第 26 卷第 10 期 电子与信息学报 Vol.26No.10
2004 年 10 月 Journal of Electronics & Information Technology Oct. 2004

一种适于非特定人语音识别的并行隐马尔可夫模型[1]

陈雁翔*** 戴蓓蒨* 周 曦* 刘 鸣*

*(中国科学技术大学电子科学与技术系 合肥 230026)
**(合肥工业大学计算机与信息学院 合肥 230009)

摘 要: 为了适合非特定人语音识别，提出了一种由多条并行马尔可夫链组成的并行 HMM(Parallel Hidden Markov Model, PHMM)，从而融合了基于分类的语音识别中为各个类别建立的模板，提高了识别性能．各条链之间允许有交叉，使得融合的多模板之间存在状态共享，同时 PHMM 可以在训练过程中自动完成聚类，且测试语音的输出结果来自所有类别，无需聚类分析和类别判断，这些都减少了存储量和计算量．汉语非特定人孤立数字的识别实验表明，PHMM 较之传统 CHMM 使识别性能及噪声鲁棒性都得到了改善．

关键词: 非特定人语音识别，连续隐马尔可夫模型，并行马尔可夫链
中图分类号: TP391.42 **文献标识码**: A **文章编号**: 1009-5896(2004)10-1601-06

An Appropriate Parallel HMM for Speaker-Independent Speech Recognition

Chen Yan-xiang*** Dai Bei-qian* Zhou Xi* Liu Ming*

*(*Dept of Electron. Sci. & Tech., Univ. of Sci. & Tech. of China, Hefei* 230026, *China*)
**(*College of Computer Science & Information, Hefei Univ. of Tech., Hefei* 230009, *China*)

Abstract In this paper Parallel Hidden Markov Model (PHMM) made up of several parallel Markov chains is proposed to fit in with speaker-independent speech recognition. The performance is improved because of the fusion of different models from classification based speech recognition. By sharing states of fused models, making classification automatically during training and getting result from all classifications, the amount of storage and operation can be decreased. The experiment for speaker-independent recognition of mandarin isolated digit shows that the PHMM improves the recognition performance and noise robustness.

Key words Speaker-independent speech recognition, Continuous Hidden Markov Model(CHMM), Parallel Markov chain

1 引言

非特定人 (Speaker-Independent, SI) 语音识别系统不再只是针对一个用户，而是适于更多的人使用，因而相对于特定人 (Speaker-Dependent, SD) 语音识别系统具有更广泛的应用领域．由于隐马尔可夫模型 (Hidden Markov Model, HMM) 用概率统计理论描述了语音信号整体上的非平稳性和局部平稳性，也就是说用状态与语音的某个平稳段相对应，而各个平稳段之间以转移概率相联系，因此在现有的非特定人语音识别系统中，HMM 已成为最为流行的语音模型，并且多采用状态输出具有连续概率分布的连续隐马尔可夫模型 (Continuous Hidden Markov Model, CHMM)．

[1] 2003-06-10 收到，2003-11-20 改回
国家自然科学基金 (No.60272039) 和安徽省自然科学基金 (No.01042205) 资助课题

当用一个 CHMM 模型描述很多人的同一发音时，为使模型中的参数能反映多个说话者的特征，通常提高 CHMM 模型中的高斯混合度的数目，但是混合度越高，所需要的训练数据就越多，而在实际应用中很难提供充足的训练数据．另外， Rabiner 等人对英文非特定人孤立数字的识别实验表明 [1] ，当高斯混合度超过一定范围后识别性能难以再提高，此时用大量说话者的语音集中训练得到的多混合度 CHMM 的输出，是语音中包含的多个人的信息混合后的统计平均，很难适合每个人．随着说话者数目的增多和语音变化范围的增大，传统多混合度 CHMM 在非特定人语音识别中的这种局限性将越发明显．

目前，解决的办法之一是采用说话者自适应，在一个由很多说话人训练出的非特定人语音识别系统中加入少量新用户的训练语音数据，通过一定的算法，使原模型很快地自适应到新用户的模式下工作．这是介于特定人和非特定人语音识别之间的一种合乎逻辑的折衷．

此外，在语音识别中，为吸收对同一发音内容不同说话人之间的发音差异，可采用基于说话者分类的语音识别的方法 [2] ．按照一定的说话者聚类准则，对参与训练的参考说话者进行聚类分析，将他们聚成 K 类，每类建立一套模板，由于同类说话者的特征较接近，因而模板建立较容易，性能也较好．识别时，可以根据说话人的少量语音来进行类别判断，然后启动最为接近的一类模板进行识别．图 1 为识别框图，可见，在采用基于说话者分类的语音识别方法时，由于聚类分析和类别判断是前提，所以这部分计算量的增加是不可避免的．另外，分类数的选择还将影响系统的建模和识别性能，分类数过少，类内差别过大，达不到易于建模和提高识别性能的目的，反之分类数过多会给系统建模带来很大的计算量和存储量．

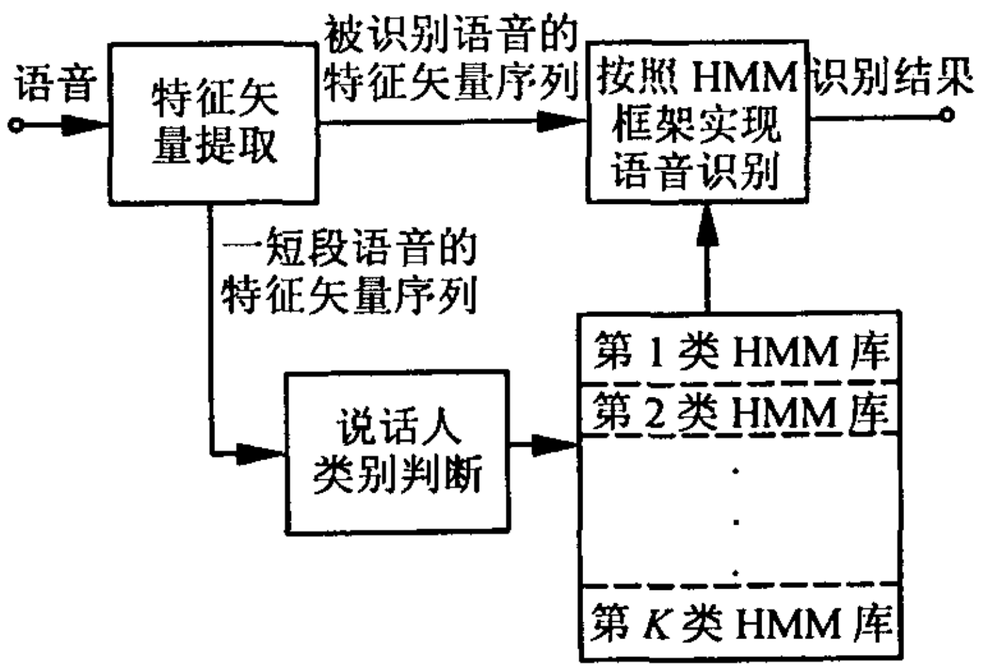

图 1 基于说话者分类的语音识别系统的框图

本文提出了一种适于非特定人语音识别的模型结构——并行 HMM(PHMM)，它由多条并行马尔可夫链组成，从而将上述基于分类建立的多个 CHMM 融合于一个模型中，提高了识别性能．同时各条链之间允许有交叉，使得各 CHMM 之间存在着状态共享，因此减少了模型参数的个数，亦即降低了模型参数估值的计算量．另外，此结构使模型的刻划更为精细，减少了发音内容相近的语音之间因混淆而造成的误识，提高了模型间的可区分性，从而使识别性能的噪声鲁棒性也得以提高．通过与传统 CHMM 在加性噪声干扰下的对比实验验证了这一点．

2 PHMM 模型

2.1 PHMM 的结构

在传统 CHMM 的基础上，一种适于非特定人语音识别的改进的隐马尔可夫模型—— PHMM 的结构如图 2 所示．它由多条并行马尔可夫链组成，各条链之间允许有交叉，即交叉线联系的两个成分之间允许有转移，这样把状态转移细化为高斯混合成分之间的转移 (同一状态的各高斯混合成分之间不能转移)．由于自左至右的状态转移描述符合人的语音特点，且在汉语中吃音现象较少，所以 PHMM 仍是无跳转的左右结构．在图 2 中假设状态数为 3，并行马尔可夫链数也为 3，在全连接的情况下对应有图 3 所示的高斯混合成分之间的转移矩阵，其中对角子矩

阵表示自转移，后面的满阵子矩阵表示全连接．训练时对模型参数进行估值，测试时输出结果来自所有链的最后一个节点的输出之和．

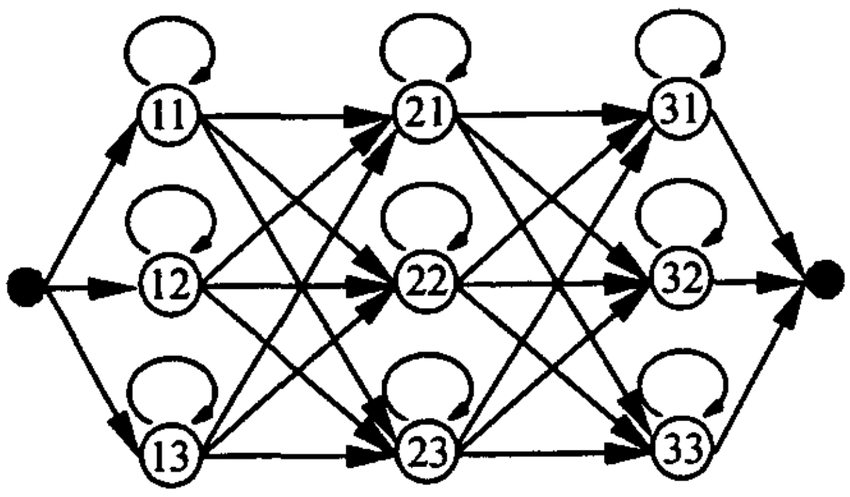

图 2　3 个状态、3 条并行链的 PHMM 的拓扑结构图

$$\begin{array}{ccc|ccc|ccc}
a_{11,11} & 0 & 0 & a_{11,21} & a_{11,22} & a_{11,23} & 0 & 0 & 0 \\
0 & a_{12,12} & 0 & a_{12,21} & a_{12,22} & a_{12,23} & 0 & 0 & 0 \\
0 & 0 & a_{13,13} & a_{13,21} & a_{13,22} & a_{13,23} & 0 & 0 & 0 \\
\hline
0 & 0 & 0 & a_{21,21} & 0 & 0 & a_{21,31} & a_{21,32} & a_{21,33} \\
0 & 0 & 0 & 0 & a_{22,22} & 0 & a_{22,31} & a_{22,32} & a_{22,33} \\
0 & 0 & 0 & 0 & 0 & a_{23,23} & a_{23,31} & a_{23,32} & a_{23,33} \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & a_{31,31} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{32,32} & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & a_{33,33}
\end{array}$$

图 3　A 修正矩阵

从另一个角度看，可把图 2 中包含数字的圆圈看成是有观察值输出的状态，且每个状态只有一个高斯混合成分，而黑圆点表示的起、终点是无观察值输出的状态，每个带箭头的线段仍表示状态间的转移，那么从起点到终点的每一条路径可被看作是基于分类时为某类建立的 CHMM(此时 CHMM 的每个状态只有一个高斯混合成分，在图 2 所示情况下共有 3^3 个这样的 CHMM 模型)，因此 PHMM 可以看作是将基于分类建立的多个模板融合于一个模型中．

为此，我们进行了汉语非特定人孤立数字的识别实验，有说话者 12 男 4 女，每人的发音 20 遍，前 10 遍作训练集，后 10 遍作测试集．实验中把他们分别聚成一类、两类 (男女各一类)、三类 (男声两类女声一类)、四类 (男女各两类) 和五类 (男声三类女声两类) 几种情况，并为每一类建立两套孤立数字 CHMM，状态数均为 5，而高斯混合度数 (M) 分别为 3 和 1．实验结果如图 4 所示．可见基于说话者分类的语音识别中，随着分类数的增加误识率降低，但当分类数达到一定值后识别性能难以再提高；再比较混合度数分别为 3 和 1 的情况，当分类数少时类内差别较大，此时模型应有较高的混合度以达到较好的识别性能，而随着分类数的增加即细致分类时，每一类中说话者的数目减少、特征更接近，因而混合度数可以降低．我们再用改进的 PHMM 模型进行了非特定人孤立数字的识别实验，PHMM 的状态数也为 5，且并行马尔可夫链数由 1 增至 5．当并行链数为 1 时，与聚成一类且 CHMM 为 5 状态 1 个混合度的情况相同．当有多条并行马尔可夫链时，并行链数较少的情况下就融合了较多的模板，此时就能接近基于分类的语音识别系统细致分类时的性能；而随着并行链数的进一步增加识别性能难以再提高．

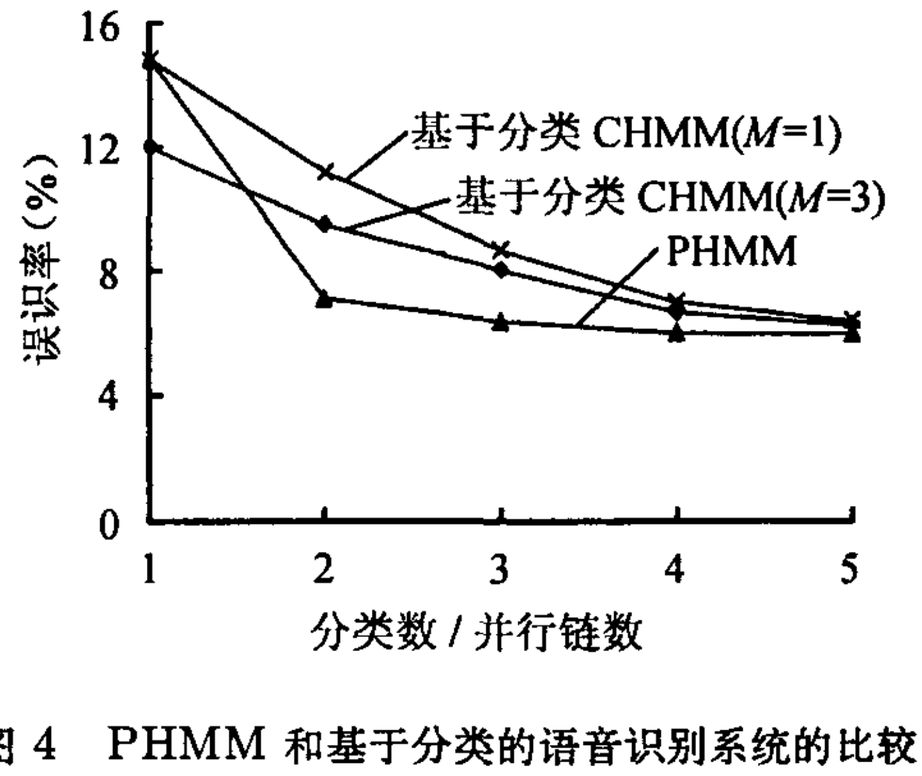

图 4　PHMM 和基于分类的语音识别系统的比较

综上所述，不难看出，虽然 PHMM 在并行链数较少时就融合了多个状态只含一个高斯成分的 CHMM，但由于各 CHMM 之间存在着状态共享，从而减少了模型参数的个数．以上述实验中分类数为 5，每类 CHMM 的状态数为 5 且混合度数为 1 的情况为例，所需的输出状态概率函数的个数为 25，自转移概率函数个数为 25，互转移概率函数个数为 20；而达到同样识别率时 PHMM 的并行马尔可夫链数为 3，此时所需的输出状态概率函数的个数为 15，自转移概率函数个数为 15，互转移概率函数个数为 36．对比可见后者减少了模型参数的个数，从而也就减少了模型参数估值的计算量．此外，基于说话者分类的语音识别中聚类分析和类别判断的

计算量是不可避免的，而 PHMM 可直接由训练数据在训练过程中完成聚类，且测试语音经过所有的类别得到输出结果，从而又在一定程度上减少了计算量.

2.2 PHMM 的实现

模型的观察值序列记为 $\boldsymbol{O}=\{\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T\}$，模型的参数为 $\lambda=(\pi,\boldsymbol{A},B)$，其中 π 为高斯混合成分的初始概率矢量，$\boldsymbol{A}$ 为高斯混合成分之间的转移概率矩阵，B 为每个状态输出的观察值 $\boldsymbol{o}_t$ 的概率密度函数组成的函数系，则定义前向概率：

$$\alpha_t(i,j)=P(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_t,q_t=i,m_t=j/\lambda) \tag{1}$$

表示在时刻 t，状态为 i、高斯混合成分为 j 时输出部分观察值序列 $\{\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_T\}$ 的概率.

从中可以看出，我们的改进就是把统计模型做得更细致，细化到状态内的各高斯混合成分，使之能更精确地描述语音的各个细节. 前向过程中的初始化：

$$\alpha_1(i,j)=\pi_{(i,j)}b_{(i,j)}(\boldsymbol{o}_1) \tag{2}$$

其中 $\pi_{(i,j)}$ 是混合成分 (i,j) 的初始概率，$b_{(i,j)}(\boldsymbol{o}_1)$ 为混合成分 (i,j) 的观察值概率密度函数在矢量 $\boldsymbol{o}_1$ 上的输出. 递推公式为

$$\alpha_t(i,j)=\left[\sum_{l=1}^{N}\sum_{k=1}^{M}\alpha_{t-1}(l,k)a_{(l,k)(i,j)}\right]b_{(i,j)}(\boldsymbol{o}_t) \tag{3}$$

其中 $a_{(l,k)(i,j)}$ 是混合成分 (l,k) 转移到 (i,j) 的概率，N 是状态数，M 是混合度数即层数. 终止

$$P(\boldsymbol{O}/\lambda)=\sum_{i=1}^{N}\sum_{j=1}^{M}\alpha_T(i,j) \tag{4}$$

其中 $P(\boldsymbol{O}/\lambda)$ 为由模型 λ 产生出 $\boldsymbol{O}$ 的概率密度.

同理有后向概率

$$\beta_t(i,j)=\sum_{l-1}^{N}\sum_{k=1}^{M}\beta_{t+1}(l,k)b_{(l,k)}(\boldsymbol{o}_{t+1})a_{(i,j)(l,k)} \tag{5}$$

改进 HMM 的 Baum_Welch 重估公式为

$$\overline{\pi}_{(i,j)}=\alpha_1(i,j)\beta_1(i,j)/[P(\boldsymbol{O}/\lambda)],$$

$$\overline{a}_{(l,k)(i,j)}=\sum_{t=1}^{T}\alpha_{t-1}(l,k)a_{(l,k)(i,j)}b_{(i,j)}(\boldsymbol{o}_t)\beta_t(i,j)\Big/\sum_{t=1}^{T}\alpha_{t-1}(l,k)\beta_{t-1}(l,k),$$

$$\overline{c}_{(i,j)}=\sum_{t=1}^{T}\gamma_t(i,j)\Big/\sum_{t=1}^{T}\sum_{j=1}^{M}\gamma_t(i,j),\quad \overline{\mu}_{(i,j)}=\sum_{t=1}^{T}\gamma_t(i,j)\boldsymbol{o}_t\Big/\sum_{t=1}^{T}\gamma_t(i,j),$$

$$\overline{U}_{(i,j)}=\sum_{t=1}^{T}\gamma_t(i,j)(\boldsymbol{o}_t-\mu_{(i,j)})(\boldsymbol{o}_t-\mu_{(i,j)})^{\mathrm{T}}\Big/\sum_{t=1}^{T}\gamma_t(i,j) \tag{6}$$

其中 $\gamma_t(i,j)$ 表示状态为 i 时出现高斯混合成分 j 的概率：

$$\gamma_t(i,j)=\left[\frac{\sum_{j=1}^{M}\alpha_t(i,j)\beta_t(i,j)}{\sum_{i=1}^{N}\sum_{j=1}^{M}\alpha_t(i,j)\beta_t(i,j)}\right]\left[\frac{c_{(i,j)}\boldsymbol{N}(\boldsymbol{o}_t,\mu_{(i,j)},U_{(i,j)})}{\sum_{j=1}^{M}c_{(i,j)}\boldsymbol{N}(\boldsymbol{o}_t,\mu_{(i,j)},U_{(i,j)})}\right] \tag{7}$$

3 实验结果和分析

孤立数字音的识别是语音识别的一个基本问题，很多语音识别的新方法往往以此作为最初的尝试，利用上述算法对汉语孤立数字识别进行 CHMM 和 PHMM 性能的对比实验．语音信号以 16 kHz 速率采样，量化精度为 16 bit，帧长 20 ms，帧移 10 ms，特征矢量取 12 维 Mel 频率倒谱参数 (MFcc)、12 维 Mel 频率差分倒谱参数 (ΔMFcc)、归一化能量和一阶差分能量共 26 维．传统 CHMM 采用状态数为 5、高斯混合度为 3 的左右无跳转结构；PHMM 的状态数也为 5，并行马尔可夫链数为 3．

实验数据集为 0~9 和幺共 11 个汉语数字发音，共用了 20 男 4 女的干净环境下的语音，从中选取 12 男 4 女作为圈内说话者，每人每音发 20 遍，前 10 遍作训练集，后 10 遍作圈内测试集．另外 8 男作为圈外说话者，只有 10 遍发音，全部作圈外测试集．此外，将圈内说话者组成 3 个集合：(1) 为圈内说话者中的 4 个男生，(2) 包括圈内所有 12 个男生，(3) 包括圈内所有 12 男 4 女．按上述集合构成训练集 (1)、(2)、(3) 和圈内测试集 (1)、(2)、(3)．

3.1 说话者人数对识别性能的影响

分别采用训练集 (1)、(2)、(3) 训练两种模型，且用相应的测试集进行圈内说话者的两种模型的识别实验，结果如表 1 所示．同样训练集的模型下圈外测试集 (测试说话者未参与训练) 的实验结果如表 2．

表 1　圈内测试集时误识率的比较 (%)

	圈内测试集 (1)	圈内测试集 (2)	圈内测试集 (3)
	训练集 (1)	训练集 (2)	训练集 (3)
CHMM	1.48	4.20	10.84
PHMM	1.14	2.61	6.25
相对误识率下降	22.97	37.86	42.34

表 2　圈外测试集时误识率的比较 (%)

	圈外测试集	
	训练集 (1)	训练集 (2)
CHMM	18.68	12.95
PHMM	17.43	9.09
相对误识率下降	6.75	29.81

从表 1 可以看出，随着说话者数目的增加，两种模型的非特定人识别系统的性能都有所下降，说话者人数越多，差别越大，则系统识别性能下降越多，但 PHMM 由于刻划精细，较之传统 CHMM 的性能一致地提高．

从表 2 可以看出，由于圈外测试集的说话者未参与训练，其识别性能较之圈内说话者有明显下降，但训练集 (2) 中由于参与的说话者人数增加，其性能优于训练集 (1)．但对比两种模型系统的结果，PHMM 较之传统 CHMM 的性能有较大的提高．

3.2 识别性能的噪声鲁棒性比较

采用训练集 (2) 分别建立两种系统的模型，对不同的加噪语音分别进行圈内测试集 (2) 和圈外测试集的识别性能比较实验，实验结果如表 3 所示．

从表中可以看出，PHMM 提高了噪声鲁棒性，使得在噪声环境下识别性能较传统 CHMM 也有所提高．

4 结束语

本文提出了一种适于非特定人语音识别的隐马尔可夫模型结构—— PHMM，它由多条并行的马尔可夫链组成，各链之间还允许有交叉，使模型的刻划更为精细，达到了“基于说话者分类的语音识别”的效果，提高了识别性能．此外，这一模型还可适当的扩展，以包容更多的可能出现的音变现象 [3]．

表 3 在噪声环境下的误识率的比较 (%)

		CHMM	PHMM
干净	圈内	4.20	2.61
	圈外	12.95	9.09
25 dB	圈内	5.10	3.02
	圈外	13.52	10.11
20 dB	圈内	9.53	6.14
	圈外	17.16	11.48
15 dB	圈内	13.56	10.59
	圈外	28.18	16.02
10 dB	圈内	24.55	18.18
	圈外	46.67	32.39
5 dB	圈内	47.16	38.64
	圈外	59.32	52.84
0 dB	圈内	59.55	56.82
	圈外	79.32	76.14

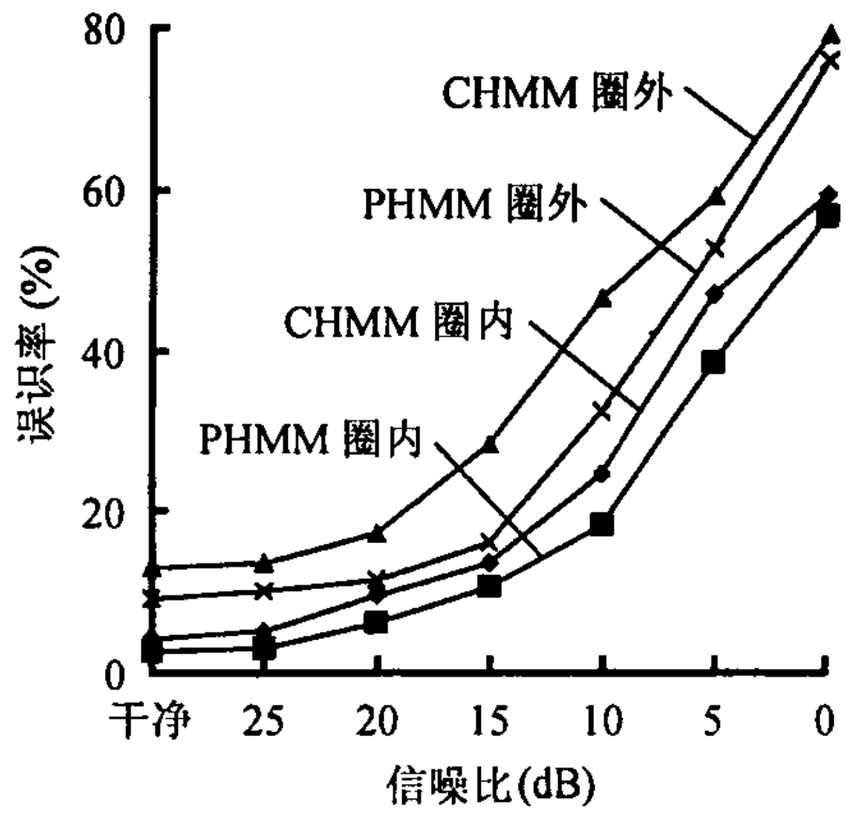

图 5 PHMM 和 CHMM 的识别性能噪声鲁棒性比较

参 考 文 献

[1] Rabiner L, Juang B-H 著，阮平望，译. 语音识别基本原理. 北京：清华大学出版社， 1999: 378–382.

[2] 戴蓓蒨，郁正庆，戴任飞，等. 基于话者分类和 HMM 的话者自适应语音识别. 中国科学技术大学学报， 1996, 26(2): 147–153.

[3] Wolfertstetter F. Ruske G. Structured Markov models for speech recognition. In Proc. of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Detroit, USA, 1995, vol.1: 544–547.

陈雁翔： 女， 1972 年生，博士生，主要研究方向：语音信号处理、模式识别与人工智能.

戴蓓蒨： 女， 1941 年生，教授，博士生导师，主要研究方向：语音信号处理、图像处理、模式识别与人工智能.

周 曦： 男， 1981 年生，硕士生，主要研究方向：语音信号处理.

刘 鸣： 男， 1976 年生，硕士，主要研究方向：语音信号处理.