INTERSPEECH 2021: Show & Tell Contribution
30 August – 3 September, 2021, Brno, Czechia

The LIUM Human Active Correction Platform for Speaker Diarization

Alexandre Flucha, Anthony Larcher, Ambuj Mehrish, Sylvain Meignier, Florian Plaut, Nicolas Poupon, Yevhenii Prokopalo, Adrien Puertolas, Meysam Shamsi, Marie Tahon

LIUM - EA4023, Le Mans Université
Avenue Olivier Messiaen, 72085 LE MANS CEDEX 9, France
anthony.larcher@univ-lemans.fr

Abstract

We developed a human assisted speaker diarization platform that enables a human annotator to correct the output of any speaker diarization system by providing a graphical view of the diarization segmentation and clustering steps while guiding the human annotator to optimize the correction process and easily improve the resulting diarization.
Index Terms: speaker diarization, human assisted learning, human machine interface

1. Introduction

Speaker diarization is the task of annotating "Who speaks when?" in an audio stream [1]. Often used as a pre-processing step for other speech related tasks (e.g., speech recognition or speaker recognition [2, 3]), speaker diarization is also necessary for data indexing. Internet platforms and broadcast archivers store an increasing amount of audio and video data that needs to be retrieved according to various criteria including the speakers identity. To be usable, the speaker annotation of audio archives need to reach a level of performance that is still higher than the one achieved by today's state-of-the-art automatic systems [1].

Manually annotating audio data is a tedious work that requires a huge amount of manpower. Lowering the cost while improving the quality of the diarization can be achieved by involving human annotators with the massive processing power of automatic systems in a so called human-in-the-loop strategy. Collaboration between humans and automatic systems can be implemented in two manners. First, the human can help initializing the diarization system in order to improve its performance. This helps addressing a well known bottleneck in speaker diarization systems related to the unknown number of speakers in the audio streams. This approach was proposed in [4]. The second manner consists of involving the human-in-the-loop while processing the audio stream. In a previous work [5], we proposed such an active correction process that has shown a great potential to reduce Diarization Error Rate (DER) while maintaining a low cost in terms of human interactions. The drawback of this option is that the automatic system should include the human expert by design. This strongly limits the possible architectures and doesn't allow to benefit from the rapidly growing diversity of systems [6, 7]. For this reason, we propose to use a third strategy that consists of post-editing the output of any speaker diarization system via an active-correction process involving a human expert. In this work we present an independent speaker diarization active correction module and its graphical user interface. This module can be used on top of any automatic speaker diarization system. The next section (Sec. 2), gives an overview of existing methods and tools for speaker diarization involving a human in the loop. Section 3 describes the platform architecture including our speaker diarization human interface (Section **??**) and the automatic system running as a back-end (Section 3.2).

[1] This conclusion is based on observations of our industrial partners and depends on the expected quality of service.

2. Related works

Different approaches involving human experts in the annotation process have been proposed to improve the quality of the automatically generated speaker diarization while limiting the human effort. Some approaches propose to optimize the starting point of the speaker diarization process by using the human expert to discover the existing speakers [4] while others propose to first run an automatic system before guiding the human expert through the correction of the resulting annotation [8]. Our work builds on top of the former approach. Considering that a chronological correction of the speaker diarization output is not efficient [8] and that the main benefit would come from the correction of the clustering step, we developed an automatic system that actively questions a human expert to correct an existing diarization.

A few user interfaces are popular for speech annotation. Amongst the most well known, Praat [9] and Transcriber [10] have been used for years and for many purposes. Those interfaces are not adapted for our task for two reasons. First, they have not been designed to specifically address speaker diarization and thus are not optimized for this task but most importantly, they don't allow the integration of an automatic system to assist the human. For those reasons we developed our own user interface that is described in Section 3.

3. The active-correction platform

Our platform relies on a Python back-end plugged into a JavaScript/HTML interface. When starting, the platform allows the user to select an audio file and a Python script that that will take this Wave file as input and output an MDTM file on disk after running an automatic diarization process. Our platform will then expose the resulting diarization through a graphical interface and guide the annotator to correct it.

3.1. The human interface for active correction

The user interface fulfills two main functions to allow the annotator to correct the segmentation and their clustering. A screenshot of the clustering correction interface is given on Figure 1. In a first step, the segments are displayed in the bottom panel on top of the waveform and the user is allowed to modify their borders, listen to the audio and add/remove segments or clusters. The segmentation is then validated and our active-correction module runs in the back-end to produce a clustering tree based

on the current diarization that is then displayed in the top-left panel. The clustering correction process is described in Sec. 3.2 and with more details in [5]. After correcting the clustering, the user interface allows the user to save the resulting diarization in an MDTM file or to reset it to the original one. The user interface also offers menus to modify the different parameters of the system and view.

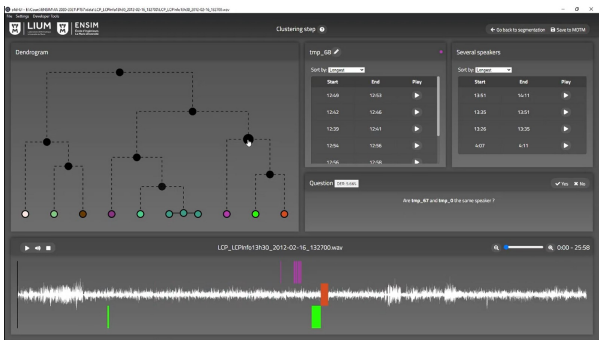

Figure 1: *Screenshot of the graphical interface of the LIUM Human Active Correction Platform for speaker diarization.*

3.2. The clustering active-correction system

In a previous work [5] we proposed an active-correction process based on a two step hierarchical agglomerative clustering. This work was limited by some assumptions made on the structure of the speaker diarization automatic system and remained a theoretical approach evaluated by a user simulation module developed for reproducibility purposes. In this work, we adapted our previous active correction process to be used on top of any automatic speaker diarization system and integrated in a user interface. Our approach takes as input an audio stream together with the diarization of this one, provided by an initial automatic system.

Building on top of our previous work [5] and considering that a clustering tree based on distances between audio segments provides an ergonomic view and a convenient framework to question the clustering process, our back-end automatic system computes one x-vectors for each audio segment and build a clustering tree by performing successive constrained hierarchical agglomerative clusterings. More details can be found in [5]. The resulting tree is exposed in the top-left panel and questions are asked to the human annotator in a sequence that is optimized w.r.t the architecture of the tree. The annotator is asked simple binary questions regarding the nodes of the clustering tree: "Should those two branches be grouped or split?". To answer the questions, the user is offered the list of segments belonging to each branch and can listen to them. Although our system ranks and limit the questions asked to the user, this one can still select other nodes from the tree and modify the clustering freely. The x-vector system used in our back-end algorithm are extracted using a ResNet architecture available in the open-source SIDEKIT platform [11] and an automatic diarization system is provided together with our platform based on S4D [12]. In test mode (when the ground truth of the diarization is available), it is possible to monitor the DER along the correction.

4. Conclusions

This work fills the gap between the academic development of speaker diarization systems and the use of those systems in commercial application where the quality of the diarization can be improved by taking advantage of a human supervision. The proposed approach enables an active-correction process that helps the human annotator maximizing the benefit of the time spent to correct automatic annotations. The current version of the system will be improved in the future to take advantage of the information provided by the human annotator along the correction process in order to improve the quality of the automatic system used in back-end in a lifelong learning manner and to enable a similar active-correction process for diarization of a collection of documents across time.

5. Acknowledgment

This project has received funding from the CHIST-ERA project ALLIES (ARN-17-CHR2-0004-01) and the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666, the Agency is not responsible for this results or use that may be made of the information.

6. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–953.

[3] S. O. Sadjadi, C. Greenberg, E. Singer, D. Reynolds, L. Mason, and J. Hernandez-Cordero, "The 2019 NIST speaker recognition evaluation CTS challenge," in *Speaker Odyssey*, vol. 2020, 2020, pp. 266–272.

[4] C. Yu and J. H. Hansen, "Active learning based constrained clustering for speaker diarization," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 11, pp. 2188–2198, 2017.

[5] Y. Prokopalo, M. Shamsi, L. Barrault, S. Meignier, and A. Larcher, "Active correction for speaker diarization with human in the loop," in *Proc. IberSPEECH 2021*, 2021, pp. 260–264.

[6] F. Landini, J. Profant, M. Diez, and L. Burget, "Bayesian HMM clustering of x-vector sequences (VBx) in speaker diarization: theory, implementation and analysis on standard tasks," *arXiv preprint arXiv:2012.14952*, 2020.

[7] N. Dawalatabad, M. Ravanelli, F. Grondin, J. Thienpondt, B. Desplanques, and H. Na, "ECAPA-TDNN embeddings for speaker diarization," *arXiv preprint arXiv:2104.01466*, 2021.

[8] P.-A. Broux, D. Doukhan, S. Petitrenaud, S. Meignier, and J. Carrive, "Computer-assisted speaker diarization: How to evaluate human corrections," in *LREC 2018, Eleventh International Conference on Language Resources and Evaluation*, 2018.

[9] P. Boersma and V. Van Heuven, "Speak and unspeak with praat," *Glot International*, vol. 5, no. 9/10, pp. 341–347, 2001.

[10] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1-2, pp. 5–22, 2001.

[11] A. Larcher, K. A. Lee, and S. Meignier, "An extensible speaker identification sidekit in python," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5095–5099.

[12] P.-A. Broux, F. Desnous, A. Larcher, S. Petitrenaud, J. Carrive, and S. Meignier, "S4D: Speaker diarization toolkit in python," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018.