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