ICEST-2K24 [International Conference on Engineering, Science and Technology]
In Association with
International Journal of Scientific Research in Science, Engineering and Technology
Print ISSN - 2395-1990 | Online ISSN : 2394-4099 (www.ijsrset.com)

Speaker Diarization Using Spectral Clustering

Ms. Priyanka N. Kokare[1], Abha S. Pathak[2], Simran S. Pardeshi[2], Rutuja A. Shinde[2]

[1]Assistant Professor, Department of Information Technology, Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India

[2]UG Students, Department of Information Technology, Vidya Pratishthan's Kamalnayan Bajaj Institute of Engineering and Technology, Baramati, Maharashtra, India

ABSTRACT

Speaker diarization is a process in which multiple speakers are getting separated and labeled from a single voice channel. It helps us to answer" Who spoke when?" in a multi-speaker environment. By using a clustering approach, we can segregate different speaker utterances. We have used k-means and spectral clustering methods to observe different clustering algorithms and analyze that spectral clustering works fine for LSTM-based d-vector embeddings while k- means gives the wrong prediction when two- speaker speaks simultaneously, they overlap when one speaker speaks more than another then the unbalancing of clusters takes place.

Both speaker diarization and speaker verification share the common goal of distinguishing between speakers. However, the primary distinction lies in their approach: speaker verification models are trained using data from target speakers, while speaker diarization lacks prior information about speakers in the recording. Speaker diarization finds utility in applications such as speaker adaptation for automatic speech recognition, audio indexing, and speaker localization. Speaker Diarization combines the LSTM-based d-vector audio embedding using spectral clustering where the segments will be converted into d- vectors. So will be using different clustering algorithms to check which clustering gives better results.

Keywords: LSTM, VAD, KNN,

I. INTRODUCTION

To develop a system capable of extracting multiple speakers from single-channel audio using a speaker diarization engine and enhancing system performance. "Diarize" refers to the act of recording or noting an event in a diary. Speaker diarization is an essential process in the domain of audio and speech analysis. It involves dividing an audio recording into segments based on the speaker's identity. Spectral clustering and LSTM (Long Short-Term Memory) networks are two distinct techniques that can be amalgamated to conduct speaker diarization, which entails segmenting an audio recording into various non-overlapping segments or clusters, each corresponding to a specific speaker or audio source. The primary objective of speaker diarization is to ascertain "who spoke when" in an audio recording, rendering it a valuable asset in numerous applications, including transcription services, voice assistants, forensic analysis, and more. Speaker diarization finds practical utility in a broad array of fields, such as transcription services, call center analytics, automatic subtitling, voice

biometrics, and forensic voice analysis. It facilitates the automation of tasks requiring the identification of speakers within audio recordings, making it a valuable asset in both research and commercial applications.

II. LITERATURE SURVEY

Reference	Task	Paper	Dataset	Method	Model	Der% /Acc%
[1]	Speaker Diarization	Speaker Diarization With Lstm	Callhome	1) K Means 2)Spectral	I-Vector, D-Vector	12.0%
[2]	Speaker Diarization	Fully Supervised Speaker Diarization	Callhome	1) Eend 2) Sc-Eend	Sc-Eend	Two Speakers 8.86% Variable Speakers: 15.75%
[3]	Speaker Diarization	Told: A Novel Two Stage Overlap Aware Framework For Speaker Diarization.	Callhome	Eend	1)Told 2)Eend- Ola	1) 10.14% 2) 12.57%
[4]	Speaker Diarization	End-To-En d Neural Speaker Diarization With Self Attention	Callhome	-	Sa-Eend (2-Spk, Adapted) Sa-Eend (2-Spk, No-Adapt)	1) 10.76 2) 12.66
[5]	Speaker Diarization	Auto-Tuning Spectral Clustering For Speaker Diarization Using Normalized Maximum Eigengap	Callhome	Spectral Clustering	1)Cos+Njw-Sc (Oracle Sad) 2) COS+AHC (Oracle SAD)	1) 24.05% 2) 21.13%

Table 1:Literature Survey

III.PROPOSED WORK

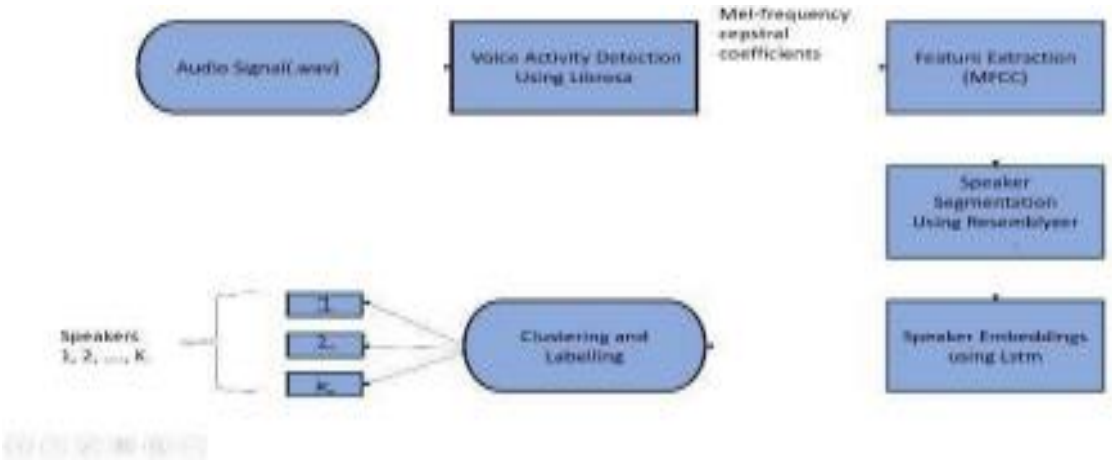

Figure 1:Proposed Architecture

Audio Signal: This will be our very first step where all the audio signals are given input to the system which are in '.wav' format. The audio files will consist of two or more speakers which are speaking respectively.

Voice Activity Detection: Using python librosa library preprocessing on audio signal done. It removes noise from the audio signal and performs voice activity detection. Output generated in the form of Mel Frequency Ceptral Coefficient.

Speaker Segmentation: Speaker Segmentation will take only the speech part of the audio input and it will separate the speech to overlaps windows. The output Speaker Segmentation is in the form of small segments.

Speaker Embedding: The output of the Speaker Segmentation is given input to the Speaker Embedding where the segments are converted into d-vectors using the LSTM algorithm. The inputs undergo through the LSTM algorithm and the outputs will be in the form of d-vectors.

LSTM: Long Short-Term Memory (LSTM) is a type of recurrent neural network (RNN) architecture that is designed to capture long-range dependencies and patterns in sequential data. It was introduced to address the vanishing gradient problem that traditional RNNs faced when trying to learn relationships between distant elements in a sequence.

Clustering: D-vectors are inputs to the Clustering. We will be using the Spectral Clustering algorithm. It is an approach of data classification that estimate show likely a data point is to be a member of one group or the other. By performing a sequence of refinement operations on input the same voice is grouped together into one single group and forwarded to the separate file.

Labelling: The group of Clusters will be saved into separate file formats. One file format will contain the speech of only one person. Thus, every speaker has its own single separate file.

IV.CHALLENGES

Scalability: Extending diarization to large datasets and real-time applications requires scalable algorithms and efficient processing Domain Adaptation: Adapting diarization systems to new domains or languages can be complex due to differences in speaking styles and acoustic conditions.

Privacy Concerns: Ethical and privacy considerations are vital when deploying diarization systems, especially in contexts where sensitive data is involved.

V. RESULT

Dataset	DER
CALLHOME	0.625
Audio 1(3 speaker)	0.221
Audio 2(7 speaker)	0.003

Table 2:Result

VI. APPLICATIONS

Speech Transcription: In scenarios like conference calls, meetings, or interviews, speaker diarization helps separate speech segments by different speakers, making it easier to transcribe each speaker's content accurately. This improves the quality and efficiency of automatic speech recognition (ASR) systems.

Content Analysis and Summarization: Speaker diarization helps understand who's speaking in a conversation. This lets us figure out who talks the most and how people interact. By knowing who says what, we can summarize discussions better. It helps find out who's the main speaker and how the conversation flows. Overall, it makes it easier to analyze and summarize what's being said in a conversation.

Medical field: During medical consultations or procedures involving multiple healthcare professionals, speaker diarization can help transcribe the conversation accurately by attributing each speaker's dialogue. This ensures that medical notes or reports are correctly documented, aiding in patient care and record-keeping.

Voice Biometrics and Forensics: Speaker diarization can aid in voice biometrics and forensic analysis by separating different speakers' voices within a recording. This can be useful in criminal investigations, authentication systems, and identifying speakers in surveillance recordings. ongoing advancements.

VII. CONCLUSION

In short, speaker diarization is the process of identifying and labelling multiple speakers in audio recordings. It's used in various applications and has evolved with technology. Challenges include handling overlapping speech and noisy environments. It has gone through various milestones, from early rule-based methods to the integration of deep learning models like LSTMs and spectral clustering for improved accuracy. Different techniques and feature extraction methods play a key role in this process, but it often requires labelled data for training. Overall, speaker diarization is important. .

VIII. REFERENCES

[1]. Wang, Quan Downey, Carlton Wan, Li Mansfield, Philip Moreno, Ignacio. “Speaker Diarization with LSTM,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 10.1109/ICASSP.2018.8462628

[2]. Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, Chong Wang. “Fully Supervised Speaker Diarization ”, ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 10.1109/ICASSP.2019.8683892

[3]. Jiaming Wang, Zhihao Du, Shiliang Zhang, “ TOLD: A Novel Two-Stage Overlap-Aware Framework for Speaker Diarization”, ICASSP2023. https://doi.org/10.48550/arXiv.2303.05397

[4]. Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Yawen Xue, Kenji Nagamatsu, Shinji Watanabe, “ End-to-End Neural Speaker Diarization with Self-attention”, ASRU 2019. https://doi.org/10.48550/arXiv.1909.06247

[5]. Tae Jin Park, Kyu J. Han, Manoj Kumar, Shrikanth Narayanan, “ Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap”, IEEE Signal Processing Letters,2020 https://doi.org/10.48550/arXiv.2003.02405