PRIVACY-PRESERVING PARTICIPATION DETECTION ON THE JETSON NANO

Alex Schneidman [1] **Bassam Bikdash** [1] **Navya Yarrabelly** [1]

ABSTRACT

We aim to implement a participation detection system using speaker diarization on the Jetson Nano 2Gb. Such a system would help instructors evaluate student performance and gain insights into how their discussions evolve. By keeping personal identifiable data on the collection device, we aim to minimize ethical and privacy concerns.

1 MOTIVATION

Many classroom settings would benefit from detecting the identity and content of speakers. Grading often involves a participation component, and such a system would automate that process. At a deeper level, educators could analyze who dominates conversations and make adjustments to accommodate different students.

2 HYPOTHESES AND TESTING

Given a video and one aligned audio recording of three participants, we aim to produce a dataset of contributions to the discussion. For each participant, we will have a database of precomputed voice embeddings that we will reference in order to identify their speech within a new test audio recording. This voice embedding will be associated with a picture from a roster. Each contribution will be tagged with the participant that made the contribution (identified by their face from a roster and a pre-computed voice embedding) and two timestamps into the video and audio recording marking the beginning and end of the contribution, respectively.

We hypothesize that this system will be able to run entirely on the Jetson Nano 2Gb. Since all data collection will be done on device, this maintains the privacy of the participants up to where the data is used.

We will test the system on a dataset of discussions that we will generate. This dataset will contain conversations between 30 seconds and 5 minutes in length between our three group members.

One of the test audio clips will be introduced to our system where we will use a sliding window to examine the signal in sections. The sections will be passed through the same audio embedding model that was used to generate the audio-embedding database of the participants. Thus, the section will have it's own embedding. This embedding will then be compared to the embeddings of each of our participants (i.e. the three members of our group) using a distance metric of our choosing (e.g. mean-squared error, Euclidean distance, etc.). We have not decided on the distance metric yet as there are several available.

Performance will be evaluated along three axes: 1) Overlap between true time segments and predicted time segments averaged over all contributions. 2) Accuracy in the predicted contributor for each contribution.

3 RELATED WORK

Speaker diarization addresses the question of who-spoke-when. Throughout the diarization process, the audio data would be divided and clustered into groups of speech segments with the same speaker identity/label. Speaker diarization involves steps 1) Speech Detection to separate speech from non-speech, 2) Speech Segmentation to extract small intervals of audio clips, 3) Embedding Extraction and 4) Clustering.

(Snyder et al., 2017) is an off-the-shelf model for text-independent speaker verification with embeddings extracted from a feedforward deep neural network. Audio ALBERT(Chi et al., 2021), a lite version of the self-supervised speech representation model, achieves superior performance on downstream tasks like Speaker Identification. Since the pre-trained representations could capture the acoustic characteristics well for similar tasks, we could further finetune the pre-trained model with speaker diarization task and generate task- specific embeddings.

4 I/O AND HARDWARE

Our project and its extensions will only require the microphone that came with the Jetson Nano 2Gb. For now, we

[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Correspondence to: Alex Schneidman <amschnei@andrew.cmu.edu>.

On Device Machine Learning, Pittsburgh, PA, USA, 15213.

will stick with the one provided however, in the future, we may need a better microphone to improve the quality of our audio embeddings. A camera is not strictly necessary but it can be used to provide a face, so to speak, for the audio embeddings we generate.

5 Off-Device Training

We will be using pre-trained models for the voice detection and embedding component, so off-device training will not be needed.

We will also test the model to gain metrics on GPU usage and number of I/O transfers.

6 Potential Challenges

The performance of our model will be heavily dependent on the fidelity of the data that we capture. Since we plan on using a voice embedding model to create a reference database for our "students in the class", our primary concern is 1) capturing good enough reference to allow for comparison to audio within a larger audio clip and 2) accounting for noise or overlapping voices. We may need to apply some filtering or noise reduction to get a clean audio signal.

Another potential will come from how we decide to engineer our voice embedding comparison. Since we will be comparing a long string of audio to our database of voice embedding, we need to 1) decide on the size of the search window for scrubbing through the audio and 2) what metric to use for comparing the two embeddings.

7 Potential Extensions

We further plan to extend to more realistic classroom scenarios and work to loosen the constraints.

- Having the system work in real time to identify the speaker could be an interesting extension. We may need to use model distillation or quantization techniques to increase the speed of inference for our audio embeddings.

- Work on a real-time transcript generation system that can output speaker name and captions by using an Automatic Speech Recognition model.

8 Potential Ethical Implications

Voice embeddings are personal identifiable information. As such, collecting this information from students without their consent could be ethically dubious. For our testing, all subjects will be just the group members. However in a production environment, the system could be paired with a student roster to match face embeddings against already collected portraits. The output from the system would just be pairings of contributions with faces from the roster, eliminating the need to offload some of the personal identifiable information to a remote server.

9 Timeline and Milestones

Date	Milestone(s)
10/07	Pick candidate set of voice embedding models, write data collection pipeline.
10/21	Collect dataset of our conversations, label timestamps and speakers for evaluation.
11/04	Implement audio signal scanning functionality to both identify when a new contribution begins and match an embedding with the contribution.
11/11	Run ablation study with variety of voice embedding models. Begin potential extension to fusion with face embeddings.
11/25	Continue potential extensions and explore real-time functionality.
12/2	Complete final report and presentation.

References

Chi, P.-H., Chung, P.-H., Wu, T.-H., Hsieh, C.-C., Chen, Y.-H., Li, S.-W., and Lee, H.-y. Audio albert: A lite bert for self-supervised learning of audio representation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 344–350. IEEE, 2021.

Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. Deep neural network embeddings for text-independent speaker verification. In *Interspeech*, pp. 999–1003, 2017.