Bayesian Distance Metric Learning on i-vector for Speaker Verification

Xiao Fang, Najim Dehak, James Glass

MIT Computer Science and Artificial Intelligence Laboratory,
Cambridge, Massachusetts 02139, USA
{fxiao, najim, glass}@csail.mit.edu

Abstract

This paper presents a new speaker verification system based on i-vector modeling as a feature extractor. In this modeling, we explore the distance constraints between i-vector pairs from the same speaker and different speakers. With an approximation of the distance metric as a weighted covariance matrix of the top eigenvectors from the data covariance matrix, variational inference is used to estimate a posterior distribution for the distance metric. Given speaker labels, we select different-speaker data pairs with the highest cosine scores to form a different-speaker constraint set. This set captures the most discriminative between-speaker variability in the training data. This Bayesian distance metric learning approach achieves better performance than state-of-the-art method. Furthermore, this approach is insensitive to score normalization, as compared to cosine scoring. Without the requirement of the number of labeled examples, this approach performs very well in the context of limited training data.

Index Terms: i-vector, score normalization, distance metric learning, channel compensation, limited training utterances

1. Introduction

Recently, many speaker verification systems based on the i-vector [1][2] have achieved significant improvements in performance. The i-vector is a compact, low-dimensional representation of any speech segment. We can generally treat an i-vector as input to common classifiers such as Support Vector Machines (SVM), or cosine distance classifiers. [1] shows cosine distance scoring to achieve state-of-the-art performance. In the i-vector training and score verification process, speaker labels are not used explicitly, suggesting that the full use of speaker labels may lead to better performance. Since the i-vector contains both speaker- and channel-variability, we'd like to carry out channel compensation before verification. Linear Discriminant Analysis (LDA) is widely used to remove the session variability.

The basic speaker recognition task seeks to determine whether the test utterance and the target utterance are from the same speaker. In other words, our goal is to determine the proximity between the test utterance and the target utterance are close enough or not. Thus we can view the speaker verification system as a distance metric learning problem: given speaker labels of training utterances, we aim to find a good distance metric that brings "similar" data points (belonging to the same speaker) close together while separating "dissimilar" data points (belonging to different speakers) [3]. There are a number of algorithms developed for supervised distance metric learning to optimize different objective functions. Most of the algorithms return a point estimation of the distance metric [4], which is sensitive to the selection of training samples. In [5], Yang and Jin introduce a Bayesian framework for distance metric learning, which aims to estimate a posterior distribution for the distance metric. The algorithm has achieved high classification accuracy in image classification. In addition, this approach is insensitive to the number of labeled examples for each class, as compared to most algorithms requiring a large number of labeled examples [5]. This advantage is particularly important for realistic speaker verification systems, in which it can be difficult to collect a sufficiently large number of samples from every speaker, even though it may be possible to collect samples from a large number of different speakers. In this paper, we present a speaker verification system based on the Bayesian distance metric learning framework.

The rest of this paper is organized as follows. Section 2 provides a background review of speaker verification based on the cosine similarity of i-vectors and channel compensation via LDA. In Section 3, we propose the speaker verification algorithm based on Bayesian distance metric learning. Some results on 2008 NIST SRE task are analyzed and explained in Section 4 to show the superior performance of the algorithm over cosine scoring, and Section 5 concludes with a discussion of future work.

2. Related work

2.1. i-vector representation

The i-vector representation, also known as total variability modeling, aims to model the utterance variability in a low-dimensional space. Total variability originates from joint factor analysis [8] but doesn't distinguish between speaker variability and channel variability [9]. Given an utterance, the speaker- and channel-dependent Gaussian Mixture Model (GMM) supervector M can be represented as

$$M = m + Tw$$

where m is the speaker- and channel-independent supervector (which can be taken to be the Universal Background Model (UBM) supervector), T is a rectangular matrix of low rank and w is a random vector having standard normal distribution $N(0, I)$. T defines the new total variability space, and the remaining variabilities not captured by T are accounted for in a diagonal covariance matrix Σ. In this modeling, all the high-dimensional supervectors lie around m in a relatively lower-dimensional subspace. w is the speaker- and channel-dependent factor in the total variability space. The mean of the posterior distribution of w corresponds to the i-vector, which can be seen as a new speaker verification feature with a relatively low dimension.

The training of parameters is based on the EM algorithm [1]. Note that all utterances from one speaker are regarded as having been produced by different speakers. Since we do not

need any speaker labels to get i-vectors, it is essentially an unsupervised training method.

For the task of speaker verification, we use a simple cosine distance classifier on the i-vectors of the target user utterance and the test utterance:

$$score(w_{target}, w_{test}) = \frac{(w_{target}^t) w_{test}}{\|w_{target}\| \cdot \|w_{test}\|} \gtrless \theta$$

where θ is the decision threshold.

2.2. Score normalization

Although cosine distance scoring is fast and robust, it suffers from the additional computation necessary for score normalization. A better generative model would simulate speech data perfectly and produce a score without normalization or calibration [10].

Z-norm and t-norm are typically used to get a calibrated score [11]. In [12], Dehak proposed a new cosine similarity scoring. It can be treated as the combination of z-norm and t-norm score normalization.

Assuming $w^{'}$ is a length-normalized i-vector, the score for a length-normalized target i-vector $w_{target}^{'}$ and a length-normalized test i-vector $w_{test}^{'}$ can be computed as below:

$$score(w_{target}^{'}, w_{test}^{'}) = \frac{(w_{target}^{'} - \overline{w^{'}})^t (w_{test}^{'} - \overline{w^{'}})}{\| C \cdot w_{target}^{'} \| \| C \cdot w_{test}^{'} \|}$$

where $\overline{w^{'}}$ is the mean of imposter i-vectors, and C is a diagonal matrix which contains the square root of the diagonal imposter's covariance matrix, $\Sigma = E[(w^{'} - \overline{w^{'}})(w^{'} - \overline{w^{'}})^t]$.

2.3. Intersession compensation

In the total variability representation, there is no explicit compensation for inter-session variability. However, the low-dimensional representation enables us to carry out compensation techniques in the new space, with the benefit of less expensive computation. We use Linear Discriminant Analysis (LDA) for session compensation. Viewing one speaker as one class, LDA attempts to define new axes that minimize the intra-class variance caused by session/channel effects, and maximize the variance between classes.

The LDA optimization problem can be defined to find the direction q that maximizes the fisher criteria

$$J(q) = \frac{q^t S_b q}{q^t S_w q}$$

where S_b and S_w are between-class and within-class covariance matrices:

$$S_b = \sum_{s=1}^{S} (\overline{w^s} - \overline{w})(\overline{w^s} - \overline{w})^t$$

$$S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (w_i^s - \overline{w^s})(w_i^s - \overline{w^s})^t$$

and $\overline{w^s} = (1/n_s) \sum_{i=1}^{n_s} w_i^s$ is the mean of i-vectors for each speaker, n_s is the number of utterances for each speaker s, $\overline{w}$ is the speaker population mean vector, S is the number of speakers.

The maximization is achieved to define a projection matrix Q composed by the top eigenvectors of the general matrix $S_w^{-1} S_b$.

3. Distance metric learning

With i-vectors as low-dimensional representations of speech utterances, the cosine distance classifier is to measure the cosine distance of the target user utterance and the test utterance. How to define the distance between vectors, which aims to find a good distance metric in feature space, is a crucial problem in classification. There has been considerable research on distance metric learning over the last few years [6]. We explore two supervised distance metric learning methods in this paper. From now on, we use an i-vector w_i to represent an utterance.

3.1. Neighborhood component analysis

Neighborhood component analysis (NCA) [4] learns a distance metric to minimize the average leave-one-out k nearest neighbor classification error under a stochastic selection rule. With a transformation matrix B, each i-vector w_i selects another i-vector w_j as its neighbor with some probability p_{ij}, which is defined over Euclidean distances in the transformed space:

$$p_{ij} = \frac{\exp(- \| Bw_i - Bw_j \|^2)}{\sum_{k \neq i} \exp(- \| Bw_i - Bw_k \|^2)}, \; p_{ii} = 0$$

The probability for the i-vector w_i selecting neighbors from the same speaker is $p_i = \sum_{j \in C_i} p_{ij}$, where C_i is the set of i-vectors from the same speaker with i. The projection matrix B is to maximize the expected number of i-vectors selecting neighbors from the same speaker:

$$B = \text{argmax}_B f(B) = \sum_i \sum_{j \in C_i} p_{ij} = \sum_i p_i$$

A conjugate gradient method is used to obtain the optimal B.

3.2. Bayesian distance metric learning framework

NCA provides a point estimation of the distance metric and can be unreliable when the number of training examples is small. [5] presents a Bayesian framework to estimate a posterior distribution for the distance metric.

Given the speaker-labels of each utterance, we can form two sets of same-speaker and different-speaker constraints S and D. The probability of two utterances w_i and w_j belonging to the same speaker or different speakers is defined under a given distance matrix A:

$$Pr(y_{ij} | w_i, w_j, A, \mu) = \frac{1}{1 + \exp\left(y_{ij}(||w_i - w_j||_A^2 - \mu)\right)}$$

where $$y_{ij} = \begin{cases} +1 & (w_i, w_j) \in S \\ -1 & (w_i, w_j) \in D \end{cases}$$

The parameter μ is the threshold to differentiate same-speaker utterances and different-speaker utterances. Two utterances are more likely to be identified from the same speaker only when their distance with respect to the distance matrix A is less than μ.

To simplify the computation, [5] models the distance metric A as a parametric form of the top eigenvectors of observed data points. Let $X = (w_1, w_2, \ldots, w_n)$ denote all the available utterances, and $\mathbf{v}_l, l = 1, \ldots, K$ be the top K eigenvectors of XX^T. Assume $A = \sum_{l=1}^{K} \gamma_l \mathbf{v}_l \mathbf{v}_l^T$, where $\gamma_l \geq 0, l = 1, 2, \ldots, K$, then the likelihood can be rewritten as:

$$\begin{aligned} Pr(y_{ij} | w_i, w_j, A, \mu) &= \frac{1}{1 + \exp\left(y_{ij}(\sum_{l=1}^{K} \gamma_l w_{i,j}^l - \mu)\right)} \\ &= \sigma(-y_{ij} \gamma^t w_{i,j}) \end{aligned}$$

where

$$w_{i,j}^{l} = [(w_i - w_j)^t \mathbf{v}_l]^2$$

$$w_{i,j} = (-1, w_{i,j}^{1}, \ldots, w_{i,j}^{K})$$

$$\gamma = (\mu, \gamma_1, \ldots, \gamma_K)$$

$$\sigma(z) = 1/(1 + \exp(-z))$$

Applying Gaussian prior distributions on the parameters $\gamma = (\mu, \gamma_1, \ldots, \gamma_K)$, the evidence function is computed as:

$$\begin{aligned} \Pr(S, D) &= \int \Pr(S, D|\gamma)\Pr(\gamma)d\gamma \\ &= \int \prod_{(i,j)\in S} \sigma(-\gamma^t w_{i,j}) \prod_{(i,j)\in D} \sigma(\gamma^t w_{i,j}) \\ &N(\gamma; \gamma_0 \mathbf{1}_{K+1}, \delta^{-1}\mathbf{I}_{K+1})d\gamma \end{aligned}$$

The transformation of the likelihood to a logistic function makes it possible to get a lower bound of the evidence, thus a variational method [16] is employed to estimate the posterior distribution for γ. The details of the algorithm can be found in [5].

After getting the posterior distribution $\phi(\gamma) \sim N(\gamma; \mu_\gamma, \Sigma_\gamma)$, we can compute the conditional probability $Pr(\pm|w_i, w_j)$ as follows:

$$\begin{aligned} Pr(\pm|w_i, w_j) &= \int \frac{N(\gamma; \mu_\gamma, \Sigma_\gamma)}{1 + \exp(\pm\gamma^T w_{i,j})} d\gamma \\ &\propto \int \exp(-l_{i,j}^{\pm}(\gamma))d\gamma \end{aligned}$$

where $l_{i,j}^{\pm}(\gamma) = \log(1 + \exp(\pm\gamma^T w_{i,j})) + \frac{1}{2}(\gamma - \mu_\gamma)^T \Sigma_\gamma^{-1}(\gamma - \mu_\gamma)$.

We first approximate the optimal solution $\gamma_{i,j}^{\pm}$ by expanding $l_{i,j}^{\pm}(\gamma)$ in the neighborhood of μ_γ, then $l_{i,j}^{\pm}(\gamma)$ by its Taylor expansion around the optimal solution $\gamma_{i,j}^{\pm}$, and compute the integral using this approximation. Thus the final estimation of the probability takes into account the full distribution of γ. The probability of identifying the target and test utterance from the same speaker $Pr(+|w_{target}, w_{test})$ is the output score.

We use NCA as a preprocessing technique to project the i-vectors into a space in which the nearest neighbor of each i-vector shares the same speaker label with a high probability. The Bayesian distance metric learning approach can model the distances between i-vectors better and more reliably in the new space. Our experimental results demonstrated the benefits of this approach.

4. Experiments

4.1. Experimental set-up

Experiments are performed on the female part of the short2-short3 condition of the 2008 NIST SRE dataset [17]. The training and test data are telephone conversational excerpt of approximately five minutes duration. The set for i-vector training contains 1,830 speakers and 21,382 utterances. It is also used for LDA and NCA training, and as the imposter set in the score normalization step. The evaluation set contains 1,678 target speakers and 24,128 test trials. A 600-dimension i-vector is extracted from each utterance. The Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) are used as metrics for evaluation.

Table 1: *Comparison of results between the cosine score and Bayes_dml w/o score normalization*

	EER	minDCF
LDA200+Cosine Score	2.54%	0.0144
LDA200+Cosine Score_combined norm	**1.791%**	**0.0098**
LDA200+Bayes_dml	2.163%	0.0108
LDA200+Bayes_dml+znorm	2.163%	0.0108
LDA200+Bayes_dml+tnorm	2.163%	0.0108

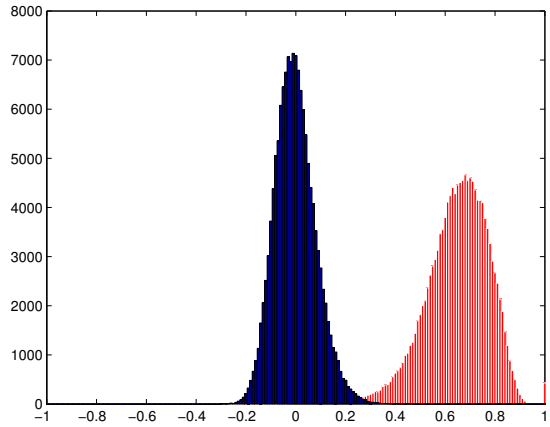

Figure 1: *Comparison of score histograms from Cosine Score blue: non-target scores, red: target scores*

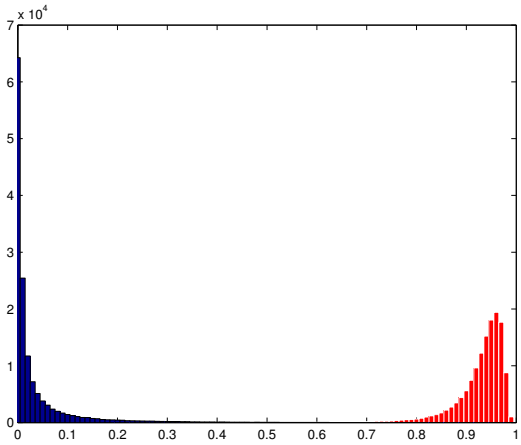

Figure 2: *Comparison of score histograms from Bayes_dml blue: non-target scores, red: target scores*

4.2. Results

The Bayesian distance metric learning algorithm is referred to as Bayes_dml, and cosine score after the score normalization described in Section 2.2 as Cosine Score_combined norm. The comparison of the results of cosine score and Bayes_dml is shown in Table 1. Given the speaker labels for training utterances, we construct the similar- and different-speaker set as follows: all possible i-vector pairs from the same speaker form the constraint S; cosine scoring is applied to all possible

Table 2: *Comparison of results between Cosine Score_combined norm and Bayes_dml with different preprocessing techniques*

Cosine Score_combined norm	EER	minDCF
LDA200	1.791%	0.0098
LDA200+NCA200	2.539%	0.0139
LDA200+NCA200+LDA200	1.780%	0.0097
LDA200+NCA200+LDA100	2.018%	0.0099
Bayes_dml	EER	minDCF
LDA200	2.163%	0.0108
LDA200+NCA200	3.032%	0.0178
LDA200+NCA200+LDA200	1.815%	0.0101
LDA200+NCA200+LDA100	1.777%	0.0096

i-vector pairs from different speakers, and those with the highest scores are selected to form the constraint D since these pairs are the most discriminative ones for a distance metric to distinguish.

We can see that Cosine Score_combined norm with LDA200 achieves the best result. However, Bayes_dml performs better than Cosine Score, i.e. cosine score without score normalization. Compared with the state-of-the-art performance from Cosine Score_combined norm, the gap of Bayes_dml is already quite small. Furthermore, there is almost no benefit to be derived from score normalization in Bayes_dml.

We can find the differences from the histograms of target scores and non-target scores from Cosine Score and Bayes_dml shown in Figure 1 and Figure 2, respectively. The target scores represent the scores of test utterances from the target speaker, and the non-target scores represent the score of test utterances not from the target speaker. The score distributions from Bayes_dml are much more concentrated than those from cosine score, and the target scores and non-target scores are better separated as well. This comparison can explain why Bayes_dml outperforms Cosine Score in table 1. As a result, there is no need to do score normalization in Bayes_dml, which makes it a more ideal model.

With a basic understanding of the difference between Cosine Score_combined norm and Bayes_dml, we compare their performances with different combinations of preprocessing techniques. The preprocessing techniques include LDA and NCA, which are applied before the scoring models. The results are shown in Table 2.

The best performance for Cosine Score_combined norm is achieved with LDA200+NCA200+LDA200, and the best performance for Bayes_dml is achieved with LDA200+NCA200+LDA100. Bayes_dml outperforms Cosine Score_combined norm, and is also the best reported result on the short2-short3 condition of the 2008 NIST SRE female data. If we only do NCA projection, the results get worse. This is because the NCA matrix is obtained under the best nearest neighbor classification criterion, without taking into consideration the clustering of i-vectors from the same speaker and the separation of i-vectors from different speakers. While LDA can achieve this goal by optimizing the Fisher criteria, generally NCA followed by LDA can project the data into a space in which i-vectors from the same speaker are close, and i-vectors from different speakers are well separated.

Table 3: *Comparison of results between Cosine Score_combined norm and Bayes_dml on limited training data (the number of training utterances for each speaker is 3)*

Cosine Score_combined norm	EER	minDCF
LDA200	4.181%	0.0210
LDA200+NCA200+LDA200	3.930%	0.0210
LDA200+NCA200+LDA100	4.664%	0.0260
Bayes_dml	EER	minDCF
LDA200	4.514%	0.0237
LDA200+NCA200+LDA200	4.190%	0.0261
LDA200+NCA200+LDA100	3.751%	0.0208

4.3. Results on limited training data

In this part, we show the advantage of Bayes_dml when the training utterances of each speaker is very limited. We select a small number (3) of utterances from each training speaker to build a made-up training set. The test set is the same as before. The best preprocessing techniques from Section 4.2 and LDA are evaluated, with the results shown in Table 3.

We can see that Bayes_dml generally achieves a better EER, which means that a lower false alarm and a lower miss probability can be achieved at the same time in Bayes_dml. The best performance of Bayes_dml is better than that of Cosine Score_combined norm. Even with only 3 utterances from each speaker, we can still get rich information from same-speaker and different-speaker i-vector pairs, whereas data sparsity can cause LDA to not fully capture the speaker variability.

5. Conclusion

In this paper, we propose a novel speaker verification framework based on the Bayesian distance metric learning algorithm. We use total variability modeling as a feature extractor. Each utterance is represented by a low-dimensional i-vector. Previous approaches try to find a class model for each speaker [9] [10], which performs poorly when the training utterances for each speaker is limited. Inspired by the successful application of distance metric learning in other areas of statistical classification [5], we explore the distance constraints between i-vector data pairs and solve this problem in a new way. The same- and different-speaker constraint sets are constructed from training data and the distance metric is learned via a Bayesian approach. NCA is used as a preprocessing technique to improve the performance combined with LDA.

Since cosine distance measure has a very competitive performance and distance metric learning uses Euclidean distance in the space projected by $A^{\frac{1}{2}}$, future work should explore incorporating cosine distance measurement into the distance metic learning framework. The performance of speaker verification in arbitrary durations has become a critical issue in the NIST evaluation protocol since 2012. It will be worth to see how the framework in this paper works for shorter duration utterances.

6. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel and P. Ouellet, "Front-End Factor Analysis for Speaker Verification", in *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788-798, 2011.

[2] P. Bousquet, D. Matrouf and J-F. Bonastre, "Intersession Compensation and Scoring Methods in the i-vectors Space for Speaker Recognition", in *Interspeech*, 2011.

[3] E. Xing, A. Ng, M. Jordan and S. Russell, "Distance Metric Learning with Application to Clustering with Side-Information", *Neural Information Processing Systems*, pp. 505-512, 2002.

[4] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov, "Neighborhood Component Analysis", *Neural Information Processing Systems*, 2004.

[5] L. Yang, R. Jin, and R. Sukthankar, "Bayesian Active Distance Metric Learning", in *Uncertainty in Artificial Intelligence*, 2007.

[6] L. Yang, "Distance Metric Learning: A Comprehensive Survey".

[7] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel,"A Study of Interspeaker Variability in Speaker Verification", in *IEEE Transaction on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980-988, July 2008.

[8] P. Kenny, G. Boulianne, P. Ouellet and P. Dumouchel,"Joint Factor Analysis versus Eigenchannels in Speaker Recognition", in *IEEE Transaction on Audio, Speech and Language Processing*, vol. 15, no. 4, pp. 1435-1447, 2007.

[9] N. Dehak, R. Dehak, P. Kenny, N. Brummer, P. Ouellet and P. Dumouchel, "Support Vector Machines versus Fast Scoring in the Low-Dimensional Total Variability Space for Speaker Verification", in *Proc. International Conference on Spoken Language Processing*, pp. 1559-1562, 2009.

[10] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors", in *Proc. IEEE Odyssey Workshop*, Brno, Czech Republic, June 2010.

[11] R. Auckenthaler, M. Carey, H. Lloyd-Thomas, "Score Normalization for Text-Independent Speaker Verification Systems", *Digital Signal Processing*, vol. 10, no. 1-3, pp. 42-54, 2000.

[12] N. Dehak, R. Dehak, J. Glass, D. Reynolds, and P. Kenny, "Cosine Similarity Scoring without Score Normalization Techniques", in *Proc. IEEE Odyssey Workshop*, Brno, Czech Republic, June 2010.

[13] P. Matejka, O. Glembek, F. Castaldo, M. Alam, O. Plchot, P. Kenny, L. Burget and J. Cemocky, "Full-covariance UBM and Heavy-tailed PLDA in i-vector speaker Verification", in *Proc. International Conference on Acoustics, Speech, and Signal Processing*, pp. 4828-4831, 2011.

[14] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification Using Adapted Gaussian Mixture Models", *Digital Signal Processing* , vol. 10, no. 1-3, pp. 19-41, 2000.

[15] D. Rubin and D. Thayer, "EM Algorithms for ML Factor Analysis", *Psychometrika*, vol. 47, no. 1, pp. 69-76, 1982.

[16] T. Jaakkola and M. Jordan, "Bayesian parameter estimation via variational methods", *Statistics and Computing*, 2000.

[17] "2008 NIST Speaker Recognition Evaluation Plan", *http://www.itl.nist.gov/iad/mig/tests/sre/2008/index.html*.