The Journal of Open Source Software

Diart: A Python Library for Real-Time Speaker Diarization

Juan Manuel Coria [1¶], **Hervé Bredin** [2], **Sahar Ghannay** [1], **Sophie Rosset** [1], **Khaled Zaouk**[3], **Ingo Fruend** [4], **Bertrand Higy** [3], **Amit Kesari**[5], **and Yagna Thakkar**[6]

1 Université Paris-Saclay CNRS, LISN, Orsay, France **2** IRIT, Université de Toulouse, CNRS, Toulouse, France **3** Ava, France **4** Verbally GmbH, Germany **5** Indian Institute of Technology, Tirupati, India **6** Tridhya Intuit Pvt Ltd, Gujarat, India ¶ Corresponding author

DOI: 10.21105/joss.05266

Software

- Review
- Repository
- Archive

Editor: Fei Tao

Reviewers:

- @sneakers-the-rat
- @mensisa

Submitted: 30 September 2022
Published: 07 July 2024

Summary

The term “speaker diarization” denotes the problem of determining “who speaks when” in a recorded conversation. Among other reasons, it has attracted the attention of the speech research community because of its ability to improve transcription performance, readability and exploitability. Speaker diarization in real-time holds the potential to accelerate and cement the adoption of this technology in our everyday lives. However, although “offline” systems today achieve outstanding performance in pre-recorded conversations, additional problems of “online” real-time diarization, like limited context and low latency, require flexible and efficient solutions enabling both research and production-ready applications. We introduce a Python package called `Diart` to address real-time speaker diarization in an efficient and flexible way.

Statement of need

`Diart` is a Python library for real-time speaker diarization. It leverages data structures and pre-trained models available in `pyannote.audio` (Bredin et al., 2020) to implement production-ready real-time inference on a variety of audio streams like local and remote audio/video files, microphones, and even WebSockets. Moreover, `Diart` was designed to facilitate research by providing fast batched inference and hyper-parameter tuning thanks to and in full compatibility with `Optuna` (Akiba et al., 2019).

`Diart` was designed with an object-oriented API fully capable of extension and customization. Streaming is powered internally by ReactiveX extensions, but available “blocks” allow users to mix and match different operations with any streaming library they choose. A prototyping tool with a CLI is also provided to quickly evaluate, profile, visualize and optimize custom systems.

`Diart` is based on previous research on low-latency online speaker diarization (Coria et al., 2021) and allows to reproduce its results. It has also participated in the recent Ego4D Audio-only Diarization Challenge (Grauman et al., 2022), outperforming the offline baseline by a large margin. We hope `Diart`’s flexibility, efficiency and customization will allow for exciting new research and applications in online speaker diarization.

Acknowledgements

This work has been funded by Université Paris-Saclay under PhD contract number 2019-089.

References

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A Next-generation Hyperparameter Optimization Framework. *Proceedings of the 25rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. https://doi.org/10.1145/3292500.3330701

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2020). pyannote.audio: neural building blocks for speaker diarization. *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*. https://doi.org/10.1109/ICASSP40776.2020.9052974

Coria, J. M., Bredin, H., Ghannay, S., & Rosset, S. (2021). Overlap-Aware Low-Latency Online Speaker Diarization Based on End-to-End Local Segmentation. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1139–1146. https://doi.org/10.1109/ASRU51503.2021.9688044

Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., Hamburger, J., Jiang, H., Liu, M., Liu, X., & others. (2022). Ego4D: Around the World in 3,000 Hours of Egocentric Video. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18995–19012. https://doi.org/10.1109/CVPR52688.2022.01842