



405, avenue Ogilvy, bureau 101  
Montréal (Québec), Canada H3N 1M3  
Telephone : (514) 840-1235 x 4624  
Fax : (514) 840-1244

Patrick.Kenny@crim.ca  
<http://www.crim.ca/perso/patrick.kenny>  
<http://www.crim.ca/>

**Report on Marcel Kockmann's PhD Thesis  
"Subspace Modeling of Prosodic Features for  
Speaker Verification"**

**Patrick Kenny**  
March 2012

The author demonstrates a clear mastery of the literature, evaluation protocols, signal processing techniques, and modeling methods in the speaker recognition field. The literature review and statement of his thesis problem is clear and succinct and could serve as a model for other PhD thesis candidates.

The central claim of the thesis, namely that major improvements in speaker recognition accuracies as measured on standard test beds (the NIST speaker recognition evaluation corpora) can be achieved using carefully extracted prosodic features combined with i-vector modeling and Probabilistic Linear Discriminant Analysis (PLDA) is amply demonstrated. The extension of i-vector modeling to handle multinomial distributions is original and the results obtained with SRI's SNERF features are excellent.

I think I can safely assert that, prior to the completion of this thesis, no expert in the field would have believed that a speaker recognition architecture designed using only prosodic features would have been capable of attaining equal error rates on the order of 5% on standard benchmarks. I would like to take this opportunity to congratulate Marcel on this outstanding result.

An important contribution is the careful evaluation in Chapter 3 of the various ways of extracting energy and pitch contour features. The author has done a much better job here than the rather slap-dash approach taken by Dehak and myself in [Dehak 2007]. I must own up to responsibility for this as I had a particular agenda to promote, namely to show that, even with a very casual approach to feature extraction, Joint Factor Analysis modeling of prosodic features was a match for the carefully engineered SNERF features developed by SRI. It is satisfying that the author's work on extracting contour features paid off but I think it would have been more helpful to the reader to highlight the final shift and overlap configuration more clearly than in the paragraph of Section 3.1.4 beginning "In later experiments ...". The author is also to be credited with getting SRI's permission to use their SNERF features so that he could address the problem of building the best possible prosodic speaker recognizer in a thoroughly systematic way.

Regarding the comparison of JFA and i-vector + PLDA modeling of contour features, the results are quite striking in that the difference between the results obtained by the two approaches (Tables 4.1 and 4.2) seem to be larger than one would expect from the results obtained with cepstral features.

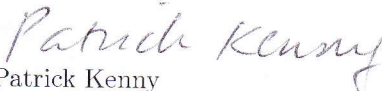
The main theoretical contribution of the thesis is the extension of the i-vector methodology to modeling "supervectors" made up of multinomial distributions. Because of its novelty, I would have liked to see a few more details in the presentation. The basic idea borrows from the way Povey and Burget handle mixture weights in the subspace GMM model for speech recognition, but the implementation seems to differ in some ways. If I remember correctly (this may be based on a recollection of an early version of their work), Povey and Burget use a simplified approach whereby they estimate the vector  $w$  in (4.71) using ordinary GMM supervectors and then use estimates obtained in this way to update the matrix  $T$  in (4.71). I think it might be helpful to clarify this point for the reader. Also, I found the "max" in (4.76) puzzling and was left wondering if this was just a crude fix to ensure positive definiteness of  $H_u$ . (If that is the

case, there is no harm in saying so, as the effectiveness of the method seems to be unequivocally demonstrated in Fig. 4.3.) Another thing I would have liked to see would be an eigenvalue profile of the matrix  $TT^*$ ; this would give some insight into the question of whether the very high dimensional collection of multinomial distributions defined by SNERFs can be reasonably modeled as low dimensional (of course this question is addressed in Fig. 4.4 but only indirectly). That said, the experimental results reported in Table 4.3 are very impressive.

The i-vector fusion technique presented in Chapter 5 has a nice appeal. I find it very interesting that the results in line 4 of Table 5.1 are uniformly better than those in line 5 (and that a similar tendency is evident in Tables 5.2 and 5.3) — in other words, that i-vector fusion is generally better than Niko Brummer's score based fusion. The issue here in my mind is that i-vector fusion is a purely generative method (so that it only gets to look at target trials in training) whereas score based fusion is discriminative (so that it gets to look at non-target trials as well as target trials). Thus the i-vector fusion method would appear to be at a disadvantage *a priori*, something that I think would be worth highlighting. (However it should also be mentioned that i-vector fusion would need to be followed by a calibration stage in practice and that would require a development set containing non-target as well as target trials.)

If the author sees fit, it might be a good idea to include some of the purely prosodic results (DCT + SNERF-iV) presented in Section 5.1.1 in the tables in Section 5.3 which sum up the final results of the thesis in order to highlight the effectiveness of the prosodic features. I would also advise supplying some historical context in the concluding section so the reader can appreciate the importance of the author's achieving an equal error rate of 5.4% on telephone speech using *only* prosodic features. If memory serves, a state of the art system in 2005 (involving a fusion of, say, a dozen heterogeneous subsystems) would not have achieved a better result.

The author has published his work extensively in the proceedings of the flagship conferences in the field (Interspeech and ICASSP) and *Speech Communication*. He has been nominated twice for best student paper awards and participated actively in several NIST evaluation campaigns and related workshops. It is my judgment that he has satisfied all of the requirements for the PhD degree and I wholeheartedly recommend that the degree be awarded.

  
Patrick Kenny