Evaluation of System for Monaural Speaker Diarization

Finnerty, Brian
bpf7056@rit.edu
Rochester Institute of Technology
USA

Matthews, Wiley
wsm8855@rit.edu
Rochester Institute of Technology
USA

Tran, Dat
dqt4873@rit.edu
Rochester Institute of Technology
USA

Wood, Dade
daw1882@rit.edu
Rochester Institute of Technology
USA

ABSTRACT

With over 4.2 billion voice assistants in the world as of 2020, there has been an ever-growing desire to capture and utilize the increasing amount of voice data shared with these systems. One such method which has been invaluable for identifying speakers in audio recordings is speaker diarization. Speaker diarization, a major part in automatic speech recognition (ASR), attempts to answer the question of "Who Spoke When". With many available and developed systems that handle this task, one major question that we pose is the accuracy of these models on increasingly short and complex audio snippets.

This paper performed a rigorous evaluation on one such diarization model: PyAnnote [3]. As a popular open source diarization library that comes with pretrained models, PyAnnote was the perfect choice to evaluate as a benchmark. The paper also provided a new dataset generation tool for diarization training and testing through the use of mixing audio files and background noise. Leveraging the the LibriSpeech dataset for data generation, the paper showed that PyAnnote could achieve a Diarization Error Rate (DER) of 25% for the simplest test case and a 41% DER for the shortest and noisiest test case. Overall, the paper provided a comprehensive system to develop and evaluate speaker diarization systems.

KEYWORDS

Speaker Diarization, Natural Language Processing, Automatic Speech Recognition, Cognitive Computing, PyAnnote

1 MOTIVATION

Speaker Diarization is a difficult problem that concerns the question of "who spoke when" in an audio clip. It handles the challenging task of taking an audio recording, often of multiple speakers with overlapping and non-overlapping voice segments, and providing labeled segments for each respective speaker. While there has been significant progress in the field over the past few decades [19], one problem for speaker diarization lies in a monaural audio settings. This is when there are multiple speakers each sharing a singular microphone. Taking a step back, Speaker Diarization is just one piece of the puzzle for an Automatic Speech Recognition system in which speech is identified in a sound clip, segmented and clustered through diarization, and finally translated into text which can be reviewed later.

Our team found its inspiration for this project when we recognized that services like Zoom and Google's transcription app had a difficult time in labeling who was speaking in any given setting. Although Zoom does have some level of speaker diarization, it struggles when multiple speakers are talking and can only be done for a cloud recording. This fails even further for us in settings with multiple microphones, which should be capable of separating and transcribing the audio data. For students that are hard of hearing it may be difficult to keep up with group meetings when a translation app cannot label the flow of conversation correctly or take too long to do so. To this end, speaker diarization systems must be both robust and efficient to maximize their usefulness in these settings.

Our project has thus settled on validating the claims that other trained models purport. While many papers take the time to push the boundaries of this field with increasingly more complex networks and models, we aim to take a step back to confirm previous findings. To achieve this end we present a similar testing method to that of Kinoshita et al. [14] to provide varying levels of mixed and overlapped voice data along with background noise to evaluate PyAnnote's ability to separate and label the different segments.

The rest of our paper will follow this structure: Section 2 will take a deeper look into the related work surrounding speaker diarization and what is popular in the field thus far. Section 3 will discuss the design of our corpus, the diarization model we are using, and how we will be constructing simulated test datasets to evaluate all the capabilities of said model. Section 4 will make a comprehensive look on how the model performed on our custom built datasets and evaluate if the model can sufficiently handle the scenarios they present. Section 5 and 6 will review some ethical and legal concerns surrounding diarization systems and why there is a pressing need to further investigate this field. Section 7 will discuss how we wish to advance our tool along with our testing methods. Finally, section 8 will cover our final thoughts and the challenges we faced.

2 RELATED WORK

Park et al. [19] provided an overview of the progress and methods of speaker diarization since the early 2000s to the present. This included both descriptions of traditional X-Vector algorithms and more recent advancements with neural networks. A traditional diarization system receives audio data, detects speech activity, segments the data, and finally performs clustering to identify speakers. Modern implementations trend towards the increased incorporation of neural models and strive to do away with explicit modality altogether; instead aiming to achieve end-to-end optimization via machine learning methods.

Ephrat et al. [4] took a deeper look into diarization by presenting an algorithm that utilized both audio and visual data to perform speech separation. Once they noticed that the human mind had a boosted level of auditory attention when visually focused on the speaker, the authors replicated this effect through the use of audio clips being paired with their respective videos. It built upon traditional diarization as it incorporated visual data through the use of pre-trained facial recognition software, which was then paired with audio features and placed through a convolutional network.

Kanda et al. [13] found that a joint approach to speaker-attributed automatic speech recognition (SA-ASR) had the potential to significantly outperform a modular approach and had also outlined the remaining issues in SA-ASR, such as the improvement of robustness in future models. Kanda et al. also addressed the issue of performing end-to-end (E2E) SA-ASR in cases where the system had no pre-existing profiles of the speakers present and where there were cases of overlapped speech [12]. It built upon prior work, which proposed a joint model of speaker counting, speech recognition and speaker identification, to address the unknown speaker case and modified the training of prior systems to better allow for identification of new speakers via clustering.

These works were useful to our own by providing a starting point to understand the design of a diarization system. With many of the current models being black box neural networks, this information was critical in selecting a sufficient model to evaluate that would provide a great starting point for future work.

Yuhi et al. documented AISHELL-4, a dataset tailored to the speech separation and diarization use cases [7]. The recorded data is 128 hours of spoken Mandarin in a conference setting. The dataset was sourced from real meetings, so many elements of real conversation were present such as short pauses, speech overlap, quick speaker change, and ambient noise. The work also included a PyTorch-based model trained on the dataset to promote reproducible research.

Takashima et al. [22] proposed an end-to-end neural speaker diarization (EEND) that was conditioned on specific subtasks, such as speech activity detection and overlap detection, in order to solve the easier subtasks first and find potentially relevant information for the ultimate speaker diarization task. This work is another potential starting point to compare with different diarization models. The idea of decomposing the problem into a simpler subtasks to separate out some of the work could be a great improvement that can provide an insightful comparison between EEND and more self-contained systems.

Zhang et al. was a somewhat older (2019) example of an online speaker diarization approach that took steps towards the more recent E2E approaches, which unified previously distinct components of the diarization process and made use of supervised learning, rather than clustering when labeled data was available [25].

Horiguchi et al. built on the work of Takashima et al. [22] through the use of unsupervised methods which allowed for E2E neural diarization on an unlimited number of speakers [11]. The paper noted that while progress had been made towards enabling neural diarization approaches to handle an arbitrary numbers of speakers, they were sill largely limited by the supervised nature of their training process and hence could only handle a maximum number of unknown voices. This work used global and local attractors in combination with speaker embedding and unsupervised methods to address this problem. This work was useful as it provided yet another approach that tackled this complex problem. In future work it would be great to perform a more comprehensive review of different diarization models.

3 DESIGN

PyAnnote utilized a wide range of features to generate a comprehensive neural system [3]. PyAnnote provides an extensive library that lays the basis for the construction of a diarization system through multiple building blocks. These building blocks, that consist of feature extraction, voice activity detection, speaker embedding, and clustering, were jointly combined which provides higher optimization. The basic implementation of the PyAnnote trained model was based on the following: first, waveform data is fed into the convolutional layers for feature extraction; afterwards, the data is fed into a recurrent layer and then into a feed-forward layer with no pooling, then a softmax activation function is applied to provide a final classification for speaker segments.

3.1 Corpus

For our corpus we develop four primary datasets to evaluate the PyAnnote diarization architecture outlined in Section 3. These datasets are generated from the LibriSpeech data [18], a collection of book reading audio snippets from thousands of unique individuals. Each piece of validation data consists of audio mixtures that are 10 seconds long, and the available list of speakers for each utterance have an even split between gender. With this, we apply techniques taken from Kinoshita et al. [14] to produce varying levels of overlap between speakers as well as our own values. Background noise is also added into the sound clips using files from Musan noise dataset [21] to better reflect a real world setting. This variety of overlapped data along with the multitude of controlled variables produces rigorous datasets for us to validate the results of PyAnnote with simulated real world data. Our selection of LibriSpeech is driven by the nature of each audio file containing a single speaker and corresponding speaker metadata, allowing for the best creation of mixtures simulating real world audio.

3.2 Experiment

Using our proposed corpus creation method, we create four validation datasets of 100 , 10 second-long mixtures each; Dataset A with few speakers, long segments, and medium probability of noise; Dataset B with more speakers, long segments, and medium probability of noise; Dataset C with many speakers, short segments, and low probability of noise; and finally Dataset D with many speakers, short segments, and high probability of noise. Datasets A and B use speaker probabilities from [14] and datasets C and D use our own probabilities. A more detailed breakdown of each dataset can be found in Table 1.

To enable the creation of the datasets described, we develop a dataset creation tool capable of creating sample datasets with specified statistics. In the spirit of Kinoshita et al. [14], this tool creates audio mixtures (audio files populated with known speakers and noise) from raw noise and speech audio files. Since these files are generated programmatically, all of the speaker activity in the

Dataset	Composition	DER
A	Two 5s segments, with noise probability of 10% each. Testing with 0-2 speakers talking in long segments.	25.32%
B	Two 5s segments, with noise probability of 10% each. Testing with 0-3 speakers talking in long segments.	32.24%
C	Four 2.5s segments, with noise probability of 5% each. Testing with 0-4 speakers talking in short segments with low noise confusion.	40.97%
D	Four 2.5s segments, with noise probability of 20% each. Testing with 0-4 speakers talking in shorter segments with high noise confusion.	41.41%

Table 1: Dataset composition and resulting Diarization Error Rate (DER).

file is known. This allows us to create fully-labeled experimental datasets of specified statistics from a corpus of unlabeled audio data. The labeling format used for speaker diarization data is the Rich Transcription Time Marked (RTTM) file format defined by NIST [9]. We can then use these datasets for both training and evaluating diarization pipelines.

The proposed datasets are created and utilized to evaluate the effectiveness of PyAnnote. After generating the LibriSpeech validation data from 3.1, we then download the pre-trained model and utilized PyAnnote's pipeline system to evaluate the model. The validation data is broken down into the balanced subsets mentioned before, with multiple levels of overlapped audio data. With the metadata provided from LibriSpeech and our own metadata saved from the dataset creation process, we have an accurate count of the number of speakers for each segment, the presence of noise, and the paths to the actual files used in each mixture, so we can compare the model's performance on each mixture to it's metadata and analyze which settings the model did good or bad on.

The metric used to measure the model's performance on individual mixtures is the Diarization Error Rate (DER) [8]. DER is the most common metric used to evaluate the accuracy of speaker diarization. DER can be understood generally as a metric which measures the percentage of time that the system misassigns speech-speaker [1]. Such misassignments can be in the form of false alarms (predicting there was speech when there was not any), misses (predicting there was no speech when there was) and confusion (predicting the wrong speaker was speaking) [1]. This metric can be understood abstractly in Equation 1:

$$DER = \frac{False\ Alarms + Miss + Confusion}{Reference\ Length} \tag{1}$$

and formally in Equation 2:

$$DER = \frac{\sum_{s=1}^{S} dur(s) \cdot (max(N_{ref}(s), N_{hyp}(s)) - N_{correct}(s))}{\sum_{s=1}^{S} dur(s) \cdot N_{ref}} \tag{2}$$

where S is the total number of speaker segments, $N(s)$ represent the number of speaker speaking in that segment [1]. The subscript refers to counts of a particular type such as reference, hypothesis, and correct assignment (reference = hypothesis). The resulting DER from evaluation on each dataset can be seen in Table 1.

DER Statistic	PyAnnote Benchmarks	Our Testing
N Datasets	9	4
Maximum	32.4%	41.4%
Minimum	8.2%	25.3%
Average	20.8%	35.0%
Std. Dev.	7.8%	6.7%

Table 2: PyAnnote.audio benchmark statistics (not finetuned) [2] compared to our testing.

4 ANALYSIS

Our testing reveals varied error rates. The rates vary both between our simulated test datasets and collectively from PyAnnote's benchmarks. The variance with regard to the test datasets provides insights into the strengths and weaknesses of the model in regard to diarization accuracy as well as provides material for conjecture regarding why our results differ from PyAnnote's benchmarks. The variance with regard to PyAnnote's benchmarks reveals that one may expect significant variance in performance when applying the model to non-benchmark datasets.

4.1 Results Comparison Between Test Datasets

As can be observed in Table 1, there are some differences in DER between each of the test datasets. Based on comparison between the composition of each dataset and its resulting error rate, one can identify a potential explanation for the rise in error rate between datasets: the number of speakers.

Before addressing the number of speakers as the cause for the change in error rate, let us discuss other factors that could be contributors, but likely are not as responsible for the change. First, segment length *could* be a factor. One can notice that shorter segment lengths tend to be correlated with higher DER. However, there is a noticeable change between DER for datasets A and B despite having the same segment length while there is only a small change between datasets C and D. Another factor similar to segment length is the number of segments. One will notice that this feature of the datasets follows the same pattern as segment length and ultimately does not explain DER as well as the number of speakers. Finally and surprisingly, background noise seems to have little effect on the diarization results, as the results show very small correlation between noise probability and DER.

This leads us to the number of speakers. The most significant difference between error rates seems to be most directly attributable

to the maximum number of speakers which could appear in the dataset. This is evidenced by the apparent pattern in the results: for every additional speaker which can appear in the dataset, the DER increases by 7%-8%. To reinforce this idea further, notice that datasets C and D have similar number of speakers and DERs despite having the widest gap in noise probability among the datasets. This accuracy issue may be explained by the idea that the complexity of the problem grows non-linearly with the number of speakers. As the number of unique speakers grows, so does the chance for the model to misidentify one speaker as another. The bulk of the problem shifts from that of speaker detection, as in the case of only one or two speakers, to the true task of speaker diarization: accurately and consistently discriminating between human voices.

4.2 Results Comparison to PyAnnote's Benchmark Datasets

As can be observed in Table 2, there is difference in DER statistics between model performance on PyAnnote's benchmark datasets and our simulated test datasets. This provides the opportunity for analysis about potential reasons for these performance differences to find possible flaws in our methodology.

First, the nature of our simulated datasets could be very different from actual real world data and the benchmark datasets PyAnnote tests on. For example, in a real world meeting it can be very rare for more than 1-3 speakers to be speaking at one time and usually it is only for a short duration of the full meeting. In our datasets, however, we have the possibility for many speakers to be speaking throughout the entire audio file which makes the simulated situation much more difficult than real world situation.

Second, the length of our mixtures and segments could be too short when compared to real world data. The length of each of mixture is at maximum 10s and the smallest segment we have is 2.5s. Because of this, our simulated data contains a high variety of speakers switching in and out over fairly short periods of time which could be more confusing than most real world scenarios. It should also be noted that our datasets are fairly small, each containing only 100 mixtures, and it could be the case that more data would result in a smaller average DER as the dataset would contain a larger pool of our controlled randomization.

Third, our audio file sampling method could be affecting the model's speaker detection. For each segment in a mixture we create, a random audio file for a particular speaker is chosen, even if it is the same speaker from the previous segment. This means that even if a speaker is included in all segments of the mixture, it is possible that each segment contains a different audio file from the speaker. Each audio file can vary in speed of speaking, tone, and other factors so this randomization could be another point of confusion in the model, especially as our segments can start and end abruptly.

5 ETHICAL CONSIDERATIONS

The ethical considerations for speaker diarization extends far beyond our work in this project. As voice activated personal assistants are becoming more and more prevalent in today's society, there are a growing number of concerns from users that their conversations are being recorded outside of the activation-command phrase. Although these concerns are unfounded, it still begs the question if the millions of microphones in the world are acting in an individuals best interests. Speaker diarization attempts to answer the question of "Who Spoke When" in an audio recording, and to do that it attempts to learn patterns of an individual's speech. This pattern recognition is how it is able to provide a label for which section was spoken in a recording.

The human voice has been found to act uniquely like a fingerprint [23], and can provide a plethora of information along with it. This information includes socioeconomic status [15], gender, mental health issues [10], and many more. With so much personal data that can possible be tracked through voice identification alone, it is no wonder those concerns plague the sphere of recording information.

With many personal assistants, like Google assistant, having the ability to learn your voice in-order to identify which user is making a request there is a stronger need for some form of standardization between applications. While the ACM Code of Ethics sets a goal to respect the privacy of users, it begs the question of when privacy can be reasonably obtained. In a real world scenario, any number of voices of passerby's could lumped into the background noise of a system and trained upon. While the first concern for those creating or evaluating speaker diarization systems would be to prevent background noise from producing another user label, a second concern would be to ensure that a reasonable expectation of is followed. Although it is unreasonable to assume that a public setting would have as much privacy as being in your own home, there is a genuine concern for personal information to be trained upon if you were a user or someone in the background.

Another ethical grey area is the necessity of informed consent and YouTube's privacy policy in the context of using datasets such as VoxCeleb for training systems. Traditionally in human subject research it is ethically and legally required to provide participants with informed consent, a way to give participants the full information of the study, their role, any related risks to participating; however, YouTube videos are considered publicly available and do not constitute as human subject research, even when utilizing biometric data like voice and facial features.

All of this culminates into requiring further investigation into the proper practices for recording, storing, and managing personal voice data. As diarization models continue to intertwine with everyday society, there needs to be a stronger consensus on acceptable practices.

6 LEGAL CONSIDERATIONS

Legally, our considerations for this project is an extension of the ethical issues above. There are various laws made to protect the private data of users and customers, and that of course include their voices: the General Data Protection Regulation (GDPR) in Europe, the Biometric Information Privacy Act (BIPA) in Illinois, California Consumer Privacy Act (CCPA), along with others. Given the analysis that could be achieved from voice data [23], there area many legal concerns that are not unfounded. On top of the ethical issues about monitoring and tracking people, security breaches could potentially cause greater damage if companies identify users and store the private data across different services.

Some companies have been brought to court for their practices in this matter, such as Google Assistant itself being found to have humans reviewing the voice data it collected. In Europe, the GDPR legally requires that phone companies have to provide an opt out for users in order to collect voice data for model training. Still, it's hard to assure people that their voices are not being used beyond what the terms and conditions say, such as for targeted marketing.

Specifically concerning our project, we will be working mainly with publicly available corpora from the internet, whose speakers (including us) have given their consent for their voices to be used in training models. We will be identifying speakers only to the extend that the corpora divulge. In the scenario that we collect voice clips from other people outside our group to test the models, a robust Privacy Policy and Consent Agreement would be written to ensure transparency of our work and non-disclosure of their identities. In reference to the VoxCeleb dataset, there is an opt out form available on their website to remove the audio-visual information of a celebrity, although this does nothing to protect them from older downloaded version of the dataset existing online.

7 FUTURE WORK

With the promising results that we have obtained thus far, we hope to continue to make progress towards creating a rigorous test suite for diarization models. One approach that we would like to pursue is to utilize a more diverse set of metrics than solely DER. While DER is currently the most popular way to evaluate these systems, other errors rates like the Jaccard Error Rate [20] or Balanced Error Rate [16] would be interesting factors to utilize.

In addition, the number of models we analyzed was limited solely to PyAnnote, but this could be done for a much broader range of systems. We aim to extend our work by performing an evaluation on a broader scope of popular open source models. This could also be extended to training and evaluating our own models using this data to compare to common benchmarks.

Finally we hope to expand the capabilities of our diarization dataset creation tool. This includes generalizing the tool to use other base datasets besides LibriSpeech, adding a flagging system to automatically identify files that a model does well or poorly on for easier analysis, and modifying the way each segment in a mixture selects a speaker's audio file to better reflect real world scenarios.

8 CONCLUSIONS

In this paper, we incorporate a new tool to produce simulated datasets with varied noise level and number of speakers to evaluate the current state-of-the-art diarization systems. With these simulated datasets, we were able to conclude that PyAnnote could not reach the same level of DER as the purported benchmarks. While this is the case, we contribute much of this loss to the uniqueness and difficulty of our simulated data due to the number of overlapped speakers at one time, how we sample random segments from clips, and the length of each sampled segment. With all of these in mind, we plan to continue to refine our dataset generation tool to produce more realistic and rigorous audio clips for diarization evaluation.

REFERENCES

[1] Xavier Anguera, Chuck Wooters, and Javier Hernando. 2007. Acoustic Beamforming for Speaker Diarization of Meetings. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 7 (2007), 2011–2022. https://doi.org/10.1109/TASL.2007.902460

[2] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*. Brno, Czech Republic.

[3] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. Barcelona, Spain.

[4] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. *ACM Trans. Graph.* 37, 4, Article 112 (jul 2018), 11 pages. https://doi.org/10.1145/3197517.3201357

[5] Hugging Face. [n.d.]. *pyannote.audio*. https://github.com/pyannote/pyannote-audio Accessed: 2022-12-01.

[6] Heather Foti. [n.d.]. *Human Subjects Research Office*. Rochester Institute of Technology. https://www.rit.edu/research/hsro/ Accessed: 2022-12-02.

[7] Yihui Fu, Luyao Cheng, Shubo Lv, Yukai Jv, Yuxiang Kong, Zhuo Chen, Yanxin Hu, Lei Xie, Jian Wu, Hui Bu, Xin Xu, Jun Du, and Jingdong Chen. 2021. AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario. In *Proc. Interspeech 2021*. 3665–3669. https://doi.org/10.21437/Interspeech.2021-1397

[8] NIST Speech Group. 2007. Spring 2007 (rt-07) Rich Transcription Meeting Recognition Evaluation Plan.

[9] NIST Speech Group. 2009. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan.

[10] Nik Wahidah Hashim, Mitch Wilkes, Ronald Salomon, Jared Meggs, and Daniel J France. 2017. Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice* 31, 2 (2017), 256–e1.

[11] Shota Horiguchi, Shinji Watanabe, Paola García, Yawen Xue, Yuki Takashima, and Yohei Kawaguchi. 2021. Towards Neural Diarization for Unlimited Numbers of Speakers Using Global and Local Attractors. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 98–105. https://doi.org/10.1109/ASRU51503.2021.9687875

[12] Naoyuki Kanda, Xuankai Chang, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2021. Investigation of End-to-End Speaker-Attributed ASR for Continuous Multi-Talker Recordings. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. 809–816. https://doi.org/10.1109/SLT48900.2021.9383600

[13] Naoyuki Kanda, Xiong Xiao, Jian Wu, Tianyan Zhou, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2021. A Comparative Study of Modular and Joint Approaches for Speaker-Attributed ASR on Monaural Long-Form Audio. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 296–303. https://doi.org/10.1109/ASRU51503.2021.9687974

[14] Keisuke Kinoshita, Marc Delcroix, Shoko Araki, and Tomohiro Nakatani. 2020. Tackling Real Noisy Reverberant Meetings with All-Neural Source Separation, Counting, and Diarization System. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 381–385. https://doi.org/10.1109/ICASSP40776.2020.9054577

[15] Jacob Leon Kröger, Otto Hans-Martin Lutz, and Philip Raschke. 2020. *Privacy Implications of Voice and Speech Analysis – Information Disclosure by Inference*. Springer International Publishing, Cham, 242–258. https://doi.org/10.1007/978-3-030-42504-3_16

[16] Tao Liu and Kai Yu. 2022. BER: Balanced Error Rate For Speaker Diarization. arXiv:2211.04304 [cs.SD]

[17] Zied Mnasri, Stefano Rovetta, and Francesco Masulli. 2022. Semi-Supervised Online Speaker Diarization using Vector Quantization with Alternative Codebooks. In *2022 30th European Signal Processing Conference (EUSIPCO)*. https://eurasip.org/Proceedings/Eusipco/Eusipco2022/pdfs/0000464.pdf

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

[19] Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J. Han, Shinji Watanabe, and Shrikanth Narayanan. 2022. A Review of Speaker Diarization: Recent Advances with Deep Learning. *Comput. Speech Lang.* 72, C (mar 2022), 34 pages. https://doi.org/10.1016/j.csl.2021.101317

[20] Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. 2018. First DIHARD challenge evaluation plan. *2018, tech. Rep.* (2018).

[21] David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. arXiv:1510.08484 arXiv:1510.08484v1.

[22] Yuki Takashima, Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Paola García, and Kenji Nagamatsu. 2021. End-to-End Speaker Diarization Conditioned on Speech Activity and Overlap Detection. In *2021 IEEE Spoken Language Technology Workshop (SLT)*. 849–856. https://doi.org/10.1109/SLT48900.2021.9383555

[23] Naresh P Trilok, Sung-Hyuk Cha, and Charles C Tappert. 2004. Establishing the uniqueness of the human voice for security applications. *Proc. CSIS Research Day, Pace University, NY, May* (2004).

[24] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield, and Ignacio Lopz Moreno. 2018. Speaker Diarization with LSTM. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5239–5243. https://doi.org/10.1109/ICASSP.2018.8462628

[25] Aonan Zhang, Quan Wang, Zhenyao Zhu, John Paisley, and Chong Wang. 2019. Fully Supervised Speaker Diarization. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6301–6305. https://doi.org/10.1109/ICASSP.2019.8683892

A EFFORT EXPENDED

An approximate breakdown of the effort expended on project work items can be seen in Table 3.

Task	Appx. Effort [hrs]
Topic Research and Decision	8
Literature Review	36
Meetings/Analysis	24
HSRO Training	20
Prototyping	8
Code Development	24
Writing (reports and presentations)	24
Total	**144**

Table 3: Approximate team effort expended on project items.

B EDUCATION MATERIALS

B.1 Human Subjects Research

Considering that the project works with data derived from the human voice, a biometric, and that some of our original ideas for the direction of the project involved testing speaker diarization systems on ourselves and others, it was our responsibility as professionals to learn more about the ethical and legal implications of these activities. Therefore, every team member completed the course "Students conducting no more than minimal risk research" course offered by the Human Subjects Research Office (HSRO) here at RIT [6].

B.2 PyAnnote.audio Github Repository

For using and evaluating a pretrained model, the github repository for pyannote.audio [5] was invaluable. It contains full tutorials and examples on how to use their models as well as the different evaluation metrics they had available.