REPORT ON DOCTORAL THESIS

Title of the thesis:
Improving Robustness of Speaker Recognition Using Discriminative Techniques

Ph.D. candidate: Ing. Ondřej Novotný

Reviewer: Doc. Ing. Petr Pollák, CSc.
Czech Technical University in Prague,
Technická 2, 166 27 Praha 6, Czech Republic

The work of Ondřej Novotný deals with speaker recognition (SRE) with focus on application of discrimative techniques to improve the robustness of speaker verification (SV) task. As author presents in the thesis, a design of text-independent SV system invariant to data domain is still challenging task of world-wide research in the field of SV or SRE generally, moreover, to increase environmental robustness of SV, we must deal with unlimited variability of signal backgrounds or distortion respectively. So the importance of this research is evident and for sure it can be accepted as a topic of PhD study.

Delivered thesis is structured into 11 chapters describing basic introduction, the state-of-the-art, general framework, particulary studied techniques including realized experiments, and in the end conclusions with brief discussion about possible future work. Within the introduction, the author presents the motivation for his work and the main goal which is summarized by him as "investigating a state-of-the-art text-independent speaker verification systems to improve their individual parts using a discriminative approach".

It can be said, that given general goal as well as particular subgoals specified and investigated further are dissertable and that the author accomplished them well within his work. Original contributions correspond to defined goals and they can be summarized as follows:

- First of all, author describes clearly in brief summary the framework in the field of SV, i.e. basic approaches of speech-feature extraction, nowadyas standards of speaker embeddings (i-vectors and x-vectors), scoring for SV purposes, available speech corpora for training of SRE systems, typical evaluation criteria used for SV evaluation, and mainly he presents the crucial difficulties which limit the performance of SV systems under real conditions.
- Within further chapters, which represent the core of the thesis, the author presents very wide study of various techniques how to improve the robustness of SV system. He studied in more detail multicondition training of i-vector extractor and PLDA scoring in standard i-vector SV, an improvement of i-vector system by the usage of SBN (stacked bottleneck features) features, DNN alignment, or discrimenatively retraind i-vector extractor, the preprocessing of noisy input of SV using DNN-based speech enhancement (so called DNN-autoencoder), the impact of score normalization techniques for better language robustness, and finally impact of data on the robustness of discriminative x-vector-based systems which usage dominates in current SRE research as the most frequently used approach of embeddings for speaker modelling. Presented solutions contribute without any doubt the research in studied field and I appreciate theoretical descriptions and discussions related to particularly described techniques. They prove the author's expertise in the field of SRE, but also its overlap into other areas (e.g. speech enhancement), where the unifying approach is the usage of DNN of various types.

- Particular approaches mentioned above were tested within very wide experimental part of this thesis. For this purpose author has defined realistic benchmark scenarios using large available speech corpora well described in the chapter 3. Special attention is devoted to data augmentation which is crucial step for multicondition training when the availability of real speech data covering various kinds of distortion is typically limited. This can be emphasized as one of the important messages of given thesis.

  Achieved results illustrate without any doubt capabilities and resonable impact of particular techniques discussed within the thesis, although the orientation in many tables with large amount of achieved numerical results is a little difficult. From this point of view, I would appreciate some overall comparison of results for the most important setups discussed in particular chapters and for selected key benchmarks in one table or figure.

  I also missed a little the benchmarks BUT-RET-* in the table 7.1 because for BUT-RET-merge we can observe rather worse results in the table 6.1, i.e. for standard i-vector system, and it could be interesting to compare the impact of pure multicondition training presented in table 6.1 with the impact of data preprocessing based on DNN-autoencoder.

Originality of this thesis is proved by author's high quality publications. The article in the journal Computer Speech and Language must be mentioned at first, however, other publications at leading international conferences and workshops as Interspeech, ICASSP, SLT, and Odyssey also represent the publication activity above the standard. I just missed a little a separate list of author's publications in the thesis, because other publications where Ondřej Novotný is not the first author are slightly hidden.

Concerning the formal issues, the thesis is written well. Discussed problems are clearly explained with good English. Maybe chapters 7 and 8 should be swapped because it is slightly less logical to discuss the impact of speech enhancement in approaches which are described in basic form later (e.g. SBN features are described in section 8.1 at the page 77, but the impact of speech enhancement to SV system with SBN in section 7.3.2 at the page 67). Typeseting of the thesis is also good et all, I have mentioned just slightly more often overfull of printing area width, typically in the case of citations, wide tables, or headings, as well as quite often near empty pages before long tables or section beginnings. I have also find some other inconsistencies (e.g. cross-reference to section 7.3.3 is probably not correct) but generally, it does not affect the overall high level of this thesis.

In the end, I would like to mention really wide bibliography. The list of cited publications is at 12 pages and it maps very well current state-of-the-art in given field. Moreover, looking at author's publications in this list and comparing them with cited publications of other authors, it is evident that author's activities represent the significant part of the mainstream within ongoing world-wide research in the field of SRE.

On the basis of facts mentioned above, I do not have any serious remark to delivered thesis and for a general discussion I have just the following additional questions:

1. You use MFCC features in two different setup within the experimental part: 1) 24 mel-filters in frequency band 120-3800 Hz including delta and double delta features and 2) 23 mel-filters in frequency band 20-3700 Hz without delta features. I have two questions related to these setups:
   - Is it really reasonable to take frequency band in the second setup from 20 Hz for

speech signal? Shouldn't be from 120 Hz in this case as well? If it is really from 20 Hz, shouldn't be the number of mel-filters higher due to significantly decreasing bandwidth of particular filters in low-frequency band?

- You do not use delta features in the second setup used for discriminative systems. I understand that some context information is taken by used TDNN stucture in x-vector system as well as by the taking context information at input of other DNN structures (e.g. for SBN). But wouldn't the use of delta features have a positive effect in these approaches as well? From the first 4 columns in tables 6.1 (7.2.) and 7.3 we see worse results for x-vector system and the differences in obtained results can be given by differences in principle structures as well as by differences in input features.

2. You have mentioned in the thesis, that standard scenarios of SV suppose the work with 8 kHz data but that nowadays the usage of 16 kHz data starts being more often. If I did not miss anything, all your experiments were realized with 8 kHz data. What changes in the setups of described systems should be realized when 16 kHz data are used? Increased frequency band is automatic, but what about increasing of number of cepstral coefficients, mel-filters, some changes in DNN structures, etc.?
3. Finally one question slightly out of the scope of your thesis: According to illustrative fig. 7.2, used DNN-autoencoder seems to map well log-magnitude spectrum of distorted speech to its enhanced variant and its contribution is proved within target SV systems where significant improvement of EER was achieved mainly for more adverse benchmark scenarions. Do you have an experience or knowledge about a performance of DNN-based autoencoder which generates really enhanced speech signal?

Finally, it can be said that the thesis of Ondřej Novotný has very high level and that it shows his capability of independent and original research activity. On the basis of these facts, **I do recommend** to accept the thesis with the aim of receiving the Doctoral degree at Brno University of Technology.

In Prague, October 26, 2021