LINCOLN LABORATORY
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
244 WOOD STREET
LEXINGTON, MASSACHUSETTS 02420-9108

23 October 2012

To Whom It May Concern:

I have finished reviewing the thesis of Ing. Ondrej Glembek, "Optimization of Gaussian Mixture Subspace Models and Related Scoring Algorithms in Speaker Verification." Overall, the thesis is an insightful, pragmatic, and incisive investigation into new methods for subspace models for speaker recognition. The topic covered is appropriate and focused on a core concern for researchers in speaker recognition. The techniques and insight are modern and show good understanding of the subject matter.

The work proposes a number of different novel contributions:

- A new linear scoring method is proposed along with analysis of "competing" methods.
- Methods for fast computation of factors are explored—methods that perform well are found.
- Discriminative techniques are applied to hyperparameter training with moderate success.

Overall, this reviewer is impressed by the scientific integrity and thoroughness of the thesis (a tradition for the BUT research group). The author presents, examines, and derives many methods throughout the thesis and evaluates them without undue advocacy for a particular approach. Examples include:

- Analysis of scoring in Chapter 4 finds a low-computation high-accuracy approach. Methods are examined systematically and the best solution using multiple criteria is found.
- Simplification of i-vectors shows novel methods of finding approximate factors. Methods are compared on the basis of accuracy, computation, and memory usage. A significantly faster and reasonably accurate approach is found.
- A discriminative training is derived which shows good potential. A straightforward evaluation of the methods shows the author is not overselling the methods.

The thesis demonstrates the proficiency of the author at speaker recognition and certainly the author's contributions have been recognized in the field. In addition, the thesis is written at the appropriate level for PhD. Minor comments and typographical errors are listed after this letter. Ing. Ondrej Glembek should address the comments to make the thesis a more complete work.

In summary, without hesitation, I conclude this thesis meets the requirements of the proceedings for the title of PhD.

Sincerely,

Dr. William M. Campbell
Senior Member of the Technical Staff
MIT Lincoln Laboratory

Comments/suggestions (page # followed by comment):

1. 2: Additional and earlier references on prosody (Kockmann2012)? How about Adami, Dehak, SNERFs, etc.
2. 5: Additional and earlier references on mean/variance normalization (Young, 2006), earlier refs., e.g. Homomorphic Signal processing (Oppenheim), CMS, etc.
3. 15: first term in (2.7), shouldn't it be e^{-s_x}, not e^{s_x^{-1}}
4. 16: "Calibration is understood as an affine transformation …" This is certainly the case for simple use cases of calibration, but not in the general forensic case (e.g., varying duration). How about "For calibration, we use a standard affine transformation of the scores" or "In our experiments, an affine transformation of scores is used for calibration."
5. 28: flat priors are -> a uniform prior is
6. 30: Most early successful subspace modeling was for channel variability. Add mentions and references to Solomonoff, 2004 and Hatch 2006. Note that subspaces for speakers was also explored in 2006/2007 where WCCN was first applied (see NAP and WCCN: Comparsion of Approaches Using MLLR-SVM Speaker Verification System, Kajarekar and Stolcke); this was only moderately successful at the time. The success of subspace modeling for speakers came 2-4 years later in JFA and i-vectors (2008 NIST SRE, JHU Workshop and 2010 NIST SRE).
7. 32: But really in equation (39), the cross entropy (in (38)) between the UBM mixture weights and the alignment estimated mixture weights is not "constant"—it varies with the utterance, correct? Isn't m'N_i\Sigma^-1m also "constant", then?
8. 35: Actually the PCA (and KPCA) estimation came out of discussion on how Solmonoff04, 05 NAP was trained; within class covariance was used for NAP for supervectors with an Arnoldi/Lanczos eigenvector iterative solver. The PCA (and KPCA) solution was borrowed for channel FA in Brummer04 and used as an alternate to the random initialization in P. Kenny's papers.
9. 36: There's a serious omission that NAP is not mentioned as a subspace modeling method. It arrived at the same time as FA for channel variability, 2004.
10. 42: Both Vair and Brummer's paper uses FA not JFA (there's no speaker factors). Actually the phrase, "their results use different scoring methods" is correct, but a little misleading. The most common method at the time was standard frame-by-frame scoring.
11. 45-46: The process of getting to (4.23) is more awkward than it should be including the explanation. There are references back to the prior chapter, a mention of ignoring constants in the Taylor expansion, and the dropping of the cross-entropy mixture term (mentioned above). Possibly it would be better to just indicate at the beginning of page 45 that the Taylor approximation will be used and terms not involving the means will be discarded as not having speaker information.
12. 45: Isn't a 1/T (T=#frames) normalization needed on 4.24 to make it usable? This should be mentioned somewhere (although it is consistent with 4.10)

13. 51: The intro is very nice and really sums up the success at the workshop. One thing that might be mentioned on the technical side is that stacking was commonly used for JFA, m = [U V][x;y], and the idea of replacing [U V] with a T and replacing the stack of factors with a "single" set of factors is also a natural progression of ideas.
14. 51-54: One thing that might disturb or confuse the reader of this work is the abrupt change in "paradigm" when going from JFA to i-vectors. Some explanation about why this was done is needed. Here's the issues to address:
    a. You spent Chapter 4 convincing us that equation (4.23) was the "right" way to score (complexity, accuracy). Yet in Chapter 5, this scoring is abandoned. A good transition explanation is needed. Some ideas:
    b. JFA versus i-vector symmetry: JFA is train/test and is asymmetric, i-vector is more train/train--treats both train and test the same (there's no reason not too).
    c. JFA scoring for ivector: Couldn't you find a model by working back from the factors the supervector space and use this in JFA scoring? One reason not to do this explicitly is that supervectors are big and factors are "small." It's easier to do scoring in the small space and it's possible to do more complex scoring techniques (as in "standard" vector classification).
15. 59: It might clarify the discussion for many readers if you point out that (5.6) is equivalent to a regularized normal equations approach (a very well-researched problem in linear algebra). The problem is that the regularization term is dependent on the data. The new approaches try to work around this issue with standard linear algebra tools—eigenvalues, etc. In the current discussion, it appears that this is an isolated problem that is being solved with new methods.
16. 69: "In speaker recognition, the use of discriminative training has been very limited." A fairer statement might be "In speaker recognition, the *successful* use of discriminative training has been very limited." All kinds of discriminative approaches have been tried—neural networks, NTNs, MMI, decision trees, polynomial classifiers--but few were widely adopted. Part of the problem is that speaker recognition enrollment has limited training, so discriminative approaches are tricky.

Typos/grammar (page # followed by correction):

- Vii: "Not only I got" -> "Not only did I get"
- 1: moder -> modern
- 3: used methods -> methods used in this thesis
- 3: standardly used -> standard
- 6: we assure -> we assume
- 9: in deep -> in detail
- 9: the JFA -> JFA
- 9: the PLDA -> PLDA
- 9: Joint Factor Analysis -> Joint Factor Analysis (JFA) [define your abbreviation]
- 11: gets familiarized -> is familiar [or "is familiarized"]
- 11: developped -> developed

- 11: positively corelate with -> be similar to [at lease correct the word “corelate” -> “correlate”]
- 13: (footnote) preferabely -> preferably
- 13: (footnote) addopted -> adopted
- 15: see the optimal -> obtain the optimal
- 15: usefull -> useful
- 21: Substituting (3.4) -> Substituting (3.3)
- 32: fist order -> first order
- 38: hiden -> hidden
- 62: algorith -> algorithm
- 63: got feasible -> became feasible
- 69: shows the situation -> shows the setup
- 90: ML estimat -> ML estimate
- Bibliography: Is there a way to organize by Author/Date? Currently finding references is moderately difficult; e.g., Burget, et. al. 2008 is before Burget, et. al. 2007.

Asides (not critical to the thesis):

- 46: If (4.24) is normalized by the number of frames (1/T), then the right most term (f-Nm-Nc) can be written as (1/T)*f – (1/T)Nm-(1/T)Nc. If N is invertible (it works out even if this isn’t true), then this can be written as Lambda*(m_spk-m_ubm-c). So the inner product looks like (Vy+Dz)’\Sigma^(-1)*\Lambda* (m_spk-m-c). So, this is interesting, since it is using the mixture weights estimated from the test utterance. The earlier inner product from Campbell06 used \Lambda from the UBM, so these scoring equations are closely related, but derived completely differently.
- 46: (4.24) Linear JFA scoring is very asymmetric. Why not use (Vy+Dz) also for the test utterance also?—in our experiments this didn’t work nearly as well. One might argue that if JFA produces better estimates, Vy+Dz should always be used for both train and test. Also, the \Lambda is from test; it’s possible to do score_new(a,b)=0.5*score(a,b)+0.5*score(b,a), but it looks very complex and it’s also not an inner product. One disturbing thing is that when going to i-vectors, the use of \Lambda goes away completely.