A SURVEY ON MULTI SPEAKER DIARIZATION SYSTEM MODELS

[1] Sethuram . V, [2] Ande Professor

[1]Reseach Scholar, [2]Professor

[1], Department of Computer Science Vikrama Simhapuri university Nellore.
[2] Department of Computer Science,Vikrama Simhapuri university Nellore,

Abstract : The era of the digital technology uses recorded speech signals for the applications such as audio broadcasting, cell phones, and the television. The speech signals recorded from the conference hall and meetings through various speakers reduces the quality of the speech signals. The speaker tracking (diarization) system allows the efficient identification of the speaker from the recorded speech signal. In this work, three novel speaker diarization systems have been proposed for improving the speaker clustering and segmentation. In the first work, the Tangent weighted Mel-Frequency Cepstral Coefficient (TMFCC) with the Lion algorithm has been proposed. In the second work, the Multiple Kernel weighted Mel-Frequency Cepstral Coefficient (MKMFCC) with WLI-Fuzzy clustering has been proposed. The third work proposes Holo-entropy with the eXtended Linear Prediction using autocorrelation Snapshot (HXLPS) model with Deep Neural Network (DNN). The final proposed model HXLPS with DNN performance analysis is obtained by comparing it with the existing XLPS with DNN model. The proposed model implementation is done by recording the speech signal from the three speakers, four speakers, five speakers, six speakers and seven speakers system. The performance analysis is done by varying the frame length and the lambda (λ) values of the speech signal. The performance analysis of the proposed model is evaluated by using the various performance metrics such as tracking distance, tracking accuracy, precision, recall, F-Measure, false alarm rate and the diarization error rate. The simulation results show that the proposed TMFCC with Lion model has better performance than the existing models such as MFCC with ILP. The proposed MKMFCC with WLI Fuzzy has better average results than the TMFCC with Lion model. The HXLPS with DNN model outperforms the TMFCC with Lion and the MKMFCC with WLI Fuzzy models with better speaker identification. The proposed HXLPS with DNN model has the best overall performance with the average values of 0.63332, 0.886, 0.886, 0.87, 0.87, 0.20708, 0.08322 for parameters such as tracking distance, tracking accuracy, precision, recall, F-Measure, false alarm rate and the diarization error rate respectively for the various lambda (λ) values. For various frame length values, the proposed HXLPS with DNN model has the best overall performance rate with the average values of 0.57058, 0.918, 0.902, 0.8938, 0.87524, 0.17698, 0.0848 for the evaluation parameters tracking distance, tracking accuracy, precision, recall, F-Measure, false alarm rate and the diarization error rate respectively

IndexTerms **– Introduction, singe speaker diarization, diarization models, multi speaker diarization system model.**

I. INTRODUCTION

Speaker diarization is the process of automatically partitioning a conversation involving multiple speakers into homogeneous segments and grouping together all the segments that correspond to the same speaker as described in fig. 1. It is also called speaker tracking. Hence, speaker segmentation followed by speaker clustering is known as speaker diarization [1] [2] [3].

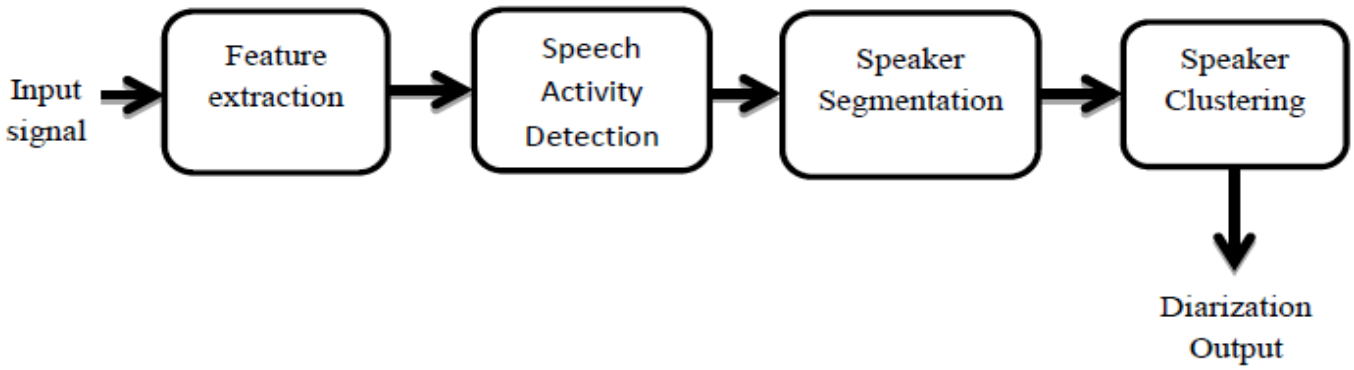

Fig. 1: A typical Speaker Diarization system

1.1. MOTIVATION

- The recognition performance of the spectral feature based speaker diarization system is affected by the limitations of the lower energy band values and stopping criteria for optimization.
- In order to resolve these issues, we propose weight based Spectral features with advanced clustering approaches.

II. RELATED WORK

SPEAKER DIARIZAION: APPROACHES

Speaker clustering is an importance task in the speaker diarization process as the final result (performance) of diarization depends greatly on the clustering process. The speaker diarization approaches are classified into five major categories based on the clustering algorithm used in the diarization process. The five major categories are: Partitioning based clustering approaches, Probabilistic model based approaches, Hierarchical clustering based approaches, Supervised learning based approaches and other approaches.

The *partitioning based* clustering approach generates a number of clusters from the given input signal [6] [7]. At the initial step, each speaker is assigned to any one of the clusters and an iterative relocation approach is employed to enhance the partitioning by moving the speakers from one cluster to another cluster. The basic constraint in partitioning based clustering is that at least one data object must be present in every cluster and every object should belong to only one cluster. The partitioning based clustering approaches for speaker diarization has several advantages such as scalability, simplicity, and suitability for datasets with compact clusters. And some of the drawbacks of partitioning based clustering are unsuitable for high dimensional problems, worse cluster descriptors, require prior specification about the number of cluster and so on.

In the *probabilistic model* based clustering, the data is assumed to be a sample obtained from a mixture model of various probability distributions. It has some significant characteristics such as it can be easily modified for handling complex problems, the intermediate mixture model can be used to assign cases at any stage of the iterative process and it results in cluster system which is easily interpretable. The major advantages of the probabilistic clustering are the interpretability of the intermediate clusters and the low-cost computation of intra-clusters measures of fit [8] [9].

In the *hierarchical* clustering approaches, the number of clusters is trained and the clusters are merged and they are reduced consecutively till one cluster for every speaker is obtained [10] [11]. Initially, the input audio signal is divided into number of segments and the closely matching clusters are grouped together, so that the number of clusters is reduced to one cluster for every speaker. Usually, GMM is used for modelling the clusters. After the clustering process, the frames are re-assigned to the clusters and the entire process repeats till the stopping criterion is reached. The stopping criteria include Bayesian Information Criterion (BIC), Kullback-Leibler (KL) based metrics and Generalized Likelihood Ratio (GLR). The Hierarchical clustering approaches has several advantages such as embedded flexibility, ease of handling and suitable for all kind of applications. But it has a major in the ambiguity of the stopping criterion.

Supervised model based clustering is based on the learning process, in which the training data consists of training example sets. In supervised speaker clustering, prior knowledge about the speaker is used in the clustering process [12] [13]. Using independent training data sets, the prior knowledge is encoded at several phases of the speaker clustering through learning the speaker discriminative acoustic feature information, learning universal speaker prior model and learning a speaker discriminative distance metric. Speaker diarization techniques based on supervised model based clustering includes Neural Networks, Linear Discriminant Analysis, Dynamic Bayesian Network based approaches.

The speaker diarization includes some other approaches such as beamforming based approaches, speaker diarization based on spatial information, Time Delay of Arrival (TDOA), Maximum a posterior (MAP) based approaches [14] [15].

III. EXISTING WORK

SPEAKER DIARIZATION SYSTEM USING TMFCC PARAMETERIZATION AND LION OPTIMIZATION

Primarily, the input to the speech diarization system is the audio signal that is to be segmented depending on the speaker identity. The audio signal is accepted as the input signal, which is then followed by pre-processing step. In the pre-processing step, the acoustic beam forming is done. The spatial filtering process makes the audio signal stable for the forthcoming diarization process, which helps in the speech detection. The acoustic beam forming helps in selecting the best audio signal from the multiple channels suitable for the diarization. Beam forming techniques are of two types such as fixed beam forming and adaptive beam forming techniques. The noise eliminating capability is poor in the fixed beam forming technique. The pre-processing step handles the sensitive errors caused by the channel delays. Once the pre-processing step is completed, the features are extracted from the audio signal.

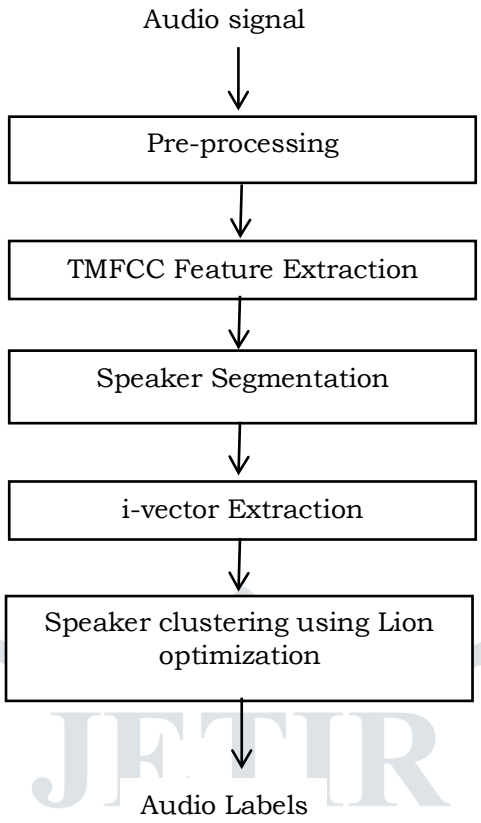

Fig. 2: Block diagram of the proposed TMFCC feature extraction and Lion optimization clustering

The features extracted from the acoustic signal gives information about the speakers signal. For extracting the features, it is obligatory to create Mel filter bank. Depending on the power spectrum of the input signal, the Mel filter bank is created. After creating the filter bank, the energy band of the filter bank is computed, and the weighted tangent function is added to get the Tangent weighted Mel filter cepstral coefficients (TMFCC). Mel Filter Cepstral (MFC) coefficients are the coefficients which jointly build up a Mel filter cepstrum. For the speaker diarization system, the 12 order TMFCC together with their first order derivatives are considered. The robustness and reliability of the TMFCC are increased by the fusion of the delta and delta-delta function.

The next step after feature extraction is the change of activity detection. Voice activity detection (VAD) is an important step in the diarization system. In general, maximum likelihood classification with the Gaussian Mixture models is used for voice activity detection. Activity change is detected, and the vectors are extracted related to the change in the signal. The Bayesian Inference criterion (BIC) is used to detect the change in the speech signal. The next step is to extract the corresponding feature i-vector from the change detected signal and is done with the consideration of the total variability in the space. The Expectation Maximization algorithm extracts the vector by employing the mean of the posterior probability. Then, Lion optimization algorithm is employed to cluster the audio signal based on the speaker identity. The optimized cluster yields the signal with separate labels corresponding to their identity.

A. TMFCC FEATURE EXTRACTION

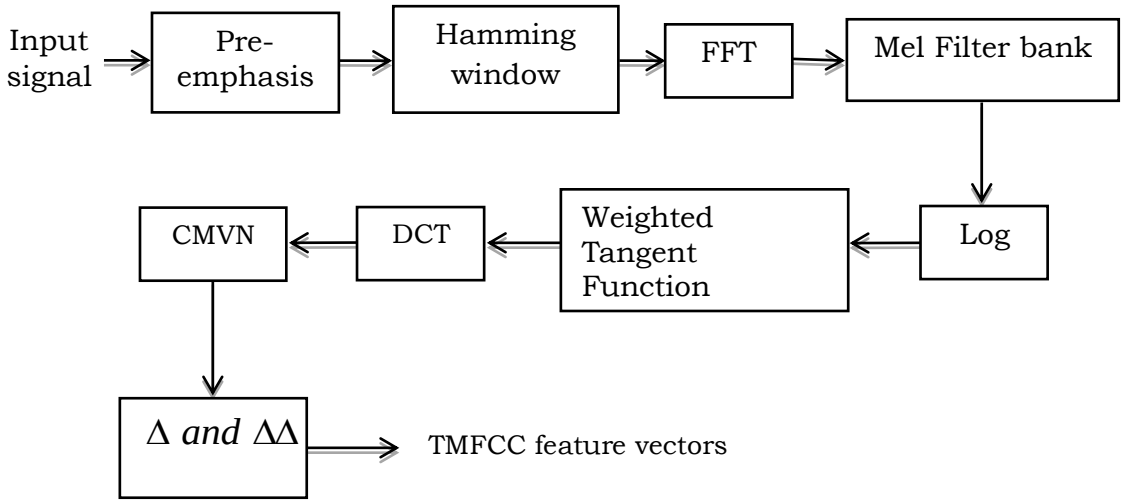

Fig. 3: Modular representation of the TMFCC feature extraction

The feature extraction steps for speaker diarization are presented below.

Step 1: Frame the signal into short frames (frame length 30ms and frame shift 10ms).

Step 2: For each frame calculate the periodogram estimate of the power spectrum.

Step 3: Apply the Mel filter bank to the power spectra, sum the energy in each filter. For the Mel filter bank creation, the low frequency (F_L) and the high frequency (F_H) components are selected from the power spectrum. Initially, the frequency is converted into Mel scale using the following formula:

$$M(f) = 1125 \ln\left(1 + {}^{f}\!/_{700}\right) \tag{1}$$

The Mel scale values are converted into the frequency components using the formula given below:

$$g(i) = (nfft+1) * h(i) / samplerate \tag{2}$$

Then using the values calculated above, the filter bank is created based on the following equation.

$$B_m(k) = \begin{cases} 0\ ; & k < g(m-1) \\ \dfrac{k - g(m-1)}{g(m) - g(m-1)}\ ; & g(m-1) \le k \le g(m) \\ \dfrac{g(m+1) - k}{g(m+1) - g(m)}\ ; & g(m) \le k \le g(m+1) \\ 0\ ; & k > g(m+1) \end{cases} \tag{3}$$

where $m = 1\ to\ M$ represents the number of filters and $g()$ is the $m+2$ Mel spaced frequencies.

Step 4: The filter bank energy is calculated using the tangent weighted function

$$E(i) = \sum_{k=0}^{\frac{N}{2}} \log |x(n)| B_m\left(K \frac{2\pi}{N}\right) * w_m \tag{4}$$

where, w_m is the proposed weighted tangent function and it is calculated as,

$$w_m = \tanh\left(\frac{-M}{2} + M \cdot \left[\frac{m-1}{M-1}\right]\right) \tag{5}$$

Step 5: The discrete cosine transform of the energy is calculated for the feature extracted coefficients, i.e. Tangent weighted Mel frequency cepstral coefficients.

$$E(i) = \tilde{E}(k) \tag{6}$$

where, $\tilde{E}(k) = \begin{cases} E(i) & ,k = k_i \\ 0 & ,Otherwise \end{cases}$

Step 6: The Cepstral Coefficient is calculated from the energy using the formula,

$$C_s(n) = \frac{1}{N'} \sum_{k=0}^{N'-1} Y(K)\, e^{jk\left(2\pi/N'\right)n} \quad (7)$$

Consequently, the TMFCC feature is extracted from the input audio signal by the consecutive steps. The Cepstral mean-variance normalization (CMVN) reduces the residual mismatch in the feature vectors. The addition of the delta and delta-delta coefficient with TMFCC enhance [16] the recognition accuracy, and it does not improve the robustness in the noise as well as the reverberation.

IV. PROPOSED WORK

The proposed research work has the following major contributions to implement in multi speaker diarization system models:

- The primary contribution of this research work is the proposal of the Tangent weighted Mel-Frequency Cepstral coefficient (TMFCC) feature parameterization using Lion optimization algorithm for speaker diarization system.
- The second contribution is proposing the Multiple Kernel Weighted Mel-Frequency Cepstral Coefficients (MKMFCC) features using WLI-Fuzzy Clustering algorithm for speaker diarization system.
- The third contribution is proposing the Holo-entropy with eXtended Linear Prediction using autocorrelation Snapshot (HXLPS) method using DNN for speaker diarization system.

V. PERFORMANCE EVALUATION OF EXISTING WORKS

Performance evaluation based on the Tracking distance

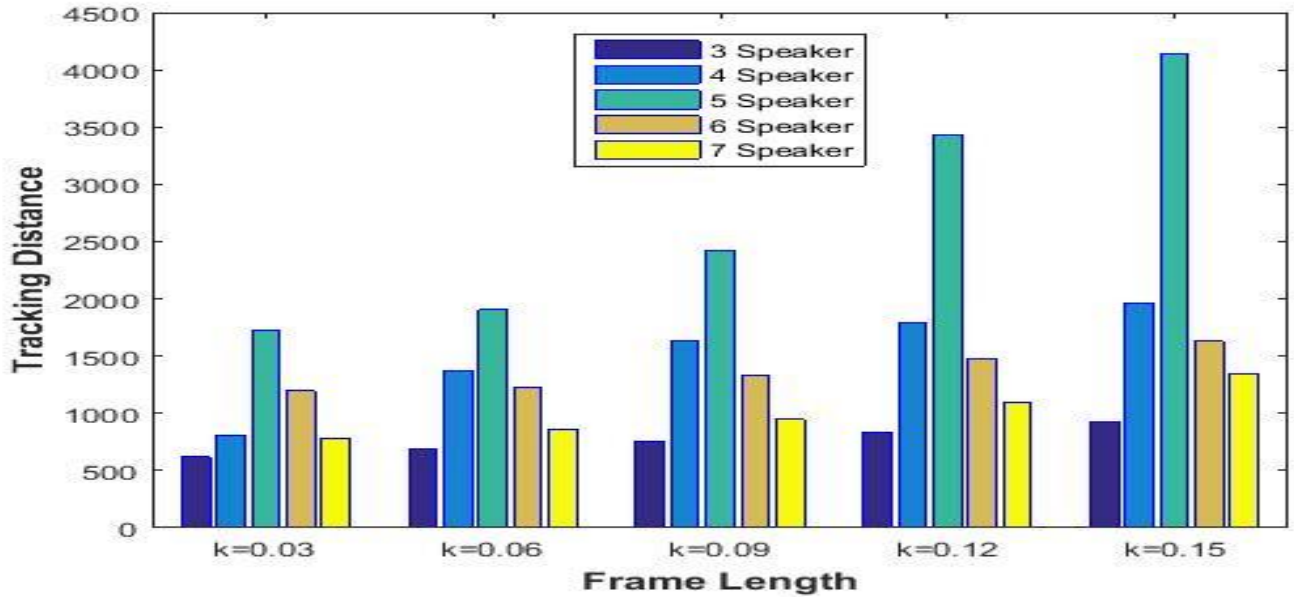

Fig. 1: Performance evaluation of the proposed HXLPS with DNN model based on the tracking distance for various speakers

Fig. 1 shows the performance evaluation of the proposed HXLPS with DNN algorithm based on the tracking distance with various speakers. The low value of tracking distance shows the model has better performance. For the system with the three speakers, the tracking distance of the proposed HXLPS with DNN model has the value of 618.4, 684.6, 756.5, 836.4, and 922.7 for the frame length values of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively. For the system with the four speakers, the tracking distance of the proposed HXLPS with DNN model has the value of 808, 1369, 1637, 1794, and 1962 for the frame length values of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively. For the system with the seven speakers, the tracking distance of the proposed HXLPS with DNN model has the value of 784.7, 863.1, 949.4, 1091.6, and 1351.2 for the frame length values of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively.

Performance evaluation based on the Tracking accuracy

Fig. 2 shows the performance analysis of the proposed HXLPS with DNN model based on the tracking accuracy. The analysis was done for the varying values of the frame length on the system with 3, 4, 5, 6, and seven speakers. For the system with the three speakers, the proposed with HXLPS with DNN model has the tracking accuracy of 0.95, 0.95, 0.9003, 0.87, and 0.87 for the frame length value of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively. For the system with the five speakers, the proposed with HXLPS with DNN model has the tracking accuracy of 0.95, 0.87, 0.87, 0.87, and 0.87 for the frame length value of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively. For the system with the six speakers, the proposed with HXLPS with DNN model has the tracking accuracy of 0.95, 0.8647, 0.87, 0.87, and 0.87 for the frame length value of 0.03, 0.06, 0.09, 0.12, and 0.15 respectively

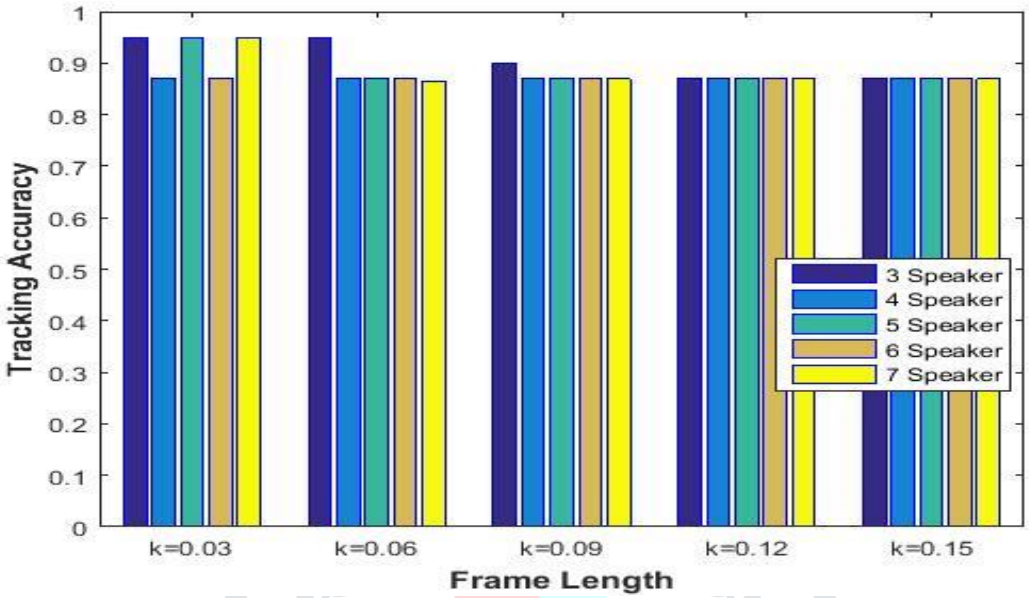

Fig. 2: Performance evaluation of the proposed HXLPS with DNN model based on the tracking accuracy for various speakers

VI CONCLUSION

In this research work, a novel method to enhance the speaker diarization system performance has been proposed. This work introduces three novel approaches such as TMFCC with Lion, MKMFCC with WLI-FUZZY, and the HXLPS with DNN for efficient speaker identification from the recorded speech signal. The audio signals obtained from the system with three speakers, four speakers, five speakers, six speakers and seven speakers are used for the performance analysis. The performance analysis is done by varying the frame length and the wavelength (lambda) values of the speech signal. The performance of the proposed models such as TMFCC with Lion, MKMFCC with WLI-FUZZY, and the HXLPS with DNN have been analyzed by comparing with the existing models such as MFCC with ILP, MFCC with Lion, TMFCC with ILP, and the XLPS with DNN models. The performance metrics such as tracking distance, tracking accuracy, precision, recall, F-measure, false alarm rate and the diarization error rate analyze the performance of the proposed model. The simulation results show that the performance of the proposed TMFCC with Lion model is better than the MFCC with ILP, MFCC with Lion, and TMFCC with ILP. The proposed MKMFCC with WLI-FUZZY model has improved performance than the proposed TMFCC with Lion model. The final proposed model HXLPS with DNN performance analysis is obtained by comparing it with the existing XLPS with DNN model. The proposed HXLPS with DNN model has the best overall performance with the average values of 0.63332, 0.886, 0.886, 0.87, 0.87, 0.20708, 0.08322 for the parameter of tracking distance, tracking accuracy, precision, recall, F-measure, false alarm rate and the diarization error rate respectively for the various lambda values. For the various frame length values, the proposed HXLPS with DNN model has the best overall performance with the average values of 0.57058, 0.918, 0.902, 0.8938, 0.87524, 0.17698, 0.0848 for the parameter of tracking distance, tracking accuracy, precision, recall, F-measure, false alarm rate and the diarization error rate respectively. The proposed HXLPS with DNN model has better performance when the frame length gets varied. The proposed model produces an improved DER when the lambda and the frame length values are small.

ACKNOWLEDGMENT

Here conveying our sincere thanks to all the authors those efforts fulfilled for this survey paper.

REFERENCES

[1]. Xavier AngueraMiro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, and OriolVinyals, "Speaker Diarization: A Review of Recent Research", In Proceedings of IEEE transaction on Audio, speech and language processing, 20(2): 356-370, 2012.

[2]. S. E. Tranter and D. A. Reynolds, "An overview of automatic speaker diarization systems", IEEE Trans. Audio, Speech, Language Processing, vol. 14(5), pp. 1557–1565, Sept. 2006.

[3]. Margarita Kotti, VassilikiMoschou, and Constantine Kotropoulas, "Speaker segmentation and clustering", Signal Processing, vol. 88, pp. 1091–1124, Dec. 2007.

[4]. M.H. Moattar, M.M. Homayounpour, "A reveiw on speaker diarization systems and approaches", Speech communications", 54(10):1065-1103, 2012.

[5]. H. Sayoud, S. Ouamour and S. Khennouf, "Virtual System of Speaker Tracking by Camera Using an Audio-Based Source Localization", In Proceedings of World Congress on Engineering. 2012.

[6]. Liping Zhu, "A modified approach to cluster refinement for speaker diarization", 4th International Conference on Computer Science and Network Technology (ICCSNT), 2015.

[7]. Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-Francois Bonastre and Laurent Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization", Computer Speech & Language, vol. 20, no. 3, pp. 303-330, 2006.

[8]. ThemosStafylakis, VassilisKatsouros, and George Carayannis, "The Segmental Bayesian Information Criterion and Its Applications to Speaker Diarization", IEEE Journal of Selected Topics in Signal Processing, 4(5), pp. 857 – 866, 2010.

[9]. Hanwu Sun, Tin LayNwe, Bin Ma, and Haizhou Li, "Speaker Diarization for Meeting Room Audio", Interspeech, 2009.

[10]. Kyu J. Han and 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.

[11]. S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, "An integrated top-down/bottom-up approach to speaker diarization", in Proc. Interspeech, 2010.

[12]. S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using auto-associative neural networks", Engineering Applications of Artificial Intelligence, vol.22, pp. 667-675, 2009.

[13]. BelkacemFergani, Manuel Davy, and AmraneHouacine, "Speaker diarization using one-class support vector machines", Speech Communication, 50, 355–365, 2008.

[14]. D. Vijayasenan, F. Valente, and H. Bourlard, "An Information Theoretic Combination of MFCC and TDOA Features for Speaker Diarization", IEEE Transaction on Audio, Speech, and Language Processing, 19, 431-438, 2011.

[15]. D. Vijayasenan and F. Valente, "Speaker Diarization Of Meetings Based On Large TDOA Feature Vectors", In Proceedings of IEEE International Conference on Audio, Speech and Signal Processing, 2012.

[16]. Kshitiz Kumar, Chanwoo Kim and Richard M. Stern, "Delta-spectral cepstral Coefficients for robust speech recognition",In Proceedings of ICASSP,pp. 4784-4787,2011.

[17]. Claude Barras, Xuan Zhu, Sylvain Meignier, and Jean-Luc Gauvain, "Multistage Speaker Diarization of Broadcast News", In Proceedings of IEEE Transactions on Audio, Speech and language processing, 14(5): 1505-1512, 2006.

[18]. P. Peeling, A. T. Cemgil, S. Godsill, "Bayesian hierarchical models and inference for musical audio processing", In Proceedings of IEEE wireless Pervasive computing, pp. 278-282, 2008.

[19]. P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, "Deep Neural Networks for extracting Baum Welch statistics for Speaker Recognition", In Proceedings of the speaker and language recognition. pp. 293- 298, 2014.

[20]. P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet and J. Alam, "Deep neural networks for Baum-Welch statistics for speaker Recognition", In proceedings of Neural Networks for speaker and language Modelling, 2014.