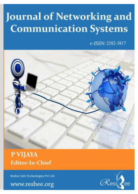
Vol.3 No.4 Oct 2020

Journal of Networking and Communication Systems

Received 8 July, Revised 19 September, Accepted 12 October



Automatic Speaker Diarization using Deep LSTM in Audio Lecturing of e-Khool Platform

Dr. Sivaram Rajeyyagari
Associate Professor, Deanship of Information Technology and E-Learning,
Shaqra University, Shaqra, Saudi Arabia.

Abstract: The speaker diarization is the process of segmentation and the grouping of the input speech signal into a region based on the identity of the speaker. The main challenge in the speaker diarization method is improving the readability of the speech transcription. Hence, in order to overcome the challenge, a speaker diarization method based on deep LSTM is proposed in this research. Initially, the pre-processing is performed for the removal of the noise from the audio lecturing of E-Khoolusers. Then, Linear Predictive Coding (LPC) is used for the extraction of the efficient features from the audio lectures of the E-Khoolusers. After the extraction of the features, the absence or presence of the speaker in the audio lecture is detected using the VAD technique which is followed by the segmentation of the speaker using the extracted features. Finally, the feature vector is determined and the speaker from the audio lecturing of the E-Khoolusers is clustered using the deepLSTM. The proposed speaker diarization method based on deep LSTM is evaluated using the metrics, such as sensitivity, accuracy and specificity. When compared with the existing speaker diarization methods, the proposed speaker diarization method based on deep LSTM obtained a minimum DER of 0.0623, minimum false alarm rate of 0.0369, and minimum distance of 2546 for varying frame length and obtained a minimum DER of 0.0923, minimum false alarm rate of 0.0869, and minimum distance of 1146 for varying Lambda.

Keywords: Speaker Diarization, Audio Signal, E-Khool Audio Lecturing, Linear Predictive Coding, Deep Learning Classifier

1. Introduction

The speaker diarization method has become the challenging task due to the development of the recorded speech that consists of audio broadcasts, voice mails, television and meeting [1]. The conversation of the speaker is found from the mixed signal that consists of speech signal of several persons [3]. The speech signals are segmented and the speech signals of the same speakers are grouped together in speaker diarization process. Hence, the main aim of the speaker diarization method lies in the identification of the speaker [1]. The speaker diarization is applied in the event scenarios, like broadcast news, reports, interviews and debates. Some of the widely used application of the speaker diarization includes the broadcast and telephone meetings, speaker detection, speaker recognition, speaker based multimedia retrieval, video segmentation and summarization [4]. The speaker diarization is performed in the audio sources, like music, background noise and speaker [11]. The component of the speaker diarization includes the speaker change detection (segmentation), VAD, re-segmentation and clustering [2]

The main component of the speaker diarization is clustering in which the segments containing the same audio sources are combined together. The solutions developed for speaker clustering includes the bottom-up approach, global optimization approaches and the top-down approach. The bottom-up approach is the most popular clustering method. The top-down clustering method forms a cluster by considering the whole video as a single model unless the criterion for termination is obtained. In the bottom-up approach, separate clusters were formed during the segmentation stage from each and every individual segment. Then, the nearby clusters are merged until the criterion for termination is obtained. Even though there are differences between the top-down based approach and bottom-up approach, they are iterative process and had error propagation. The distance metrics also have an impact on the speaker clustering as the distance between the segments that belong to the same class needs to be evaluated effectively. Some of the distance metrics includes generalized log-likelihood ratio (GLR) [13], Bayesian information criteria (BIC) [12], i-Vector based distances, probabilistic linear discriminant analysis (PLDA) [14] and Kullback-Leibler (KL) divergence [15] [2].

https://doi.org/10.46253/jnacs.v3i4.a3

The major contribution of the research is the development of the speaker diarization method based on deep LSTM. Initially, the noise in the audio lecturing of E-Khool users is suppressed by pre-processing. Then, the LPC is used for the extraction of the efficient features from the audio lectures of the E-Khoolusers which is followed by the detection of speech using VAD. After the detection of speech, the speaker is segmented by the extracted features. Finally, the speaker from the audio lecturing of the E-Khool users is clustered using the deepLSTM.

The organization of the paper is as follows, section 1 introduces the speaker diarization methods, section 2 reviews the existing speaker diarization methods, section 3 describes the proposed speaker diarization method based on deepLSTM, section 4 discusses the result of the proposed speaker diarization method based on deepLSTM and section 5 concludes the paper.

2. Literature Review

The literature review of the speaker diarization methods are as follows: Ramaiah, V.S. and Rao, R.R.[1] developed a speaker diarization method based on Holoentropy with the eXtended Linear Prediction using autocorrelation Snapshot (HXLPS) along with deep neural network (DNN). This method provided better performance in diarization and had lower Diarization Error Rate (DER) but this method failed to improve the tracking accuracy. Yu, C., & Hansen, J. H. L.[2] designed a speaker diarization method using bottom-up speaker clustering algorithm. The clustering algorithms had two stages such as constrained clustering and explore clustering. The explore clustering improved the speaker clustering process by finding the atleast one sample whereas the constrained clustering lasted until there was one cluster remaining from the explore stage. This method provided lower DER rate along with the inclusion of more queries. However, this method failed to remove the human errors.

Subba Ramaiah, V., & Rajeswara Rao, R.[3] modeled a speaker diarization method based on Lion optimization and Tangent weighted Mel frequency cepstral coefficient (TMFCC). In this method, the Lion optimization was used for clustering the audio stream that was detection in the voice activity into particular speaker groups. Although this method had better tracking accuracy, it had low tracking distance. Park, T. J.et al.[4] developed a speaker diarization method based on the values of normalized maximum eigengap (NME). During the spectral clustering, the threshold parameter for the elements in the affinity matrix was determined by the NME besides estimating the number of clusters. This method evaluated the threshold parameter without tuning the parameter on the development set. This method reduced the error rate but failed to analyze the ratio between the tuning parameter and the NME.

2.1. Challenges

The challenges faced in the speaker diarization methods are as follows,

- The speaker diarization method using the HXLPS and DNN assigned the label for each speaker signal to discriminate the speech signal thus providing lower DER but the main challenge lies in improving the tracking accuracy [1].
- In [2], the challenge of the speaker clustering algorithm was the exclusion of the human errors. This method considered that the perfect answers were provided by the human assistance for any query pair which was the major drawback.
- The speaker diarization method based on Lion optimization and TMFCC provided better tracking accuracy but had lower tracking distance. Thus, the challenge lies in improving the tracking distance of the speech signal [3].
- The challenge of the speaker diarization based on NME was analyzing the ratios of both the tuning parameter to the values of NME and the generalization of the different data on the production system [4].

3. Proposed Speaker Diarization Method based on deepLSTM

In this research, the speaker diarization is performed in the audio lecturing of E-Khool users and clustered based on the speaker identity. The first step of the speaker diarization is the pre-processing in which the noise in the audio lecturing of E-Khool users is suppressed to obtain the audio signal with better quality. The second step is the extraction of the features in which the efficient features from the audio signals are extracted using the LPC which is followed by the detection of the speech. In the speech detection step, the absence or presence of the speaker in the audio signal is determined using the VAD technique. The next step is the segmentation of the speaker in which the audio signal is segmented by the extracted features, which is followed by the determination of the feature vector. Finally, the speaker from the audio lecturing of the E-Khool users is clustered using the deepLSTM [6]. Figure 1 shows the block diagram of the speaker diarization method based on deepLSTM.

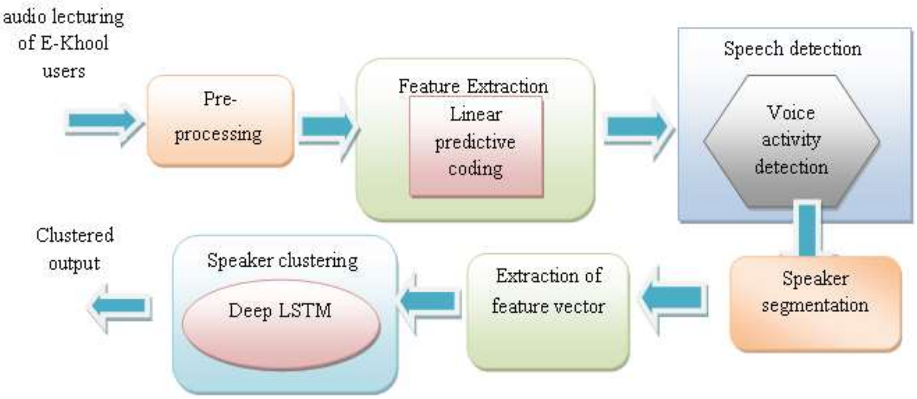


Fig. 1. *Block diagram of the speaker diarization method based on deepLSTM*

3.1. Pre-Processing of the Audio Lecturing of E-Khool Users

The initial process of the speaker diarization method is the pre-processing of the input audio lecturing of E-Khool users. The pre-processing is done for the input audio lecturing signal of E-Khool users for suppressing the noisy regions from the signal. Thus, the audio signal with best quality is obtained through the pre-processing step.

3.2 Extraction of features using Linear predictive coding

The features from the input audio lecturing signal of E-Khool users are extracted using the LPC from which different activities were clearly distinguished by extracting the features. The data in the audio signal consists of time-varying frequency components that are compressed effectively using the LPC. In the LPC method, the spectral envelope having low data dimensionality was represented as the poles of the spectrum are determined by the roots of the coefficients of linear predictor (LP). Thus, the LPC converts the audio speech signal into spectral envelope. At time p, the signal $y(p)$ is represented by the LP model through the linear combination of the number of samples in the past as,

$$\hat{y}(p) = \sum_{u=1}^{g} c_u . y(p-u) \tag{1}$$

where, the prediction of $y(p)$ and the LP coefficients are represented as, c_u and $\hat{y}(p)$. The prediction error $h(p)$ is represented as,

$$h(p) = y(p) - \hat{y}(p) = y(p) - \sum_{u=1}^{g} c_u . y(p-u) \tag{2}$$

While minimizing the generated error, the $y(p)$ signal is modeled using the previous values. The prediction coefficients that reduce the expected value is obtained using the MSE criterion. The MSE criterion is represented as,

$$E\left[h^2(p)\right] = E\left[\left(y(p) - \sum_{u=1}^{g} c_u . y(p-u)\right)\right] \tag{3}$$

$$E\left[y^2(p)\right] - 2\sum_{u=1}^{g} c_u . E\left[y(p).y(p-u)\right] + \sum_{u=1}^{g} c_u \sum_{v=1}^{g} c_u . E\left[y(p-u).y(p-v)\right] \tag{4}$$

$$s_{yy}(0) - 2s_{yy}^{H}.c + c^{H}.B_{yy}^{H}.c \tag{5}$$

where, the autocorrelation vector is represented as, $s_{yy} = E\left[y(p)y\right]$ the coefficient vector of the prediction is given as, $c^H = \left[c_1, c_2, ... c_a\right]$, the input vector's autocorrelation matrix is given as, $B_{yy} = E\left\lfloor yy^H \right\rfloor$. Then, the MSE's gradient in correspondence to the coefficient vector of the prediction is considered. From the equation (5), the MSE value is set as zero and rearranged to represent as below.

$$s_{yy} = c.B_{yy} \tag{6}$$

By re-arranging the above equation for solving the coefficient vector of the prediction we get,

$$c = B_{yy}^{-1}.s_{yy} \tag{7}$$

The autocorrelation matrix and its invertion are responsible for the computational complexity in LPC. The evaluated LP coefficient is given to the classifier as the input vector.

3. 3. Speech Detection using Voice Activity Detection

After the extraction of the features, the speech is detected using the VAD technique. In the VAD method, the existence of the speaker is detected from the audio lecturing signal. According to the identity of the speaker, the audio signal of the speaker is determined. The non-speech regions were also removed from the audio signal. The Bayesian Information Criterion (BIC) were used for the detection of the speech activity.

At first, the features and audio signal is given to the GMM model, which is followed by the detection of the audio lecturing signal using the BIC criteria. In case of large number of speakers, the audio signal is posed with the vocal tracts. In the audio signal, the probabilistic model is catered implicitly using the GMM model. There are two aspects in the GMM model, such as the identification of the speaker by integrating the Gaussian components with the speech signals for providing the configuration of the vocal tract and the Gaussian component mixture for achieving smooth approximation. In the BIC, the identification of the speaker and the likelihood measures are evaluated. For the acoustic signal, the GMM model is expressed as,

$$d(k) = \sum_{m=1}^{e} b_m D(\alpha_m, \beta_m, k) \tag{8}$$

where, mean and covariance of the Gaussian function is denoted as, $D(\alpha_m, \beta_m, k)$, the weight and the Gaussian component is given by the term, b and e . For the detection of the activity, the BIC is used for improving the likelihood. The BIC criterion is generally expressed as,

$$J(C) = \log W(G/C) - \mu \frac{1}{2} A \log(T_i) \tag{9}$$

where, the G's frame size is given as, T_i , the measure of the log likelihood and the penalty weight is represented as, $\log W(G/C)$ and μ . The desired and the perplexity model is represented as, c and A . Two hypothesis are necessary for determining the distance between the segments of the audio that depends on BIC in order to detect the signal speaker. The hypothesis of the BIC is given as,

$$J(C) = \log W(G/C) - \mu \frac{1}{2} A \log(T) \tag{10}$$

$$J(C_m, C_n) = \log W(G_m / C_m) + \log W(G_m / C_m) - \mu \frac{1}{2}(2A) \log(T) \tag{11}$$

The input signal is used for the detection of the activity of the speaker by predetermining the threshold value. In addition, the BIC is evaluated for determining the score value of the audio lecturing signal. The voice activity is performed for the threshold value is lower than the BIC score. If the threshold value is higher than the BIC score then, there is no speaker activity. The score of the BIC is computed by the below expression,

$$\Delta J = J(C_m, C_n) - J(C) \tag{12}$$

The BIC score value is evaluated for the detection of the speech activity. Hence, the feature extracted from the audio lecturing signal is used for the evaluation of the voice, which is given as,

$$M = \{M_1, M_2, ...M_a\} \tag{13}$$

3.4. Extraction of the Feature Vector using i-Vector Representation

The feature vector is determined for the audio lecturing signals as some features failed to segment the audio lecturing signals effectively. In this research, i-vector representation is obtained through the Universal Background Model (UBM) for extracting the vector. In the UBM, the first and zeroth order statistics are used for obtaining i-vector and T-matrix. The i-vector representation is provided by the UBM model by integrating with the GMM model. For the segmented audio signal, the feature information obtained in i-vector representation is expressed as,

$$L = \{L_1, L_2, ...L_j\} \tag{14}$$

3.5. Speaker Clustering using Deep LSTM

After segmenting the speaker from the audio lecturing signal of the E-Khool users, the segmented signal for the same speaker is clustered using deepLSTM [6]. The input to the deepLSTM is the output from the i-vector representation. The advantage of the deepLSTM is the presence of larger transitional kernel for capturing the fast motions, which helps in the effective clustering of speakers. When the large transitional kernel is employed for the clustering, effective feature patterns are obtained in the deepLSTM. The deepLSTM is organized as the 3D tensors having inputs $L_1,.....,L_t$, hidden states $X_1,.....,X_t$, outputs, $R_1,.....,R_t$ gates U_t, E_t, and Z_t. The Hadamard product 'o' and Convolutional operator '*' are used for predicting the future using the inputs and the previous states that corresponds to the neighbors. The LSTM comprises of memory units that constitutes cell and gates. The flow of information is controlled using the memory cell and gates. In the deepLSTM, when the memory cell is updated with the new input, the unnecessary contents from the past are removed and the output is provided for the new input. Initially, the input gate's output is represented as,

$$U_t = \rho\left(\beta_U^L * P_t + \beta_U^R * R_{t-1} + \beta_U^X o X_{t-1} + \gamma^U\right) \tag{15}$$

where, the activation function of the gate is given as, ρ, the input vector is given as L_t weight between the input gate and the layer is given as, β_U^L, the weight between the output of the cell and the input layer is represented as, β_U^X, the weight between the output of the memory and the input layer is represented as, β_U^R. The output from the past output of the cell and the input layer is given as, X_{t-1} and R_{t-1}. The input layer's bias is given as, γ^U. The element-wise multiplication and the convolutional operator is given as, o and *. The forget gate's ouput is represented as,

$$E_t = \rho\left(\beta_E^L * P_t + \beta_E^R * R_{t-1} + \beta_E^X o X_{t-1} + \gamma^E\right) \tag{16}$$

where, the weight between the forget gate and the input layer is given as, β_E^L. The weight between the cell and the output gate is represented as, β_E^X, the forget gate's bias is represented as, γ^E, the weight between the previous layer's memory unit and the output gate is denoted as, β_E^R. The output gate's is computed as,

$$Z_t = \rho\left(\beta_Z^L * P_t + \beta_Z^R * R_{t-1} + \beta_Z^X o X_{t-1} + \gamma^Z\right) \tag{17}$$

where, the weight that connects both the input layer and the output gate is represented as, β_Z^L. The weight that linked the memory unit with the output gate is represented as, β_Z^R, the weight that connects both the cell and the output gate is denoted as, β_Z^X. The activation function of the weight that corresponds to the cell is used for determining the temporary Cell state's output, which is represented as,

$$\widetilde{X}_t = \tanh\left(\beta_Y^L * L_t + \beta_Y^R * R_{t-1} + \gamma^Y\right) \tag{18}$$

where, the weight that connects both the input layer and the cell, the weight that linked the memory unit and cell is given as, β_Y^L and β_Y^R. The cell's output is represented as,

$$X_t = X_t o X_{t-1} + U_t o \widetilde{X}_t \tag{19}$$

$$X_t = X_t o X_{t-1} + U_t o \tanh\left(\beta_Y^L * L_t + \beta_Y^R * R_{t-1} + \gamma^Y\right) \tag{20}$$

The memory unit's output is computed as,

$$R_t = Z_t o \tanh\left(X_t\right) \tag{21}$$

where, the output gate and the memory block's output is given as, Z_t and R_t. The output from the output layer is represented as,

$$K_t = \eta\left(\beta_K^R . R_t + \gamma^K\right) \tag{22}$$

where, the weight that connects the memory unit and the output vector K_t is given as, β_K^R, the output layer's bias is represented as, γ^K. Thus, the output from the LSTM output layer is used for the clustering of the speaker.

4. Results and Discussion

The result of the proposed speaker diarization method based on deep LSTM is discussed along with the comparative analysis with the existing speaker diarization methods in this section.

4.1. Experimental Setup

The proposed speaker diarization method using the deep LSTM classifier uses the audio lecturing signals available from the E-Khool platform [16].

4.2 Performance Metrics

The performance metrics used by the proposed speaker diarization method based on deep LSTM for the analysis are DER, false alarm rate and tracking distance

Tracking distance:

The tracking distance is the measure of distance between the output and the input signal from the speaker. The tracking distance is calculated as,

$$\text{Trackingdis}\tan\text{ce} = \sqrt{\left(W_m^F - W_m^g\right)^2} \tag{23}$$

where, the original signal and the output signal from the speaker is given as, W_m^F and W_m^g.

Diarization error rate (DER):

The DER established a map between the error evaluation in the audio signal and the speaker tags. The DER is evaluated by the below term as,

$$\text{DER} = \frac{E_C + E_M + A_F}{T_R} \tag{24}$$

where, the confusion error, miss error and the false alarm are represented as, E_C, E_M and A_F. The total reference speech time is given as, T_R.

False alarm rate (FAR):

The FAR is the measure of the incorrect evaluation of the non-speech segment. The false alarm rate is calculated as,

$$\text{FAR} = 1 - \lambda \tag{25}$$

where, λ is defined as specificity and the specificity is denoted with respect to the false positive and true negative.

4.3. Comparative Methods

The comparative methods used for the analysis of the proposed speaker diarization method based on deep LSTM are TMFCC with ILP [7], multi-kernel based MFCC (MKMFCC) with Wu-Li-Fuzzy clustering (WLI fuzzy clustering) [9], TMFCC with Lion [8], XLPS with DNN [10], HXLPS with DNN [1].

4.4. Comparative Analysis

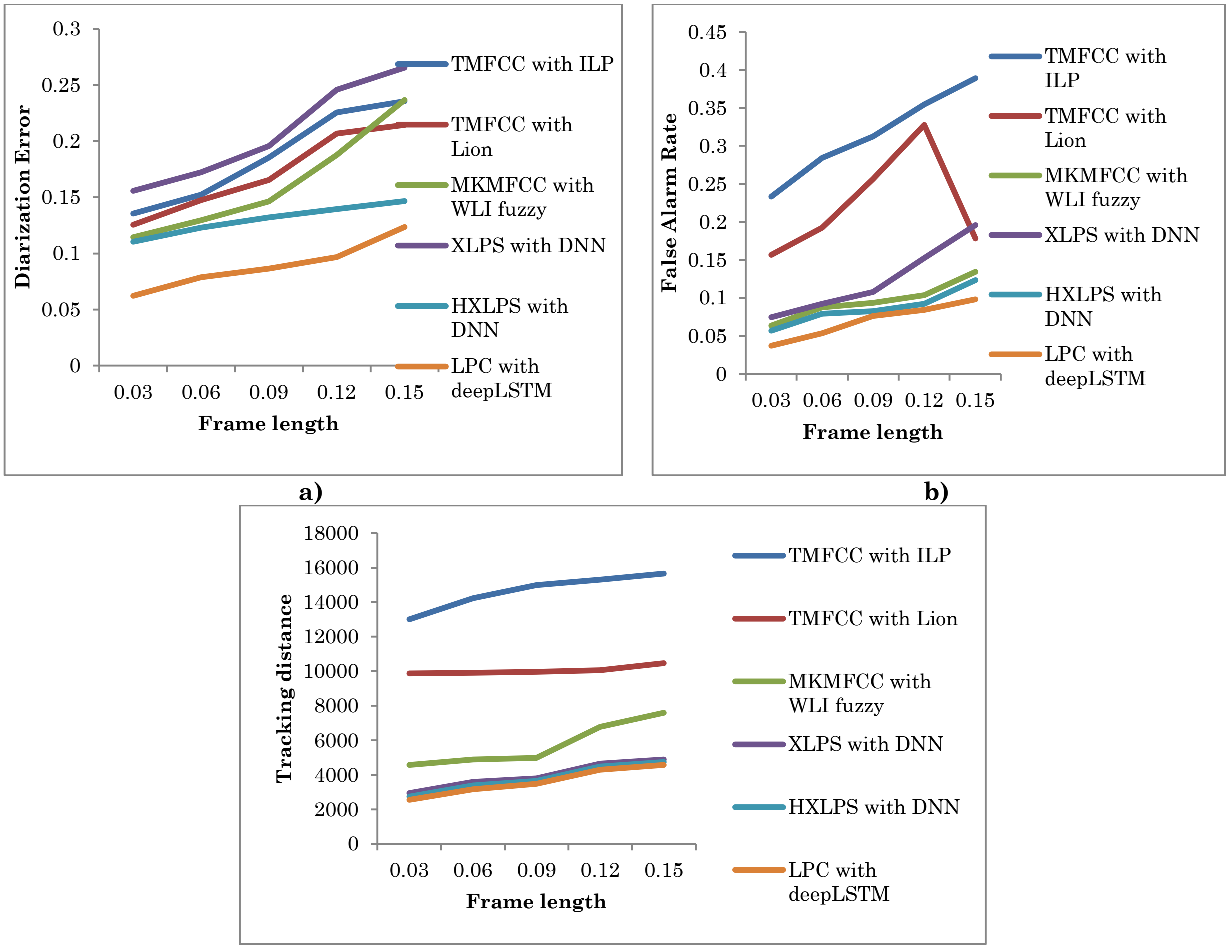
The comparative analysis of the proposed speaker diarization method based on deep LSTM is performed based on the lambda value and frame length for the metrics, like DER, tracking distance and FAR by varying the frame length.

4.4.1. Comparative Analysis based on Frame Length

Figure 2 shows the analysis of the speaker diarization methods based on Frame length for varying tracking distance for dataset-2. Figure 2 a) shows the analysis of the speaker diarization method using DER. The DER obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM for the frame length of 0.03 is 0.1356, 0.1145, 0.1255, 0.1556, 0.1105, and 0.0623 respectively. For the frame length of 0.15, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the DER of 0.2355, 0.2367, 0.2145, 0.2654, 0.1465, and 0.1234 respectively.

Figure 2 b) shows the comparative analysis of the speaker diarization method using false alarm rate. For the frame length of 0.03, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the false alarm rate of 0.2334, 0.0635, 0.1565, 0.0745, 0.0567 and 0.0369, respectively. The false alarm rate obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM for the frame length of 0.15 is 0.3894, 0.1342, 0.1782, 0.1959, 0.1236 and 0.0981, respectively.

Figure 2 c) shows the comparative analysis of the speaker diarization method using tracking distance. The tracking distance obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM for the frame length of 0.03 is 13000, 4567, 9874, 2939, 2741 and 2546, respectively. For the frame length of 0.15, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the tracking distance of 15659, 7589, 10456, 4876, 4734 and 4572, respectively.



a) b)

c)

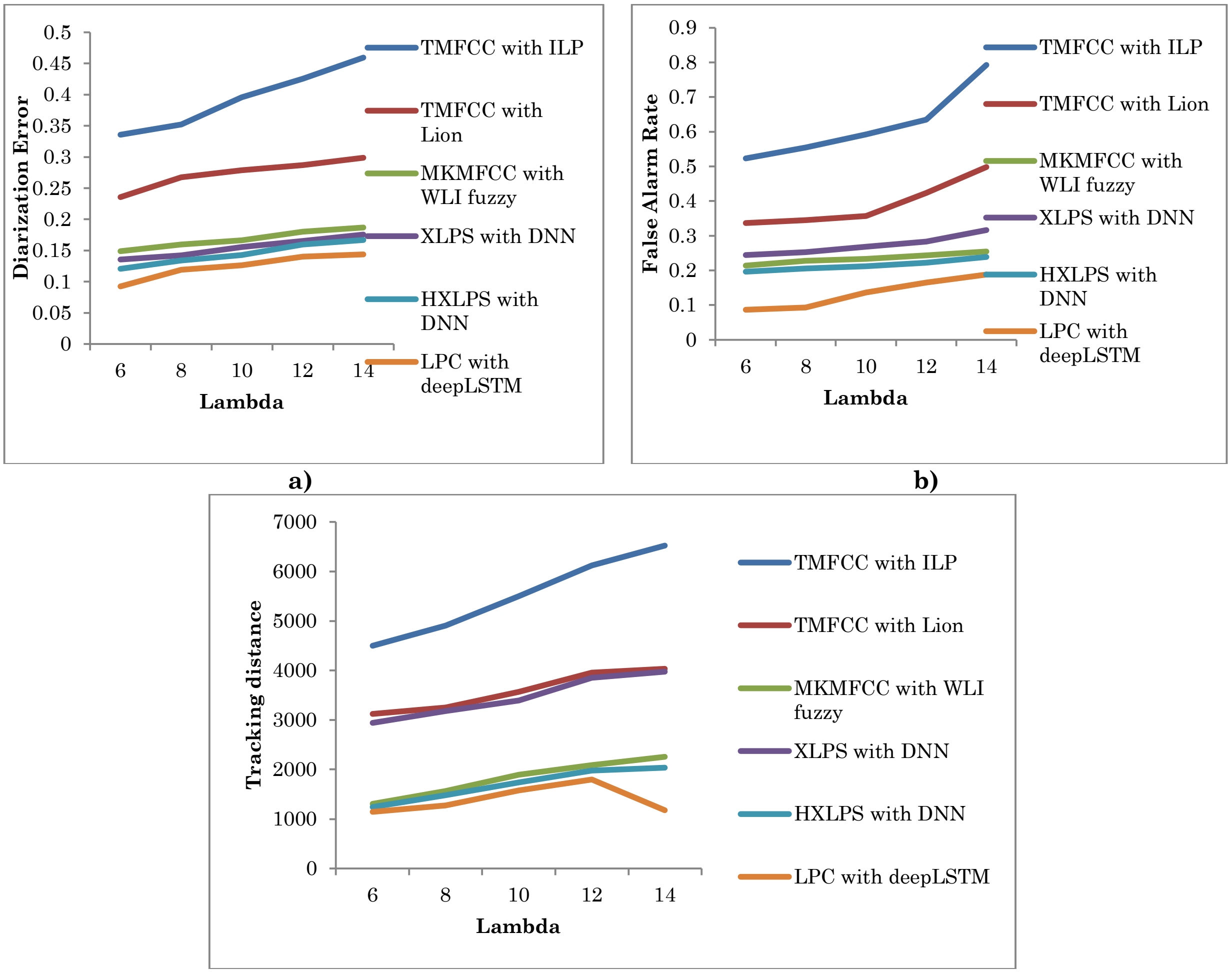
***Figure2.** Comparative analysis based on Frame length for a) DER, b) false alarm rate and c) tracking distance*

4.4.2Comparative Analysis based on Lambda

Figure 3 shows the analysis of the speaker diarization method based on Lambda for varying tracking distance. Figure 3 a) shows the analysis of the speaker diarization method using DER. The DER obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, XLPS with DNN, TMFCC with Lion, HXLPS with DNN and the proposed LPC with deep LSTM for the Lambda of 6 is 0.3356, 0.1485, 0.2355, 0.1356, 0.1205 and 0.0923, respectively. For the Lambda value=14, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the DER of 0.4595, 0.1867, 0.2985, 0.1754, 0.1665 and 0.1434, respectively.

Figure 3 b) shows the analysis of the speaker diarization method using false alarm rate. For the Lambda value=6, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the false alarm rate of 0.5234, 0.2135, 0.3365, 0.2445, 0.1967 and 0.0869, respectively. The false alarm rate obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed for the Lambda value of 14 is 0.7923, 0.2542, 0.4982, 0.3159, 0.2386 and 0.1881, respectively.

Figure 3 c) shows the analysis of the speaker diarization method using tracking distance. The tracking distance obtained by the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM for the Lambda of 6 is 4500, 1308, 3125, 2939, 1241 and 1146, respectively. For the Lambda value=14, the TMFCC with ILP, MKMFCC with WLI fuzzy, TMFCC with Lion, XLPS with DNN, HXLPS with DNN and the proposed LPC with deep LSTM obtained the tracking distance of 6523, 2258, 4034, 3976, 2034 and 1172, respectively.



a)

b)

c)

Figure3. *Comparative analysis based on Lambda a) DER, b) false alarm rate and c) tracking distance*

5. Conclusion

In this research, a speaker diarization method is developed based on deepLSTM for clustering the audio lecturing of E-Khool users based on the identity of the speaker. At first, the audio lecturing of E-Khool users are pre-processed for suppressing the noise in the signal. Then, the efficient features from the audio lectures are extracted using the LPC which is followed by the speaker detection. The speaker is detected using the VAD technique and the detected speaker is segmented with the help of the extracted features. After the segmentation, the feature vector is determined using i -vector representation model. Finally, the speaker from the audio lecturing of the E-Khool users is clustered using the deepLSTM. The proposed speaker diarization method based on deep LSTM is evaluated using the metrics, such as sensitivity, accuracy and specificity. The proposed speaker diarization method based on deep LSTM obtained a minimum DER of 0.0623, minimum false alarm rate of 0.0369, and minimum distance of 2546 for varying frame length and obtained a minimum DER of 0.0923, minimum false alarm rate of 0.0869, and minimum distance of 1146 for varying Lambda. The proposed speaker diarization method can be further enhanced using advanced methods for speaker segmentation and clustering.

Compliance with Ethical Standards

Conflicts of interest: Authors declared that they have no conflict of interest.

Human participants: The conducted research follows the ethical standards and the authors ensured that they have not conducted any studies with human participants or animals.

References

[1] Ramaiah, V.S. and Rao, R.R., "Speaker diarization system using HXLPS and deep neural network," Alexandria Engineering Journal, vol.57, no.1, pp.255-266, 2018.

[2] Yu, C., & Hansen, J. H. L., "Active Learning Based Constrained Clustering For Speaker Diarization," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25(11), 2188–2198, 2017.

[3] Subba Ramaiah, V., & Rajeswara Rao, R. , "A novel approach for speaker diarization system using TMFCC parameterization and Lion optimization," Journal of Central South University, vol.24, no.11, pp.2649–2663, 2017.

[4] Park, T. J., Han, K. J., Kumar, M., & Narayanan, S., "Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap," IEEE Signal Processing Letters, pp.1–1, 2019.

[5] Javier, R. J., & Youngwook Kim., "Application of Linear Predictive Coding for Human Activity Classification Based on Micro-Doppler Signatures," IEEE Geoscience and Remote Sensing Letters, vol.11, no.10, pp.1831–1834, 2014.

[6] The ELSDSR dataset for speaker diarization system, <http://cogsys.compute.dtu.dk/soundshare/elsdsr.zip>.

[7] R. Kumara Swamy, K. Sri Rama Murty, B. Yegnanarayana, "Determining number of speakers from multi-speaker speech signals using excitation source information," IEEE Signal Process. Lett. Vol.14, no.7, 2007.

[8] B. Rajakumar, "The Lion0s Algorithm: a new nature-inspired search algorithm," Procedia, vol.6 pp.126–135, 2012.

[9] Chih-Hung Wu, Chen-Sen Ouyang, Li-Wen Chen, Li-Wei Lu, "A new fuzzy clustering validity index with a median factor for centroid-based clustering, IEEE Trans. Fuzzy Syst. Vol.23, (no.3), pp.1–16, 2013.

[10] Jouni Pohjalainen, Rahim Saeidi, Tomi Kinnunen, Paavo Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," INTERSPEECH, 2010.

[11] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke et al., "The icsi meeting project: Resources and research," in Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop, 2004.

[12] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in Proc. DARPA Broadcast News Transcription and Understanding Workshop, vol. 8. Virginia, USA, pp. 127–132, 1998.

[13] A. Solomonoff, A. Mielke, M. Schmidt, and . Gish, "Clustering speakers by their voices," in Acoustics, Speech and Signal Processing, Proceedings of the 1998 IEEE International Conference on, vol. 2., pp. 757–760, 1998.

[14] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in Computer Vision, 2007. ICCV 2007, pp. 1–8, 2007.

[15] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in Proc. DARPA speech recognition workshop, vol. 1997, 1997.

[16] Ekhool-Top learning management system from "https://ekhool.com/".