

不认人可扩展词表汉语语音识别系统

何晓冬 周健来 刘建 俞铁城

中国科学院声学研究所

北京 2712 信箱 5 室 100080

Email: {hexd, zjl, lj, tcyu}@speech1.ioa.ac.cn

摘要:

本文介绍了一个基于隐马尔可夫模型 (HMM) 的不认人词汇量汉语语音识别系统。本系统具有词表可动态扩展的特点。可根据需要修改词表, 无须重新训练, 即可应用于不同的专业领域中。文中描述了系统的框架, 并针对在实际中出现的协同发音和声韵过渡问题, 根据汉语发音的特点, 从语音模型, 识别单元选取, 训练和识别等方面给出了解决办法。文中给出的系统测试结果表明, 在 1200 词汇量时, 其识别率高于 92%。在 800 词汇量时, 其识别率接近 98%。

SPEAKER INDEPENDENT VOCABULARY EXTENSIBLE CHINESE SPEECH RECOGNITION SYSTEM

Abstract:

In this paper, a HMM based speaker independent middle vocabulary Chinese speech recognition system is introduced. It has a extensible vocabulary, which can be changed easily according to different applications without training. The main framework of the system was described. With the consideration of the character of Chinese, we gave the solution for the effect of coarticulation and the transition between voice and unvoiced sound in practical speech respectively from phonetic rule, selection of recognition units, training and recognition. The test result, based on 1200 words spoken by 10 speakers, showed that the recognition accuracy is more than 92%. When the size of vocabulary is 800 words, a higher correction rate, approximate 98%, can be achieved.

关键字: 语音识别, 隐马尔可夫模型, 可扩展词表

Key words: Speech Recognition, Hidden Markov Model, Extensible Vocabulary

1. 引言

汉语语音识别技术已有近五十年的发展历史, 早期的语音识别系统的研究主要集中在中小词表, 特定人的语音识别技术。经典模式识别中的聚类技术于模板匹配技术被应用于语音识别领域, 而识别技术的研究主要集中在如何解决在时域上对不定长的语音信号进行规整的问题。比较成功的算法有动态时间规整算法 (DTW) [1], 声刺激法[2]等。但在不认人连续语音的识别方面, 以上算法结果并不理想。自七十年代以来, Baker[3]和 Jelinek[4]等人将隐马尔可夫模型 (HMM) 技术引入语音处理中, 并在大词汇连续语音识别问题上取得了突破。近十年来, 隐马尔可夫模型 (HMM) 技术被广泛地应用

于各种语音命令系统，听写机，语音交互系统中，并取得了成功。

我们在文中介绍了一个基于 HMM 的不认人，词表可动态扩展，中词汇量汉语孤立词识别系统。在本文的第二节描述了系统的基本结构，且详细给出了系统各部分的实现。在第三节中给出了系统的测试结果。并在文章最后给出一个结论。

2. 系统构成

本识别系统的结构见图 1。

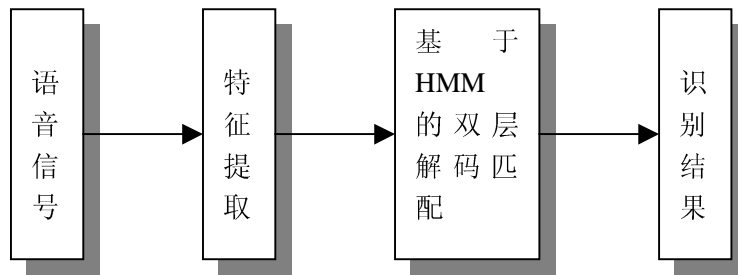


图 1. 系统结构框图

2. 1 特征参数的选取

我们在系统中采用 LPC-CEP 作为语音特征参数。其有计算快捷，可直接采用欧氏距离测量等优点。本系统中 LPC 的阶数定为 18，CEP 的阶数定为 14。这是因为系统中语音信号的采样率为 16KHz，信号的频宽大致在 7~8KHz 之间，要想真实地描述信号的频谱包络，LPC 的阶数需要高些，但 LPC 的阶数也不宜太高，否则信号的低频区域的谐波可能拆分过细，增加了一定的激励源特性描述，反而使系统的识别率降低。

2. 2 识别单元的选取

由于本系统有可动态扩展词表的要求，不可能以整个词作为一个模板来建立隐马尔可夫模型，而应考虑以更小的单元，如全音节或半音节（声母，韵母），为基本识别单元来建模。虽然以全音节为基本单元建模可更精细地刻画声韵母间的过渡信息，但考虑到全音节的数目比半音节数多好几倍，它要求有非常充分的训练语料，否则将使识别效果大大降低。因此我们决定采用半音节为基本识别单元。

2.2.1 原始语音单元：

韵母模型 38 个

a) 单元音：9 个

a, o, e, u, v, i, i1, i2, er

其中 'i2' 代表知韵的 'i'，'i1' 代表资韵的 'i'。

b) 复合元音 13 个：

ai, ei, ao, ou, ia, ie, iao, iu, ua, uo, uai, ui, ve

其中：'iu' 代表 'iou'，'ui' 代表 'uei'。

c) 复合鼻元音 16 个：

an, en, ang, eng, ong, ian, in, iang, ing, iong

uan , un , uang , ueng , van , vn

其中：‘un’代表‘uen’，‘weng’代表‘ueng’。

声母模型 21 个：

b, d, g, p, t, k, z, c, s, zh, ch, sh, j, q, x, f, h, r, m, n, l。

寂静模型 1 个：

h#。

2.2.2 声母的细化：

一般来说，在字词连读时存在协同发音的现象，声母受所接韵母的影响而变化，因而要求声学模型能够描述这种声韵过渡信息。一种常用的方法是对每一种可能的声韵组合都相应的建一个过渡段模型，但这样将使得识别单元过多，降低识别速度，使识别过程复杂化。这里我们采取了一种折中的办法。在 14 个声母模型中，b, d, g, r, m, n, l, 等七个音发音很短，或者与其后所接的韵母之间没有明确的分界，其受后接韵母的影响很大，因此我们应该根据其后接韵母来对其进行细化，并把过渡信息加入模型中。考察汉语的韵母共有 38 个，按其起始发音特点可分为四类[5]：开口呼类（以 a, o, e 开头的以及 er, -i 在内），齐齿呼类（以 i 开头的），合口呼类（以 u 开头的）和撮口呼类（以 v 开头的）。其中还有两个例外：ong 算合口呼，iong 算撮口呼。因此我们根据这七个声母与这四类韵母的组合对声母细化，细化声母均带有过渡段信息。

考虑有几种组合只有极少的音，如 b 与合口呼类组合只有一个合法拼音“bu”，这样将使这个音得不到充分训练，我们把这些细化声母并入其他相近的细化声母模型中。最后得到细化声母如下：

‘b1’（‘b’+开口呼），‘b2’（‘b’+齐齿呼），（‘bu’组合中的‘b’并入‘b1’中。）

‘d1’（‘d’+开口呼），‘d2’（‘d’+齐齿呼），‘d3’（‘d’+合口呼），

‘g1’（‘g’+开口呼），‘g3’（‘g’+合口呼），

‘l1’（‘l’+开口呼），‘l2’（‘l’+齐齿呼），‘l3’（‘l’+合口呼），

（‘lv’和‘lve’中的‘l’并入‘l2’中。）

‘m1’（‘m’+开口呼），‘m2’（‘m’+齐齿呼），（‘mu’组合中的‘m’并入‘m1’中。）

‘n1’（‘n’+开口呼），‘n2’（‘n’+齐齿呼），‘n3’（‘n’+合口呼），

（‘nv’和‘nve’中的‘n’并入‘n2’中。）

‘r1’（‘r’+开口呼），‘r3’（‘r’+合口呼），（‘ri’组合中的‘r’并入‘r1’中。）

最后我们一共选取了 70 个基本声学识别单元。考虑到开口呼与合口呼类，齐齿呼与撮口呼类发音相似，容易混淆，我们允许它们互换，并采用多词典（multi-Lexicon）语音模型来描述这种互换的可能性。

2.3 隐马尔可夫模型的选择

隐马尔可夫模型主要有三类：离散隐马尔可夫模型（DHMM），连续隐马尔可夫模型（CHMM），和半连续隐马尔可夫模型（SCHMM）。一般来说，在训练语料充分的情况下，连续隐马尔可夫模型对语音信号的刻画较佳，能获得较高的识别率，但由于它所估计的参数较多，因此当训练语料不足时，将出现退化。另一方面，连续隐马尔可夫模型的计算量非常大，将严重影响语音识别系统的实时性，而离散隐马尔可夫模型对语料的要求较低，并且计算量很小，比较适于用在实时系统中。因此，在本系统中，我们采用离散隐马尔可夫模型，码书的大小为 256。

2.4 训练方法

由于采用半音节而非整个词条作为识别单元，我们采用了对连续语音的 segmental K-means 训练方法，训练步骤如下：

1. 训练寂静模型：得到刻画背景噪声的 **Silence** 模型。
2. 初始化语音模型：如前述，为了采用多词典语音模型，需要一个初始模型，以分辨同一个音素的不同发音。在这里，我们利用标住信息按孤立词训练方式生成一套初始模型。
3. 模型迭代：采用 **Segmental K-means** 方法训练，迭代过程中采用多词典和 **silence-option** 的解码方法。

2.5 识别

令 A 代表声学特征向量， Γ 代表所有可能识别结果， W 代表语言模型，则识别结果可表示为：

$$\operatorname{argmax}_{W \in \Gamma} P(W|A) = \operatorname{argmax}_{W \in \Gamma} \frac{P(A|W)P(W)}{P(A)}。$$

为了求 $P(A|W)$ ，采用了基于 **HMM** 的多词典双层解码算法，其结构见图 2：

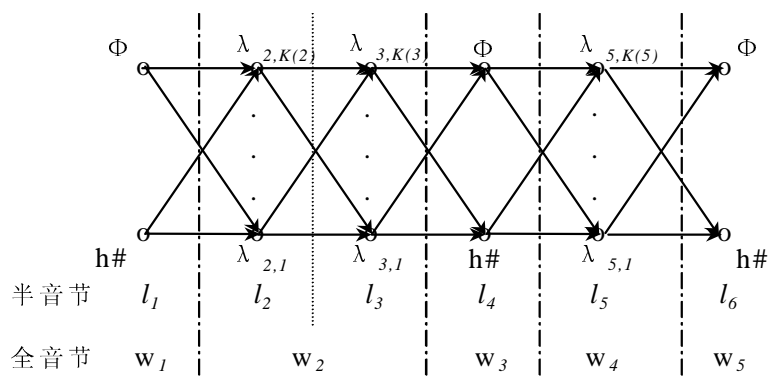


图 2. 基于 HMM 的多词典双层解码算法
(其中 Φ 为空模型，不产生实际输出)

我们采用的识别算法为一个限于此框架的帧同步搜索算法。

在孤立词识别问题中，可以认为 $P(W)$ 与 $P(A)$ 均为等概率的，所以识别结果为：

$$W^* = \operatorname{argmax}_{W \in P} P(A|W)$$

这实际上是一个在严格语言模型限制下的连续语音识别策略，词表中每一个词条表示了一个有限状态网 (FSN) 中的一条或几条路径。

2.6 词表的扩展

由于以半音节为基本识别单元，并采用了类似连续语音识别方法的识别策略，本系统有非常好的扩展性，可根据需要随意增加、删除和修改词表，而无须生成新的模板和进行新的训练。扩展词表的方法如图 2：

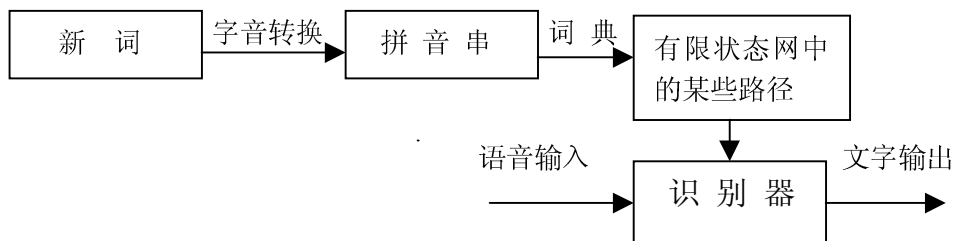


图 3. 词表的扩展框图

3. 测试结果

我们对上述系统进行了测试，词表大小为 1200 词，语料为 50 男 50 女，每人 1200 个词，其中 40 男 40 女的语料用于训练，其余作为测试集。测试平台为 Pentium II 300MHz。为了测试词表大小的影响，我们从这 1200 词中抽出一些词，组成几个大小不同的词表，分别给出了测试结果。关于词表信息见表 1，测试结果见表 2。

词表大小	500 词	800 词	1000 词	1200 词
二字词数	88	333	524	701
三字词数	130	174	182	203
四字词数	282	293	294	296

表 1. 词表信息

词表大小	500 词	800 词	1000 词	1200 词
首选正识率	99. 1%	97. 5%	94. 8%	92. 3%
前五选正识率	99. 7%	99. 6%	99. 4%	99. 3%
前十选正识率	99. 9%	99. 9%	99. 9%	99. 8%
平均每句耗时	0. 6 秒	1. 4 秒	2. 2 秒	3. 2 秒

表 2. 系统测试结果

4. 结论

以上我们描述了一个不认人、可扩展词表、中词汇汉语语音识别系统。本系统的最大特点是用训练连续语音识别系统的方法来构造命令识别系统，从而获得非常大的灵活性，并且在某些特定的场合和工作条件下，可得到比通用连续语音识别系统高得多的性能。用这种方法作为系统内核，可以构造多种专门用途的语音识别系统或应用。

参考文献:

- [1] H. Sakoe, S.Chiba. "Dynamic Programming Algorithm Optimization for Speech Word Recognition", *IEEE Trans. ASSP-26*, pp.43-49, 1978
- [2] 俞铁城, "用图样匹配算法在计算机上自动识别语音", 《物理学报》, 1977 年 9 月, Vol. 20, No.5
- [3] J.K. Baker, "The dragon system-An overview", *IEEE Trans. Acoustics, Speech, Signal Proc.*, ASSP-23(1):pp.24-29, Feb. 1975
- [4] F. Jelinek, "A fast sequential decoding algorithm using a stack", *IBM J. Res. Develop.*, 13: pp.675-685, 1969
- [5] 陈永彬, 王仁华, 《语言信号处理》, 中国科学技术大学出版社, pp.37-65, 1990