CSLT TECHNICAL REPORT-20160001 [Monday 11th April, 2016]

Max-margin metric learning for speaker recognition

Lantian Li[1,2,3], Chao Xing[1,2] and Dong Wang[1,2*]

[*]Correspondence: wangdong99@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 discriminant 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 (MMML) 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 show that the MMML and PLDA models have respective advantages under different training/test conditions. With the number of utterances for each speaker increasing, the MMML has an advantage over PLDA model in the beginning, and then the PLDA model surpasses the MMML model when more utterances are available. On the other hand, as the number of speakers increases, MMML tends to deliver better performance.

Keywords: Max-margin; metric learning; PLDA; speaker recognition

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 [1]; 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 within-class covariance normalization (WCCN) [3], nuisance attribute projection (NAP) [4], linear discriminant analysis (LDA) [5], and its Bayesian counterpart, probabilistic linear discriminant analysis (PLDA) [6].

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 that 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 [7] 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 [8]. Burget, Cumani and colleagues proposed a pair-wised discriminative model that discriminates true speakers and imposters [9, 10]. 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 [10]), 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 [11], which optimizes the discriminative model on a subspace and does not require any feature expansion. In [12], 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 [13, 14], though it has not been extensively studied in speaker recognition. Besides, we hope to investigate the respective advantages of MMML and PLDA under different training conditions, and try to identify the conditions that each method is mostly suitable.

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.

[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.

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 [15] 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 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 [16] is more relevant to our proposal. The authors formulated the training task as a semidefinite program (SDP) [17] 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 accommodates 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 [13]. 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:

$$d(w_1, w_2) = \frac{< \mathbf{w}_1, \mathbf{w}_2 >}{\sqrt{||\mathbf{w}_1||||\mathbf{w}_2||}}. \tag{1}$$

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:

$$\mathcal{L}(M) = \sum_{(\mathbf{w},\mathbf{w}^+,\mathbf{w}^-)\in S} \max\{0, \delta - d(M\mathbf{w}, M\mathbf{w}^+) + d(M\mathbf{w}, M\mathbf{w}^-)\} \tag{2}$$

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 the 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:

$$M^t = M^{t-1} + \epsilon \frac{\partial \mathcal{L}}{\partial M} \tag{3}$$

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 [18] 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.

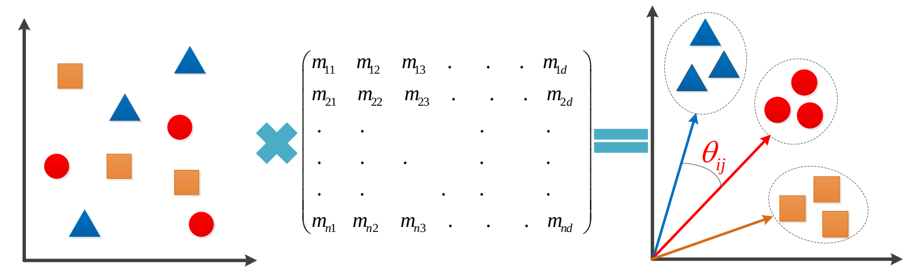
m_{11} m_{12} m_{13} . . . m_{1d}
m_{21} m_{22} m_{23} . . . m_{2d}
m_{n1} m_{n2} m_{n3} . . . m_{nd}
θ_{ij}

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

This section first presents the data used and the experimental setup, and then reports the results in terms of equal error rate (EER) and DET curves.

4.1 Database

In order to ensure the effectiveness and robustness of the proposed MMML method, two databases *Fisher telephone speech database*(Fisher) and *NIST 2005 speaker recognition evaluation*(SRE05) are used as the development sets to train the i-vector systems. For the evaluation, *NIST 2008 speaker recognition evaluation*(SRE08) is

Table 1 Evaluation conditions reproduced from [19]

Trial condition	Number of trials	Description
c1	19,776	only interview speech in training and test
c2	996	interview speech from the same microphone type in training and test
c3	18,780	interview speech from different microphones types in training and test
c4	6,693	interview training speech and telephone test speech
c5	4,408	telephone training speech and noninterview microphone test speech
c6	23,385	only telephone speech in training and test
c7	11,146	English language telephone speech in training and test
c8	5,233	only English language telephone speech spoken by a native U.S. English speaker in training and test
Total	59,343	All trials in evaluation set

used to evaluate the proposed method. Note that all the data are recordings of females.

- **Development sets**:
  - Fisher: 7196 female speakers with 13287 utterances are used to train the i-vector, LDA and PLDA models. The same data is also used to conduct the metric learning.
  - SRE05: 476 female speakers with 6677 utterances are used as the development set similar as *Fisher*.
- **Evaluation set**:
  - SRE08: The short2-short3 condition of SRE08 [19] is used as the evaluation set. The evaluation set consists of 1997 female enrollment utterances and 3858 test utterances, and it is based on the pair-wised 59343 trials, including 12159 target trials and 47184 imposter trials. Table 1 presents the test conditions, produced from [19]

4.2 Experimental setup

We largely follow the Kaldi SRE08 recipe to conduct the experiments. The acoustic feature is 19-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 60-dimensional feature vectors. The UBM involves 2048 Gaussian components and was trained with about 8000 female utterances randomly selected from either the Fisher or SRE05 database. The dimensions of the i-vector space and the LDA projection space are set to 400 and 150, respectively. For the metric learning, utterances either in the Fisher or SRE05 database are sampled randomly to build the contrastive triples and are used to train the projection matrix. In order to compare with the LDA model, the dimension of MMML projection space is chosen as 150. The margin δ is set to 1, the mini-batch size is set to 100, and the learning rate ϵ is set to 0.2. For each each i-vector w, 15 positive w^{+}s and negative w^{-}s are sampled to construct 15 contrastive triples. The value 15 is the optimal choice in our experiments. Note that a larger number of samples slows the training down, but resulted in more stable MMML models in our experiments.

4.3 Basic results

Two group of experiments are conducted with the two development sets (*Fisher* and *SRE05*) respectively. 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 [19]. The EER results are reported in Table 2 and Table 3.

It can be observed that there is significant discrepancy on the results of two groups of experiments. For the results in Table 2 where the development set is the *Fisher* corpus, the proposed MMML approach significantly improves the discriminative capability of raw i-vectors, and it outperforms both LDA and PLDA in condition 1-4 (which takes the major proportion of the test data). In condition 5-8, the PLDA wins the competition. Nevertheless, since condition 1-4 takes a large proportion of the data, the MMML approach gets the best overall performance.

For the results in Table 3 where the development set is the *SRE05* corpus, the PLDA is overwhelmingly superior to the purely discriminative MMML. Certainly, the proposed MMML still contributes: it outperforms the traditional cosine scoring approach in a significant way. In condition 6-8, MMML even outperforms LDA.

Condition	Cosine	LDA	PLDA	MMML
C1	28.65	22.34	19.63	**15.63**
C2	4.78	1.49	1.79	**1.19**
C3	28.60	22.29	19.96	**16.18**
C4	19.67	12.61	15.47	**12.61**
C5	20.79	14.18	**11.66**	12.98
C6	11.20	10.42	**8.31**	10.92
C7	7.35	6.08	**4.31**	6.34
C8	7.37	5.53	**4.74**	5.53
Overall	24.65	20.58	19.30	**16.02**

Table 2 *EER results on NIST SRE 2008 core test under the* Fisher *Development set. The best results are shown in bold face for each condition.*

Condition	Cosine	LDA	PLDA	MMML
C1	23.84	13.69	**13.59**	19.54
C2	5.97	2.09	**1.49**	3.88
C3	23.06	**13.28**	13.94	19.94
C4	19.67	**13.81**	15.32	16.82
C5	22.12	15.50	**14.42**	22.24
C6	14.02	11.36	**9.20**	10.70
C7	11.41	8.49	**6.21**	7.86
C8	12.63	10.00	**6.84**	8.16
Overall	22.02	**15.94**	15.95	18.05

Table 3 *EER results on NIST SRE 2008 core test under the* SRE05 *Development set. The best results are shown in bold face for each condition.*

4.4 Experimental validation

From the basic experiments, we observe that the three discriminative methods (LDA/PLDA/MMML) behave very differently when the models are developed based on different data (*Fisher* and *SRE05*). To discover the root of the difference, we analyze the two databases thoroughly and found that a clear difference between them is that the Fisher database consists of more speakers and each speaker just contains about 1-2 utterances, while the SRE05 database has less speakers but each speaker contains more utterances (On average, each speaker has about 13 utterances).

In order to verify if this discrepancy on data profiles (number of speakers and number of utterances per speaker) caused the different comparative advantages of MMML compared to other models, we conduct a serious of experiments in this section.

4.4.1 Utterance-sensitive test

In this experiment, the number of speakers is fixed and the number of utterances per speaker varies from 2 to 5. The *SRE05* database is used as the development set to train the LDA/PLDA/MMML models (Fisher does not contain such a large number of per-speaker utterances). For each speaker, the required n utterances are randomly selected. According to different n, four development datasets are constructed as shown Table 4.

Training Sets	SRE05-2	SRE05-3	SRE05-4	SRE05-5
Speakers	476	476	476	476
utterances	950	1404	1841	2271

Table 4 *Statistics of Utterance-sensitive development datasets.*

The experiment uses the UBM model and i-vector model previously trained using the entire SER05 database (used already to produce the results in Table 3). The four development sets are used to train the LDA, PLDA and MMML models respectively. The results are presented in Fig. 2 and Fig. 3.

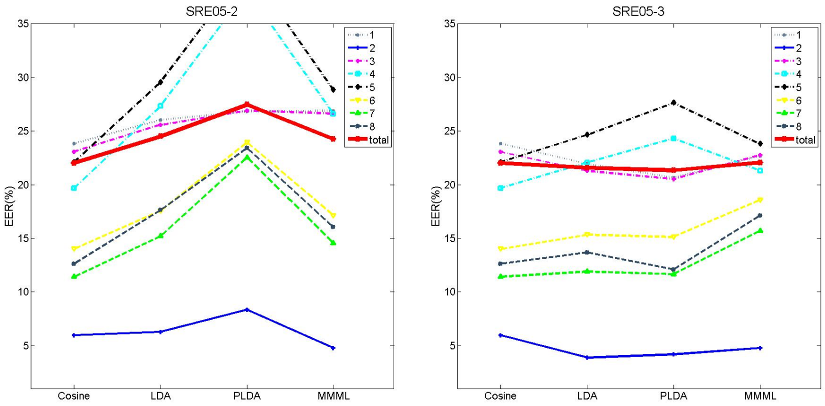

Figure 2 *Performance of NIST SRE 2008 core test with four discriminative model. Each line is represented by each test condition. For the left figure,* ***two*** *utterances are randomly selected from each speaker in SRE05. For the right figure,* ***three*** *utterances are randomly selected from each speaker in SRE05.*

It can be seen that with the number of per-speaker utterances increasing, the performance with the three discriminative model (LDA, PLDA and MMML) are all improved, however the contribution of extra utterances is significantly different. Focusing on the test condition 'total', we find that with 2 utterance per speaker, all the discriminative models are not effective. With 3 utterances, all the discriminative models become effective and outperform the baseline cosine scoring, and MMML shows clear advantage compared to LDA and PLDA. With more utterances, the improvement for MMML and LDA is marginal, while PLDA still benefits from the extra data and outperforms MMML.

This observation are true for individual test conditions: with more utterances, the improvement with PLDA is much more significant than LDA and MMML. This means PLDA is 'utterance-sensitive'. The difference between LDA and MMML is not significant.

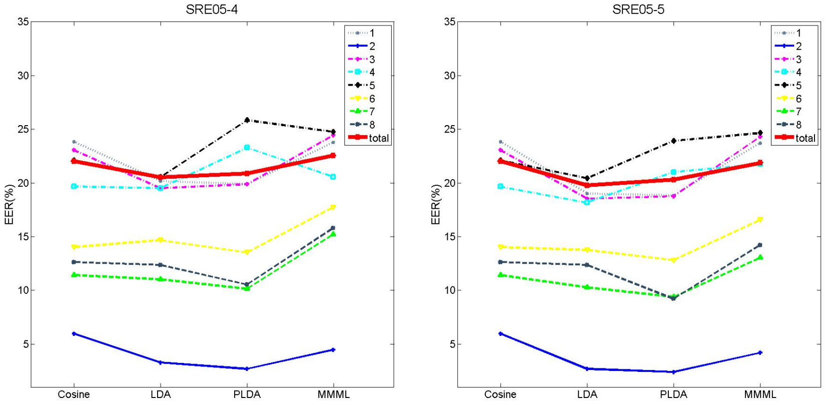

Figure 3 *Performance of NIST SRE 2008 core test with four discriminative model. For the left figure,* ***four*** *utterances are randomly selected from each speaker in SRE05. For the right figure,* ***five*** *utterances are randomly selected from each speaker in SRE05.*

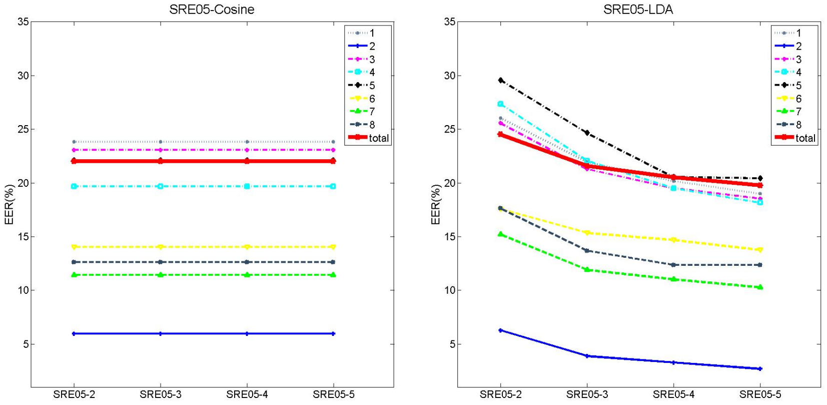

Figure 4 *Utterance-sensitive test based on* ***Cosine*** *and* ***LDA*** *models.*

Moreover, from the results with SRE05-2 and SRE05-3, we can see the advantage of MMML mostly exhibits when the EERs are high, i.e., condition 1, 3, 4, 5. Another condition that MMML wins is condition 2. All these conditions share the same property: there are some channel mismatch between model training and enrollment/test. Note that the SRE05 data are mostly telephone speech, while the enrollment and/or test data in all the above conditions involve microphone speech. This suggests that the probable reason for the disadvantage of PLDA we observed in this experiment is two-fold: limited per-speaker utterances for model training, and channel mismatch between model training and enrollment/test. This is reasonable since PLDA is a hierarchical linear Gaussian model and requires sufficient data to estimate model parameters correctly. With few utterances, most of them are telephone speech, and so the variation can not be well modeled in the population level (priors for speaker means) as well as speaker level (inter-session variation).

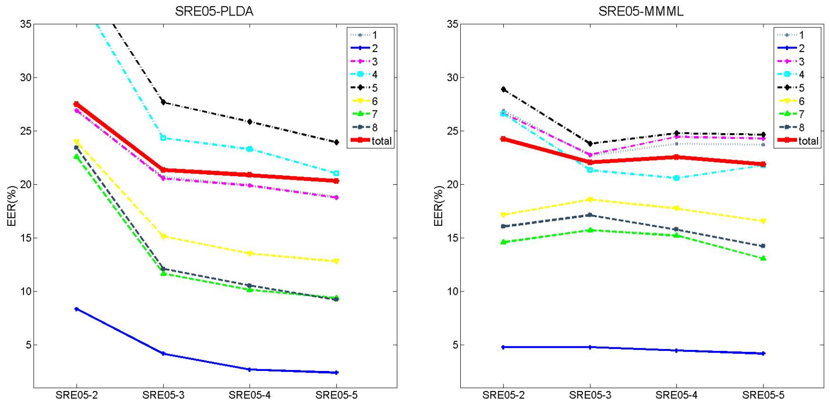

Figure 5 *Utterance-sensitive test based on* ***PLDA*** *and* ***MMML*** *models.*

This leads to bad performance in conditions where the enrollment/test utterances involve microphone speech. With more data available, microphone data are more sampled and the channel variation is addressed by PLDA.

4.4.2 Speaker-sensitive test

In this experiment, we fix the number of utterances per speaker and change the number of speakers in the development set. Thanks to the large number of speakers of the *Fisher* database, we can use it to conduct the experiments. Specifically, we randomly select n speakers from the Fisher database, where 99% speakers just have 1 to 3 utterances. Four datasets are constructed according to n, as shown in Table 5. Note that larger n leads to more utterances in total, but less utterances per speaker.

Training Sets	Fisher-1	Fisher-2	Fisher-3	Fisher-4
Speakers	1867	2396	4803	7196
utterances	4557	4429	8859	13287

Table 5 *Parameters of speaker-directed condition. Note that in order to compare with the results from SRE05-2 and SRE05-3, we randomly selected* 1867 *speakers from the Fisher and each speaker contains* 2 *or* 3 *utterances. Besides, Fisher-2, Fisher-3, Fisher-4 are selected according to the utterance number.*

The UBM and i-vector models used here are trained with the Fisher database and are the same as those used in Table 2. The four datasets are used to train the LDA, PLDA and MMML models. The EER results are presented in Fig 6 and Fig 7.

From the test condition 'total', it can be seen that the advantage of MMML is more clear compared to PLDA/LDA with more speakers involved in model training. More speakers seem does not change the performance of PLDA, but it provides significant performance on MMML. (We suspect that this advantage may be caused by the less per-speaker utterances with more speakers?????)

In more details, MMML outperforms PLDA in condition 1-4, while in condition 5-8 the PLDA wins. This results are consistent to the results in 2, except condition 5 for which MMML performs better better than PLDA in the utterance-sensitive test. A possible reason is that the Fisher database is purely telephone....

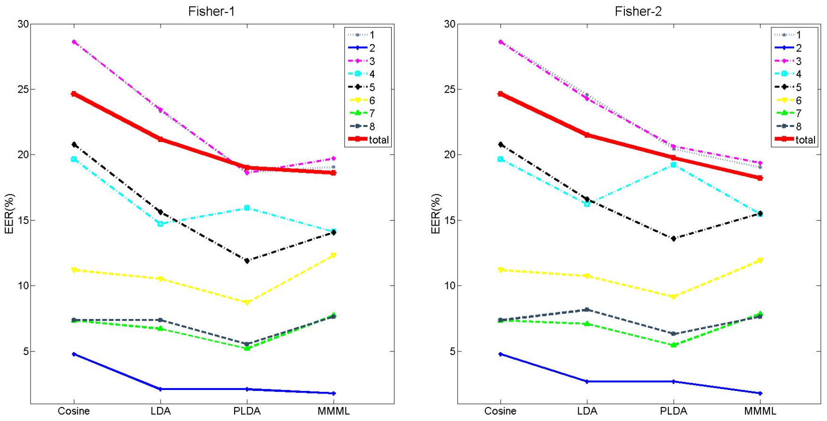

Figure 6 *Performance of NIST SRE 2008 core test with four discriminative model. For the left figure,* ***Fisher-1*** *as the training set is used for LDA, PLDA and MMML training. For the right figure,* ***Fisher-2*** *is used as the training set.*

4.5 Combination

In this section, two combination approaches have been proposed.

- **Tandem composition**

We note that both MMML and LDA learn a linear projection, though they are based on different learning criteria: LDA uses Fisher discriminant while MMML uses max-margin. The results in Section 4.3 and Section 4.4 show that MMML and LDA have respective advantages. An interesting question is whether the two criteria can be composed in a tandem way. The results are shown in Table 6 and Table 7, where the system 'LDA+MMML' involves a 400 × 150 dimensional LDA projection followed by a 150×150 dimensional MMML projection, while the system 'MMML+LDA' involves a 400 × 150 dimensional MMML and a 150 × 150 dimensional LDA. From these results, we find that the *last* projection is the most important. For example, in Table 6, LDA is inferior to MMML and the 'MMML+LDA' has the same performance as LDA. Meanwhile, the results of 'LDA+MMML' is analogous to MMML. It seems that the tandem composition of the two linear projection methods does not dig out more discriminative information. However, the 'MMML-PLDA' tandem composition achieves fairly good performance. We attribute it to ...

Condition	LDA	PLDA	MMML	MMML + LDA	LDA + MMML	MMML + PLDA
C1	22.34	19.63	15.63	21.39	16.45	**13.13**
C2	1.49	1.79	1.19	1.79	1.19	**0.90**
C3	22.29	19.96	16.18	21.38	17.07	**13.09**
C4	12.61	15.47	12.61	13.06	13.21	**11.11**
C5	14.18	11.66	12.98	12.74	13.10	**10.34**
C6	10.42	**8.31**	10.92	10.31	11.03	9.81
C7	6.08	**4.31**	6.34	6.21	6.21	5.20
C8	5.53	4.74	5.53	6.05	5.53	**4.47**
Overall	20.58	19.30	16.02	19.85	16.27	**15.86**

Table 6 *EER results with tandem composition under the* Fisher *Development set. The best results are shown in bold face for each condition.*

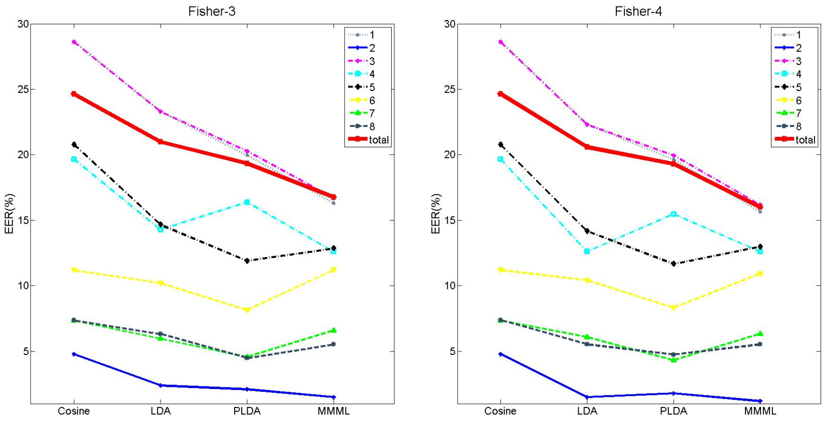

Figure 7 *Performance of NIST SRE 2008 core test with four discriminative model. For the left figure,* ***Fisher-3*** *as the training set is used for LDA, PLDA and MMML training. For the right figure,* ***Fisher-4*** *is used as the training set.*

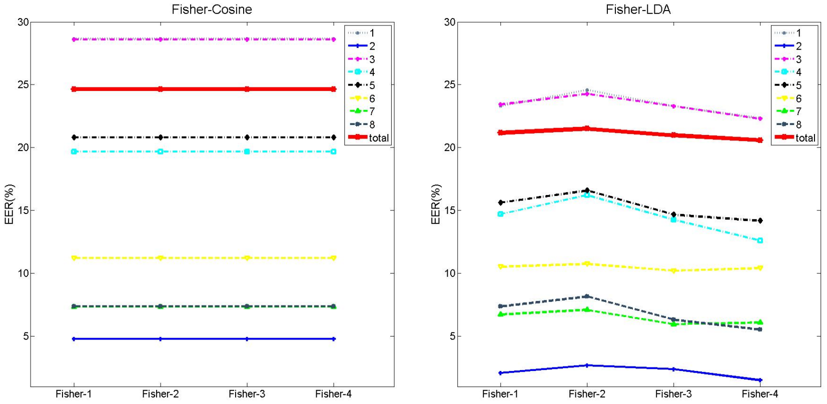

Figure 8 *Speaker-sensitive test based on* ***Cosine*** *and* ***LDA*** *models.*

The DET curves on the overall test condition with the six models are presented in Fig 10. It is clearly observed that the 'MMML+PLDA' approach outperforms the others.

- **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. Combining 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 fusion function is $\alpha s_{mmml} + (1-\alpha)s_{lda/plda}$, where s_{mmml} and $s_{lda/plda}$ are scores from the MMML and LDA/PLDA systems respectively, and α is the interpolation factor. The results are presented in Figure 11 and Figure 12.

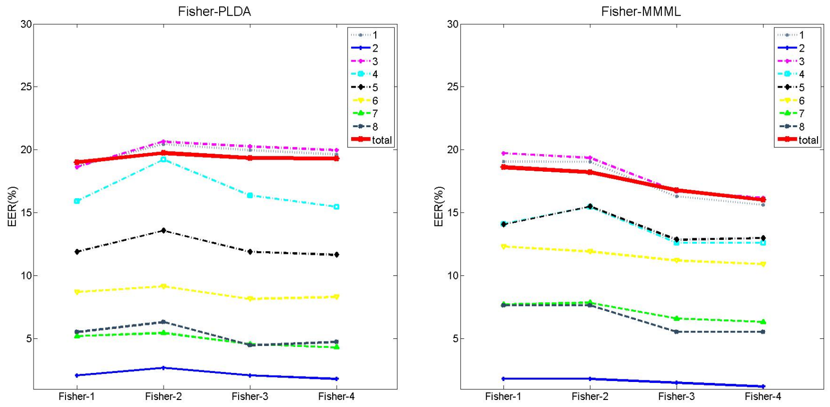

Figure 9 *Speaker-sensitive test based on* ***PLDA*** *and* ***MMML*** *models.*

Condition	LDA	PLDA	MMML	MMML + LDA	LDA + MMML	MMML + PLDA
C1	13.69	13.59	19.54	14.32	17.19	**11.61**
C2	2.09	**1.49**	3.88	2.39	3.28	2.39
C3	13.28	13.94	19.94	13.80	17.45	**11.60**
C4	13.81	15.32	16.82	14.71	16.67	**12.61**
C5	15.50	14.42	22.24	14.54	20.31	**11.78**
C6	11.36	**9.20**	10.70	10.98	10.92	9.53
C7	8.49	**6.21**	7.86	7.73	7.73	6.97
C8	10.00	**6.84**	8.16	9.21	8.16	8.16
Overall	15.94	15.95	18.05	15.58	17.47	**12.62**

Table 7 *EER results with tandem composition under the* SRE05 *Development set. The best results are shown in bold face for each condition.*

We observe that the score 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. If the interpolation factor α had been tuned for each condition separately, the fusion system would obtain the best performance in all the conditions.

5 Conclusions

In this paper, we proposed a max-margin metric learning approach for speaker recognition. This approach is a simple linear transform that is trained with the criterion of max-margin between true speakers and imposters based on cosine distance. Besides, from both 'utterance-directed' condition and 'speaker-directed' condition, we explored the performance tendency between MMML, LDA and PLDA and explicitly interpreted the application scenarios of each method. Moreover, two system combination schemes were proposed to further improve recognition performance. Future work will investigate metric learning with non-linear transforms, and study better approaches to combining PLDA and MMML.

Acknowledgement

This research was supported by the National Science Foundation of China (NSFC) under the project No. 61371136, and the MESTDC PhD Foundation Project No. 20130002120011. It was also supported by Sinovoice and Huilan Ltd.

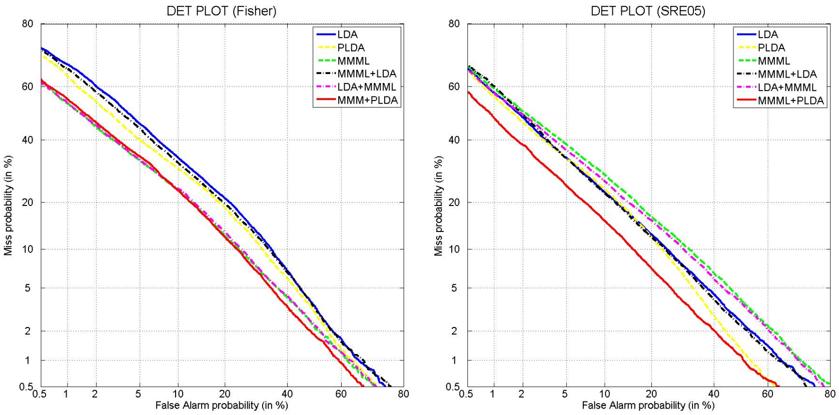

Figure 10 *The DET curves on the NIST SRE 2008 overall test condition under the* Fisher *or* SRE05 *development set.*

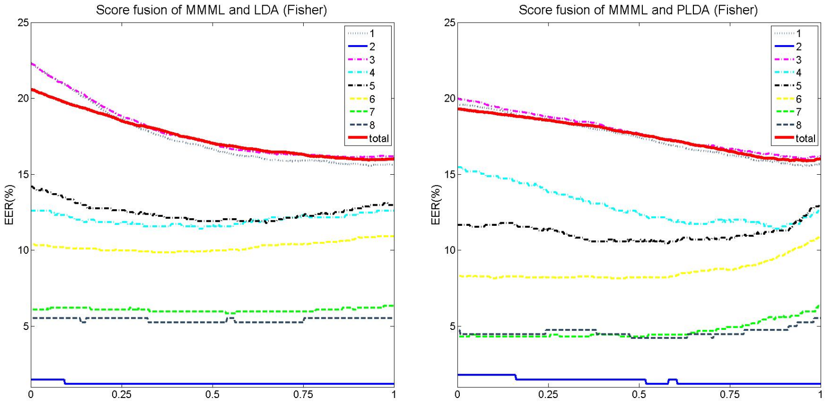

Figure 11 *EER results with score fusion under the* Fisher *development set.*

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, 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, 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. Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in *INTERSPEECH*, 2006.
4. 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.
5. Najim Dehak, Patrick J. Kenny, Reda 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.

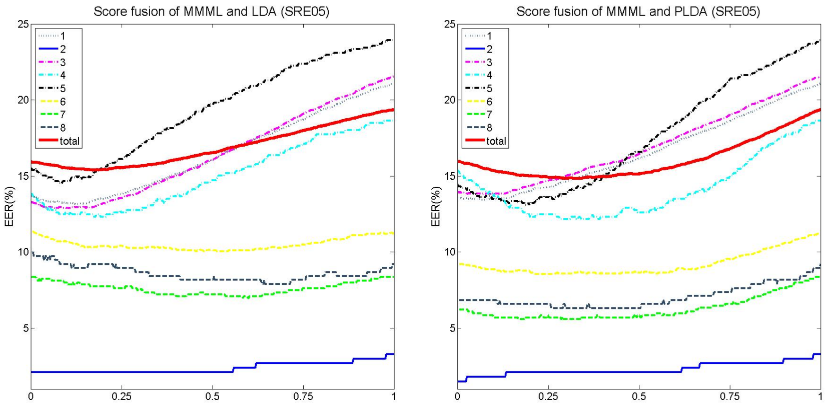

Figure 12 *EER results with score fusion under the* SRE05 *development set.*

6. Sergey Ioffe, “Probabilistic linear discriminant analysis,” *Computer Vision ECCV 2006, Springer Berlin Heidelberg*, pp. 531–542, 2006.
7. Patrick Kenny, “Bayesian speaker verification with heavy-tailed priors,” in *Odyssey'2010: The Speaker and Language Recognition Workshop*, 2010.
8. Daniel Garcia-Romero and Carol Y Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems,” in *INTERSPEECH*, 2011, pp. 249–252.
9. 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.
10. 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.
11. 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.
12. 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.
13. Liu Yang, “An overview of distance metric learning,” *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2007.
14. Matthew Schultz and Thorsten Joachims, “Learning a distance metric from relative comparisons,” *NIPS*, p. 41, 2004.
15. James Glass Xiao Fang, Najim Dehak, “Bayesian distance metric learning on i-vector for speaker verification,” *INTERSPEECH*, 2013.
16. 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.
17. 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.
18. James Bergstra, Olivier Breuleux, Frederic 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.
19. 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.