BRNO FACULTY
UNIVERSITY OF INFORMATION
OF TECHNOLOGY TECHNOLOGY

Review of doctoral dissertation
Towards Discriminative Speaker Representations for Speaker Recognition and Diarization
Submitted by Magdalena Marta Rybicka at AGH University of Krakow

The thesis investigates into neural representations of speakers in automatic speaker recognition (SR) and speaker diarization (SD) systems, touching also speaker separation task. These topics fall under the area of automatic speech processing, which is itself considered as part of artificial intelligence. The investigated topics and drawn conclusions are up-tp-date and relevant for both academic and industrial R&D community.

The thesis has 4 chapters, and 131 pages and is based on five candidate's publications – three Interspeech (CORE A) conference papers and two papers in important journals within the speech processing community – IEEE Transactions on Audio, Speech and Language Processing (T-ASL) and IEEE Signal Processing Letters (SPL). This review first deals with the technical content of the thesis, then summarizes its technical quality, comments on the formal points and finally presents overall conclusion and recommendation to the committee.

Technical content of the thesis and remarks to chapters and papers

Chapter 1 introduces the thesis, provides a general overview of the problems and definition of the main tasks. It also presents the main achievements and structure of the thesis, as well as a summary of work not covered by the 5 core publications. It documents extensive cooperation of the candidate with Johns Hopkins University (Center of Language and Speech Processing - CLSP, and Center of Excellence - CoE), a top US academic laboratory dealing with speech and natural language processing. Minor comments to this chapter: in section 1.3 (Contribution), it would be nice to provide references to respective sections of the thesis and papers and in in the paragraph covering projects, I would appreciate knowing who were their principal investigators.

Chapter 2 "Reseqrch background" is a "heavy-weight" text presenting the tasks, State-of-the-Art (SotA), data and evaluation metrics. Its coverage is broad and it documents excellent orientation of the candidate in the domain. It gives a solid base for understanding the rest of the thesis. I especially appreciated a comprehensive overview of angular methods for computation of the objective functions. Small comments:

1. Structure of the chapter – in my opinion, it would be better to present an overview of **all** tasks at the beginning of the chapter, and give their evaluation metrics and then only attack the SotA – this would give the reader better structuring of the acquired know-how. Also, the chapter is rather long, I would recommend to split it into three shorter ones: 2. Tasks and metrics, 3. SotA and 4. Data.
2. In some places, especially in the description of "legacy" techniques such as i-vectors, I would appreciate a more mathematical rigor (probabilities vs. likelihoods, matrices and vectors in bold). In later parts dealing with neural architectures, the notation is flawless.
3. The data section could be complemented by summary tables, giving the purpose of individual datasets, technical details and references, resp. citations to sections/papers where these were used.

Chapter 3 covers candidate's contribution and introduces the five core papers: Section 3.1 and Papers I and II deal with discriminative speaker representations, mainly with the structure and training of neural architectures for their extraction. Paper I concentrates on making the parameters of angular loss function computation adaptable – the task is well explained, solid hypothesis is proposed, experiments are well executed and the solution works better than the baseline. The second part of Paper I investigates changes in the NN structure

– combining blocks of TDNN scheme and ResNet. While the combined architecture works better, I would appreciate deeper insight into what the main problem of the simply architectures is and how the proposed changes address it rather than presenting only the structural changes and results. In this regard, Paper II on replacing ResNet with SpineNet and on introducing temporal processing (Squeeze and Excitation Blocks) into the processing chain does much better job and the reasoning behind is very clear.

Section 3.2 and Papers III and IV propose strengthening speaker representations in neural end-to-end (EEND) based speaker diarization, addressing mainly the embedding clustering step. The introduction od Non-Autoregressive Attractors (NAA) defined in Paper III is a very successful piece of research – several systems were proposed, evaluation was performed with positive results and the work attracted the interest of the research community (Paper III has 19 citations on Google Scholar). Journal paper IV is an extension of the conference one and brings several innovations and more extensive experimental evaluation (including an ablation study). I appreciate mainly the introduction of the Single Speaker Activity Detection (SSAD) block significantly improving the results of all schemes that include it. I would still appreciate a more detailed and intuitive explanation, what led to introduction of several alternatives (from EEND-NAA-Overest to EEND-NAA-1step) – i.e. what was wrong with the simple approaches and how the following variants fixed it. Also, in case of iterative search of the attractors (Paper III), it would be good to comment on the computing requirements of this approach, as it seems that the iterations take place also at inference time, and about how the optimum number of iterations was set.

Finally, Section 3.3 and SPL Paper V deal with the joint task of diarization and speech separation – this is a natural extension of previous work into signal processing domain relying on good speaker representations. This generalization worked very well and the experiments document superiority of the proposed attractors also for speech separation task. A small inconvenient is the need to know beforehand whether the source signal has not much speaker overlaps (clustering attractors work better) or high amount speaker overlaps (diarization attractors showing superiority) – for further research, it would be interesting to suggest one scheme working in a variety of scenarios.

Chapter 4 is the standard summary and future work –I can agree with all mentioned points (good engineering plans), but I am lacking a bit of more global picture, with more "far-flung" goals, as well as comments on the current trends in speech processing (SSL pre-trained models and speech LLMs).

Summary of the technical content of the thesis

The thesis clearly demonstrates the qualities of the candidate – capability to study non-trivial literature from several fields, suggest own novel solutions, implement them, carefully test and discuss the results of experiments. It demonstrates a broad overview of both speech and signal processing. machine learning and modern neural architectures. The experiments are well conducted, well documented, well analyzed and well commented. Several of the findings (such as NAA attractors and combination for diarization and speech separation) already have impact on the broad research community or have a strong potential to have it in near future.

Comments on the formal aspects

While I generally prefer "linear" theses over ones assembling several publications, I liked this one – the author carefully chose a limited number of the most relevant publications, chapter 3 introducing their content was sufficiently "light" and the content in the papers did not overlap much. The structure of the thesis (except for lengthy Chapter 2) is good and the author did her best to present the work done in logical sequence. The thesis is in flawless English, with only a very limited number of stylistic issues – I will be happy to hand the candidate a commented version of the document.

The quality of presenting the results is excellent, the tables and figures are well readable, well annotated, and provide straightforward information on what was achieved. The mathematical writing (except minor exceptions) is clear and coherent. The usage of literature is excellent, the lists of references in the thesis and in the papers are extensive and document a broad scope of candidate's knowledge. I would just suggest alphanumerical style for references in the thesis – [Watanabe2020] is more readable than [193]

Summary and recommendation

I have carefully examined the doctoral thesis of Ms. Magdalena Marta Rybicka. Despite minor critical points raised above, in my opinion, it is a solid work that contributes to progress in neural speaker representation research. I also examined candidate's publication track and find it meeting the standards for a PhD candidate at a respected University.

I conclude that in the doctoral dissertation submitted for review by Ms. Magdalena Marta Rybicka entitled "Towards Discriminative Speaker Representations for Speaker Recognition and Diarization", the Ph.D. candidate demonstrated knowledge of speech processing, neural architectures and machine learning. I consider the original contribution to the scientific field described in the doctoral dissertation, as well as the publication record, to be significant. In my opinion, the presented work meets the requirements for doctoral dissertations in the current law on academic degrees and titles. Therefore, I request that the Discipline Council admits Ms. Magdalena Marta Rybicka to the next stages of the doctoral procedure.

Considering the technical quality of the thesis, quality of experiments conducted on open data-sets, validity of results and their importance for the international scientific community, as well as international cooperation with a prestigious US speech/NLP laboratory, I suggest that this thesis obtains a distinction.

I suggest the following questions to be answered at the PhD defense:

1. Explain the reasoning and intuition behind combining blocks of TDNN scheme and ResNet in section 3.1 and Paper I.
2. Provide an intuitive explanation, what led to introduction of several alternatives (from EEND-NAA-Overest to EEND-NAA-1step) in Section 3.2 and Paper IV – i.e. what was wrong with the simple approaches and how the following variants fixed it.
3. At several places in the thesis and in Paper IV, you state that the positional encoding was not used – please give the reason behind.
4. The Separation/Diarization scheme in Section 3.3 and Paper V is trained with an objective function combining several criteria. Do you think that it would be feasible to train it only with the speech separation criterion, that should contain all the information about the diarization (i.e. when the speaker is not active, his/her signal should be zero)?

In Brno, January 14th 2026

Prof. Dr. Jan "Honza" Černocký
Head of Department of Computer Graphics and Multimedia
Responsible of BUT Speech@FIT group
Faculty of Information Technology, Brno University of Technology
Božetěchova 2, 612 66 Brno, Czech Republic
Tel: +420 5 41141284 Cell: +420 604738324,
mailto:cernocky@fit.vutbr.cz, http://www.fit.vutbr.cz/~cernocky
http://www.fit.vutbr.cz/ https://cs-cz.facebook.com/FIT.VUT
http://speech.fit.vutbr.cz https://www.facebook.com/BUT-Speech/