Unraveling the Techniques for Speaker Diarization

Ganesh Pechetti, Anakapalli Rohini Durga Bhavani, Abhinav Dayal(✉), and Sreenu Ponnada

Computer Science and Engineeering Department, Vishnu Institute of Technology, Bhimavaram, Andhra Pradesh, India
{21pa1a05d1,21pa1a0508,abhinav.dayal}@vishnu.edu.in

Abstract. This research paper aims to contribute to the field of speaker diarization by providing an in-depth analysis of existing audio datasets and evaluating prominent models. The study focuses on the suitability of these datasets for studying speaker diarization tasks and examines the performance of models such as pyannote-speaker diarization and NVIDIA NeMo speaker diarization. For aspiring researchers in the field, this paper serves as a solid foundation, offering valuable guidance and resources for experimentation in speaker diarization. The evaluation of the models reveals important insights. While each model has its advantages, their limitations must be considered. Overall, this research paper provides valuable insights into audio dataset analysis, model evaluation, and selection considerations for speaker diarization tasks. It equips researchers with essential knowledge to make informed decisions and lays the groundwork for further advancements in the field.

Keywords: Speaker Diarization · Segmentation · Voice Activity Detection · Pyannote · Kaldi · NeMo

1 Introduction

Speaker Diarization is the task of dividing an audio sample, which contains multiple speakers, into segments that belong to individual speakers based on their homogeneous characteristics [1]. Throughout the years, numerous speaker diarization models have been proposed, each with its distinctive approach and underlying techniques. As the demand for accurate and efficient speaker diarization systems continues to grow, it becomes essential to compare and evaluate the existing models.

The main steps involved in the speaker diarization are VAD(Voice Activity Detection), segmentation, feature extraction, clustering, and labeling. VAD identifies voice activity regions in an audio sample, while segmentation splits the large audio into smaller samples. Feature extraction techniques are applied to these smaller chunks to extract features and convert them into embeddings. Clustering techniques group the embeddings into clusters based on the extracted features. Finally, the audio sample is annotated,

P. Pareek et al. (Eds.): IC4S 2023, LNICST 536, pp. 300–308, 2024.
https://doi.org/10.1007/978-3-031-48888-7_25

assigning labels to the clusters [1]. To have a deeper understanding of these steps go through the blog at[1].

Speaker diarization is a very important step in speaker identification. Because it allows for the accurate identification by separating the speakers within audio recordings. It has a lot of applications some of them are transcription service [2] which refers to the process of converting audio recording into written text, used in forensic investigations [3], call centers [4, 5], speaker identification [6, 7] etc.

Initially, traditional clustering methods like Gaussian Mixture Models (GMMs) [8] and Hidden Markov Models (HMMs) [9] were used for audio sample diarization. Subsequently, the Bayesian Information Criterion(BIC) [10]. Later, an i-vector [11] based method was proposed. The field experienced a breakthrough with the application of AI and Deep Learning networks, leading to the proposal of revolutionary architectures for diarization. Now-a-days, the majority of models with lower Diarization Error Rate(DER) make use of Deep Neural Network (DNN) based architectures.

In this study, we focused on exploring various models for speaker diarization and we selected some models which have less error rate namely DER. With these models we try to compare their performance. Later, we evaluated and compared the selected models based on their ability to accurately separate speakers.

The key contributions of this research work are as follows. Firstly, an in-depth analysis of existing audio datasets is provided, specifically focusing on their suitability for studying the speaker diarization task. Secondly, a comprehensive examination and evaluation of prominent models, including pyannote-speaker diarization and NVIDIA NeMo speaker diarization, are conducted. Moreover, for aspiring researchers embarking on speaker diarization research, this paper serves as a solid foundation, offering guidance and resources for experimentation in this domain. By consolidating knowledge and highlighting relevant tools and methodologies, this work facilitates a smoother initiation and exploration of speaker diarization research endeavors.

2 Related Work

This section introduces the existing mechanisms to perform speaker diarization. Speaker diarization can be performed with the help of i-vectors [11], x-vectors [12] and through deep neural networks [13]. With the advent of deep learning technology, significant advancements have been made in the field of speaker diarization. These advancements have propelled the development of more advanced techniques and methodologies.

There have been various methods developed over time for different steps in diarization which includes VAD, segmentation, feature extraction, and clustering etc. Traditional approaches like GMMs [8] and HMMs [9] which are basically clustering models struggled with handling overlapped speech, leading to inaccurate segmentation and speaker assignment. They also have limitations in capturing the full range of speaker characteristics. Landini et al., [10] proposed BIC clustering was introduced to determine the optimal number of clusters, but it also faced challenges with overlapping speech. Subsequently, Wang et al., [11] proposed i-vectors and Kim et al., [12] x-vectors

[1] https://medium.com/@21pa1a05d1/speaker-diarization-fec87f839f52.

which are used for extracting the features from an audio sample, were introduced as low-dimensional representations using neural networks. Garcia-Romero et al., [13] proposed deep neural networks embeddings. However, extracting i-vectors and x-vectors requires significant computational power. Nowadays, advanced methods such as neural speaker segmentation, multi-modal approaches, and deep neural networks techniques for VAD and Multi-Scale Diarization Decoder (MSDD) [14] are being used.

3 Methodology

This section provides the methods adopted to carry out the study. The architecture used to perform this study is presented through Fig. 1. It provides insights on the adopted datasets, models, pre-processing approaches used and performance evaluation.

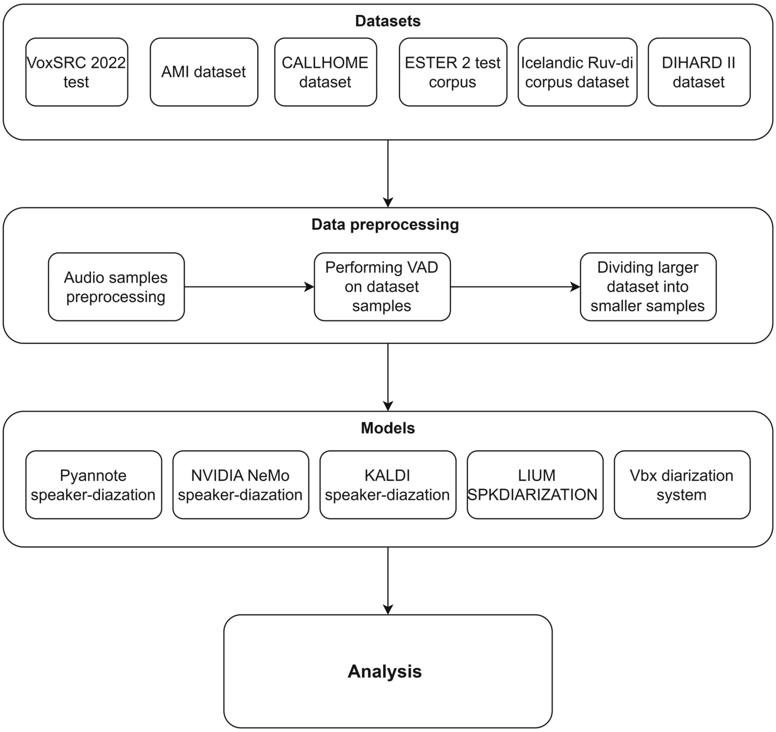

Fig. 1. The framework used to perform the study

The first block of the Fig. 1 showcases the datasets that are used by most of the researchers. The second block includes the preprocessing steps we performed on the datasets. Next block includes the various models used to evaluate the performance of the speaker diarization task on the datasets and the final block includes the analysis section where we compare different models.

In our study, we conducted preprocessing steps to enhance the quality of our data. Firstly, we performed noise removal to eliminate unwanted noise from the dataset [14]. Additionally, we applied VAD [16] to identify and remove silent portions of the audio samples. Subsequently, we segmented the audio samples to create segments that are suitable for the diarization process.

To evaluate the performance of a model the commonly used metrics are DER [1] and JER(Jaccard Error Rate) [1]. In our study, we have chosen to focus on DER as it is the preferred metric used by the majority of researchers in the diarization field. DER is the sum of speaker error, false alarm speech, missed speech.

$$\mathrm{DER} = \frac{\mathrm{SER} + \mathrm{FA} + \mathrm{Miss}}{\mathrm{Total\_speech}} \tag{1}$$

3.1 Corpora

VoxSRC-2022 Test Dataset: The VoxSRC-2022 dataset[2], created for the VoxCeleb Speaker Recognition Challenge in 2022, is a valuable collection of speech obtained ‘in the wild.’ It features audio from diverse sources like celebrity interviews, news shows, talk shows, and debates, representing real-world scenarios. The dataset includes professionally edited videos and casual conversational audio, offering a wide range of speech styles and acoustic conditions.

With 5,994 speakers and 1,092,009 utterances, it provides a substantial amount of data for analysis and model training. The dataset’s inclusion of background noise, laughter, and other natural artifacts adds realism to the evaluation of speaker recognition methods. Researchers can leverage this dataset to enhance the performance and reliability of speaker recognition systems, making them more applicable in practical applications.

Icelandic Ruv-di Corpus Dataset: The Icelandic Ruv-di corpus dataset[3] consists of speech data sourced primarily from the Icelandic national broadcasting service, RÚV. It offers authentic recordings from various programs, interviews, and news broadcasts, making it valuable for speech and language processing tasks in Icelandic. The dataset captures natural speech patterns, accents, and styles and represents variation in these characteristics.

Depending on the release, it may include annotations like transcriptions or speaker identities. The dataset’s focus on Icelandic ensures targeted solutions for language-specific challenges and its availability facilitates research and enables the development of speech processing models tailored to Icelandic. In summary, the Icelandic Ruv-di corpus dataset offers authentic and diverse speech data focused on Icelandic with the potential for annotations.

CALLHOME Dataset: The CALLHOME American English Speech dataset[4] was developed by the Linguistic Data Consortium (LDC). It consists of 120 unscripted 30-min telephone conversations between native speakers of English language. This means

[2] http://mm.kaist.ac.kr/datasets/voxceleb/voxsrc/competition2022.html.

[3] https://clarin.is/en/resources/j_ruv/.

[4] https://catalog.ldc.upenn.edu/LDC97S42.

that there are a total of 240 speakers in the dataset, with two speakers in each audio recording. The total duration of the data in the dataset is 60 h. However, it only contains telephone conversations between native speakers of English and was collected in the 1990s, which may limit its applicability to research on other languages or current speech patterns. Despite these limitations, the availability of this dataset for research purposes can facilitate the development and evaluation of speech processing technologies.

AMI Dataset: The AMI Meeting Corpus is a dataset[5] that combines multiple modes of data and comprises 100 h of recorded meetings. It includes both elicited and naturally occurring meetings and provides a rich source of data for research. The dataset consists of 171 meetings recorded at 4 locations, with each meeting having 4–5 speakers. However, the dataset primarily consists of recordings from controlled meeting environments and may not fully capture the diversity of real-world multi-party interactions. Additionally, the scale of the dataset may be relatively small compared to other speech or audio datasets and the annotation process can be resource-intensive. Despite these limitations, the AMI Meeting Corpus is a valuable resource for research on multi-party interactions.

DIHARD II Dataset: The DIHARD II dataset[6] is designed for evaluating speaker diarization systems. It includes diverse audio recordings from various sources and provides manual annotations for ground truth evaluation. With approximately 144 unique speakers and around 44 h of audio data, it serves as a valuable benchmark for advancing speaker diarization technology. The dataset offers advantages such as its evaluation focus, varied acoustic conditions, diverse data sources, and established evaluation metrics. However, using the dataset may require significant computational resources, updates may be infrequent, and there could be copyright or licensing restrictions.

ESTER 2 Test Corpus: The ESTER 2 test corpus dataset[7], designed for evaluating French speech transcription systems, consists of approximately 150 h of audio data. This substantial amount of data enhances the dataset's utility for comprehensive training and evaluation of automatic speech recognition (ASR) models. With its diverse audio sources, manual transcriptions, evaluation metrics, and language-specific focus, the ESTER 2 test corpus serves as a valuable resource for advancing ASR technology in the context of the French language.

3.2 Models

Pyannote Diarization Model: The pyannote.audio provides a neural speaker diarization pipeline, which is available through Hugging Face. The pipeline contains neural speaker segmentation which is a method for automatically detecting speaker changes in an audio recording using a neural network. The neural network is trained to analyze the acoustic characteristics of the speech signal and to identify points in time where the speaker changes. This can be done by sliding a fixed-length window over the speech signal and predicting, for each window, whether it contains a speaker change or not. The

5 https://groups.inf.ed.ac.uk/ami/download/
6 https://dihardchallenge.github.io/dihard2/.
7 https://catalogue.elra.info/en-us/repository/browse/ELRA-S0338/.

output of the neural network can then be post-processed to obtain a final segmentation of the audio recording into speaker-homogeneous segments, the SpeechBrain implementation of the ECAPA-TDNN model for extracting feature embeddings. Agglomerative hierarchical clustering is used for clustering embeddings.

It reaches a DER = 5.6% on VoxSRC 2022 test dataset [17].

Nvidia NeMo: NVIDIA NeMo's speaker diarization system consists of several modules: a Voice Activity Detector (VAD) model namely MarbleNet model which is a deep 1D neural network, detects the presence or absence of speech to generate timestamps for speech activity from the given audio recording; a Speaker Embeddings model namely TitaNet, which extracts speaker embeddings on speech segments obtained from VAD time stamps; and a Multi-Scale Diarization Decoder (MSDD), which is a speaker diarization model based on initializing clustering and multi-scale segmentation input. This model has two significant improvements to enhance diarization importance of each scale at each step.

It reaches a DER = 3.92% dataset and DER = 1.05% on CALLHOME and AMI datasets [14].

Kaldi: Kaldi is an open-source toolkit for speech recognition that includes support for speaker diarization and also consists of models for x-vectors. Kaldi's speaker diarization system uses x-vectors, a type of speaker embedding, to represent speech segments. The x-vector extractor is a Time Delay Neural Network (TDNN) that is trained on a large amount of labeled speech data to learn a mapping from speech segments to a fixed-dimensional embedding space. The extracted x-vectors are then used in combination with clustering algorithms such as Agglomerative Hierarchical Clustering (AHC) to group speech segments by speaker. The kaldi model used x-vectors, MFCCS, PLDA trained on Althingi Parliamentary Speech corpus.

It reaches a DER = 26.27% on Icelandic Ruv-di corpus dataset [18].

LIUM SPKDIARIZATION: LIUM SpkDiarization is an open-source toolkit for speaker diarization developed by the LIUM (Le Mans University). It inlcudes a fullset of tools that facilitates the creation of an entire speaker diarization system, starting from the audio signal and progressing towards speaker clustering using CLR/NCLR metrics. These toolset encompasses MFCC computation, speech/non-speech detection, and various speaker diarization methods. The LIUM_SPKDIARIZATION model uses hierarchical agglomerative clustering methods using measures such as BIC and CLR(Classification Likelihood Ration).

It reaches a DER = 10.01% on ESTER 2 test corpus [19].

VBX: VBx is a recently proposed speaker diarization method that uses a Bayesian Hidden Markov Model (BHMM) to cluster x-vectors, which are fixed-dimensional representations of variable-length speech segments. The VBx method utilizes a BHMM to group x-vectors and identify speaker clusters within a sequence of x-vectors. The VBx model applies the same BHMM approach to detect speaker clusters in a sequence of x-vectors.

It reaches a diarization error rate of DER = 21.77% on CALLHOME dataset and DER = 18.55 on DIHARD II dataset [10].

4 Results and Discussion

Our study adopted six different corpora and compared them based on several parameters namely the number of hours of available data, the number of speakers, language and pricing type. The results of this comparison are tabulated in Table 1. Among the corpora, the AMI and ESTER 2 datasets contained the largest amount of audio data in terms of hours. For scenarios requiring a high number of speakers, the Icelandic Ruv-di corpus emerged as a suitable choice.

Fortunately, there are also freely available datasets that can be utilized to initiate diarization research. These include the VoxSRC-2022 test dataset, the Icelandic Ruv-di corpus, and the AMI dataset. These datasets were sourced from various multimedia platforms such as YouTube, radio broadcasts, news shows, debates, and celebrity interviews. Additionally, some of these datasets encompass telephonic recordings, such as the CALLHOME datasets, while others focus on meeting recordings, such as the AMI dataset.

Table 1. Comparison of datasets

S. No	Name of the dataset	No of hours	No of speakers	Style	Language	Pricing type
1	VoxSrc-2022 test	50	4–6	Celebrity, interviews, news, shows, talk, debates	English	Free
2	Icelandic Ruv-di corpus	46 (min)	20	Programs, Interviews, News broadcasts	Icelandic	Free
3	CALLHOME	60	2	Telephone conversations	English	Paid
4	AMI	100	4–5	Meeting recordings	English	Free
5	DIHARD II	44	1–10	YouTube, court rooms, meetings	English	Paid
6	ESTER 2 test corpus	150	1–3	Radio broadcast	French	Paid

Our study adopted five different speaker diarization models, and their performance was evaluated on various datasets. The results, including the DER, are presented in Table 2. The models selected for our study were Pyannote-speaker diarization, NVIDIA NeMo speaker diarization, Kaldi speaker diarization, LIUM speaker diarization, and VBX speaker diarization.

Table 2. Comparison of different models.

S.No	Name of the model	Dataset	DER%
1	Pyannote-speaker diarization	VoxSRC 2022 test	5.6
2	Nvidia Nemo speaker diarization	CALLHOME, AMI	3.92, 1.05
3	Kaldi speaker diarization	Icelandic Ruv-di corpus	26.27
4	LIUM speaker diarization	ESTER 2 test corpus	10.01
5	Vbx speaker diarization	CALLHOME, DIHARD II	21.77, 18.55

Among these models, Pyannote-speaker diarization demonstrated better performance with a DER of 5.6, while NVIDIA NeMo achieved a DER of approximately 3.92. In some of the rows two dataset names are provided in the dataset column, along with two corresponding DER values in the DER column. This signifies that the first dataset name corresponds to the first DER value, while the second dataset name corresponds to the second DER value.

While each model has its unique advantages, it is important to consider their limitations as well. If the primary goal is to minimize the DER and time is not a limiting factor, the pyannote diarization model can be a suitable choice. However, it should be noted that this model might require more time to generate the output. The NVIDIA NeMo speaker diarization model is designed to work with specific audio requirements, including bitrate, duration, and other properties. It is crucial to ensure that the original audio sample meets these specifications in order to effectively utilize the NVIDIA NeMo model. When the properties are matched, selecting the NVIDIA NeMo model can lead to lower DER compared to the pyannote-diarization model. Additionally, the NVIDIA model exhibits faster processing times compared to pyannote-diarization, offering a more efficient solution for speaker diarization tasks. We also mentioned some other models which are good to use but have more DER.

5 Conclusion

In addition to the models mentioned earlier, there are other models available for speaker diarization, although they may have a higher DER compared to the previously discussed options. It is important to consider that the DER changes depending on the dataset used. Therefore, it is crucial to select a model that best fits the dataset is very important. The choice of model should be based on evaluation of its performance for the given dataset, considering factors such as accuracy, efficiency, and compatibility. This ensures that the selected model aligns well with the dataset and maximizes the effectiveness of speaker diarization outcomes.

References

1. Tae Jin, P., et al.: A review of speaker diarization: recent advances with deep learning. Comput. Speech Lang. **72**, 101317 (2022)

2. Claude, B., et al.: Multistage speaker diarization of broadcast news. IEEE Trans. Audio Speech Lang. Process. **14**(5), 1505–1512 (2006)
3. Joyanta, B., et al.: An overview of speaker diarization: approaches, resources, and challenges. In: 2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA). IEEE (2016)
4. Zajíc, Z., Kunešová, M., Müller, L.: Applying EEND diarization to telephone recordings from a call center. In: Karpov, A., Potapova, R. (eds.) SPECOM 2021. LNCS (LNAI), vol. 12997, pp. 807–817. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87802-3_72
5. Rosenberg, Aaron E., et al.: Unsupervised speaker segmentation of telephone conversations. In: INTERSPEECH (2002)
6. Aleksandar, M., Gerazov, B., Ivanovski, Z.: Delay based optimization of an integrated online call recording speaker diarisation and identification system. In: IEEE EUROCON 2017–17th International Conference on Smart Technologies. IEEE (2017)
7. Lakshmana Rao, A., Bonthu, S., Dayal, A.: Multiclass spoken language identification for Indian Languages using deep learning. In: 2020 IEEE Bombay Section Signature Conference (IBSSC). IEEE (2020)
8. Tantan, L., Liu, X., Yan,Y.: Speaker Diarization System Based on GMM and BIC. In: International Symposium on Chinese Spoken Language Processing, Singapore (2006)
9. Jeremy, H.M.W., Xiao, X., Gong, Y.: Hidden markov model diarisation with speaker location information. In: ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE (2021)
10. Federico, L., et al.: Bayesian hmm clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation, and analysis on standard tasks. Comput. Speech Lang. **71**, 101254 (2022)
11. Wei, W., et al.: I-vector features and deep neural network modeling for language recognition. Procedia Comput. Sci. **147**, 36–43 (2019)
12. Myungjong, K., Apsingekar, V.R., Neelagiri, D.: X-Vectors with Multi-Scale Aggregation for Speaker Diarization. arXiv preprint arXiv:2105.07367 (2021)
13. Garcia-Romero, D., et al.: Speaker diarization using deep neural network embeddings. In: 2017 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). IEEE (2017)
14. Park, T.J., et al.: Multi-scale speaker diarization with dynamic scale weighting. arXiv preprint arXiv:2203.15974 (2022)
15. Fu, S.-W., et al.: MetricGAN+: an improved version of metricgan for speech enhancement. arXiv preprint arXiv:2104.03538 (2021)
16. Hao, Z., Deming, L.: Research of voice activity detection algorithm. In: 2011 International Conference on Computational and Information Sciences. IEEE (2011)
17. Bredin, Hervé: pyannote. audio speaker diarization pipeline at VoxSRC 2022
18. Fong, J.Y., Gudnason, J.: RÚV-DI Speaker Diarization (20.09) (2020)
19. Sylvain, M., Merlin, T.: LIUM SpkDiarization: an open source toolkit for diarization. In: CMU SPUD Workshop (2010)