Review of Master's Thesis

Student: Lukáč Peter, Bc.
Title: Speaker Verification without Feature Extraction (id 23746)
Reviewer: Rohdin Johan A., Dr., DCGM FIT BUT

1. Assignment complexity — **more demanding assignment**

The student has to learn and understand state-of-the-art neural network architectures for speaker recognition and obtain good results from them. Such architectures are typically very complex and require substantial experience to use successfully. Moreover, the student is expected to improve upon them.

2. Completeness of assignment requirements — **assignment fulfilled with enhancements**

The student fulfilled all tasks. In particular, he applied many state-of the-art neural network techniques (loss functions, architectures, etc) to the RawNet model. The experiments were exhaustive and systematic and the baseline result was outperformed substantially.

3. Length of technical report — **within minimum requirements**

The thesis is exactly 50 pages and therefore meets the minimum requirements.

4. Presentation level of technical report — **90 p. (A)**

The thesis is very well written. The contents of the chapters are well balanced. Whenever material that has been explained earlier in the thesis is mentioned, there is a reference to the relevant section. The argumentation is clear and logical. From the comments on the results it is clear that the student has good understanding of the topic.

5. Formal aspects of technical report — **85 p. (B)**

The English is generally good. A few, but really not many, sentences are were a bit unclear to me. The thesis is well edited. I spotted only very few typographical mistakes.

6. Literature usage — **85 p. (B)**

The student has surveyed several papers with state-of-the art techniques which are cited clearly and correctly. He has also utilized existing tools when available and clearly stated this. The information about image sources could, however, have been clearer. Often (e.g. on p9) it is not clear from putting the reference in the caption that the image is taken from that paper.

7. Implementation results — **85 p. (B)**

The proposed solutions have been evaluated according to standard protocols. Occasionally, the statistical significance of the result could have been better considered. (For example the improvement of the final fusion 1.45% over the system with feature extraction EER 1.46% might not be statistically significant.)
The accompanying code is well documented. As far as I can judge, external tools have been used according to the licences.

8. Utilizability of results

The work extends RawNet with more powerful architectures and losses used in other methods. The work therefore combines several existing techniques. The experimental work is very systematic and the improvement over the original RawNet is substantial. Therefore this work can serve as good baseline for future research on speaker verification without explicit feature extraction.

9. Questions for defence

On p6 you say that 2D convolutions such as in ResNet are ideal when the input is a feature and that 1D convolutions are ideal for processing raw waveforms. But what about having first one or more 1D convolutions that extracts "features" from the raw waveform and then continue to process them with 2D convolutions?

What kind of patterns do you think the RawNet can extract from the waveform that are missing in standard features such as fbank or MFCC?

10. Total assessment — **85 p. very good (B)**

The student have carefully studied and implemented many different techniques and done exhaustive experiments. The thesis is well written, and it is clear that the student understands well what he is doing. For a "A mark" it would have been necessary to do something beyond combining existing techniques.

In Brno 9 June 2021

Rohdin Johan A., Dr.
reviewer