Deep Embedding Clustering for Speaker Diarization

Aditya Singh
Department of Mechanical Engineering
Indian Institute of Technology Kanpur
adisingh@iitk.ac.in

Shashi Kant Gupta
Department of Electrical Engineering
Indian Institute of Technology Kanpur
shashikg@iitk.ac.in

Vivek Agrawal
Department of Electrical Engineering
Indian Institute of Technology Kanpur
vivekagr@iitk.ac.in

Abstract

Speaker diarization has received considerable interest within the speech community due to its promises to considerably improve automatic speech transcription. Embedding vectors such as d-vectors, i-vectors, or x-vectors clustered using Spectral Clustering are the commonly used approach to this problem. We propose to use of Unsupervised Deep Embedding Clustering, known to provide sharper cluster boundaries and generalize well with data imbalance, to improve the results. Residual AutoEncoders have been used in place of Denoising AutoEncoders for improved learning. We use VoxConverse and AMI Corpus datasets to test our model. Our model shows considerable improved over Spectral Clustering approach. The project can be accessed at https://github.com/shashikg/speaker_diarization_ee698.

1 Introduction and Problem Statement

With the increase in digital media data volumes, speaker diarization finds significant use in cases like multimedia information retrieval, meeting transcription analysis, film and movie transcription, speaker turn analysis, and speaker adaptive processing.

Most of the speaker diarization systems include two crucial modules. First module performs preprocessing on the raw audio signal and extracts high dimensional features known as embedding vector. These embeddings are extracted for multiple sliding window segments of the sampled input audio file. The second module takes these embedding vectors and groups them into different clusters to assign unique speaker labels to each of them [1–5] (See Fig. 1). Advanced diarization systems also use additional modules to handle the cases for speaker overlap and produce refined clusters [6].

Many of the state-of-the-art speaker diarization systems use recent deep learning frameworks to extract the feature embeddings called as d-vectors[1, 2] and x-vectors[3–5], and cluster them using conventional clustering methods like k-means, spectral[7], AHC, and affinity propagation clustering algorithm [8]. Recent developments in robust deep neural networks based clustering method [9] has motivated its application in speaker diarization for improved label assignment. Along with the deep embedding clustering method, we used permutation-free diarization objective [6], originally proposed to train an end-to-end speaker diarization model, to further refine the diarization results on an x-vector system.

2 Baseline

For baseline, we use the algorithm described in [1] to annotate the audio files. It is based upon on x-vector embeddings instead of the originally proposed d-vectors embeddings because of superior performance of x-vector in speaker verification tasks [10].

Prior to x-vector extraction, we separate speech segments from raw audio signal using Silero Voice Activity Detection module [11]. These speech segments are then split into overlapping windows as depicted in Fig 1. SpeechBrain ECAPA-TDNN (Time Delay Neural Network) implementation [12] is next used for embedding features extraction on these windows. It takes time invariant sampled audio window, extracts 80 dimensional Mel Frequency Cepstral Coefficents (MFCC) features, and passes it through pretrained ECAPA-TDNN network to get 192 dimensional x-vector.

The feature stack for all the audio windows are then clustered using three unsupervised techniques namely kmeans with known number of clusters (oracle kmeans), spectral clustering with known number of clusters (oracle spectral clustering), and spectral clustering with number of clusters determined using max eigen-value gap (eigengap spectral clustering).

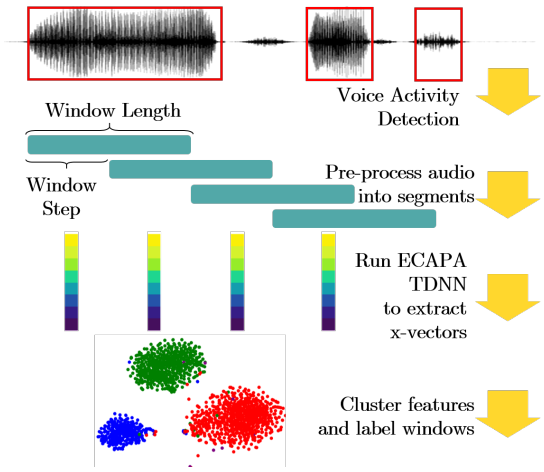

Figure 1: Basic steps involved in ECAPA x-vector based diarization system

Table 1: Diarization performance of our proposed and spectral clustering on different dataset

VoxConverse Dataset	
Method	DER
Spectral	17.76
Ours	12.99
Spectral (Oracle VAD)	17.98
Ours (Oracle VAD)	**11.70**

AMI Dataset	
Method	DER
Spectral	21.99
Ours	23.93
Spectral (Oracle VAD)	14.96
Ours (Oracle VAD)	**13.14**

3 Proposed Model

We used deep embedding based clustering (DEC) [9] analysis to perform diarization on the obtained x-vectors. The motivation behind doing this is that DEC was found to be more robust against imbalance data. Since diarization data used to be highly imbalanced, DEC can substantially benefit in obtaining better diarization results.

At first we trained an autoencoder on the extracted ecapa-x-vectors from the training split of the VoxConverse dataset. The autoencoder was trained using unsupervised reconstruction loss in a residual fashion, in which we forced each of the decoder layers to reconstruct the corresponding encoder layers. The hidden dimensions of the autoencoder was fixed to 500–500–2000–30. Where 30 is the latent space dimension. Post training we extracted the encoder module from the autoencoder and used to extract latent features.

To perform the diarization on a given audio file. The model first extracted ecapa-x-vectors at a step length of 750 ms with window length of 1500 ms. The model first uses those x-vectors to fine-tune the pre-trained autoencoder to get a better reconstruction on that specific audio file. After fine-tuning, those extracted x-vectors were passed through the encoder module to extract the latent space feature vectors. These latent space feature vectors were used to estimate the optimal number of clusters. After finding the optimal number of clusters, those latent space feature vectors were used to further tune the encoder weight to get a better cluster by using a clustering loss, in which the model computes

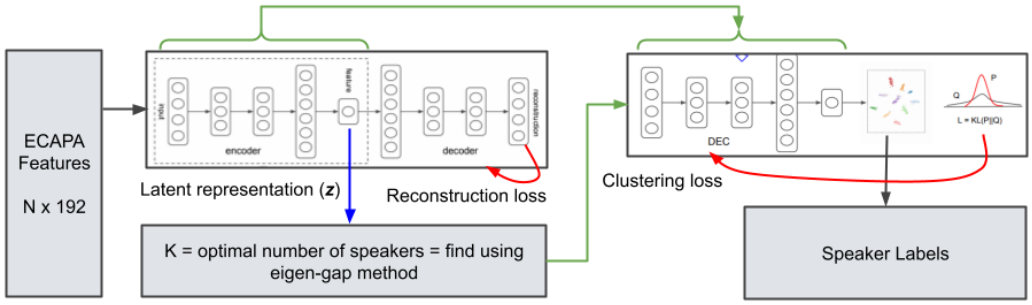

Figure 2: Our proposed clustering method flowchart

a auxiliary target distribution and then uses it minimise the KL divergence loss [cite]. Originally, the DEC paper suggested initializing the centroid point using K-Means, instead of that we used spectral clustering. **Figure 2** shows the overall flowchart of the method.

4 Experiment

4.1 Dataset

We use the VoxConverse [13] and AMI Corpus test split [14] speaker diarization dataset for training and evaluation. VoxConverse dataset consists of 216 multi-speaker audio files with over 20 hours of speech covering more than 8200 speakers. The VoxConverse data is split into two sets of size 166 (train) and 50 (test) for consistent comparison of results with our DEC model that requires training. AMI Corpus test split contains 16 files with a total length of over 13 hours. It is used only for testing.

4.2 Evaluation Criteria

We calculate the Diarization Error rate on the test dataset to evaluate our approach against Spectral Clustering method. The window length is kept at 1500 milliseconds, with window step of 750 milliseconds. We sample the audio at 16kHz and use a batch size of 512. The overlapped region is ignored while calculating the DER since out model is incapable of returning multiple labels for a single window. The collar of 250 milliseconds of the overlapping window is kept at 250 milliseconds. We train the DEC for 150 iterations to minimize the KL Divergence for each audio file using Adam Optimizer with a learning rate of 1e-4.

5 Results and Discussions

Table 1 lists the diarization error rate (DER) for spectral clustering and our proposed method. We have shown the result using both with and without using oracle voice-audio-detection (VAD). All the DER were calculated only for non-overlapping regions with collar of 250 ms, which is a standard evaluation criteria used by most diarization work. When using oracle VAD, our method better performed than spectral clustering for both of the dataset. But when using VAD predicted using the Silero model, our method clearly showed better results on VoxConverse but a slightly poor result on AMI when compared against the spectral clustering method. If we only compare the results on AMI dataset for oracle VAD vs predicted VAD, we get an interesting observation. Since, the predicted VAD is not accurate there will be several data points corresponding to non-speech regions which could definitely interfere with cluster assignment. And since, our proposed method is more robust to imbalance dataset, its quite possible that those non-speech region will have more effect on our method as compared to spectral clustering method. So, this could be a possible issue with using our propose method. But despite that we take an overall look at the performance on both the dataset, our proposed method is a clear winner. Moreover, since our method using oracle VAD produced better DER that means if we can improve the prediction of VAD model then the result for our method will be better than the spectral clustering.

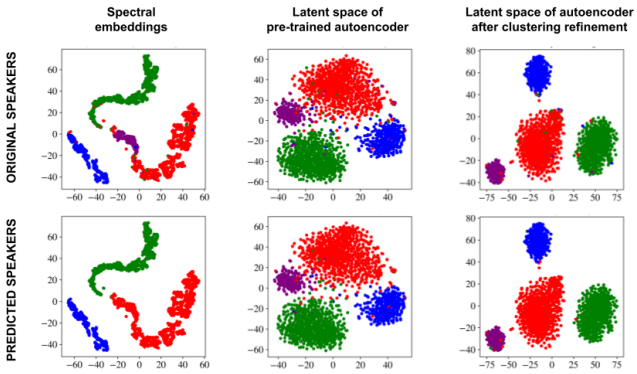

Figure 3: Comparison of t-SNE clusters on spectral embeddings and deep embeddings learned by our proposed method. This example shows a clear example of imbalanced speaker diarization data. Here the purple color speaker is present only for a short duration of time.

We also looked at the number of audio files on which our method gave better DER than spectral clustering. The numbers were surprisingly good for our method. For oracle VAD version, out of 50 audio files in voxconverse our method showed better DER on 35 examples and same DER on 7 examples. On AMI dataset, out of 16 files our method showed better DER on 13 examples and same DER on 1 example. Even for the predcited VAD version, out of 50 audio files in voxconverse our method showed better DER on 32 examples and same DER on 6 examples. On AMI dataset, out of 16 files our method showed better DER on 8 examples and same DER on 1 example.

To take a closer look on the speaker assignment by our method as compared to spectral clustering, we looked at the clusters formed by the spectral embeddings vs deep embeddings formed by our method (refer to Figure 3). The shown example is a perfect example for imbalanced diarization speech segments with all the speakers having different size of speech segments. The spectral clustering merged the smaller purple cluster with the red cluster while our proposed method detected that lower cluster as well. The second coloumn of the figure shows the cluster formed by the initial latent space vectors extracted using the pre-trained encoder while the third column shows the latent space vectors extracted after fine-tuning the encoder module to form better clusters. Clearly the the method improved the cluster formation by forming more clear and concentrated clusters.

6 Acknowledgement

We would like to thank Dr. Vipul Arora for providing us the opportunity and guidance to work on this project as part of the course EE698R.

References

[1] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno. Speaker diarization with lstm. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5239–5243, 2018. doi: 10.1109/ICASSP.2018.8462628.

[2] Li Wan, Quan Wang, Alan Papir, and Ignacio Lopez Moreno. Generalized end-to-end loss for speaker verification, 2020. arXiv: 1710.10467.

[3] D. Garcia-Romero, D. Snyder, G. Sell, D. Povey, and A. McCree. Speaker diarization using deep neural network embeddings. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4930–4934, 2017. doi: 10.1109/ICASSP.2017.7953094.

[4] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018. doi: 10.1109/ICASSP.2018.8461375.

[5] Joon Son Chung, Jaesung Huh, Arsha Nagrani, Triantafyllos Afouras, and Andrew Zisserman. Spot the conversation: Speaker diarisation in the wild. *Interspeech 2020*, Oct 2020. doi: 10.21437/interspeech.2020-2337. URL http://dx.doi.org/10.21437/Interspeech.2020-2337.

[6] Yusuke Fujita, Naoyuki Kanda, Shota Horiguchi, Kenji Nagamatsu, and Shinji Watanabe. End-to-end neural speaker diarization with permutation-free objectives, 2019. arXiv: 1909.05952.

[7] Ulrike von Luxburg. A tutorial on spectral clustering, 2007. arXiv: 0771.0189.

[8] Brendan J. Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007. ISSN 0036-8075. doi: 10.1126/science.1136800. URL https://science.sciencemag.org/content/315/5814/972.

[9] Junyuan Xie, Ross Girshick, and Ali Farhadi. Unsupervised deep embedding for clustering analysis, 2016. arXiv: 1511.06335.

[10] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Interspeech 2020*, Oct 2020. doi: 10.21437/interspeech.2020-2650. URL http://dx.doi.org/10.21437/Interspeech.2020-2650.

[11] Silero Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad, 2021.

[12] Mirco Ravanelli, Titouan Parcollet, Aku Rouhe, Peter Plantinga, Elena Rastorgueva, Loren Lugosch, Nauman Dawalatabad, Chou Ju-Chieh, Abdel Heba, Francois Grondin, William Aris, Chien-Feng Liao, Samuele Cornell, Sung-Lin Yeh, Hwidong Na, Yan Gao, Szu-Wei Fu, Cem Subakan, Renato De Mori, and Yoshua Bengio. Speechbrain. https://github.com/speechbrain/speechbrain, 2021.

[13] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. *Interspeech 2017*, Aug 2017. doi: 10.21437/interspeech.2017-950. URL http://dx.doi.org/10.21437/Interspeech.2017-950.

[14] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner. The ami meeting corpus. In *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research. L.P.J.J. Noldus, F. Grieco, L.W.S. Loijens and P.H. Zimmerman (Eds.), Wageningen: Noldus Information Technology*, 2005.