Optimal Trained Artificial Neural Network for Telugu Speaker Diarization

[1]Sethuram.V, [2]Ande Prasad and [3]R. Rajeswara Rao

[1]Assistant Professor, Rashtriya Sanskrit Vidyapeetha

[2.3]Assistant Professor, VikramaSimhapuri University

Abstract
Speaker indexing or diarization is the process of automatically partitioning the conversation involving multiple speakers into homogeneous segments and grouping together all the segments that correspond to the same speaker. So far, certain works have been done under this aspect; still, the need of accurate partitioning process gets lagged under certain criteria. With this in mind, this paper aims to introduce a new speaker indexing or diarization model (Telugu language) that initially involves Mel Frequency Cepstral coefficient (MFCC) based feature extraction. Subsequently, a new Optimized Artificial Neural Network (ANN) is introduced for clustering process. The novelty behind the clustering process is: the training of ANN takes place through optimization logic that updates the weight of ANN by a hybrid concept of Artificial Bee Colony (ABC) and Lion Algorithm (LA). Thereby, the proposed model is named as ANN-ABC-LA model. Finally, the performance of the proposed ANN-ABC-LA model is compared over the state-of-the-art models with respect to different performance measures.

Keywords:
Speaker Diarization; Feature Extraction; Neural Network; Lion Algorithm; Artificial Bee Colony

1. Introduction

With the innovation of smart technologies in the field of engineering, a lot of intelligent and efficient methodologies were emerged to enhance the standard of human life. When it comes to human machine interaction, there is an ever increasing demand to develop automated human language recognition models that enables proficient and intellectual ways to communication [9]. For this purpose, a capable exploring, indexing, and retrieving techniques are necessitated for audio signals interaction. In addition, extortion of the speech signals recorded using speech recognition system gives a deep base for the chores yet, the words are obviously complex to read and cover each data involves in the audio signal [11]. All these difficulties urge the introduction of audio diarization technique. Generally, speaker diarization can be defined as the method of interpreting an input audio signal into the data which overlaps temporal areas of signal energy with its particular sources [10][12]. Moreover, it divides the audio signal into homogenous segments with respect to the audio recognition. The different sources of audio input can have music, speakers, signals, background noises, diverse channel properties, etc [13]. In addition, diarization is utilized in assisting speech recognition, promoting audio search facilities, and audio archives indexing, and further maximizing the quality of automated dictations as well [15].

Some methods focus on the speech signals that include speech quotes by means of speakers. Particularly, the spoken signals are single-channel inputs which contain several audio sources [14][16]. Furthermore, these are obtained from various speakers, noises, music, and so on and the formats and information about the audio inputs can be application-specific with respect to the sources. In addition to this, the term speaker diarization can be referred as audio indexing, segmentation, clustering, etc according to the trend of the researching people [17][19]. Specifically, speaker diarization identifies the speech signals such as input from speakers (i.e., male or female) and non-speech signals (music, noises, etc) and partitions the audio into similar segments [18][22].

There are a lot of techniques were established to enhance the performance of audio diarization such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), etc., which enables the audio segmentation and clustering in broadcasting audio streams [20]. However, the removal of unnecessary audio signals such as noises is being a complex task in audio diarization techniques. Further, the limitations such as determining the structure of

Manuscript received February 5, 2026
Manuscript revised February 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.2.10

the broadcasting application and audio signal heterogeneity, quality of audio signals from various sources, etc., bound the usage of audio diarization techniques [21][23]. Other methods are necessitating to improve the performance of the speaker diarization techniques that allows the dictations more interpretable and efficient. For this reason, optimization algorithms are utilized to accomplish the existing challenges and provide intellectual technologies [24][25]. Thus, there exists a large scope for the researchers to introduce novel technologies to improve the efficiency of speaker diarization.

The main contribution of this paper is as follows.

1. As a novelty, this paper addresses a new speaker indexing or diarization model, which is accomplished via speaker clustering.
2. Generally, speaker clustering plays an essential role in speaker diarization.
3. For this purpose, the speaker clustering process is optimally done using optimized ANN. Moreover, the proposed ANN is optimally trained via the efficiency of the metaheuristic algorithms named ANN-ABC-LA.
4. Finally, the efficiency of the proposed ANN-ABC-LA model is compared over the state-of-the-art models.

The organization of this paper is in this order: Section II presents the literature regarding speaker diarization models. Proposed speaker diarization model using the ANN model under optimization-based training is illustrated in Section III. Proposed clustering process via optimized ANN is demonstrated in Section IV. Section V gives a short description of used optimization algorithms. Section VI represents the attained results, and Section VII concludes the paper.

2. LITERATURE REVIEW

2.1 Related Works

In 2010, He *et al.* [1] have proposed a speaker indexing model by integrating the efficiency of Bayesian Information Criterion (BIC) with GMM and Jensen's inequality named GMM-BIC to enhance the performance of audio model creation. From the empirical results, it was revealed that the proposed GMM-BIC achieved improved results and outperformed the conventional Single Gaussian Model-BIC (SGM-BIC).

In 2013, Vaquero *et al.* [2] have developed a speaker characterization approach using HMM that used a group of confidence metrics to estimate the superiority hypothetical diarization results according to identify the audio signals which were perfectly partitioned. Moreover, the confidence metrics were employed to receive the expected signals and eliminate unnecessary signals. The simulation analysis confirmed the efficacy of the speaker characterization technique.

In 2014, Yella and Bourlard [3] have established a speaker diarization approach using HMM/GMM model by taking the features with a maximum level of data like silence and speaker vary statistics for enhancing the acoustic features in single speaker speech signals. In addition, the overlapping probability was also determined using a long-term window with various corpora. The empirical outcomes verified the performance in terms of Diarization Error Rate (DER) and laughter overlap identification.

In 2018, Ramaiah, and Rao [4] have established a speaker diarization model through Deep Neural Network (DNN) and Holoentropy-eXtended Linear Prediction with autocorrelation Snapshot (HXLPS) methods to enhance the performance by extracting and classifying the features and the extracted features were identified using Voice Activity Detection (VAD) technique. The performance of the HXLPS and DNN was verified and attained efficient results by means of DER and false alarm rate.

In 2009, Jothilakshmi *et al.* [5] have addressed speaker diarization technique through Auto-Associative Neural Networks (AANN), which included partitioning a speech into homogeneous fragments and further grouped as speaker clusters. Furthermore, it utilized Mel Frequency Cepstral Coefficients (MFCC) to take the speaker data and was tested by various speaker datasets. The simulation outcomes validated the performance through comparative work with the existing methods.

In 2012, Vijayasenan *et al.* [6] have presented a diarization system using HMM/GMM method with MFCC as well as Time Delay of Arrivals (TDOA) features to enhance the efficiency of multi-stream diarization of audio signals. Besides, the log-likelihood integration model was employed to generalize the space of related parameters. In addition, a non-parametric multi-stream diarization technique was proposed using Information Bottleneck (IB) technique. The empirical work revealed the efficiency by means of error sensitivity.

In 2008, Fergani *et al.* [7] have developed a One-class Support Vector Machines (1-SVM) to construct and organize an efficient speaker diarization scheme. In this model, it utilized n-dimensional heterogeneous acoustic feature vectors and the proposed 1-SVM classified the segments of various features. The efficiency of 1-SVM approach was verified through a comparison work and outperformed the conventional models.

In 2007, Gupta *et al.* [8] have proposed a speaker diarization technique using BIC, Speaker Identification

(SID), and GMM techniques to perform various clustering approaches and Viterbi re-segmentation. Moreover, these methods were performed in MFCC and in Gaussianized MFCC features. From the empirical results, it was confirmed that the GMM method produced minimized DER value and proved its efficiency.

2.2 Review

Table 1 summarizes the features and challenges of conventional models for speaker indexing system implemented through various approaches which possess the potential to be applied in real-world applications. The GMM-BIC [1] provided more speaker data and attained better Speaker Indexing Accuracy (SIA). However, MFCC features extraction was complex and it required high computational time. The HMM [2] model achieved reliability and attained minimized DER rate but segmenting the audio signal were difficult and resulted in performance degradation due to huge overlapped speech. The HMM/GMM [3] provided automated feature extraction and gained minimized DER yet, in overlap labeling, it produced high errors and extracting the acoustic features were complex. The HXLPS and DNN [4] model attained better performance for both speech and non-speech signals and accomplished minimized DER still, the performance was efficient only for MFCC features and the feature extraction was complex. The HMM/GMM [5] model attained better segmentation performance and obtained reduced DER but it has poor performance for acoustic features and ineffective in the removal of background noises. The HMM/GMM [6] model attained improved performance for MFCC and TDOA features and provided faster computation. However, it has poor performance for the speech signals in microphone and low scalability. The 1-SVM [7] model attained minimized error and offered simpler implementation still, it has low robustness against a large dataset and increased computational load decreased the efficiency. The GMM [8] model attained minimized error rate and highly noise sensitive yet, it has complex segmentation process and training the dataset was difficult. Thus, there exists a vast opening for the researchers and scholars to develop new methods and models for accomplishing the above-mentioned challenges and provide improved performance in speaker diarization system.

Table 1: Features and challenges of conventional models for Speaker Indexing System by various methods

Author [Citation]	Method	Features	Challenges
He *et al.* [1]	GMM-BIC	• Provided more speaker data • Attained better SIA	• MFCC features extraction was complex • Requires high computational time
Vaquero *et al.* [2]	HMM	• Achieved reliability • Attained minimized DER rate	• Segmenting the audio signal were difficult • Performance degradation due to huge overlapped speech
Yella and Bourlard [3]	HMM/GMM	• Provided automated feature extraction • Gained minimized DER	• In overlap labeling, it produced high errors • Extracting the acoustic features were complex
Ramaiah, and Rao [4]	HXLPS and DNN	• Attained better performance for both speech and non-speech signals • Accomplished minimized DER	• The performance was efficient only for MFCC features • Feature extraction was complex
Jothilakshmi *et al.* [5]	HMM/GMM	• Attained better segmentation performance • Obtained reduced DER	• Poor performance for acoustic features • Ineffective in the removal of background noises
Vijayasenan *et al.* [6]	HMM/GMM	• Attained improved performance for MFCC and TDOA features • Provided faster computation	• Poor performance for the speech signals in microphone • Poor scalability
Fergani *et al.* [7]	1-SVM	• Attained minimized error • Offered simpler implementation	• Low robustness against a large dataset • Increased computational load decreased the efficiency
Gupta *et al.* [8]	GMM	• Minimized error rate • Highly noise sensitive	• Complex segmentation process • Training the dataset was difficult

3. Proposed Speaker Diarization Model using ANN Model under Optimization-based Training

3.1 Overall Structure of Proposed Model

Fig. 1 depicts the architecture of proposed speaker diarization model. The step-by-step process of speaker diarization model involves Feature Extraction, Speech Activity Detection (SAD), Speaker Segmentation and Speech Clustering process. Initially, the input audio signal is given to the system. All the aforementioned steps are carried out to attain the diarization output. Here, this paper plans to establish a new speaker indexing or diarization model using Telugu language via **Optimal trained ANN** for clustering process, where the training is carried out using a new hybrid algorithm (hybrid ABC and LA). Consider the input audio signal for the proposed speaker diarization system is S with multiple speakers, i.e., $S=\{S_1,S_2,\ldots S_M\}$, where M indicates the count of speakers. From the given input signal S, the MFCC features get extracted i.e., $F_x(N)=\{F_x(1),F_x(2),\ldots\}$ where $N=1,2,\ldots Z$. Subsequently, the speech activity detection takes place under two means such as silence removal and music removal. Besides, the segmentation process happens, which is explained in the next section. Finally, the clustering is attained using the optimally trained ANN into k clusters based on the identity of the speaker, which is represented as $C=C_1,C_2,\ldots C_c$ where $1\le c\le N$.

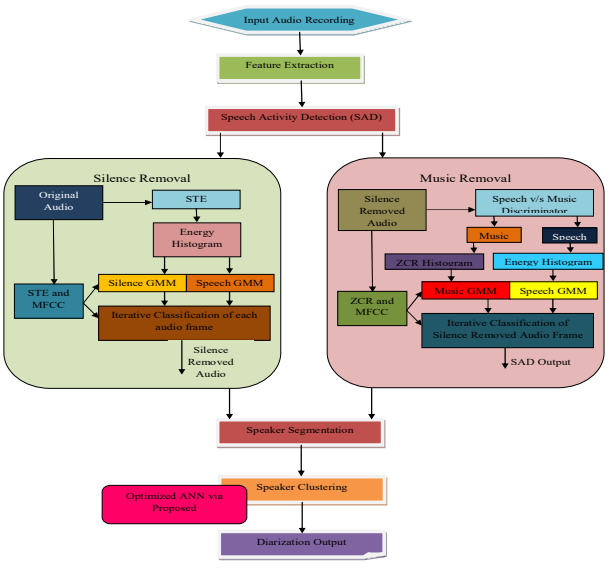


Figure 1 Geometrical Representation of Proposed Speaker Diarization Model

3.2 Feature Extraction

The MFCC features are extracted in this speaker diarization model. Here, the audio signal used for speaker indexing is transformed as framed signal with the frame reference i. Consider $y(v)$ as the input signal and the framed signal $y_i(v)$ in which v specifies the sample count and i indicates the frame count. The power spectrum is calculated as given in Eq. (1), in which $Y_i(z)$ be the discrete Fourier transform for the framed signal and it is given in Eq. (2).

$$p_i(z)=\frac{1}{V}|Y_i(z)|^2 \tag{1}$$

$$Y_i(z)=\sum_{v=1}^{V} y_i(v)h(v)e^{-j2\Pi zn};\ 1\le z\le Z \tag{2}$$

Consider L_f and H_f is the low and high frequency and the frequency is transformed as Mel scale as stated in Eq. (3). Eq. (3) is reconstructed using Eq. (4).

$$M(F)=1125+lv\left(1+\frac{F}{700}\right) \tag{3}$$

$$d(i)=(nfft)*h(i) \tag{4}$$

These values are used to produce the filter bank as represented in Eq. (5), in which $m=1\,to\,M$ indicating the filter count and $d(\ )$ refers to $m+2$ Mel spaced frequencies.

$$G_m(z)=\begin{cases}0 & z<d(m-1)\\ \dfrac{z-d(m-1)}{d(m)-d(m-1)} & d(m-1)\le z\le d(m)\\ \dfrac{d(m-1)-z}{d(m+1)-d(m)} & d(m)\le z\le d(m+1)\\ 0 & z>d(m-1)\end{cases} \tag{5}$$

3.3 Speech Activity Detection

In an audio recording process, an approach is implemented for partitioning speech from non-speech signals. It involves two main challenges such as (1) minimum missed speech as well as (2) minimum false alarm speech. The percentage of speech signal misclassified as non-speech signals via SAD are termed as Missed Speech Rate (MSR), while the percentage of non-speech signals identified as speech signals are referred as False Alarm Speech Rate (FASR). These two rates are the analytical measures for SAD. The model-based classifier is used in the implemented SAD subsystem. Furthermore, it is free from

training the speech as well as non-speech data. Here, the SAD subsystem is implemented via 2 decoupled stages. Initially, the silence present in the entire recording is eliminated by energy based bootstrapping continued with repetitive classification. Secondly, from the recording, music, as well as extra perceptible non-speech signals, are recognized. Herein, music elimination process takes the silence eliminated audio for music vs. speech bootstrap discriminator. The music audio signals having greater confidence are utilized for training the music model that is repeatedly refined.

Silence Removal: It is implemented using nineteen MFCC features combined to STE as well as its 1st and 2nd order derivatives. A confidence value is allotted for each frame by the bootstrap segmentation to both silences as well as speech classes. Besides, Gaussian mixtures with size 4 from 60 dimensional feature spaces are used to train the bootstrap silence model. Similarly, speech model is trained with the equal size. In this iterative step, all frames are classified as 2 classes such as speech and silence. In fact, the high confidence silence, as well as speech frame, is exploited for training the silence as well as speech signals for successive iterations. The count of 60 dimensional Gaussian utilized for modeling the silence as well as speech Gaussian Mixture Model (GMM) are increased with the increase in the iteration count. The optimal outcomes were attained while the GMM size of speech is 32 as well as non-speech is 16. Now, the silences, as well as pauses, are eliminated, still the music and jingles i.e., the audible non-speech is available.

3.4 Speaker Segmentation

The speaker segmentation algorithm used in this system is a growing window size w via Bayesian Inference Criterion ΔBIC distance. Initially, an investigation is carried out for a single speaker change from the starting of the audio and for each encounter change; the investigation is restarted on the next frame. Here, the search window is declared as well as a ΔBIC distance is estimated for all frames located within the window. When the maxima exceeds the threshold value ψ, the maxima notes it as a change. When no maxima are found on the window, the window size is incremented and this process is repeated till a change is encountered. After the removal of non-speech frames by the SAD, the speech signals are processed and after detecting the change points from the audio signals, its respective positions from the original audio are identified and noted as change points. Previously, the segmentation process is implemented via 2 stages. Initially, the ΔBIC based change recognition is carried out based on aforementioned threshold value. Secondly, the combination of successive segments takes place only for those having positive ΔBIC score. These 2 stages are implemented because of the over-segmentation problem in zero threshold ΔBIC based segmentation. With the intention of eliminating this 2 stage process, the maxima value that exceeds the threshold ψ is considered for further processes and effectively minimizes the over-segmentation as stated in Eq. (6).

$$\Delta BIC(y_i) = N\log|E| - N_1\log|E_1| - N_2\log|E_2| - \frac{\gamma}{2}\left(b + \frac{1}{2}b(b+1)\right)\log N \tag{6}$$

Consider L_f and H_f is the low and high frequency and the frequency is transformed as Mel scale as stated in Eq. (3). Eq. (3) is reconstructed using Eq. (4).

4. Proposed Clustering Process: Optimized Artificial Neural Network

4.1 Proposed Optimized ANN-base d Clustering Process

Clustering process involves collecting and combining segments of similar speakers. For this purpose, ANN [28] is implemented to cluster the segments, in which the training the done via the proposed model. Moreover, it is a well-known technique exploited in many applications due to its flexibility when compared to several other classifiers. In this paper, the selected features are utilized with ANN for significant feature classification. ANN is modeled as a network of artificial/biological neurons to compute Artificial Intelligence (AI) problems. Usually, the weights in ANN are indicated as connections of neurons, in which positive weight is termed as excitatory connection and negative one is inhibitory connection. Moreover, every input is transformed as weights and everyone is added. Eventually, the amplitude of the output is regulated through an activation function. Generally, ANN contains 3 layers like input layer, hidden layer, and output layer employed for training the output. The input and output neuron is represented as in_i and h_i in order. The hidden layer output H is stated in Eq. (7), in which, F is the input to ANN, af_1 specifies the nonlinear activation function, wi_{bh_i} points to the bias weight to h_i^{th} hidden layer and wi_{bo_i} indicates the bias weight to o_i^{th} output neuron. Furthermore, the output of the network is expressed in Eq. (8), in which af_2 specifies an activation function, Ei_2^* specifies output of o_i^{th} output neuron, $wi_{in_i h_i}$ denotes the weight from in_i^{th} input neuron to h_i^{th} hidden neuron and $wi_{h_i o_i}$ refers to the weight from h_i^{th} hidden neuron to o_i^{th} output neuron. Finally, Eq. (9) portrays the

entire output of the network (predicted). Moreover, the error among the predicted and actual values Ei_2^* is specified in Eq. (9), in which si signifies the actual output and $\hat{si}$ points to derived output.

$$H = af_1\left[wi_{bh_i} + \sum_{in_i=1}^{N_{in_i}} \left(F \times wi_{in_i h_i} \right) \right] \quad (7)$$

$$si = af_2\left[wi_{bo_i} + \sum_{h_i=1}^{N_{h_i}} \left(H \times wi_{h_i o_i} \right) \right] \quad (8)$$

$$Ei_2^* = \underset{\{wi_{bh_i}, wi_{bo_i}, wi_{in_i h_i}, wi_{h_i o_i}\}}{\arg\min} \sum_{o_i=1}^{N_{o_i}} \left| si - \hat{si} \right| \quad (9)$$

The main contribution goes with the training process (offline process) of ANN, where the weights $wi_{bh_i}, wi_{bo_i}, wi_{in_i h_i}, wi_{h_i o_i}$ together termed W is updated by the optimization concept. More particularly, a hybrid based updating process takes place with the introduction of specific logic:

Step 1: (i) Initialize a random value R
(ii) Fixing a threshold value $t = 0.5$.

Step 2: If the random value R falls less than the threshold $t = 0.5$, the weight W gets updated using the LA update as per Eq. (10). The detailed description of the LA model is given in the subsequent section.

$$s_n^{female+} = \begin{cases} s_m^{female+}; & when\ n = m \\ s_n^{female+}; & otherwise \end{cases} \quad (10)$$

Step 3: In the else condition, if the random value R is greater than the threshold $t = 0.5$, the weight W gets updated using ABC update as per Eq. (11). The clear explanation of the ABC algorithm is given in the next section. The pseudo-code of the proposed hybrid based training process is as follows.

$$S_u = bc_j + r(0,1) \cdot \left(ab_k - bc_k \right) \quad (11)$$

Table 1 Pseudocode of Proposed Hybrid based training

Pseudocode of Proposed Hybrid based training		
Initializing a random number ‘R’		
Set threshold ‘t=0.5’		
	If (R<t)	
		Weight ‘W’ is updated as per Eq. (10) //LA update
	else	
		Weight ‘W’ is updated as per Eq. (11) // ABC update
	End if	

In fact, the update process is done based on the defined objective function that is given in Eq. (12), where e indicates the actual and predicted value.

$$Obj = Min(e) \quad (12)$$

Thus the clustering process takes place in an efficient manner with respect to time consumption and accuracy rate. The diagrammatic representation of the proposed clustering process is shown in Fig. 2.

5. Short Description on Used Optimization Algorithms

5.1 Traditional LA

Traditionally, LA [26] is developed via the social behavior of lions which includes the territorial defense and takeover with 6 analytical phases such as (i) Pride generation, (ii) Fertility estimation, (iii) Mating, (iv) Territorial defense, (v) Territorial takeover and (vi) Termination.

Pride Generation: The initialization process defined in Eq. (13) represents the male S^{male} , female S^{female} , and a nomad lion S^{nomad} and Xi refers to the length of the lion where a and b specifies the numerical values to compute the length of the lions.

$$Xi = \begin{cases} a; & a > 1 \\ b; & o.w \end{cases} \quad (13)$$

Eq. (14), (15), and (16) states the binary values utilized to indicate the vector components of LA as 1 or 0. For instance, if $a = 1$, the searching functions begin with binary-encoded lion.

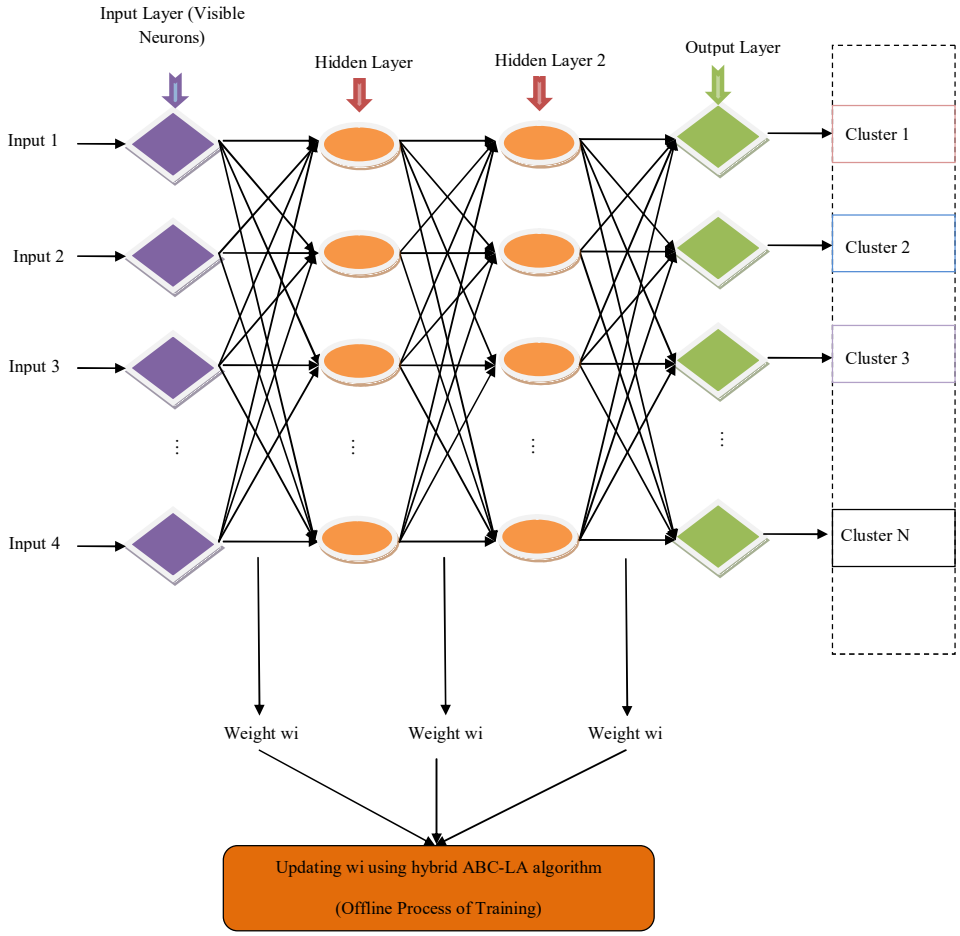


Figure 1 Geometrical Representation of Proposed Speaker Diarization Model

$$f(s_{xi}) \in \left(s_{xi}^{\min}, s_{xi}^{\max}\right) \quad (14)$$

$$n\%2 = 0 \quad (15)$$

$$f(s_{xi}) = \sum_{xi=1}^{Xi} s_{xi} 2^{\left(\frac{Xi}{2} - xi\right)} \quad (16)$$

Fertility Estimation: Typically, fertility estimation is performed to evade the local optimum issues. Consider S^{female} as fertile one to generate the cubs as expressed in Eq. (17) and (18) and the LA update is given in Eq. (10).

$$s_m^{female+} = \min\left[s_m^{\max}, \max\left(s_m^{\min}, \nabla_m\right)\right] \quad (17)$$

$$\nabla_m = \left[s_m^{female} + (0.1c_2 - 0.05)\left(s_m^{male} - c_1 s_m^{female}\right)\right] (18)$$

Mating: In mating, crossover and mutation are the 2 prime phases and gender clustering is an auxiliary phase. The lions S^{male} and S^{female} generates cubs to a maximum of 4 after mating (crossover and mutation).

LA Operators: Usually, the territorial defense is initiated to evade local optimum issues and assist to resolve various issues through same fitness which normally operates through creating nomad coalition, survival fight, pride, and nomad coalition updates in order. In addition, the nomad coalition $S^{e-nomad}$ is chosen using the Eq. (19)-(21).

$$fit\left(S^{e-nomad}\right) < fit\left(S^{male}\right) \quad (19)$$

$$fit\left(S^{e-nomad}\right) < fit\left(S^{b-cub}\right) \quad (20)$$

$$fit\left(S^{e-nomad}\right) < fit\left(S^{gc-cub}\right) \quad (21)$$

If $S^{e-nomad}$ is defeated to select S^{nomad}, then update nomad coalition. Moreover, S^{male} and S^{female} are updated if $\left(S^{b-cubs} and\, S^{gc-cubs}\right) > Ai^{\max}$ (Age Ai) in a territorial takeover.

Termination: Eventually, LA ended through accomplishing any 1 of the criteria given in Eq. (22) and (23).

$$iter > iter_{\max} \quad (22)$$

$$\left| fit\left(S^{male}\right) - fit\left(S^{optimal}\right) \right| \le er(ti) \quad (23)$$

5.2 Traditional ABC

Typically, ABC [27] model is inspired based on the searching behavior of honey bees, in which a group of bees (swarm) proficiently accomplish the solution through social cooperation. Basically in traditional ABC, 3 varieties of bees exist in the swarm-like (a) employed bees, (b) onlooker bees, and (c) scout bees. Normally, the prim e half of the pack in ABC approach contains employed bees and another half involves onlooker bees. Consider $S_u = \{s_1, s_2, \ldots, s_n\}$ indicates the n^{th} solution of group as well denotes the dimension size. In Eq. (24), all employed bees S_u create a novel candidate solution T_u on its neighborhood, in which, S_v refers to arbitrarily chosen candidate solution $u \neq v$, $\phi_{v,k}$ portrays the arbitrary value within $[-1,1]$, and k indicates arbitrary dimension index chosen in the set $1,2,\ldots,n$.

$$t_{u,k} = s_{u,k} + \phi_{u,k} \cdot \left(s_{u,k} - s_{v,k}\right) \quad (24)$$

Besides, when a novel candidate solution T_u is created, a greedy selection approach is exploited. Update S_u through T_u, while the fitness of T_u exceeds S_u, else sustain S_u as unchangeable. Eq. (25) specifies the food source picked by the onlooker bee through the nectar amount computed based on the nectar information gathered by each employed bees through the roulette wheel selection approach, in which, fit_u represents the fitness value of u^{th} solution in the pack. The update evaluation is as per Eq. (11), in which $r(0,1)$ specifies an arbitrary value within $[-1,1]$ through normal distribution, ab and bc represents the upper and lower boundaries of k^{th} dimension.

$$Pi_u = \frac{fit_u}{\sum_{k=1}^{nn} fit_k} \quad (25)$$

Termination: Eventually, LA ended through accomplishing any 1 of the criteria given in Eq. (22) and (23).

6. RESULTS AND DISCUSSIONS

6.1 Simulation Setup

The proposed speaker indexing or diarization model was implemented in MATLAB and the experimental investigation was carried out. The given input audio data was partitioned into five test cases such as Test case 1, Test case 2, Test case 3, Test case 4 and Test case 5. The efficiency of the proposed ANN-ABC-LA algorithm for speaker diarization model was compared with the traditional models such as Levenberg Marquardt (LM) [29], Firefly (LL) [30], Grey Wolf Optimization (GWO) [31], LA [26] and ABC [27] algorithms and obtained the results based on accuracy performance, diarization error, False Discovery Rate (FDR), False Negative Rate (FNR), and False Positive Rate (FPR), respectively.

6.2 Performance Analysis under Accuracy

In this section, the accuracy performance analysis of the proposed speaker diarization model is discussed over other existing models. Fig. 3 shows the accuracy performance of proposed ANN-ABC-LA model over the conventional models and the learning percentage ranges from 0 to 100% for Test case 1, Test case 2, Test case 3, Test case 4, and Test case 5. Fig. 3(a) shows the accuracy performance of Test case 1, which is 2.94%, 7.69%, 9.37%, 2.94%, and 7.69% better than LM, FF, GWO, ABC, and LA respectively at learning percentage 50. From Fig. 3(b), it is clear that the proposed ANN-ABC-LA model achieved better accuracy, which is 20.68% superior to LM, 25% superior to FF, 22.8% better than GWO, 40% superior to ABC and 16.66% better than LA at learning percentage 70.

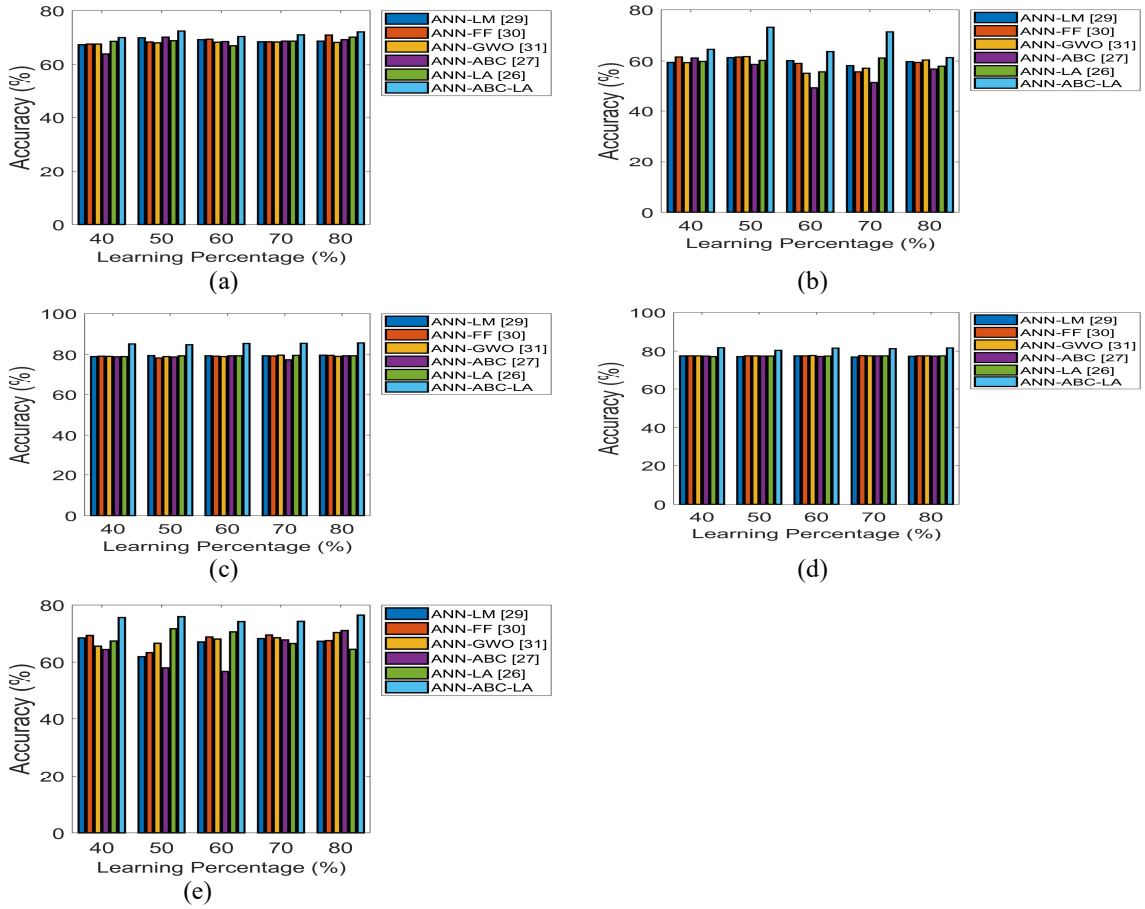
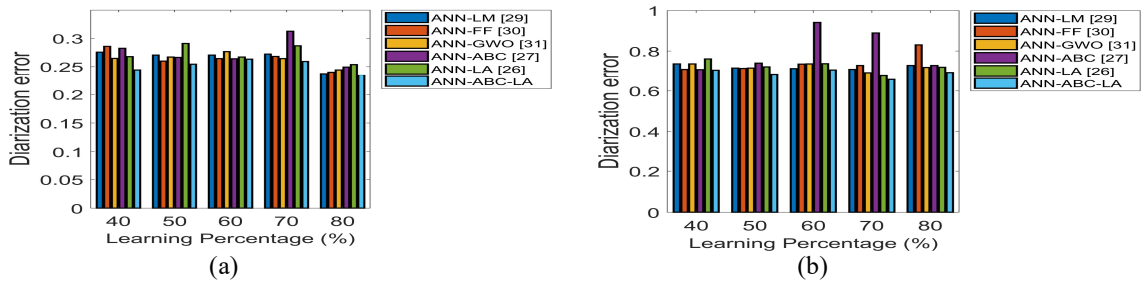


Figure 3 Accuracy Performance of Proposed ANN-ABC-LA Model over the Conventional Models for (a) Test case 1, (b) Test case 2, (c) Test case 3, (d) Test case 4, and (e) Test case 5

6.3 Diarization Error Analysis

Fig 4 depicts the diarization errors of five test cases. The proposed model attained minimized errors over the conventional models. From Fig. 4(c), the proposed model attained better results for Test case 3, which is 4.76%, 11.11%, 4.76%, 6.97%, and 6.85% better than LM, FF, GWO, ABC, and LA respectively at learning percentage 50. Through Fig. 4(d), the proposed ANN-ABC-LA model obtained minimized errors for Test case 4, which is 5.88% superior to LM, 11.11% superior to FF, 3.03% better than GWO, 20% superior to ABC and 3% better than LA at learning percentage 60.



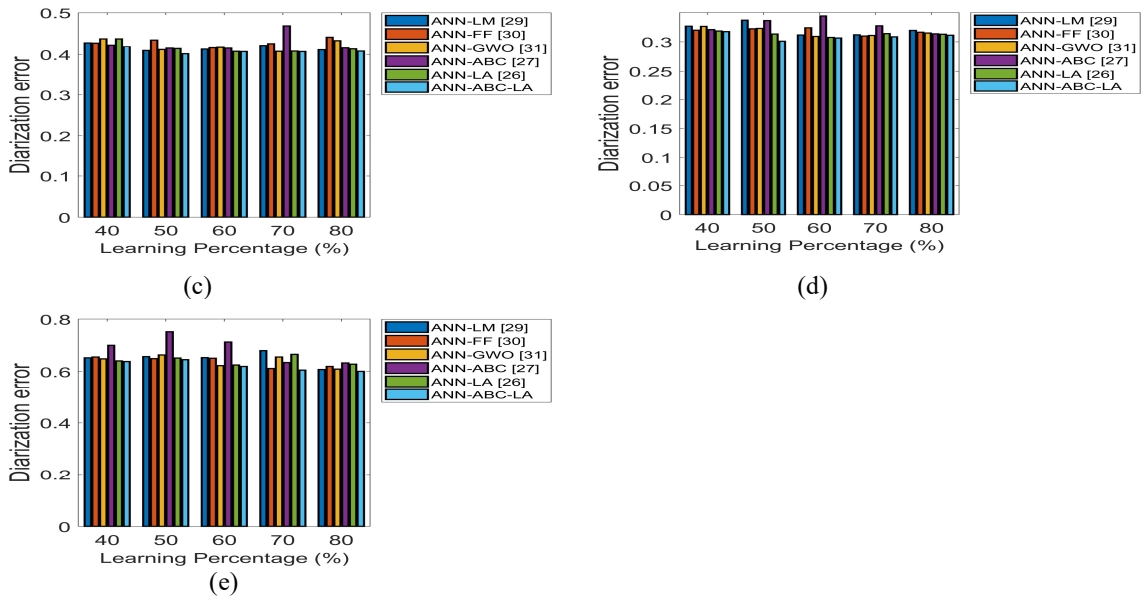


Figure 4 Error Analysis of proposed ANN-ABC-LA Model over the Conventional Models for (a) Test case 1, (b) Test case 2, (c) Test case 3, (d) Test case 4, and (e) Test case 5

6.4 Performance Analysis Over Conventional Model

The performance analysis of the conventional speaker diarization model is stated in this section. Table VIII tabulates the performance measures of the existing model and the diarization error. The accuracy performance of proposed ANN-ABC-LA is 42%, 40%, 59%, 67.53%, and 27.48% better than conventional method in terms of all Test cases respectively. The diarization error of proposed ANN-ABC-LA model attained minimized errors and improved results, which is 62%, 30.57%, 36.52%, 39.3, and 24.75% better than conventional model for al Test cases respectively. Thus, from the comparative study, it is clear that the proposed ANN-ABC-LA model attained better results and proved its efficiency.

Table 2: Overall Performance Analysis of Conventional MOdel

Performance Measures	Test case 1		Test case 2		Test case 3		Test case 4		Test case 5	
	Conventional Method [32]	Proposed ANN-ABC-LA	Conventional Method [32]	Proposed ANN-ABC-LA	Conventional Method [32]	Proposed ANN-ABC-LA	Conventional Method [32]	Proposed ANN-ABC-LA	Conventional Method [32]	Proposed ANN-ABC-LA
Accuracy	0.5	0.71204	0.5	0.70099	0.83333	0.88256	0.5	0.83767	0.6	0.76491
Sensitivity	0.4	0.71204	0.25	0.55125	0.66667	0.81759	0.5	0.77339	0.4	0.64516
Specificity	0.6	0.71204	0.75	0.77586	1	0.94724	0.5	0.90166	0.7	0.82479
Precision	0.5	0.71204	0.5	0.55151	1	0.93912	0.5	0.88674	0.4	0.64803
FPR	0.4	0.28796	0.25	0.22414	0	0.05276	0.5	0.09834	0.3	0.17521
FNR	0.6	0.28796	0.75	0.44875	0.33333	0.18241	0.5	0.22661	0.6	0.35484
NPV	0.6	0.71204	0.75	0.77586	1	0.94724	0.5	0.90166	0.7	0.82479
FDR	0.5	0.28796	0.5	0.44849	0	0.06088	0.5	0.11326	0.6	0.35197
Diarization Error	0.6	0.22391	1	0.69427	0.6667	0.42318	0.5	0.30346	0.8	0.60193

7. Conclusion

In this paper, a new speaker indexing or diarization model via Optimized ANN has been introduced. For this purpose, the proposed speaker diarization model accomplished the phases like speaker segmentation and clustering, for which the clustering process was performed via the proposed optimized ANN and the training was carried out through a new hybrid algorithm by choosing the optimal weight using a hybrid algorithm named proposed ANN-ABC-LA. The accuracy performance proposed ANN-ABC-LA for Test case 1, which was 2.94%, 7.69%, 9.37%, 2.94%, and 7.69% better than LM, FF, GWO, ABC, and LA respectively. The overall accuracy performance of proposed ANN-ABC-LA model attained 1.22 %, 5.57%, 0.88%, 1.13%, and 5.21% better than LM, FF, GWO, ABC, and LA respectively for Test case1. The accuracy performance of proposed ANN-ABC-LA was 42%, 40%, 59%, 67.53%, and 27.48% better than conventional method in terms of all Test cases respectively. Moreover, the diarization error of proposed ANN-ABC-LA model attained minimized errors and improved results, which is 62%, 30.57%, 36.52%, 39.3, and 24.75% better than conventional model for al Test cases respectively. Therefore, the performance of proposed ANN-ABC-LA model proved its betterment and efficiency over the conventional models and attained improved results.

References

[1] S. Cumani and P. Laface, "Analysis of Large-Scale SVM Training Algorithms for Language and Speaker Recognition," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 5, pp. 1585-1596, July 2012.

[2] Z. Liu, Z. Wu, T. Li, J. Li, and C. Shen, "GMM and CNN Hybrid Method for Short Utterance Speaker Recognition," IEEE Transactions on Industrial Informatics, vol. 14, no. 7, pp. 3244-3252, July 2018.

[3] Emma Jokinen, Rahim Saeidi, Tomi Kinnunen, and Paavo Alku, "Vocal effort compensation for MFCC feature extraction in a shouted versus normal speaker recognition task", Computer Speech & Language, vol. 53, pp 1-11, 2019.

[4] Mansour Alsulaiman, Awais Mahmood, and Ghulam Muhammad, "Speaker recognition based on Arabic phonemes", Speech Communication, vol. 86, pp 42-51, 2017.

[5] Omid Ghahabi, and Javier Hernando, "Restricted Boltzmann machines for vector representation of speech in speaker recognition", Computer Speech & Language, vol. 47, pp 16-29, 2018.

[6] Javier Franco-Pedroso, and Joaquin Gonzalez-Rodriguez, "Linguistically-constrained formant-based i-vectors for automatic speaker recognition", Speech Communication, vol. 76, pp 61-81, 2016.

[7] Chang Huai You, Haizhou Li, and Kong Aik Lee, "Relevance factor of maximum a posteriori adaptation for GMM–NAP–SVM in speaker and language recognition", Computer Speech & Language, vol. 30, no. 1, pp 116-134, March 2015.

[8] Abbas Khosravani, and Mohammad M. Homayounpour, "A PLDA approach for language and text independent speaker recognition", Computer Speech & Language, vol. 45, pp 457-474, September 2017.

[9] M. Sahidullah and G. Saha, "A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition," IEEE Signal Processing Letters, vol. 20, no. 2, pp. 149-152, Feb. 2013.

[10] F. Richardson, D. Reynolds and N. Dehak, "Deep Neural Network Approaches to Speaker and Language Recognition," IEEE Signal Processing Letters, vol. 22, no. 10, pp. 1671-1675, Oct. 2015.

[11] T. May, S. van de Par and A. Kohlrausch, "Noise-Robust Speaker Recognition Combining Missing Data Techniques and Universal Background Modeling," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 1, pp. 108-121, Jan. 2012.

[12] T. Stafylakis, P. Kenny, M. J. Alam and M. Kockmann, "Speaker and Channel Factors in Text-Dependent Speaker Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1, pp. 65-78, Jan. 2016.

[13] S. Cumani and P. Laface, "Speaker Recognition Using e–Vectors," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 26, no. 4, pp. 736-748, April 2018.

[14] L. Li, D. Wang, C. Zhang and T. F. Zheng, "Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 6, pp. 1129-1139, June 2016.

[15] Z. Tang, L. Li, D. Wang and R. Vipperla, "Collaborative Joint Training With Multitask Recurrent Model for Speech and Speaker Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, no. 3, pp. 493-504, March 2017.

[16] M. McLaren and D. van Leeuwen, "Source-Normalized LDA for Robust Speaker Recognition Using i-Vectors From Multiple Speech Sources," IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 3, pp. 755-766, March 2012.

[17] M. I. Mandasari, R. Saeidi, M. McLaren and D. A. van Leeuwen, "Quality Measure Functions for Calibration of Speaker Recognition Systems in Various Duration Conditions," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 11, pp. 2425-2438, Nov. 2013.

[18] M. Ferràs, S. Madikeri, P. Motlicek, S. Dey and H. Bourlard, "A Large-Scale Open-Source Acoustic Simulator for Speaker Recognition," IEEE Signal Processing Letters, vol. 23, no. 4, pp. 527-531, April 2016.

[19] X. Zhang, X. Zou, M. Sun, T. F. Zheng, C. Jia and Y. Wang, "Noise Robust Speaker Recognition Based on Adaptive Frame Weighting in GMM for i-Vector Extraction," IEEE Access, vol. 7, pp. 27874-27882, 2019.

[20] T. Stafylakis, M. J. Alam and P. Kenny, "Text-Dependent Speaker Recognition With Random Digit Strings," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 7, pp. 1194-1203, July 2016.

[21] Seyedali Mirjalili, Seyed Mohammad Mirjalili and Andrew Lewis, “Grey Wolf Optimizer”, Advances in Engineering Software, vol. 69, pp 46-61, 2014.

[22] Parthe Pandit, and Preeti Rao, "Speaker Diarization of Broadcast News Audios", July 2015.