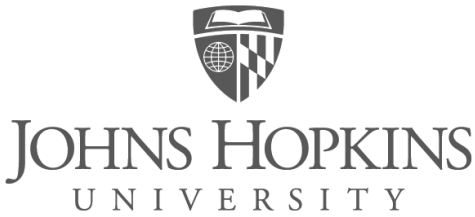


The Johns Hopkins University
Department of Electrical and
Computer Engineering
3400 N. Charles Street
304 Barton Hall
Baltimore, MD 21218-2686

Brno University of Technology
Faculty of Information Technology
Božetěchova 2, 612 66
Brno, Czech Republic

November 30, 2022

Review of Doctoral Thesis

Exploiting Uncertainty Information in Speaker Verification and Diarization submitted by **Anna Silnova**

This thesis investigates the topics of speaker verification and diarization. Speaker recognition is a biometric modality consisting of identifying people by the characteristics of their voices. Speaker recognition derives in several tasks and applications. Speaker verification is the task of deciding if a given target speaker is present in an audio recording. Meanwhile, speaker diarization is the task of saying "who spoke when" and involves splitting an audio into single speaker turns and assigning unique speaker labels to those turns. These technologies have applications on security, access control, law enforcement, forensics, speech analytics, etc. Therefore, these are topics worth investigating.

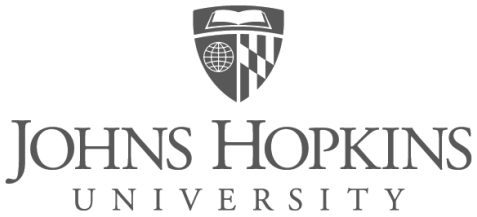
Most speaker verification and diarization pipelines involve computing some form of representation (a.k.a. embedding) intended to summarize the biometric information of the speaker while discarding other types of information. This thesis focuses on modeling the uncertainty about the value of this embedding, which is disregarded by most previous works. This work shows that modeling this uncertainty can improve performance in certain conditions. Experiments were done on standard datasets like VoxCeleb, SRE and Dihard.

The thesis consists of seven chapters plus appendices with a total of 121 pages. 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

Chapter 1 introduces the concepts of speaker verification and diarization, motivates the need of introducing uncertainty when making decisions, and summarizes the structure of the thesis. It also enumerates the claims of the thesis.

Chapter 2 introduces the technical background on speaker verification and diarization. The chapter starts by describing the assumptions of speaker recognition models using latent representations to encode the



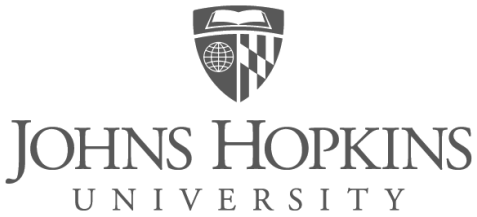
speaker characteristics. Then, it formalizes the concept of speaker verification by log-likelihood ratios and introduces the metrics. In the second part of the chapter, it describes the typical speaker diarization pipeline using agglomerative clustering and diarization metrics. Finally, it includes a short literature review on embedding extractors, and describes the datasets used along the thesis. I find this chapter a good introduction to concepts required to understand the rest of the thesis. However, I would have appreciated a more extended review of the state-of-the-art. For example, including a more detailed description of the pipelines, features, alternative network architectures and approaches before i-vectors. Regarding diarization, I would consider pipelines with spectral clustering and Variational Bayes PLDA. Also, this chapter would be a good point to include some results comparing the baseline models.

Chapter 3 is dedicated to the probabilistic linear discriminant analysis back-ends. After introducing the well-known Gaussian PLDA model, the author dives into the proposed Heavy Tailed PLDA model. From this chapter, I meanly enjoyed the math derivations, which lead to efficient evaluation of the log-likelihood ratio. Here, I only missed some more detailed explanation of why length normalization is a good substitute for HT-PLDA.

Chapter 4 deals with the generative training of the HT-PLDA model. It starts defining the Variational Bayes procedure to get to approximate posteriors of the latent variables and the corresponding mathematical derivations. In this part, it is not clear how to go from $Q(\alpha)$ in eq. (4.5) to (4.6), so the author could elaborate more on that. In the experimental section, the HT-PLDA is evaluated using i-vectors and SRE datasets, or TDNN x-vectors and VoxCeleb/SITW datasets. My questions here would be:

- Why not evaluate SRE and VoxCeleb using more modern x-vector architectures like ResNet and Res2Net.
- Rather than VoxCeleb, which has little mismatch between training and test, why not evaluate on more challenging datasets like SRE20 or 21. May be there would be more improvement of those datasets.
- What is the motivation to use different degrees of freedom in training and test?

Chapter 5 shows how to train the HT-PLDA model discriminatively. This is a very extensive chapter where several strategies are explained in detail. The first strategy is binary cross-entropy which is common in the literature, and it based comparing pair of recordings and maximize the same/different label posterior. More interestingly, later, the author approaches the problem of how to optimize the posterior of the correct label partition of the whole dataset. Since this objective is intractable, the author proposed several approximations using the pseudolikelihood or sampling strategies. Here, we find the most novel part of the thesis since, to my knowledge, all works in the speaker or face recognition literature are restricted to binary or categorical cross-entropy objectives. The trained models are evaluated on speaker verification and diarization tasks showing some improvements in terms of performance and calibration. From the theoretical part of the chapter, I would like to comment about an apparent mismatch between the equations in the regular Split-Merge sampling and the Smart/Dumb Split/Merge. I understand that the “Dumb” merge/split proposals in eq. (5.29-30) should match the ones in (5.25) and eq. (5.33-34) should match (5.24) but I may be mistaken. From the experimental part, I would ask the author to comment why $\sigma = 0$ is chosen as diarization threshold for calibrated scores, I think it would better to use -logit P where P is the prior prob of finding a pair of vectors from the same speaker in the AHC affinity matrix, i.e., P=1/(expected-num-speakers).



Chapter 6 changes the topic from PLDA to embedding extraction. This chapter proposes to modify the embedding extractor to predict not only a point estimate (mean) of the embedding but also the uncertainty about the embedding in the form a diagonal precision vector. This involves adding some extra layers to the extractor to predict the precision and append a simplified PLDA model —trained jointly with the extractor-- to compute the likelihoods of the data given the label partitions. Previous works have tried to do this using the variational autoencoder framework predicting approximate posteriors for the latent variable. The novelty of this work is that it is based on predicting the likelihoods of the data. This method is evaluated on a diarization task. Since diarization uses embeddings extracted from short segments, they should have large uncertainty. The results show improvement in terms of calibration w.r.t. regular embeddings. However, Table 6.1 only show results using G-PLDA. I think that table should compare also with HT-PLDA results, and I wonder if HT-PLDA could be used on top of embeddings with uncertainty. Would we need to train that HT-PLDA jointly with the extractor or could be trained a posteriori? Also, I wonder whether the author has done some analysis of how the uncertainty is related to the length/quality of the audios.

Chapter 7 summarizes the content of the thesis and proposes some future directions.

Appendices include mathematical derivations and discussions about how to sample speaker partitions and train discriminative HT-PLDA in an efficient manner.

Summary of the technical content of the thesis

This thesis clearly demonstrates the qualities of the candidate – capability to propose novel solutions, implement them, carefully test, and discuss the results of experiments. I highly appreciate the formal mathematical derivatoins, quantity and quality of experiments done on several standard public datasets and thorough discussions. I think the contributions of thesis, i.e., fast HT-PLDA, discriminative training strategies for PLDA and Probabilistic Embeddings will be greatly appreciated by the international research community.

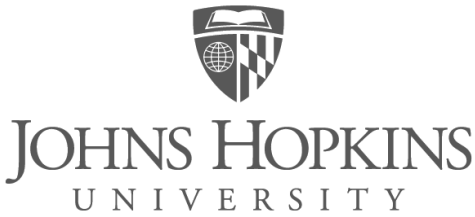
Comment on the formal aspects

The thesis is well written in correct English, it is well structured easy to follow. I greatly appreciate the formalism in the mathematical definitions and derivations. Tables and figures are of good quality and easy to read and understand. Something that I miss is a summary at the end of each chapter. Some of them have it but others do not.

Summary and Recommendation

I have carefully examined the doctoral thesis of Ms. Anna Silnova. Despite the minor criticism raised above (many points are rather recommendations than critique), in my opinion, it is a solid work that contributes to progress in Speaker Verification and Diarization. I also examined his publication track and find it exceeding the standards for a PhD candidate at a respected University.

To conclude, I find the dissertation ready for publication and recommend accepting it as a partial requirement for granting Ms. Anna Silnova the Doctoral degree at Brno University of Technology.



In Baltimore, November 30th, 2022

Jesus Antonio Villalba Lopez
Assistant Research Professor
Department of Electrical & Computer Engineering