Review of a Doctoral Thesis at FIT BUT

Doctoral thesis (hereinafter referred to as "thesis"), title of the thesis:

From modular to end-to-end speaker diarization

Name of the doctoral student (hereinafter referred to as "student"), name and surname:

Federico Nicolás LANDINI

Name and institution of the reviewer (full name of the reviewer, full name, and country of the institution):

Dr. Hervé BREDIN, Centre National de la Recherche Scientifique (CNRS), France

I. Thesis

Appropriateness and relevance

The thesis entitled "From modular to end-to-end speaker diarization" describes Mr. Landini's research on the timely topic of speaker diarization, which can be summarized as answering the question "who speaks when?" in the audio recording of a conversation, without prior knowledge on the identity nor the number of speakers.

Throughout his thesis, Mr. Landini provides a comprehensive and up-to-date overview of the very prolific field of speaker diarization, selecting and summarizing the most relevant papers to the specific problems addressed in each chapter.

Compared to other speech processing tasks -- such as automatic speech recognition or speaker verification -- which have witnessed huge improvements in the last decade, thanks to (supervised) deep learning and the availability of large amount of labeled data, speaker diarization remains one of the most challenging tasks in the general field of speech processing, for two main reasons: the lack of large labeled datasets and the fact that it can be seen as an (at least partially) unsupervised clustering task that deep learning still struggles with. Mr. Landini accurately pinpoints and describes those two challenges in his thesis and proposes novel solutions for both.

A summary of the contributions of the thesis

The first problem addressed by Mr. Landini is the clustering step on which modular speaker diarization approaches rely -- and this is where I believe the most impactful contribution of his thesis lies. The proposed Bayesian HMM clustering of x-vector speaker embedding (dubbed VBx) is today widely considered by the community as the best clustering approach for modular (as opposed to end-to-end) speaker diarization pipelines: it has been instrumental in the winning system of DIHARD2 challenge, is being used as baseline in most (if not all) recent international speaker diarization challenges since then, and the corresponding 2021 publications already has more than 150 citations. Backed by strong theoretical foundations, the proposed VBx algorithm also shines in practice and I would like to highlight here that, though most results reported in Mr. Landini's thesis are for 2-speakers telephone conversations, he nevertheless made the very rare (and often humbling) effort to also evaluate the approach on a wide range of "in-the-wild" multi-speaker benchmarks.

The second problem addressed by Mr. Landini is the lack of large labeled datasets needed for training end-to-end (as opposed to modular) speaker diarization approaches. While there exists a huge amount of single-speaker recordings for training automatic speech recognition systems, only a handful of multi-speaker datasets are available with speaker labels for training speaker diarization (mostly because of the inherent difficult and high cost of the manual annotation process). Mr. Landini's proposed solution to this problem is to leverage the huge amount of single-speaker data and use it to generate synthetic (but as realistic as possible) multi-speaker data (dubbed Simulated Conversations, or SC) for training end-to-end speaker diarization systems. Training on simulated conversations is compared to training on similar (but less realistic) Simulated Mixtures (or SM), proposed by another research group a few years before. Results show that SC-based training clearly outperforms SM-based training. One thing that I believe is missing in this part of the thesis is a more detailed comparison of the proposed two-stages SC-based training procedure (training on SC, then fine-tuning on real data) with a simple baseline of training directly on real data. Though the latter might fail completely, it would be nice to report those numbers anyway, to give a better idea of the actual gain brought by the proposed approach. I also appreciate the honesty of Mr. Landini's regarding the limitations of the proposed approach for multi-speaker "in the wild" benchmarks.

The third main contribution of Mr. Landini's is at the crossroads of both problems: bringing better clustering (more precisely, improving the estimation of the number of speakers) to end-to-end speaker diarization models. As a matter of fact, while end-to-end approaches (EEND) have become increasingly popular in the last few years, they do struggle with the automatic estimation of the number of speakers, especially when the number of speakers at test time is greater than that at training time. Many approaches have been proposed to tackle this problem and Mr. Landini’s third contribution (DiaPer) falls into the same category: estimating attractors in the latent space of the model to predict the number of speakers. What makes it stand out is that it relies on non-auto-regressive attractors thanks to the recently proposed Perceiver architecture. Though it does not solve all problems of EEND systems (in particular, the upper bound on the number of speakers must be fixed *a priori*), results suggest it leads to better estimation of the number of speakers than previous approaches.

Novelty and significance:

The three main contributions of Mr. Landini's mentioned above are novel and sorted in order of significance for the scientific field of speaker diarization.

- VBx already has had a strong impact on the community and remains one of the strongest baseline any new research should compare to.
- Despite their current limitations for wideband "in the wild" data, the use of SC is a very promising research directions that will be (and already have been) explored by other research groups.
- Finally, the proposed DiaPer is probably the less impactful and most incremental of the three contributions but still a significant one, as it improves the estimation of the number of speakers in end-to-end speaker diarization systems.

Review of a Doctoral Thesis at FIT BUT

Evaluation of the formal aspects of the thesis:

Mr. Landini's thesis reads very well and is easy to follow. The structure of the thesis is well-organized and follows what I expect from a thesis in the field of automatic speech processing: introduction, literature review, main contributions (including experimental results and discussions) and conclusions. The bibliography is complete and well-organized. It is well illustrated with figures and tables, which are always mentioned and explained in the text. Large tables of results might have sometimes been advantageously replaced by more readable figures, but this is a minor point and a matter of preference.

Quality of publications

Mr. Landini's co-authored 15 publications (+ 1 under review) since 2018 at high level speech processing conferences and journals, including :

- ICASSP (CORE rank B, GGS rating A) in 2024 (2x), 2023, 2021, and 2020 (2x)
- Interspeech (CORE rank A, GGS rating A): 2023, 2022, and 2018
- IEEE/ACM Transactions on Audio, Speech, and Language Processing (impact factor 5.4): 2024, and 2019
- Computer Speech & Language: 2021

The core contributions of the thesis have all been published (or submitted for DiaPer) as first author with a high percentage of contribution (VBx: 30%, SC: 70%, DiaPer: 80%).

II. Student's overall achievements

Overall R&D activities evaluation:

Mr. Landini's thesis clearly demonstrate that he can conduct high-quality and novel research in the field of automatic speech processing. He also successfully participated to a few international benchmarks (DIHARD, CHiME, or NIST SRE) either as a member or a leader of his team, reaching top positions in most of them, hence reaching state-of-the-art performance.

Assessment of other characteristics:

His publications track record also indicates that he can work in a team and collaborate with other researchers. He also spent a few months as research intern at MetaAI, FAIR, and Apple Inc. through the course of his thesis, showing his ability to adapt and willingness to address real-world problems.

Finally, I would like to highlight one aspect that is often overlooked: reproducible research and code sharing. Most (if not all) publications of Mr. Landini are accompanied by an open-source code repository, which is a very good practice that I would like to see more often in the research community. Like in many fields, it is often difficult to compare the results of different papers because of the lack of details, making results impossible to compare (let alone reproduce). The speaker diarization field is no exception to this rule. Not only did Mr. Landini share his code, but he also shared the evaluation protocol and ground truth labels used in his experimental results - which now serve as reference for many (especially the AMI and CALLHOME benchmarks).

III. Conclusion

For all these reasons, I hereby state that, in my opinion, Mr. Landini's doctoral thesis and achievements meet the generally accepted requirements for the award of an academic degree (in accordance with Section 47 of Act No. 111/1998 Coll., on higher education institution).

Toulouse (France) 17.05.2024

Signature of the reviewer: