CSLT TECHNICAL REPORT-20150027 [Monday 19th October, 2015]

Max-Margin Metric Learning for Speaker Recognition

lantian Li[1,3], Chao Xing[1], Dong Wang[1*] and Thomas Fang Zheng[1]

*Correspondence: wang-dong99@mails.tsinghua.edu.cn
[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

Probabilistic linear discriminative analysis (PLDA) is among the most popular methods that accompany the i-vector model to deliver state-of-the-art performance for speaker recognition. A potential problem of the PLDA model, however, is that it essentially assumes strong Gaussian distributions over i-vectors as well as speaker mean vectors, and the objective function is not directly related to the goal of the task, e.g., discriminating true speakers and imposters.

We propose a max-margin metric learning approach to solve the problem. It learns a linear transform with the criterion that target trials and imposter trials are discriminated from each other by a large margin. Experiments conducted on the SRE08 core test show that this new approach achieves a performance comparable to or even better than PLDA, though the scoring is as simple as a cosine computation.

Keywords: speech recognition; deep neural network; noise injection

1 Introduction

The i-vector model represents the state of the art for modern speaker recognition [1, 2]. By this model, a speech segment is represented as a low-dimensional continuous vector (i-vector), so that speaker recognition (and other tasks) can be performed based on the vector representations.

A particular property of the i-vector model is that both the speaker and session variances are embedded in a single low-dimensional subspace. This is an obvious advantage since more speaker-related information is retained compared to other factorization models, e.g., JFA [3]; however, since the speaker-related information is buried under others, raw i-vectors are not sufficiently discriminative with respect to speakers. In order to improve the discriminative capability of i-vectors for speaker recognition, various discriminative models have been proposed, including with-in class covariance normalization (WCCN) [4], nuisance attribute projection (NAP) [5], linear discriminant analysis (LDA) [6], and its Bayesian counterpart, probabilistic linear discriminant analysis (PLDA) [7].

Among these models, PLDA plus length normalization is regarded to be the most effective and delivers state-of-the-art performance. The success of this model is largely attributed to two factors: one is the training objective function that reduces the intra-speaker variation while enlarges inter-speaker variation, and the other is the Gaussian prior it assumes over the speaker mean vectors, which improves robustness on speakers with little or no training data.

These two factors, however, are also the two main shortcomings of the PLDA model. As for the objective function, although it encourages discrimination among

speakers, the discrimination is based on Euclidian distance, which is inconsistent with the normally used cosine distance that has been demonstrated to be more effective.[1] Additionally, our task in speaker recognition is to discriminate true speakers and imposters, which is a binary decision, instead of the multi-class discrimination in PLDA training. As for the Gaussian assumption, it is often over strong and can not be held in practice, leading to a less representative model.

Some researchers have noticed these problems. For example, to go beyond the Gaussian assumption, Kenny proposed a heavy-tailed PLDA [8] which assumes a non-Gaussian prior over the speaker mean vector. Garcia-Romero et al. found that length normalization can compensate for the non-Gaussian effect and boost performance of Gaussian PLDA to the level of the heavy-tailed PLDA [9]. Burget, Cumani and colleagues proposed a pair-wised discriminative model that discriminates true speakers and imposters [10, 11]. In their approach, the model accepts a pair of i-vectors and predicts the probability that they belong to the same speaker. The input features of the model are derived from the i-vector pairs according to a form derived from the PLDA score function (further generalized to any symmetric score functions in [11]), and the model is trained on i-vector pairs that have been labelled as identical or different speakers. A particular shortcoming of this approach is that the feature expansion is highly complex. To solve this problem, a partial discriminative training approach was proposed in [12], which optimizes the discriminative model on a subspace and does not require any feature expansion. In [13], we proposed a discriminative approach based on deep neural networks (DNN), which holds the same idea as the pair-wised training, while the features are defined manually.

Although promising, the discriminative approaches mentioned above seem rather complex. We hope a model as simple as LDA and the inference as simple as a cosine computation. This paper presents a max-margin metric learning (MMML) approach, which is a simple linear projection trained with the objective of discriminating true speakers and imposters directly. Once the projection has been learned, simple cosine distance is sufficient to conduct the scoring. This approach belongs to the simplest metric learning which has been studied for decades in machine learning [14, 15], though it has not been extensively studied in speaker recognition.

The rest of this paper is organized as follows. Section 2 discusses some related work, Section 3 presents the max-margin learning method. The experiments are presented in Section 4, and Section 5 concludes the paper.

2 Related work

Some of the related works, particularly the pair-wised discriminative model, have been discussed in the previous section. This section presents some researches on metric learning for speaker recognition, which are related to our study more directly. A representative work proposed in [16] employs neighborhood component analysis (NCA) to learn a projection matrix that minimizes the average leave-one-out k-nearest neighbor classification error. Our model differs from the NCA approach in

[1]This inconsistency is more serious for the LDA model for which cosine distance is used in evaluation. For PLDA, the training and evaluation are with the same Euclidian distance, though cosine distance is potentially more suitable.

that we use max-margin as the training objective and cosine distance as the distance measure, which is more suitable for speaker recognition.

The cosine similarity large margin nearest neighborhood (CSLMNN) model proposed in [17] is more relevant to our proposal. The authors formulated the training task as a semidefinite program (SDP) [18] which moves i-vectors of the same speaker closer by maximizing the cosine distance among them, while penalizing the criterion of separating the data of different speakers by a large margin. Our approach uses a similar objective function, though employs a simpler solver based on stochastic gradient descendent (SGD), which supports mini-batch learning and accomodates large scale optimization.

3 Max-margin metric learning

This section presents the max-margin metric learning for speaker recognition. Metric learning has been studied for decades. The simplest form is to learn a linear projection M so that the distance among the projected data is more suitable for the task in hand [14]. For speaker recognition, the most popular used distance metric is the cosine distance and the goal is to discriminate true speakers and imposters, we therefore optimize M to make the projected i-vectors more discriminative for genuine and counterfeit speakers measured by cosine distance.

Formally, the cosine distance between two i-vectors $\mathbf{w}_1$ and $\mathbf{w}_2$ is given as follows:

$$d(w_1, w_2) = \frac{< \mathbf{w}_1, \mathbf{w}_2 >}{\sqrt{||\mathbf{w}_1||||\mathbf{w}_2||}}.$$

where $< \cdot, \cdot >$ denotes inner product, and $||\cdot||$ is the l-2 norm. Further define a contrastive triple (w, w^+, w^-) where the i-vectors w and w^+ are from the same speaker, and w and w^- are from different speakers. Letting S denote all the contrastive triples in a development set, we can define the max-margin objective function that encourages i-vectors of the same speaker moving close while penalizing i-vectors from different speakers, given by:

$$\mathcal{L}(M) = \sum_{(\mathbf{w}, \mathbf{w}^+, \mathbf{w}^-) \in S} \max\{0, \delta - d(M\mathbf{w}, M\mathbf{w}^+) + d(M\mathbf{w}, M\mathbf{w}^-)\}$$

where δ is a hyperparameter that determines the margin. Note that minimizing this function results in maximizing the margin between i-vectors of the same speaker and different speakers.

Note that optimizing $\mathcal{L}(M)$ directly is often infeasible, because size of S is exponentially large. We choose the SGD algorithm to solve the problem, where the training is conducted in a mini-batch style. In a mini-batch t, a number of contrastive triples are sampled from S, and these triples are used to calculate the gradient $\frac{\partial \mathcal{L}}{\partial M}$. The projection M is then updated with this gradient as follows:

$$M^t = M^{t-1} + \epsilon \frac{\partial \mathcal{L}}{\partial M}$$

where M^t is the projection matrix at mini-batch t, and ϵ is a learning rate. This learning iterates until convergence is obtained. In this study, the Theano package [19] was used to implement the SGD training.

Once the matrix M has been learned from the development data, an i-vector $\mathbf{w}$ can be projected to its image $M\mathbf{w}$ in the projection space, where true speakers and imposters are more easily to be discriminated, according to the training objective. Note that the max-margin metric learning is based on cosine distance, which means that the simple cosine distance is the theoretically correct choice when scoring trials in the projection space. This is a big advantage compared to PLDA, which requires complex matrix computation.

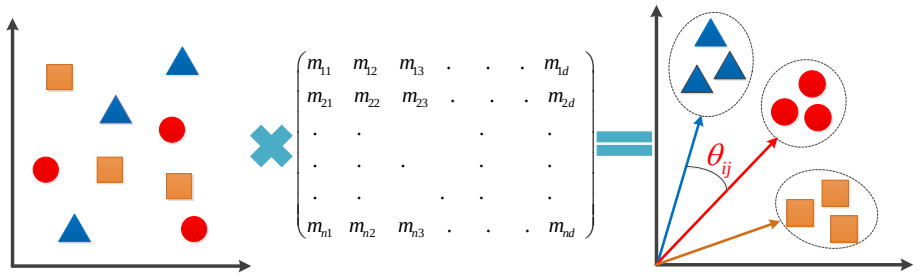

Figure 1 *Illustration of the improved discrimination with the max-margin metric learning. Each speaker is represented by a shape and a particular color. After applying the projection that is learned from data, speakers that congest together in the original i-vector space become separable.*

Fig.1 illustrates the concept of the max-margin metric learning for speaker recognition. The i-vectors from the same speaker are labeled as the same color and shape. In the input space, i-vectors of all the speakers are congested together. After applying the learned projection, i-vectors of the same speaker are moved closer, while those of different speakers are moved apart. Note that there is a margin measured by angle θ_{ij} between a speaker pair i and j.

4 Experiments

We evaluate the proposed method on the SRE08 core test. This section first presents the data used and the experimental setup, and then report the results in terms of equal error rate (EER) and DET curves.

4.1 Database

The Fisher database is used to train the i-vector system. We selected 7196 speakers to train the i-vector model, the LDA model and the PLDA model. The same data are also used to conduct the metric learning. The NIST SRE 2008 evaluation database [20] is used as the test set. We selected 1997 female utterances from the core evaluation data set (short2-short3) and based on that constructed 59343 trials, including 12159 target trials and 47184 imposter trials.

4.2 Experimental setup

The acoustic feature is 12-dimensional Mel frequency cepstral coefficients (MFCCs) together with the log energy. The first and second order derivatives are augmented

to the static feature, resulting in 39-dimensional feature vectors. The UBM involves 2048 Gaussian components and was trained with about 4000 female utterances selected from the Fisher database randomly. The dimensionality of the i-vectors is 400. The LDA model was trained with utterances of 7196 female speakers, again randomly selected from the Fisher database. The dimensionality of the LDA projection space is set to 150. For the metric learning, utterances in the Fisher database are sampled randomly to build the contrastive triples and are used to train the projection matrix.

4.3 Basic results

We first present the basic results obtained with various discriminative models: raw i-vectors with cosine scoring (Cosine), LDA, PLDA, max-margin metric learning (MMML). The test is based on the NIST SRE 2008 core task, which is divided into 8 test conditions according to the channel, language and accent [20]. The EER results are reported in Table 1.

It can be observed that the proposed MMML significantly improves the discriminative capability of raw i-vectors, and it outperforms both LDA and PLDA in conditions 1-4 (which takes the major proportion of the test data). In condition 5-8, the PLDA wins the competition. We attribute this discrepancy to the data imbalance in the development set: condition 5-8 involves complex patterns (e.g., multilingual speakers, different accents) that were not involved in the Fisher database that was used to train the discriminative models. This leads to performance degradation on these conditions with the MMML approach that we found heavily relies on large training data. For LDA and PLDA, the Gaussian assumption improves generalizablility on unseen conditions, thus resulting to superior performance than MMML, a purely discriminative approach. Nevertheless, since C1-C4 takes a large proportion of the data, the MMML approach get the best overall performance.

Condition	Cosine	LDA	PLDA	MMML
C1	29.34	22.11	18.57	**13.65**
C2	4.78	**1.19**	1.79	**1.19**
C3	29.66	22.65	18.70	**14.12**
C4	18.92	12.91	14.41	**10.66**
C5	20.31	14.42	**10.58**	11.42
C6	12.47	10.75	**9.42**	11.25
C7	7.73	5.58	**4.06**	6.08
C8	7.37	5.52	**4.21**	5.26
Overall	25.58	20.96	19.13	**15.64**

Table 1 EER results on NIST SRE 2008 core test. The best results are shown in bold face for each condition .

The DET curves on the overall test condition with the four models are presented in Fig 2. It is clearly observed that the MMML approach outperforms the other three.

4.4 Tandem composition

We note that both LDA and MMML learn a linear projection, though they are based on different learning criteria: LDA uses Fisher discriminant while MMML uses max-margin. The results in Table 1 show that the max-margin criterion is clearly superior. An interesting question is if the two criteria can be composed in a tandem way. The results are shown in Table 2, where the system 'LDA+MMML'

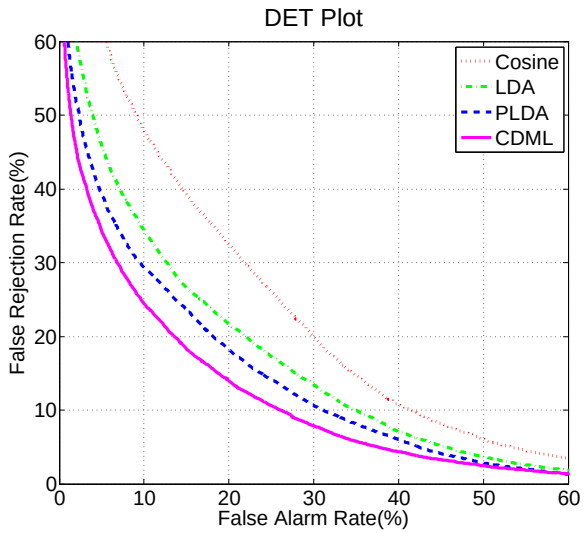

Figure 2 *The DET curves on the NIST SRE 2008 overall test condition.*

involves a 500×150 dimensional LDA projection followed by a 150×150 dimensional MMML projection, while the system 'MMML+LDA' involves a 500×150 dimensional MMML and a 150×150 dimensional LDA. From these results, we find that the *last* projection is the most important: if it is an MMML, the performance is always good. The 'MMML+LDA' system seems a bit superior than the original LDA, which is perhaps because the advantage of the max-margin training has been consolidated in the process of dimension reduction, which benefits the subsequent LDA.

Condition	LDA	MMML	MMML + LDA	LDA + MMML
C1	22.11	13.65	20.82	14.66
C2	1.19	1.19	1.19	0.90
C3	22.65	14.12	21.49	15.29
C4	12.91	10.66	11.86	10.96
C5	14.42	11.42	13.70	11.66
C6	10.75	11.25	11.03	11.14
C7	5.58	6.08	5.70	5.96
C8	5.52	5.26	4.47	5.26
Overall	20.96	15.64	20.49	15.47

Table 2 *EER results with tandem composition.*

4.5 Score fusion

The LDA/PLDA model and MMML model are complementary: LDA/PLDA are generative models and so better generalizable to rare conditions where little training data are available, whereas MMML is purely discriminative and is superior for matched conditions. Combing these two types of models may offer additional gains. We experimented with a simple score fusion approach that linearly interpolates the scores from LDA/PLDA and MMML. The results are presented in Table 3, where the interpolation factor for the MMML system is chosen to be 0.4. Compared to Table 1, we observe that the fusion leads to consistently better performance than the original LDA and PLDA systems. Interestingly, the performance on condition

5-8 is also improved, although the MMML approach does not work well individually in these conditions. Note that the performance degradation on condition 1,3,4 compared to the original MMML system is simply because we used a global interpolation factor. If the factor had been tuned for each condition separately, the fusion system would obtain the best performance in all the conditions.

Condition	LDA + MMML	PLDA + MMML
C1	16.45	16.22
C2	0.60	0.90
C3	17.04	16.53
C4	10.06	10.96
C5	11.54	9.38
C6	10.31	9.03
C7	5.32	4.06
C8	5.00	3.68
Overall	17.84	17.67

Table 3 EER results with score fusion, where the interpolation factor for MMML is chosen to be 0.4.

5 Conclusions

In this paper, we proposed a max-margin metric learning approach for speaker recognition. This approach is a simple linear transforms that is trained with the criterion of max-margin between true speakers and imposters based on cosine distance. It is as simple as LDA, but the performance is comparable or even better than PLDA, especially with large training data on matched conditions. Future work will investigate metric learning with non-linear transforms, and study better approach to combining PLDA and MMML.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grant No. 61371136 and No. 61271389, it was also supported by the National Basic Research Program (973 Program) of China under Grant No. 2013CB329302.

Author details

[1]Center for Speech and Language Technology, Research Institute of Information Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [2]Center for Speech and Language Technologies, Division of Technical Innovation and Development, Tsinghua National Laboratory for Information Science and Technology, ROOM 1-303, BLDG FIT, 100084 Beijing, China. [3]Department of Computer Science and Technology, Tsinghua University, ROOM 1-303, BLDG FIT, 100084 Beijing, China.

References

1. Patrick J. Kenny, G. Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1435–1447, 2007.
2. Patrick J. Kenny, G. Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Speaker and session variability in GMM-based speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, pp. 1448–1460, 2007.
3. Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
4. Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *INTERSPEECH'06*, 2006.
5. A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc Odyssey, Speaker Language Recognition Workshop 2004*, 2004, pp. 57–62.
6. Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Ouellet, and Pierre Dumouchel, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
7. Sergey Ioffe, "Probabilistic linear discriminant analysis," *Computer Vision ECCV 2006, Springer Berlin Heidelberg*, pp. 531–542, 2006.
8. Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey'2010: The Speaker and Language Recognition Workshop*, 2010.
9. Daniel Garcia-Romero and Carol Y Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *Interspeech*, 2011, pp. 249–252.
10. Lukas Burget, Oldrich Plchot, Sandro Cumani, Ondrej Glembek, Pavel Matejka, and Niko Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*. IEEE, 2011, pp. 4832–4835.
11. Sandro Cumani, Niko Brummer, Lukas Burget, Pietro Laface, Oldrich Plchot, and Vasileios Vasilakakis, "Pairwise discriminative speaker verification in the-vector space," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 6, pp. 1217–1227, 2013.
12. I. Hirano, Kong Aik Lee, Zhaofeng Zhang, Longbiao Wang, and A. Kai, "Single-sided approach to discriminative PLDA training for text-independent speaker verification without using expanded i-vector," in *Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on*, Sept 2014, pp. 59–63.
13. Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng, and Frank Soong, "Discriminative scoring for speaker recognition based on i-vectors," in *Asia-Pacific Signal and Information Processing Association, 2014 Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–5.
14. Liu Yang, "An overview of distance metric learning," *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2007.
15. Matthew Schultz and Thorsten Joachims, "Learning a distance metric from relative comparisons," *NIPS*, p. 41, 2004.
16. James Glass Xiao Fang, Najim Dehak, "Bayesian distance metric learning on i-vector for speaker verification," *INTERSPEECH*, 2013.
17. Waquar Ahmad, Harish Karnick, , and Rajesh M. Hegde, "Cosine distance metric learning for speaker verification using large margin nearest neighbor method," *Advances in Multimedia Information Processing*, pp. 294–303, 2014.
18. Kilian Q Weinberger, John Blitzer, and Lawrence K Saul, "Distance metric learning for large margin nearest neighbor classification," in *Advances in neural information processing systems*, 2005, pp. 1473–1480.
19. James Bergstra, Olivier Breuleux, Frédéric Bastien, Pascal Lamblin, Razvan Pascanu, Guillaume Desjardins, Joseph Turian, David Warde-Farley, and Yoshua Bengio, "Theano: A CPU and GPU math compiler in python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 3 – 10.
20. NIST, "The NIST year 2008 speaker recognition evaluation plan," *Online: http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf*, 2008.