Interested readers can refer to [14] for more information about the MFCC, PNCC and CMVN.

B. i-vector Approach

According to Dehak et al., i-vectors are derived as shown in the following "(1)" and "(2)" , [1].

$$S = u + U_x + V_y + D_z \quad (1)$$

$$S = u + T_V i \quad (2)$$

where: *D, V , U* are the diagonal residual, the eigen speaker variability, and eigen channel variability, respectively. *S* is the speaker and channel dependent supervector; *u* is the speaker and channel independent supervector; In addition, the factors *x, y* and *z* represent the channel, the speaker and residual vector, respectively. The total factor identity vector is *i*, while the low-rank matrix is total variability *Tv*. Finally, i-vectors can be calculated [1] as explained in "(3)":

$$i = (I + (T_v)^t \sum^{-1} \widehat{N} T_v)^{-1} (T_V)^t \sum^{-1} \hat{F} \quad (3)$$

where: $\sum$ is $(cd \times cd)$ dimension which represent the diagonal covariance matrix, I is the identity matrix, $\widehat{N}$ is $(cd \times cd)$ a diagonal matrix where c represents the number of mixture components and d is the dimension of the feature vectors, $\hat{F}$ is a supervector $(cd \times 1)$ of dimension achieved by concatenating the first-order Baum–Welch statistics, and $(.)^t$ denotes transpose [1].

The i-vector has been integrated with the feature extraction to represent the speech utterance in a compact way regardless of the length of the utterance. The MSR Identity Toolbox is used to calculate the i-vector. The main steps for extracting the i-vector can be summarized as follow [14].

Step 1: Forming a UBM from training data using the EM algorithm and Gaussian mixture components for the speakers.

Step 2: Extract the sufficient statistics for the training features using the Baum Welch (BW) algorithm.

Step 3: Learning a total variability subspace.

Step 4: Extract the i-vector.

C. Cosine Distance Scoring (CDS)

Cosine similarity is a determination of similarity between two vectors (non zero vectors) of an inner product space that calculates the cosine of the angle between them [2].

$$Similarity = \frac{A.B}{||A||.||B||} = \quad (4)$$

$$= \frac{\sum_{i=1}^{n} A_i.B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}} \quad (5)$$

where: A_i represents the target i-vector and B_i is the test i-vector.

III. EXPERIMENTS RESULTS AND DISCUSSION

The TIMIT database is used as a benchmark is our experiments where 64 speakers are randomly selected from 8 dialects. The parameters of the selected samples are 1- channel PCM, 16 KHz sampling rate with 8 major dialects of American English. We selected 4 speech files for males and 4 for females from each dialect region. In addition, each speaker has 10 speech utterances, five of them are selected for training and the rest are left for the testing purpose. Furthermore, in this work, four i-vector feature types are employed namely: feature warping-*MFCC; CMVN MFCC; feature warping PNCC, and CMVNPNCC.*

Table I shows the performance measurements including the EER, DCF, running time corresponding to different i-vector features for 100-dimension (FW MFCC- CMVN MFCC- FW

PNCC-CMVN PNCC) with various mixture sizes (128, 256, 512). It clear that the best results were achieved with CMVNMFCC where the lowest EER of 15.219% is achieved at mixture size of 512 compared with other features of the i-vector. The relatively high EER (15%-20%) can be attributed to using eight dialect regions some of these dialects are very close to each other which made the speaker recognition process challenging.

Figure 2 shows the performance measurements of the MFCC with feature warping as in the example at mixture size 128 and i-vector dimension 100. In addition, Fig. 2 includes three sub-figure representing the FAR vs. FRR, DET curve, and ROC curve. In Fig. 2a, the relationship between FAR and FRR illustrates the error with a variable threshold while the DFT curve represents the relation between the miss probability and false alarm probability. Finally, the ROC curve represents the relationship between the genuine speakers (1-FRR) with the imposter FAR.

TABLE I. THE EQUAL ERROR RATE (EER) OF DIFFERENT FEATURES OF I-VECTORS AT MIXTURE SIZES OF 128, 256, 512 AT 100 I-VECTOR DIMENSION.

I-vector with Feature Warping and MFCC features			
Mixture Size	**EER**	**Running Time (sec)**	**I-Vector dimension**
128	**19.359%**	**2772**	**100**
256	**16.057%**	**18106**	**100**
512	**15.846%**	**725773.5**	**100**
I-vector with CMVN and MFCC features			
128	**17.318**	**1645**	**100**
256	**15.688**	**8872.35**	**100**
512	**15.219**	**11668.53**	**100**
I-vector with Feature Warping and PNCC features			
128	**19.451**	**8965.54**	**100**
256	**16.184**	**11936.79**	**100**
512	**15.938**	**34667.32**	**100**
I-vector with CMVN and PNCC features			
128	**17.7148**	**11118.455**	**100**
256	**16.343**	**11935.04**	**100**
512	**16.478**	**13190.84**	**100**

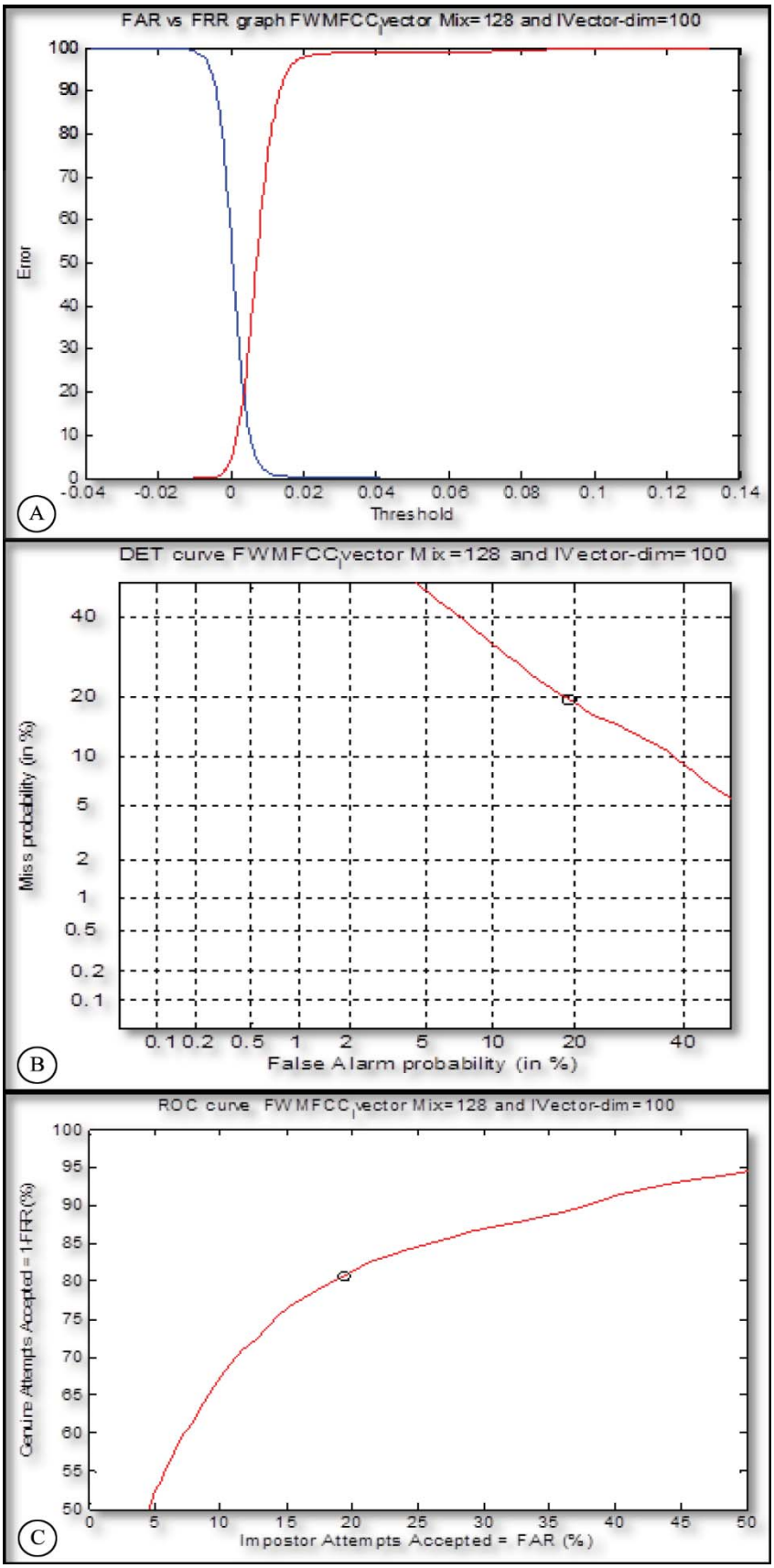

Fig. 2. System performance curves at a mixture size of 128 and I-vector dimension= 100: (A) FAR (blue) vs. FRR (red); (B) DET curve; (C) ROC curve where the white dot represents the ERR

IV. CONCLUSIONS

In this paper, CDS with the i-vectors was exploited for the speaker verification task. In addition, four features were utilized to construct the i-vectors namely: FW-MFCC, CMVNMFCC, FW-PNCC, CMVN-PNCC. Moreover, FAR and FRR, DET Curve, ROC and speaker detection are exploited to measure the performance of speaker verification. According to the results, the lowest EER was achieved using the CMVNMFCC features. Furthermore, the reported EER is rather high and this because difficult and close dialects were used to judge the performance of our system to simulate real-life scenarios.

REFERENCES

[1] 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.

[2] S. Hourri and J. Kharroubi, "A novel scoring method based on distance calculation for similarity measurement in text-independent speaker verification," Procedia computer science, vol. 148, pp. 256–265, 2019.

[3] X. Zhang, X. Zou, M. Sun, T. F. Zheng, C. Jia, and Y. Wang, "Noise robust speaker recognition based on adaptive frame weighting in GMM for i-vector extraction," IEEE Access, 2019.

[4] A. Irum and A. Salman, "Speaker verification using deep neural networks: A," International Journal of Machine Learning and Computing, vol. 9, no. 1, 2019.

[5] A. Kanagasundaram, D. Dean, S. Sridharan, J. Gonzalez- Dominguez, J. Gonzalez-Rodriguez, and D. Ramos, "Improving short utterance i-vector speaker verification using utterance variance modelling and compensation techniques," Speech Communication, vol. 59, pp. 69–82, 2014.

[6] P. Verma and P. K. Das, "i-vectors in speech processing applications: a survey," International Journal of Speech Technology, vol. 18, no. 4, pp. 529–546, 2015.

[7] M. I. Mandasari, M. McLaren, and D. A. van Leeuwen, "Evaluation of i-vector speaker recognition systems for forensic application," 2011.

[8] N. Dehak, R. Dehak, J. R. Glass, D. A. Reynolds, and P. Kenny, "Cosine similarity scoring without score normalization techniques.," in Odyssey, p. 15, 2010. [9] M. T. S. Al-Kaltakchi, "Robust text independent closed set speaker identification systems and their evaluation," 2018.

[10] M. T. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and . A. Chambers, "Speaker identification evaluation based on the speech biometric and i-vector model using the TIMIT and NTIMIT databases," in 2017 5th International Workshop on Biometrics and Forensics (IWBF), pp. 1–6, IEEE, 2017.

[11] M. T. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Multi-dimensional i-vector closed set speaker identification based on an extreme learning machine with and without fusion technologies," in 2017 Intelligent Systems Conference (IntelliSys), pp. 1141–1146, IEEE, 2017.

[12] M. T. Al-Kaltakchi, W. L. Woo, S. S. Dlay, and J. A. Chambers, "Comparison of i-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments," in 2017 25th European Signal Processing Conference (EUSIPCO), pp. 533–537, IEEE, 2017.

[13] M. T. Al-Kaltakchi, R. R. O. Al-Nima, M. A. M. Abdullah, & H. N. Abdullah, (2019). Thorough evaluation of TIMIT database speaker identification performance under noise with and without the G. 712 type handset. International Journal of Speech Technology, 22(3), 851-863.

[14] S. O. Sadjadi, M. Slaney, and L. Heck, MSR identity toolbox: A Matlab toolbox for speaker-recognition research, Speech and Language Processing Technical Committee Newsletter, vol. 1, no. 4, 2013..