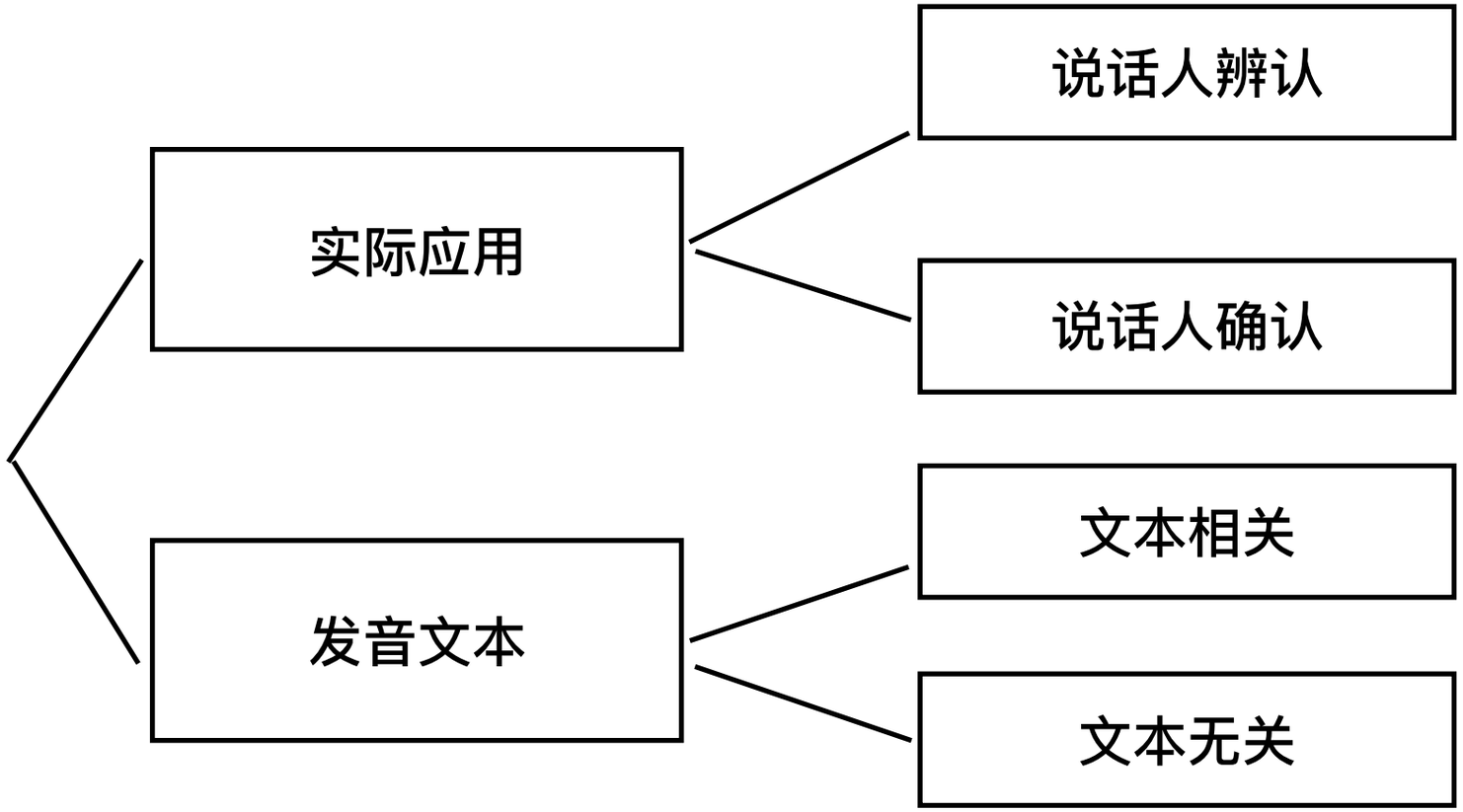
说话人识别学习报告

王雪仪 CUMTB

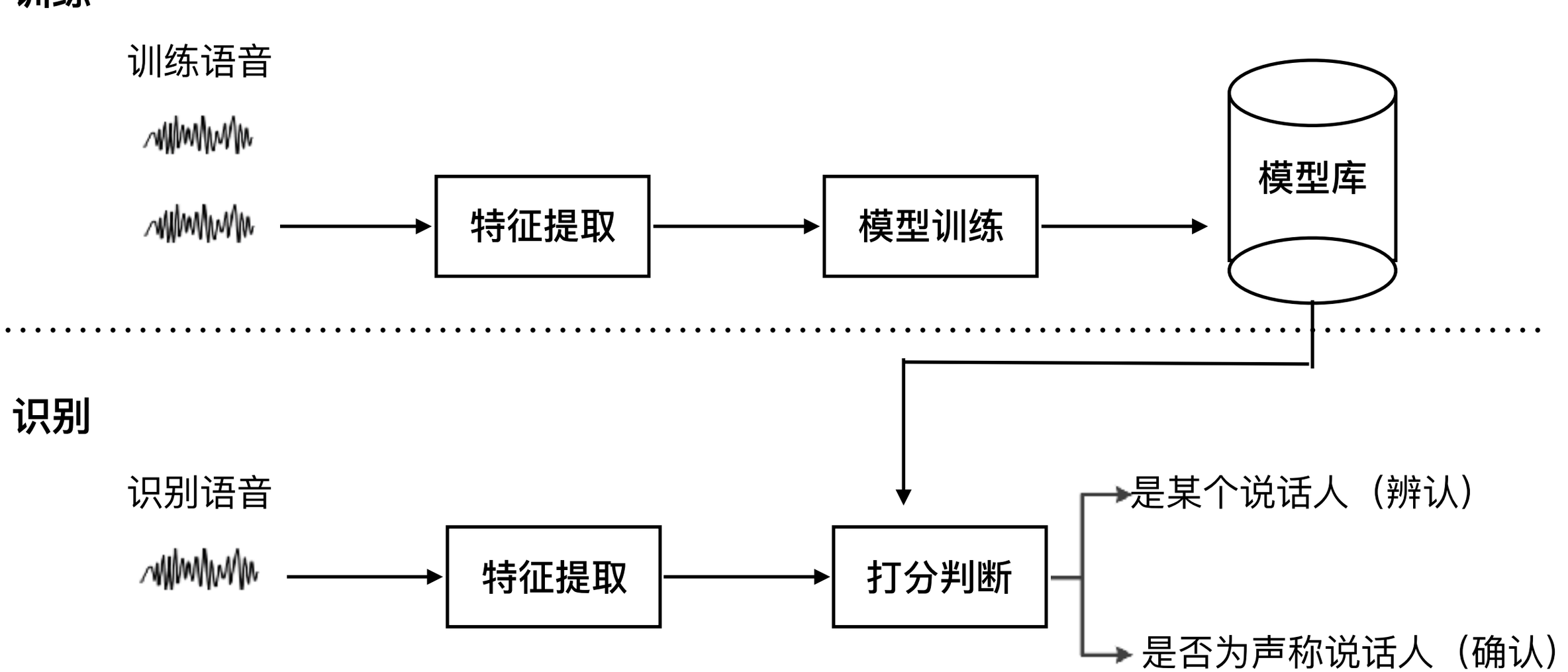
目录

说话人识别

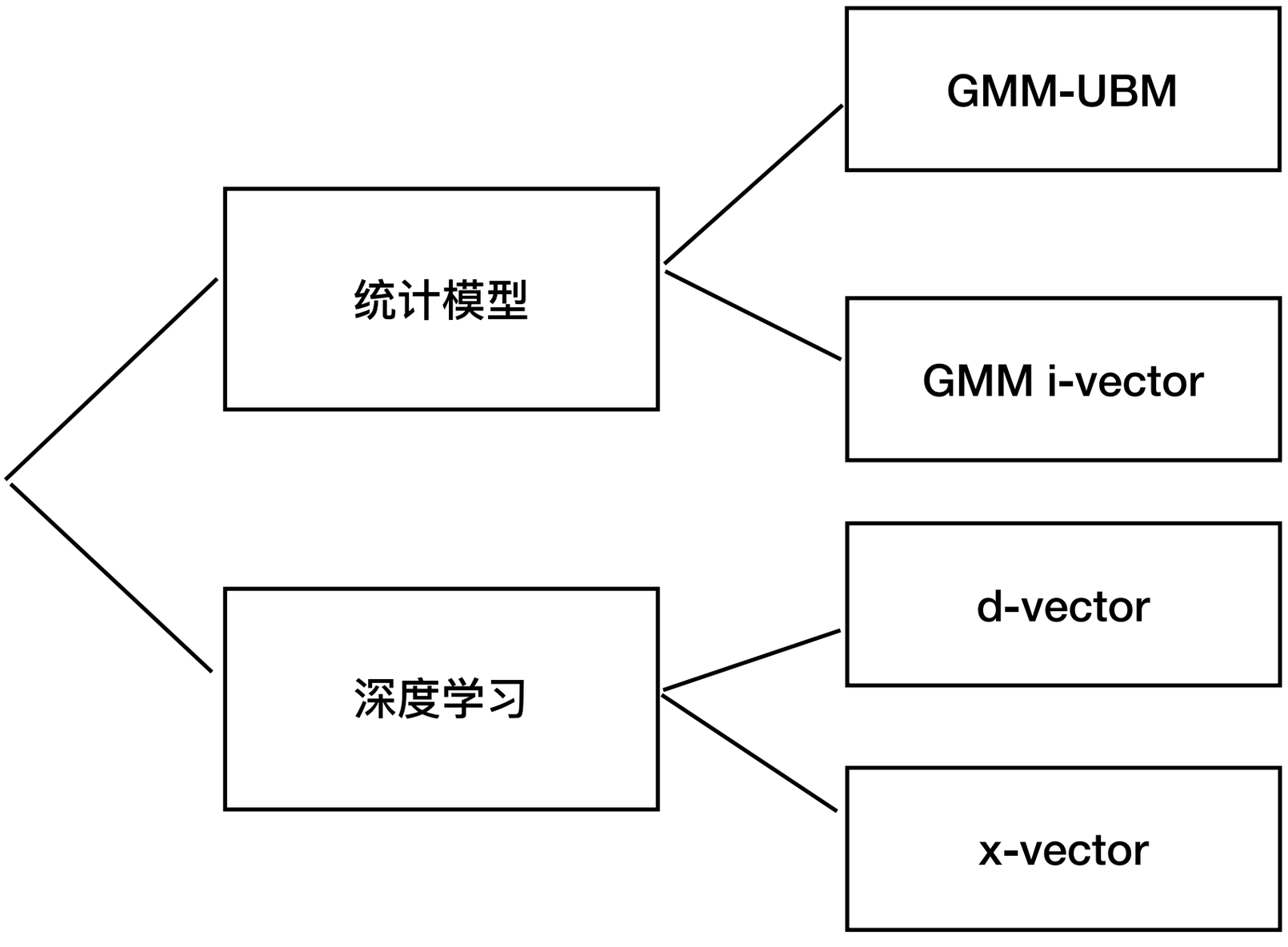
- 定义：根据说话人语音信号确认说话人身份
- 分类



说话人识别系统结构

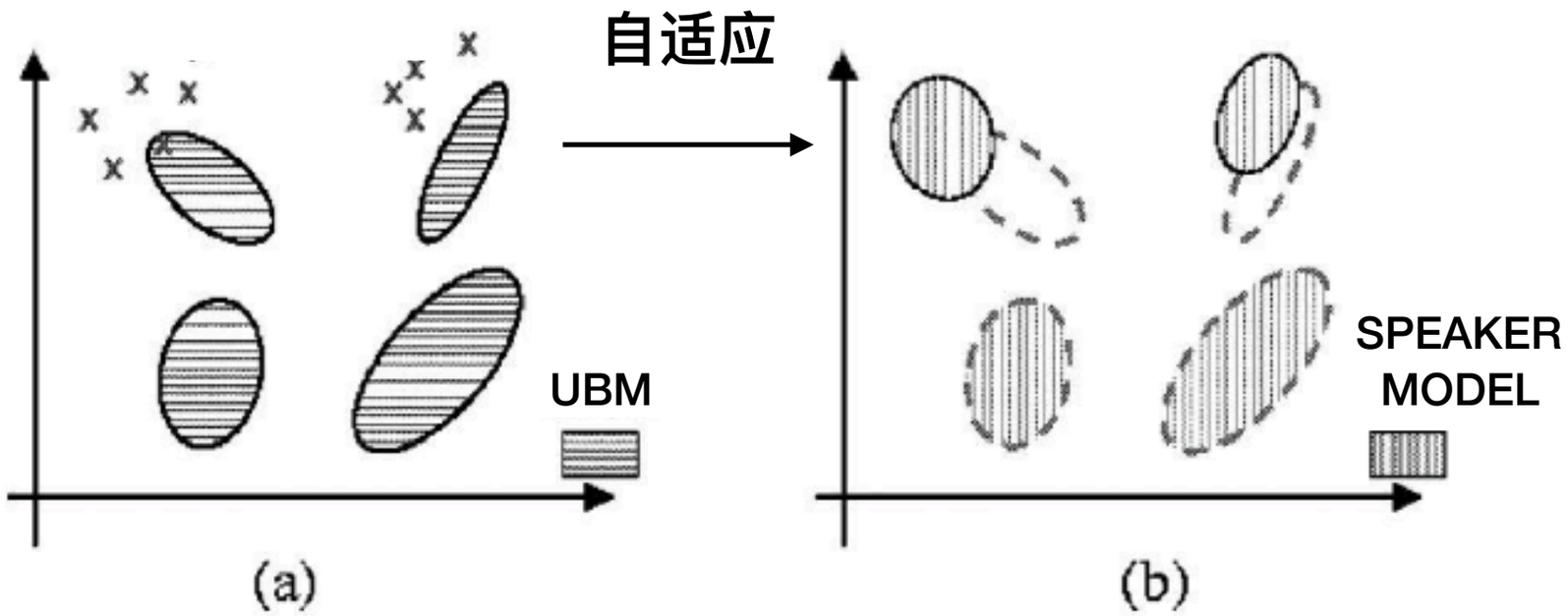


方法



统计模型

GMM-UBM



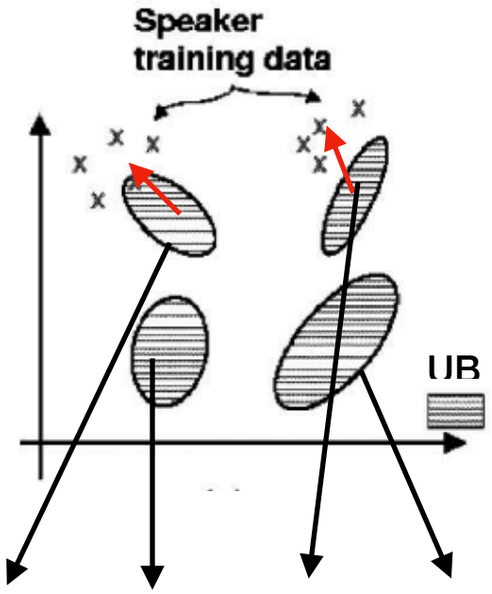
说话人建模：

- **使用大量的其他不同说话人语音数据训练出UBM**
- **基于最大后验估计算法，用说话人语音数据自适应得到说话人的GMM**

i-vector的提出

GMM-UBM: $M = \mathrm{m} + \mathrm{s}$

缺点：独立同分布假设过强，分量含有大量冗杂；既含有说话人相关信息，又包含了大量说话人无关信息

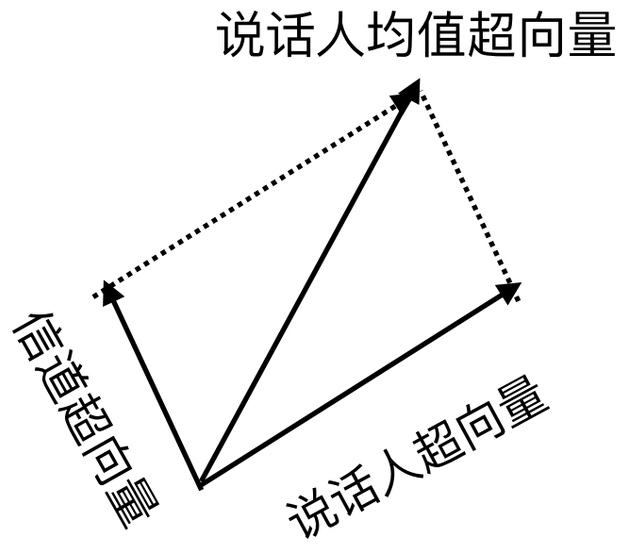


说话人均值超向量：

JFA: $M = \mathrm{m} + \mathrm{V}y + \mathrm{U}x + Dz$

说话人因子　信道因子　残差因子

缺点：很难将说话人空间和信道空间划分出来



i-vector

均值超矢量：

i-vector：

400～600维

公式：$M = m + Tw$　　T：全局差异空间矩阵

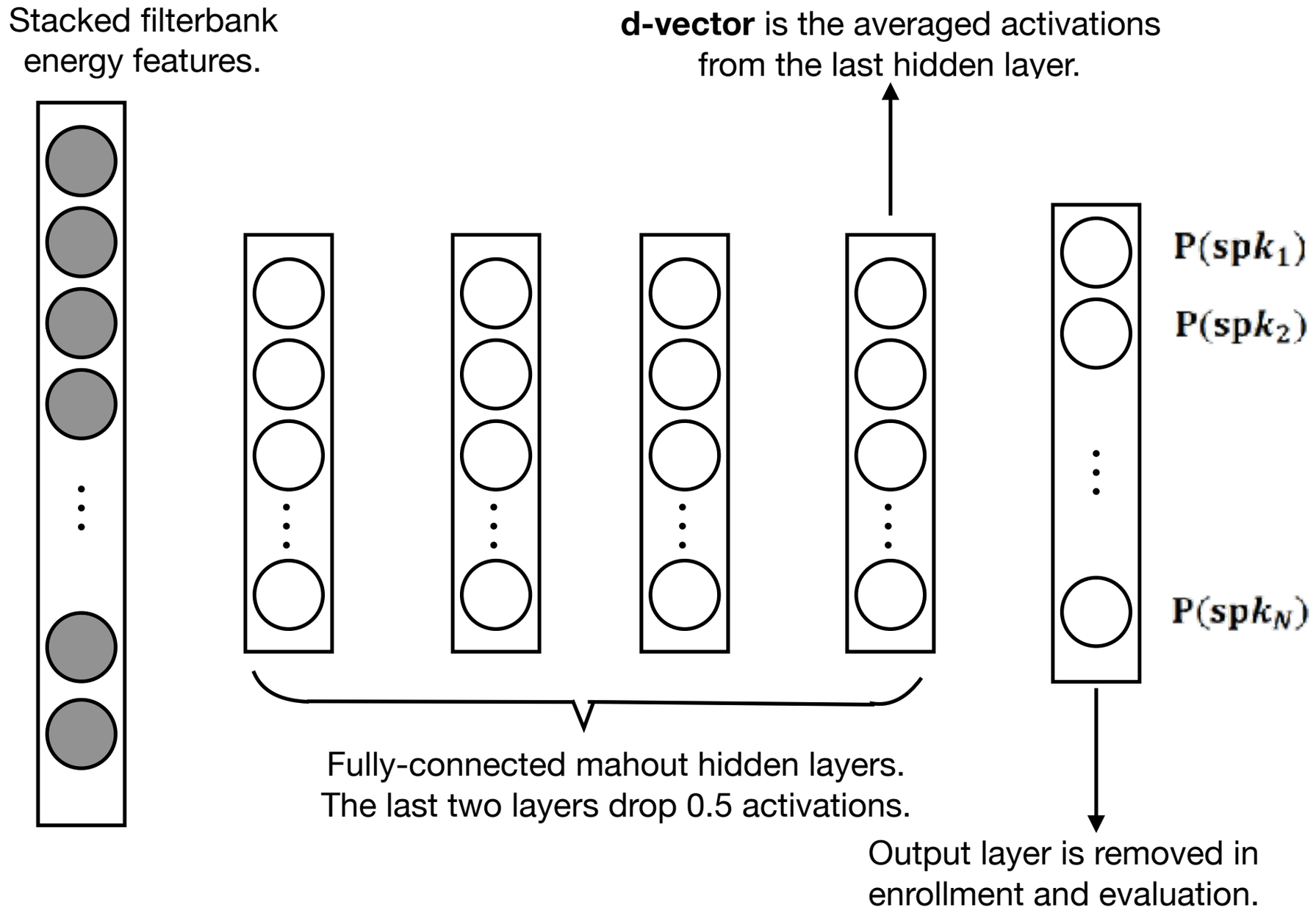
思想：基于因子分析，将均值超向量映射到一个低维的子空间，该子空间同时描述说话人信息和信道信息

深度学习

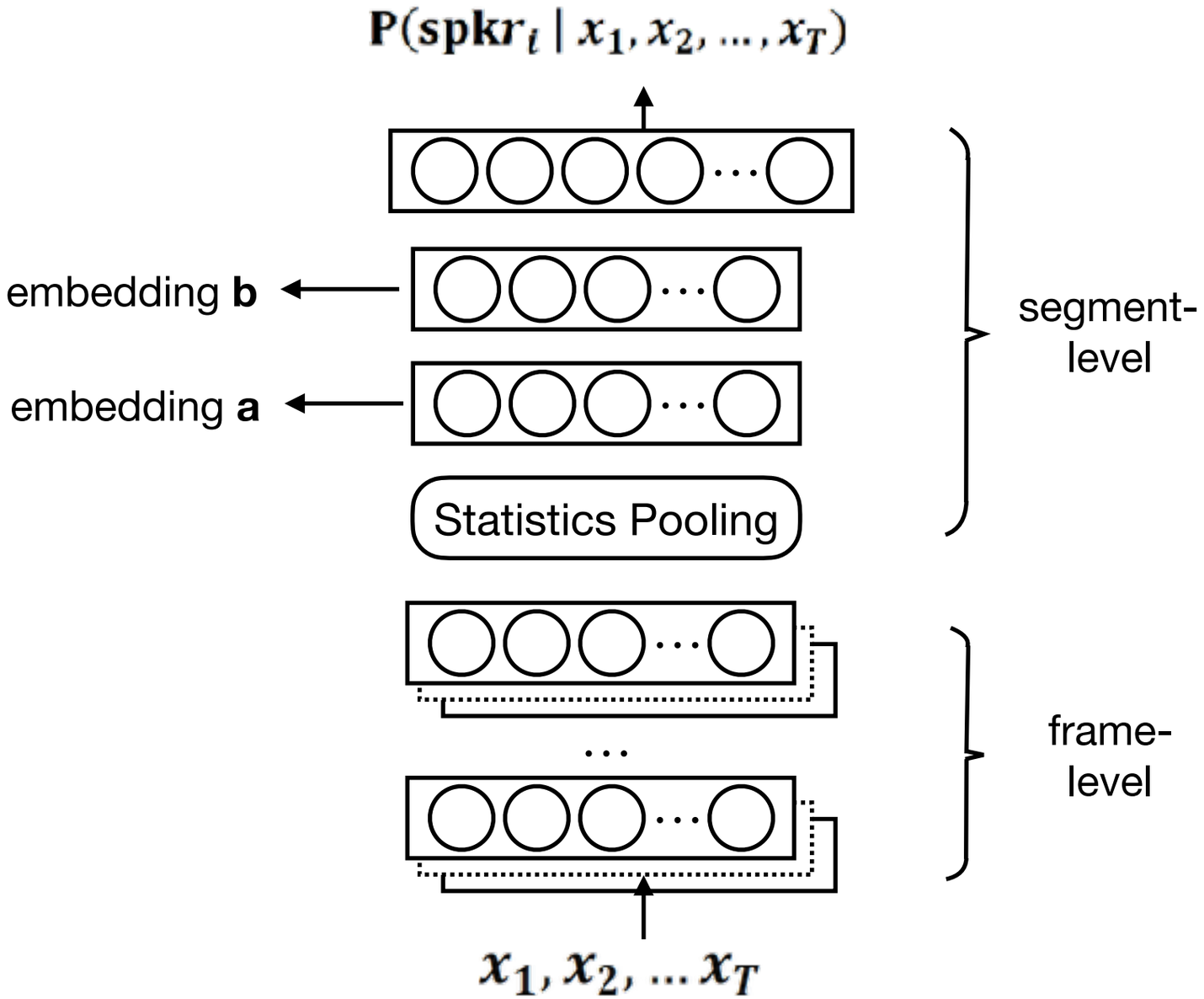
深度学习的理由

- **统计模型的缺点**
  - 原始特征（MFCC）受非说话人因素影响显著，变动性强
  - 线性高斯模型先验假设过强
- **深度学习的优点**
  - 无需人为设计特征，任务导向
  - 函数表达能力强
  - 多层结构，特征学习能力强

d-vector


Stacked filterbank energy features.
d-vector is the averaged activations from the last hidden layer.
P(spk1)
P(spk2)
P(spkN)
Fully-connected mahout hidden layers.
The last two layers drop 0.5 activations.
Output layer is removed in enrollment and evaluation.

x-vector


$\mathbf{P}(\mathbf{spkr}_i \mid x_1, x_2, \ldots, x_T)$
embedding **b**
embedding **a**
Statistics Pooling
segment-level
frame-level
$x_1, x_2, \ldots x_T$

Kaldi实验

i-vector

数据准备

UBM训练

i-vector extractor训练

数据增强

提取i-vector

打分模型训练

打分

x-vector

数据准备

数据增强

x-vector特征准备

网络训练

打分模型训练

打分

数据集

数据集：thchs30

data set	utterance
eval	2495
dev	893
train	10000

utt	train set
before augmenttation	10000
after augmentation	50000

实验结果

EER%	i-vector
Cosine	9.54
LDA	7.12
PLDA	6.27

EER%	x-vector
Cosine	12.09
LDA	10.62
PLDA	7.08

现象与分析

- **实验现象**
  - x-vector和i-vector的表现并不是很好
- **原因分析**
  - x-vector：训练时未收敛，而是从中选择了一个模型作为最终测试的模型
  - i-vector：由于电脑配置问题，分量数选择得比较小
  - 使用的数据集比较小，换用更大的数据集效果可能会更好一些

参考文献

[1] J. P. Campbell, “Speaker recognition: A tutorial,” Proceedings of the IEEE, vol. 85, no. 9, pp. 1437–1462, 1997.

[2] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust dnn em- beddings for speaker recognition,” in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

[3] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[4] D. Snyder, G Chen, and D. Povey, “MU- SAN: A Music, Speech, and Noise Corpus,” 2015, arXiv:1510.08484v1.

[5] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouel- let, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[6] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudan- pur, “Deep neural network embeddings for text-independent speaker verification,” *Proc. Interspeech*, pp. 999–1003, 2017.

[7] D.Snyder,D.Garcia-Romero,andD.Povey,“Timedelaydeep neural network-based universal background models for speaker recognition,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 92– 97.

参考文献

[8] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Proc*. *Odyssey*, 2014.

[9]V. Ehsan, L. Xin, M. Erik, L. M. Ignacio, and G.-D. Javier, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, vol. 28, no. 4, 2014, pp. 357–366.

[10] P.Kenny, G.Boulianne, P.Ouellet, and P.Dumouchel, "JointFactor Analysis versus Eigenchannels in Speaker Recognition," *IEEE Trans- action on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1435–1447, May 2007.

[11] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Speaker and Session Variability in GMM-Based Speaker Verification," *IEEE Transaction on Audio Speech and Language Processing*, vol. 15, no. 4, pp. 1448–1460, May 2007.

[12] M. Senoussaoui, P. Kenny, N. Dehak, and P. Dumouchel, "An I-Vector Extractor Suitable for Speaker Recognition with Both Microphone and Telephone Speech," in *IEEE-Odyssey*, Brno, Czech Republic, 2010.

[13] S.J.D. Prince, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV-11th*. IEEE, 2007, pp. 1–8.

[14] D. A. Reynolds, T. F. Quatieri, and Dunn. R. B., "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19 – 41, 2000.

Thanks