第17卷 第6期 声 学 学 报 Vol. 17, No. 6
1992 年 11 月 ACTA ACUSTICA Nov., 1992

认人的限定主题的连续汉语语音识别系统的研究

林道发 杨家沅 罗万伯 王跟东
(四川大学模式识别研究室，成都 610064)

1991 年 4 月 15 日收到

摘要 本文描述一个基于矢量量化 (VQ)、隐马尔可夫模型和有限态文法的认人的限定主题的连续汉语语音识别系统. 引入跨零幅度差函数作为判定语音有无的特征参量之一，HMM 训练用的各单个词语的语音数据由连续话句的语音数据经自动切分而得，识别过程中，每帧都考虑多个可能过渡到其它模型的文法节点. 这些技术措施显著地提高了识别系统的准确率. 这类系统能用于特定人操作的、特定主题的信息查询任务. 待进一步解决非特定人的连续语音识别问题后，可用于特定主题的公用信息查询系统.

Study on continuous Chinese speech recognition system for speaker-dependent and topic-constrained applications

LIN Daofa, YANG Jiayuan, LUO Wanbo and WANG Gendong
(Sichuan University Pattern Recognition Lab., Chengdu)

Received April 15, 1991

Abstract In this article a speaker-dependent, topic-constrained and continuous Chinese speech recognition system based on vector quantization (VQ), hidden Markov models (HMM's) and finite states syntactic analysis is studied. The difference of zero-crossing amplitudes was used as one of parameters to determine the position of voice start and end. The words-data used in HMM's training was obtained by machine segmentation of continuous spoken sentences. Multipossible transitions of models nodes were considere in each frame. So the recognition accuracy was improved. The system can be applied in speaker-dependent and topic constrained information request.

一、引 言

近年来，在语音识别这个研究领域中，孤立字词识别技术日趋完善，连续语音识别越来越受到人们的重视. 在国内开始出现一些研究结果. 俞一彪等[1]采用预先对汉语音节进行切分，对音节进行识别得出二维音节序列格形表，再利用句法结构知识对格形表进行搜索，得出话句识别的最后结果.

对汉语来说，由于字的单音节特性，各字音的分割比起西方语言各单词的分割容易性，预

先切割出单个汉字字音再进行识别，无疑是一种值得探索的方案．但在说话较快时，有些相邻字音连成一片，要准确可靠地切分是相当困难的．特别是，若分割中出现“漏字”或“冒字”的情况，往往会使最后识别结果差错很大．本文讨论不预先把待识句进行切分而实现的连续汉语语音识别，介绍识别系统的组成，讨论提高性能的技术．

二、语音识别系统概况

本语音识别系统是一个中等词汇量、特定话者的连续语音识别系统，是针对火车订票及信息查询这一特定主题而设计的．按该主题的常用话句，拟定句式及简化句共 17 种，使用 89 个词语，其中包括 14 个城市名和 10 个数字．通过词汇代换，可构成数万个不同的句子．该系统在 386 微机上用 FORTRAN 和汇编语言实现．

系统的组成如图 1 所示．

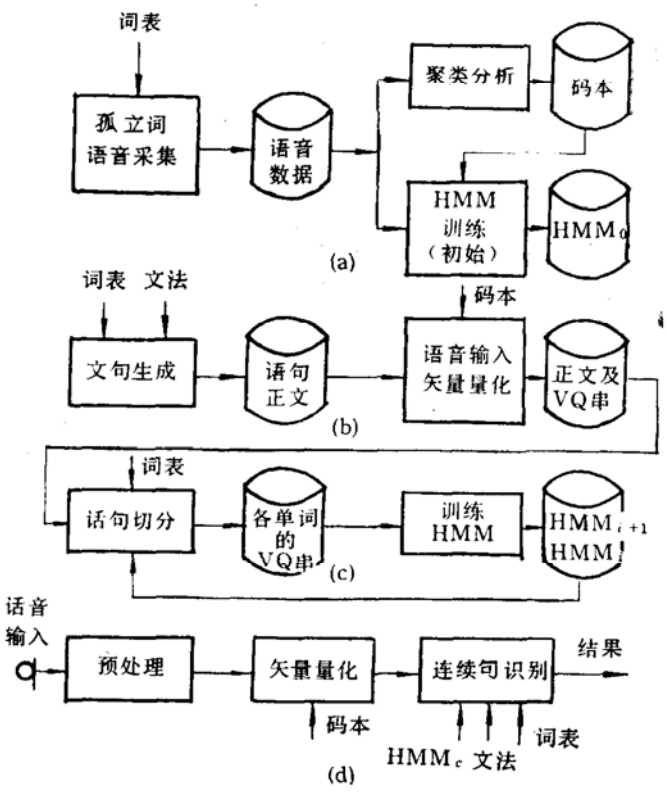

图 1　连续语音识别系统的组成．
(a) 用孤立词语音作成码本及 HMM 初值；(b) 训练用连续句话音数据的获取；(c) 连续句切分及 HMM 训练；(d) 连续语音的识别

系统由四个部分组成．第一部分使用孤立词的语音数据来作成系统的码本及 HMM 的初值 HMM_0，这个部分基本上与我们原有的英语话句识别系统[2]中的训练部分相同．第二部分是 HMM 训练用的连续语音的获取及存储，首先根据词表和句式及对训练用数据的要求生成一批文本 (text) 形式的语句，然后由讲话人以连续自然的方式口述这批句子．由于计算机的存储容量有限，我们把这些话音数据进行矢量量化，只存储每帧语音经量化之后的结果．第三部分用记录的连续话音的 VQ 符号串进行 HMM 的训练，首先把连续句进行自动切分，把各单词的 VQ 串分别汇总，然后把每个词语的多个 VQ 串，用 Baum-Welch 算法[3]进行 HMM 的训练．如图 1(c) 所示，切分和训练以迭代方式进行，用 HMM_i (最初 $i=0$) 对 VQ 串进行切分，经训练得 HMM_{i+1}，然后用 HMM_{i+1} 代替 HMM_i，再次进行切分和训练．按此方式进行迭代，直到获得一个较适宜于表征连续话音中的单词的模型 HMM_e 为止．从我们的实验看，迭代两三次就足够了．第四部分就是进行识别的部分，它接收连续自然的语音，话者发音速度约为每秒钟 5 个音节．语音信号经预处理后，用码本对其进行矢量量化，得 VQ 串，用连续识别算法对 VQ 串进行处理，得出识别结果．从发音结束到输出识别结果之间的时间间隔约为发音时间的 60%．

以下对本连续语音识别系统中的几个重要技术措施进行讨论.

三、无话音段的切除

在识别系统对话句进行处理时，要预先把无话音段切除掉. 一般使用短时能量和短时过零率来进行. 但是过零率是一个很不稳定的参数,容易受各种因素的影响,如环境噪声、电路噪声、话筒灵敏度及模拟通道的增益变化都可能会使过零率发生较大的变化，难于掌握应用. 为此很多研究者都进行了探索，提出了多门限过零率、能频值等特征参数作为语音切割的判据. 我们引入短时跨零幅度差 Z_M 作为分割的特征参数，配合短时振幅函数使用. Z_M 定义为:

$$Z_M = \sum_{s_i与s_{i+1}异号} |s_i - s_{i-1}|$$

Z_M 可以综合地反映信号中有无语音，而且计算量较小. 容易看到，在清音段由于过零次数多，Z_M 会较大,在浊音段,由于过零点附近波形升降速度较高，s_i 和 s_{i-1} 的差值较大,也使 Z_M 较大. 在无话音段,信号幅值较小,而且过零次数不多，Z_M 较小. 我们用短时平均振幅和短时平均过零幅度差作为有无话音的判定参数,较好地解决了无话音段的切除问题.

四、连续句的识别

用各个词语的 HMM_c 相连接构成图 2 所示网络，对这个网络实行帧同步型最优状态转移路径的搜索,实现对连续句的识别.

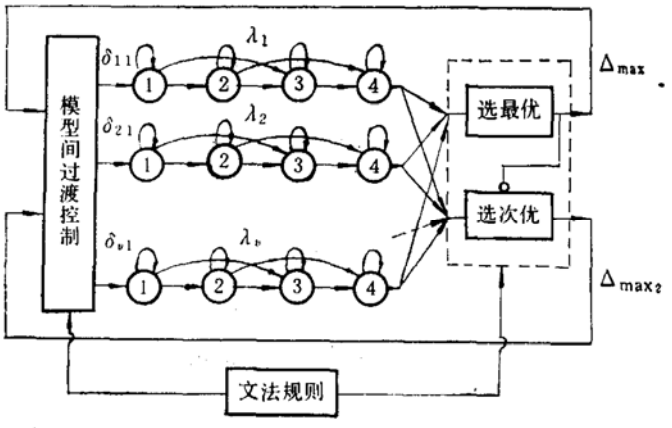

图 2 HMM 状态转移网络

在图 2 的网络中，有与每个词表项 w_i 相对应的隐马尔可夫模型 λ_i 的 N 个状态及状态间的转移线. 各个模型的第 N 态在选优逻辑及有限态文法[2]的控制下，有可能过渡到各个模型的第 1 态. 这种过渡就构成一个可能的单词切分点. 用 Viterbi 算法[3]对图 2 所示网络进行处理,在输入VQ串的全部符号处理完后，选取符合文法规则的终态概率最高的模型，再经过回溯就获得识别的结果.

在图 2 的 HMM 状态转移图中,模型之间的过渡受有限态文法的控制，"选最优"部分在 V 个 HMM 中选出第 N 态概率最高(记为 Δ_{max}) 的模型 λ_{max}，认为它最可能转移到其它单词的 HMM 的第 1 态. 在下一时刻,若 Δ_{max} 大于某 δ_{i1}，且文法规则允许(若 λ_{max} 与 λ_i 分别是某文法节点 D 的前后迁移线)的话,就发生由 λ_{max} 的终态向 λ_i 的首态的转移. "选次优"部分也实行类似的操作,只不过它仅对那些不是文法节点 D 的输入模型者进行选择,也就是说这些候选模型与后接词的模型之间的文法节点不是 D.

采用这种多次选优技术，大大减少了“拒识”的可能性，提高了识别的准确性。

五、从连续句中切分出训练用的语音

如图 1(b)(c) 所示，先对输入的连续句话音进行矢量量化，得出 VQ 串，再利用已有的近似模型 HMM_i 实现对 VQ 串的切分。从状态转移的意义上看，这里实现的过程类似于图 2 所示的识别阶段的过程。与识别的差异是，此处具有一个有利条件，就是我们预先知道这个 VQ 串所对应的话句的内容，我们利用这个条件来实现其较良好的词语分割。

设待切分的话句包含 K 个词语，记为 $w_{s(1)}$、$w_{s(2)}$，…，$w_{s(K)}$。我们对此话句可画出 HMM 状态迁移图如图 3 所示。它共有 KN 个状态，模型间的过渡只发生在 $\lambda_{s(i)}$ 的第 N 态向 $\lambda_{s(i+1)}$ 的第 1 态过渡。对这个图的 NK 个状态使用 Viterbi 算法，寻求出一个最优状态转移路径，判定出此最优路径内发生模型间过渡的时刻，就求出了句子中各个词语的切分点。

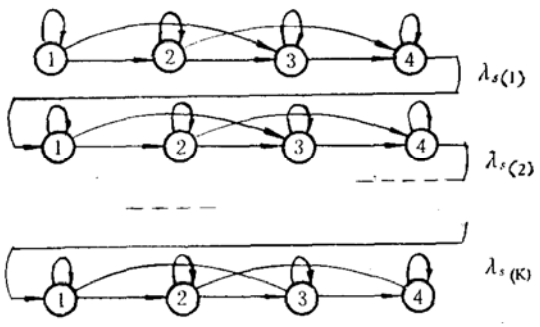

图 3　连续句切分处理的 HMM 状态图

六、实 验 结 果

用两组话句的 VQ 串对识别部分进行了测试。第 1 组话句有 313 个句子，共包含 2838 个词语，其中部分词语经自动切分后用于训练 HMM_c。第 2 组句子是专用于测试的，有 233 个话句的 VQ 串，含词语 1781 个。识别部分对每个句子的 VQ 串进行处理，得到识别结果，把这个结果和已知的话句文本用一个动态规划型字词对准算法进行对比，计算出误识的词语数，若对一个句子未得出识别结果，则认为每个词语都识别错了。测试结果如表 1 所示。由表

表 1　对识别系统的测试结果

话句组号	帧移/帧长	模型	词语误识率(%)		
			选最优	二次选优	三次选优
1	1/3	HMM_0	43.0	19.5	12.8
		HMM_c	6.9	1.7	1.7
2	1/3	HMM_0	47.4	18.5	14.1
		HMM_c	13.3	2.9	2.9
	2/3	HMM_0	73.4	33.2	23.5
		HMM_c	22.9	3.0	2.9
	3/3	HMM_0	80.6	61	40.7
		HMM_c	35.4	13.1	6.6

1 可观察到下列情况:

1. 由于第 1 组话句中的部分单词被用于对 HMM 进行训练,因此对这组句子的识别错误率较低。对第 2 组话句,使用 HMM_c 和两次选优时,词语误识率为 2.9%。

2. HMM_c 比 HMM_0 的误识率低得多。

3. 对于 HMM_0, 三次选优可显著降低误识率;对于 HMM_c 而言,除帧移等于帧长的情况之外,三次选优比起二次选优无明显提高。

4. 帧移由 1/3 帧长加大到 2/3 帧长时,误识率变化不大,但帧移等于帧长时, 性能显著变坏。

七、结 束 语

对于认人的、限定主题的、中等词表的受有限态文法约束的连续语音识别系统而言, 采用本文描述的技术可以使识别准确率得到大幅度的提高。首先, 使用连续话句进行切分得到隐马尔可夫模型的训练数据,可使 HMM_c 更能代表连续语音中各单词的特征。第二, 在识别过程中,在用有限态文法控制下的网络搜索过程中,使用"多重选优"算法可以大大降低拒识和其它误识的发生。第三,我们的初步试验表明,在语音信号的预处理阶段, 短时平均跨零幅度差对于判定语音的有无是一个较好的参量。最后,适当加大帧移,可以在不显著降低识别性能的条件下取得加快处理速度的效果。

参 考 文 献

[1] 俞一彪、袁保宗,"连续语音识别中句法结构知识的利用",电子学报,**18**(1990), No. 6,68—74.

[2] 林道发、罗万伯、杨家沅, "用矢量量化和隐马尔可夫模型实现英语话句的识别", 四川大学学报, **28**(1991), No. 3, 296—301.

[3] Rabiner, L. R., *Proceedings of IEEE*, **77**(1989), No. 2 257—285.