A tutorial on speaker verification

Student: Jun Wang
Supervisor: Thomas Fang Zheng
Tsinghua University, 2014,4

Outline

- Introduction
- GMM-UBM framework of speaker verification
- The ivector methodology of speaker verification
- Intersession compensation and scoring method for ivector
- Toolkits and database
- Some of my previous work
- References

1 Introduction

- Speaker recognition is a technique to recognize the identity of a speaker from a speech utterance.

spk recognition

- spk identification
  spk verification
- text dependent
  text independent
- close set
  open set

- My research area focus on the open-set, text-independent speaker verification.

A multitude of researches have been conducted to address the following three fields:

- Speech parameterization
- Pattern matching
- Scoring method

fig1 main research fields in speaker recognition

Speech parameterization (feature extractor)

Speech parameterization consists in transforming the speech signal to a set of feature vectors. Most of the speech parameterizations used in speaker verification systems relies on a cepstral representation of speech.[F. Bimbot, 2004]

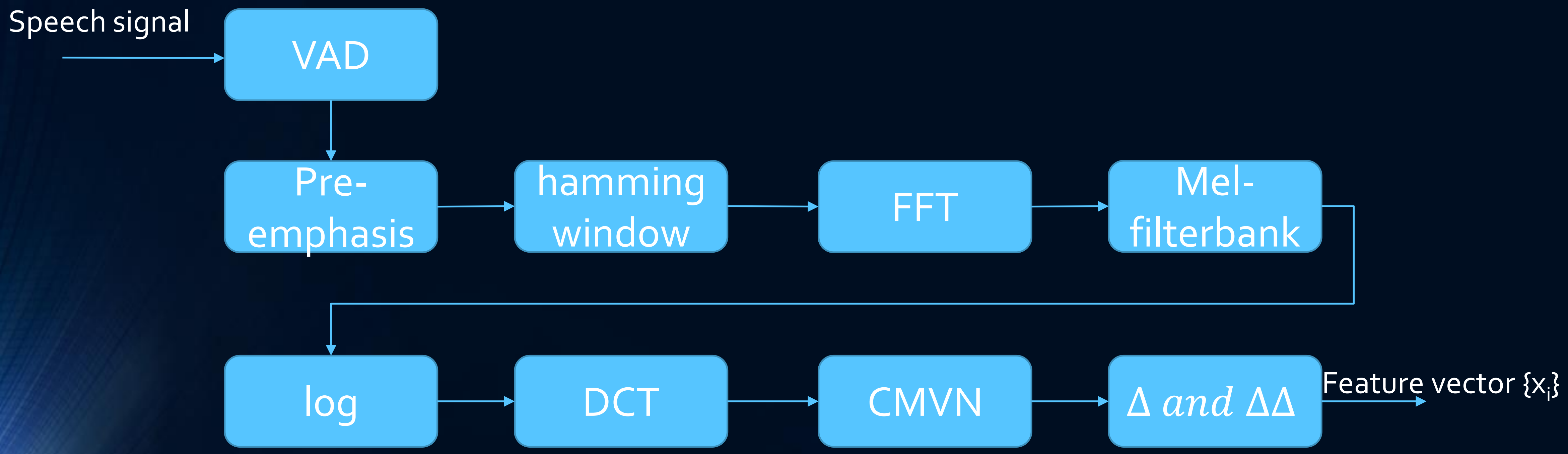

fig2 modular representation of mfcc feature extractor

- Main approaches in pattern matching for speaker recognition

main approach
- Template matching
  - Nearest neighbor [A. Higgins, 1993]
  - Vector quantization [F. Soong, 1985]
- Probabilistic model
  - Gaussian Mixture Model [A. Reynolds, 2003]
  - Joint factor analysis [P. Kenny, 2006]
  - ivector [N. Dehak, 2011]
- Artificial Neural Network
  - time delay neural work [Y. Bennani , 1991]
  - decision tree [K. R. Farrell, 1991]

Performance measure

- For speaker identification:

$$Recognition\ Rate = \frac{number\ of\ correct\ recognition}{total\ number\ of\ trials}$$

- For speaker verification:

$$False\ Reject\ Rate = \frac{number\ of\ rejective\ true\ speaker}{total\ number\ of\ true\ speaker}$$

$$False\ Acceptance\ Rate = \frac{number\ of\ accepted\ imposter}{total\ number\ of\ imposter}$$

$$\mathrm{EER} = False\ Reject\ Rate = False\ Acceptance\ Rate$$

Detection error tradeoff (DET) curve is often used to describe the performance.

Cost function (C_{DET}) is also defined as a weighted sum of FAR and FRR. [NIST, 2008]

2 GMM-UBM framework of speaker verification

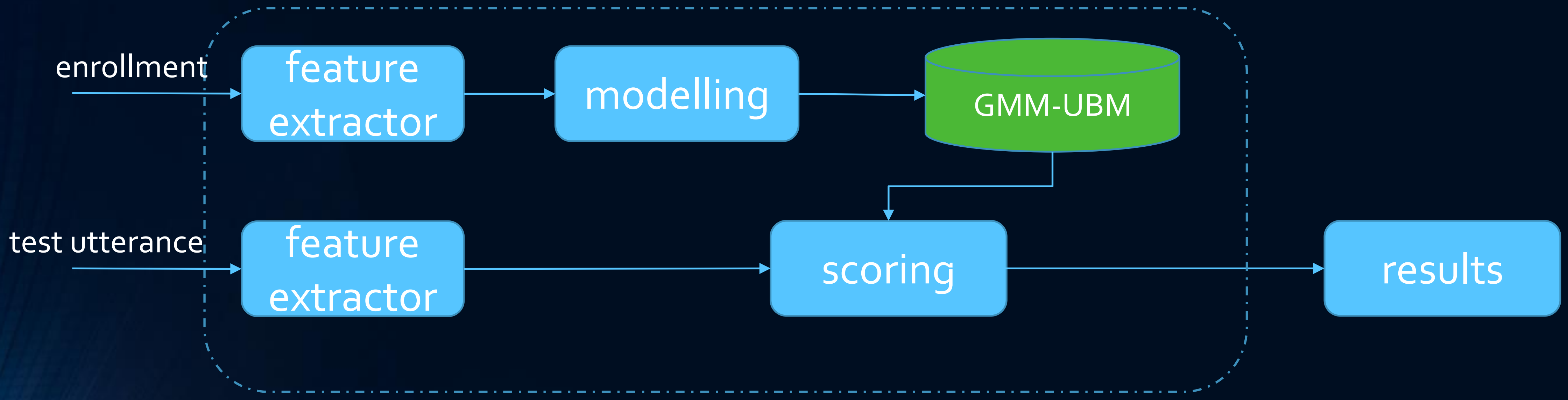

fig3 speaker verification framework

Speaker verification[S. Furui, 1981; D. A. Reynolds, 2003]：to verify a speech utterance belongs to a specified enrollment, accept or reject.

- GMM-UBM framework [D. A. Reynolds, 2000]
  - Gaussian Mixture Model is used to modeling the probability density function of a multi-dimensional feature vector.
  - Given a speech feature vector X={x_i} of dimension F, the probability density of x_i given a C GMM speaker model λ is given by:

$$p(x_i|\lambda) = \sum_{c=1}^{C} w_c g(x_i, \mu_c, \Sigma_c)$$

$$\sum_{c=1}^{C} w_c = 1$$

- The UBM is trained using EM algorithm and a speaker GMM is estabilished by adjusting the UBM parameters by MAP.

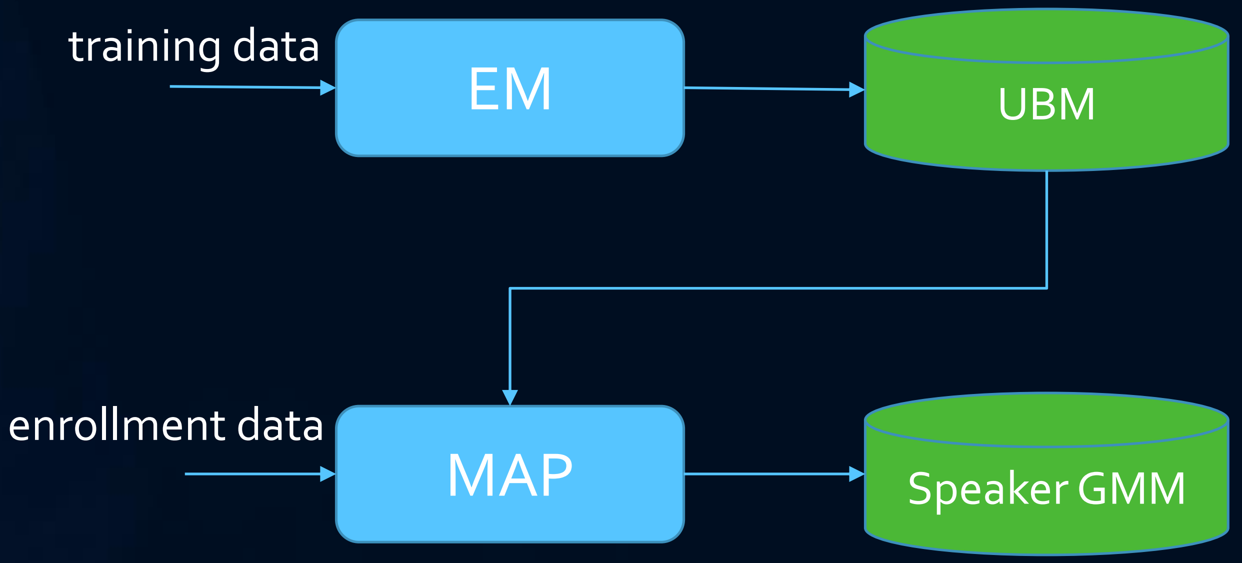

fig4 modeling methods for GMM-UBM

- From distribution:
  - A speaker utterance is represented by GMM which is adapted from the UBM via MAP.

$$M=m+Dz$$

  - UBM m represents all acoustic and phonetic variations in speech data where m is a supervector with dimension CF.
  - D is diagonal matrix in full space (CF×CF) and z is normally distributed random vector with dimension CF。
  - $M \sim N(m, DD^T)$。

3 ivector methodology of speaker verification

- Over recent years, ivector has demonstrated state-of-the-art performance for speaker verification.

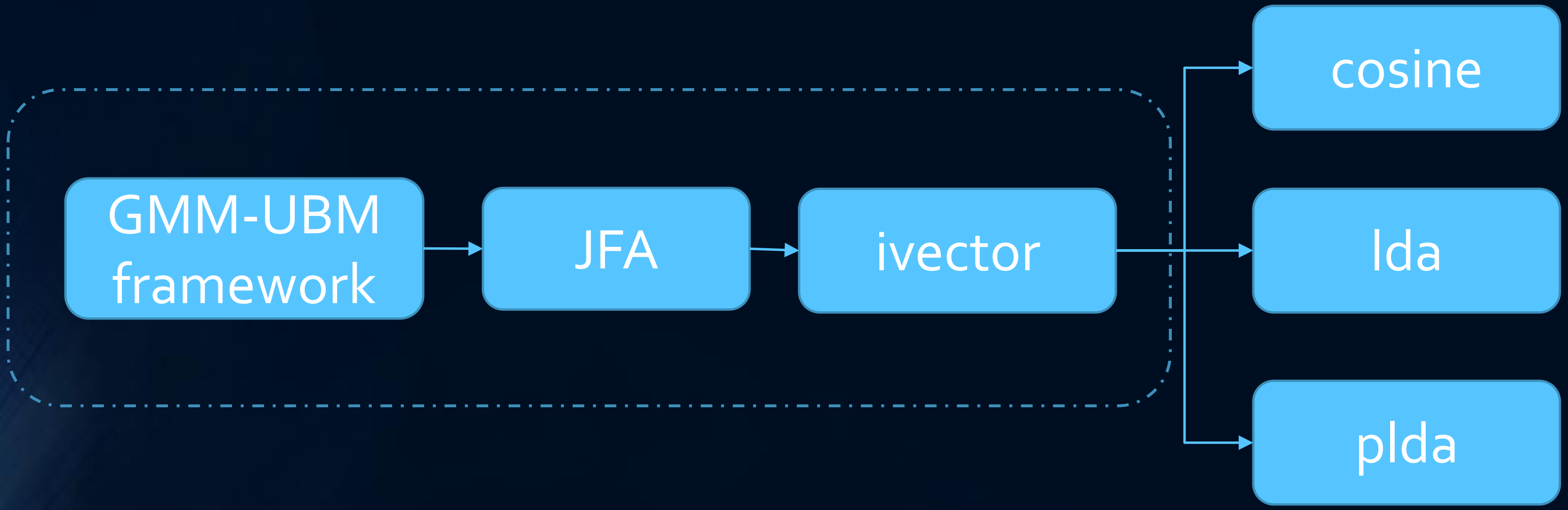

fig5 ivector methodology for speaker verification

- Jonit factor analysis [P. Kenny, 2007]
  - JFA is a model of speaker and session variability in GMMs.

$$M = m + Vy + Ux + Dz$$

  - where m is a speaker- and session-independent supervector with CF dimension. (UBM)
  - M is a speaker- and channel- dependent supervector.

$$m = [\vdots]_{CF\times 1} \quad M = [\vdots]_{CF\times 1}$$

- $M = m + Vy + Ux + Dz$
- V and D define a speaker subspace, and U defines a session subspace。
- $V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_c \end{bmatrix}_{CF \times R} \quad U = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_c \end{bmatrix}_{CF \times L} \quad D = \begin{bmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_c \end{bmatrix}_{CF \times CF}$
- The vector y, z and x are assumed to be a random variable with a normally distribution $N(0, I)$.
- z is a normally distributed CF dimension random vector.

- i-vector [N. Dehak, 2011]
  - make no distinction between speaker effects and session effects in GMM supervector space.
  - define a total variability space, contains speaker and session variabilities simultaneously.

$$M = m + Tw$$

  - $M \sim N(m, TT^T)$
  - $w \sim N(0, I)$

- $M = m + Tw$

- $T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_c \end{bmatrix}_{CF \times R}$, $m = [\vdots]_{CF \times 1}$, M$= [\vdots]_{CF \times 1}$, $w = [\vdots]_{R \times 1}$

- T is a low rank $CF \times R$ subspace that contains the eigenvectors with the largest eigenvalues of total variability covariance matrix.

- $w \sim N(0, I)$

Training and testing procedure for ivector

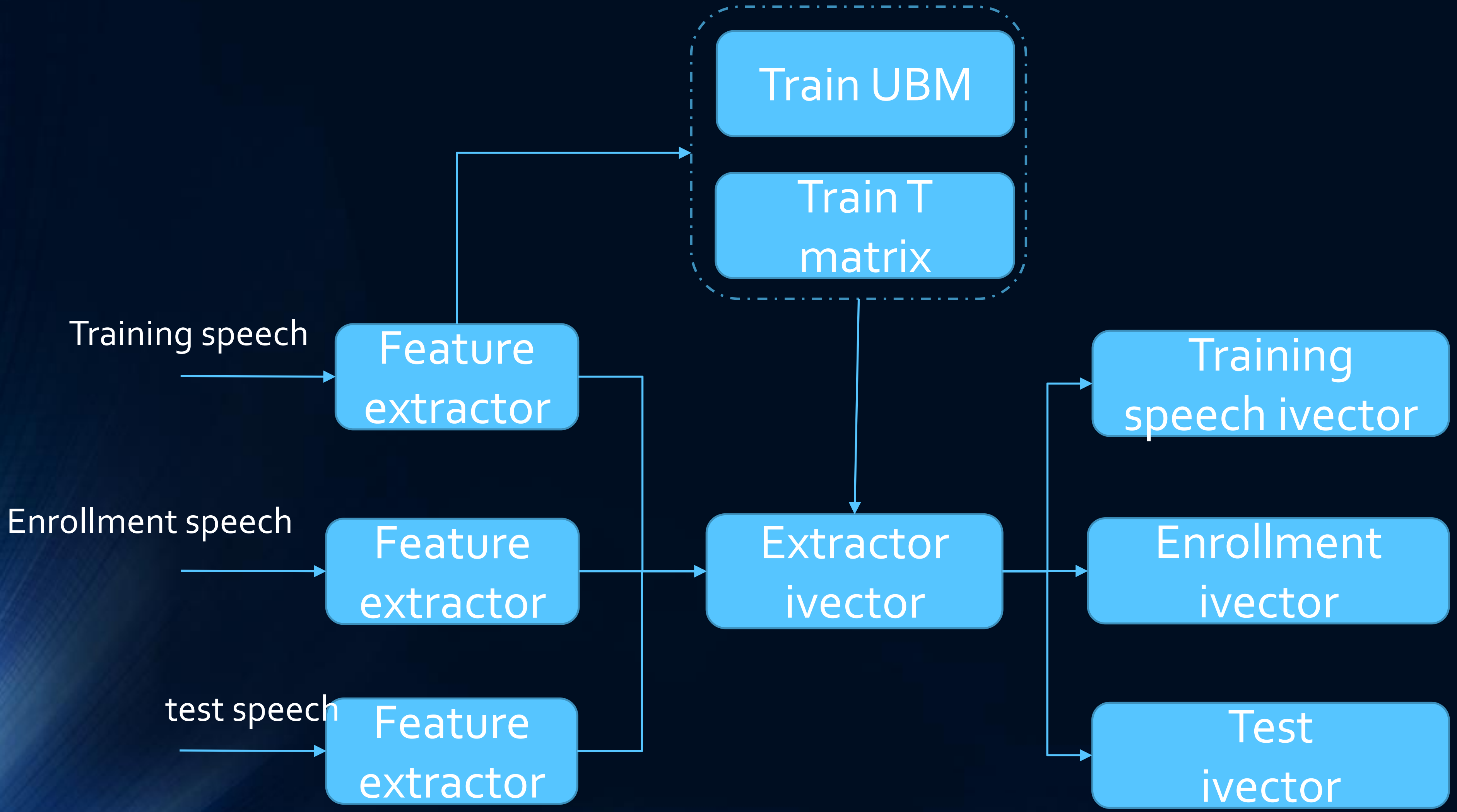

fig6 training and testing procedure for i-vector

- Object function
  - $M = m + Tw$
  - $M \sim N(m, TT^T)$
  - Suppose $x_i \sim N(M, \Sigma)$, $x_i = m + Tw + \varepsilon$
  - For Gaussian Mixture Model, $x_{i,c} = m_c + T_c w + \varepsilon_c$
  - $\mathcal{L} \sim p(x_i|\lambda)$
  - Define object function: $\mathcal{L} = \prod_c p\left(x_{i,c}|\lambda\right)$

- i-vector extraction [N. Dehak, 2011]
  - The Baum Welch statistics needed to estimate a given speech utterance:
  - $N_c = \sum_t P(c|x_t)$
  - $F'_c = \sum_t P(c|x_t)x_t$
  - $F_c = \sum_t P(c|x_t)(x_t - m_c)$

- i-vector extraction [N. Dehak, 2011]
  - The ivector of a speech segment X is computed as the mean of the posterior probability P(w|X).
  - $P(w|X) \sim N(\bar{w}, \Xi)$
  - $\bar{w} = \Xi T^T \Sigma^{-1} F$
  - $\Xi = (I + \sum_c T_c^T \Sigma_c^{-1} N_c T_c)^{-1}$

- T matrix training [N. Dehak, 2011]
  - T matrix can be trained by an EM procedure.
    - E steps computes the posterior probability P(w|X).
    - M step optimizes T by updating following formula:
    - $T_c = (\sum_u F_c(u)\overline{w}^T)(\sum_u N_c(u)(\overline{w}\overline{w}^T + \Xi)$

- T matrix training [N. Dehak, 2011]
  - $T_c = (\sum_u F_c(u)\overline{w}^T)(\sum_u N_c(u)(\overline{w}\overline{w}^T + \Xi)$
  - $T_c = \begin{bmatrix} \cdots \\ \cdots \\ \vdots \\ \cdots \end{bmatrix}_{F\times R} \quad T = \begin{bmatrix} T_1 \\ T_2 \\ \vdots \\ T_C \end{bmatrix}_{CF\times R}$

4 Intersession compensation and scoring method for ivector

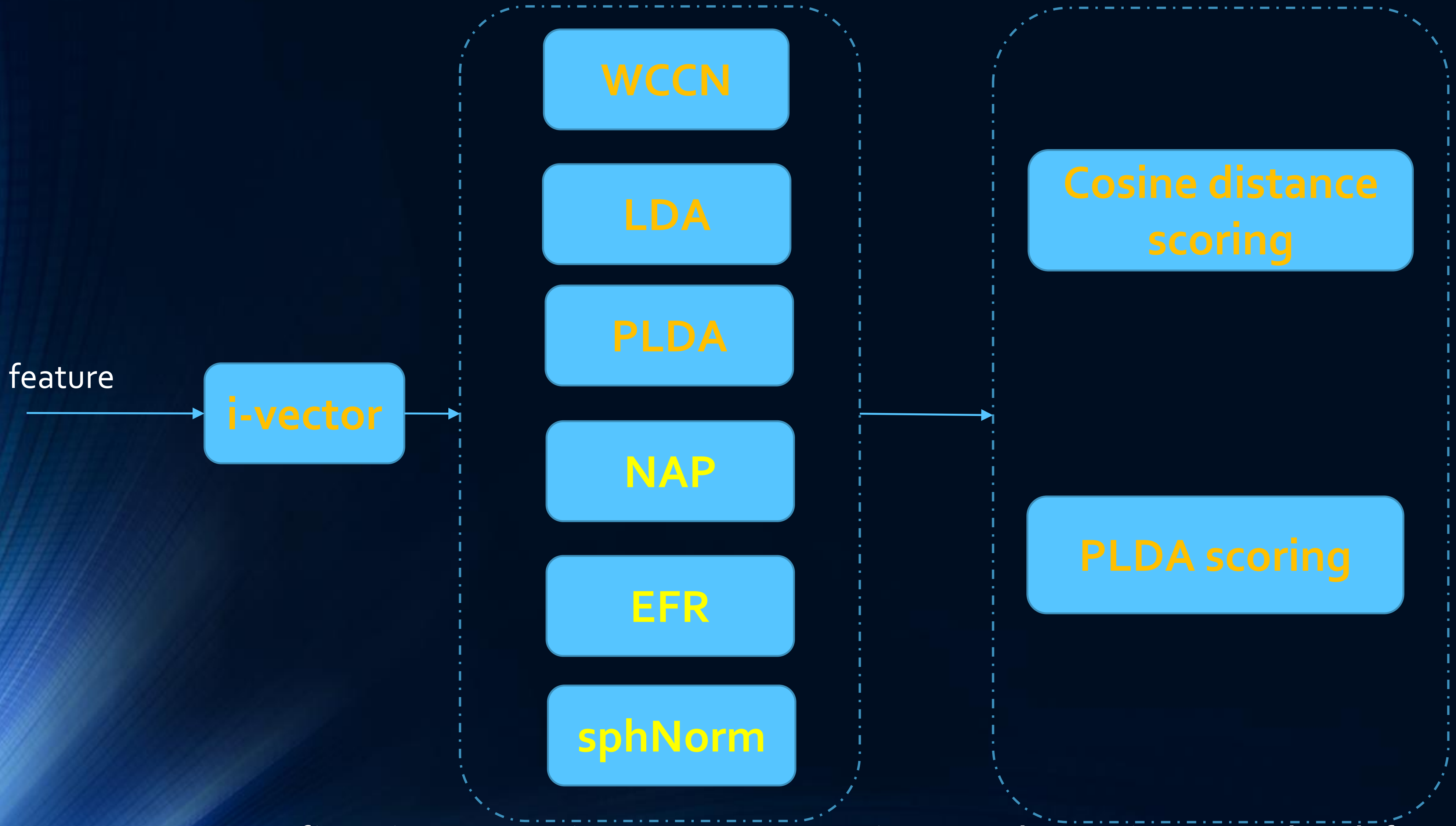

fig7 intersession compensation and scoring method for ivector

- Cosine distance [N. Dehak, 2009]
  - Using cosine kernel between the target speaker ivector and test speaker ivector.
  - $score(\omega_1, \omega_2) = \frac{\omega_1^t \omega_2}{\sqrt{\omega_1^t \omega_1}\sqrt{\omega_2^t \omega_2}}$

- WCCN [A. Hatch, 2006]
  - to minimize the classification error.
  - $k(\omega_1, \omega_2) = \omega_1^t R \omega_2$
  - $R = W^{-1}$ $\qquad W^{-1} = BB^T$
  - $W = \frac{1}{S}\sum_{s=1}^{S}\frac{1}{n_s}\sum_{i=1}^{n_s}(\omega_i^s - \overline{\omega_s})(\omega_i^s - \overline{\omega_s})^t$
  - $\omega' = B^t\omega$

- LDA [K. Fukunaga, 1990; N. Dehak, 2009]
  - to seek new orthogonal axes to better discriminate different classes.
  - a linear transformation that maximizes the between-class variation while minimizing the within-class variances.
  - fisher criterion is used for this purpose.

- LDA [K. Fukunaga, 1990; N. Dehak, 2009]
  - S_b is between-class covariance matrix, and S_w is the within-class covariance matrix. The solution v is generalized eigenvectors.

- $J(v) = \frac{v^t S_b v}{v^t S_w v}$ Reyleigh coefficient
- $S_b = \sum_{s=1}^{S} (w_s - \bar{w})(w_s - \bar{w})^t$
- $S_w = \sum_{s=1}^{S} \frac{1}{n_s} \sum_{i=1}^{n_s} (\omega_i^s - \overline{\omega_s})(\omega_i^s - \overline{\omega_s})^t$
- $S_b v = \lambda S_w v$
- $\omega' = A^t \omega$

- PLDA [S. J. D. Prince, 2007]
  - Technically, assuming a factor analysis (FA) model of the i-vectors of the form:

    $$\mathrm{w} = \mu + Fh + Gy + \varepsilon$$, in practice G always equals to zero

  - First computes the maximum likelihood estimate (MLE) of the factor loading matrix F (the Eigenvoice subspace).
  - Here, w is the i-vector, μ is the mean of training i-vectors, and $\mathrm{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a vector of latent factors. The full covariance residual noise term ε explains the variability not captured through the latent variables.

- PLDA [S. J. D. Prince, 2007]
  - Given a pair of ivectors D={w_1,w_2}, H_1 means two vectors from the same speaker and H_0 means two vectors from different speakers.[P. Kenny, 2010]
  - the verification score is computed for all possible model-test i-vector trials. The scores are computed as the log-likelihood ratio between the same (H_1) versus different (H_0) speaker models hypotheses:

$$llr = \ln \frac{p(\boldsymbol{w}_1, w|H_1)}{p(\boldsymbol{w}_1|H_0) \cdot p(\boldsymbol{w}_2|H_0)}$$

5 Toolkits and database

- Kaldi toolkits [D. Povey, 2011]
- database:

trials: NIST SRE08 female core test, contains 1997 females, 59343 trails.

lda/plda training data: fisher English database, contains 7196 females, 13827 sessions.

UBM training data: fisher English database, 6000 sessions female speech data.

- setup:

mfcc features, extracting with 20ms hamming window, every 10ms, 19 mel-frequency cepstral coefficient together with log energy were used. Delta and delta-delta coefficient were then calculated to produce 60-dimensional feature vector.

2048 Gaussian Mixtures, gender-dependent.

400-dimensional ivector.

150-dimensional lda/plda.

- SRE 8 results with kaldi: core test, female

EER(%)	1	2	3	4	5	6	7	8
cosine	28.77	4.78	28.60	21.32	20.43	11.36	7.35	7.63
LDA	24.10	1.79	24.18	14.56	14.42	10.25	6.46	6.58
PLDA	20.09	2.09	20.43	17.87	13.34	8.37	4.44	4.74

condition :
1 All trials involving only interview speech in training and test
2 All trials involving interview speech from the same microphone type in training and test
3 All trials involving interview speech from different microphones types in training and test
4 All trials involving interview training speech and telephone test speech
5 All trials involving telephone training speech and noninterview microphone test speech
6 All trials involving only telephone speech in training and test
7 All trials involving only English language telephone speech in training and test
8 All trials involving only English language telephone speech spoken

6 Some of my previous work

- Sequential Model adaptation for Speaker Verification
- Block-wise training for ivectors
- Phone-based alignment for channel robust speaker verification ……
- Mlp classification for ivector ……
- ……

References

[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] F. Bimbot, J. F. Bonastre, C. Fredouille, et al. A tutorial on text-independent speaker verification[J]. EURASIP journal on applied signal processing, 2004, 2004: 430-451.

[3] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang. A vector quantization approach to speaker recognition. International Conference on Acoustics, Speech and Signal Processing. 1985, 387–390.

[4] A. Higgins, L. Bhaler, and J. Porter. Voice identification using nearest neighbor distance measure. International Conference on Acoustics, Speech and Signal Processing. 1993, 375–378.

[5] Y. Bennani and P. Gallinari. On the use of tdnn-extracted features information in talker identication. International Conference on Acoustics, Speech and Signal Processing. 1991, 385–388.

[6] K. R. Farrell, R. J. Mammone, and K.T. Assaleh. Speaker recognition using neural networks and conventional classiers. IEEE Transactions on Speech and Audio Processing. 1994, 2:194–205.

[7] 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.

[8] 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.

[9] 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.

[10] 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.
[11] 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.
[12] The NIST Year 2008 Speaker Recognition Evaluation Plan, http://www.nist.gov/speech/tests/spk/2008/sre-08_evalplan-v9.pdf.
[13] 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
[14] 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.
[15] K. Fukunaga, Introduction to Statistical Pattern Recognition. 2nd ed. New York: Academic Press, 1990, ch. 10.

THANK YOU!