ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446231

APOLLO'S UNHEARD VOICES: GRAPH ATTENTION NETWORKS FOR SPEAKER DIARIZATION AND CLUSTERING FOR FEARLESS STEPS APOLLO COLLECTION

Meena M. C. Shekar[1], *John H. L. Hansen*[1]

[1]Center for Robust Speech Systems,
The University of Texas at Dallas, USA

ABSTRACT

Speaker diarization has traditionally been explored using datasets that are either clean, feature a limited number of speakers, or have a large volume of data but lack the complexities of real-world scenarios. This study takes a unique approach by focusing on the Fearless Steps APOLLO audio resource, a challenging data that contains over 70,000 hours of audio data (A-11: 10k hrs), the majority of which remains unlabeled. This corpus presents considerable challenges such as diverse acoustic conditions, high levels of background noise, overlapping speech, data imbalance, and a variable number of speakers with varying utterance duration. To address these challenges, we propose a robust speaker diarization framework built on dynamic Graph Attention Network optimized using data augmentation. Our proposed framework attains a Diarization Error Rate (DER) of 19.6% when evaluated using ground truth speech segments. Notably, our work is the first to recognize, track, and perform conversational analysis on the entire Apollo-11 mission for speakers who were unidentified until now. This work stands as a significant contribution to both historical archiving and the development of robust diarization systems, particularly relevant for challenging real-world scenarios.

Index Terms— Fearless Steps APOLLO, Graph Networks, Speaker Diarization, Historical Archiving

1. INTRODUCTION

Speaker diarization aims to identify 'who spoke when' in an audio stream by segmenting the signal into distinct, speaker-specific clusters. Traditional systems include voice activity detection, followed by speaker embedding extraction. These embeddings are then used for clustering algorithms to group speaker segments and assign speaker labels to each cluster.

Recent advances in the field of Graph Neural Networks (GNNs)[1, 2] have garnered significant attention. However, the application of GNNs in speaker diarization tasks remains relatively underexplored. Notable studies such as Wang et al. [3] and Singh et al. [4] have applied GNNs to speaker diarization. However, their methodologies are primarily validated either on datasets featuring limited numbers of speakers in controlled meeting scenarios or in datasets that require an extensive amount of training data. This highlights a common issue in speaker diarization research where most algorithms rely on datasets derived from simulated environments, meeting scenarios with limited speakers, or clean conditions. There is a need for speaker diarization algorithms that can manage the complexities of real-world team-based acoustic environments without compromising on performance.

Our work focuses on the Fearless Steps Apollo 11 (FSteps) audio corpus, a challenging dataset with over 10,000 hours of mostly unlabeled audio data. The Fearless Steps APOLLO collection will contain all missions, with over 150k hrs of audio and meta-data. Manual annotation for this corpus is an impractical task. The dataset presents challenges such as, overlapped speech, variable acoustic conditions, background noise, and a varying number of speakers per channel ranging from 3 to 65. Each speaker also has highly varying utterance duration [5, 6]. Existing diarization algorithms struggle to perform well under these complex conditions, making the dataset a unique challenging benchmark [7, 8].

Previous work on FSteps corpus [6] focused on the variability in utterance duration and addressing data imbalance by omitting speakers with limited speech duration. Current work here includes all speakers, even those that have limited utterances/duration. We propose a robust network architecture that maintains a high performance even under these challenging conditions.

The main contributions of this study are: [i] introducing a novel framework based on dynamic Graph Attention Networks, which is optimized through data augmentation and weight decay; [ii] Evaluation encompasses both supervised and unsupervised graph edges, relating to speaker classification and diarization respectively: [iii] Tracked and analyzed conversational dynamics of nine key speakers of interest integral to Apollo-11 mission; [iv] Our work stands as one of the first initiatives to identify and extract meta-labels for speakers contributing to historical archiving, and serves as a lasting tribute to the unsung heroes of the Apollo mission.

This work was supported by National Science Foundation under Grant Award number 2016725, J. H. L. Hansen(PI)

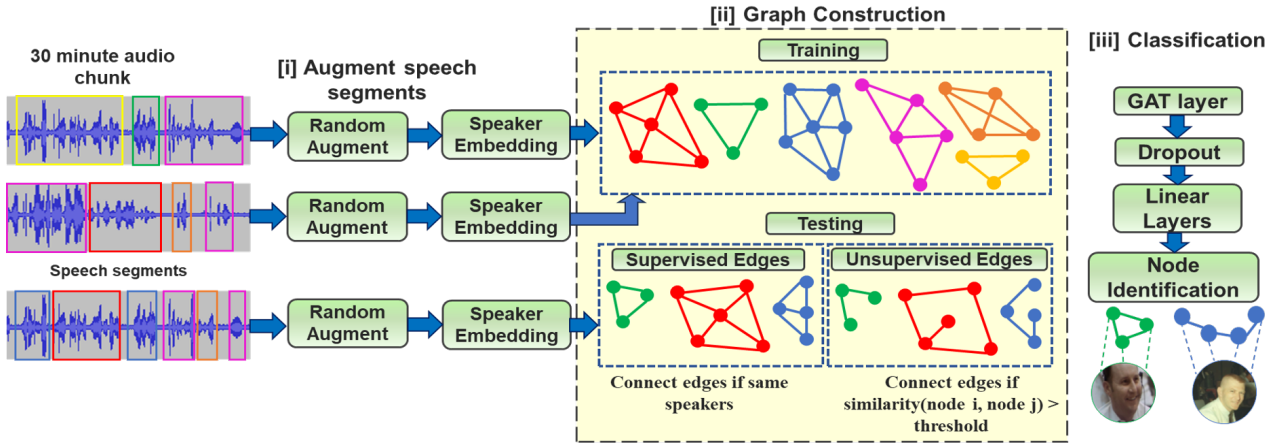

Fig. 1. Flowchart of speaker diarization pipeline with [i] audio augmentation and embedding, followed by [ii] graph construction for train/test phases with supervised and unsupervised edges, and finally [iii] classification via DGAT layer.

2. DATASET DESCRIPTION

For our study, we employ the FSteps A-11 corpus, which contains over 10,000 hours of audio across 30 channels (part of Fearless Steps APOLLO collection, +150k hrs data). Both train and test are conducted on 100-hour labeled subset that includes three mission-critical events: Lift-Off (25 hours), Lunar Landing (50 hours), and Lunar Walk (25 hours). Even though human annotators have manually labeled this dataset, it still has inaccuracies and inconsistencies. To the best of our knowledge, 172 speakers have been identified [9, 10, 11], with 5 out of 30 channels considered. Unlike prior work [6], we include all speakers in the dataset, even those with limited speech utterance duration.

2.1. Data Augmentation

We observe a significant data imbalance, with approximately 30 speakers having fewer than 3 utterances. To address this issue, we propose using data augmentation for each speaker class, wherein 'm' augmentations are applied for each speaker class; in our experiments, 'm' is set to 10. We use the augly toolkit [12] to implement five types of audio augmentations: volume adjustment, reverberation, speed alteration, tempo modification, and pitch shifting. For each randomly chosen audio segment, up to 5 of these augmentations are randomly applied, each with a variable factor. This process is repeated to generate a total of 10 augmented segments for each speaker class.

3. GRAPH STRUCTURE

A graph is defined by its node set $V = \{v_1, ...v_n\}$ and edge set $E \subseteq \{(v_i, v_j) \mid v_i, v_j \in V\}$, where $(i, j) \in E$ denotes an edge from node j to node i. In our context, a single graph corresponds to a 30-minute audio segment. A node v_i is represented by its speaker embedding with F dimensions and $x_i, i \in 1, 2, ..., N$ where N is the total number of nodes. Note that the number of nodes in each graph can vary according to the number of speakers and their utterances. Message passing between neighboring nodes occurs during each iteration via a propagation function f(.) [13].

In training, nodes are connected by edges if they belong to the same speaker class. For test, we employ two types of edges: supervised and unsupervised. In supervised, we have prior knowledge of speaker clusters, and edges are connected between nodes within these predefined clusters. Conversely, in the unsupervised setting, edges are connected between nodes if the cosine similarity between those nodes cross a threshold of 0.65. The Diarization Error Rate (DER) is calculated for all baseline systems. Human annotated SAD (Speech Activity Detection) labels are used here in both train and test phases.

4. BASELINE SYSTEMS

For all our experiments, we use ECAPA TDNN as our speaker embedding, producing a 192-dim vector pretrained on Voxceleb1+Voxceleb2 [14, 15, 16, 17]. Baseline methods include cosine distance, K-means, and Agglomerative Clustering. We also test GNN variants including Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT) [1, 2].

5. PROPOSED SYSTEMS

Dynamic Graph Attention Networks (DGAT) [18] uses a dynamic attention mechanism to model the interaction between nodes in a graph. DGAT inherently provides flexibility wherein the attention mechanism adapts to features and con-

Data Augmentation	Clustering Type	Accuracy	Precision	Recall	F1	DER
No	Cosine Similarity	50.47	69.19	50.47	54.25	42.25
No	K-means	45.50	71.23	45.5	50.66	52.59
No	Agglomerative	52.32	73.99	52.33	57.52	45.35

Data Augmentation	Network Type	Supervised edges	Unsupervised edges				
		Accuracy	Accuracy	Precision	Recall	F1	DER
No	GCN	77.57	62.19	69.88	62.19	61.3	33.6
Yes	GCN	87.48	71.5	76.39	71.5	71.4	25.5
No	GAT	79.09	63.41	72.3	63.41	64.23	32.9
Yes	GAT	89.79	77.61	**80.27**	77.61	77.81	20.3
No	2-layer DGAT [6]	80.11	63.35	73.47	63.49	63	32.7
Yes	2-layer DGAT [6]	88.46	73.26	77.37	73.26	72.76	24.5
Yes	**1-layer DGAT (Ours)**	**90.49**	**78.37**	80.17	**78.36**	**78.2**	**19.6**

Table 1. Comparative evaluation of baseline systems and proposed method. Metrics for supervised edges and unsupervised edges considered. Proposed method shows superior performance

text of the graph. The propagation function is defined as:

$$\mathbf{h}'_i = \sum_{j \in \mathcal{N}i} \alpha_{ij}.\mathbf{W}\mathbf{h}_j, \tag{1}$$

where $\mathcal{N}_i$ is the set of neighboring nodes of node i, and $\mathbf{h}'_i$ is the updated feature representation of node i. Let $\mathbf{h}_i \in R^F$ be the input features of node i and $\mathbf{h}_j \in R^F$ be the input features of a neighboring node j. To learn the node representation for each node, the attention co-efficient function can be given as:

$$\alpha_{ij} = \frac{\exp\left(\overrightarrow{\mathbf{a}}^T \,\text{LeakyReLU}\left(\left[\mathbf{W}\vec{h}_i || \mathbf{W}\vec{h}_j\right]\right)\right)}{\sum_{k \in \mathcal{N}_i} \exp\left(\overrightarrow{\mathbf{a}}^T \,\text{LeakyReLU}\left(\left[\mathbf{W}\vec{h}_i || \mathbf{W}\vec{h}_k\right]\right)\right)}. \tag{2}$$

Attention coefficients compute the weights contributed by each neighbor between pairs of nodes. The attention score e_{ij} on the nodes can then be computed as:

$$e_{ij} = a(\mathbf{W}\vec{h}_i || \mathbf{W}\vec{h}_j). \tag{3}$$

This equation indicates the importance of node j's features to node i. e_{ij} is only computed for nodes $j \in N_i$, where N_i is some neighborhood of node i in the graph. If a node is generally less reliable or not very informative, the attention mechanism, during training, can adjust to have less significance for such nodes. Furthermore, data augmentation can improve the robustness of the model and handle variation between embeddings. The DGAT layer can more accurately differentiate between useful and noisy information. In previous work [6], two-layer DGAT architecture was utilized to overcome limitations of data imbalance by excluding speakers with limited speech duration. However, this omits speaker information from at least 30 speakers. In our current study, we address these limitations by introducing data augmentation and consider all available speakers for training. As a result, we find that a single-layer DGAT leads to both robust and discriminative embeddings. This architecture yields better performance compared to a two-layer approach, as can be seen in Table 1, thereby validating our design choices.

In terms of implementation, our proposed framework leverages a single DGAT constructed using PyTorch Geometric [19]. The input dimension is set to 192, while the hidden dimension is configured at 256, and the number of heads is fixed at 4. Following the DGAT layer, a dropout layer with a dropout rate of 0.3 is added. Furthermore, two linear layers are added: the first has an output dimension of 256 and the second layer with 172. The DGAT layer incorporates an edge softmax function to normalize the attention weights. For the loss function, we employ cross-entropy. To regulate the network and mitigate the risk of overfitting, weight decay is introduced with a factor of 0.0001. To provide a holistic view of our model, we evaluate using multiple metrics. Classification accuracy is measured for both supervised and unsupervised edges which is relevant only for graph-based models. For unsupervised edges and other baseline models, we measure precision, recall, and F1 scores. The Diarization Error Rate (DER) is also measured. All experiments are conducted using human annotated SAD.

	Sup Edges	Unsup Edges				
Num of Augs	**Acc**	**Acc**	**Prec**	**Recall**	**F1**	**DER**
None	81.60	65.58	74.59	65.58	65.86	31.9
5	89.90	77.12	79.11	77.12	77.08	21.2
10	*90.49*	**78.37**	**80.17**	**78.36**	**78.2**	**19.6**
20	89.33	76.49	78.64	76.49	76.32	22.1
30	90.25	77.25	78.48	77.25	76.52	21
Parameter	**Acc**	**Acc**	**Prec**	**Recall**	**F1**	**DER**
No dropout	88.38	75.62	77.75	75.62	75.7	22.3
No weight decay	**90.79**	77.66	79.74	77.66	77.34	20.4
Hidden dim	**Acc**	**Acc**	**Prec**	**Recall**	**F1**	**DER**
128	90.33	76.79	79.08	76.79	76.51	21.09
512	90.71	78.01	79.83	78.01	77.8	19.9

Table 2. Ablation studies on supervised and unsupervised edges, examining variations in augmentations, network architecture, and hidden dimensions

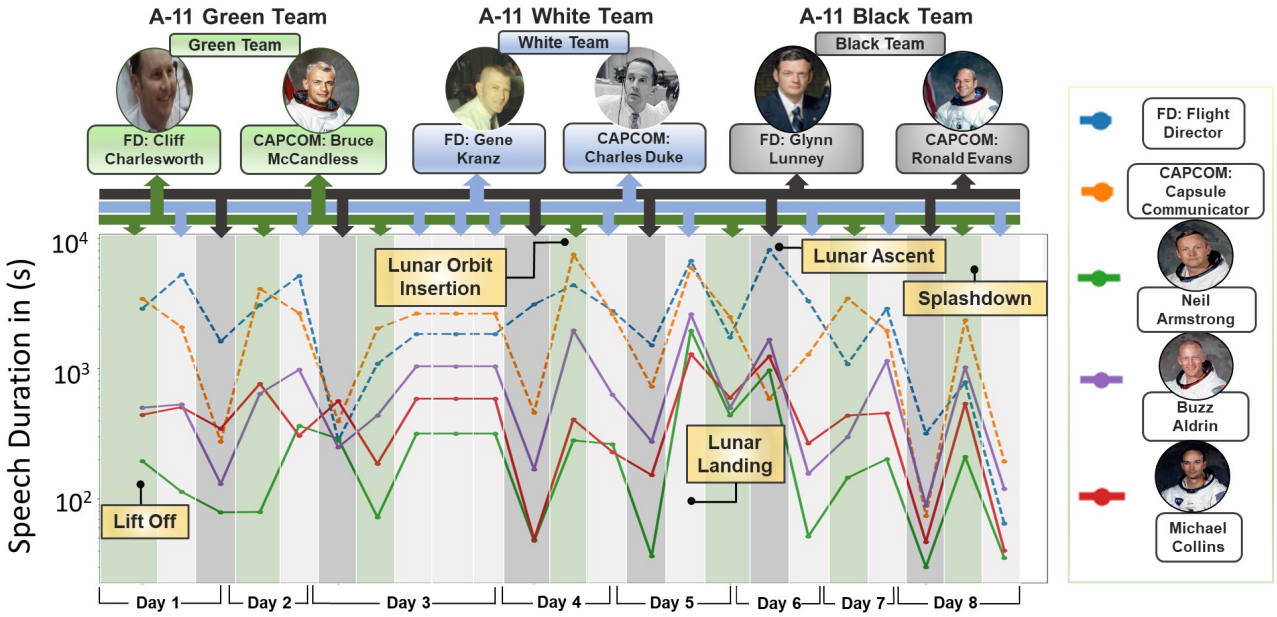

Fig. 2. Tracking nine key speakers over entire Apollo 11 mission, with background colors indicating active Apollo 11 teams and their corresponding Flight Director and CAPCOM

6. RESULTS

The challenge posed by FSteps A-11 corpus is the imbalanced distribution of speaker utterances and their varying utterance duration. Utilizing data augmentation can help in alleviating this imbalance. Hence, we measure performance with and without data augmentation even in baseline systems to show the effectiveness of data augmentation. Table 1 shows that our method outperforms traditional clustering systems and popular graph networks in terms of performance. We conduct ablation studies with our initial experiments focusing on the number of augmentations per speaker class, ranging from 5 to 30. Although there are small variations, we see that 10 augmentations per speaker class yields the best overall performance. The model is also assessed based on different hidden dimensions. While variations in hidden dimensions do not have a significant impact on results, using 256 hidden dimensions offers a slight performance advantage. A stark difference is observed when data augmentation is excluded, reaffirming its importance for enhancing the model in both our proposed and baseline systems. We also found no difference in performance when we varied the number of heads in the DGAT layer.

To extract speaker identities, it is important to understand Apollo-11 communications within the team. The Apollo-11 mission involved four teams: Green, White, Black, and Maroon, each taking time shifts over the 8 day mission with their corresponding set of speakers. Our research focuses exclusively on the labeled data generated from the Green, White, and Black teams. We use our proposed model to generate meta-labels for nine speakers of interest: three Flight Directors, three CAPCOMs[1], and three Astronauts, each originating from one of three teams(except Astronauts as they are present throughout 8 day mission). In Fig 2, active teams were highlighted with corresponding background color during each interval over 8 days.

For the purposes of this analysis, we generated meta-labels and only included those that met a high confidence threshold according to our model. As most of the data is completely unlabeled, we use a SAD system trained on the FSteps corpus [20]. Additionally, we have annotated periods of elevated speech duration, which coincide with key moments during the mission. Finally, all curated audio, metadata, and models will be shared on [21]

7. CONCLUSION

This study introduced a novel speaker diarization framework that can address challenges present in the naturalistic FSteps corpus. To handle data imbalance, we proposed data augmentation and leveraged a DGAT network which achieved a DER of 19.6%. Unlike previous work, our approach incorporates all speakers and makes broader analysis feasible. Moreover, we identified, tracked and analyzed nine key yet previously unidentified speakers throughout the entire Apollo-11 mission contributing to advancements in historical archiving and serving as a tribute to the unsung heroes of Apollo

[1]Note, CAPCOM is the only NASA specialist authorized to communicate with the astronauts

8. REFERENCES

[1] Thomas N. Kipf and Max Welling, "Semi-supervised classification with graph convolutional networks," *Inter Conf on Learning Representations*, 2017.

[2] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, Yoshua Bengio, et al., "Graph attention networks," *Statistics*, vol. 1050, no. 20, pp. 10–48550, 2017.

[3] Jixuan Wang, Xiong Xiao, Jian Wu, Ranjani Ramamurthy, Frank Rudzicz, and Michael Brudno, "Speaker diarization with session-level speaker embedding refinement using graph neural networks," *IEEE ICASSP*, pp. 7109–7113, 2020.

[4] Prachi Singh, Amrit Kaul, and Sriram Ganapathy, "Supervised hierarchical clustering using graph neural networks for speaker diarization," *IEEE ICASSP*, pp. 1–5, 2023.

[5] Meena Chandra Shekar, "Knowledge based speaker analysis using a massive naturalistic corpus: Fearless steps apollo-11," *Master's Thesis, The University of Texas at Dallas*, 2020.

[6] Meena M. C Shekar and John H.L. Hansen, "Speaker Tracking using Graph Attention Networks with Varying Duration Utterances across Multi-Channel Naturalistic Data: Fearless Steps Apollo-11 Audio Corpus," *ISCA INTERSPEECH*, pp. 1459–1463, 2023.

[7] Aditya Joglekar, Seyed Omid Sadjadi, Meena Chandra-Shekar, Christopher Cieri, and John H.L. Hansen, "Fearless Steps Challenge Phase-3 (FSC P3): Advancing SLT for Unseen Channel and Mission Data Across NASA Apollo Audio," *ISCA INTERSPEECH*, pp. 986–990, 2021.

[8] Meena M Chandra Shekar and John H.L. Hansen, "Historical Audio Search and Preservation: Finding Waldo Within the Fearless Steps Apollo 11 Naturalistic Audio Corpus ," *IEEE Signal Processing Magazine*, vol. 40, no. 3, pp. 30–38, 2023.

[9] John H.L. Hansen, Aditya Joglekar, Meena Chandra Shekhar, Vinay Kothapally, Chengzhu Yu, Lakshmish Kaushik, and Abhijeet Sangwan, "The 2019 inaugural Fearless Steps Challenge: A giant leap for naturalistic audio," *ISCA INTERSPEECH*, pp. 1851–1855, 2019.

[10] Aditya Joglekar, John H.L. Hansen, Meena Chandra Shekar, and Abhijeet Sangwan, "FEARLESS STEPS Challenge (FS-2): Supervised Learning with Massive Naturalistic Apollo Data," *ISCA INTERSPEECH 2020*, pp. 2617–2621, 2020.

[11] John H.L. Hansen, Aditya Joglekar, Szu-Jui Chen, Meena Chandra Shekar, and Chelzy Belitz, "Fearless Steps APOLLO: Advanced Naturalistic Corpora Development," *LREC*, pp. 14–19, 2022.

[12] Zoe Papakipos and Joanna Bitton, "Augly: Data augmentations for robustness," *arXiv:2201.06494*, 2022.

[13] Jiaxuan You, Jure Leskovec, Kaiming He, and Saining Xie, "Graph structure of neural networks," *Inter Conf on Machine Learning*, pp. 10881–10891, 2020.

[14] Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck, "ECAPA-TDNN: emphasized channel attention, propagation and aggregation in TDNN based speaker verification," *ISCA INTERSPEECH*, pp. 3830–3834, 2020.

[15] Mirco Ravanelli et. al, "SpeechBrain: A general-purpose speech toolkit," 2021, arXiv:2106.04624.

[16] Hervé Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," *ISCA INTERSPEECH*, August 2017.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[18] Shaked Brody, Uri Alon, and Eran Yahav, "How attentive are graph attention networks?," *arXiv preprint arXiv:2105.14491*, 2021.

[19] Matthias Fey and Jan Eric Lenssen, "Fast graph representation learning with pytorch geometric," *arXiv preprint arXiv:1903.02428*, 2019.

[20] Aditya Joglekar and John HL Hansen, "DeepComboSAD: Spectro-Temporal Correlation based Speech Activity Detection for Naturalistic Audio Streams," *IEEE Signal Processing Letters*, accepted Sept 8, 2023.

[21] CRSS-UTDallas community resource website, "Explore Apollo Website," https://app.exploreapollo.org/. Accessed: 2023-09-12.