International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

Speaker Diarization: A Review

Krishna Kumar[1]

[1] *RACE REVA University*

Abstract - Speaker diarization is the task of determining "who spoke when?" in an audio or video recording that contains an unknown amount of speech and an unknown number of speakers. It is a challenging task due to the variability of human speech, the presence of overlapping speech, and the lack of prior information about the speakers. It is the process of labeling a speech signal with labels corresponding to the identity of speakers. It is a crucial task in audio signal processing and speech analysis. A recent review of speaker diarization research since 2018 can be found in this paper which discusses the historical development of speaker diarization technology and recent advancements in neural speaker diarization approaches.

Key Words: speaker diarization, speaker clustering, speaker embeddings

1.INTRODUCTION

Speaker diarization, also known as speaker segmentation and clustering, is a crucial task in audio signal processing and speech analysis. It involves partitioning an audio stream into homogeneous segments corresponding to different speakers. This literature review examines recent research contributions in speaker diarization over the past five years. This literature review examines recent research contributions in speaker diarization, highlighting key methodologies, techniques, and performance improvements.

2. RELATED WORK

There are different studies in the literature on evaluating the performances of speaker diarization systems.

A typical speaker diarization system consists of four components:

(1) Speech segmentation

(2) Audio embedding extraction

(3) Clustering

(4) Re-segmentation

Quan et el in paper [13] SPEAKER DIARIZATION WITH LSTM builds a d-vector based speaker verification systems to develop a new d-vector based approach to speaker diarization. They combine LSTM-based d-vector audio embeddings with nonparametric clustering to obtain a state-of-the-art speaker diarization system. This system is evaluated on three standard public datasets, suggesting that d-vector based diarization systems offer significant advantages over traditional i-vector based systems. They achieved a 12.0% diarization error rate on NIST SRE 2000 CALLHOME, while the model is trained with out-of-domain data from voice search logs. Fig 1 shown below is the flow chart of the system from paper [13].

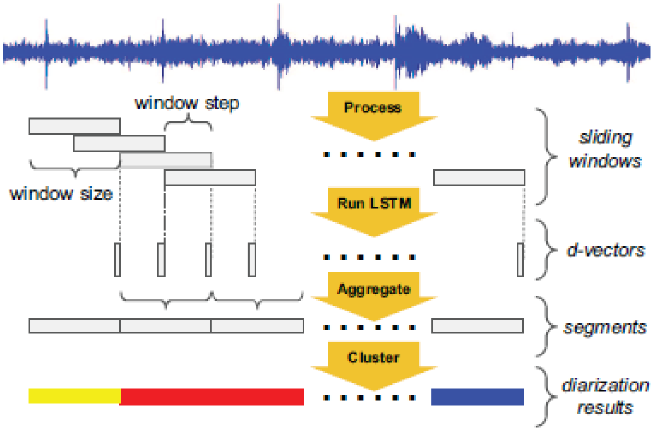


Fig -1: A flowchart of d-vector based diarization system.

Matthew et el in paper [6] CHARACTERIZING PERFORMANCE OF SPEAKER DIARIZATION SYSTEMS ON FAR-FIELD SPEECH USING STANDARD METHODS used AMI Meeting Corpus data and showed, there is a degradation of diarization performance on far-field speech. They had used an i-vector/PLDA (Probabilistic Linear Discriminant Analysis)-based diarization system and compared performance on near-field, far-field, and signal-enhanced conditions. Agglomerative Hierarchical Clustering (AHC) was used. AHC is a “bottom-up” clustering approach where each i-vector is assigned to a cluster and each cluster is merged according to the PLDA score until a stopping criterion is met.

Aonan et el in paper [14] FULLY SUPERVISED SPEAKER DIARIZATION propose a fully supervised speaker diarization. They call it unbounded interleaved-state recurrent neural UIS-RNN. The clustering module is replaced by a trainable unbounded interleaved-state RNN. The sequence generation model was composed of one layer of 512 GRU cells with a tanh activation, followed by two fully connected layers each with 512 nodes and a ReLU activation. Beam search of width 10 was used for decoding. Evaluation was done on single channel audio CALLHOME data. When this UIS-RNN was trained on both in domain and out of domain dataset with a large corpus, 7.6% DER was achieved. d-vector embedding was used as input features.

Tae Jin Park et el in paper [15] Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap had proposed a framework using normalized maximum eigengap (NME) values to estimate the number of clusters and the parameters for the threshold of the elements of each row in an affinity matrix during spectral clustering, without the use of parameter tuning on the development set. Experiments on CALLHOME data show and improvement in DER. Using System SAD (speech activity detection) 5.41% DER and Oracle SAD 7.29% DER.

Andreas et el in paper [3] DOVER: A METHOD FOR COMBINING DIARIZATION OUTPUTS had proposed an algorithm for weighted voting among diarization hypotheses. It was found that DOVER (diarization output voting error reduction), consistently reduces diarization error rate over the average of results from individual channels, and often improves on the single best channel. Any comparison on the CALLHOME dataset is not present but it is expected to give better results when diarization is performed on multiple channels of audio data.

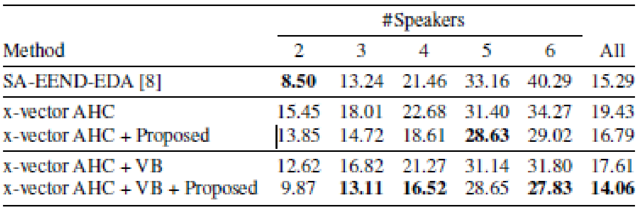
G. Sun et el in paper [8] SPEAKER DIARISATION USING 2D SELF-ATTENTIVE COMBINATION OF EMBEDDINGS proposes a generic framework to improve performance by combining i-vectors and d-vectors into a single embedding, referred to as a c-vector. Two types of DNN models, a feedforward TDNN and a high order recurrent neural network (HORNN) system are studied as example d-vector systems. Two approaches are studied, a simultaneous attention approach where the annotation matrix produced by a single attention model and a second consecutive attention approach where a separate attention across time is performed inside each system before the final attention across systems. Experiments show that penalty term improves d-vector extraction by a clear margin. c-vectors with both 2D attentive structures outperform the individual d-vector systems. The consecutive structure gives the lowest error rate. AMI corpus data set is used and so the results cannot be compared with others with CALLHOME data.

Vivek et el in paper [10] DESIGNING AN EFFECTIVE METRIC LEARNING PIPELINE FOR SPEAKER DIARIZATION using empirical studies proposes a better sampling strategy and loss function and margin parameter selection for speaker diarization using deep learning. Study shows that Choice of sampling strategy could be Distance-Weighted Sampling or Semi-Hard Mining while Random does not perform well in any case. Similarly, Loss function could be Triplet or Quadruplet. Margin can be Fixed or Adaptive. A best DER of 12.47% was achieved with combination of DWS Quadruplet Fixed on CALLHOME dataset. MFCC features were used in the experiments.

Latan´e et el in paper [4] OVERLAP-AWARE DIARIZATION: RESEGMENTATION USING NEURAL END-TO-END OVERLAPPED SPEECH DETECTION proposes a LSTM (Long Short-Term Memory) based architecture for overlap detection. The paper tries to address the problem of effectively handling overlapping speech in a diarization system. The author introduces a neural architecture for overlap detection and a re-segmentation module that assigns two speakers in frames detected as overlapping speech. The module has been tested on AMI, DIHARD and ETAPE datasets. Evaluation of model on AMI shows a 20% relative DER reduction over the baseline system.

Shota et el in paper [2] END-TO-END SPEAKER DIARIZATION AS POST-PROCESSING proposes a two-speaker end-to-end diarization method as post-processing of the results obtained by a clustering-based method. For experiments CALLHOME, AMI and DIHARD II datasets are used. The paper focuses on the problem of multiple speakers in a speech segment or overlapping speech. First obtain the initial diarization result using x-vector clustering, which does not produce overlapping results. Iterate on i) frame selection to contain only two speakers and silence and ii) overlap estimation using a two-speaker EEND model.

Table -1: DERs (%) on CALLHOME. All the results include overlapped regions and are NOT based on oracle AD. Collar tolerance of 0:25 s is allowed.



	#Speakers					
Method	2	3	4	5	6	All
SA-EEND-EDA [8]	**8.50**	13.24	21.46	33.16	40.29	15.29
x-vector AHC	15.45	18.01	22.68	31.40	34.27	19.43
x-vector AHC + Proposed	13.85	14.72	18.61	**28.63**	29.02	16.79
x-vector AHC + VB	12.62	16.82	21.27	31.14	31.80	17.61
x-vector AHC + VB + Proposed	9.87	**13.11**	**16.52**	28.65	**27.83**	**14.06**

Table 1 is taken from the paper [2] shows 1.23% DER improvement on CALLHOME dataset on SA-EEND-EDA method.

Herv´e et el in paper [5] PYANNOTE.AUDIO: NEURAL BUILDING BLOCKS FOR SPEAKER DIARIZATION introduces a python library pyannote.audio. This is an open-source toolkit for speaker diarization. pyannote.audio is based on PyTorch machine learning framework, it provides a set of trainable end-to-end neural building blocks that can be combined and jointly optimized to build speaker diarization pipelines. It also comes with pre-trained models covering a wide range of domains for voice activity detection, speaker change detection, overlapped speech detection and speaker embedding. pyAudioAnalysis is another python library that can be used for audio signal analysis and speaker diarization. There is no need for any feature extraction and a waveform is the direct input. The results are reproducible.

Zili et el in paper [5] SPEAKER DIARIZATION WITH REGION PROPOSAL NETWORK proposes a Region Proposal Network based Speaker Diarization (RPNSD). The model can handle overlapped speech. Model combines the segmentation, embedding extraction and re-segmentation into one stage as seen in Fig -2. This has been taken from paper [5].

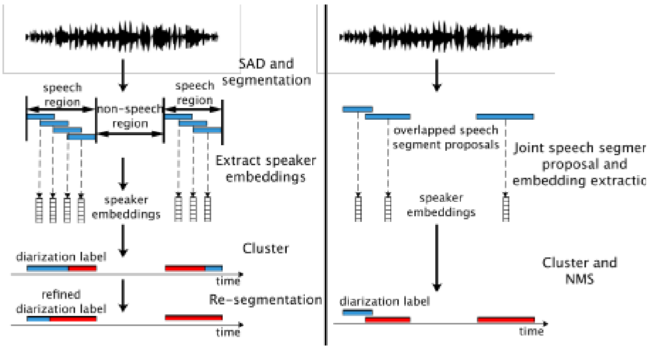


Fig -2: Pipelines of the standard diarization system (left) and the RPNSD system (right)

Feature maps are the input and RPN predicts speech segment proposals. The RPN also predict scores and refine boundaries for each speech segment proposal with convolution layers. RoIAlign pools the features into fixed dimension. During post-processing, speech segment proposals whose fg probability is lower than a threshold is removed and clustering is applied to group the segments and NMS is applied to remove the highly overlapped segment proposals to get the diarization prediction. As per the table below from the paper [5] RPNSD performs better on CALLHOME dataset. The DER includes Miss Error (MI), False Alarm Error (FA), and Confusion Error (CF).The SAD error includes Miss (MI) and False Alarm (FA).

Table -2: The DER composition of different diarization systems on CALLHOME dataset.

System	DER	DER breakdown			SAD error	
		MI	FA	CF	MI	FA
x-vector	32.20	18.6	5.1	8.6	4.2	5.3
x-vector (+VB)	29.54	18.6	5.1	5.9	4.2	5.3
RPNSD	25.46	12.8	7.5	5.2	5.2	3.2

Enrico et el in paper [9] SUPERVISED ONLINE DIARIZATION WITH SAMPLE MEAN LOSS FOR MULTI-DOMAIN DATA propose qualitative modification to paper [14]. The clustering module is replaced by a trainable model called unbounded interleaved-state RNN. A novel loss called Sample Mean Loss function is proposed that significantly improve the learning efficiency and the overall diarization performance.

Jixuan et el in paper [16] SPEAKER DIARIZATION WITH SESSION-LEVEL SPEAKER EMBEDDING REFINEMENT USING GRAPH NEURAL NETWORKS presents the first use of graph neural networks (GNNs) for the speaker diarization problem, utilizing a GNN to refine speaker embeddings locally using the structural information between speech segments inside each session. Fig -3 is taken from paper [16], shows the GNN. Experiments with x-vectors on CALLHOME dataset show a DER of 7.24% and experiment with d-vector for speaker number detection achieves a 73.7% relative DER reduction compared to the model with the original embeddings.

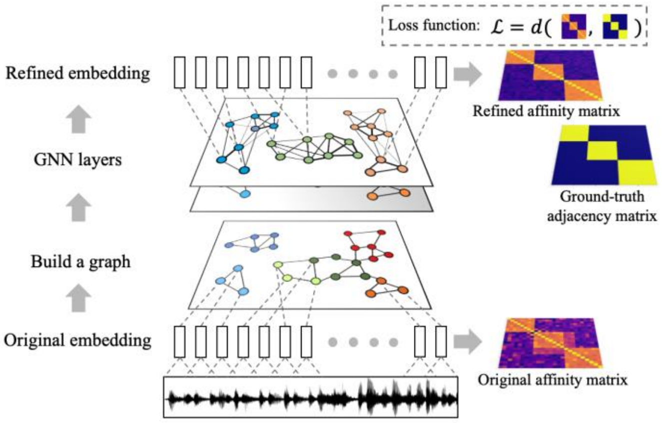


Fig -3: A GNN model is applied to remap the original embedding into another embedding space.

Federico et el in paper [1] BUT SYSTEM FOR THE SECOND DIHARD SPEECH DIARIZATION CHALLENGE describes the winning systems developed by the BUT team for the four tracks of the Second DIHARD Speech Diarization Challenge. Diarization Challenge. The tracks 1 and 2 mono channel and agglomerative hierarchical clustering (AHC) of x-vectors, followed by another x-vector clustering based on Bayes hidden Markov model and variational Bayes inference worked well. Tracks 3 and 4 were multi-channel data and so the best performance was achieved with applying AHC on the fusion of per channel probabilistic linear discriminant.

Bowen et el in paper [12] TSUP Speaker Diarization System for Conversational Short-phrase Speaker Diarization Challenge describes the new evaluation metric called conversational diarization error rate (CDER) to the ISCSLP 2022 conversational short-phrase speaker diarization (CSSD) challenge. Fig -4 below taken from paper [12] sows the relation between segment duration and DER and CEDR.

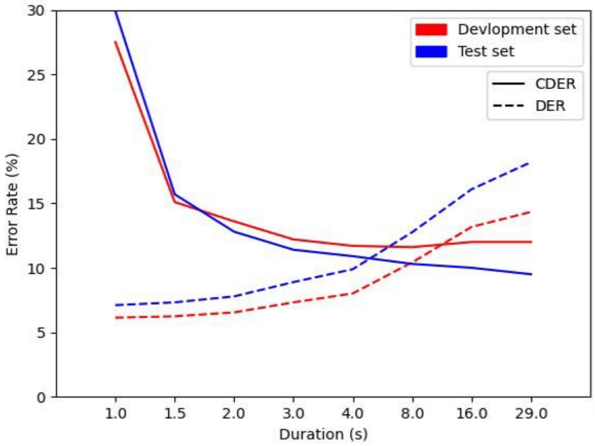


Fig -4: The influence of different sub-segment duration to DER (%) and CDER (%)

It is evident that CDER falls with increasing duration. Also, the paper [12] shows that SC, TS-VAD and EEND are explored and then DOVER-LAP was applied to fuse the RTTM outputs inferred from the above three systems to get the result. Paper [12] shows that spectral clustering-based speaker diarization on metric CDER performs best with 12.0% and

9.5% on the dev set and test set respectively and hence is competitive with the new CDER metric.

Tao et el in paper [11] The X-Lance Speaker Diarization System for the Conversational Short-phrase Speaker Diarization Challenge 2022 describes their submission. X-Lance team achieved CDER of 13.2% and 8.0% in the CSSD dev and unseen CSSD eval set respectively. The system developed outputs the ensemble results of the four modules: self-attentive-based VAD, uniform segmentation, ECAPA-TDNN-based embedding extractor, and spectral clustering. This is the same challenge as in paper [12] above. CSSD stands for conversational short-phrase speaker diarization dataset and has three features. CSSD focuses on daily conversation, which contains plenty of spontaneous speech, some even not recognizable. Second, the CSSD dataset contains many short-phrase speeches. Third, metric used is CDER stands for short for conversational diarization error rate, is designated for better evaluating short segments, where the result of DER and JER cannot measure well.

3. CONCLUSIONS

In this paper, we have presented a comprehensive analysis of the speaker diarization approach by describing various feature extraction and feature modelling methods and recent research contributions, highlighting key methodologies, techniques, and performance improvements.

Few take away points are:

- i-vectors and d-vectors are embeddings extracted from speech.
- Compared to near field, far field speech has higher diarization error.
- For clustering RNN trained on in domain and out of domain data performs better.
- NME-based spectral clustering method is competitive in terms of performance, while not requiring any hyper-parameter tuning.
- C-vector with consecutive structure gives the lowest error rate.
- A re-segmentation module can be used to get a lower DER.
- pyAudioAnalysis and pyannote.audio are 2 different python open source library for speech analytics.
- RPNSD system can be applied to the overlapping speech problem.
- GNN can be used to refine embedding before applying clustering.

ACKNOWLEDGEMENT

This literature review has been done after studying speech analytics subject as part of M. Tech in AI. I am thankful to professor Dr. J B Simha for imparting the knowledge.

REFERENCES

1. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
2. S. Horiguchi, P. García, Y. Fujita, S. Watanabe, and K. Nagamatsu, “End-to-end speaker diarization as post-processing,” in ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 7188–7192. doi: 10.1109/ICASSP39728.2021.9413436.
3. IEEE Signal Processing Society and Institute of Electrical and Electronics Engineers, 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) : ASRU 2019 : proceedings : December 15-18, 2019, Guadeloupe, West Indies.
4. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
5. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
6. IEEE Signal Processing Society, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada.
7. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
8. IEEE Signal Processing Society, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada.
9. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.
10. IEEE Signal Processing Society, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada.
11. T. Liu, X. Xiang, Z. Chen, B. Han, K. Yu, and Y. Qian, “The X-Lance Speaker Diarization System for the Conversational Short-phrase Speaker Diarization Challenge 2022,” in 2022 13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 498–501. doi: 10.1109/ISCSLP57327.2022.10037955.
12. B. Pang et al., “TSUP Speaker Diarization System for Conversational Short-phrase Speaker Diarization Challenge,” in 2022 13th International Symposium on Chinese Spoken Language Processing, ISCSLP 2022, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 502–506. doi: 10.1109/ISCSLP57327.2022.10037846.
13. IEEE Signal Processing Society, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada.
14. IEEE Signal Processing Society, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing : proceedings : April 15-20, 2018, Calgary Telus Convention Center, Calgary, Alberta, Canada.
15. T. J. Park, K. J. Han, M. Kumar, and S. Narayanan, “Auto-Tuning Spectral Clustering for Speaker Diarization Using Normalized Maximum Eigengap,” IEEE Signal Process Lett, vol. 27, pp. 381–385, 2020, doi: 10.1109/LSP.2019.2961071.
16. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020.