March 27, 2009

Mgr. Sylva Sadovská
DEAN FIT BUT
Božetěchova 2
612 66 Brno
Czech Republic

To Whom It May Concern:

I am happy to present my review of the doctoral thesis of Martin Karafiat entitled "Study of Linear Transformations Applied to Training of Cross-domain Adapted Large vocabulary Continuous Speech Recognition Systems" submitted to Brno University of Technology in 2008.

The contributions of the thesis fall in two categories and may be summarized as follows. First, the author examines the widely used HLDA feature transformation technique and proposes three improvements. The first is a smoothing approach to obtain more robust variance estimates from the data. Second, he suggests deriving these statistics by MAP-adaption for cases where out-of-domain background training data is involved. Finally, he shows that a simple strategy of deweighting non-speech frames in the training data can improve results. While the improvements obtained with these techniques are small they are certainly worthwhile given that they can be implemented with low overhead.

The second, and main contribution of the thesis consists of a method to deal with bandwidth mismatch between training and data in the training of acoustic models. The author shows that one can do better than the standard downsampling of all data to a common bandwidth, by estimating a linear transform (either in model space or feature space) that maps from high-bandwidth features to low-bandwidth features. Such a transform can be estimated using MLLR or constrained MLLR (CMLLR), and is particularly convenient to apply as a feature-space transform. It is showns that the feature mapping approach gives on the order of 5% relative improvement over the downsampling approach. The details of the feature mapping approach in the context of HLDA, speaker-adaptive training, and discriminative training with the MPE criterion are carefully studied and presented clearly, in a way that will be highly useful to practitioners in the field.

These two contributions are presented in chapters 5 and 7, respectively, of the thesis. Chapters 1 through 4 and Chapter 6 present the underlying speech recognition algorithms and modeling techniques, and Chapter 8 summarizes and suggests future work. The work is well-written and -presented. I have a few detailed comments and suggestions that I append to this letter.

SRI International

Andreas Stolcke • EJ119 • 333 Ravenswood Ave. • Menlo Park, CA 94025 • (650) 859-2544 • Fax (650) 859-5984 • http://www.sri.com

The scholarship of the thesis is of high quality, and the relevant prior work is referenced appropriately. I consider both of the lines of work summarized above as original and making a significant contribution to the advancement of the field of speech recognition. The second contribution especially – dealing with bandwidth mismatch – is of great practical importance, since this kind of mismatch occurs frequently and otherwise limits the spread of technology to real-life applications. Meeting recognition, the domain considered in the experiments, is an especially hard problem, and any methods improving recognition accuracy are of great interest.

The candidate has published the main contributions in a series of papers papers at major international conferences. The work is of sufficiently high quality and interest that I would strongly suggest a journal publication as well.

In summary, I believe that Martin Karafiat's thesis meets the highest standards and merits conferment of a Ph.D. degree.

Sincerely,

Andreas Stolcke, Ph.D.
Sr. Research Engineer
SRI International /
International Computer Science Institute

Appendix: Comments and Suggestions

1. Section 4.3: It is not clear what segmentation algorithm was applied to the test data, or if manually provided segmentations were used, since this can have a large effect on meeting recognition accuracy.
2. Chapter 7: It would be important to know if the NB-WB feature transformation approach applies equally well to MFCC and PLP features, since MFCC are widely used, yet the thesis looks only at PLP features.
3. Table 7.9 and following text: The table gives a WER of 26.5%, but the text has 26.6%. Please fix this inconsistency.
4. Table 7.11 and Figure 7.14 and surrounding text: the designations "NBWB" and "WBNB" seem to be used variously. It would be better to make all system labels consistent.