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

基于分类特征空间高斯混合模型和神经网络融合的说话人识别 [1]

黄 伟 戴蓓蒨 李 辉

(中国科学技术大学电子科学与技术系 合肥 230026)

摘 要: 该文提出了一种基于分类高斯混合模型和神经网络融合 (FS-GMM/NN) 的说话人识别方法，通过对特征矢量进行聚类分析，将说话人的训练语音分成若干类，然后根据各个类中含特征矢量的多少采用不同的模型混合度，训练建立分类高斯混合模型，并采用神经网络实现各个分类高斯混合模型输出的融合．在 100 个男性话者的与文本无关的说话人识别实验中，基于分类高斯混合模型和神经网络融合的方法在识别性能及噪声鲁棒性上都优于不分类的 GMM 识别系统，并具有较高的模型训练效率，且可以有效地降低话者模型的混合度和测试语音长度.

关键词: 说话人识别，分类特征空间，高斯混合模型，神经网络融合

中图分类号: TP391.42 **文献标识码:** A **文章编号:** 1009-5896(2004)10-1607-06

Speaker Identification Based on Classify Feature Sub-space Gaussian Mixture Model and Neural Net Fusion

Huang Wei Dai Bei-qian Li Hui

(*Electron. Sci. and Tech. Dept, Univ. of Sci. and Tech. of China, Hefei* 230026, *China*)

Abstract In this paper, a speaker identification system is proposed based on classify Feature Sub-space Gaussian Mixture Model and Neural Net fusion (FS-GMM/NN) . With clustering analysis of the feature vectors, the speaker's training feature vectors can be classified to some subsets and training classify Gaussian Mixture Models (GMM) with different mixtures according to the subset's feature vectors's number. Finally, the outputs of every classify GMM will be fused by Neural Net (NN). In the experiment of text-independent speaker identification of 100 speakers (male), the system based on FS-GMM/NN overmatch the Baseline Gaussian Mixture Model (B-GMM) in identification performance and noise robustness with fewer mixtures and shorter test speech. Moreover, the training of FS-GMM/NN is more effective.

Key words Speaker identification, Classified feature-subspace, GMM, Neural Net(NN) fusion

1 引言

近年来，在与文本无关的 (Text-independent) 说话人识别系统中大多采用高斯混合模型 (Gaussian Mixture Model, GMM) 作为说话人模型，取得了比隐马尔可夫模型 (Hidden Markov Model, HMM) 更好的识别性能 [1,2] ．一些当今最高水平的系统都采用了 GMM ，如目前在 NIST(美国国家标准技术局) 举办的一年一度的说话人识别评测中，领先的系统基本上都是基于 GMM[3] ．高斯混合模型是一种基于短时谱的统计模型，通过多个多维高斯密度函数的加权和

[1] 2003-05-16 收到， 2003-12-04 改回
国家自然科学基金项目 (60272039) 和安徽省自然科学基金项目 (01042205) 资助课题

来描述说话人特征信息在特征空间中的分布，因此高斯混合模型的混合度越高，对说话人特征信息分布的描述越细致．与所有的统计模型一样，高斯混合模型的性能依赖于模型训练的训练语音数据是否充分，以及较长的测试语音．显然，这种对训练语音量和测试语音长度的要求不太适宜于说话人识别技术的实用化．因此，在保持相同识别性能的基础上，如何减少模型对语音量及测试语音长度的要求，成为本文研究的主要目的．

通过对说话人特征信息在特征空间中分布情况的观察与研究，我们发现对于采用 D 维 Mel 倒谱参数 (MFCC) [4−6] 作为特征信息时，说话人一段语音的 K 组 MFCC 矢量在特征空间中的分布是不均匀的，若能通过聚类技术将这种不均匀表示为几类，每个子类包含的特征矢量组变少了，可以用较低的模型混合度来描述类特征空间的分布，减少模型对训练语音量以及测试语音长度的要求．鉴于此，本文提出了一种基于分类高斯混合模型 (FS-GMM) 和神经网络融合的说话人辨识方法，先通过对特征矢量的聚类分析，将说话人的训练语音特征矢量集分成若干类，根据各个类中含特征矢量的多少采用不同的混合度，训练建立分类高斯混合模型，用神经网络实现各个分类高斯混合模型输出的融合．实验显示，采用本文提出的方法在总混合度相同情况下，较不分类的 GMM 具有更好的识别性能及鲁棒性，且模型的训练效率高，在较少训练语音量及较短测试语音长度时仍具有较好的识别性能．

2 分类高斯混合模型 (FS-GMM)

2.1 分类高斯混合模型

由于说话人的一段语音发音中的各种音素发音特征的出现频度是不相同的，加上不同音素发音特征的短时谱在特征空间的分布不同，造成话者特征信息在特征空间分布的不均匀，稀密程度也不同，而且特征空间中不同区域对话者识别性能影响也有所不同．因此我们采用 VQ 聚类分析的方法，对说话人的训练语音集进行特征矢量分类，将特征空间分为若干个子类，在每个子类内单独训练建立分类高斯混合模型．第 i 个子类 GMM 的概率密度函数表示为 $P_i(\boldsymbol{x}|\lambda_i)=\sum_{j=1}^{M_i} w_{ij}N(\boldsymbol{x};\mu_{ij},\boldsymbol{v}_{ij})$, $\sum_{j=1}^{M_i} w_{ij}=1$, $(0<w_{ij}<1)$．式中，$\boldsymbol{x}$ 是 D 维特征矢量，M_i 为混合高斯成分个数．μ_{ij} 和 $\boldsymbol{v}_{ij}$ 分别表示第 i 个分类的第 j 个混合高斯成分的均值和方差矩阵．因此第 i 个分类高斯混合模型的参数 λ_i 描述为 $\lambda_i=\{M_i,w_{ij},\mu_{ij},\boldsymbol{v}_{ij}\}$, $j=1,2,\cdots,M_i$．对于某个说话人的训练集语音，经 VQ 聚类分析，分为 N 个子类，N 个训练语音子集分别训练得到 N 个子类高斯混合模型，由它们组成该说话人的 FS-GMM，其结构如图 1 所示．

FS-GMM 由 N 个子 GMM 组成“或”的关系．对于组成 FS-GMM 的各个子类模型而言，由于其仅代表了总的特征空间中的一个部分，因此子类 GMM 所需要的混合度就可以比较低．不同子类所包含的矢量个数不同，可以采取不同的模型混合度．本文中采用了限定总混合度 M，

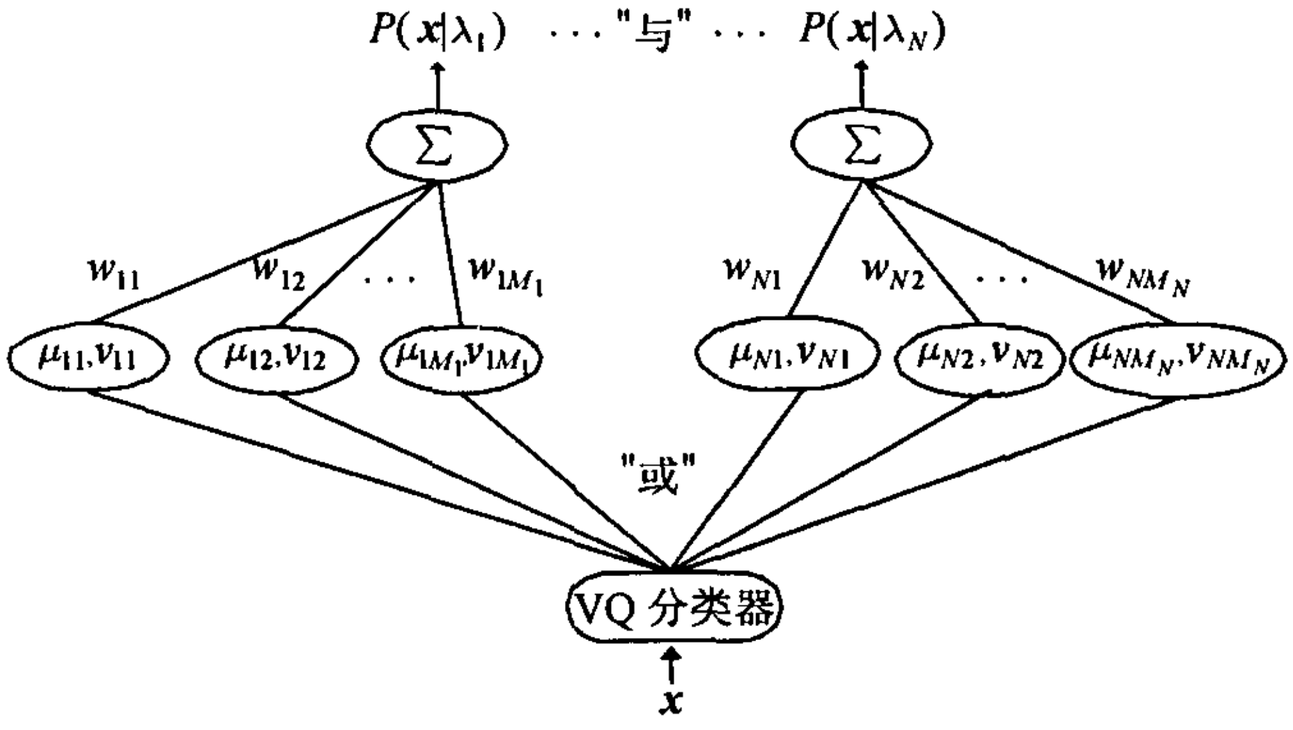

图 1 FS-GMM 模型结构图

并按各训练语音矢量子集的大小分配子类 GMM 混合度 M_i 的方法，即 $M_i = T_i/T_{\text{all}} \times M$，$i = 1, 2, \cdots, N$，其中，$M_i$ 和 T_i 分别满足约束条件 $\sum_{i=1}^{N} M_i = M$ 和 $\sum_{i=1}^{N} T_i = T_{\text{all}}$.

2.2 FS-GMM 的训练效率

经分类后建立的子类 GMM 模型的混合度较低，因此即使是训练 N 个子 GMM 模型，与不分类 GMM 相比，仍具有更快的训练效率. 我们以对 GMM 模型高斯分量系数的重估为例，若限定话者模型总的混合度个数为 M，训练特征矢量个数为 T_{all}，训练的迭代次数为 L，则 GMM 系数重估的计算次数约为 $Q = M \times T_{\text{all}} \times M \times L$；而分类的 GMM 模型，若分类空间的个数为 N，总的混合度个数为 M，则 N 个子类模型的训练时间为 $\sum_{i=1}^{N} M_i \times T_i \times M_i \times L_i$.

若假设每个子类的高斯混合度相同，即 $M_i = M/N, T_i = T_{\text{all}}/N$，则对分类 GMM 的重估的计算次数大约为 $N \times M/N \times T_{\text{all}}/N \times M/N \times L_i = L_i/N^2 L \times Q$. 另外，由于分类后每个子类的特征矢量相对较少，而且每个子类中的特征矢量相对集中，因而训练的收敛较快，也即 $L_i < L$.

3 基于 FS-GMM 和神经网络融合的说话人系统构成

3.1 系统结构

基于 FS-GMM 和神经网络融合的说话人识别方法的系统框图如图 2 所示. 在系统的训练阶段，首先对每个话者的训练语音特征矢量集进行 VQ 聚类分析，将话者训练语音集矢量分为 N 个子类，并记录下每个子类的类心，组成 VQ 分类器，然后再按分类的结果训练建立 N 个子类 GMM . 最后，对训练语音集的各个子模型输出训练一个进行数据融合的神经网络. 因此，每个话者模型的参数将由三部分组成：(1) 组成 VQ 分类器的 N 个类心特征矢量参数；(2) N 个子类 GMM 参数 $\lambda = \{\lambda_i\}$；(3) 神经网络权值参数.

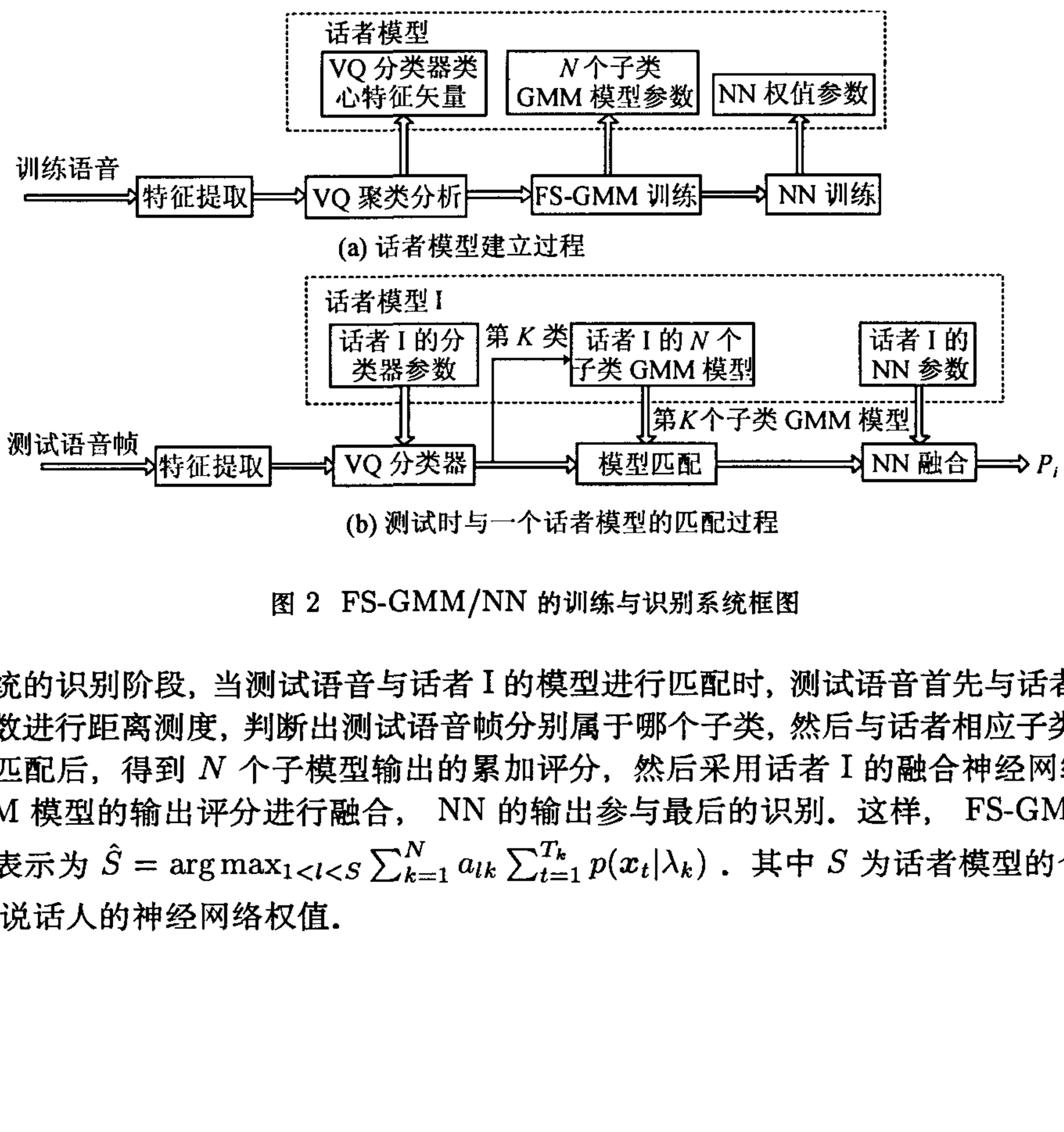

图 2　FS-GMM/NN 的训练与识别系统框图

在系统的识别阶段，当测试语音与话者 I 的模型进行匹配时，测试语音首先与话者 I 的分类器类心参数进行距离测度，判断出测试语音帧分别属于哪个子类，然后与话者相应子类的 GMM 模型进行匹配后，得到 N 个子模型输出的累加评分，然后采用话者 I 的融合神经网络对 N 个子类 GMM 模型的输出评分进行融合，NN 的输出参与最后的识别. 这样，FS-GMM/NN 的识别可以表示为 $\hat{S} = \arg\max_{1<l<S} \sum_{k=1}^{N} a_{lk} \sum_{t=1}^{T_k} p(\boldsymbol{x}_t|\lambda_k)$. 其中 S 为话者模型的个数，a_{lk} 为第 L 个说话人的神经网络权值.

3.2 神经网络融合

各个子类特征空间中包含了说话人的不同特征信息，它们对于系统识别性能的影响也是不一样的．因此，如何有效地对各子类 GMM 模型的输出进行数据融合是十分重要的．由于神经网络具有很好的非线性映射能力，因此在本文中我们采用神经网络对 FS-GMM 的多个子模型的输出进行融合，将融合后的结果作为最终的评分．

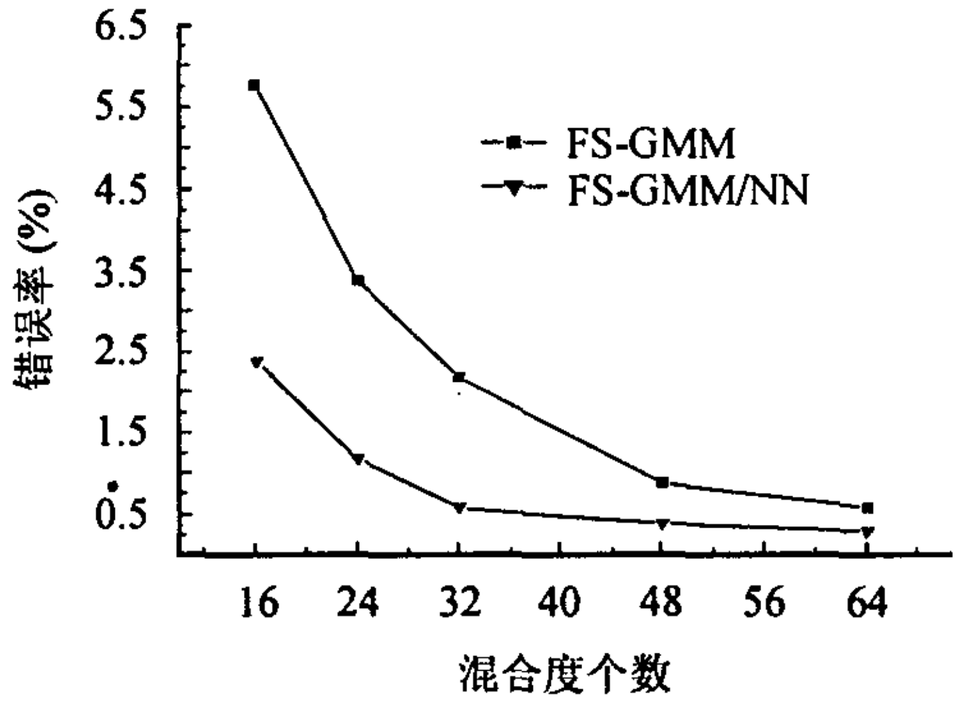

图 3 融合方式对系统识别性能的影响

我们通过实验对直接融合与神经网络融合的效果进行了比较，实验结果如图 3 所示， FS-GMM 表示直接相加融合，实验采用分类空间个数为 4 ，测试语音时间长度为 1 s(实验条件见第 4 部分所述) ．从图 3 看出，对于总模型混合度从 16~64 的几种取值，神经网络融合的效果优于直接相加融合，尤其是在模型的混合度较低的情况下，神经网络融合的效果更为显著，这样就有利于我们采用较低的模型混合度和较短的测试语音的策略．

4 实验结果及分析

为了进一步验证基于 FS-GMM/NN 的说话人识别系统的识别性能以及运算效率，我们进行了对比实验，为此我们建立了两个实验系统．第 1 个实验系统是不分类的 GMM(B-GMM) ；第 2 个实验系统是基于分类高斯混合模型和神经网络融合的 (FS-GMM/NN) ．实验采用微软亚洲研究院提供的 100 个男性话者的麦克风语音数据库 [6] ，每个说话人有 20 条语音 (每条语音时长约为 5~7 s) ，取其中 10 条组成训练集，另 10 条组成测试集．语料库是在干净环境下的随意发音方式，实验时的采样频率为 16 kHz, 16 bit 量化，短时分析的帧长为 20 ms ，帧移 10 ms ，选用 12 阶 Mel 倒谱参数 (MFCC) 作为特征参数．

4.1 类空间个数与模型混合度的关系

对上述两种实验系统进行的不同模型混合度 (M) 和不同分类个数 (N) 情况下的说话人识别实验，在测试语音长度为 1 s 时的实验结果如表 1 所示，表中所示为误识率．可以看出： (1) 在总的混合度个数 M 相同的情况下，采用分类的 GMM(FS-GMM/NN) 的识别性能均优于不分类的 GMM (B-GMM) ； (2) 采用 FS-GMM 和神经网络融合的策略可以选用较低的高斯混合度获得较好的识别性能，例如选取 $M = 16, N = 2$ 时，其误识率为 2.4% ，优于 B-GMM 选用 $M = 64$ 时的性能．

表 1 不同类空间个数和模型混合度时的误识率 (%)

M	N			
	B-GMM	FS-GMM/NN		
	1 sub-space	2 sub-space	3 sub-space	4 sub-space
64 Mix	2.9	0.3	0.3	0.2
48 Mix	3.3	0.5	0.6	0.3
32 Mix	3.7	0.7	0.7	0.5
24 Mix	4.8	1.3	1.2	1.1
16 Mix	7.6	2.4	2.5	2.3

4.2 类空间个数与测试语音长度的关系

对于基于 GMM 的说话人识别系统，较长的测试语音会有较好的识别结果．对于 FS-GMM/NN 系统，由于采用分类的策略，使得 M_i 变低，从而可以在较短的测试语音时就具有较好的效果．为此，我们进行了总 $M = 32$ 时，不同 N 和不同 T 时对比实验，结果如表 2 所示．当测试语音长度 T 变短时，系统性能都会下降，但是当测试语音长度减少为 1 秒时， FS-GMM/NN 系统仍有较好的识别性能．

表 2　不同类空间个数和模型混合度时的误识率 (%)

T	N			
	B-GMM	FS-GMM/NN		
	1 sub-space	2 sub-space	3 sub-space	4 sub-space
1 s	3.7	0.7	0.7	0.5
2 s	1.2	0	0.1	0.2
3 s	0.9	0	0	0.1
4 s	0.8	0	0	0

4.3 话者模型训练效率的比较

对于 FS-GMM/NN 系统，由于各个子模型的混合度 M_i 较低，同时参与训练的语音样本也是经过了分类后的对应的语音样本，模型训练的收敛较快，在相同的总混合度 M 下，所需要的训练时间比不分类的要少．

图 4 绘出了模型的总混合度 $M = 64$ 时，不同特征空间分类情况下的训练 100 个男性话者模型得出的每个话者模型训练所需要的平均时间．可以看出， FS-GMM 模型训练所需的时间远少于不分类方法．

4.4 类空间个数对识别系统噪声鲁棒性的影响

当话者模型是用干净环境下的语音训练的，而测试用的语音是在一定的环境噪声背景下的带噪语音，这两种环境的不匹配将导致话者识别系统性能的下降．我们对 B-GMM 进行了 SNR=20 dB 加性白噪声下的识别实验 ($M = 64$)，识别系统的误识率从干净测试语音下的 2.9% 降到 14.25% ．而对分类方法的同样条件下的测试结果如图 5 所示．当取 $N = 2$ 时，系统在 SNR=20 dB 时的误识率为 9.25% ，仍优于不分类系统．

5 结论

从实用性出发，为了在较低的模型混合度以及较短的测试语音的情况下获得较好的识别性能，本文提出了一种基于分类高斯混合模型和神经网络融合 (FS-GMM/NN) 的说话人识别方

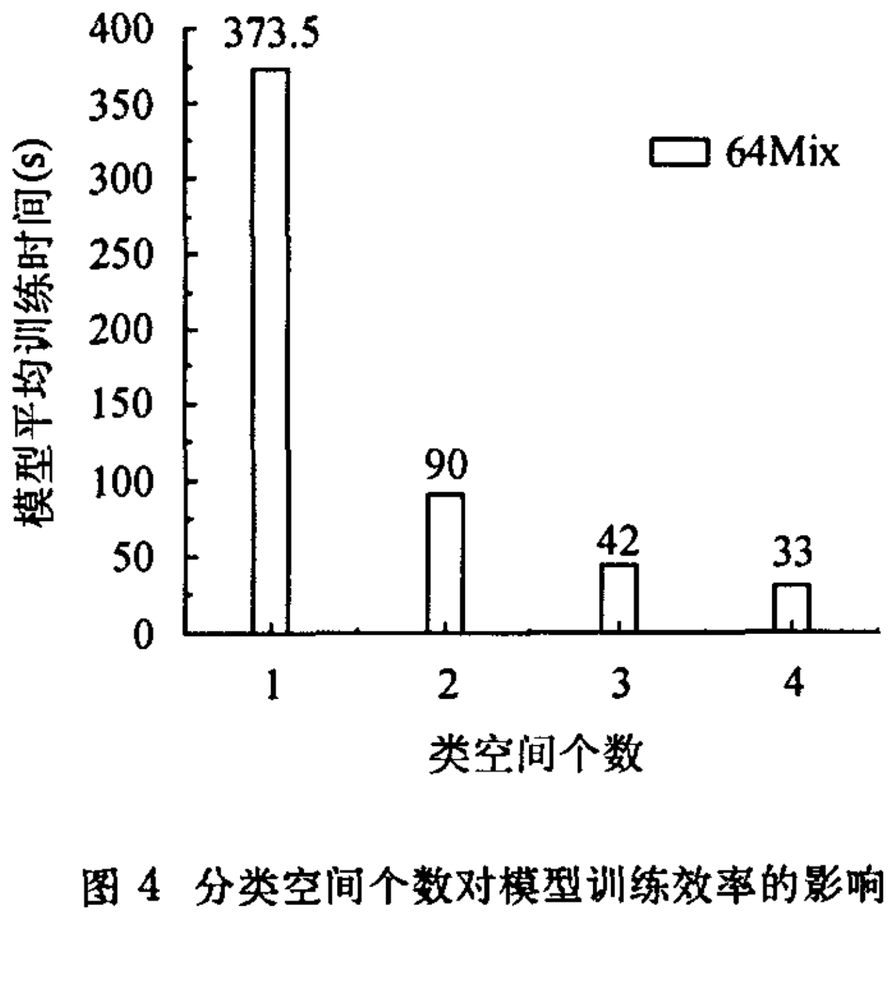

图 4　分类空间个数对模型训练效率的影响

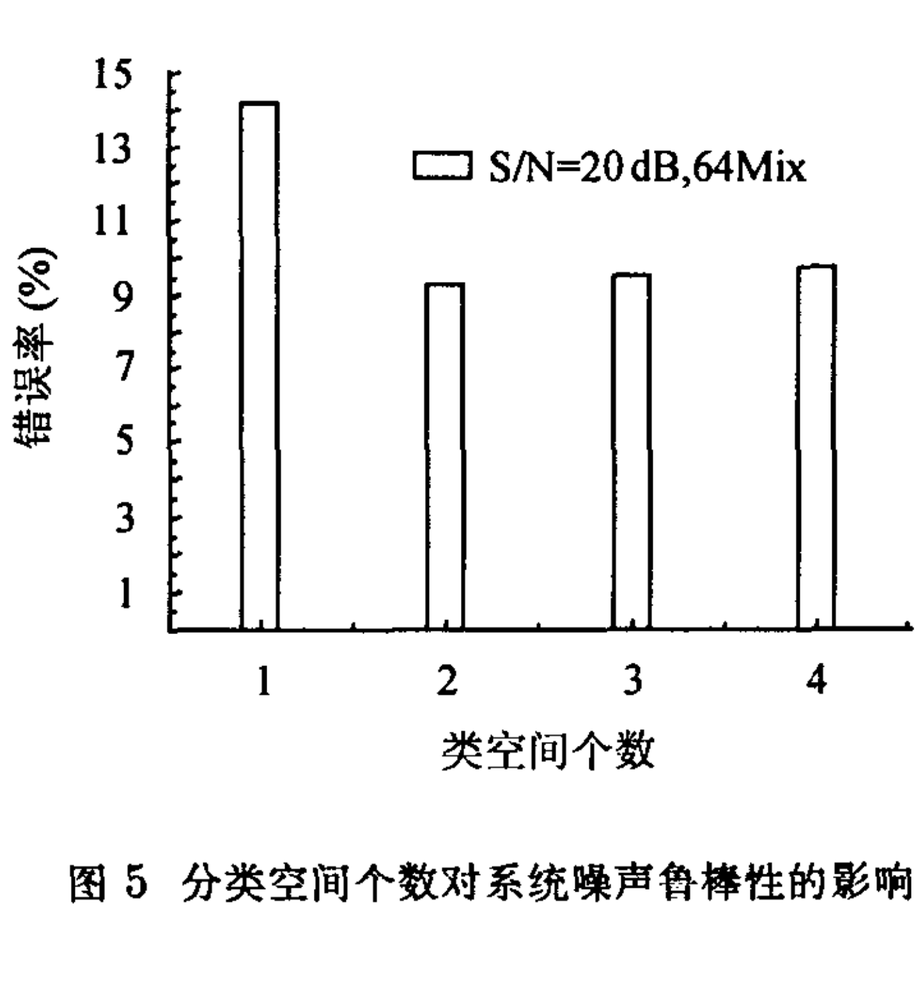

图 5　分类空间个数对系统噪声鲁棒性的影响

法，采用聚类分析的方法将训练语音特征矢量集分成若干个类，根据各个类中特征矢量分布的多少采用不同的模型混合度建立类特征空间的混合高斯模型，并采用神经网络对各个分类模型的输出进行融合.

较之不分类的方法，(1) FS-GMM 中的每个子类 GMM 可用较低的模型混合度，同时使模型的训练以及匹配时间减少，并降低对测试语音长度的要求. (2) 各个子类模型输出经 NN 融合，使不同子类对系统性能的影响不尽相同，突出某些重要子类的重要性. (3) 在总的模型混合度相同的情况下， FS-GMM 具有更好的识别性能以及鲁棒性，模型的训练效率也较高.

参 考 文 献

[1] Reynolds D A, Rose R C. Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. on Speech Audio Process*, 1995, 3(1): 72–83.

[2] Reynolds D A. Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, 1995, 17(1–2): 91–108.

[3] Reynolds D A. Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, 2000, 10(1–3): 19–41.

[4] Deller J R, Proakisa J G, Hansenm J H L. Discrete-Time Processing of Speech Signals. New York: Macmillan Publishing Company, 1993.

[5] Reynolds D A. Experimental evaluation of features for robust speaker identification. *IEEE Trans. on Speech Audio Process*, 1994, 2(4): 639–643.

[6] Chang E, Shi Y, Zhou J, Huang C. Speech lab in a box: A mandarin speech toolbox to jumpstart speech related research. in EUROSPEECH, Aalborg, Denmark, 2001: 192–199.

黄 伟： 男， 1976 年生，博士生，研究方向为模式识别与人工智能、语音信号处理.

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

李 辉： 男， 1959 年生，高级工程师，硕士生导师，主要研究方向为语音信号处理、现代电子系统设计.