Institute of Information Theory and Automation
Pod Vodárenskou věží 4, 182 08 Praha 8, Czech Republic

Prague 30.11.2022

Reviewer report on dissertation thesis: "EXPLOITING UNCERTAINTY INFORMATION IN SPEAKER VERIFICATION AND DIARIZATION"

Author: Anna Silnová

The main topic of the thesis is investigating whether more accurate modeling of uncertainty positively impacts the accuracy of the speaker verification or diarization problem.

Topic selection: The chosen topic of the thesis is very important and demanding, with lots of space for potential improvement. While all methods are demonstrated in the audio domain, the proposed techniques can be easily applied in other domains requiring verification and diarization. Also, the field is typically dominated by maximum likelihood approaches, so the question of potential improvement from proper uncertainty modeling is very relevant. Therefore, I consider the topic to be well chosen.

Content: The thesis is essentially split into three key areas: i) generative training of HT-PLDA model, ii) discriminative training of HT-PLDA model, and iii) probabilistic embeddings. The common component is the HT-PLDA model, which models embedding as a linear transformation of the latent space (I would call it a factor analyzer model). However, different treatment of the model in those parts requires fundamentally different methods for uncertainty processing. I would focus my review primarily on the methodological part, which is my own research background.

Methods: Due to the focus on the linear generative model, many operations are analytically tractable, and the author makes proper use of this advantage. On the other hand, it raises suspicion if linearity is not the limiting factor in relatively minor performance improvements reported in the chosen application. Bayesian analysis of the model in Chapter 4 is based on the Variational Bayes approach, which is well studied for this class of models. Thus, I consider this part to be mainly application oriented. The methodology of the discriminative training in Chapter 5 is much more interesting since the advantageous analytical properties cannot be used due to an excessive number of terms in the analytical formula. The author studies multiple approaches to approximate the problem, ranging from pseudo-likelihood to advanced Monte Carlo methods. While many of the methods that are used in this part are also known, their application is innovative, and the results interesting. The last part of the theses in Chapter 5 is based on the extension of the model to speech segments rather than previously used embeddings. The methodology of this section is less clear to me, it looks like a form of

marginal likelihood estimate, using again the analytical properties of the linear model. I would appreciate more details and a summarization of the final optimization problem.

Style and language: The thesis is written in very good English with minimal typos and grammar mistakes. The author paid attention to a detailed introduction to each topic, so the flow of the argument is easy to follow. A had minor difficulty following derivation in Chapter 5 where parameter θ is sometimes omitted in the conditioning. Also, some sub-routines are described in details without explicit

Outcome: The main outcome of the thesis is a number of high quality publications, two of them in journals and many conference proceedings. I consider this to be a significant achievement since many of them collected significant amount of citations. I have no doubt about the scientific erudition of the author.

I have the following remarks and questions:

- The basic PLDA model is based on linear transformation (3.1). Can you comment on methods that would be suitable for analysis of uncertainty if this transformation is non-linear?

- The use of sampling to approximate normalization in (5.17) essentially turns (5.18) into an "EM-like" objective function. You mention on page 53 that sampling will be performed with a fixed parameter $\hat{\theta}$. How is the full algorithm running? I presume that you are optimizing θ using GD. Does it mean that the sampler is run from an initial state T iterations and then terminates? How long was the chain within each of the GD steps?

- I am wondering about the validation protocol. In my experience, variance among human subjects is sometimes surprising, and the choice of the validation/test set dramatically influences the outcome. You mention on page 43 that the dataset was split to 90% training 10% test. Was it an i.i.d sampling, was it repeated? Do you see any need to use e.g. group cross-validation in this application?

The author of this thesis demonstrated his ability to work independently and propose novel ideas in the field of computer science. In my opinion, the proposed thesis meets standards imposed on dissertation thesis in this field. I recommend its acceptance.

Doc. Ing. Václav Šmídl, Ph.D.
Institute of Information Theory and Automation