Evaluation of System for Monaural Speaker Diarization

Brian Finnerty, Wiley Matthews, Dat Tran, Dade Wood
Rochester Institute of Technology, Rochester, New York, USA
[bpf7056, wsm8855 dqt4873, daw1882]@rit.edu
Advisor: Dr. Rajendra K. Raj

Motivation

- The aim of our project is to research speaker diarization systems in order to first understand the current state of the art, then to evaluate the performance of an existing system using methods described in the literature.
- The systems we prioritized researching are designed to handle varying numbers of speakers, speaker overlap, and the presence of noise.

Proposed Solution

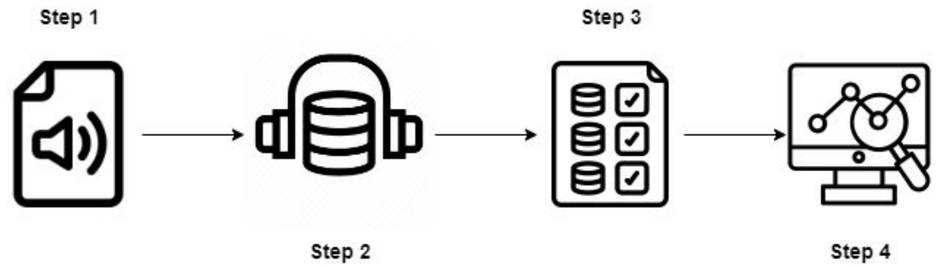
- Create a tool for generating diarization datasets by combining individual speaker and noise files.
- Generate datasets that grow increasingly noisy and contain more overlapping speech.
- Evaluate existing diarization models on our datasets and identify areas of improvement.

Background

- Park et al. [3] provides an overview of the progress and methods of speaker diarization systems starting from early 2000s to the present. It serves as a useful source for gaining familiarity with the domain and as a guide for our research into specific systems.
- Takashima et al. [4] implements End-to-End Neural speaker Diarization (EEND) - a speaker diarization system that meets two of of our three research interests: varying number of speakers and overlap handling. This system is all-neural and is trained in a supervised manner.
- Bredin et al. [1, 2] implements pyannote.audio, an offline speaker segmentation system based on Takashima et al. [4] that is available via Hugging Face. This system augments EEND by only working on short audio segments, but at a higher temporal resolution.

Design

- **Step 1:** Gather audio data from LibriSpeech.
- **Step 2:** From that data, generate diarization datasets with custom audio mixtures that control for multiple speakers and the addition of noise.
- **Step 3:** Use the datasets to evaluate the pyannote.audio pretrained diarization pipeline using DER as an evaluation metric.
- **Step 4:** Perform analysis to identify areas of improvement for the model.



Results

Dataset	Diarization Error Rate (DER)
A	25.32%
B	32.24%
C	40.97%
D	41.41%

- **Dataset A:** Two 5s segments, with noise probability of 10% each. Testing with 0-2 speakers talking in long segments.
- **Dataset B:** Two 5s segments, with noise probability of 10% each. Testing with 0-3 speakers talking in long segments.
- **Dataset C:** Four 2.5s segments, with noise probability of 5% each. Testing with 0-4 speakers talking in short segments with low noise confusion.
- **Dataset D:** Four 2.5s segments, with noise probability of 20% each. Testing with 0-4 speakers talking in short segments with high noise confusion.

Legal and Ethical Considerations

- Legal and ethical considerations share significant overlap with other issues in the Automatic Speech Recognition (ASR) domain.

Legal

- General Data Protection Regulation (GDPR)
- Biometric Information Privacy Act (BIPA)
- California Consumer Privacy Act (CCPA)

Ethical

- Privacy: Voice is a biometric
- Diversity/bias: Diversity of voices and languages
- Human Subject Research: HHS 45 CFR 46 “Common Rule”

Conclusion and Future Work

- Our tool can be used to generate new diarization datasets.
- Overlapping speakers affect pyannotate.audio’s model significantly more than the presence of noise.
- Future work includes:
  1. Add a flagging system to mark files based on the model’s performance on them and the ability to compare models automatically.
  2. Evaluate models using additional diarization evaluation metrics.
  3. Train diarization models on our datasets for performance comparisons.

References

[1] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In Proc. Interspeech 2021. Brno, Czech Republic.
[2] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and MariePhilippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing. Barcelona, Spain.
[3] 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
[4] 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