HOSTED BY

ELSEVIER

Alexandria University

Alexandria Engineering Journal

www.elsevier.com/locate/aej
www.sciencedirect.com

Speaker diarization system using HXLPS and deep neural network

V. Subba Ramaiah [a,*], R. Rajeswara Rao [b]

[a] *Mahatma Gandhi Institute of Technology, Kokapet, Hyderabad, Telangana 500075, India*
[b] *JNTUK-UCEV, Kakinada, Andhra Pradesh 535002, India*

Received 1 September 2016; revised 17 November 2016; accepted 15 December 2016

KEYWORDS

Speaker diarization;
HXLPS feature extraction;
Voice activity detection;
Deep neural network;
Speaker clustering;
Diarization Error Rate (DER)

Abstract In general, speaker diarization is defined as the process of segmenting the input speech signal and grouped the homogenous regions with regard to the speaker identity. The main idea behind this system is that it is able to discriminate the speaker signal by assigning the label of the each speaker signal. Due to rapid growth of broadcasting and meeting, the speaker diarization is burdensome to enhance the readability of the speech transcription. In order to solve this issue, Holoentropy with the eXtended Linear Prediction using autocorrelation Snapshot (HXLPS) and deep neural network (DNN) is proposed for the speaker diarization system. The HXLPS extraction method is newly developed by incorporating the Holoentropy with the XLPS. Once we attain the features, the speech and non-speech signals are detected by the Voice Activity Detection (VAD) method. Then, i-vector representation of every segmented signal is obtained using Universal Background Model (UBM) model. Consequently, DNN is utilized to assign the label for the speaker signal which is then clustered according to the speaker label. The performance is analysed using the evaluation metrics, such as tracking distance, false alarm rate and diarization error rate. The outcome of the proposed method ensures the better diarization performance by achieving the lower DER of 1.36% based on lambda value and DER of 2.23% depends on the frame length.

1. Introduction

With a rapid growth of recorded speech, which constitutes voice mails, audio broadcasts, meeting and television, speaker diarization technique becomes the facilitating and challenging task. Speaker diarization is defined to segment the speech signal and then grouped for the same speaker. Thus, the core intent of speaker diarization system is to identify the speaker by their audio signals [1]. In other words, it is used to estimate the audio signal as "who speak what and when". Some event scenarios, such as reports, broadcast news, debates, and interviews are the useful applications of the speaker diarization. Then, the diarization is widely used in several applications such as speaker detection, telephone and broadcast meetings and also then auxiliary video segmentation, speaker recognition, multimedia summarization and speaker based retrieval of multimedia [4]. This type of audio sources includes music, speaker, and background noises, where the signal of the same

[*] Corresponding author.
E-mail addresses: subbubdl@gmail.com (V. Subba Ramaiah), raob4u@yahoo.com (R. Rajeswara Rao).

Peer review under responsibility of Faculty of Engineering, Alexandria University.

http://dx.doi.org/10.1016/j.aej.2016.12.009
1110-0168 © 2016 Faculty of Engineering, Alexandria University. Production and hosting by Elsevier B.V.

speaker is detected or classified [2,3]. In general, the speaker segmentation and speaker clustering are considered as the two main components of speaker diarization system [5,7,8,11–13].

The speaker clustering followed by the speaker segmentation process is termed as the speaker diarization. The speaker clustering is defined as the grouping or clustering the segmented signals of the same speaker. Hence, the speaker clustering has been the great significant part since it provides the final diarization performance [14]. Some of the clustering mechanism, such as bottom-up approach, top-down approach, neural network, K-means, and self organizing maps are developed. In bottom-up approach, the Agglomerative Hierarchical Clustering (AHC) [15] is the most popular method for speaker clustering, where the number of clusters is generated based on speaker identity simultaneously. On the other hand, Hidden Markov Model (HMM) becomes the prominent method for the top-down approach [3]. Thus, the speaker detection can be performed by such algorithms, such as step by step approach, integrated approach and mixed approach.

In this paper, HXLPS and deep neural network are proposed for speaker diarization. The audio signal includes multi-speaker (i.e., five speakers and seven speakers) is considered as the input signal for the proposed methodology. The two main contributions of this paper are as follows:

- The new HXLPS feature extraction method is developed by incorporating the Holo-entropy with eXtended Linear Prediction using autocorrelation Snapshot (HXLPS). Thus, the acoustic features are used for the speaker segmentation.
- Once we attain the segmented speakers using i-vector representation, the DNN is utilized to cluster the audio signals of the respective speaker.

This paper is structured as follows: Section 2 discusses about the speaker diarization system from eight research papers. The problem statement and challenges behind the speaker diarization are presented in Section 3. Section 4 is briefly explained about the speaker diarization using HXLPS and DNN. Consequently, Section 5 provides the experimental results and performance analysis. Finally, Section 6 concludes this paper.

2. Literature review

This section discusses about the speaker diarization based on speaker segmentation and speaker clustering from eight research papers. Jothilakshmi et al. [1] explained the approach for the speaker diarization using AutoAssociative Neural Network (AANN). The AANN model was employed to segment the speaker's audio signal and then, grouped for the same speaker. Here, the features of the signal were extracted by the Mel Frequency Cepstral Coefficients (MFCC). Finally, the experimental results of the speaker diarization were evaluated and attained the better performance, when compared to the existing speaker diarization method. Bigot et al. [2] demonstrated the speaker diarization using feature extraction and classification method. It was mainly used to detect the speaker roles, such as Anchor, Journalist and so on. Then, the speech features were extracted from the segmented audio signals based on the temporal, prosodic and basic signal. Thus, the 36-vector representation of feature was obtained. It was then fed into the classifier model, such as GMM, KNN and SVM.

In [3], Shum et al. developed the probabilistic approach for speaker clustering using Bayesian Gaussian Mixture Model (BGMM) to principal component analysis for i-vector extraction. Based on the various temporal resolutions, the segmentation and clustering improved the speaker cluster assignments and segmentation boundaries. Thus, the probabilistic based model achieved the better performance in the state-of-the-art benchmark dataset. Xu et al. [4] improved i-vector representation using DNN for speaker diarization. The UBM was utilized instead of Gaussian Mixture Model (GMM) which was used to calculate the posterior information. Thus, the zeroth-order and first-order statistic was estimated by the DNN and MFCC, where i-vector was obtained. Thus, the speaker diarization system attained the better speaker recognition performance.

In [6], Zelenak et al. presented three spatial cross correlation-based features together with spectral information for speaker overlap detection. Then, the overlap segments were removed which lead to assign two labels for each speaker with the aid of Viterbi decoding. Using beamforming and TDOA features, higher performance was achieved. Evans et al. [9] presented the top-down and bottom-up approach for the speaker diarization system. The bottom-up approach was utilized to detect the purer model which was more sensitive to nuisance variation. On the other hand, the top-down approach provided the better normalization performance against variation. Thus, the experimental results were validated and yielded the lower error rate of 21% and 22% for the bottom-up and top-down approach.

Pertila [10] deliberated online method for speaker detection, speech separation and speaker direction tracking. The speech signal was segmented by the multiple acoustic source tracking, assisted by the Bayesian filtering and time-frequency masking. The reverberation measurement of various amounts using two different designs was evaluated to separate the four active speakers. Thus, the results were analysed by the ideal binary masking and oracle tracking used to determine the effect of number of microphones and their spacing. Madikeri [16] developed the PPCA's EM algorithm for the speaker diarization. Initially, the covariance matrix was computed based on the PPCA framework. Then, the optimization exploited framework to prevent the inversion of precision matrix. With the baseline i-vector extraction procedure showed that the speed was improved in terms of the Equal Error Rate (EER). Finally, the speaker recognition performance was studied on the telephone conditions of the benchmark NIST SRE 2010 dataset.

3. Motivation behind the approach

3.1. Problem description

The main problem in speaker diarization system is that to detect the same speaker signal from the audio or speech signal. Consider the input signal as audio signal includes u number of speakers. The input signal for the proposed speaker diarization is given as follows:

$$X = \{ x_i; \quad 1 \leqslant i \leqslant u \}$$

where X defines the input signal and i represents the number of speaker. Here, the challenge is to group the input signal into different speaker's signal using feature extraction and speaker clustering.

3.2. Challenges

- Due to the annotation problem of unlabelled audio file, speaker segmentation where the speech activity is detected and speaker clustering of the same speaker signal become the challenging tasks for the speaker diarization system [3].
- The extraction of acoustic feature is another challenge [7] since the feature exhibits the significant representation of the audio signal. Then, the feature is used to easily determine the speaker locations.
- The noisy factors affect the diarization performance heavily so, developing speaker clustering methods which are adaptable to noisy environment is the challenging one [14].
- Commonly, speaker diarization is performed using LP residual which is not considered the time delay between the users' signal. This time delay will heavily affect the tracking performance because it assumes the time delay as constant. On the other hand, in conference or meeting, the speaker will not be in same position. They may move in different directions and positions. So, the handling of movement to recognize and determine the speakers should be considered in the future works.

4. Proposed methodology: speaker segmentation using HXLPS feature extraction and deep neural network for speaker clustering

The ultimate aim of this paper was to discriminate the speech signal with respect to its corresponding speaker from the input signal. In general, the speech signal or the audio signal poses with multiple numbers of speakers. Fig. 1 depicts the block diagram representation of the proposed methodology. The proposed methodology constitutes speaker segmentation and speaker clustering. Initially, the input speech signal is undergone for the preprocessing step, where the noisy regions are suppressed to acquire the best quality of audio signal. This signal is used for the further step to improve the efficiency of the feature extraction and clustering stage. Subsequently, the speech signal is fed as input to the feature extraction process and thereby, the audio feature space is achieved through HXLPS. After the features are extracted, it is then used to segment the audio signal assisted by the VAD, where the speaker signals are segmented. Then, the feature information is represented by the i-vector of the every segmented signal. Finally, DNN is utilized in this paper for speaker clustering. The i-vector representation of the signal is trained in the DNN and then, we identifies the speaker class label.

4.1. Feature extraction using proposed HXLPS

The feature information of the speech signal is extracted by the proposed HXLPS. Normally, the feature space exhibits the significant representation of the audio signal. According to the feature information, the speaker activity detection is performed on the segmented speech signals. The proposed feature extraction is developed by integrating the holoentropy function into the XLPS method. Firstly, the XLPS [18] feature extraction is explained as follows: The extended linear prediction (XLP) comprises of both Linear Prediction (LP) and Weighted Linear Prediction (WLP) mitigates the error energy using the partial weight values. The advantage of LP is that it contains less bandwidth consumption while increasing the number of speakers and also, the size of the transmitting signal gets reduced. On the other hand, the WLP has the benefits of robustness measure for irrelevant noises. By combining these advantages, the XLP model is developed. Thus, the prediction error using XLP is derived as follows:

$$E^{XLP} = \sum_{j} \left(x_j Y_{j,0} - \sum_{k=1}^{p} \alpha_k^{XLP} x_{j-k} Y_{j,k} \right)^2$$

where $Y_{j,0}$ and $Y_{j,k}$ are defined as the partial weights for XLP. Then, the normal equation for XLP is defined by,

$$\sum_{k=1}^{p} \alpha_k^{XLP} \sum_{j} Y_{j,k} x_{j-k} Y_{j,l} x_{j-l} = \sum_{j} Y_{j,0} x_j Y_{j,l} x_{j-l;} \quad 1 \leqslant l \leqslant p$$

When the partial weight is $Y_{j,k} = \sqrt{W_j}$, then the WLP prediction model is obtained whereas the LP is achieved through $Y_{j,k} = d$. At each time instant, the weight is used separately for

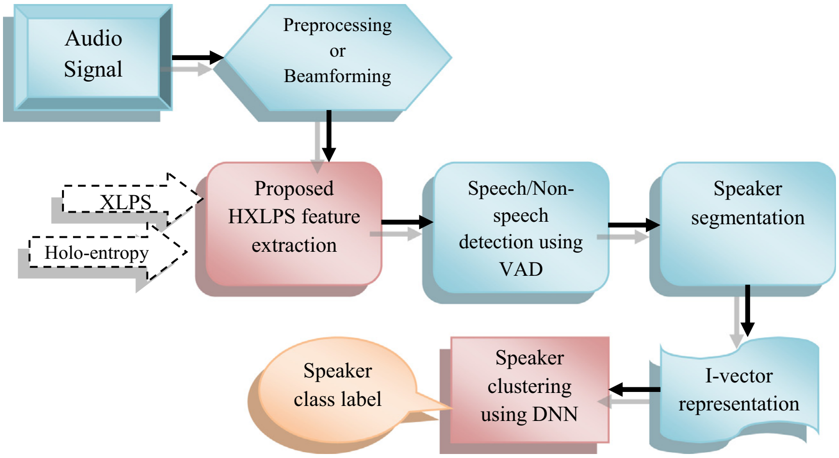

Audio Signal
Preprocessing or Beamforming
XLPS
Holo-entropy
Proposed HXLPS feature extraction
Speech/Non-speech detection using VAD
Speaker segmentation
I-vector representation
Speaker clustering using DNN
Speaker class label

Fig. 1 Block diagram of proposed methodology.

each lagged sample speech signal. This model is termed as the XLP. Thus, the formulation of the normal equation of XLP is given below that decreases the error energy.

$$\sum_{k=1}^{p}\alpha_k\sum_{j}M_{j,k,l}x_{j-k}x_{j-l}=\sum_{j}M_{j,l,0}x_jx_{j-l;}\quad 1\leqslant l\leqslant p$$

where $M_{j,k,l}$ and $M_{j,l,0}$ are the weights in XLP model which is defined by the multiplication of the weight of the lagged sample speech signal.

Then, the weighting function is represented in the form of matrix using autocorrelation snapshot vectors and the term α. It is represented by,

$$\left(\sum_{j}M_j\otimes\left(x_jx_j^T\right)\right)\alpha=\sum_{j}M_{j,l,0}x_jx_j$$

In the above equation, the element wise multiplication is done between the weights and autocorrelation snapshot vectors along with term $\alpha=(\alpha_1,\alpha_2,\ldots,\alpha_p)^T$. Due to autocorrelation snapshot vectors, the error energy gets dwindled well which leads to provide the most significant part of the audio signal for the speaker segmentation. The snapshots of each lagged speech signal are distinctly weighted with respect to time instants.

Due to some low bit rates of the acoustic signal, the performance of XLP [18] gets degraded. To resolve this issue, we develop the novel feature extraction method called, HXLPS. Here, weighting scheme plays a prominent role in the proposed HXLPS feature extraction. The new feature extraction method is designed by integrating the holoentropy into the XLPS method. Basically, the holoentropy measures the audio signal globally and also reduces the computational complexity. Then, the feature information of the signal is modelled by the holoentropy based XLPS. The holoentropy function is used to estimate the new signal features for the speech activity detection. Thus, the mathematical formulation for feature extraction using proposed HXLPS is deliberated below.

$$F=HE*Y_{j,l}$$

where HE is the holoentropy function and $Y_{j,l}$ represents the Absolute Value Sum (AVS) weighting function, which is mainly used to reduce the error energy while extracting the feature information. The AVS function is defined as follows:

$$Y_{j,l}=\frac{p-1}{p}Y_{j-1,l}+\frac{1}{p}\left(|x_j|+|x_{j-l}|\right)$$

The holoentropy function is then estimated by the product of entropy measure and weight function. Due to the novel weighting based holoentropy XLPS extraction, it attains the low diarization error rate which ensures the better diarization performance. Every feature vector of the speech signal is altered by the entropy and weight function.

$$HE(r)=En(r)\times W(r)$$

Initially, the holoentropy of the r^{th} feature vector is determined by the product of entropy value and weights. Then, the entropy measure and weight are expressed by,

$$En(r)=-\sum_{r=1}^{n}P(r)\times\log_2 P(r)\quad\text{and}$$

$$W(r)=2\times\left(1-\left[\frac{1}{1+\exp(-1*En(r))}\right]\right)$$

where $P(r)$ defines the probability measure of r^{th} feature vector. Similarly, the holoentropy is evaluated for the n number of feature vectors. Thus, every feature vector of XLP method is estimated by the holoentropy measure which leads to provide the n number of feature vectors of the acoustic signal. Depending on the features, the speaker signal is segmented from the input audio signal. The voice activity detection is employed to perform the segmentation of the signal.

4.2. Voice activity detection (VAD)

The VAD is a detection technique, where the presence or absence of speech signal of the speaker is determined. It is majorly used for speaker diarization or recognition system. The speaker of audio signal is determined according to speaker identity, which is further used for the speaker segmentation. It is also used to remove the nonspeech regions from the audio signal. To perform the speech activity detection, the Bayesian Information Criterion (BIC) [19] is utilized.

Before the BIC criteria for the detection, the audio signal and HXLPS features are fed into the GMM model. Normally, the audio signal poses with the vocal tracts for large number of speakers. Thus, the GMM model is used to cater the probabilistic model of the audio signal implicitly. The GMM [20] contains two aspects: (i) the Gaussian components are integrated with speech signals which provide the vocal tract configuration to identify the speaker and ii) the smooth approximation is achieved through the Gaussian mixture of components. Then, the likelihood measures and speaker identification are deliberated using BIC. Thus, the GMM model for the acoustic signal is represented by,

$$q(a)=\sum_{i=1}^{g}w_iG(\mu_i,\sigma_i,a)$$

where $G(\mu_i,\sigma_i,a)$ represents the Gaussian function with mean and covariance, g is the Gaussian component and w represents the weight.

(ii) The BIC [19] is mainly used to improve the likelihood measure for the activity detection. It has the main advantage of decreasing the number of false alarms. The general form of BIC criterion is given below.

$$B(D)=\log L(X|D)-\xi\frac{1}{2}P\log(F_x)$$

where $\log L(X|D)$ is the log likelihood measure, F_x denotes the frame size of X, ξ defines the penalty weight and P is the perplexity model and D represents the desired model.

(iii) To detect the speaker from the signal, two hypothesis are required to evaluate the BIC based distance between audio segments. Thus, h_0 and h_1 are considered as the two hypothesis which is modelled by the signal. Thus, the BIC for two hypotheses is followed by,

$$B(D)=\log L(X|D)-\xi\frac{1}{2}P\log(F)\text{and}$$

$$B(D_i,D_j)=\log L(X_i|D_i)+\log L(X_j|D_j)-\xi\frac{1}{2}(2P)\log(F)$$

The threshold value is predetermined to be used for the speaker activity from the input acoustic signal. Also, the score value of the signal is calculated from the BIC. If the BIC score value is greater than the threshold value, the voice activity detection is performed. Or else, there is no activity by the speaker. Thus, the BIC [19] score computation is evaluated by,

$$\Delta B = B(D_i, D_j) - B(D)$$

The speech activity detection is done through the BIC score value. Thus, the voice activity is detected by the feature extracted speech signal which is represented as follows:

$$S = \{S_1, S_2, \ldots, S_d\}$$

4.3. Feature vector extraction using i-vector representation

Using GMM model, the features are not good enough to segment the speakers from the audio signal. In order to solve this difficulty, the Universal Background Model (UBM) [21] is employed. The i-vector representation [4] of the acoustic signal plays a vital role for the speaker diarization system. Using UBM model, the T-matrix and i-vector are obtained by the zeroth and first order statistics. Thus, the UBM model is merged with the GMM for providing the i-vector representation of the audio signal. The zeroth and first order Baum-Welch statistics [24] is used to obtain i-vector representation for each Gaussian mixture g. It is determined as follows:

$$B_g = \sum_t \lambda_t(g)$$

$$C_g = \sum_t \lambda_t(g)(X_t - m_g)$$

where m_g is the sub-vector of the mixture component, $\lambda_t(g)$ represents the posterior probability and t defines the observation time. These values are used to train the UBM [21] model where the T-matrix and i-vector are acquired. Fig. 2 shows the schematic representation of i-vector feature extraction [4]. Then, while using the Joint Factor Analysis (JFA) [4,22], the speaker and channel space are grouped altogether to perform the speaker segmentation. Thus, we intend to extract the feature vector with the aid of GMM super vector and mean super vector,

$$U = U_0 + Tn$$

where U and U_0 are defined as the UBM and mean vector, n defines the low dimensional matrix with respect to the normal distribution and T denotes the total variability. Thus, the feature information of i-vector representation for every segmented speaker signal is given as $I = \{I_1, I_2, \ldots, I_y\}$.

4.4. Speaker clustering using deep neural network

The speaker clustering is the second significant aspect for the speaker diarization system. Once we segment the speakers from the input or audio signal, the deep neural network (DNN) is utilized in this paper to cluster the segmented signals for the same speaker. Rather than the other neural network, the DNN [23] includes multiple numbers of hidden layers for both training phase and testing phase. It is also a multilayer perceptron network where the acoustic features are given as input to the visible layer. The major concern of DNN is the deep learning process, where the technical term "deep" represents the many hidden layers. The DNN has the tendency to handle the high dimensional data and unstructured data. Here, the segmented speaker signals are trained and grouped together with respect to the speaker identity. The core advantage of the DNN is that it has ability to cluster the signals or unlabelled data which provides the better diarization performance. The extracted i-vector representation is subjected as input to the DNN. Thus, the DNN network includes input layer or visible layer, hidden layer and output layer which are apparently deliberated below. Fig. 3 depicts the processing model for speaker clustering using DNN.

(a) Restricted Boltzmann Machines (RBM)

Since the DNN contains the multi layers, the each pair of layer is trained or processed by the help of restricted Boltzmann machine (RBM). Every layer in the DNN network constitutes the number of stochastic units or neurons which is used to make the connection between the layers. Furthermore, the units in the visible layer are denoted by the Gaussian or Bernoulli distribution, whereas in the hidden layer, the stochastic units are processed by the Bernoulli distribution. Thus, before performing the joint optimization of input and hidden layers, the RBM function processes the stochastic units. Thus, the mathematical formulation is given below.

Due to the Bernoulli and Gaussian distribution, the units in the visible layer represent the discrete values and hidden stochastic variable has binary values. Let the i-vector of the speaker signal is defined as $I = \{I_1, I_2, \ldots, I_y\}$ is given as input to the deep learning model, where the probability measure is done through the i-vector feature v and hidden units h along with the model parameters ϕ. Thus, the RBM [23] of the stochastic visible and hidden unit is expressed by,

$$R(v, h; \phi) = \frac{\exp(-e(v, h; \phi))}{K}$$

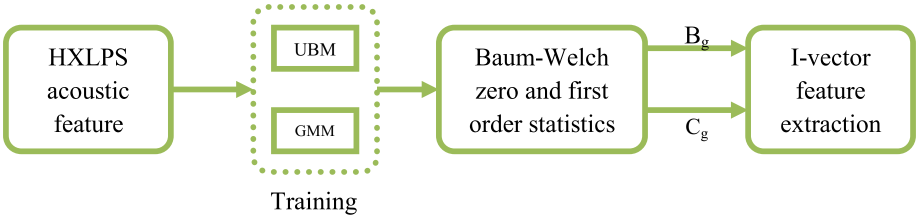

HXLPS acoustic feature
UBM
GMM
Training
Baum-Welch zero and first order statistics
B_g
C_g
I-vector feature extraction

Fig. 2 Schematic representation of i-vector feature extraction.

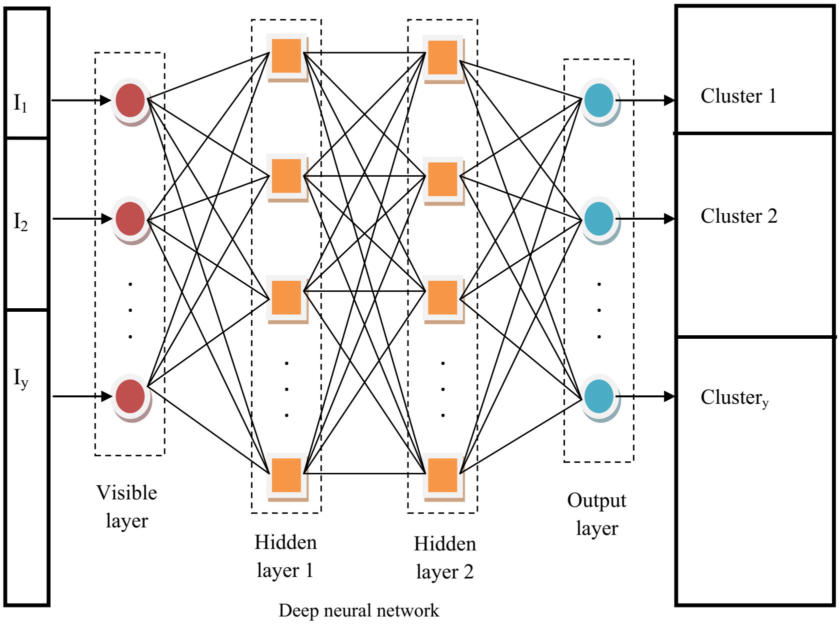


Fig. 3 Processing model for speaker clustering using DNN.

where $e(v,h;\phi)$ is the energy function and partition function K is defined as $K=\sum_I\sum_H \exp(-e(v,h;\phi))$. Then, the marginal probability of the visible stochastic units is given as,

$$R(v;\phi)=\frac{\sum_H \exp(-e(v,h;\phi))}{K}$$

Since the i-vectors in the visible layer exploit the Gaussian or Bernoulli distribution, the energy and conditional probability is determined by two combinations of distribution, which are (1) Bernoulli – Bernoulli RBM and (2) Gaussian – Bernoulli RBM. Therefore, the energy measure of Bernoulli of visible unit and Bernoulli of hidden unit is formulated as below.

$$e(v,h;\phi)=-\sum_{m=1}^{I}\sum_{n=1}^{H}\omega_{mn}v_m h_n-\sum_{m=1}^{I}\beta_m v_m-\sum_{n=1}^{H}\alpha_n h_n$$

where ω_{mn} represents the weight used to connect the visible and hidden units. Then, α_n and β_m are referred as the bias term weights for the hidden and visible layer and I,H denotes the number of i-vector feature and hidden stochastic variables. Then, the conditional probability is derived by,

$$R_c(h_n=1|v;\phi)=\mu\left(\sum_{m=1}^{I}\omega_{mn}v_m+\alpha_n\right)$$

$$R_c(v_m=1|h;\phi)=\mu\left(\sum_{n=1}^{H}\omega_{mn}h_n+\beta_m\right)$$

where $\mu(b)=1/1+\exp(-b)$.

Consequently, the energy measure and conditional probability for Gaussian (visible units) and Bernoulli (hidden units) [23] is evaluated as follows:

$$e(V,H;\phi)=-\sum_{m=1}^{I}\sum_{n=1}^{H}\omega_{mn}v_m h_n+\frac{1}{2}\sum_{m=1}^{I}(v_m-\beta_m)^2-\sum_{n=1}^{H}\alpha_n h_n$$

Then, the conditional probability of Gaussian – Bernoulli RBM is given as follows:

$$R_c(h_n=1|v;\phi)=\mu\left(\sum_{m=1}^{I}\omega_{mn}v_m+\alpha_n\right)$$

$$R_c(v_m=1|h;\phi)=\eta\left(\sum_{n=1}^{H}\omega_{mn}h_n+\beta_m,1\right)$$

where the hidden units are represented by the Bernoulli distribution function. Contrary, the Gaussian distribution of mean and variance one is employed to measure the conditional probability of the visible neurons.

(b) Speaker clustering

The learning process of the DNN [23] is that the features of the acoustic signal are trained by the Gaussian and Bernoulli RBM activation function. Thus, the output layer provides the probability measure of the i-vector belonging to the speaker using softmax operation. It is defined by,

$$R(t=s|h;\phi)=\frac{\exp\left(\sum_{n=1}^{H}\delta_{is}h_i+\alpha_s\right)}{K(h)}$$

where the visible unit t has been classified through the Gaussian and Bernoulli RBM function which provides the output for the s^{th} class label, and δ_{is} is the weight between the last hidden layer and the output layer. Hence, each signal feature is processed through the many hidden layers using the RBM activation function. On the other hand at the testing phase of the DNN, the testing audio signal is classified by the deep learning method. Then, the classifier acquires the class labels of speaker signals. Based on the speaker identity, the segmented signals are clustered together in the same speaker which enhances the speaker diarization performance.

5. Results and discussion

The experimental results of the proposed system and performance analysis are deliberated in this section using MATLAB implementation. Then, the performance of the speaker diarization system is compared with the existing system.

5.1. Experimental setup

(i) Dataset Description: The ELSDSR [17] dataset is utilized for experimenting the proposed speaker diarization system using HXLPS and DNN. English Language Speech Database for Speaker Recognition (ELSDSR) is used to provide audio signals for the evaluation of the automatic speaker recognition system. This database constitutes the audio signals of 22 speakers, which includes 12 male speakers and 10 female speakers.

(ii) Evaluation parameters: The performance of the speaker diarization system is analysed by the diarization error rate, false alarm rate and tracking distance. The evaluation metrics are determined as below:

Diarization error rate: The DER is the main metric for the performance evaluation. It consists of establishing a mapping between the speaker tags and evaluating the errors of the audio signals. The DER is calculated as follows:

$$Diarization\ Error\ rate,\ DER = \frac{Confusion\ error + Miss\ error + false\ alarm}{total\ reference\ Speech\,time}$$

False alarm rate (FAR): It is defined as the measure of determination of non-speech segment incorrectly. Thus, the false alarm rate is determined as,

$$False\,alarm\,rate, = 1 - Specificity$$

where specificity is expressed in terms of true negative (TN) and false positive (FP).

Tracking distance:

Due to lower value of tracking distance, the proposed method provides the better speaker diarization performance. Thus, it is evaluated by the distance measure between the original and speaker output signal. It is given by,

$$Tracking\ Distance = \sqrt{(x_i^O - x_i^g)^2}$$

(iii) Comparative methods

MFCC: It is widely used in the speaker diarization system to extract the acoustic features. The features are then used to segment the speaker signals from the input audio signal. Thus, the signals are segmented by the MFCC features which are then grouped by the Integer Linear Programming (ILP) [26] based clustering mechanism and Lion algorithm [24].

TMFCC: In TMFCC, the MFCC feature is incorporated with the tangent weighted function developed for the speaker diarization system. Thus, the acoustic features are extracted by the TMFCC method. Then, the signals are segmented with respect to the speaker identity. Finally, the speaker clustering is performed with the aid of ILP [26] clustering mechanism and Lion algorithm [24].

MKMFCC with WLI Fuzzy: The signal features are extracted by the multi kernel based MFCC extraction method. After the features are obtained, the speech activity detection is employed to segment the signals. Then, the WLI fuzzy clustering [25] is utilized to group the signals of the same speaker based on the centroid.

XLPS with DNN: The [18] eXtended Linear Prediction with autocorrelation Snapshot (XLPS) is the feature extraction method to provide the audio features of the speaker. Then, the VAD is used to determine the speech and non-speech regions of the input audio signals. Then, the signals are clustered using the deep learning method of DNN [23] network.

Proposed HXLPS with DNN: The proposed HXLPS is developed by integrating the Holoentropy function into XLPS method. Thus, the acoustic features are obtained. The VAD is used to segment the signal which then leads to acquire the i-vector representation of the signal. This representation is fed as input to the DNN [23] network, where the grouping of the signal of the same speaker is performed.

(iv) Experimental datasets: The performance analysis over the audio signal contains multiple numbers of speakers used for the proposed speaker diarization system. Here, the dataset-1 is generated by the input speech signal which constitutes five different speakers. Also, in dataset-2, the speech signal is generated using seven different speakers.

(v) Experimental results: Fig. 4 represents the experimental results of the HXLPS based speaker diarization system. The speaker diarization is performed by the audio signals of the multi number of speaker. Fig. 4a shows the input signal of the proposed system containing five different speakers used for speaker segmentation and clustering. The feature of the input signal is extorted using the proposed Holoentropy based XLPS feature extraction method. The segmented signals that are represented in terms of amplitude with varying number of audio frames are shown in Fig. 4b.

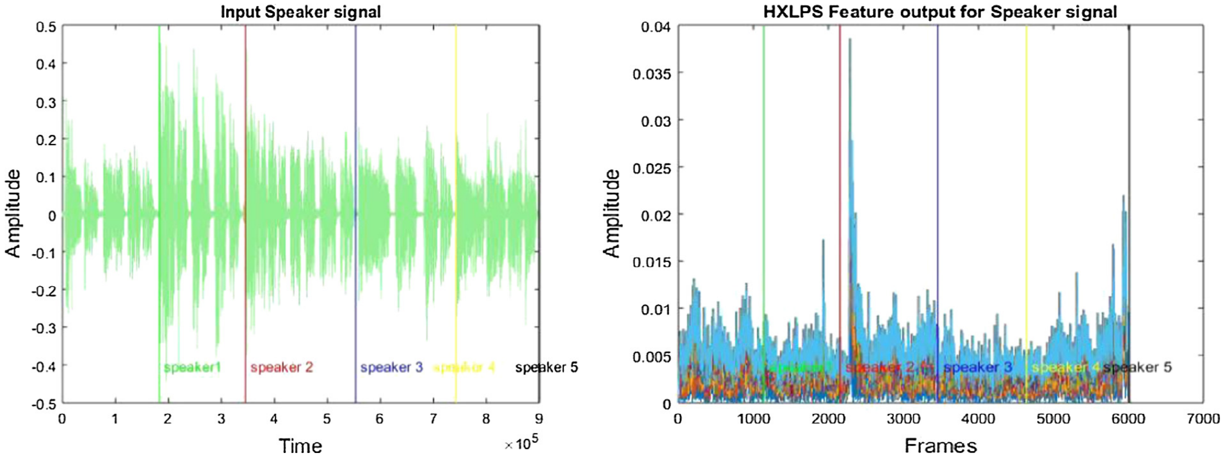
Similarly, Fig. 5 depicts the experimental results using seven different speaker signals. The input speech signal is shown in Fig. 5a. Thus, the input signal includes seven speakers used for the performance evaluation. Then, Fig. 5b shows the extracted feature representation of the audio signal using novel HXLPS extraction method. Thus, the voice activity of speaker signal is detected based on the acoustic features.

5.2. Performance analysis

This section presents the performance analysis of the proposed speaker diarization system based on the metrics, such as tracking distance, FAR and DER. Then, the analysed performance is compared with the existing systems.

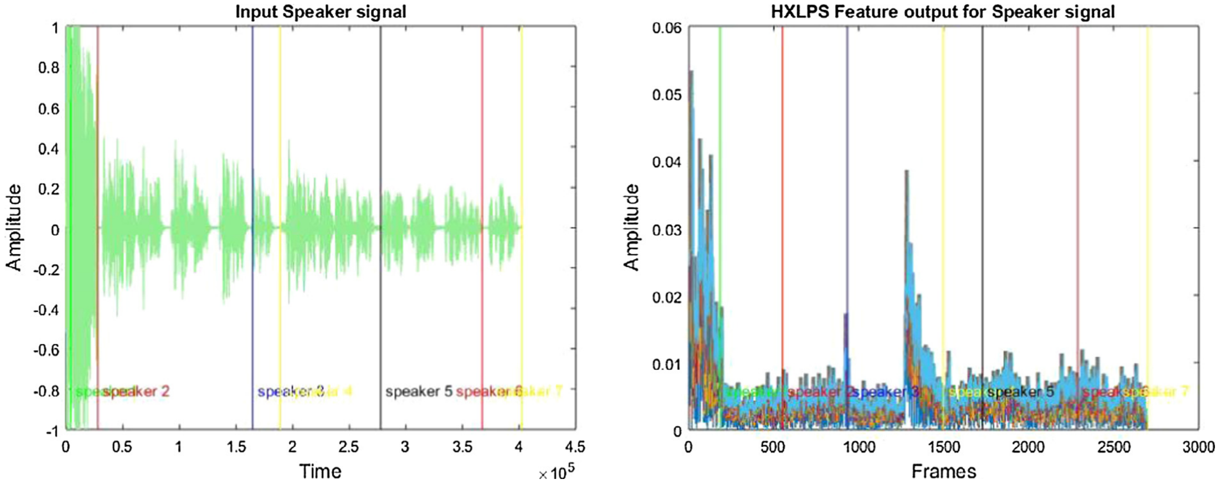
(i) Analysis based on Lambda

(a) Tracking distance: Fig. 6 represents the tracking distance performance of the proposed speaker diarization system. Fig. 6a shows the tracking distance performance for the dataset-1. When the lambda value is eight, the existing MFCC with Lion algorithm has the tracking distance of 4928, the TMFCC with ILP has the tracking distance of 5188 and also, MKMFCC with WLI fuzzy has the tracking distance of 1696.4 between the original and speaker output signal. But, the proposed system achieves the tracking distance of 1527.3 which is 169.1 lesser than the existing system. Similarly, Fig. 6b represents the tracking distance performance for the dataset-2. When the lambda value is six, the distance measure of 6119 and 4675.2 is achieved for MFCC with ILP and MFCC with Lion, and 5830 and 3886.7 for TMFCC with ILP and Lion algorithm and 3086.7 is attained for XLPS with DNN. But, the proposed HXLPS with DNN attains the tracking distance 1389 which is 188.2 lower than the existing MKMFCC with WLI fuzzy clustering mechanism which is demonstrated in



(a) Input speech signal (b) HXLPS feature representation

Fig. 4 Plot representation of the five different speaker signals.



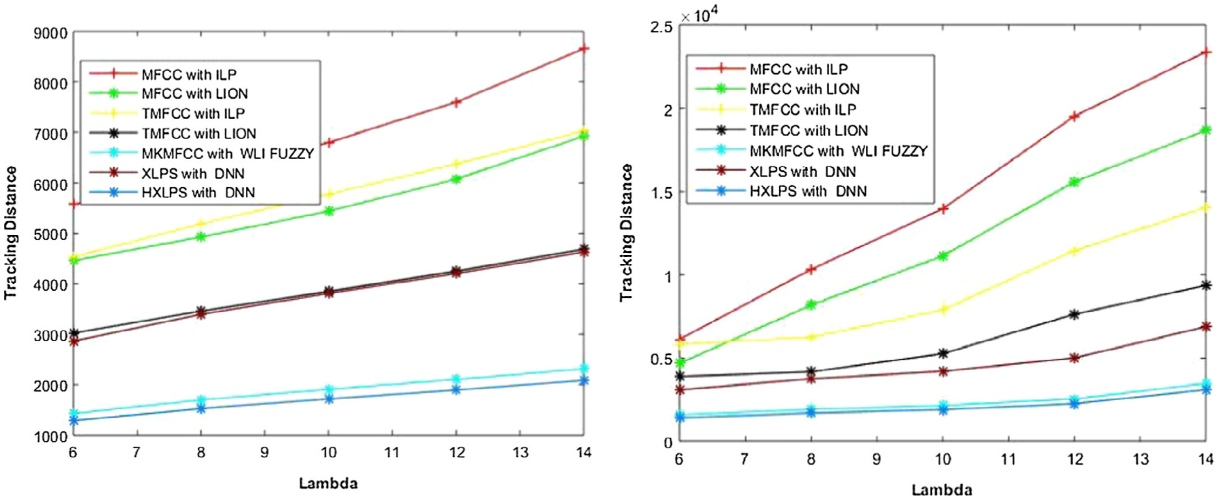
(a) Input speech signal (b) HXLPS feature representation

Fig. 5 Plot representation of the seven different speaker signals.

Fig. 6b. The reduced tracking distance is achieved which ensures the efficiency of proposed speaker diarization system.

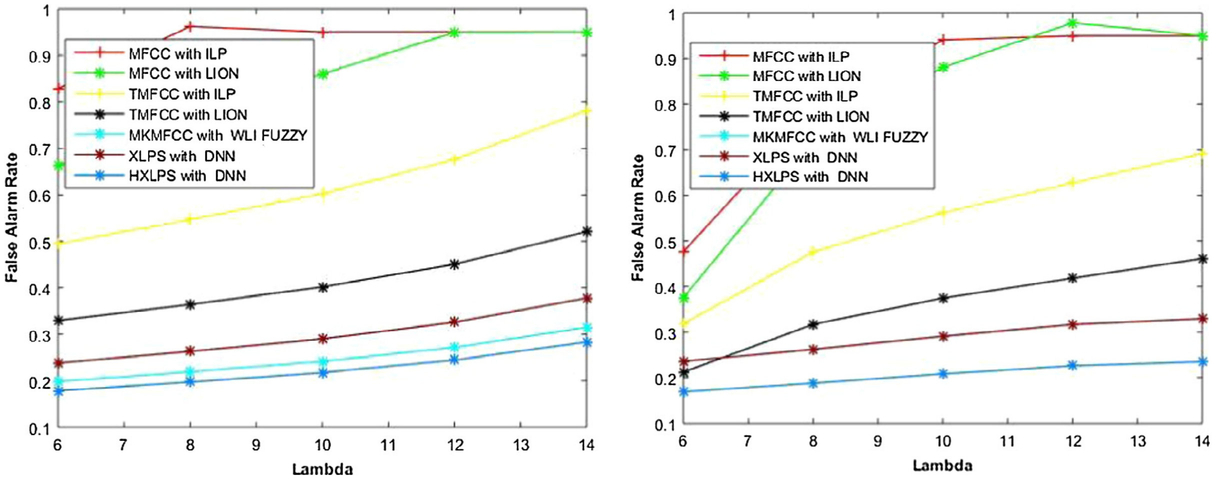
(b) FAR: Fig. 7 shows the performance of the FAR based on the lambda value. The lambda is the measure which is used to detect the speech and non-speech segments of the input audio signal. Fig. 7a represents the false alarm measure for the dataset-1. The false alarm is the measure of incorrectly labelled non-speech segments. When the lambda value is ten, FAR of 95% for MFCC with ILP clustering, 85.9% for MFCC with Lion algorithm, and FAR of 60.25% and 40.16% are obtained by TMFCC with ILP and Lion algorithm. Also, the existing of multi kernel MFCC achieves 24.16% false alarm rate and 29% for XLPS with DNN network. Comparing to the existing methods, the proposed HXLPS with DNN network attains the minimal FAR of 21.75% value. Similarly, Fig. 7b depicts the FAR performance analysis for the dataset-2. While using six lambda values for the voice activity detection, the existing TMFCC with ILP clustering mechanism acquires FAR of 31.83%. The FAR is then gradually increased to 46.12%. But, the proposed method obtains FAR of 16.9% while increasing the lambda value, the FAR of 23.49% is obtained rather than the existing methods.

(c) DER: Fig. 8 shows the DER performance for the speaker signal. Fig. 8a depicts the performance analysis for the input signal of five different speakers. The existing method of MKMFCC feature with WLI fuzzy clustering achieves DER of 13.2% which is then moderately increased to 19.72%. But, the proposed method acquires the minimum DER of 9.9% based on the lambda value compared to the MFCC with ILP and Lion algorithm, TMFCC with ILP and Lion algorithm, MKMFCC with WLI fuzzy clustering and XLPS with DNN network. Then, the DER performance for the dataset-2 is represented in Fig. 8b. While using the existing MFCC with Lion algorithm, DER of 6.64% is obtained when lambda value is six. It is then greatly increased to 95%. However, the proposed HXLPS method attains the DER of 1.3% which is slightly increased to 5.12% while increasing the lambda value. Thus, due to the lower value of DER, the proposed speaker diarization caters the better performance.



(a) Dataset-1 (b) Dataset-2

Fig. 6 Performance analysis for tracking distance.



(a) Dataset-1 (b) Dataset-2

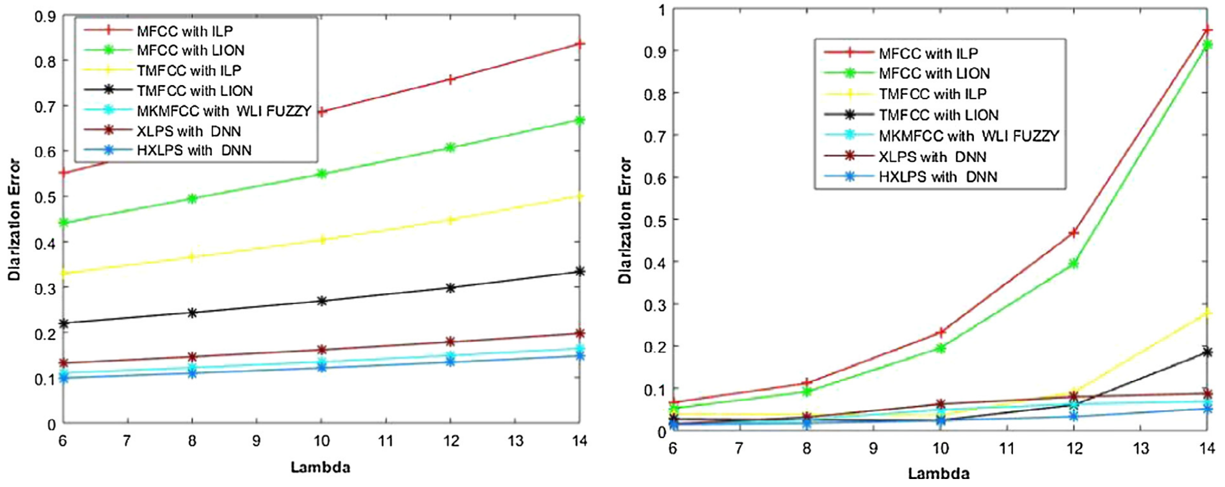
Fig. 7 Performance analysis of false alarm rate.

(ii) Analysis based on frame length

(a) Tracking distance: Fig. 9 shows the performance analysis of the tracking distance measure. The distance measure between the original and speaker output signal is analysed by the five different speaker signal deliberated in Fig. 9a. The existing XLPS with DNN method attains the tracking distance measure of 3825 which is gradually increased to 9199. But, the proposed method of speaker diarization acquires the distance measure of 1721.2 which is also moderately increased to 4139.6 while increasing the frame length of the input speech signal. Consequently, Fig. 9b depicts the performance measure for the dataset-2. In this dataset, the input speech signal contains the signal of seven different speakers. The distance measure of 3917.1, 4100.8, 4674.7, 5704.8 and 6550.6 is obtained by the existing method of TMFCC with Lion algorithm. But, the proposed method achieves the tracking distance of 784.71 which is lower than the XLPS with DNN when the frame length is 0.03. Also, the proposed method attains the tracking distance of 1091.6 which is lesser than the existing MKMFCC with WLI clustering method which is demonstrated in Fig. 9b. From Fig. 9, the proposed method obtains the minimal tracking distance when compared to the MFCC with ILP and Lion algorithm, TMFCC with ILP and Lion, MKMFCC with WLI fuzzy clustering and XLPS with DNN.

(b) FAR: The performance analysis using dataset-1 and dataset-2 is shown in Fig. 10. Based on the frame length, the FAR is evaluated for the different speakers. Fig. 10a demonstrates the FAR for the dataset-1. When the speech signal contains six frame lengths, the existing XLPS with DNN method achieves the FAR of 9.33%. It is then gradually increased to FAR of 22.53% while increasing the number of frame lengths. But, the proposed method attains FAR of 5.13% which is then slightly increased to 16.9%. However, compared to the existing

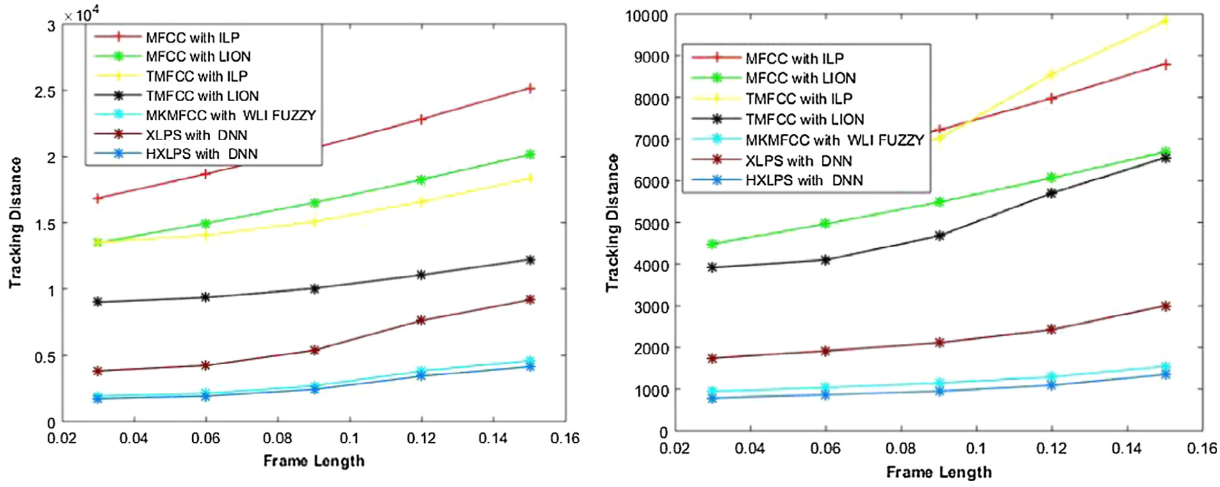
Please cite this article in press as: V. Subba Ramaiah, R. Rajeswara Rao, Speaker diarization system using HXLPS and deep neural network, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2016.12.009



(a) Dataset-1

(b) Dataset-2

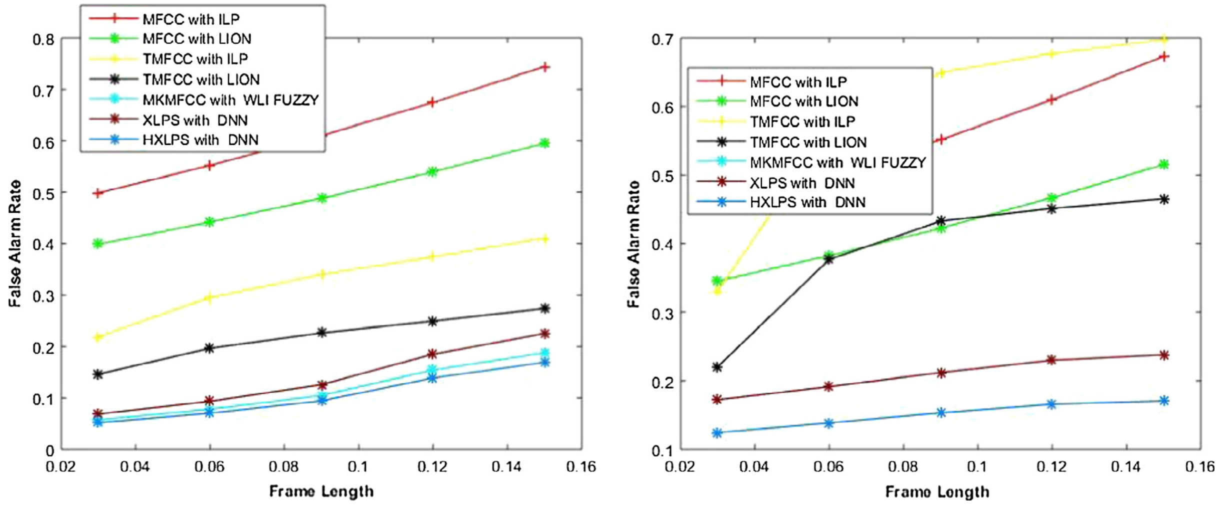
Fig. 8 Performance analysis of diarization error rate.



(a) Dataset-1

(b) Dataset-2

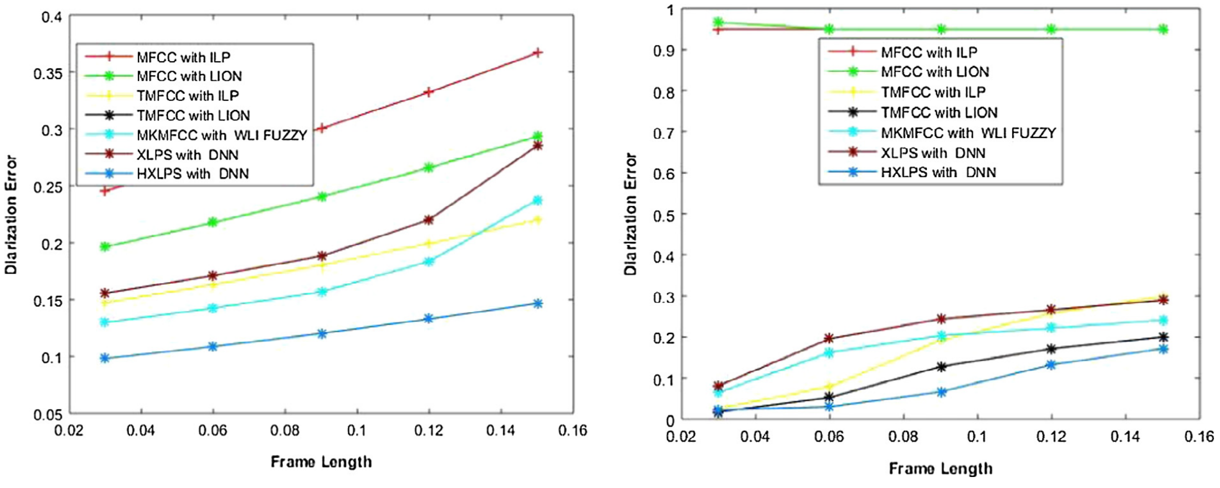
Fig. 9 Performance analysis for tracking distance.



(a) Dataset-1

(b) Dataset-2

Fig. 10 Performance analysis of the false alarm rate.



(a) Dataset-1

(b) Dataset-2

Fig. 11 Performance analysis of DER.

system, the proposed speaker diarization method attains the minimal FAR. Fig. 10b depicts the performance analysis of the seven different speakers of the input speech signal. The proposed HXLPS with deep neural network method achieves 12.4%, 13.82%, 15.31%, 16.57% and 17% depending on the frame length when compared to the MFCC with ILP, MFCC with LION, TMFCC with ILP, TMFCC with Lion algorithm, MKMFCC with WLI fuzzy clustering and XLPS with DNN.

(c) DER: Fig. 11a shows the performance analysis based on the frame length for dataset-1. The existing MFCC with Lion algorithm acquires the DER of 19.64%, 21.74%, 24.03%, 26.56% and 29.13% by varying the frame length from 3 to 15. But, the proposed method of HXLPS with DNN caters the minimal DER of 9.82% rather than the existing system, like ILP and Lion algorithm for MFCC and TMFCC feature extraction, MKMFCC with WLI and finally XLPS with DNN. Subsequently, the DER performance analysis for the seven different speakers of the input speech signal is demonstrated in Fig. 11b. When the frame length of the input speech signal is 12, 95% of DER is obtained by both MFCC with ILP clustering and Lion algorithm, 25.74% for TMFCC with ILP, and 17.16% for TMFCC with Lion algorithm; 22.17% and 26.59% DER is achieved by MKMFCC with WLI fuzzy clustering and XLPS with DNN. Comparing to the existing system, the proposed holoentropy based XLPS with DNN attains minimal DER of 2.23%. Thus, we infer from Fig. 11 that the lower value of DER leads to provide the better speaker diarization performance.

6. Conclusion

In this paper, we have presented the speaker diarization system using proposed HXLPS feature extraction and DNN. The novel method of HXLPS was designed by extending the XLPS with holoentropy function. After the features are extracted, the signals are segmented which were then used to detect the speech and non-speech signals. Consequently, the features were used to provide the i-vector representation of the every segmented signal. For speaker clustering, the DNN was employed to perform the clustering or grouping the audio signals of the same speaker. Finally, the experimental results were evaluated and the performance was analysed by the evaluation metrics which was then compared with the existing systems. Thus, the proposed method achieved the lower DER of 1.36% and 2.23% based on lambda value and frame length which proves the effectiveness of the speaker diarization system.

References

[1] S. Jothilakshmi, V. Ramalingam, S. Palanivel, Speaker diarization using autoassociative neural networks, Eng. Appl. Artif. Intell. 22 (2009) 667–675.

[2] Benjamin Bigot, Isabelle Ferrane, Julien Pinquier, Regine Andre-Obrecht, Detecting individual role using features extracted from speaker diarization results, Multimedia Tools Applicat. 60 (2) (2012) 347–369.

[3] Stephen H. Shum, Najim Dehak, Reda Dehak, James R. Glass, Unsupervised methods for speaker diarization: an integrated and iterative approach, IEEE Trans. Audio Speech Lang. Process. 21 (10) (2013) 2015–2028.

[4] Yan Xu, Ian McLoughlin, Yan Song, Kui Wu, Improved i-vector representation for speaker diarization, Circ. Syst. Signal Process. 35 (9) (2016) 3393–3404.

[5] Vishwa Gupta, Speaker change point detection using deep neural nets, Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (2015).

[6] Martin Zelenák, Carlos Segura, Jordi Luque, Javier Hernando, Simultaneous speech detection with spatial features for speaker diarization, IEEE Trans. Audio Speech Lang. Process. 20 (2) (2012) 436–446.

[7] Deepu Vijayasenan, Fabio Valente, Herve Bourlard, An information theoretic combination of MFCC and TDOA features for speaker diarization, IEEE Trans. Audio Speech Lang. Process. 19 (2) (2011) 431–438.

[8] Félicien Vallet, Slim Essid, Jean Carrive, A multimodal approach to speaker diarization on TV talk-shows, IEEE Trans. Multimedia 15 (3) (2013) 509–520.

[9] Nicholas Evans, Simon Bozonnet, Dong Wang, Corinne Fredouille, Raphael Troncy, A comparative study of bottom-

Please cite this article in press as: V. Subba Ramaiah, R. Rajeswara Rao, Speaker diarization system using HXLPS and deep neural network, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2016.12.009

up and top-down approaches to speaker diarization, IEEE Trans. Audio Speech Lang. Process. 20 (2) (2012) 382–392.

[10] P. Pertila, Online blind speech separation using multiple acoustic speaker tracking and time–frequency masking, Comput. Speech Lang. 27 (3) (2013) 683–702.

[11] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-Francois Bonastre, Laurent Besacier, Step-by-step and integrated approaches in broadcast news speaker diarization, Comput. Speech Lang. 20 (3) (2006) 303–330.

[12] Margarita Kotti, Vassiliki Moschou, Constantine Kotropoulos, Speaker segmentation and clustering, Signal Process. 88 (2008) 1091–1124.

[13] Mathieu Hu, Dushyant Sharma, Simon Doclo, Mike Brookes, Patrick A. Naylor, Speaker Change Detection and speaker diarization using spatial information, Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (2015).

[14] Liping Zhu, A modified approach to cluster refinement for speaker diarization, Proc. IEEE Int. Conf. Computer Sci. Netw. Technol. (2015).

[15] Kyu J. Han, Shrikanth S. Narayanan, A robust stopping criterion for agglomerative hierarchical clustering in a speaker diarization system, in: Proceedings of INTERSPEECH Eighth Annual Conference of the International Speech Communication, 2007.

[16] Srikanth R. Madikeri, A fast and scalable hybrid FA/PPCA-based framework for speaker recognition, Digit. Sign. Process. 32 (2014) 137–145.

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

[18] 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).

[19] V.B. Le, O. Mella, D. Fohr, Speaker diarization using normalized cross likelihood ratio, INTERSPEECH 7 (2007) 1869–1872.

[20] D.A. Reynolds, R.C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, IEEE Trans. Speech Audio Process. 3 (1) (1995) 72–83.

[21] Douglas Reynolds, Universal background models, Encycl. Biomet. (2009) 1349–1352.

[22] S. Madikeri, I. Himawan, P. Motlicek, M. Ferras, Integrating Online I-vector extractor with Information Bottleneck based Speaker Diarization system, Idiap, 2015.

[23] Sabato Marco Siniscalchi, Dong Yu, Li Deng, Chin-Hui Lee, Exploiting deep neural networks for detection-based speech recognition, Neurocomputing 106 (2013) 148–157.

[24] B. Rajakumar, The Lion′s Algorithm: a new nature-inspired search algorithm, Procedia 6 (2012) 126–135.

[25] 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. 23 (3) (2013) 1–16.

[26] 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. 14 (7) (2007).

Please cite this article in press as: V. Subba Ramaiah, R. Rajeswara Rao, Speaker diarization system using HXLPS and deep neural network, Alexandria Eng. J. (2017), http://dx.doi.org/10.1016/j.aej.2016.12.009