Report on the dissertation

“Semi-supervised Training of Deep Neural Networks for Speech Recognition

presented by Karel Vesely

Reviewer: Dr. Lori Lamel
CNRS-LIMSI, BP 133, 91405 Orsay Cedex, France

This thesis addresses the challenge of reducing the development cost of acoustic models for speech recognition. The document is very well written and a pleasure to read. It is organized in 10 Chapters including the introduction and concluding remarks. The core of the thesis is comprised of Chapters 2-5 providing background and setting the context for the experiemental work presented in Chapters 6-9. In particular I appreciate the author’s self-reflections on his own previously published work as his research evolved. This self-assessment is a demonstration of his maturity as a researcher.

I would also like to emphasize the contribution of the PhD candidate to the wider speech community. His nnet1 training procedure has been integrated in the widely used Kaldi toolkit and is therefore availalbe to the worldwide speech community.

The introduction to the thesis highlights some of the main progress and challenges in automatic speech recognition, motivating the thesis work and explicitly listing a number of questions addressed in the thesis.

The second chapter is dedicated to the use of Neural Networks in speech recognition. The author highlights that although these approaches were originally proposed about 30 years ago, it is only during the last decade that they have seen wide use by the community due to the increase in computing power and GPUs. This overview reads quite nicely, and highlights points that are of particular relevance to the thesis research. It might have been nice in some places to use terms like ’generally’ or ’typically’ to let the reader know that there may be other options than what is mentioned.

Chapter 3 describes the ’nnet1’ DNN training recipe develeped for and implemented in the Kaldi open-source toolkit. The strategies implemented to improve the training efficiency are presented. This toolkit, along with the contributed training algorithm and some more recent versions, is used by researchers and students worldwide. The paper cited on the Kaldi website for the nnet1 training setup, “Sequence-discriminative training of deep neural networks”, published in Interspeech 2013, has already been cited over 400 times. After an analysis of the computational speed up due to algorithmic and hardware improvements, this chapter ends with a comparison with other training implementations used by other state-of-the-art NN training toolkits.

The corpora used in the experimental work are described in Chapter 4. These are data from the IARPA Babel program (principally conversational telephone speech in Vietnamese) and the widely used Switchboard corpus of conversational telephone speech in English. These corpora are publicly available allowing this work to be easily compared with or replicated by other researchers.

The fifth chapter is devoted to an introduction to semi-supervised training. While this chapter provides a nice overview of related work, a more comprehensive literature review would have cite some of the early work on this topic, such as Zavaliagkos & Colthurst, 1998; Kemp & Waibel, 1999, Lamel, Gauvain & Adda, 2000, as well as the work of Yu, Gales, Wang, & Woodland (2010) which explored the use of unsupervised discriminative training of GMM/HMM models even though the techniques were generally applied to broadcast data rather than conversational telephone speech. In particuler the paper by Yu et al., 2010 would be appropriate to cite in section 5.1 in relation to active learning. In this same section, it might be nice to remind the reader that WER recovery can only be measured if the full set of data are transcribed, which is not the case for many practical applications where semi-supervised training is of interest to shorten the system development period and/or costs. Two other related works that could be cited are lattice-based unsupervised training that by Fraga-Silva et al, 2011 and the thesis of Christian Gollan (RWTH, 2014) that considered some of the issues explored in this work. As a general comment, I particularly appreciate the comments provided about the techniques in other proposed work and how these relate to this thesis research.

Initial experiments with semi-supervised training are reported in Chapter 6. Here different usages of confidence score are explored (frame vs sentence level, data selection/weighting) to assess their impact on semi-supervised training using the iARPA Babel Vietnamese corpus. Training using the automatic transcripts is shown to improve the WER, with an additional small improvement when filtering out about 20% of the data with the lowest frame-level confidence scores. In one of the early experiments a small gain was obtained by duplicating transcribed data, but this was later found to not always be helpful when using weighted training data. A natural question is whether or not similar experiments were made with data perturbation? It also would have been interesting to compare the LLP and LLP+SST AMs using the same LM (FullLP). The callibration of confidence scores is explored with the aim of improving frame-weighted training. Absolute WER improvements of 2-3% are obtained for 5 Babel languages with the proposed semi-supervisd training scheme with confidence score calibration and sMBR tuning using the supervised data. The author however points out that these results are not replicated in other studies he later carried out.

Chapter 7 returns to the questions initially posed about semi-supervised training. These include Oracle experiments to access which confidence level should yield the largest performance improvement, followed by exploration of confidence score calibration, data selection and weighting. This chapter includes extensive comparative experiments and many nice illustrative figures to enhance understanding and concludes with some guidelines for semi-supervised training, which is the topic of next chapter (8). The overall finding is that word selection (confidence ordered) to roughly match the system accuracy is a reasonable generic approach.

In the final chapter several interesting conclusions are drawn from the extensive experimental work performed, with an objective view of the utility of the different research questions which were explored.

Below I explicitly address each of the guidelines provided for the thesis review.

- *Is the topic appropriate to the particular area of dissertation and is it up-to-date from the viewpoint of the present level of knowledge?*

  Yes, completely. The thesis addresses an important challenge in training state-of-the-art acoustic models.

- *Is the work original and does it mean a contribution to the area - specify where the original contribution lies?*

  There are several original and lasting contributions of the thesis work. One notable contribution is the nnet1 training procedure distributed as part of the Kaldi toolkit. Semi- and unsupervised have received growing attention in recent years for HMM/GMM systems, but the community as a whole has less experience in the context neural network based systems which have taken over as the state-of-the-art.

- *Has the core of the doctoral thesis been published at an appropriate level?*

  The thesis work has been published in the most prestigious and relevant international conferences in the field (IEEE ICASSP, ASRU, SLT, ISCA Interspeech).

- *Does the list of the candidates publications imply that he is a person with outstanding research erudition?*

  Yes. Karel Vesely has 30 publications, four of which have over 100 citations and one almost 2000. He has an h-index of 15, which is quite impressive and rare for a doctoral student.

  In 2017, he contributed to a journal article in *Computer, Speech & Language* and a chapter in the book *New Era for Robust Speech Recognition*.

- *More characteristics of the candidate may be added here.*

In conclusion, it is without any reserve that I consider that this doctoral thesis meets the requirements of the proceedings leading to PhD title conferment.

Lori Lamel, PhD, HDR
Directeur de recherche (DR1), CNRS
Orsay, 20 February 2018