HMM 非特定人连续语音识别的嵌入式实现

杜利民　　谢凌云　　刘　斌

（中国科学院声学研究所语音交互技术实验室　北京　100080）

摘　要：　嵌入式系统正逐渐成为语音识别实际应用的首选平台。该文在嵌入式平台上研究 HMM 连续语音识别的计算复杂度要素，提出特征系数屏蔽方法和综合剪枝相结合的“瘦身”计算方法，降低计算复杂度并保持识别率。该方法在嵌入式平台上研究的实验数据表明，HMM 连续语音识别瘦身系统与基线系统相比，计算时间从基线系统的 100%降低到 27.91%，识别率仅从基线系统的 89.65%下降到 89.41%。

关键词：　嵌入式系统，语音识别，搜索算法，特征屏蔽

中图分类号：TP391.42　　**文献标识码：**A　　**文章编号：**1009-5896(2005)01-0060-04

Embedded Implementation of HMM Speaker-Independent Continuous Speech Recognition System

Du Li-min　　Xie Ling-yun　　Liu Bin

(*Lab for Speech Interaction Technology*,

Institute of Acoustics, *Chinese Academy of Sciences*, *Beijing* 100080, *China*)

Abstract　The embedded systems are gradually becoming the first choice of platforms which should be used for real-time speech recognition system. This paper discusses the computation complexity factors of HMM-based continuous speech recognition for embedded system. An optimized way integrating feature masking and pruning is presented to reduce the computation complexity and keep the recognition accuracy. The experiments for embedded system show that, comparing with the base-line system, the computation time is reduced from 100% to 27.91%, and the recognition accuracy is degraded only from 89.65% to 89.41%.

Key words　Embedded system, Speech recognition, Search algorithm, Feature masking

1　引言

随着半导体和集成电路技术的突飞猛进，基于嵌入式系统的便携式移动设备（例如手机、掌上电脑等）正在成为人们生活中不可缺少的电子产品。作为人机交互最自然的方式，语音识别技术在嵌入式系统上的应用也是当前的热点。但是目前的实用产品多是基于使用 DTW 的模板匹配技术的孤立词识别系统，而对于中大词汇量的 HMM 非特定人连续语音识别系统而言，由于嵌入式系统的资源配置与 PC 有着极大的不同，移植起来有相当的困难，最主要的瓶颈在于计算复杂度。

HMM 连续语音识别系统的计算复杂度主要有 3 个组成部分：（1）特征提取；（2）对数概率计算；（3）路径搜索。本文侧重从（2），（3）两个部分研究降低计算复杂度的瘦身方法，在保持或略降识别准确度的前提下，达到显著提高识别速度的目的。首先介绍本文研究的基线系统和嵌入式平台，分析系统各部分的计算复杂度占比，在此基础上给出屏蔽观测矢量和综合剪枝等瘦身计算的优化算法，最后给出方法佐证的实验数据。

2　HMM 基线系统和嵌入式平台

本文研究的基线系统是汉语通用词汇非特定人 HMM 连续语音识别系统[1,2]。连续延展的语音信号先被分成 25ms 等长的帧，相邻帧的间隔为 10ms。在特征提取前，先对分帧的语音信号进行预加重 $H(z)=1-0.97z^{-1}$，接着进行汉明窗加窗处理消除分帧的边界效应，然后进行 FFT 变换，并构成按照 Mel 刻度均匀分布的 26 通道滤波输出，最后对 26 个通道滤波输出进行 DCT 变化，保留 DCT 的前 12 个系数，即 MFCC 系数。这 12 个 MFCC 系数和能量构成当前帧的 13 维静态特征。语音识别系统的观测矢量由当前帧的 13 维静态特征加上其一阶动态特征和二阶动态特征组成，共 39

2003-08-04 收到，2003-12-26 改回

维，速率为 100 帧/s。声学模型采用由左至右可跳转的前后语境相关的三音子(triphone) HMM，状态数为 5 个，状态输出为对角线协方差矩阵的连续高斯密度分布，由 5 个单高斯密度分布加权混合构成。任务模型是出租车对话系统，包含北京的比较有名的机构和景点名称、司机与乘客交流的常用会话。任务模型转换成词图，通过 Viterbi 算法进行束搜索解码。

嵌入式平台的处理器为 Intel StrongARM 的 SA-1110 32 位 RISC，206MHz 主频，32MB SDRAM，16MB ROM，嵌入式 Linux 操作系统。

3　计算复杂度分析及优化

HMM 连续语音识别系统的计算复杂度主要有 3 个组成部分：（1）特征提取过程；（2）对数概率计算过程；（3）路径搜索过程。在一个系统的实现中，这 3 个部分的计算量占比是不同的，而且是不固定的。它随着系统的运行环境、声学模型、识别任务等因素的变化而变化。图 1 和图 2 给出了本系统在 PC 机和嵌入式平台上的计算量占比。

由于 PC 和嵌入式平台进行浮点计算的机制不同，所以它们在浮点计算部分的占比的差异较大。我们研究降低复杂度的着眼点，首先放在计算量占比最大的对数概率计算和搜索计算两个方面。

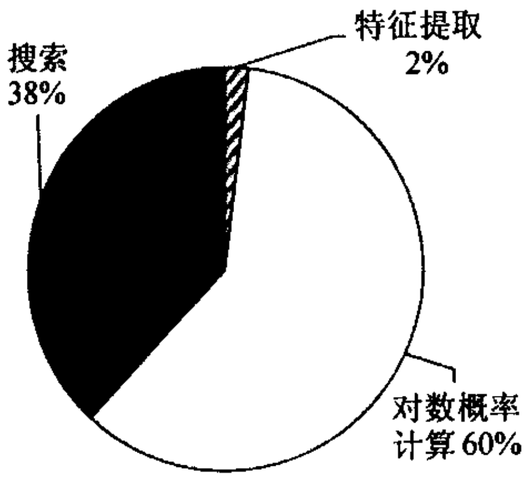

图 1　PC 平台计算量占比

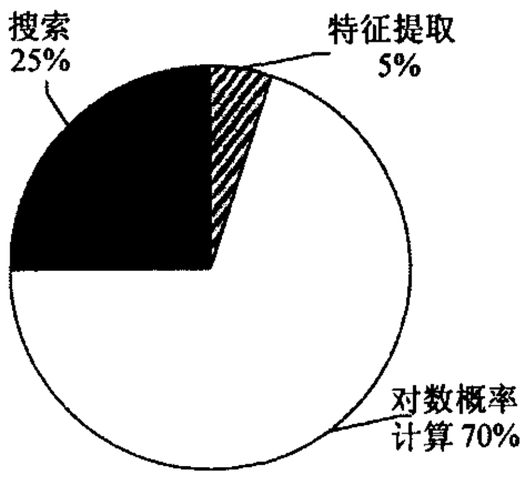

图 2　嵌入式平台计算量占比

3.1 降低高斯密度的混合数

连续密度的 HMM 状态输出由多个单高斯密度分布加权混合构成。一般来说，增加混合的单高斯密度分布的数目，有助于提高语音识别的识别率[3]，但增加的计算复杂度往往很大。对一个具体任务的识别系统，可以通过权衡二者的利弊进行特定目的的优化。对于我们的基线系统，瘦身后的 HMM 状态输出采用单高斯概率分布函数，对数概率计算占空比从 70%下降到 40%左右，识别性能变化甚微。

在下面的研究中，基线系统使用瘦身后的单高斯概率分布函数 HMM。

3.2 屏蔽观测矢量的分量

HMM 的观测矢量共计 39 维，包含当前帧的 13 维静态特征和 13 维一阶动态特征以及 13 维二阶动态特征。为了降低计算复杂度，以当前帧的 13 维静态特征为基础，另加 13 维一阶动态特征或 13 维二阶动态特征构成 26 维观测矢量，虽然实时性提高显著，但是，识别性能的下降也很明显。事实上，在静态特征和一阶二阶动态特征中，每一个系数对正确识别的贡献是不一样的，简单舍弃整个二阶或者一阶动态特征的处理策略是不妥当的。本文实验研究了观测矢量的各维系数对正确识别的贡献，据此确定对观测矢量的一些分量的屏蔽，使语音识别系统在原观测矢量的子空间工作，达到降低计算复杂度的目的。

设观测矢量为 $\boldsymbol{O}$，则单高斯 HMM 的状态 j 的输出概率为

$$b_j(O)=\frac{1}{\sqrt{(2\pi)^n|\Sigma|}}e^{(-1/2)(O-\mu)'\Sigma^{-1}(O-\mu)} \tag{1}$$

其中 n 是观测矢量的维数，μ 是观测矢量的均值向量，Σ 是观测矢量的协方差矩阵。由于协方差矩阵采用逆对角矩阵，可以改写式（1）如下：

$$\ln[b_j(O)]=-\frac{n}{2}\ln(2\pi)+\frac{1}{2}\sum_{i=1}^{n}\left[\ln(\mathrm{cov}[i])-(O[i]-\mu[i])^2\times\mathrm{cov}[i]\right] \tag{2}$$

其中 $O[i]$，$\mu[i]$，$\mathrm{cov}[i]$分别是观测矢量 $\boldsymbol{O}$，均值向量 μ，协方差矩阵（逆对角阵形式）Σ 的第 i 维分量。

从式（2）中可以提取出第 i 维观测矢量对输出概率贡献的表达式，即：

$$a_i=-\frac{1}{2}\left[\ln(2\pi)-\ln(\mathrm{cov}[i])+(O[i]-\mu[i])^2\times\mathrm{cov}[i]\right] \tag{3}$$

为了对每一维特征的贡献进行定量评估，我们定义第 i 维的贡献比率的度量

$$\delta_i=\frac{a_i}{\ln[b_j(O)]} \tag{4}$$

为了研究本系统中观测矢量的每个分量对识别的影响，我们进行了大量的统计实验。设贡献比率的下限阈值为 0.01，上限阈值为 0.1，实验中，分别统计观测矢量的每个分量的贡献比率低于下限阈值和超过上限阈值的概率，结果如图 3 所示。由于超过上限阈值的概率普遍较低，为了与低于下限阈值的概率在同一个图中表现出来，我们在图中对它进行了放大 10 倍的操作，这只是为了让图表的效果更加清楚，并不会影响到我们对结果的分析。图中横坐标轴的观测矢量表示方式为：C1~C12，E0 表示 12 个 MFCC 系数以及能量，D1~D12，E1 和 A1~A12，E2 分别表示其一阶和二阶差分系数。

根据统计数据，我们将屏蔽那些贡献比率低于下限阈值的概率较大，而高出上限阈值的概率又较小的特征矢量，以

减少概率计算时的计算量，又不会对识别率产生过大的影响。在我们的系统中，被屏蔽的分量共 11 维，特征系数为：C12，D10~D12，A5，A6，A8~A12。表 1 为 C12，D10~D12，A5，A6，A8~A12 等 11 个被屏蔽的特征系数的贡献比率低于下限阈值和高出上限阈值的概率数据。

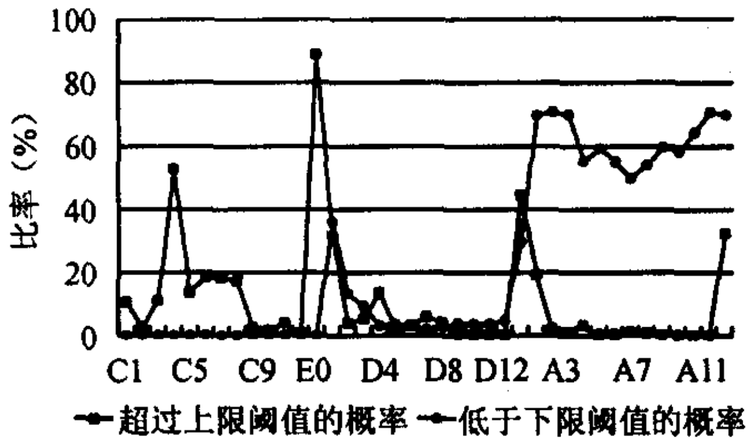

图 3 各维特征矢量的贡献比率统计

表 1 贡献较少的特征系数的贡献比率

特征系数	δ_i >0.1	δ_i <0.01	特征系数	δ_i >0.1	δ_i <0.01
C12	0.0005	0.0021	A8	0.0008	0.5412
D10	0.0002	0.0375	A9	0.0002	0.5982
D11	0.0004	0.0380	A10	0.0003	0.5828
D12	0.0001	0.0467	A11	0.00003	0.6436
A5	0.0004	0.5936	A12	0.00003	0.7073
A6	0.0004	0.5521			

3.3 缩减识别搜索的空间

在一个复杂任务的语音识别系统中，对每一帧观测矢量可能存在成千上万条待延展的路径，每一条待延展的路径都要进行对数概率的计算，然后通过 Viterbi 算法从中找出对数概率最大的路径作为结果输出。式（5）是 Viterbi 算法的基础，其中 $P_j(t)$ 为 t 时刻在状态 j 的最大似然概率，tran_{ij} 为从状态 i 到状态 j 的转移概率，$b_j(o_t)$ 为 t 时刻在状态 j 观测到输入向量 o_t 的输出概率。

$$P_j(t)=\max_i\left\{P_i(t-1)+\lg(\mathrm{tran}_{ij})\right\}+\lg(b_j(o_t)) \qquad (5)$$

如果路径的数目过多，对式（5）的计算意味着庞大的计算量，同时也占用了嵌入式系统宝贵的内存资源。一个有效的解决方法就是剪枝，利用剪枝技术，把那些最终胜出概率非常小的路径早早地舍弃掉。

剪枝可以采取两种直接的策略[4]。（1）路径总数剪枝：先对所有的路径数目设定一个上限阈值 *n*Path，这个阈值一般为靠经验取得的固定值。当每一帧的搜索开始的时候，如果待延展的路径数目超过了 *n*Path，则把所有路径按照当前各自的对数概率由大到小排序，舍弃掉那些排名靠后的，保留排名靠前的 *n*Path 条路径。（2）束剪枝：利用当前路径的对数概率进行剪枝。取得每一步搜索的当前所有路径的最大的对数概率值，并与其他的路径的对数概率值进行比较，当两者的差超过一个预先设定的阈值的时候，就舍弃掉该路径。

嵌入式平台上，我们综合应用上述两种剪枝技术，针对识别率和识别速度的变化，调整两个阈值的大小，以求获得最佳的经验阈值。其中，路径总数的剪枝在每一步搜索刚开始的时候进行，而对数概率的剪枝在每一步搜索刚结束的时候进行，这样能减少下一步搜索开始时路径排序的计算量。

4 实验结果和讨论

语音数据的信号采样率为 16kHz，每个采样值为 16bit 量化。语音帧长为 25ms，帧移为 10ms。测试语音数据库为出租车对话系统的常用会话，说话人为 12 人，8 男 4 女，每人说 20 句话，共 240 句话，其中有 80 句话还带有噪声。

表 2 为特征系数屏蔽实验的测试结果，其中贡献较少的系数是指 C12，D10~D12，A5，A6，A8~A12 等 11 个贡献比率较低的特征系数，计算时间栏是以 39 维特征矢量的基线系统的计算时间为基准。从在嵌入式系统上的实验结果可以看出，与基线系统相比较，屏蔽贡献较少的系数反而让识别率提高了 0.81%，而同时又能把计算时间下降 10.54%，减少了计算复杂度。对于本系统而言，相比起其它几种舍弃特征系数的方式，这种方法保留了贡献较大的系数，舍弃了贡献较少甚至容易降低识别率的系数，更加准确快速。

表 2 特征系数屏蔽实验

屏蔽的特征系数	基线系统	13 维二阶差分	13 维一阶差分	26 维差分	贡献较少的系数
计算时间(%)	100	87.77	87.77	76.31	89.46
系统识别率(%)	89.65	85.24	72.66	0.29	90.46

表 3 为搜索剪枝实验的测试结果。其中，“L”表示对数概率剪枝，“N”表示路径总数剪枝，其后的数字为阈值大小，基线系统不使用搜索剪枝。综合运用两种剪枝技术的搜索算法大大地减小了搜索空间，使得在嵌入式系统上的识别速度有了大幅度的提高，相比较基线系统而言，计算时间的减少超过了 70%，识别率最好的情况比基线系统仅下降了 0.26%。

表 3 搜索剪枝实验

剪枝参数	基线系统	L400 N780	L400 N680	L400 N340	L200 N340	L200 N150	L150 N340
计算时间(%)	100	28.82	28.08	25.04	23.22	21.67	22.34
系统识别率(%)	89.65	89.42	89.36	85.29	82.45	77.76	81.81

表 4　综合方法的实验结果

剪枝参数	基线系统	L400 N780	L400 N680	L400 N340	L200 N340	L200 N150	L150 N340
计算时间(%)	100	27.91	27.35	24.82	22.97	21.40	22.12
系统识别率(%)	89.65	89.41	89.04	85.48	83.56	79.75	83.03

表 4 是综合特征系数屏蔽和搜索剪枝两种方法的实验结果，为了便于对比，其中搜索剪枝的参数设置与表 3 一样。与单用剪枝的方法相比较，综合方法的计算时间普遍减少。表 3 中，部分剪枝实验的参数设置过严，导致识别率有所下降，而从表 4 的实验结果来看，它们的句子识别率又有所回升。由此可以看出，综合了特征系数屏蔽的方法能够帮助这部分剪枝参数过严的识别系统提高句子识别率，同时计算时间也减少了。

5　结论

本文研究说明，在嵌入式平台实现 HMM 非特定人连续语音识别，下面的技术可以帮助我们在提高计算速度和保持识别率之间取得适当的平衡：（1）采用特征系数屏蔽的方法，舍弃贡献比率小的特征系数，可以大幅度降低计算复杂度，同时保持或微降系统的识别率。（2）采用综合剪枝的方法，将路径剪枝和束剪枝相结合，可以有效控制搜索路径的数目，大幅度降低计算复杂度，同时保持或微降系统的识别率。（3）特征系数屏蔽与综合剪枝相结合，可以更加有效地降低语音识别的复杂度，同时保持或微降系统的识别率。

本研究是将 HMM 非特定人连续语音识别技术从实验室环境移植到实时应用环境的一次有益尝试。上述研究结果的意义是积极的，但还需要进一步优化性能，提高鲁棒性才能达到实用化的目标。

参 考 文 献

[1] Du Limin, Feng Junlan, Song Yi, Sun Jinchen. A Chinese-English speech translation prototype system: CEST-CAS1.0. ICSPAT'99, Orlando, USA, 1999.

[2] Du Limin, Feng Junlan, Song Yi, Wang Heng. Speech translation on internet CEST-CAS2.0. Proc. of ISIMP2001, Hong Kong, 2001: 189 – 192.

[3] Rabiner L, Juang B H. Fundamentals of Speech Recognition. New Jersey, USA, Prentice Hall, 1993: 350 – 352.

[4] Ney H, Ortmanns S. Dynamic programming search for continuous speech recognition. *IEEE Signal Processing Magazine*, 1999, 16(5): 64 – 83.

杜利民：男，1957 年生，博士，1996 年美国麻省理工学院（MIT）访问科学家，1999 年美国 AT&T 访问研究员，主任研究员，主要从事信号处理和语音交互技术的研究.

谢凌云：男，1977 年生，博士生，研究方向为语音识别的快速算法和优化.

刘　斌：男，1979 年生，博士生，研究方向为语音识别的实时系统实现.