

DEVELOPMENT OF THE PRIMARY CRIM SYSTEM FOR THE NIST 2008 SPEAKER RECOGNITION EVALUATION

Patrick Kenny, Najim Dehak, Pierre Ouellet, Vishwa Gupta and Pierre Dumouchel

Centre de recherche informatique de Montréal (CRIM)

{Patrick.Kenny, Najim.Dehak, Pierre.Ouellet, Vishwa.Gupta, Pierre.Dumouchel}@crim.ca

ABSTRACT

We describe how we modified the CRIM factor analysis speaker verification system to handle the new cross-channel conditions encountered in the 2008 NIST speaker recognition evaluation. Using the 2006 evaluation data for development, we obtained results on a broad spectrum of test conditions that are uniformly better than the best results that have been published in the literature.

Index Terms: speaker verification, factor analysis

1. INTRODUCTION

Cross-channel tests, where telephone speech was used to enroll target speakers and microphone speech for verification trials, were introduced in the NIST 2005 and 2006 speaker recognition evaluations (SRE) [1]. In the 2008 evaluation, both types of speech are encountered in enrollment as well as testing. This paper describes how we modified our factor analysis system to handle this diversity of channel effects.

We begin with a brief summary of factor analysis based speaker recognition. Let C be the number of components in a Universal Background Model (UBM) and F the dimension of the acoustic feature vectors. We use the term supervector to refer to the CF -dimensional vector obtained by concatenating the F -dimensional mean vectors in the GMM corresponding to a given utterance.

We assume firstly that such a supervector M can be decomposed into a sum of two supervectors, a speaker supervector s and a channel supervector c :

$$M = s + c, \quad (1)$$

where s and c are statistically independent and normally distributed.

Secondly, we assume that the distribution of s has a hidden variable description of the form

$$s = m + vy + dz \quad (2)$$

where m is a $CF \times 1$ supervector; v is a rectangular matrix of low rank and y is a normally distributed random vector; d is a $CF \times CF$ diagonal matrix and z is a normally distributed CF -dimensional random vector. We will refer to the columns of v as eigenvoices and we will refer to the components of y as speaker factors.

Thirdly, we assume that the distribution of c has a hidden variable description of the form

$$c = ux, \quad (3)$$

where u is a rectangular matrix of low rank and x is a normally distributed random vector. We refer to the components of x as channel factors and we use the term eigenchannels to refer to the columns of u .

Equation (2) can be interpreted as saying that speaker supervectors are normally distributed with mean $\mathbf{0}$ and covariance matrix $d^2 + vv^*$. We interpret this normal distribution as a prior distribution in the sense in which this term is used in Bayesian statistics. Given an enrollment utterance and the hyperparameters m , u , v and d , we enroll a target speaker by calculating the posterior distribution of the hidden variables x , y and z , using the maximum a posteriori estimate of $m + vy + dz$ as a point estimate of the speaker's supervector. (We do not use the point estimate of x .) Details can be found in Section III of [2].

At verification time, we match a speaker supervector s with a given test utterance using equation (3); that is, we assume that the supervector for the test utterance has the form $s + ux$ where x is random.

So far we have implicitly assumed that the same channel effects obtain both at enrollment time and at verification time, so that the same matrix u can be used in both cases. In this paper, we will explain how we handled the diversity of channel effects in the 2008 SRE by modifying the matrices u used at enrollment and verification times on a case-by-case basis. (Similarly, we modified the imposter cohorts for zt -norm.) We also present new results on the 10 second test conditions of the 2006 SRE.

2. FACTOR ANALYSIS TRAINING

2.1. Acoustic features

We extracted 19 cepstral coefficients together with a log energy feature using a 25 ms Hamming window and a 10 ms frame advance. These were subjected to feature warping using a 3 s sliding window. Δ and $\Delta\Delta$ coefficients were then calculated giving a total of 60 acoustic features.

There is one circumstance in which we and other authors have found $\Delta\Delta$ coefficients to be unhelpful, namely the NIST 10sec-10sec condition where only 10 seconds of speech data are available for enrolling target speakers [3]. Thus we only used 40 dimensional features for the 10sec-10sec tests reported in this paper; on the other hand we used the full 60 dimensional feature set for the other 10 second tests in which larger amounts of data is available for enrolling target speakers.

2.2. Factor analysis configurations

We trained two gender-dependent UBM’s having 2048 Gaussians and gender-dependent factor analysis models having 300 eigenvoices estimated from telephone speech, 100 eigenchannels estimated from telephone speech and 100 eigenchannels estimated from the auxiliary microphone development data provided by NIST prior to the 2006 SRE.

Again we made an exception for the 10sec-10sec tests where we reduced the number of Gaussians to 1024 (but kept the configuration otherwise unchanged).

2.3. Factor analysis training data

For training UBM’s and factor analysis models we used the LDC releases of Switchboard II, Phases 2 and 3; Switchboard Cellular, Parts 1 and 2; and the NIST 2004 SRE and 2005 SRE data. We used only those speakers for which five or more recordings were available for factor analysis training.

We used the 2006 SRE data to determine the decision thresholds which minimize the value of the NIST detection cost function (DCF) on the various enroll/test conditions. These are the decision thresholds that we used for our primary system in the 2008 SRE.

The evaluation protocol allows each site to submit up to two additional “contrastive” systems. To produce a contrastive system, we retrained the factor analysis models by adding the 2006 SRE data to the factor analysis training set and we used the same decision thresholds as in our primary system. (We do not report any results obtained with this system in this paper since we would need a test set disjoint from the 2006 SRE data to obtain meaningful results.)

2.4. Decoupled estimation of v and d

Equation (2) is not a realistic model of inter-speaker variability and joint maximum likelihood of v and d is not the most effective estimation procedure. As explained in [4], we have been able to obtain much better results by estimating v and d on disjoint subsets of the factor analysis training data. Thus we estimated v by fitting a pure eigenvoice model with 300 eigenvoices to the subset of the factor analysis training data defined by withholding the NIST 2004 SRE data; we then estimated d in such a way as to account for the speaker variability in the NIST 2004 SRE data that this estimate of v could not account for.

To implement this, we pooled all of the recordings of each speaker in the factor analysis training set and ignored channel effects. (The rationale here is that channel effects can be averaged out if sufficiently many recordings are available for each speaker.)

A consequence of using decoupled estimation is that factor analysis training runs very quickly. This enabled us to work with the large factor analysis configurations described in Section 2.2.

2.5. Eigenchannel estimation

We decoupled the estimation of the eigenchannels from that of v and d as in [5]. We estimated a set of 100 eigenchannels from the factor analysis training data described in 2.3 (which consists solely of telephone speech) and we estimated another set of 100 eigenchannels on the development data released by NIST for the auxiliary microphone tests in the 2006 SRE.

3. SPEAKER RECOGNITION

3.1. Channel modeling for enrollment and verification

As summarized in Table 1, we used the telephone and microphone eigenchannels in different ways depending on whether the enrollment data and test data consisted of telephone or microphone speech. For example, in situations where we were given telephone speech for enrollment and microphone speech for testing we used 100 telephone eigenchannels to model channel effects in the enrollment data and 200 eigenchannels (100 telephone and 100 microphone) to model channel effects in the test data.

Table 1. Numbers of channel factors used for enrollment and testing.

enroll/test	enroll	test
tel/tel	100 tel	100 tel
tel/mic	100 tel	100 tel + 100 mic
mic/tel	100 tel + 100 mic	100 tel
mic/mic	100 tel + 100 mic	100 tel + 100 mic

3.2. Likelihood computation for verification decisions

At verification time, likelihoods were evaluated according to (19) in [5]. (We did not use the correction (20) in [5]. This is a minor technical issue which is discussed at length in [6].) Thus, we account for channel effects in test utterances by integrating over the channel factors x in (3) rather than by using a point estimate of the channel factors for each test utterance as other authors do.

This point seems to be worth stressing. If a test utterance is sufficiently long, the posterior distribution of the channel factors will be sharply peaked and using a point estimate of the channel factors (either a MAP estimate or a maximum likelihood estimate) will give essentially the same result as integrating over the channel factors. But in the case of short test utterances (say 10 seconds of speech), integrating over channel factors seems to be the right thing to do. (Since the integral in question is Gaussian there is no difficulty in evaluating it in closed form.) It was reported in [3, 7, 8] that channel factors are unhelpful for tasks involving short test utterances but this does not agree with our experience. In this paper we will present some good results on 10 second test conditions; we believe that our success can be traced to *not* attempting to obtain point estimates of channel factors under these conditions.

3.3. Imposters

The likelihoods obtained at verification time were normalized using z -norm. As explained in [6], we found it useful to use exceptionally large numbers of imposters for this purpose.

The way we selected t -norm models and z -norm utterances for the various enrollment and test conditions is summarized in Table 2. For example, in situations where we were given telephone speech for enrollment and microphone speech for testing we used 300 telephone speech utterances to create t -norm models and 1000 microphone speech utterances for z -norm. Note that we made no attempt to match the t -norm models with the quantity of data available for enrollment (which can vary from 10 seconds to 20 minutes) nor the z -norm utterances with the quantity of data available at verification time. (Our experience in the past has been that this is not helpful.)

Table 2. *Imposters used for zt-norm (numbers are approximate).*

enroll/test	<i>t</i> -norm	<i>z</i> -norm
tel/tel	300 tel	1000 tel
tel/mic	300 tel	1000 mic
mic/tel	300 mic	1000 tel
mic/mic	300 mic	1000 mic

4. TESTS ON WHOLE CONVERSATION SIDES

The principal difference between the NIST 2006 and 2008 SRE’s is that cross-channel tests — where channel conditions vary markedly from enrollment time to verification time — are mandatory in 2008 but optional in 2006. In 2006, the only cross-channel tests involved telephone speech for enrollment and microphone speech for verification (tel/mic for short) but in 2008, the mic/tel and mic/mic conditions are also encountered.

In most cases we were able to simulate the conditions of 2008 SRE using the index files that served to define the various conditions of the 2006 SRE but in some cases (such as mic/mic) we had to construct our own test sets.

4.1. Core condition

The term “core condition” refers to the situation where a whole side of a telephone conversation is available for enrolling each target speaker and the verification test data also consists of whole conversation sides. In the past, it was used to refer to the tel/tel case exclusively. This condition was also referred to as 1conv-1conv; in the tel/mic case the terminology 1conv-1convmic was used.

The results we obtained on the 1conv-1conv (tel/tel) condition of the NIST 2006 SRE with the primary system described in Section 2.2 are summarized in Table 3.¹

For comparison, the best results that have been published on this task are those of STBU [9]; this system achieved an EER of 2.3% (English language trials only, results pooled over male and female speakers) by fusing 10 subsystems (cepstral and MLLR).

Table 3. *1conv-1conv NIST 2006 SRE*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	1.4%	0.009	2.4%	0.018
All trials	2.9%	0.016	3.5%	0.022

In the 1conv-1convmic (tel/mic) condition in the 2006 SRE, the enrollment data for each target speaker consists of a conversation side extracted from a recording of a telephone conversation but the test data consists of recordings made using one of 8 different microphones. (The identity of the microphone is not given. The task is described in detail in [10].) Our results are summarized in Table 4.

For comparison, MIT’s results are the best that have been published on this task [10]; an EER of 4.0% was obtained by MIT by

¹All results in this paper were obtained using version 4 of the 2006 SRE answer key. This ensures that the comparisons with results reported by other sites are fair but NIST is expected to release a new version of the key in April 2008.

Table 4. *1conv-1convmic NIST 2006 SRE*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	2.4%	0.009	2.5%	0.015
All trials	2.5%	0.009	2.9%	0.016

fusing two cepstral systems (a support vector machine with nuisance attribute projection and a GMM/UBM system with channel factors). Speech enhancement played an important role in reducing error rates but our system makes use of no special-purpose signal processing.

There was no 1convmic-1conv (mic/tel) condition in the 2006 SRE but this condition is easy to simulate by interchanging the roles of enrollment and test utterances in the 1conv-1convmic condition. Under these circumstances we obtained the results reported in Table 5.

Table 5. *1convmic-1conv NIST 2006 SRE*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	2.9%	0.011	3.1%	0.021
All trials	2.9%	0.011	3.3%	0.021

Nor was there a 1convmic-1convmic (mic/mic) condition in the 2006 SRE but Doug Reynolds kindly provided one (consisting of 5 K target and 148 K non-target trials) derived from the test utterances in the 1conv-1convmic condition. The results we obtained are reported in Table 6.

Table 6. *1convmic-1convmic*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	4.8%	0.034	5.2%	0.034

4.2. Extended data condition

In 2006 SRE, the 8conv-1conv (tel/tel) and 8conv-1convmic (tel/mic) conditions were offered; “8conv” indicates that 8 conversation side were available for enrollment. It happens that some of the subjects in these tests were exposed in 2005 and so found their way into our factor analysis training set (Section 2.3). We exclude such speakers in reporting the results of these tests.

Our results on the 8conv-1conv test are summarized in Table 7. For comparison, the best results on this task in the literature are those reported by MIT/IBM [11] where EER’s of 1.5% (English language trials, male and female results pooled) and 2.6% (all trials) were obtained by fusing 9 subsystems (cepstral, MLLR and higher level). It is interesting to note that, although the extended data task was intended to encourage research into higher level systems, and higher level systems (including an MLLR system) play an important role in reducing the error rates in [11], we were able to obtain better results using cepstral features alone.

Our results on the 8conv-1convmic test are summarized in Table 8.

Table 7. *8conv-1conv NIST 2006 SRE. Speakers exposed in 2005 and recycled in 2006 excluded.*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	0.5%	0.003	1.7%	0.017
All trials	1.1%	0.007	1.8%	0.010

Table 8. *8conv-1convmic NIST 2006 SRE. Speakers exposed in 2005 and recycled in 2006 excluded.*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	0.6%	0.003	1.1%	0.003
All trials	0.7%	0.003	1.3%	0.009

5. TESTS ON 10 SECOND UTTERANCES

Despite their evident practical interest, little has been published on the 10 second test conditions apparently because they are perceived as being too difficult. On the whole conversation side test conditions, major progress has been made in recent years thanks largely to powerful channel compensation techniques such as channel factors and nuisance attribute projection. Some authors have found that these methods are ineffective on 10 second test conditions [3, 7, 8] but, as we mentioned in Section 3.2, we disagree with this conclusion.

Results on the 10sec-10sec (tel/tel), 1conv-10sec (tel/tel), and 8conv-10sec (tel/tel) conditions of the 2006 SRE (obtained with 100 channel factors) are presented in Tables 9, 10, and 11. They show that while the 10sec-10sec condition remains beyond our reach, fairly respectable performance can be achieved if adequate amounts of enrollment data are available. For comparison, the best

Table 9. *10sec-10sec NIST 2006 SRE.*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	13.4%	0.058	16.2%	0.069
All trials	16.8%	0.065	18.0%	0.075

results that have been published on the 10sec-10sec task are an EER of 20.8% and a DCF of 0.081 (all trials, male and female results pooled) in [7]. These results were obtained by fusing two GMM systems (one with eigenvoices, the other without) and an SVM system.

6. REFERENCES

- [1] (2006) The NIST year 2006 speaker recognition evaluation plan. [Online]. Available: <http://www.nist.gov/speech/tests/spk/2006/index.htm>
- [2] P. Kenny, "Joint factor analysis of speaker and session variability: Theory and algorithms, Tech. Report CRIM-06/08-13," 2005. [Online]. Available: <http://www.crim.ca/perso/patrick.kenny>
- [3] B. Fauve, N. Evans, N. Pearson, J.-F. Bonastre, and J. Mason, "Influence of task duration in text-independent speaker

Table 10. *1conv-10sec NIST 2006 SRE.*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	5.7%	0.030	8.5%	0.039
All trials	7.5%	0.040	10.4%	0.047

Table 11. *8conv-10sec NIST 2006 SRE. Speakers exposed in 2005 and recycled in 2006 excluded.*

	Male speakers		Female speakers	
	EER	DCF	EER	DCF
English trials	2.6%	0.014	5.6%	0.039
All trials	3.1%	0.019	6.5%	0.030

verification," in *Proc. Interspeech 2007*, Antwerp, Belgium, Aug. 2007.

- [4] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, "A study of inter-speaker variability in speaker verification." [Online]. Available: <http://www.crim.ca/perso/patrick.kenny>
- [5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, May 2007.
- [6] P. Kenny, N. Dehak, R. Dehak, V. Gupta, and P. Dumouchel, "The role of speaker factors in the NIST extended data task," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008.
- [7] B. Fauve, N. Evans, and J. Mason, "Improving the performance of text-independent short-duration SVM- and GMM-based speaker verification," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008.
- [8] R. Vogt, C. Lustri, and S. Sridharan, "Factor analysis modeling for speaker verification with short utterances," in *Proc. IEEE Odyssey Workshop*, Stellenbosch, South Africa, Jan. 2008.
- [9] N. Brummer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiat, D. van Leeuwen, P. Matejka, P. Schwarz and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2072 – 2084, Sept. 2007.
- [10] D. E. Sturim, W. M. Campbell, D. A. Reynolds, R. B. Dunn, and T. F. Quatieri, "Robust speaker recognition with cross-channel data: MIT-LL results on the 2006 NIST SRE auxiliary microphone task," in *Proc. ICASSP 2007*, Honolulu, HI, Apr. 2007, pp. IV–49–IV–52.
- [11] W. M. Campbell, D. E. Sturim, W. Shen, D. A. Reynolds, and J. Navratil, "The MIT-LL/IBM 2006 speaker recognition system: High-performance reduced-complexity recognition," in *Proc. ICASSP 2007*, Honolulu, HI, Apr. 2007, pp. IV–217 – IV–220.