Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2020) 28: 1236 – 1245
© TÜBİTAK
doi:10.3906/elk-1906-118

Comparisons of extreme learning machine and backpropagation-based i-vector approach for speaker identification

Musab T.S. Al-KALTAKCHI[1,*], **Raid R. O. AL-NIMA**[2], **Mohammed A.M. ABDULLAH**[3]
[1]Department of Electrical Engineering, College of Engineering, Mustansiriyah University, Baghdad, Iraq
[2]Technical Engineering College of Mosul, Northern Technical University, Mosul, Iraq
[3]Department of Computer and Information Engineering, College of Electronics Engineering, Ninevah University, Mosul, Iraq

Received: 18.06.2019 • **Accepted/Published Online:** 26.12.2019 • **Final Version:** 08.05.2020

Abstract: The extreme learning machine (ELM) is one of the machine learning applications used for regression and classification systems. In this paper, an extended comparison between an ELM and the backpropagation neural network (BPNN)-based i-vector is given in terms of a closed-set speaker identification task using 120 speakers from the TIMIT database. The system is composed of the mel frequency cepstral coefficient (MFCC) and power normalized cepstral coefficient (PNCC) approaches to form the feature extraction stage, while the cepstral mean variance normalization (CMVN) and feature warping are applied in order to mitigate the linear channel effect. The system is utilized with equal numbers of speakers of both genders with 120 speakers with eight dialects from the TIMIT database. The results demonstrate that the combination of the i-vector with the ELM for different features has the highest speaker identification accuracy (SIA) compared with the combination of the BPNN with the i-vector. The results also show that the i-vector with ELM approach is faster than the BPNN-based i-vector and it has the highest SIA.

Key words: Speaker recognition, extreme learning machine, TIMIT database, i-vector

1. Introduction

There are several open issues in machine learning techniques, such as intensive human interference, slow learning speed, and poor learning [1]. The extreme learning machine (ELM) appears to solve these problems [1]. The researchers in [2–5] presented a review in terms of ELM and dealt with theory, algorithms, trends, and applications. In the aforementioned papers, the mathematical model of the ELM was explained. In addition, the previous works reported that the ELM is efficient, simple, and widely used in various domains such as system identification, computer vision, biomedical engineering, robotics, and control. The ELM algorithm is utilized to attain more accurate results and to save time during regression and classification.

On the other hand, the backpropagation neural network (BPNN) is widely used as a supervised learning technique in machine learning to extract information. However, in [6] the BPNN showed poor time and space complexity. In [7], a description of the algorithm as well as an example of a nonlinear model-based BPNN classifier including the MATLAB code were given. In [8] the authors employed three methods to overcome the detection of nontechnical losses (NTLs) using an online sequential EML (OS-ELM), BPNN, and support vector machine. A comparative study was conducted for two supervised machines learning the BPNN and the

[*]Correspondence: musab.tahseen@gmail.com

OS-ELM. The results showed that the OS-ELM outperformed the BPNN with hit rates of 51.38% and 36.07%, respectively.

An overview in terms of the state of the art using the i-vector was presented in [9]. In addition, a survey of different applications of the i-vector approach in the speech processing domain was also given. Although the ELM was successfully employed for speaker identification in our previous studies in [10–12], the depicted models are time-consuming and comparisons with other neural networks methods were not considered.

This paper provides a combination of the i-vector approach with the ELM for speaker identification and the main motivations for this combination are as follows: This combination gives higher speaker identification accuracy (SIA) than each technique alone. In addition, the ELM has a faster training performance compared with other neural network techniques such as BPNN and deep learning, and the ELM has the ability to randomly generate the weight and the bias. Moreover, the performance with the ELM does not suffer from local error problems such as with the BPNN. Furthermore, the ELM has faster learning speed compared with the BPNN or deep learning, where it is straightforward, so no feedback is required and it thereby reduces the training and testing time required. On top of that, combining the ELM with the i-vector gives better accuracy for speaker identification. We compare two classifiers based on ELM and BPNN in terms of their speed and SIA for evaluating closed-set speaker identification performance. In summary, our contributions are as follows:

- The MFCC and PNCC are exploited to form a robust feature extractor model while the CMVN and feature warping are used for normalization.
- The ELM and BPNN-based i-vector approaches are employed for classification.
- A fair comparison is given between the BPNN and ELM-based i-vector approaches under the same conditions.

This paper is organized as follows: Section 2 discusses the acoustic features. Section 3 introduces the concept of the BPNN. Section 4 presents the concept of the ELM. Section 5 presents the main concept of the i-vector. Experiments and results with discussion are given in Section 6. Finally, Section 7 concludes this paper.

2. The main block diagram

Figure 1 shows the main block diagram using MFCC and PNCC for feature extraction, CMVN and feature warping for feature normalization, the i-vector as an acoustic model, and the BPNN and ELM as the classifiers.

In the proposed system, a preemphasis filter with 0.96 emphasis coefficients was applied for both features to compensate the high frequency that was suppressed while producing the speech signal. A Hamming windowing with 16 ms frame duration is used with 8 ms overlap between the frames. Then MFCC and PNCC features with 16 feature dimensions were used in the feature extraction stage.

The MFCC extraction operator can be divided into five stages: preemphasis, frame blocking and windowing, fast Fourier transform, mel-scaled filter bank, and cepstrum evaluation. Likewise, the PNCC implementation starts by preemphasizing the speech signal, and then a short-time Fourier transform (STFT) is performed using a Hamming window of 16 ms duration with 8 ms frame period. The squared magnitude of the STFT outputs is passed to a 40-channel Gammatone filter bank. The center frequencies of the Gammatone filter are linearly spaced in the equivalent rectangular bandwidth (ERB) auditory frequency scale. The output is then temporally masked and spectrally smoothed. Finally, the mean power is estimated for each frame and the running average for time-frequency normalization is used.

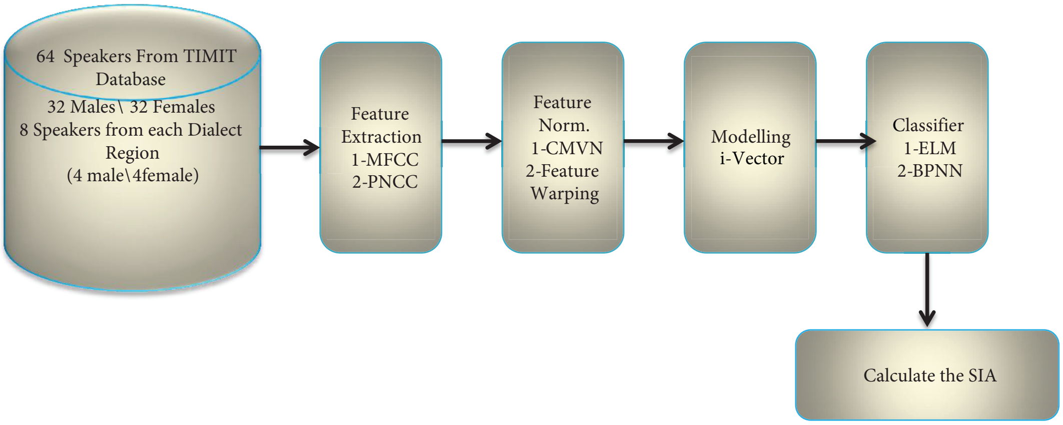

Figure 1. The main block diagram.

Feature normalization is also adopted by using FW and CMVN for both MFCC and PNCC features. FW and CMVN approaches are used to improve the SIA for the system and reduce the sensitivity between the types of telephone handsets and could also help to reduce the linear channel effects. The concepts of the BPNN, ELM, and i-vector are presented in the next sections, respectively.

3. The concept of BPNN

In machine learning, the BPNN is used to construct a classifier in which the set of decision rules is formulated to model the nonlinear functions [7]. The BPNN model involves three sections: the input layer, hidden layers (one or more layer), and the output layer. In the input layer, the number of neurons is equal to the number of attributes (i-vector dimension), while in the output layer, the number of neurons is equal to the number of nonlinear functions in order to construct the classifier from the decision rules. On the other hand, in the hidden layers, the number of neurons is randomly selected [7]. Figure 2 shows the structure of the BPNN.

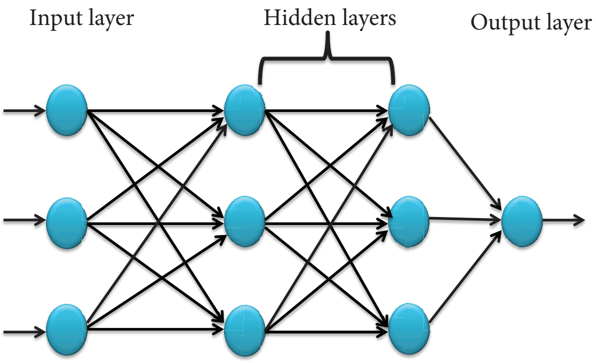

Figure 2. The general structure of backpropagation neural network (BPNN).

4. The concept of ELM

The ELM was originally proposed for a single hidden layer based on a feedforward network with randomly generated hidden neurons. In the ELM, no iterations or tuning are needed in the hidden layer [13]. In addition, all hidden nodes are independent with training data and target functions and have universal approximation capability. The output weights of the ELM may be determined with and without iterations. The ELM can be applied for multiclass classification and regression and it is efficient in sequential, batch, and incremental learning. The ELM has been widely used in different applications such as in signal processing, biometrics, image processing, bioinformatics, and brain/computer interface [13]. Figure 3 shows the structure of the ELM.

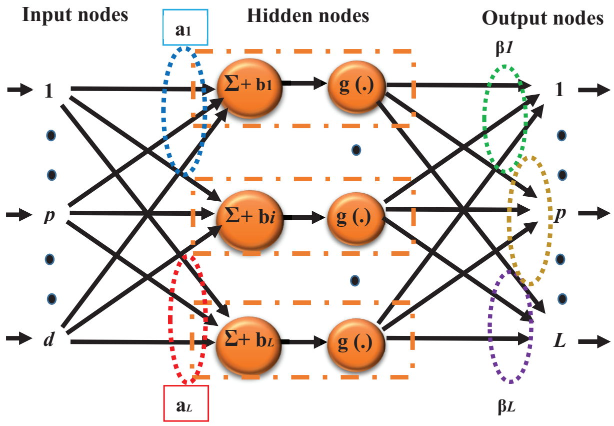

Figure 3. Structure of a single-layer feedforward extreme learning machine with d input dimension, *L* hidden nodes, and *L* outputs.

According to Figure 3, the implementation for the ELM network scheme is illustrated as shown in [1, 14, 15]. The number of input nodes is equal to the i-vector dimension. In addition, different neuron numbers are also selected when it is necessary to achieve higher performance accuracy. However, the number of output neurons is equal to the number of classes, and in this work we employ 120 classes to represent 120 speakers. The sigmoid function is used as the activation function in this paper. To calculate the output weights of the ELM, the following equations are used to find regularized values [1, 14, 15]:

$$\boldsymbol{H}\boldsymbol{\beta} = \check{\boldsymbol{T}}, \tag{1}$$

$$\boldsymbol{\beta} = \boldsymbol{H}^{\dagger}\check{\boldsymbol{T}}, \tag{2}$$

$$\boldsymbol{\beta} = \left(\frac{\boldsymbol{I}}{\boldsymbol{r}} + \boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\right)^{-1} \boldsymbol{H}^{T}\check{\boldsymbol{T}}, \tag{3}$$

where $\boldsymbol{\beta}$ is the output weight matrix $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_L]^{\boldsymbol{T}}$, $H = \left[h^T(\boldsymbol{x}_1), ..., h^T(\boldsymbol{x}_N)\right]^T$, $h(\mathbf{x}) = [\, g\left(a_1^T\mathbf{x} + b_1\right)$, ..., $g\left(a_L^T\mathbf{x} + b_L\right) \,]$ is the hidden node outputs, $\mathbf{x}$ is the input vector, and $g\left(a_j^T\mathbf{x} + b_j\right)$ is the output of the jth hidden node. N is the number of training samples, $\check{\boldsymbol{T}}$ is the target label matrix, $\check{\boldsymbol{T}} = \left[\check{t}_1, ..., \check{t}_N\right]^T$, $\boldsymbol{H}^{\dagger}$ is the

Moore–Penrose generalized inverse of matrix $\boldsymbol{H}$, and $\boldsymbol{r}$ is the regularization factor. The target label matrix $\check{\boldsymbol{T}}$ is represented by $(no.\ of\ speakers \times no.\ of\ training\ samples\ (L \times N))$. Each i-vector with 100 dimensions is entered into the ELM classifier. The outputs represent the speaker classes in this work with 120 speakers. In addition, the actual outputs are real numbers, while the maximum is selected for the output vector and this maximum refers to the identified speaker position. Ultimately, the SIA is determined.

The main steps of the ELM classifier and SIA calculation can be summarized as follow:

Step 1: Compute the i-vectors from training and testing data.

Step 2: Assume the number of output neurons is equal to the number of speakers.

Step 3: Form the target matrix for training.

Step 4: Randomly generate the input weights and biases for the hidden neurons.

Step 5: Suppose the number of hidden neurons is equal to the i-vector dimension.

Step 6: Calculate the hidden neurons output, matrix H.

Step 7: Calculate the output weights.

Step 8: Apply the testing data to the trained ELM.

Step 9: Calculate the genuine speaker identified for each training example by the position of the maximum in the output of the ELM.

Step 10: Calculate the SIA.

Table 1 shows the main differences between the BPNN and ELM.

Table 1. Comparisons of BPNN and ELM.

BPNN	ELM
Takes a long time for training	Faster learning speed
May suffer from local minimum (the point that has the lowest mean square error (MSE), where the network starts to train the neurons and the rate of MSE decreases)	Does not suffer from local minimum
Iterations required during training	No need to iterate or tune
Cannot easily add or remove trained subjects	Can easily add or remove trained subjects

5. The concept of i-vector

The main stages of constructing the i-vector are as follows [16]: first, forming the UBM from training data using the expectation maximization (EM) algorithm and Gaussian mixture components for the speakers; second, extracting the sufficient statistics for the training features using the Baum–Welch (BW) algorithm; and third, learning the total variability subspace. Finally, we extract the i-vector. The i-vector (identity vector) has a fixed length with a low dimension, which is used for modeling speakers for both MFCC and PNCC features. The coupling between the i-vector and ELM and BPNN is utilized in this paper in order to improve the identification rate. The mathematical model to establish the i-vector is proposed in [17] and is explained in equations (4), (5), and (6), where equation (4) is the model of joint factor analysis and equation (5) illustrates the relation between the total factor and total variability matrix. In addition, equation (6) is used to calculate the i-vector.

$$\boldsymbol{S} = \boldsymbol{\mu} + \boldsymbol{U}\boldsymbol{x} + \boldsymbol{V}\boldsymbol{y} + \boldsymbol{D}\ \boldsymbol{z}, \tag{4}$$

$$\boldsymbol{S} = \boldsymbol{\mu} + \boldsymbol{T_V}\ \ \boldsymbol{i}, \tag{5}$$

where $\boldsymbol{S}$ is a dependent supervector for both speaker and channel; $\boldsymbol{V}$, $\boldsymbol{U}$, and $\boldsymbol{D}$ are the speaker, channel, and diagonal residual matrices, respectively; and $\boldsymbol{\mu}$ is the independent supervector for the speaker and channel. In addition, factors $\boldsymbol{y}$, $\boldsymbol{x}$, and $\boldsymbol{z}$ represent the speaker, channel, and residual factor, respectively. Furthermore, the total factor identity vector is $\boldsymbol{i}$ and $\boldsymbol{T_V}$ is the total variability low rank matrix. Finally, i-vectors can be determined as in [17] and are described in ((6):

$$\boldsymbol{i} = (\boldsymbol{I} + (\boldsymbol{T_V})^{\boldsymbol{T}}\ \boldsymbol{\Sigma^{-1}}\ \hat{\boldsymbol{N}}(\boldsymbol{u})\ \boldsymbol{T_V})^{-1}\ (\boldsymbol{T_V})^T\ \boldsymbol{\Sigma^{-1}}\check{\boldsymbol{F}}(\boldsymbol{u}), \tag{6}$$

where $\boldsymbol{i}$ is the identity vector (i-vector); $\boldsymbol{I}$ is the identity matrix; $\boldsymbol{\Sigma}$ is the $(CF \times CF)$ diagonal covariance matrix, where C is the number of mixture components; $\boldsymbol{\hat{N}(u)}$ is a diagonal matrix of dimension $(CF \times CF)$; and the dimension of the feature vectors is F. Moreover, $\check{\boldsymbol{F}}$ is the $(CF \times 1)$ dimension supervector, which is acquired by concatenating all first-order Baum–-Welch statistics, and $\boldsymbol{u}$ is the given speech utterance and $(.)^T$ denotes transpose.

6. Experimental results and discussion

In this paper, 120 speakers with equal numbers from both genders are selected randomly from eight dialects of American English for the TIMIT database. The data source was microphone speech with sampling frequency equal to 16 kHz. The dialect regions (DR) include New England, Northern, North Midland, Southern, New York City, Western, and Army Brat, corresponding to: DR1, DR2,..., DR8, respectively. Each speaker has ten speech utterances, five of them for training and the remaining used for testing.

For the 120 speakers, 720 utterances are used for training while 480 utterances are used for testing. Hence, the i-vector is constructed and classified using the ELM to identify speakers by employing 120 classes to represent 120 speakers.

In addition, instead of the ELM, BPNN is used for comparison purposes. Furthermore, the best hidden layer is selected to achieve higher SIA, which is proved experimentally to be 100. In addition, the best SIA is obtained with UBM mixture size of 256. Table 2 shows the SIA for the i-vector ELM approach where the best results are empirically attained under 100 hidden neurons with 100 i-vector dimensions where the sizes of the UBM are 128, 256, and 512. The highest SIA is reported to be 92.95% using the i-vector with CMVN-MFCC features and mixture size of UBM equal to 256. In addition, Table 2 illustrates the processing time of the current method, which is 40-47 s. Furthermore, the results with MFCC are better than the corresponding results with MFCC. On the other hand, Table 3 shows the results for the i-vector using the BPNN approach with UBM mixture sizes of 128, 256, and 512. The results show that the highest SIA obtained is equal to 86.235% with CMVN MFCC features. It has been noticed that the result with MFCC also outperformed the results of PNCC features. Moreover, the comparisons between the ELM and BPNN and the i-vector are explained through comparisons between Table 2 and Table 3. The results obtained by employing the ELM are better than those obtained with the BPNN. In addition, the timing requirements of the BPNN are about 10 times larger than the

Table 2. The SIA for the i-vector based on ELM Approach at UBM Mixture Sizes 128, 256, 512 with 100 hidden neurons and 100 i-vector dimension.

Method	Features	UBM Mixture size	SIA	Consuming Time
i-vector-ELM	MFCC+Feature Warping	128	91.13%	42.168 Sec
i-vector-ELM	PNCC+Feature Warping	128	85.41%	40.346 Sec
i-vector-ELM	MFCC+CMVN	128	92.77%	40.233 Sec
i-vector-ELM	PNCC+CMVN	128	87.95%	46.94Sec
i-vector-ELM	MFCC+CMVN	256	**92.95%**	41.091 Sec
i-vector-ELM	PNCC+CMVN	256	87.81%	47.156 Sec
i-vector-ELM	MFCC+Feature Warping	256	92.29%	41.4 Sec
i-vector-ELM	PNCC+Feature Warping	256	87.16%	49.468 Sec
i-vector-ELM	MFCC+Feature Warping	512	89.42%	42.625 Sec
i-vector-ELM	PNCC+Feature Warping	512	84.72%	46.103 Sec
i-vector-ELM	MFCC+CMVN	512	91.75%	42.599 Sec
i-vector-ELM	PNCC+CMVN	512	84.71%	46.841 Sec

Table 3. The SIA for the i-vector based on BPNN Approach at UBM Mixture Sizes 128, 256, 512 with 100 hidden neurons and 100 i-vector dimension.

Method	Features	UBM Mixture size	SIA	Consuming Time
i-vector-BPNN	MFCC+ Feature Warping	128	64.69%	179.849 Sec
i-vector-BPNN	PNCC+Feature Warping	128	59.375%	256.029 Sec
i-vector-BPNN	MFCC+CMVN	128	69.375%	286.954 Sec
i-vector-BPNN	PNCC+CMVN	128	58.75%	371.909 Sec
i-vector-BPNN	MFCC+Feature Warping	256	74.437%	311.694 Sec
i-vector-BPNN	PNCC+Feature Warping	256	71.375%	524.828 Sec
i-vector-BPNN	MFCC+CMVN	256	**86.235%**	414.908 Sec
i-vector-BPNN	PNCC+CMVN	256	73.75%	491.187 Sec
i-vector-BPNN	MFCC+Feature Warping	512	82.23%	914.975 Sec
i-vector-BPNN	PNCC+Feature Warping	512	80.312%	985.730 Sec
i-vector-BPNN	MFCC+CMVN	512	84.27%	776.145 Sec
i-vector-BPNN	PNCC+CMVN	512	81%	813.46 Sec

timing corresponding to the ELM. Hence, using the ELM gives faster learning performance while maintaining classification accuracy.

Table 4 shows the SIA for clear speech of 120 speakers from the TIMIT database with a wide range of Gaussian mixture components. In addition, Table 5 depicts the SIA for noisy speech of 120 speakers from the TIMT database using the i-vector-ELM approach. Figure 4 shows the SIA versus Gaussian mixture components for both the i-vector-ELM approach and i-vector-BPNN approach. It can be seen from Figure 4 that the i-vector ELM approach outperforms the I-vector BPNN approach. In summary, all the parameters of the setups are illustrated in Table 6.

Table 4. The SIA for Clean Speech of 120 Speakers from TIMIT Database with 100 Dimension of I-vector-ELM Approach.

The SIA for Clean Speech to 120 Speakers of TIMIT Database for I-vector – ELM Approach (I-vector Dimension is 100)						
Mix 8	Mix 16	Mix 32	Mix 64	Mix 128	Mix 256	Mix 512
35%	61.67%	87.5%	90%	93.33%	95.83%	95.83%

Table 5. The SIA for Noisy Speech of 120 Speakers from TIMIT Database with 100 Dimension of I-vector-ELM Approach.

The SIA for 120 Speakers with 100 Dimensions of I-vector ELM Approach Under AWGN						
0dB	5dB	10dB	15dB	20dB	25dB	30dB
5%	7.5%	19.17%	33.33%	47.5%	59.17%	76.67%
The SIA for 120 Speakers with 100 Dimensions of I-vector ELM Approach Under Street Traffic NSN						
13.33	23.33%	41.67%	61.67%	77.5%	81.67%	87.5%
he SIA for 120 Speakers with 100 Dimensions of I-vector ELM Approach Under Bus Interior NSN						
53.33%	63.33%	78.33%	85.83%	86.67%	90.83%	92.5%
The SIA for 120 Speakers with 100 Dimensions of I-vector ELM Approach Under Crowd Talking NSN						
11.67%	23.33%	47.5%	64.17%	74.17%	84.17%	85.83%

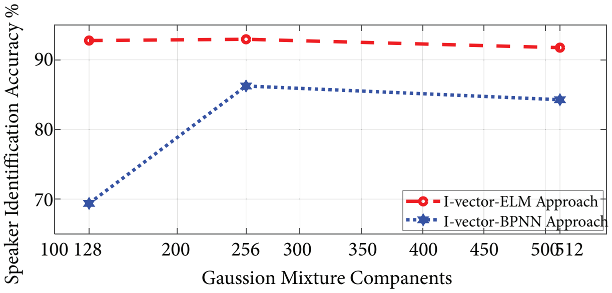

Figure 4. Comparisons between the i-vector ELM and BPNN approaches in terms of the Gaussian mixture components.

7. Conclusions

In this paper, a fair comparison between the ELM and BPNN based the i-vector was given in terms of a closed-set speaker identification task using the TIMIT database. The results showed that the ELM approach achieved better timing performance with higher SIA compared to the BPNN. In addition, the results obtained using the MFCC features gave better identification rates compared with the PNCC features. Moreover, various mixture sizes of UBM are considered and the highest accuracy is achieved using the UBM mixture of size equal to 256.

Table 6. Parameters and setup used for 120 speakers selected from TIMIT database.

Aspects	Parameters and experimental setup
Sampling frequency	16000
Window type	Hamming
Frame length	16 ms
Frame shift	8 ms
Preemphasis factor	0.96
Databases	TIMIT
Number of speakers	120 speakers for each database, total 480 speakers for all databases
Total speech utterances used	1200 for each database
Language	English
Data source(s)	Microphone speech for TIMIT
No. of samples per speaker	10 for each of TIMIT
Testing samples for each database	Total 480 utterances
Training samples for each database	Total 720 utterances
Average sample duration	8 seconds in length, 129250 (for each speech utterance in both training and testing); all speech samples were taken with fixed length of 129250 samples; concatenation is applied where necessary
Features	MFCC and PNCC
Feature dimensions	16
Feature normalization	FW and CMVN
Modeling	i-vector
Classifier	ELM and BPNN
UBM mixture sizes	{8, 16, 32, 64, 128, 256, 512}
System environment	Original speech recordings, AWGN with street traffic, bus interior, and crowd talking NSN
SNR levels in dB	{0, 5, 10, 15, 20, 25, 30}

References

[1] Huang GB, Zhu QY, Siew CK. Extreme learning machine: theory and applications. Neurocomputing 2006; 70 (1-3): 489-501. doi: 10.1016/j.neucom.2005.12.126

[2] Huang G, Huang GB, Song S, You K. Trends in extreme learning machines: a review. Neural Networks 2015; 61: 32-48. doi: 10.1016/j.neunet.2014.10.001

[3] Ding S, Zhao H, Zhang Y, Xu X, Nie R. Extreme learning machine: algorithm, theory and applications. Artificial Intelligence Review 2015; 44 (1): 103-115. doi: 10.1007/s10462-013-9405-z

[4] Albadra MA, Tiuna S. Extreme learning machine: a review. International Journal of Applied Engineering Research 2017: 12 (14): 4610-4623.

[5] Ding S, Xu X, Nie R. Extreme learning machine and its applications. Neural Computing and Applications 2014; 25 (3-4): 549-556. doi: 10.1007/s00521-013-1522-8

[6] Dhanani J, Mehta R, Rana D, Tidke B. Back-Propagated Neural Network on Map Reduce Frameworks: A Survey. New York, NY, USA: Springer, 2019.

[7] Gopi ES. Digital Speech Processing Using MATLAB. New York, NY, USA: Springer, 2014.

[8] Yap KS, Tiong SK, Nagi J , Koh J, Nagi F. Comparison of supervised learning techniques for non-technical loss detection in power utility. International Review on Computers and Software 2012; 7 (2): 1828-6003.

[9] Verma P, Das PK. i-Vectors in speech processing applications: a survey. International Journal of Speech Technology 2015; 18 (4): 529-546. doi: 10.1007/s10772-015-9295-3

[10] Al-Kaltakchi MT, Woo WL, Dlay SS, Chambers JA. Speaker identification evaluation based on the speech biometric and i-vector model using the TIMIT and NTIMIT databases. In: 5th IEEE International Workshop on Biometrics and Forensics; London, UK; 2017. pp. 1-6.

[11] Al-Kaltakchi MT, Woo WL, Dlay SS, Chambers JA. Comparison of i-vector and GMM-UBM approaches to speaker identification with TIMIT and NIST 2008 databases in challenging environments. In: 25th IEEE European Signal Processing Conference; Kos, Greece; 2017. pp. 533-537.

[12] Al-Kaltakchi MT, Woo WL, Dlay SS, Chambers JA. Multi-dimensional i-vector closed set speaker identification based on an extreme learning machine with and without fusion technologies. In: 2017 Intelligent Systems Conference; London, UK; 2017. pp. 1141-1146.

[13] Huang GB, Cambria E, Toh, Widrow B, Xu Z. New trends of learning in computational intelligence. IEEE Computational Intelligence Magazine 2015; 10 (2): 16-17. doi: 10.1109/MCI.2015.2405277

[14] Cambria E, Liu Q, Li K, Leung VCM, Feng L et al. Extreme learning machines. IEEE Intelligent Systems 2013; 28 (6): 30-59. doi: 10.1109/MIS.2013.140

[15] Lan Y, Hu Z, Soh YC, Huang GB. An extreme learning machine approach for speaker recognition. Neural Computing and Applications 2013; 22 (3-4): 417-425. doi: 10.1007/s00521-012-0946-x

[16] Sadjadi SO, Slaney M, Heck L. MSR Identity Toolbox. Seattle, WA, USA: Microsoft, 2013.

[17] Dehak N, Kenny PJ, Dehak R, Dumouchel P, Ouellet P. Front-end factor analysis for speaker verification. IEEE Transactions on Audio, Speech, and Language Processing 2011; 19 (4): 788-798.