Maximum Likelihood i-vector Space Using PCA for Speaker Verification

Zhenchun Lei [1], *Yingchun Yang*[2]

[1] School of Computer and Information Engineering, Jiangxi Normal University, Nanchang, China
[2] College of Computer Science, Zhejiang University, Hangzhou, China

Zhenchun.lei@hotmail.com, yyc@zju.edu.cn

Abstract

This paper proposes a new approach to training the i-vector space using a variant of PCA with the Baum-Welch statistics for speaker verification. In eigenvoice the rank of variability space is bounded by the number of training speakers, so a variant of the probabilistic PCA approach is introduced for estimating the parameters. But this constraint doesn't exist in i-vector model because the number of utterances is much bigger than the rank of total variability space. We adopt the EM algorithm for PCA with the statistics to train the total variability space, and the maximum likelihood criterion is used. After WCCN, the cosine similarity scoring is used for decision. These two total variability spaces will be fused at feature-level and score-level. The experiments have been run on the NIST SRE 2008 data, and the results show that the performances in two total variability spaces are comparable. The performance can be improved obviously after feature fusion and score fusion.

Index Terms: speaker verification, i-vector, principal component analysis

1. Introduction

Joint Factor Analysis (JFA) [1, 2] has achieved the state of the art for text-independent speaker recognition in recent years. The main idea in traditional JFA, introduced by Kenny [2], is to find two subspaces which represent the speaker and channel-variabilities respectively. Dehak [3] proposed a single space that models the two variabilities and named it the total variability space, which is a low-dimension space of the Gaussian Mixture Model (GMM) [4] supervector space. The vectors in the low-dimensional space are called i-vectors. The i-vectors are smaller in size and can get recognition performance similar to that obtained by JFA.

In the total variability space, the approach training the total variability matrix is processed by following a similar process to that of learning the eigenvoice matrix of JFA and is fully detailed in [1]. The main difference between these two is that in training the eigenvoice of JFA, all recordings of a given speaker are considered to belong to the same person, where in train the total variability matrix, each instance of a given speaker's set of utterances is regarded as having been produced by a different speaker.

In training eigenvoice, a variant of the Probabilistic Principal Component Analysis (PPCA) [5] approach is introduced for estimating the parameters. The probabilistic approach has an advantage over other approaches because it enables us to estimate as many eigenvoices as there are speakers in the training set. But in i-vector model, the number of training recordings is much larger than the dimension of eigenspace. So the constraint existing in training eigenvoice doesn't exist in training the i-vector space.

In this paper, we train the total variability space by combining the Principal Component Analysis (PCA) and the Baum-Welch statistics. Our approach and Kenny's approach are similar, and an EM algorithm for PCA [6] is used in processing thousands of datapoints in hundreds of thousands of dimensions. And we fuse these two approaches at feature level and score level for pursuing the better performance.

This paper is organized in the following way: In section 2 we explain our EM algorithm for training the total variability space and two fusion approaches. Section 3 we present the results of our experiments. Finally, section 4 is devoted to the main conclusions and our future work.

2. Methods

Given speaker- and channel-dependent GMM supervector M can be modeled as follows:

$$M = M_0 + Tw \tag{1}$$

where M_0 is a speaker- and channel-independent supervector, T is a low rank matrix, which represents a basis of the reduced total variability space and w is a normal distributed vector which are referred to as i-vector. The M is assumed to be normally distributed with mean vector M_0 and covariance matrix $T \cdot T^T$. This model can be viewed like a principal component analysis of the larger supervector space that allows projecting the speech utterances in the total variability space.

A variant of the PPCA approach is used in training the total variability matrix T. The feature vector associated with a given recording is the MAP estimation of w, and the matrix T is estimated using the EM algorithm described in Kenny's paper [1].

In our method, the likelihood function as the estimation criterion is:

$$\prod_s \max\ P(O(s) \mid M_0 + Tw(s), \Sigma) \tag{1}$$

where s ranges over the recordings in the training set, O(s) is the recording data and Σ is the covariance matrix of GMM. Although we can get all supervectors in training dataset, using PCA directly to estimate the total variability matrix T is not feasible. The dimension of supervector is very big and the number of recordings is big also, and computing the sample covariance is very costly.

We adopt the idea of EM algorithm for PCA [6], and estimate the model parameters with the Baum-Welch statistics. The optimization proceeds by iterating the following two steps:

1) For each training recording s, use the estimate of T to find the i-vector which maximizes the GMM likelihood.
2) Estimate a new total variability space T given the old space and the new i-vectors over all recordings in the training set.

Each of the optimization steps is described in detail in the next two subsections.

2.1. Estimating the i-vector

The ML estimation of the i-vector for a given recording is similar to the MLED (Maximum Likelihood eigen-decomposition) in [7, 12] and estimating the cluster weights in [8]. According to the E-step computing in [6], we can estimate the i-vector for recording s with the Baum-Welch statistics.

If m is a Gaussian distribution in GMM for a given recording, first we normalize each observation for m by subtracting the corresponding component of mean vector M_0.

To maximize the likelihood of observation, we maximize an auxiliary function $Q(\lambda,\hat{\lambda})$, where λ is current model and $\hat{\lambda}$ is estimated model. Ignoring the standard constants and terms independent of λ, we have:

$$Q(\lambda,\hat{\lambda}) = -\frac{1}{2}\sum_{s}\sum_{m}\sum_{t}\gamma_m(t)(o_t - T_m w_s)^T \Sigma_m^{-1} (o_t - T_m w_s) \tag{2}$$

where $\gamma_m(t)$ is the occupation probability, and T_m is the submatrix of T for mixture component m.

Consider the basis vector e(j) in T with j = 1 … K:

$$T = \left[e(1), e(2), ..., e(j), ... e(K)\right] \tag{3}$$

$$e(j) = \begin{bmatrix} e_1(j) \\ e_2(j) \\ ... \\ e_M(j) \end{bmatrix} \tag{4}$$

where $e_m(j)$ represents the subvector of basis vector j corresponding to the mean vector of mixture Gaussian m.

By differentiating with respect to the i-vector of a particular recording and equating to zero, the i-vector for recording s can be got. For each recording s, we set $(\partial Q / \partial w_s(j)) = 0, j = 1...K$. Assuming the basis are independent, $(\partial w_s(i) / \partial w_s(j)) = 0, i \neq j$. One obtain for j = 1…K

$$\sum_{m}\sum_{t}\gamma_m(t)(e_m(j))^T \Sigma_m^{-1} o_t = \sum_{m}\sum_{t}\sum_{k=1}^{K}\gamma_m(t) w_s(k)(e_m(k))^T \Sigma_m^{-1} e_m(j) \tag{5}$$

There are K equations to solve for the K unknown weights ($w_s(j)$ values).

2.2. Estimating the total variability space

The ML estimation of the i-vector space is similar to the estimating model-based clusters in [8]. Two statistics are required to estimate the space basis:

$$G^{(m)} = \sum_{s}\sum_{t}\gamma_m(t) w(s) w(s)^T \tag{6}$$

$$K^{(m)} = \sum_{s}\sum_{t}\gamma_m(t) w(s) o(t)^T \tag{7}$$

To maximize the likelihood of observation, we maximize an auxiliary function $Q(\lambda,\hat{\lambda})$. After differentiating Q with respect to T(m) and equating to zero, the total variability space may be estimated using:

$$T^{(m)T} = G^{(m)-1} K^{(m)} \tag{8}$$

2.3. Comparing to Kenny's method

The PPCA is applied by Kenny, and the PCA by our method. Their formulas are very similar. Using PPCA, the i-vector for utterance s can be obtained by the following:

$$w = (I + T^T \Sigma^{-1} N(s) T)^{-1} \cdot T^T \Sigma^{-1} \widetilde{F}(s) \tag{9}$$

where $N(s)$ is the diagonal matrix of dimension MF × MF whose diagonal block are $N_m(s)I$, (m=1,…,M) and $\widetilde{F}(s)$ is a supervector of dimension MF × 1 obtained by concatenating all the centralized first order Baum-Welch statistic. If we remove the I in (9), the formula is same to ours.

In Kenny's method, the total variability space can be obtained by the following:

$$\sum_{s} N(s) \cdot T \cdot E[w(s) w(s)^T] = \sum_{s} \widetilde{F}(s) \cdot E[w(s)^T] \tag{10}$$

If we make

$$E[w(s) w(s)^T] = E[w(s)] E[w(s)^T] = w(s) w(s)^T \tag{11}$$

then the formula is same to ours also.

2.4. WCCN and cosine similarity scoring

WCCN is introduced by Andrew Hatch [9] in the context of SVM classifiers. The idea is to scale the total variability space by a factor that is inversely proportional to an estimate of the within-class covariance matrix. The within-class covariance matrix is estimated using the total factor vectors form a set of development speakers as

$$W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(w_s^{(i)} - \overline{w}_s)(w_s^{(i)} - \overline{w}_s)^T \tag{12}$$

where $\overline{w}_s$ is the mean of the i-vectors for each speaker s with n_s corresponding to the number of utterances for that speaker, and S is the total number of speakers.

The simple cosine similarity metric [10, 11] has been applied successfully in the total variability space to compare two supervectors for making a speaker detection decision. Given two i-vectors via the projection of two supervectors in the total variability space and the WCCN compensation for inter-session variabilities, a target w'_{target} from a know speaker and a test w'_{test} from an unknown speaker, the cosine similarity score is given as:

$$score(w'_{target}, w'_{test}) = \frac{(w'_{target})^T w'_{test}}{\left\| w'_{target} \right\| \left\| w'_{test} \right\|} \underset{<}{\geq} \theta \tag{13}$$

where θ is the decision threshold.

2.5. Feature fusion and score fusion

There are two fusion methods in our experiments: feature fusion and score fusion. In feature fusion, the i-vectors in two total variability spaces are concatenated to a new vector for a given recording, and then the cosine classifier is used. For a given recording, $w^{(PCA)}$ and $w^{(PPCA)}$ are two i-vectors in two total variability spaces, the new i-vector is given as:

$$w^{fusion} = \begin{bmatrix} w^{(PCA)} \\ w^{(PPCA)} \end{bmatrix} \tag{14}$$

In score fusion, two classifiers are separately used with the i-vectors in two total variability spaces, and then a simple linear function is used to get the result. The new score is:

$$\begin{aligned} score(w'_{target}, w'_{test}) &= \alpha \cdot score(w^{PCA}_{target}, w^{PCA}_{test}) \\ &+ (1-\alpha) \cdot score(w^{PPCA}_{target}, w^{PPCA}_{test}) \end{aligned} \tag{15}$$

where α is between 0 and 1.

3. Experiments

3.1. Experiment set-up

The features were derived from the waveforms using 19 mel-frequency cepstral coefficients on a 20 millisecond frame every 10 milliseconds. Delta and delta-delta coefficients were computed making up a thirty nine dimensional feature vector. And the band limiting was performed by retaining only the filter bank outputs form the frequency range 300-3400 Hz. Mean removal, preemphasis and a hamming window were applied, and energy-based end pointing eliminated nonspeech frames.

Our experiments were performed on the 2008 NIST SRE dataset. NIST SRE2004 1side training corpus was used to train two gender-dependent UBMs with 512 Gaussian components. The rank of the total variability matrix T was chosen to be 400. NIST SRE2004, SRE 2005, and SRE 2006 telephone datasets were used for estimating the total variability space and the projection matrix in WCCN for inter-session compensation. There were 11656 female utterances and 8615 male utterances in the training datasets. Finally the cosine similarity scoring was used for the decision.

For measuring the performance, we used equal error rate (EER) and the minimum decision cost function (DCF). The score normalization [13] did not used in our experiments, although it is effective for improving performance.

3.2. Results on telephone condition

The first experiment was run on the 1conv-1conv 2008 SRE core telephone condition. As show in Table 1, performances of the two i-vector spaces are similar. In our experiment, the total variability space generated by PPCA is a little better than that by PCA for female, which is opposite for male. But we cannot say one is better than another.

Table 1: *Comparison results between the total variability spaces generated by PCA and PPCA. The results are on the 1conv-1conv 2008 SRE core telephone condition*

model	gender	EER(%)	DCF
i-vector(PCA)	female	10.57	0.051
i-vector(PPCA)	female	10.48	0.051
i-vector(PCA)	male	7.10	0.034
i-vector(PPCA)	male	7.19	0.033

3.3. Results on cross-channel condition

The second experiment was run on the 1conv-1conv 2008 SRE core cross-channel condition. Training the total variability space and the projection matrix in WCCN were still on the NIST 2004, 2005, 2006 telephone datasets.

Table 2 shows the results of experiments on female part of the 2008 SER data. The performance of the total variability space generated by PCA is worse than that by PPCA on cross-channel condition. Table 3 shows the result of experiments on male data, we can get the opposite result.

Table 2: *Comparison results between the total variability spaces generated by PCA and PPCA. The results are on the female portion of the 1conv-1conv 2008 SRE core cross-channel condition*

model	training	testing	EER(%)	DCF
PCA	phonecall	phonecall	10.57	0.051
	phonecall	interview	12.11	0.057
	interview	phonecall	13.17	0.055
	interview	interview	16.40	0.058
	all	all	14.37	0.061
PPCA	phonecall	phonecall	10.48	0.051
	phonecall	interview	11.72	0.054
	interview	phonecall	12.32	0.055
	interview	interview	15.95	0.059
	all	all	14.16	0.062

Table 3: *Comparison results between the total variability spaces generated by PCA and PPCA. The results are on the male portion of the 1conv-1conv 2008 SRE cross-channel condition*

model	training	testing	EER(%)	DCF
i-vector	phonecall	phonecall	7.10	0.034
PCA	phonecall	interview	8.48	0.040
	interview	phonecall	8.56	0.035
	interview	interview	12.08	0.050
	all	all	10.21	0.047
i-vector	phonecall	phonecall	7.19	0.033
PPCA	phonecall	interview	10.05	0.041
	interview	phonecall	9.16	0.039
	interview	interview	12.12	0.051
	all	all	10.29	0.048

3.4. Feature fusion

The third experiment was about the feature fusion, which had been run on the 1conv-1conv 2008 SRE core cross-channel condition. The rank of the total variability matrix T was 400, so the size of new i-vector was 800. We can see from table 4 that the performance has been improved a little overall compared to any single total variability space in table 2 or table 3.

Table 4: *Results on the i-vector model fusing the total variability spaces generated by PCA and PPCA at feature level. The results are on the 1conv-1conv 2008 SRE cross-channel condition*

gender	training	testing	EER(%)	DCF
female	phonecall	phonecall	10.41	0.050
	phonecall	interview	11.23	0.054
	interview	phonecall	12.52	0.054
	interview	interview	16.01	0.057
	all	all	14.05	0.061
male	phonecall	phonecall	7.01	0.033
	phonecall	interview	8.98	0.039
	interview	phonecall	8.39	0.036
	interview	interview	11.89	0.050
	all	all	10.10	0.047

3.5. Score fusion

Finally, we fused two total variability spaces at score level, and table 5 shows the results. We can see that the performance is improved obviously compared any single total variability space. Compared the feature fusion, the score fusion also get the better performance.

Table 5: *Results on the i-vector model fusing the total variability spaces generated by PCA and PPCA at score level. The results are on the 1conv-1conv 2008 SRE cross-channel condition*

gender	training	testing	EER(%)	DCF
female	phonecall	phonecall	10.31	0.050
	phonecall	interview	10.98	0.053
	interview	phonecall	12.21	0.054
	interview	interview	16.00	0.057
	all	all	14.01	0.061
male	phonecall	phonecall	6.95	0.032
	phonecall	interview	8.38	0.038
	interview	phonecall	8.27	0.035
	interview	interview	11.80	0.049
	all	all	10.03	0.047

4. Conclusions

We propose a method to train the i-vector model parameters. The main idea is adopting the EM algorithm for PCA and working with the Baum-Welch statistics of utterances. Our method is compared with the Kenny's method, and we find that they are very similar. Finally, the WCCN is used for inter-session compensation and the cosine similarity scoring is used for decision. We can fuse two methods at feature level and score level. Our experiments were run on NIST SRE dataset. The results show that the performance of our method is comparable with Kenny's method. After feature fusion, the performance can be improved a little overall. And the score fusion can improve the performance obviously.

Although the model shows the advantages, there are some new problems which still need to solve. First, the covariance matrix Σ is not updated in our method, and this will be tested in near future. Second, PCA is a widely used unsupervised dimensionality reduction technique in data analysis. And some discriminative training methods show better performance in many applications, such as LDA [14]. We will try to combine the Baum-Welch statistics and LDA to get a more discriminative total variability space. JFA has been reinterpreted as signal coding using overcomplete dictionaries in Danile's paper [15], and the i-vecor model can also be reinterpreted in the same way. Some theories and algorithms in overcomplete dictionaries have been developed rapidly in recent years. We will also do more research on the basis of the total variability space.

5. Acknowledgements

This work is supported by National Natural Science Foundation of P.R.China (60970080), Educational Commission of Jiangxi Province of P.R.China (GJJ09139) and by Youth Foundation of Jiangxi Normal University.

6. References

[1] Patrick Kenny, Gilles, B. and Pierre, D., "Eigenvoice Modeling With Sparse Training Data", IEEE Trans. Speech and Audio Proc., 13(3):345-354, 2005.

[2] Patrick Kenny, et al. "Joint Factor Analysis Versus Eigenchannels in Speaker Recognition", IEEE Trans. Speech and Language Proc., 15(4):1435-1447, 2007.

[3] N.Dehak, et al. "Support vector machines versus fast scorring in the low-dimensional total variability space for speaker verification", in Interspeech, Brighton, UK, Sept 2009.

[4] D.A.Reynolds, T.Quatieri, and R.Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol.10, no.1-3, 2000.

[5] M. Tipping and C. Bishop, "Mixtures of probabilistic principal component analyzers", Neural Computation, vol.11, pp.435-474, 1999

[6] Sam Roweis, "EM Algorithms for PCA and SPCA", in Advances in neural information processing systems, 1998.

[7] Roland Kuhn, et al. "Rapid Speaker Adaptation in Eigenvoice Space", IEEE Trans. Speech and Audio Proc., 8(6):695-707, 2000.

[8] Mark J.F. Gales, "Cluster Adaptive Training of Hidden Markov Models", IEEE Trans. Speech and Audio Proc., 8(4):417-427, 2000.

[9] A.Hatch, S.Kajarekar, and A.Stolcke, "Withinclass covariance normalization for svm-based speaker recognition", in Interspeech – 9th International Conference on Spoken Language Processing-ICSLP, vol.3, pp.1471-1474, 2006.

[10] Stephen Shum, et al, "Unsupervised Speaker Adaptation based on the Cosine Similarity for Text-Independent Speaker Verification", Odyssey, The Speaker and Language Recognition Workshop, 28 June-1 July 2010.

[11] Najim Dehak, et al. "Cosine Similarity Scoring without Score Normalization Techniques", Odyssey, The Speaker and Language Recognition Workshop, 28 June-1 July 2010.

[12] Patrick Nguyen, Christian Wellekens, and Jean-Claude Junqua, "Maximum Likelihood Eigenspace and MLLR for Speech Recogniton in Noisy Environments", in Sixth European Conference on Speech Communication and Technology, Eurospeech, Budapest, Hungary, September 5-9, 1999

[13] R.Auckenthaler, M.Carey and H.L.Thomas, "Score Normalization for Text-Independent Speaker Verification Systems" Digital Signal Processing 10, 42-54, 2000.

[14] A.M.Martinez and A.C.Kak, "PCA versus LDA", IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(2), pp.228-233, 2004

[15] Daniel G.R. and Carol Y.E.W, "Joint Factor Analysis for Speaker Recognition reinterpreted as Signal Coding using Overcomplete dictionaries", Odyssey, The Speaker and Language Recognition Workshop, 28 June-1 July 2010.