MMCE 算法在 FAGMM 中的应用

吴延渠，曾以成，蒋阳波
(湘潭大学光电工程系，湖南湘潭 411105)

摘要：提高说话人模型的识别性能一直是语音识别领域的一个重要课题。因子分析高斯混合模型(FAGMM)是因子分析方法与高斯混合模型(GMM)结合而成的多维概率统计模型，能更好地表征语音特征矢量的相关性，然而模型参数过多导致不能实现很好的分类。把改进的最小分类错误(MMCE)算法应用于该模型，形成一种新的 FAGMM+MMCE 模型，能解决前述问题，而且克服了传统的最小分类错误(MCE)算法在系统训练时不灵活、训练速度慢的缺点。实验结果表明，在 30 个说话人的识别应用中，本模型的识别率随着高斯混合数的增加而提高，较传统的 MCE 算法，识别率平均提高了 3%，训练时间也平均节省了 20%，说明该方法是有效的。

关键词：因子分析高斯混合模型(FAGMM)；改进的最小分类错误(MMCE)算法；FAGMM+MMCE 模型
中图分类号：TN912.34 **文献标识码**：A **文章编号**：1000-3630(2010)-01-0083-04
DOI 编码：10.3969/j.issn1000-3630.2010.01.019

Application of MMCE algorithm to FAGMM

WU Yan-qu, ZENG Yi-cheng，JIANG Yang-bo
(*Department of Photoelectric Engineering, Xiangtan University, Xiangtan* 411105, *Hu'nan, China*)

Abstract: To improve performances of speaker models is a significant research subject in the field of speech recognition. The factor analyzed Gaussian mixture model(FAGMM) is a multi-dimensional probability statistical model through the combination of factor analysis and GMM, and can reflect the intra-frame correlation of feature vectors well. However, it has too many model parameters to classify. In this paper, a modified minimum classification error (MMCE) algorithm is applied to the model, which forms a new FAGMM+MMCE model. The new model not only realizes the optimized classification of FAGMM model parameters, but also has the better flexibility and the faster training speed over the traditional MCE algorithm. The experimental results show that the identification rate of the new model continuously increases with Gaussian mixture number, and compared with the conventional MCE, it increases by an average of 3%. Besides, the training time also reduces by an average of 20% with a 30 speaker population. Such proves the new model to be effective.

Key words: factor analyzed Gaussian mixture model(FAGMM); modified minimum classification error (MMCE) algorithm; FAGMM+MMCE model

1 引 言

因子分析高斯混合模型(FAGMM)是因子分析方法[1-3]与高斯混合模型(GMM)[4]相结合的一种有效模型[5]，它能够较好地描述语音特征矢量的帧内相关问题，而且在一定程度上降低了计算复杂度，提高了收敛速度，较好地模拟了携带说话人特征的语音发音过程。因此，其在说话人识别领域得到应用，且达到了较好的识别效果[6]。但是，FAGMM 模型的缺点是模型参数较多，训练时占用存储空间大，并且传统的期望最大化(EM)训练算法又没有考虑各类模型参数之间的相似程度，忽略了参数估计的区分性问题，因而不能实现很好的分类。

最小分类错误(MCE)算法是一种实现最佳分类效果的学习方法[7]，它克服了 EM 算法只在类内判别且易陷入局部最优的不足，较好地实现了系统模型训练中的参数最优分类问题[8]，收敛速度快且算法简捷，成功应用于说话人识别领域[9]；然而对于一个系统，MCE 算法需要多次对比判别，当说话人数变化时系统还要重新训练模型参数，使系统训练很不灵活。通过改进 MCE 算法，得到修正的最小分类错误(MMCE)算法[10]，它在系统训练时利用简捷的排序来代替多次损失函数的判别，而且说话人增减不用重新训练原有的模型参数，只需把新增说话人的参数集加入有序序列即可，因此可以解决传统的 MCE 算法的不灵活性问题，训练时间也会相应缩短。

收稿日期：2009-01-09；**修回日期**：2009-03-26
基金项目：湖南省自然科学基金(08JJ5031)
作者简介：吴延渠(1979-)，男，山东鄄城人，硕士研究生，研究方向为语音信号处理及说话人识别。
通讯作者：曾以成，E-mail: yichengz@xtu.edu.cn

本文把 MMCE 算法用于 FAGMM 模型，形成新的 FAGMM+MMCE 模型，试图综合因子分析方法和 MMCE 算法的优点，研究新模型的识别性能。将给出算法与模型的结合实现过程。通过实验室环境下的自制语音库及说话人识别实验，考察 MMCE 算法在 FAGMM 模型中不同高斯混合数下的识别情况，对比 MCE 与 MMCE 训练算法在系统中的识别率和训练速度。

2 因子分析高斯混合模型

FAGMM 是将因子分析方法与 GMM 结合的一种多元系统模型，它模拟了携带说话人特征的语音发音过程。如图 1 所示。

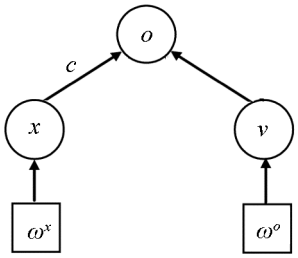

图 1 因子分析高斯混合模型

Fig.1 The factor analyzed Gaussian mixture model

图 1 中 k 维因子矢量 $\boldsymbol{x}$ 由 GMM 产生，因子矢量 $\boldsymbol{x}$ 为决定说话人身份的各个因素的集合构成的矢量(该矢量是隐藏的)，对因子矢量 $\boldsymbol{x}$ 进行线性变换，再叠加上观测噪声 $\boldsymbol{v}$，从而产生了 p 维特征矢量 $\boldsymbol{o}$，其中，观测噪声 $\boldsymbol{v}$ 通过高斯混合分布模拟在发音过程中产生的噪声。FAGMM 模型可以表示为：

$$\boldsymbol{x}\sim\boldsymbol{\lambda}_{GMM},\ \boldsymbol{\lambda}_{GMM}=\{\boldsymbol{p}_n,\boldsymbol{\mu}_n^x,\boldsymbol{\Sigma}_n^x\} \tag{1}$$

$$\boldsymbol{o}=\boldsymbol{C}\boldsymbol{x}+\boldsymbol{v},\ \boldsymbol{v}\sim\sum_{m=1}^{M}\boldsymbol{c}_m\boldsymbol{N}(\boldsymbol{\mu}_m^o,\boldsymbol{\Sigma}_m^o) \tag{2}$$

这里式(1)所描述的数学空间称为因子空间，式(2)所描述的数学空间称为观测空间，式(1)、(2)中 $\boldsymbol{p}_n$、$\boldsymbol{\mu}_n^x$、$\boldsymbol{\Sigma}_n^x$ (n=1,2,⋯,N)分别表示因子混合元 $\omega^x=n$ 时(因子空间中产生因子矢量 $\boldsymbol{x}$ 的第 n 个高斯混合元)的权重，均值矢量和协方差矩阵，N 为因子空间的高斯混合数。$\boldsymbol{c}_m$、$\boldsymbol{\mu}_m^o$、$\boldsymbol{\Sigma}_m^o$ (m=1,2,⋯,M)分别表示观测混合元 $\omega^o=m$ 时(观测空间中产生观测噪声 $\boldsymbol{v}$ 的第 m 个高斯混合元)的权重、均值矢量和协方差矩阵，M 为观测空间的高斯混合数。因此，就形成了完整的 FAGMM 模型，用一组参数集合来表示，即：$\boldsymbol{\lambda}_{FAGMM}=\{\boldsymbol{p}_n,\boldsymbol{\mu}_n^x,\boldsymbol{\Sigma}_n^x,\boldsymbol{C},\boldsymbol{c}_m,\boldsymbol{\mu}_m^o,\boldsymbol{\Sigma}_m^o\}$。

由式(1)、(2)可知，给定 $\omega^x=n$ 时，产生因子矢量 $\boldsymbol{x}$ 的概率为：

$$p(\boldsymbol{x}|n)=N(\boldsymbol{\mu}_n^x,\Sigma_n^x) \tag{3}$$

假设因子矢量 $\boldsymbol{x}$ 已知，给定 $\omega^o=n$ 时，产生特征矢量 $\boldsymbol{o}$ 的条件概率为：

$$p(\boldsymbol{o}|\boldsymbol{x},m)=N(\boldsymbol{C}\boldsymbol{x}+\boldsymbol{\mu}_m^o,\boldsymbol{\Sigma}_m^o) \tag{4}$$

因此，通过对因子矢量 $\boldsymbol{x}$ 积分可以得到当 $\omega^x=n$、$\omega^o=n$ 时，产生特征矢量 $\boldsymbol{o}$ 的概率为：

$$p(\boldsymbol{o}|m,n)=N(\boldsymbol{C}\boldsymbol{\mu}_n^x+\boldsymbol{\mu}_m^o,\boldsymbol{C}\boldsymbol{\Sigma}_n^x\boldsymbol{C}'+\boldsymbol{\Sigma}_m^o) \tag{5}$$

在训练和识别阶段，均要计算式(5)中的 $p\times p$ 阶协方差矩阵 $\boldsymbol{C}\boldsymbol{\Sigma}_n^x\boldsymbol{C}'+\boldsymbol{\Sigma}_m^o$ 的逆矩阵和对应的行列式，当 p 比较大时，这项计算需要大量的内存空间和计算时间，会严重降低系统的实时性能，因此，本文采用下列矩阵和行列式等式进行计算

$$(\boldsymbol{C}\boldsymbol{\Sigma}_n^x\boldsymbol{C}'+\boldsymbol{\Sigma}_m^o)^{-1}=\boldsymbol{\Sigma}_m^{o-1}-\boldsymbol{\Sigma}_m^{o-1}\boldsymbol{C}(\boldsymbol{C}'\boldsymbol{\Sigma}_m^{o-1}\boldsymbol{C}+\boldsymbol{\Sigma}_n^{x-1})^{-1}\boldsymbol{C}'\boldsymbol{\Sigma}_m^{o-1} \tag{6}$$

$$|\boldsymbol{C}\boldsymbol{\Sigma}_n^x\boldsymbol{C}'+\boldsymbol{\Sigma}_m^o|=|\boldsymbol{\Sigma}_m^o||\boldsymbol{\Sigma}_n^x||(\boldsymbol{C}'\boldsymbol{\Sigma}_m^{o-1}\boldsymbol{C}+\boldsymbol{\Sigma}_n^{x-1})| \tag{7}$$

由于 $\boldsymbol{\Sigma}_m^o$ 和 $\boldsymbol{\Sigma}_n^x$ 是对角矩阵，求其逆矩阵速度很快，矩阵 $\boldsymbol{C}'\boldsymbol{\Sigma}_m^{o-1}\boldsymbol{C}+\boldsymbol{\Sigma}_n^{x-1}$ 则为 $k\times k$ 阶矩阵。因此，采用上述两个等式后，当 $k<p$ 时，可以利用一组低维的隐藏的因子矢量就能很方便地描述高维观测矢量的协方差矩阵结构，但是 FAGMM 的模型参数较多，占用存储空间较大，其训练时所采用的 EM 算法[11]不能较好地实现模型参数最优分类。

3 MCE 算法

首先，对于有 K 个说话人的模型，输入特征矢量序列为 $\boldsymbol{x}$，定义模型产生的区分性函数为：

$$\theta_k(\boldsymbol{x};\lambda)=\ln p(\boldsymbol{x}|\lambda_k) \tag{8}$$

式(8)中条件分布 $p(\boldsymbol{x}|\lambda_k)$ 可以用一个 FAGMM 来表示，λ_k 为一个说话人的密度模型的参数。有了区分性函数后，对于给定的任意一个观察矢量序列 $\boldsymbol{x}$，可以通过式(9)所示的规则判断出该观察矢量序列是哪一个说话人说的。

$$\hat{k}=\arg\max_s\theta_k(\boldsymbol{x};\lambda) \tag{9}$$

然后，定义某个说话人 k 的错误分类距离为：

$$d_k(\boldsymbol{x};\lambda)=\frac{1}{K}\sum_{i=1}^{K}[-\theta_k(\boldsymbol{x}_i;\lambda)+\max_{y\neq k}\theta_y(\boldsymbol{x}_i;\lambda)] \tag{10}$$

错误分类距离是模型参数的连续函数，如果错误分类距离 $d_k(\boldsymbol{x};\lambda)>0$, 则意味错误的分类, 相反 $d_k(\boldsymbol{x};\lambda)\leq 0$, 则意味着正确的分类。

最后，定义系统的损失函数为：

$$l_k(\boldsymbol{x};\lambda)=l(d_k(\boldsymbol{x};\lambda))=\frac{1}{1+\exp(-\gamma\cdot d_k(\boldsymbol{x};\lambda))} \tag{11}$$

这是一个值域在 0 到 1 之间的 sigmoid 函数，其中 λ 表示所有模型参数的合集，大于 0 的常数 γ 表示 sigmoid 函数的中心斜率。

那么，系统的整体平均损失函数为：

$$L(\lambda)=E_x(l(\boldsymbol{x},\lambda))=\int l(\boldsymbol{x};\lambda)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}=\int\sum_{k=1}^{K}l_k(\boldsymbol{x};\lambda)p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \tag{12}$$

MCE 训练的目的是要将系统错误分类带来的损失降到最低，通常采用基于梯度下降的 GPD 算法来最小化分类错误，如式(13)所示。

$$\lambda(r+1)=\lambda(r)-\varepsilon_r\nabla L(\lambda)|_{\lambda=\lambda_r} \tag{13}$$

其中，$\lambda(r)$表示第 r 次迭代时的参数集，ε_r 是反映收敛速度的常数，$\nabla L(\lambda)|_{\lambda=\lambda\gamma}$ 代表 $L(\lambda)|_{\lambda=\lambda\gamma}$ 的梯度。

4　MCE 的改进及其在 FAGMM 中的实现

在用传统的 MCE 算法训练模型参数时，对于有 K 个说话人的模型系统，每一个类别的分类错误都需要计算 K–1 类的判别函数，随着 K 的增加，使得计算量大量增加。而且，当说话人数增减时，还要重新计算判别函数，于是导致系统重新训练，降低了说话人识别系统灵活性、增加了计算量。

于是，把改进的 MCE 算法用于训练 FAGMM 模型。首先，对于给定的观察矢量序列 $\boldsymbol{x}$，我们设计一个新的错误分类距离函数，可表示为

$$d_{kj}(\boldsymbol{x};\lambda)=-\ln\left(p\left(\boldsymbol{x};\lambda_j\right)\right)+\ln\left(p(\boldsymbol{x};\lambda_k)\right),\ k\neq j \tag{14}$$

其中 $p(\boldsymbol{x};\lambda_j)$和 $p(\boldsymbol{x};\lambda_k)$分别表示特征序列 $\boldsymbol{x}$ 是由模板 λ_j、λ_k 产生的概率。

其次，对于每一个说话人，定义一组错误分类距离函数，即给出 K–1 个 d_{kj}，按照 d_{kj} 的大小，把每一个说话人 k 定义为一个数值从小到大的 d_{kj} 序列 $\{d_{1j}, d_{2j}, \cdots, d_{Kj}\}(k\neq j)$，表示除了 k 以外的相似度。从而可以定义一个有序的参数集合 $S(\boldsymbol{x}, k, N)$，用该参数集合来求分类距离，其中 $N<k$–1 表示对于每一个说话人，用 N 个最接近说话人的模型来生成模型参数。于是，得到相应的损失函数为：

$$l_{kj}(\boldsymbol{x};\lambda)=l_{kj}\left(d_{kj}(\boldsymbol{x};\lambda)\right)=\frac{1}{1+\exp\left(d_{kj}(\boldsymbol{x};\lambda)\right)} \tag{15}$$

相应的的平均损失函数为：

$$l(\boldsymbol{x};\lambda)=\sum_k\sum_{j\in K(O<K<N)}l_{kj}\left(d_{kj}(\boldsymbol{x};\lambda)\right) \tag{16}$$

采用梯度下降法实现函数 $l(\boldsymbol{x};\ \lambda)$的最小化。对于第 j 个说话人，模型参数 λ_j 调整如下：

$$\frac{\partial l_{kj}(x;\lambda)}{\partial\lambda_j}=\frac{\partial l_{kj}}{\partial d_{kj}}\frac{\partial d_{kj}}{\partial\lambda_j} \tag{17}$$

$$\frac{\partial l_{kj}}{\partial d_{kj}}=l_{kj}\left(d_{kj}\right)\left(1-l_{kj}\left(d_{kj}\right)\right) \tag{18}$$

$$\frac{\partial d_{kj}}{\partial\lambda_j}=\begin{cases}-\dfrac{1}{p\left(\boldsymbol{x}|\lambda_{kj}\right)}\dfrac{\partial p\left(\boldsymbol{x}|\lambda_{kj}\right)}{\partial\lambda_j},\ (k=kj)\\ \dfrac{1}{p\left(\boldsymbol{x}|\lambda_{kj}\right)}\dfrac{\partial p\left(\boldsymbol{x}|\lambda_{kj}\right)}{\partial\lambda_j},\left[\begin{array}{c}k\neq kj,\\ \text{且}\ln\left(p\left(\boldsymbol{x}|\lambda_{kj}\right)\right)=\\ \max\limits_{k=1,\ldots,N,k\neq kj}\ln\left(p\left(\boldsymbol{x}|\lambda_{kj}\right)\right)\end{array}\right]\\ 0\qquad,(\text{其它})\end{cases} \tag{19}$$

5　实验结果

我们用一个与文本无关的说话人识别系统来测试 FAGMM+MMCE 模型的识别性能。实验基于实验室环境下 30 个人的语音库的闭集测试，每个说话人说出 50 段文本，每段时间为 5s。取每个人的前 30 段语音数据进行训练，其余的 20 段语音数据用来测试。语音采样率为 16kHz，经过 $1–0.95Z^{-1}$ 预加重后用汉明窗进行分帧，帧长为 32ms，帧移 16ms。所有特征矢量为 12 阶的基于 Mel 频率的倒谱参数(Mel Frequency Cepstral Coefficients, MFCC)及其一阶差分 MFCC。

图 2 给出了 FAGMM+MMCE 模型的说话人识别过程的流程。

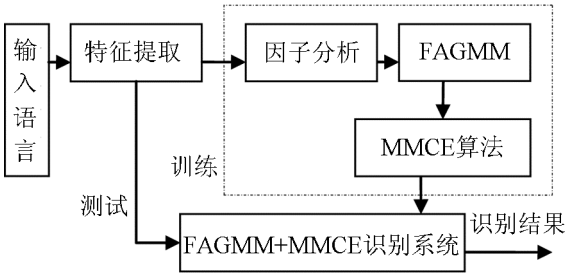

图 2　FAGMM+MMCE 的说话人识别系统

Fig.2　Speaker identification system based on FAGMM+MMCE

表 1 给出了不同高斯混合数 N 和因子混合数 K 的情况下，FAGMM+MMCE 模型的识别率情况，从表 1 可见，随着 N 和 K 的增加，识别率都是不断增加的，表明 FAGMM+MMCE 能在低的高斯混合数下实现单纯的 GMM 在高维的混合数下达到的识别性能，即因子分析方法具有降低维数的作用，而且从它的算法实现过程中可以知道，它的收敛速度较快，这也体现了 MMCE 算法的优点。

表 2 给出了该模型在 MCE 和 MMCE 训练时识别性能比较的情况，基本的 FAGMM 模型是采用对角协方差矩阵形式，由表 2 可见，随着混合数的增加，FAGMM+MCE 和 FAGMM+MMCE 的识别率

表 1 K 和 N 取不同值时的说话人识别率
Table 1 The identification rate for different values of K and N

	N=2	N=4	N=6	N=8	N=16	N=32	N=64	N=128
K=2	86.62%	93.33%	95.67%	96.33%	96.33%	96.87%	97.06%	97.14%
K=4	88.33%	94.56%	96.23%	97.12%	97.26%	97.73%	98.12%	98.33%
K=6	90.67%	95.56%	97.33%	98.46%	98.57%	98.76%	98.96%	99.23%
K=8	92.00%	96.06%	98.11%	98.78%	98.82%	98.93%	99.33%	99.56%

表 2 FAGMM+MCE 与 FAGMM+MMCE 的识别率比较
Table 2 The identification rate comparison between FAGMM+MCE and FAGMM+MMCE

高斯混合数 N	2	4	6	8	16	32	64	128
FAGMM+MCE (K=6)	84.52%	92.38%	95.62%	96.42%	96.77%	96.93%	97.32%	98.06%
FAGMM+MMCE (K=6)	90.67%	95.56%	97.33%	98.46%	98.57%	98.76%	98.96%	99.23%

都不断增加；在相同高斯混合数目下，采用 FAGMM+ MMCE 方法的识别率较 FAGMM+MCE 有较明显的提高，平均提高了 3%左右。

表 3 为 MCE 与 MMCE 训练时间的比较情况，在不同高斯混合数下，用 MMCE 来训练模型参数都比 MCE 节省训练时间，训练占用的 CPU 时间较 MCE 平均节省了约 20%，所以训练速度加快了。

表 3 MCE 与 MMCE 的训练时间比较
Table 3 The training time comparison between MCE and MMCE

高斯混合数 N	模型参数训练占用的 CPU 时间(s)	
	MCE 训练	MMCE 训练
2	10.34	8.47
4	20.53	15.62
6	29.48	21.29
8	34.41	27.33
16	40.08	30.06
32	199.85	145.24
64	383.93	267.92
128	724.21	512.78

6 结 论

本文把 MMCE 算法用于 FAGMM 模型，形成了一种说话人识别的新的 FAGMM+MMCE 模型。在新模型中，既取得了 FAGMM 模型中因子分析方法的优点，又引入 MMCE 算法实现了参数分类器的优化，较 MCE 算法提高了灵活性，降低了运算量，同时在一定程度上提高了识别率。对于新的 FAGMM+ MMCE 模型，随着混合因子数 K 和高斯混合数 N 的增加，识别率不断提高，尤其在 K=8、N=128 时，系统的识别率达到了 99.56%；与 MCE 算法相比，引入 MMCE 算法后系统的识别率得到比较明显地提高，而且 MMCE 算法的训练时间比 MCE 节省 20%，训练速度也相应加快。

新的模型明显提高了模型参数训练的灵活性，较好地解决了特征矢量的帧内相关问题，实现了参数的最优分类，降低了运算量和系统开销，识别率得到较大程度地提高，训练速度也相应加快。实验结果证明了 MMCE 算法应用于 FAGMM 模型的有效性和实用性。

参 考 文 献

[1] Rosti A-V I, Gales M J F. Factor analyzed hidden Markov models[A]. Proc. of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)[C]. 2002, **1**: 949-952.

[2] Yao K, Paliwal K K, Lee T W. Generative factor analyzed HMM for automatic speech recognition[J]. Speech Communication, 2005, **45**: 435-454.

[3] LEI Xiongguo, LI Ling, ZENG Yicheng. Text-independent speaker identification using factor analyzed hidden Markov model[A]. NCMMSC'05[C]. 2005, **24**: 222-225.

[4] Reynolds D A, Quatieri T, Dunn R. Speaker verification using adapted Gaussian mixture models[J]. Digital Signal Processing, 2000, (10): 19-41.

[5] 岳博, 焦李成. 混合因子分析的重新抽样方法[J]. 电子学报, 2002, **12**(12): 1873-1875.
YUE Bo, JIAO Licheng. The resampling method for mixtures of factor analyzers[J]. Acta Electronica Sinica, 2002, **12**(12): 1873-1875.

[6] Lawrence Saul, Mazin Rahim. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 2000, **8**(2): 115-125.

[7] 王成儒, 王金甲. 基于 MCE 训练算法的说话人辨认系统[J]. 计算机工程, 2003, **29**(13): 105-107.
WANG Chengru, WANG Jinjia. Efficient training of MCE for speaker identification[J]. Computer Engineering, 2003, **29**(13): 105-107.

[8] Saul L, Rahim M. Maximum likelihood and minimum classification error factor analysis for automatic speech recognition[J]. IEEE Transactions on Speech and Audio Processing, 1999, **8**(2): 115-125.

[9] 李晓宇. 利用 MCE 算法提高说话人识别性能[J]. 电路与系统学报, 2000, **5**(3): 46-49.
LI Xiaoyu. Using MCE algorithm to improve the performance of speaker recognition[J]. Journal of Circuits and Systems, 2000, **5**(3): 46-49.

[10] 邱政权, 尹俊勋. 结合重叠子帧的 KLT 和 MMCE 的说话人辨认[J]. 声学技术, 2007, **26**(4): 660-663.
QIU Zhengquan, YIN Junxun. Combination of KLT and overlap sub-frame in MMCE speaker identification[J]. Technical Acoustics, 2007, **26**(4): 660-663.

[11] 成新民. 基于修正 EM 算法的说话人识别的研究[J]. 电声技术, 2004, (12): 51-53.
CHENG Xinmin. Research recognition based on modified EM algorithm[J]. Audio Engineering, 2004, (12): 51-53.