Performance Analysis of Feature sets in Speaker Diarization techniques

C Sailaja¹, Suman Maloji², Kasiprasad Manepalli³
Research scholar, Department of ECE, KLEF(Deemed to be University), Vijayawada, Andhrapradesh, India
HoD, ECE Department, KL University, KLEF, Vijayawada, Andhrapradesh, India
Associate Professor, ECE Department, KL University, KLEF, Vijayawada, Andhrapradesh, India

sailaja.cheruvupalli@gmail.com

Abstract: Speech is the most important communication among humans. Processing of speech signal has many strategies including speech coding, speaker recognition, speaker verification, etc. Speaker diarization is the pre-processing stage for many applications of speaker recognition systems. Speaker Diarization is the mission of determining “who Spoke when” for any audio recording that carries an unknown quantity of records and an unknown variety of audio systems. Speaker diarization has come to be an era for many tasks like navigation, retrieval, or higher-level interference on audio data. It mainly performs three operations feature extraction, voice activity detection, and classification. In this paper, we’ve reviewed the few speaker diarization Techniques. The trendy speaker diarization structures finished nice outcomes. In this paper, few speaker diarization device performances are evaluated for Diarization mistakes, Tracking time, and False alarm.

Keywords: Speaker Diarization, MFCC, Multi-kernel MFCC, Tangent weighted MFCC, XLPS, HXLPS.

1. Introduction:
Nowadays, speech processing technology is an important technology in searching for or automated indexing and retrieval of records [11]. For this application extraction of metadata is required, this can be done by several technologies. Speaker diarization is one of the emerging technologies which consist of most important stages segmentation and clustering [18]. In segmentation first task is the extraction of features from the speech signal and indexing them. Indexing of speech signal in its simple form indexing speech, music, or no speech data [12]. In its complex form, indexing the speaker changes is involved.

The main goal of a good speaker diarization system is selecting a better feature extraction technique to improve the accuracy of the speaker diarization system [13]. Theoretically, it needs to be feasible to apprehend speech without delay from the digitized waveform [16]. However, due to the large variability of the speech signal, its miles better to carry out some features extraction that might lessen that variability.

Speaker Diarization is finding lot of applications in several signal processing applications where more than one speaker is involved. Speaker diarization is the pre-processing stage of any speaker
recognition system. Speaker diarization system consists of feature extraction, voice activity detection, segmentation, and clustering [14]. Clustering is the main task of the diarization system. In clustering there are two types of techniques, top-down approach and bottom-up approach [2].

In Bottom up approach, the quantity of methods is reduced by merging clusters in each iteration until it reaches the optimum quantity of models. In Top down approach, the number of models is increased at each iteration by cluster splitting method. Among these two methods, top down approach is less popular than bottom up approach [3]. Both methods are opposing each other. Speaker diarization system consist two main sections, Feature extraction and Classification.

2. Speaker Diarization Techniques:

Speaker diarization is a technique that deals with “who speaks when.” The process involved in speaker diarization is shown in Fig. 1. The important blocks are as follows, Feature extraction, Speech activity detection, Segmentation, and clustering.

The selection of good feature extraction techniques is the fundamental task in any of the speech processing applications. The speech signal is a largely varying signal, to process any speech signal, it is better to extract the features from the speech signal. The extraction of features from any speech signal gives the benefit of reducing the processing complexity by removing the unvoiced part from the signal.

![Fig.1. Speaker Diarization Method](image)

Voice activity detection or Speech activity Detection is an important block in speech processing to detect speech and silence detection and avoid the unwanted part from the signal. In Segmentation, there are two types of segmentation methods: Blind method and Aided segmentation [4]. Clustering is one of the key blocks in speaker diarization where a grouping of the same speaker features is clustered.

The challenges faced in speaker diarization are mainly two. First is optimizing the speaker sequence as the total number of speakers is unknown. Secondly clustering, although we extract features from the input speech signal it also depends on several other factors [3].

The following speaker diarization techniques are discussed in this paper,

2.1. Mel Frequency Cepstral Coefficients (MFCC):

To extract acoustic features from the given speech signal the method used most widely is MFCC. Using the extracted features speech signals are segmented for different speakers [17]. The features of input signal are extracted through MFCC algorithm and clustered using Integer Linear Programming (ILP) [5] based totally clustering mechanism and Lionalgorithm.

In this algorithm, Hilbert envelope of LP residual is considered and cross-correlation between Hilbert envelope and multisource preprocessed HE of LP residual is calculated as follows,
Hilbert envelope of LP residual \( H[n] = \sqrt{e^2(n) + e_k^2(n)} \)

Preprocessed HE is, \( G_i(n) = \frac{h_i^2[n]}{\sum_{m=n-M}^{n+M} h_i^2[m]} \)

The cross-correlation function is

\[
R_{12}[l] = \sum_{n=-N/2}^{N/2} g_1(n)g_2(n-l) \sum_{n=-N/2}^{N/2} g_1^2(n)[ \sum_{n=-N/2}^{N/2} g_2^2(n-l)]^2
\]

2.2. **Tangent weighted MFCC (TMFCC):**

In this algorithm, the MFCC features are integrated with the tangent weighted function to develop a feature extraction technique for speaker diarization system. Thus, the feature of input signal are extracted using TMFCC algorithm and clustered using ILP [5] clustering mechanism and Lion algorithm[7].

Filter bank energy is calculated by “tangent weighted function” as

\[
E(i) = \sum_{k=0}^{N/2} \log|x(n)| Bm(k, \frac{2N}{N}) Wm
\]

Where

\[
Wm = \tanh(-M + M \left[ \frac{m-1}{M-1} \right]) \]

The cepstral coefficients are calculated using,

\[
C_i(n) = \frac{1}{N} \sum_{k=0}^{N-1} Y(k) e^{j(2\pi)n}
\]

2.3. **MKMFCC with WLI Fuzzy:**

In this method, multi kernel MFCC algorithm is used to extract features of input signal and after VAD segmentation for clustering the signals, WLI Fuzzy clustering [8] algorithm is used. The below equation is proposed in [10] to cluster the same speaker.

\[
WLI(k) = \sum_{n=0}^{N} Y(n) e^{j(2\pi)n}
\]

2.4. **XLPS with DNN:**

The “eXtended Linear Prediction (XLPS)” algorithm is used to extract the features of input signal after VAD segmentation, DNN technique is used for grouping of same speaker segments.

\[
E_{XLPS} = X_n \sum_{k=1}^{N} (C_k S_{n+k} Z_{n+k})^2
\]

2.5. **HXLPS with DNN:**

The Holoentrophy extended LP with auto-correlation snapshot is designed by integrating the Holo-entropy function into extended LPS method [10]. In this algorithm features are obtained by hybrid feature extraction technique. The VAD is used to segment the signal then the results are used in collecting the i-vector representation of the signal. These i-vectors are input to the DNN network, where same speaker grouping is obtained.

\[
E_{HXLPS} = \sum_{k=1}^{P} (X_k Y_j, 0 - \sum_{k=1}^{N} \alpha_k (x_k p) X_j - k Y_j, k)^2
\]

3. **Performance Measurements:**

For the performance evaluation of the diarization systems, “English Language Speech Database for Speaker Recognition (ELSDSR)”[15] and Catalan 3/24 database is used to provide audio signals for the evaluation of the speaker diarization system.

3.1. **Tracking distance:**

The tracking distance is the measure of difference between original and speaker output signals. The tracking distance is measured to analyze the efficiency of Speaker diarization
From Tracking distance we can conclude the efficiency of speaker diarization system as if the value of tracking distance is less then performance of the system is better. The Tracking distance is measured using the below equation,

$$\text{Tracking Distance} = \sqrt{(X_{i}^{\text{o}} - X_{i}^{\text{g}})^2}$$

Where,

- $X_{i}^{\text{o}}$ = The speaker output signal
- $X_{i}^{\text{g}}$ = The original Signal

### 3.2. Diarization error rate (DER):

To compare the diarization systems the most widely used metric is Diarization error rate. It is defined as the ratio between sum of false alarm, missed detection and confusion to total duration of speech. DER can be calculated using,

$$\text{DER} = \frac{\text{confusion error} + \text{miss error} + \text{false alarm}}{\text{total response speech time}}$$

### 3.3. False Alarm rate:

It is the defined as the duration of non speech where it is recognized as speech signal incorrectly. This can be calculated as the number of false alarms in total number of warnings.

### 3.4. Missed alarm:

It is the duration of speech incorrectly detected as non speech. This can be calculated as the number missed speech part in total duration of speech.

### 3.5. Confusion:

It is the duration of speaker confusion.

### 4. Results and discussions:

![Tracking distance comparison graph](image)

The above fig.2 shows the graph between Tracking distance and frame length. When the frame length is 0.06, the diarization algorithms MFCC is showing 1.79, TMFCC is showing 1.47, MKMFCC-WLIFUZZY is showing 1.30, XLP-DNN is showing 1.19, and HXLP-DNN is showing 0.61.
Fig. 3 Diarization error rate
The above plot shows the graph between Diarization error rate and lambda. When Lambda is 8, MFCC is 0.49, TMFCC is 0.40, MKMFCC-WLIFUZZY is 0.24, XLP-DNN is 0.11, HXLP-DNN is 0.12.

Fig. 4 False Alarm Rate
The above figure shows the plot between False alarm rate with Frame Length. Frame length from 0.02 to 0.16 is considered for evaluating different diarization systems like MFCC (Mel Frequency Cepstral Coefficients), TMFCC (Tangent-Mel Frequency Cepstral Coefficients), MKMFCC-WLIFUZZY (Multi Kernal MFCC and WliFuzzy clustering), XLP-DNN (Extended Linear prediction with DNN), HXLP-DNN (Holoentrophy eXtended Linear Prediction with DNN).
When the frame length is six, MFCC is 41.3%, TMFCC is 23.4%, MKMFCC-WLIFUZZY is 27.8%, XLP-DNN is 15.8%, HXLP-DNN is 9.5%. Among these all algorithms Holoentrophy with eXtended Linear Prediction is showing lowest false alarm rate.

Comparative Analysis:

The below table Table. 1 gives a comparative analysis of speaker diarization systems with three comparison parameters Tracking distance, Diarization error rate and False alarm.
| Parameters                        | MFCC | TMFCC | MKMFCC | WLIFUZZY | XLP-DNN |
|----------------------------------|------|-------|--------|----------|--------|
| Tracking Distance @Framelength   | 1.79 | 1.47  | 1.30   | 1.19     | 0.61   |
| 0.06                             |      |       |        |          |        |
| Diarization error rate @ Lamda 8 | 0.49 | 0.40  | 0.24   | 0.11     | 0.12   |
| False alarm rate @Framelength 6  | 41.3%| 23.4% | 27.8%  | 15.8%    | 9.5%   |

Table 1. Comparison between the Algorithms

5. Conclusion:
   In this paper we analysed different diarization systems and compared their performance by means of different parameters such as diarization error rate, tracking distance, and false alarm. From the above analysis, we can conclude that, HXLP feature extraction with DNN clustering is showing good results compared with the remaining algorithms.

6. References:
[1] S. Traner and D. Reynolds, ”An overview of automatic speaker diarization systems,” IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 5, pp. 1557-1565, Sep. 2006.
[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, ”Speaker diarization: A review of recent research,” IEEE Trans. Audio, Speech, Lang. Process. Special Iss. New Frontiers in Rich Transcript., 2012.
[3] Nicholas Evans, Dong Wang and Raphael Troncy, ”A Comparative Study of Bottom-Up and Top-Down Approaches to Speaker Diarization,” vol. 20, No. 2, Feb 2012.
[4] Md. Mijanur Rahman and Md. Al-Amin Bhuiyan, ”Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches,” International Journal of Advanced Computer Science and Applications, Vol. 3, No. 11, 2012.
[5] R. Kumara swamy, k. Sri ramamurty, and b. Yegnanarayana, ” determining number of speakers from multispeaker speech signals using excitation source information”, ieee signal processing letters, vol. 14, no. 7, July 2007 pgno.481-484
[6] Subba Ramaiah, V., and R. Rajeswara Rao, ”A Novel Approach For Speaker Diarization System Using TMFCC Parameterization And Lion Optimization”. Journal Of Central South University, vol 24, no. 11, 2017, pp. 2649-2663.
[7] B.R. Rajkumar,” The Lion’s Algorithm: A New Nature-Inspired Search Algorithm”, Procedia 6 (2012) 126-135.
[8] Chih-Hung Wu, Chen-Sen Ouyang, Li-Wen Chen, And Li-Wei Lu, “A New Fuzzy Clustering Validity Index With A Median Factor For Centroid-Based Clustering”, IEEE Transactions On Fuzzy Systems 2013.
[9] Jouni Pohjalainen , Rahim Saeidi , Tomi Kinnunen and Paavo Alku, “Extended Weighted Linear Prediction (XLP) Analysis of Speech and its Application to Speaker Verification in Adverse Conditions “, INTERSPEECH (2010).
[10] Zelenák, Martin, Henrik Schulz, and Javier Hernando. "Speaker diarization of broadcast news in Albayzin 2010 evaluation campaign." EURASIP Journal on Audio, Speech, and Music Processing 2012, no. 1 (2012).
[11] Alluri K.N.R.K.R., Achanta S., Prasath R., Gangashetty S.V., Vuppala A.K. (2017) A Study on Text-Independent Speaker Recognition Systems in Emotional Conditions Using Different
Pattern Recognition Models. In: Prasath R., Gelbukh A. (eds) Mining Intelligence and Knowledge Exploration. MIKE 2016. Lecture Notes in Computer Science, vol 10

[12] 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).

[13] Mathieu Hu, Dhusyanth Sharma, Simon Doclo, Mike Brookes, Patrik A. Naylor, Speaker change Detection and Speaker Diarization using spatial information, Proc., IEEE Int. conf. Acoust. Speech signal processing. (2015).

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

[15] The ELSDR dataset for speaker diarization http://www2.imm.dtu.dk/~lfen/elsdr

[16] Kasiprasad Mannepalli, Panyam Narahari Sastry and Maloji Suman,” Analysis of Emotion Recognition System for Telugu Using Prosodic and Formant Features” In Advances in Intelligent Systems and Computing (Vol. 664, pp. 137–144). Springer Verlag.(2018).

[17] Mannepalli, K., Sastry, P. N., & Suman, M. (2016). MFCC-GMM based accent recognition system for Telugu speech signals. International Journal of Speech Technology, 19(1), 87–93.

[18] Mannepalli, K., Sastry, P. N., & Suman, M. (2017). “Accent recognition system using deep belief networks for telugu speech signals”, In Advances in Intelligent Systems and computing (Vol. 515, pp. 99–105). Springer Verlag.