i-vector空间下intersession的补偿及打分方法综述

王　军

2013年11月03日

Center for Speech and Language Technologies
GROUPING

提 纲

- 说话人确认系统框架
- i-vector空间下intersession的补偿及打分方法
- ALIZE3.0测试实验
- 参考文献

一、说话人确认系统框架

- 说话人确认[S. Furui, 1981; D. A. Reynolds, 2003;]：确定一段说话人的语句是否与所声明的参考说话人相符，接收或是拒绝。

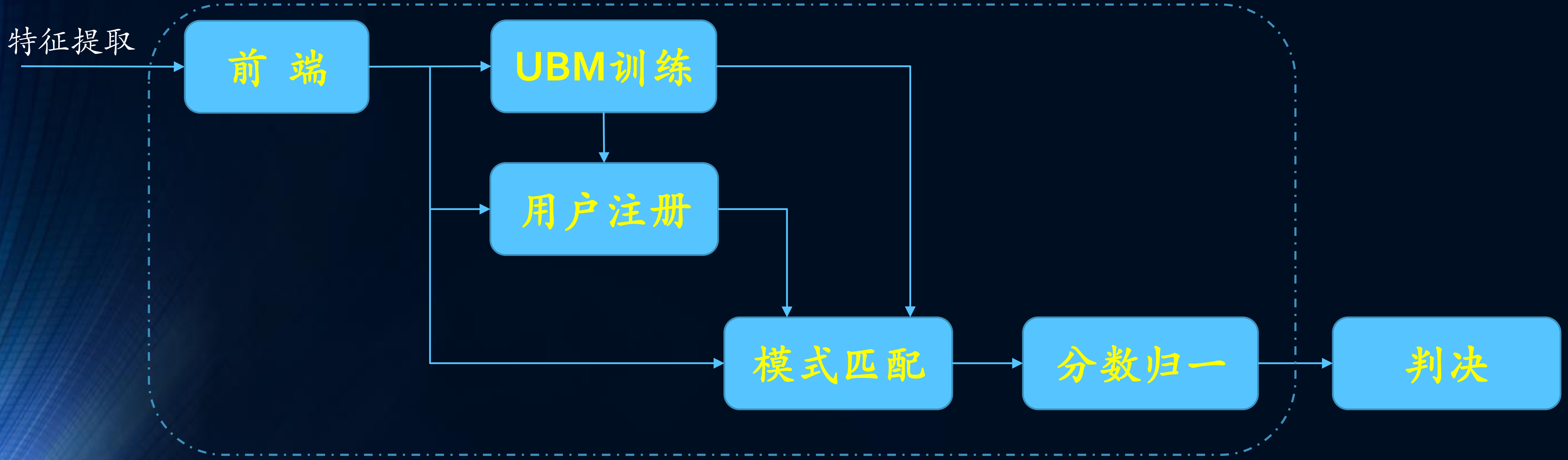


图1 说话人确认系统框架

- i-vector [N. Dehak, 2011]
  - 利用因子分析定义了低维度的total variability空间，在此空间中utterance表示成一个低维度向量，信道补偿就可以在低维度空间进行。
  - M=m+Tw
  - session variability用仅包含session信息的vectors的方差矩阵建模。
  - 依赖于session的vector可以通过对一个给定说话人的i-vector，减去此人所有sessions的vector的均值得到。
  - 这种思想与数值分析领域紧密联系在一起。

二、ivector空间下主流的intersession补偿及打分方法

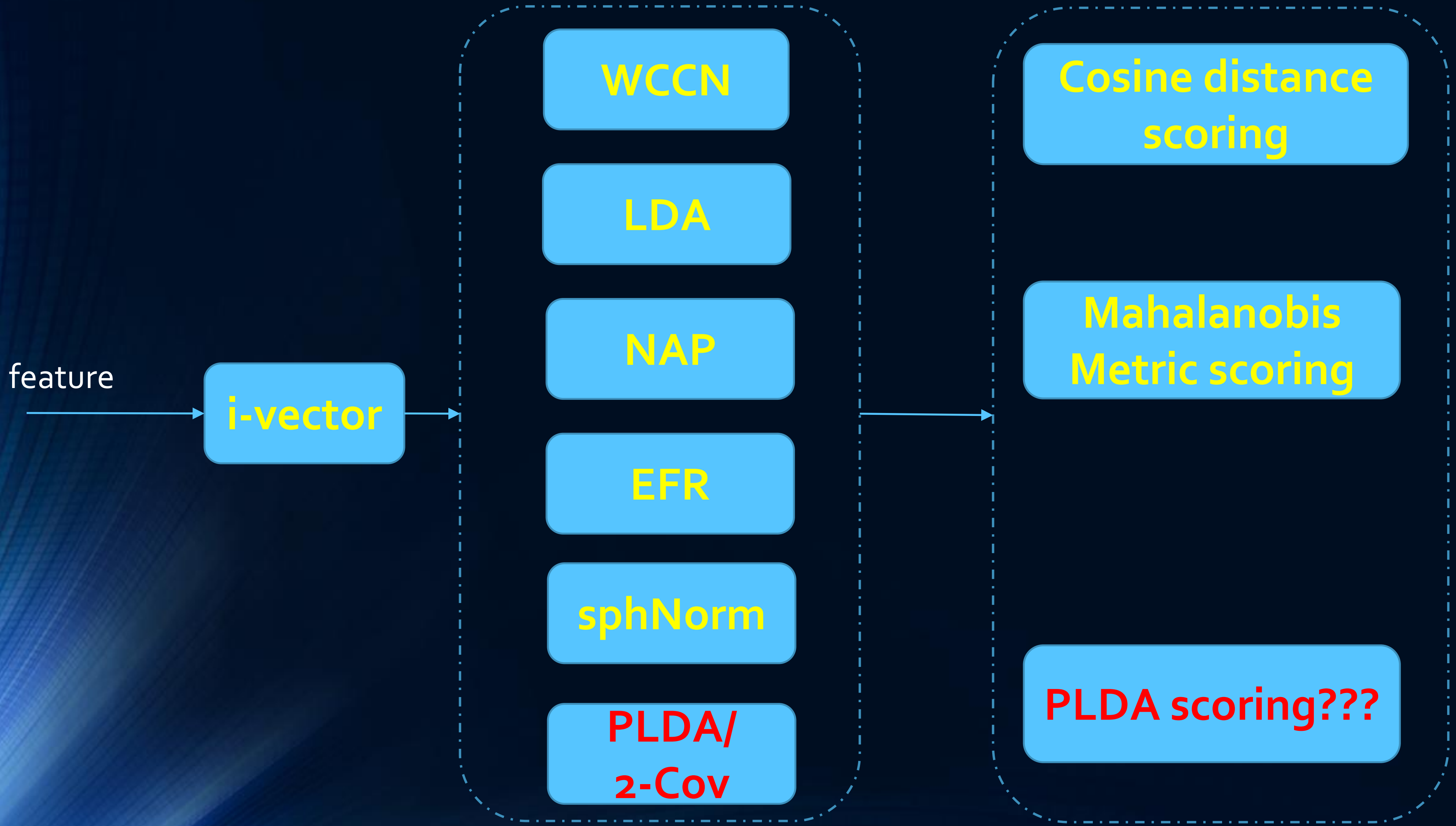


图2 i-vector下主流intersession补偿及打分方法

Cosine Distance Scoring [N. Dehak, 2011]

- 采用Cosine打分，Dehak认为非说话人信息影响i-vector的幅度。
- Factor analysis仅充当特征提取的角色。
- 打分过程速度更快。
  - $score(\omega_1, \omega_2) = \frac{\omega_1^t \omega_2}{\sqrt{\omega_1^t \omega_1}\sqrt{\omega_2^t \omega_2}}$

- WCCN (Within-Class Covariance Normalization) [A. Hatch, 2006]

◆WCCN 的思想是在SVM训练时最小化错误接受和错误拒绝的期望。目的是补偿intersession variability.

- $k(\omega_1, \omega_2) = \omega_1^t R \omega_2$
- $R = W^{-1}$
- $W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(\omega_i^s - \overline{\omega_s})(\omega_i^s - \overline{\omega_s})^t$
- $\omega' = B^t\omega$
- $score(\omega_1, \omega_2) = \frac{(\omega_1')^t\omega_2}{\sqrt{(\omega_1')^t\omega_1}\sqrt{(\omega_2')^t\omega_2}}$

LDA (Linear Discriminant Analysis) [N. Dehak, 2011]

- LDA 寻找具有更好类区分度的坐标系.
  - $J(v) = \frac{v^t S_b v}{v^t S_w v}$ Reyleigh coefficient
  - $S_b = \sum_{s=1}^{S} (w_s - \overline{w})(w_s - \overline{w})^t$
  - $S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_i^s - \overline{\omega_s})(\omega_i^s - \overline{\omega_s})^t$
  - $S_b v = \lambda S_w v$
  - $\omega' = A^t \omega$
  - $score(\omega_1, \omega_2) = \frac{(\omega_1')^t \omega_2}{\;}$

- NAP (Nuisance Attribute Projection) [N. Dehak, 2011; W. M. Campbell, 2006]

- LDA 寻找具有更好类区分度的坐标系.

  - $P = I - RR^t$
  - R的列是W的前k个eigenvector
  - $\omega' = P\omega$
  - $score(\omega_1, \omega_2) = \frac{(\omega_1')^t \omega_2}{\sqrt{(\omega_1')^t \omega_1}\sqrt{(\omega_2')^t \omega_2}}$

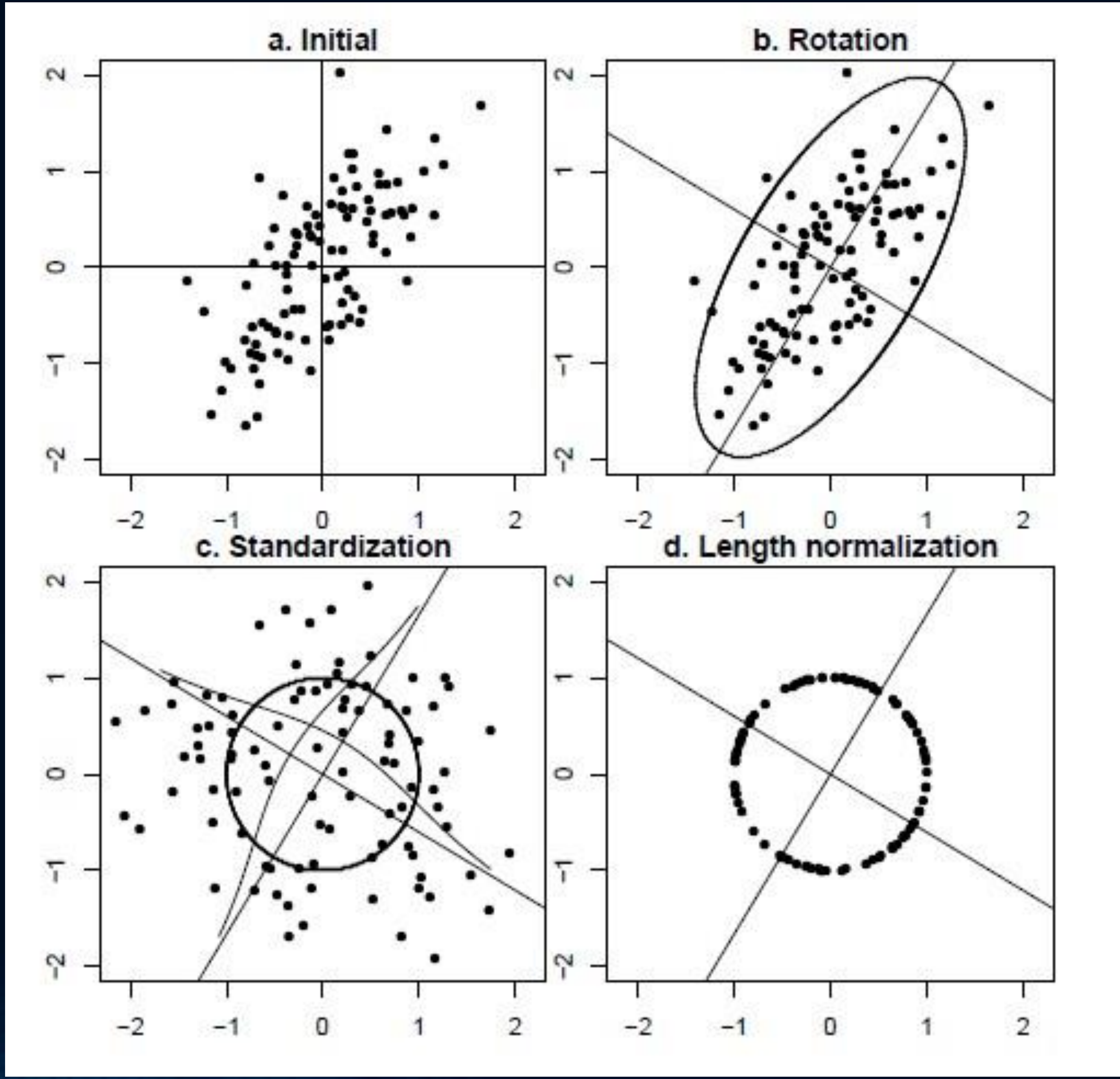
EFR (Eigen Factor Radial) [P.-M. Bousquet, 2011]

- i-vector理论上需满足$\mathcal{N}(0, I)$。
- channel或session的影响不仅是线性而且有非线性。LDA去除线性影响同时降维。
- 由T矩阵降维得到的i-vector直接通过LDA映射至低维空间。作者认为在T的满秩空间进行区分性变换比在LDA降维后变换更好。

$$\omega' = \frac{D^{-\frac{1}{2}} P^t (\omega - \overline{\omega})}{\sqrt{(\omega - \overline{\omega})^t V^{-1} (\omega - \overline{\omega})}}$$

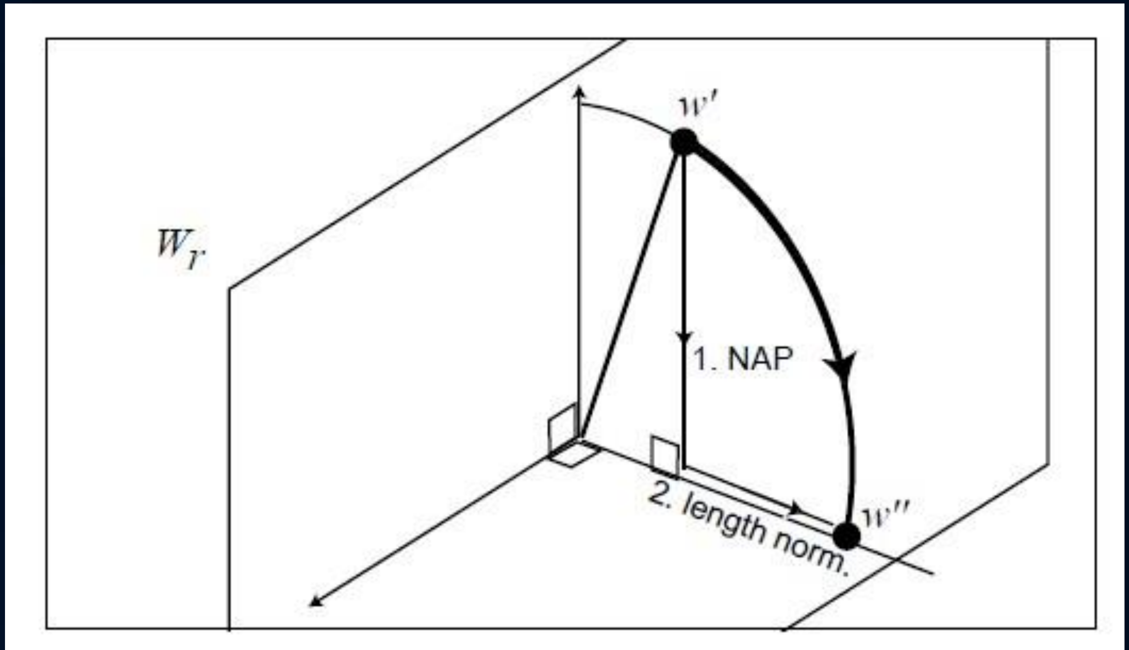
- EFR (Eigen Factor Radial) [P.-M. Bousquet, 2011]

➢ $\omega' = \dfrac{D^{-\frac{1}{2}} P^t (\omega - \overline{\omega})}{\sqrt{(\omega - \overline{\omega})^t V^{-1} (\omega - \overline{\omega})}}$



- EFR-NAP [P.-M. Bousquet, 2011]

$$\omega' = \frac{(\omega - \mathrm{P}\omega)}{\sqrt{(\omega - \mathrm{P}\omega)^t(\omega - \mathrm{P}\omega)}}$$



- Mahalanobis metric scoring [P.-M. Bousquet, 2011]

$$score(\omega_1, \omega_2) = (\omega_1 - \omega_2)^t W^{-1} (\omega_1 - \omega_2)$$

三、ALIZE3.0的测试实验[A. Larcher, 2013]

- 实验采用ALIZE3.0开源的说话人识别的工具包。[A. Larcher, 2013]
- 实验数据如下表。测试为NIST SRE06 core test。

	Switchboard	NIST2004	NIST2005
UBM	-	x	x
T	-	x	x
WCCN	-	x	x
LDA	-	x	x

表一 本实验训练UBM, T, WCCN,LDA的数据库

- 实验配置：
  - 50维度MFCC特征（19 MFCC， 19 delta, 11 delta-delta. E delta）
  - 2048 UBM。 （6974 sessions）
  - 500 rank T。 （6974 sessions，最新加入switchboard库，15290 sessions）
  - WCCN, ivNorm, PLDA （517 speakers，4000左右sessions; 最新采用1410+517 speakers， 19890 左右sessions）
  - Target 462人，测试语音2192条，共计30637次测试

实验结果

NIST SRE2010	EER%
WCCN+Cosine	5.81
WCCN+LDA(150)+Cosine	3.65
EFR+Mahalanobis	2.53
SphNorm+2Cov	2.23
Plda(400, 0)	4.90
LengthNorm+Plda(400, 0)	2.33
SphNorm+Plda(400, 0)	2.24

NIST SRE06	EER%
SVM FA	5.39
Cosine	3.84
WCCN+Cosine	2.76
LDA(250)	3.31
WCCN+LDA(200)+Cosine	2.72

表二 论文中的实验结果

实验结果

	EER%
GMM-UBM（without norm）	14,78
Cosine	10.09
WCCN+Cosine	9.69
WCCN+LDA(400)+Cosine	9.35
WCCN+LDA(250)+Cosine	9.20
EFR+Mahalanobis	8.46%
SphNorm+2Cov	9.20
SphNorm+Plda(400, 0)	8.01

表三 测试的实验结果

存在问题：训练数据及人数严重不足！！！

参考文献

[1] S. Furui. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Processing, 1981. 29(2):254-272.

[2] D.A. Reynolds. Channel robust speaker verification via feature mapping. In ICASSP, 2003, (2): 53-56.

[3] N. Dehak, P. Kenny, R. Dehak, et al. Front-end factor analysis for speaker verification[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(4): 788-798.

[4] A. Larcher, J. Bonastre and B. Fauve, et al. "ALIZE 3.0 - Open Source Toolkit for State-of-the-Art Speaker Recognition", in Proc. Interspeech 2013.

[5] P.-M. Bousquet, D. Matrouf, and J.-F. Bonastre, "Intersession compensation and scoring methods in the i-vectors space for speaker recognition," in Annual Conference of the International Speech Communication Association (Interspeech), 2011, pp. 485– 488.

[6] A. Hatch and A. Stolcke, "Generalized linear kernels for one-versus-all classification: application to speaker recognition," in to appear in proc. of ICASSP, Toulouse, France, 2006.

[7] A. Hatch, S. Kajarekar, and A. Stolcke, "Within-class covariance normalization for SVM-based speaker recognition," in Proc. Int. Conf. Spoken Lang. Process., Pittsburgh, PA, Sep. 2006.

[8] The NIST Year 2006 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2006/sre-06_evalplan-v9.pdf.

[9] W. M. Campbell, D. E. Sturim, D. A. Reynolds and A. Solomonoff. SVM based speaker verification using a GMM supervector kernel and NAP variability compensation. ICASSP, 2006. 97-100

[10] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in International Conference on Computer Vision. IEEE, 2007, pp. 1–8.