Supervised and Self-supervised clustering algorithms for Speaker Diarization in the Context of Conversational Speech

Prachi Singh

Department of Electrical Engineering, Indian Institute of Science, Bangalore.
prachisingh@iisc.ac.in

Abstract

Speaker diarization is the task of finding ”who spoke when” in conversational speech recording containing multiple talkers. It is an important first step in analysing conversational speech for tasks like rich speech transcription. The conventional approach to this task involves segmenting the audio into short segments of 1-2s, extracting representations and finally performing unsupervised clustering. In this modular approach, each step is optimized separately based on separate training sets which is completely different from the test set in terms of domain, speakers, background noise etc. Therefore, the method performs suboptimally on the test set. In my work, I hypothesize that self-supervised learning can bridge this gap between clustering and representation learning. I have proposed self-supervised clustering which uses clustering output as pseudo targets and trains a representation learning network . Then the learnt representations are used to update the clustering output. This increases inter-speaker distance and decreases intra-speaker distance between representations called as speaker embeddings. In order to introduce learning based on clustering performance directly, I have proposed supervised hierarchical clustering using graph neural networks which further improves the performance. Our approach shows improvement on various benchmark datasets.

1. Motivation and research questions

Till recent years, speech processing research was more focussed on improving the performance of various techniques for single talker speech recordings. In contrast to this, real world speech/audio recordings feature multiple talkers e.g. in office meetings, restaurant conversations, clinical interviews, audio podcasts, and so on. In such recordings the talkers speak with intonations, there are short turns, and often speech of one talkers overlaps with the other. This has made machine processing, transcription, and interpretation of such audio recordings a challenge. My thesis aims at understanding some of these challenges and designing solutions. Specifically, I have focussed on speaker diarization, that is, automatic segmentation of the given audio recording into regions corresponding to different speakers. We recognize this as a crucial first step in information extraction from conversational speech. The challenges in speaker diarization arise from short speaker turns, background noise, variable number of speakers in natural conversations. The detailed discussion of the recent advancements in the field is given in [1]. Since, deep neural networks requires large amount of training data and compute, I have focussed my research on improving the performance in a low resource setting without requiring huge computation. Therefore, the key research questions of my thesis are: 1) Can we utilize potential of self-supervised learning to improve clustering performance? 2) Can we incorporate metric learning along with representation learning to make the algorithm more robust? 3) Can we train a clustering algorithm using supervision to improve performance on challenging datasets containing large number of speakers? 4) Can we incorporate overlapping speech prediction and multi-speaker prediction at a given time instant? In order to address these questions, we have used recent machine learning and deep learning approaches like representation learning, self-supervised learning and graph based clustering.

2. Methodologies

The dominant approach to speaker diarization involves a two step process [2, 3]. In the first step, the speech recording is divided into short segments and feature vectors are extracted from them. Second step involves forming groups of these feature vectors from different regions of audio using a similarity metric. This process is called clustering. Regions of audio belonging to the feature vectors of same cluster are then assigned with a unique speaker label. In this way, we can map the clusters to their corresponding duration in audio recording. The feature vectors, commonly known as embeddings, are extracted from a deep neural network which is trained to identify speakers from a large dataset. Clustering e.g. agglomerative hierarchical clustering (AHC)[4], Spectral Clustering[5] is performed in unsupervised manner based on the similarity scores. The two steps of the approach are independently done. In my work, I have connected the two steps by taking output of clustering to train the representation learning network and then perform clustering. My initial works are focussed on self-supervised learning to allow training in low resource setting. Later I have introduced supervised clustering to update the embeddings and shown to improve the performance.

1. **Self-supervised Clustering (SSC):** In order to learn robust speaker embeddings, we leverage the clustering output to improve the feature representation. For each recording, we update the representations based on clustering labels so that same cluster feature become more closer compared to different speaker clusters. Then we update clustering results based on the learned representations. This approach referred to as self-supervised clustering (SSC), can provide effective representations for speech recordings containing out-of-set speakers without requiring the actual speaker labels. Another advantage of this approach is that it can be applied separately to each recording and does not require additional training data.
2. **Self-supervised Path Integral Clustering (SSC-PIC):** An extension of this work was to introduce a more robust clustering algorithm called as path integral clustering (PIC) [6]. It is an agglomerative clustering algorithm which encodes the structure of the embedding space in the form of graph

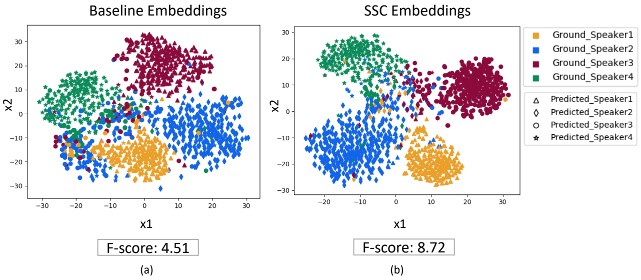

Figure 1: *t-SNE based visualization of baseline and SSC embeddings extracted from the AMI set recording. Higher F-score indicate better separability.*

$G = (V, E)$ where V is the set of vertices/nodes and E is the set of edges connecting the vertices. In our work, the embeddings are the vertices and similarity scores are the edge weights of the graph. PIC forms clusters in the graph, based on nearest-neighbour and then merge two clusters at each step based on the affinity measure between them which is path integral between two clusters. Path integral is the sum of all possible paths in a cluster and measures it's stability. Merging is performed to attain higher stability of cluster. This approach has shown to improve the performance while also being computationally efficient. Using graph based clustering in self-supervised setting, diarization performance is improved over the baseline system by around 59% on meeting dataset and around 13% on telephone conversations.

3. **Self-supervised PLDA PIC (Self-Sup PLDA PIC):** The proposed approach discussed in previous sections uses cosine similarity scores as it is a non-parametric method and does not contain any learnable parameters. However, existing works use Probabilistic Linear Discriminant Analysis (PLDA) scoring [3]. PLDA is a generative model which performs factor analysis on the embedding to extract speaker factor and compute log-likelihood score between a pair of embeddings from a recording. In our work, we introduce a metric learning network along with representation learning inspired from PLDA model. This helps to provide more flexibility to improve the similarity scores.
4. **Supervised Hierarchical Graph Clustering (SHARC)**: One limitation of self-supervised approaches was that it shows degradation in performance as the number of speakers increases in a recording ($>= 7$) because the initial clustering algorithm fails to capture all the speakers. This led us to develop an supervised approach called as supervised hierarchical graph clustering using graph neural networks (GNN). The major contributions of this approach are: 1. Introducing supervised hierarchical clustering using Graph Neural Networks (GNN) for diarization. 2. Developing the framework for joint representation learning and clustering using supervision.

3. Results

The performance is evaluated in terms of diarization error rate (DER) which is the sum of false alarm rate, speaker confusion rate and missed speech rate (lower the better). I have evaluated all my performances on multiple speaker diarization datasets covering different domains e.g. AMI [11], Voxconverse [12], DIHARD [13], CALLHOME telephone datasets.

Table 1: *DER (%) comparison on the AMI datasets with the baseline methods. OVP: overlap, COL: collar, AHC: Agglomerative Hierarchical Clustering, SC: Spectral clustering*

AMI System	**with OVP + no COL**		**w/out OVP + COL**	
	Dev.	Eval.	Dev.	Eval.
Baseline1 with AHC [7]	24.50	29.51	7.61	14.59
Baseline2 with SC	19.8	22.29	4.1	5.76
Prop. SHARC	**19.71**	21.44	**3.91**	4.88
Prop. SHARC+ Refinement [8]	**19.35**	**19.82**	**3.46**	**2.73**

Table 2: *DER (%, w/out overlap + with collar) comparison with state-of-the-art on AMI subset.*

AMI subset System	Dev.	Eval.
x-vec(ResNet101)+AHC+VBx [9]	2.78	3.09
ECAPA-TDNN [10]	3.66	**3.01**
Prop. SelfSup-PLDA-PIC (+Refine.)	5.38 (**2.18**)	4.63 (3.27)
Prop. SHARC (+Refine.)	3.58 (3.72)	**2.29 (2.11)**

The Augmented multi-party interaction (AMI) dataset comprises of meetings recorded in restricted environment containing 3-5 speakers with each recording ranging between 15-60 mins. Figure 1 shows the 2d plot of embeddings before (Baseline) and after (Self-supervised clustering) training. The separability of between speakers have increase compared to baseline which is also indicated using F-score metric.

The results based on the AMI dataset for all the proposed work are shown in Table 1 and Table 2. Table 1 shows that the proposed approach after applying refinement to smooth the clustering output achieved 53% relative improvement over the best baseline. Similarly, Table 2 shows the performance comparison with the state-of-the-art approaches. Our proposed approaches shows significant improvement on the Eval set.

4. Current and future research directions

I am in the final year of my PhD and currently working on overlap speech detection using graph neural networks. In the real conversational speech, multiple speakers often speak at the same time resulting in overlaps. Therefore, identification of the overlapping regions and prediction of the speakers present in the corresponding regions is crucial for diarization systems. Our first step in achieving this is to improve the existing state-of-the-art methods to predict overlapping regions more accurately. We hypothesize that graph neural networks can help in improving the predictions as they are capable of capturing the neighborhood information based on similarity and further bring the representations closer in the latent space. Next step is to predict the two speakers present in the overlapping regions correctly. In order to build a single system, all the steps will be performed using the GNN model.

5. Summary

My work has addressed challenges related to clustering performance. The proposed approaches introduced a robust clustering algorithm. The method also introduced self-supervised learning to perform cluster refinement without requiring additional data. Joint metric learning and representation learning model helps to update the metric to compute better scores. Finally I have introduced end-to-end supervised clustering which performs hierarchical clustering using a single graph neural network model which enables to improve the purity of the predicted clusters representing the speakers and also improves prediction of number of speakers.

6. References

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, “A review of speaker diarization: Recent advances with deep learning,” *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[2] G. S. et. al., “Diarization is hard: Some experiences and lessons learned for the jhu team in the inaugural dihard challenge,” in *Proc. Interspeech 2018*, 2018, pp. 2808–2812.

[3] G. Sell and D. Garcia-Romero, “Speaker diarization with plda i-vector scoring and unsupervised calibration,” in *2014 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 413–417.

[4] W. H. E. Day and H. Edelsbrunner, “Efficient algorithms for agglomerative hierarchical clustering methods,” *Journal of Classification*, vol. 1, pp. 7–24, 1984.

[5] A. Ng, M. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” *Advances in neural information processing systems*, vol. 14, 2001.

[6] W. Zhang, D. Zhao, and X. Wang, “Agglomerative clustering via maximum incremental path integral,” *Pattern Recognition*, vol. 46, no. 11, pp. 3056–3065, 2013.

[7] N. Ryant *et al.*, “The Third DIHARD Diarization Challenge,” in *Proc. INTERSPEECH*, 2021, pp. 3570–3574.

[8] F. Landini *et al.*, “BUT system for the second DIHARD speech diarization challenge,” in *IEEE ICASSP*, 2020, pp. 6529–6533.

[9] ——, “Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks,” *arXiv preprint arXiv:2012.14952*, 2020.

[10] N. Dawalatabad *et al.*, “ECAPA-TDNN embeddings for speaker diarization,” *arXiv preprint arXiv:2104.01466*, 2021.

[11] W. Kraaij, T. Hain, M. Lincoln, and W. Post, “The AMI meeting corpus,” pp. 137–140, 2005.

[12] J. S. Chung *et al.*, “Spot the Conversation: Speaker Diarisation in the Wild,” in *Proc. INTERSPEECH 2020*, 2020, pp. 299–303.

[13] N. e. a. Ryant, “Third dihard challenge evaluation plan,” *arXiv preprint arXiv:2006.05815*, 2020.