ISSN (Online): 2456-0448
International Journal Of Innovative Research In Management, Engineering And Technology
Vol. 1, Issue 5, June 2016

SPEAKER DIARIZATION USING DEEP LEARNING AND HMM SPEAKER MODELS

[1] Archana S M, [2] Amal Babu, [3] Raseek C
[1] [2]Dept of Computer Science and Engineering,Govt. Engineering College Sreekrishnapuram, Palakkad, Kerala, India
[3]Asst. Prof. Dept. Computer Science and Engineering. Govt. Engineering College Sreekrishnapuram,Palakkad, Kerala, India
[1]1111archa@gmail.com [2]amalbabuputhanpurayil@gmail.com [3] raseek.c@gmail.com

Abstract: ***Speaker diarization finds continuous speaker segments in an audio stream and clusters them by speaker identity. In this paper we propose a method for Speaker diarization by using a new area of machine learning, i.e Deep learning. For Speaker segmentation we trained one of the Deep Learning Network by short-term spectral features to predict given speech segments belongs to same or different speaker and for Speaker clustering HMM speaker models has been used for speaker recognition from the given set of speakers.***

Keywords: **speaker diarization; speaker segmentation, speaker clustering; deep learning; Hidden markov model (HMM).**

I. INTRODUCTION

The task of speaker diarization is to determine 'Who spoke when' in a speech signal/Audio Wave. This information can be useful for various applications, such as annotating and indexing multimedia data, carrying out speaker recognition of multiple speaker recordings or identifying speaker-specific speech events for improving speech recognition applications [1]. Actually the annotating "who spoke when" is performed without any prior information: neither the number of speakers, nor the identities of the speakers, nor samples of their voices are needed. The labels representing the speakers are automatically generated without matching the real identities of the speakers.
A common approach for this task consists in detecting homogeneous audio segments, in which each contain the voice of only one speaker. Then the generated segments are grouped into clusters, one speaker at a time.
Audio segmentation, in general, is the task of segmenting a continuous audio stream in terms of acoustically homogeneous regions, where the rule of homogeneity depends on the task. Here the tasks related to Speech Segmentation are Speaker Segmentation and Speaker clustering. Speaker segmentation will detect homogeneous audio segments which contain only one speaker and Speaker clustering will group those segments according to speaker identities.

II. AUTOMATED SPEAKER DIARIZATION

The basic speaker diarization process generally comprises of three main components, namely speech activity detection, speaker segmentation and speaker clustering.

A. Speech Activity Detection

The first stage of automated speaker diarization is also known as end-point detection or voice activity detection (VAD or SAD). It is the process of classifying the audio into speech and non-speech regions. VAD is considered as a preprocessing step in many speech processing applications, including speaker diarization, speaker verification and speech recognition. This stage is very important because of identifying and discarding of non-speech regions such as music and noise in the diarization process, will avoid hindering subsequent speaker segmentation and clustering processes. So it will remove only prolonged periods of music or noise, rather than targeting short speaker pauses in the middle of speaker turns and thus breaking up homogeneous speaker segments.

B. Speaker Segmentation

Speaker segmentation deals with partitioning the speech signal according to speaker identities. So that after this stage we will get all speaker change points within a speech. The speaker segmentation stage ideally produces pure, homogeneous speaker segments containing one speaker each.

C. Speaker Clustering

At this last step of speaker diarization, it will cluster the speech segments belonging to the same speaker together. So at the last, one cluster will be having for each speaker.
Most of present state-of-the-art speaker diarization systems fit into one of two categories: the bottom-up and the top-down approaches. The top-down approach is initialized with very few clusters (usually one) whereas the bottom-up approach is initialized with many clusters (usually more clusters than expected speakers). In both cases the aim is to iteratively converge towards an optimum number of clusters [2].

The bottom-up approach is most common in the literature. Also known as agglomerative hierarchical clustering (AHC or AGHC), the bottom-up approach trains a number of clusters or models and aims at successively merging and reducing the number of clusters until only one remains for each speaker.
In all cases the audio stream is initially over-segmented into a number of segments which exceeds the anticipated maximum number of speakers. The bottom-up approach then iteratively selects closely matching clusters to merge, hence the number of clusters in each iteration is reduced by one.

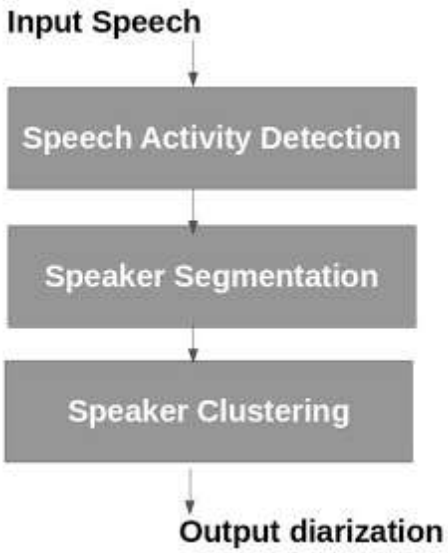

Fig. 1. Automated Speaker Diarization

Clusters are generally modeled with a GMM and, upon merging, a single new GMM is trained on the data that was previously assigned to the two individual clusters. Standard distance metrics, such as BIC, CLR, GLR are used to identify the closest clusters [2].

III. PROPOSED METHOD

Here speaker diarization is doing on the dataset where speaker identities are known. In proposed method Speaker Segmentation can be done using Deep Neural Networks and Speaker clustering done by identifying speaker identities from HMM models.

Speech Activity Detection or Voice Activity Detection is performed by Energy based detector from pyAudioanalysis python library. Noise has been removed from each speech signal by Audacity tool then given to VAD. Actually VAD in pyAudioAnalysis is a semi-supervised approach where SVM model is trained to distinguish lower and higher energy frames. For each signal SVM applied and dynamic threshold used to detect the active segments. Then the non-speech regions discarded from the original speech signal and then given to the following stages of speaker diarization.

Deep neural network architecture is trained to perform next tasks related to speaker diarization, i.e. speaker segmentation. Deep Belief Networks (DBNs) are neural networks consisting of a stack of Restricted Boltzmann Machine (RBM) layers. Here, a deep belief network is trained for classifying two given speech segments whether belongs to the same or different speakers. A deep belief network composed of multiple layers of hidden units, with connections between the layers but not between units within each layer [6]. When trained on a set of examples in unsupervised way, a DBN can learn to probabilistically reconstruct its inputs. The layers then act as feature detectors on inputs. After this learning step, a DBN can be further trained in a supervised way to perform classification.

Speech signal after VAD splits into frames of 25ms with a step of 10ms. And for each frame 13 MFCC features has been extracted. Mel-frequency Cepstral Coefficients (MFCC) is a widely used feature in Automatic Speech and Speaker recognition. These features have been given to the network architecture as feature vectors. Then the deep belief network has been trained with four hidden layers to predict the given two speech frames belongs to same speaker or not. It results a model of having best F1 score.

At the core of every speaker diarization and linking approach there is a speaker modeling and clustering stage that carries out a major role in identifying and clustering spoken segments from the same speaker identities [2, 3].

HTK is a toolkit for building Hidden Markov Models (HMMs). However, HTK is primarily designed for building HMM-based speech processing tools, in particular recognizers. The HTK training tools are used to estimate the parameters of a set of HMMs using training utterances and their associated transcriptions (speaker labels). Secondly, unknown utterance's identity can be found using the HTK recognition tools [6].

First create HMM models of speakers on training data of four speakers and the segmented speech given as test waves to identify the speaker identities. The HTK program HCopy can convert a waveform into MFC acoustic feature types. HCompV tool used to initialize the models with the training data, which includes speech waves of each speaker identity. Re-estimate the models using HERest i.e. Re-estimated the HMM parameters then returns the updated model. Finally Hvite tool used to perform recognition using the Viterbi algorithm. The waves obtained after speaker segmentation is used for testing. So that the speakers in an audio and the time at which they are active can be identified.

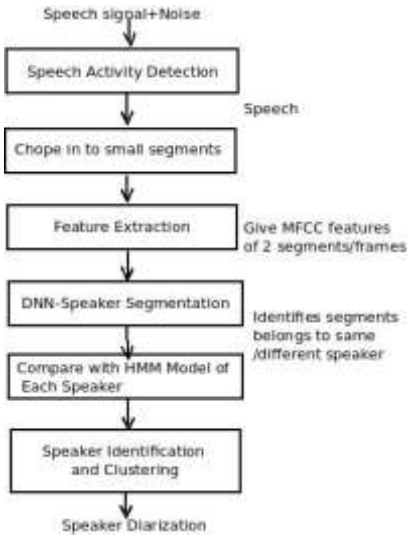

Fig. 2. Flow Diagram

CONCLUSION

This paper proposes a method for speaker diarization using deep learning and HMM modeling. The training set for speaker segmentation used was 30 speech signals of 4 speakers. The efficiency can be improved by expanding dataset with more number of speech and speakers for training the DNN and HMM.

ACKNOWLEDGMENT

I would like to thank my guide Mr. Raseek C, Mr. Amal Babu and all others for the support and guidance given.

REFERENCES

[1] Houman Ghaemmaghami, David Dean, Sridha Sridharan, " A cluster-voting approach for speaker diarization and linking of Australian broadcast news recordings" , In Proceedings of the 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2015, IEEE, Brisbane, Australia, pp. 4829-4833.

[2] Xavier Anguera, Simon Bozonnet, Nicholas W. D. Evans, Corinne Fredouille, Gerald Friedland, and Oriol Vinyals, "Speaker diarization: A review of recent research," IEEE Transactions on Audio, Speech & Language Processing, pp.356–370, 2012.

[3] Sue E Tranter and Douglas A Reynolds, "An overview of automatic speaker diarization systems," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 5, sep 2006.

[4] David I-Chung Wang, "Speaker Diarization - Who Spoke When", Thesis, Queensland University of Technology, Speech and Audio Research Laboratory Science and Engineering Faculty, October 2012

[5] https://en.wikipedia.org/wiki/Deep_belief_network

[6] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain et.al "The HTK Book", March 2009