基于汉语语音特点的大词表语音识别系统的研究

李建民　赵彤青　郑　方　方棣棠　吴文虎

（清华大学计算机科学与技术系语音实验室，北京 100084）

APPROACHES TO LARGE-VOCABULARY SPEECH RECOGNITION BASED ON CHINESE SPEECH CHARACTERISTICS

Li Jianmin, Zhao Tongqing, Zheng Fang, Fang Ditang and Wu Wenhu

(Speech Laboratory, Dept. of Computer, Tsinghua University, Beijing 100084)

Abstract This paper discusses some approaches to Chinese speech recognition and briefly introduces a practical Large-Vocabulary Chinese Connected speech recognition system. In this system, some characteristics of Chinese speech are taken into consideration, such as Chinese being syllable language, the syllable simply composed of initial consonant and final vowel, and Chinese words being the basic units in conversation. One of the most important characteristics of this system have a convenient addition of new words, which makes the system have a good user interface. With more than 10,000 words, the system has achieved an average accuracy of 93.1% (top candidate).

Key words Syllable segmentation, short-time energy-frequency-value, segmental vector quantization, segmental probabilitical model.

摘要　本文探讨了汉语语音识别的若干问题，并简单介绍了一个大词表汉语语音识别系统，该系统充分考虑了汉语语音的特点，其中主要是汉语语音具有音节性比较强的特点、音节的简单声韵母结构以及汉语以词/词组为语音交流基础的特点。该系统一个显著的特点是系统可以不进行任何训练地添加新词汇，从而使得系统具有比较好的用户接口。

现在系统具有 10,000 多个词汇，实时测试的平均识别结果是 93.1%。

关键词　音节切割，短时能频值，分段矢量量化，分段概率模型。

本文 1990 年 9 月 18 日收到。本课题得到75重点攻关项目基金资助。李建民，清华大学在读博士生，从事大字表汉语语音识别等方面的研究工作。

一、汉语语音识别的若干问题的研究

1. 汉语语音建模单位的选择

汉语语音建模单位通常有声韵母、音节和词这三种．虽然声韵母的个数比较少，但由于声韵母包含的信息太少，声韵母的分割十分困难，同时声韵母的分割又损失了声韵母的关联信息，所以汉语语音建模单位很少选择声韵母；以词为汉语语音建模单位的系统虽然有很高的系统性能，但词的个数太多，很难扩大系统的词表，所以以词为汉语语音建模单位的系统一般都是中小字表，它不适合作为大词表的语音建模单位．汉语音节则是一个很好的语音建模单位，汉语音节的个数适中，同时又是汉语发音的基本单位，比较适合于汉语大词表的语音识别系统．

汉语是一个音节性比较强的语言．汉语的音节（一个汉字对应着一个音节）是汉语发音的基本单位，而词则是汉语语言交流的基本单位；汉语的音节包括音和调两部分．汉语中有着成千上万的汉字和词，然而它们所对应的音节却只有 1200 个左右，如果不考虑汉语声调（无调音节），则只有 400 个左右的无调音节，而且现在汉语声调的识别已取得了比较满意的结果．汉语的词几乎全部是由一至四个汉语音节构成的，分别为单音节词、双音节词、三音节词、四音节词．另外，由于汉语音节性比较强的特点，虽然词是汉语语言交流的基本单位，词以连接词的方式（音节之间稍有停顿）输入并不是不可接受的．基于汉语的这些特点，我们选择了汉语无调音节作为语音建模单位，设计了以音节识别为基础、以词识别为主的汉语语音识别系统．这样连接词汉语语音识别系统结构的提出，必须要解决一个关键的问题，即汉语音节之间的协同发音是否导至连接词识别的失败？我们认为音节之间的协同发音会影响连接词的识别，但其影响不是很大，连接词词法信息的利用弥补了音节之间的协同发音的影响．我们这个大词表语音识别系统的研制成功，说明了这种连接词汉语语音识别的系统结构是一个可行的、很有吸引力的方法[1,3,10]．

开始时，我们选用了出现在小学一到五年级统编教材中出现的所有词汇，其中包括汉语音节 1145 个，单音节词 2075 个，双音节词 4176 个，三音节词 545 个，四音节词 717 个．后来，为了使系统更具有实用性，我们根据常用词的词频统计，选择了 403 个无调音节作为汉语语音建模单位，加上四声判别就形成了 1181 个有调音节的模型，这些模型对应了 6763 个国标一、二级汉字．系统中已有 8940 个多字词．

2. 采样与切字

从系统框图图 1 和图 2 可以看到，采样与切字是系统比较关键的部分，因为该系统是基于音节识别的连接词系统．要保证系统的可靠性，首先应要求采样对环境有较强的适应能力，其次应保证切字的正确率．本系统在采样中引入了学习机制，也就是说，系统能自动统计环境噪声的特性，并利用统计结果来自动确定语音采样阈值．这样就在一定程度上保证了系统对环境的适应性．

在采样和切字方法的研究中，我们引入了一个新的参数短时能频值（EFV，Energy-Frequency-Value），能频值定义为短时能量乘上短时过零率．因为汉语音节具有简单的声韵母结构，而声母具有比较高的过零率、能量比较低，韵母则具有比较高的能量、过零率

比较低，这样能频值既顾及了声母的高过零率又顾及了韵母的高能量，从而提高了语音信号与背景噪声的分辨力．实验结果表明能频值是一个很好的参数，有着较好的稳定性和较高的语音——噪声分辨力．在传统的采样系统中，一般将能量与过零率单独使用来判别语音的头尾或进行切字．这时，采样的阈值与噪声的均值之比值约为十几倍，而使用能频值时，其比值可高达五六百倍，这样一来，我们就可以很好地从噪声中取出语音段了．使用能频值进行语音端点检测和音节切割使得我们的系统更接近于连续语音识别[10]．

以能频值为基础的采样切字方法，给我们的语音识别系统打下了良好的基础．

3. 语音模型的选择

汉语的音节本身也有特点，汉语的音节具有简单的声韵母结构，正确识别一个汉语音节必须要正确识别其对应的声、韵母部分；音节的声韵母部分在时、频域上的特性是不同的．汉语的韵母部分通常具有比较高的能量、而过零率比较低，并且韵母的持续时间也比较长；相反地，汉语的声母部分通常具有比较高的过零率、而能量比较低，并且声母的持续时间也很短(约 40—70 毫秒)；另外，韵母的信号具有比较好的准周期性，而声母信号通常是非周期性信号，韵母信号要比声母信号稳定的多；另外，在汉语音节识别中，声母部分所起的作用常常受到韵母部分的压制、干扰．因此，韵母部分要比声母部分容易识别，声母的识别是汉语音节识别的主要困难．实验结果也证实了这一结论，我们分别进行了声母(24 个)、韵母(35 个)的识别实验，我们采用了 7*1181 个音节的语音（即七遍全音节语音)，取其中的三遍(每次取一遍)作为待识别语音，而其他的六遍作为训练语音，平均识别结果是声母识别率为 79.87%（前二名为 90.04%，前三名为 92.44%)，而韵母识别率为 91.56%（前二名为 98.37%，前三名为 98.62%)．这样，汉语音节的声母信号在音节识别中的作用应该得以加重和提高，对汉语语音识别的模型选择必须充分考虑研究汉语音节的特点．

由于汉语的韵母部分比较容易识别，于是可以用韵母进行粗识别，缩少语音识别的搜索范围．实验表明，利用韵母进行粗识别时，当搜索空间被压缩到 1/4 时，正确判别率为 99.6%（前十名)，因此，这是一个实用的粗判方法．

近十几年来，矢量量化和隐马尔可夫模型（VQ-HMM，Vector Quantization and Hidden Markov Model）正在语音识别领域中得到广泛的应用，可以说 VQ-HMM 方法是现在语音识别领域中的主要方法，尤其适用于连续语音、非特定人的语音识别．然而对具有简单声韵结构的汉语音节识别，VQ-HMM 不是十分适合，因为汉语音节识别必须同时正确识别声母和韵母两部分，然而 VQ-HMM 的语音模型常常使得韵母部分压制了声母部分在音节识别的作用．考虑汉语音节的特点，我们提出了更适合于汉语音节识别的分段矢量量化和分段概率模型(SVQ-SPM，Segmental Vector Quantization and Segmental Probabilitical Model)，我们认为 SVQ-SPM 比 VQ-HMM 更适合于汉语语音音节的识别，采用 SVQ-SPM 对汉语音节进行建模，可以通过分段矢量量化 SVQ 过程减少韵母信号对声母信号的压制干扰，另外，而分段概率模型 SPM 则可以给声母信号加更大的权，这些都有利于音节的声母信号的识别[9,13,21]．

SVQ-SPM 是一种分段的语音识别方法，比较适合汉语音节的识别．SVQ-SPM 方法建立于语音序列的分段算法基础上，在我们的系统中，我们提出并采用了一种非线性的

等特征变化分段方法（NLP, Nonlinear Partition Algorithm）. NLP 算法是根据特征变化的大小把沿着时间轴上的语音序列分成若干段，其中各段中所包含的特征变化量相等，这样各分段点是相对稳定的，对发音的速度变化并不十分敏感，这一点对 SVQ-SPM 十分重要.

矢量量化 VQ 是一种数据压缩技术. 矢量量化过程就是把矢量映射到某个码字(也叫聚类中心，它是一类相近/相似矢量的代表)，并以此码字的序号来表示输入矢量的过程，进行矢量量化时，码本 CODEBOOK 起着重要的作用. 我们希望码本中的每个码字所对应的矢量集具有比较好的类聚性，能反映出各矢量类的特性. 分段矢量量化 SVQ 是结合汉语音节的特点对 VQ 进行的改进. 就如以上提到的那样，汉语音节具有简单的声韵母结构，而声韵母信号在时，频域上的特性是不一样的；我们认为声韵母的特征参数在特征空间上具有不同的分布，这样我们对各段语音信号分别通过他们自己的矢量量化器进行量化——分段矢量量化. 分段矢量量化使得各段语音信号在具有更好类聚性的特征空间中进行映射量化，从而减少不同性质语音信号之间的相互干扰，尤其是元音信号对辅音信号的干扰、压抑.

分段概率模型 SPM 十分类似于"从左向右"的 HMM，它保留了 HMM 在语音识别中起主要作用的输出概率，而对 HMM 的状态转移概率进行了修改. 在 SPM 中，状态的转移取决于语音序列的等特征变化分段的结果. 在 SPM 中，每段语音序列对应于一个状态. 采用 SPM 方法，一个重要的优点在于 SPM 可以对声母进行加权，从而提高声母部分在音节识别中的作用，最后提高汉语音节的识别率.

假设通过等特征变化分段算法 NLP 把语音音节序列分成相对稳定的 N 段，对应于 N 个 SPM 状态；然后这 N 段分别进行分段矢量量化；假设各段矢量量化器的码本大小都是 M(实际上，各段码本的大小可以不同. 如声母的码本大，而韵母的码本小). 我们把 SVQ-SPM 定义为：

$$\mathrm{SPM} = \{B_1, B_2, \cdots, B_N\}.$$

其中

$$B_i = \{b_i(1), b_i(2), \cdots, b_i(M)\},$$

$$b_i(j) = \frac{\sum_{k=1}^{K} \text{第 } k \text{ 遍发音中第 } i \text{ 块中出现第 } j \text{ 个码字的帧数}}{\sum_{k=1}^{K} \text{第 } k \text{ 遍发音中第 } i \text{ 块中的总帧数}}$$

表示第 i 个状态的输出概率.

从以上 SVQ-SPM 的定义可以看出，对 SVQ-SPM 的模型训练可以使用最大似然比估计 MLE 方法十分简便地进行，而语音识别过程也十分简单. 这样，使用 SVQ-SPM 方法大大地简化了语音训练和识别过程，便于在微机上实现[7,8,10].

4. 构词

上文中我们叙述了汉语单音节识别的一些问题，在我们这个基于音节识别的连接词识别系统中，关键是词的识别.

词识别策略我们采用了由音节连接组词的方法，该方法虽然简单，但很实用和有效．词识别所用到的数据主要有以下几部分：403 个无调音节的音节表及所对应的音节模型 SVQ-SPM，双音节词表(7686 个)，三音节词表(542 个)及四音节词表(725 个)．

为缩小搜索范围，在音节表中记录了该音节是否在上述四个词表中出现的信息(称为音节信息字)．比如音节信息字为 (MSB)＊＊＊＊＊＊0000111010(LSB) 表示该音节在单字词中不出现，在双字词中的第一个字出现，第二个字不出现，而在三音节中所有字都出现，但在四音节词中每个字都不出现．这样，在识别过程中，对应于信息位为零的模型，就可以不进行匹配概率的计算，这样就减小了模型匹配的计算量，提高词识别的速度．

每个词表，除了记录了该词的汉字显示信息外，还记录了词中的每个字与音节表的对应信息，即每个词的每个字都有一个指针，该指针指向该字在全音节表中的索引位置，这指针称为音节索引项．也就是说，词表的内容包括两部分：词组的汉字代码和音节索引项．有了词组的音节索引项，我们就可以进行基于音节的词识别．

设有一发音信号，经过音节切割后判定是 K 个音节的词，它们的语音序列分别为 $S(1),S(2),\cdots,S(K)$．首先进行音节识别，即让语音序列中的每个音节 $S(1)\cdots S(K)$ 分别与音节的 SVQ-SPM 进行匹配，可以由音节识别方法求出匹配的概率：

$$P(S(1)|M(v)),\cdots,$$
$$P(S(K)|M(v)),$$
$$v=1,2,\cdots,403.$$

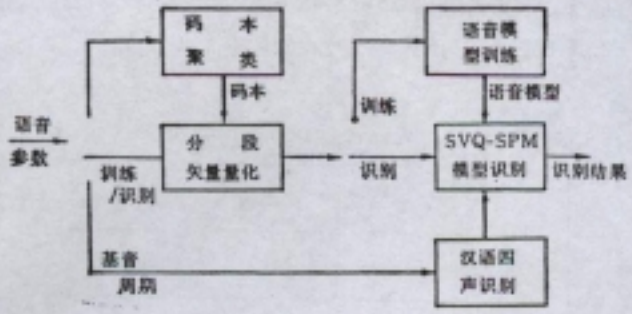

图1 音节识别框图

图2 语音识别系统框图

然后进行词的匹配，K 音节的词表中的每一个词 $C(m)(m=1,2,\ldots,V(K),V(K)$ 是 K 音节词表的大小) 是音节 $C(m,1)$，$C(m,2)$，$\cdots$，$C(m,K)$ 构成的，由词表中的音节索引项可以得到音节 $C(m,i)(i=1,\cdots,K)$ 在全音节表中的位置 $I(C(m,i))$，也就是可以求出语音序列音节中的 $S(i)$ 与词 $C(m)$ 中的第 i 个音节 $C(m,i)$ 的匹配概率 $P(S(i)|C(m,i))$．这样，词表中的每个词 $C(m)$ 产生发音是 $S(1)\cdots S(K)$ 的概率为：

$$P(S|C(m))=P(S(1)|C(m,1))*P(S(2)|C(m,2))\cdots$$
$$=\prod_{k=1}^{K}P(S(k)|C(m,k)),m=1,2,\cdots,V(K).$$

取 $P(S|C(m))$ $(m=1,2,\cdots,V(K))$ 的最大者所对应的 K 音节词表中的词作为识别结果，这样，我们就完成了词的识别，也就完成了构词[9]．

二、大字表连接词汉语语音识别系统简介

该系统建于 GW386 上，外加一块 TMS320C25 板（进行语音采样、LPC-CEP 计算、音节切割、分段矢量量化以及四声判别）和二块 TMS32010 板（进行 SVQ-SPM 样本匹配），该系统的流程框图见图 1 和图 2 .

语音信号以 9.6KHz 的采样率采集，并且用传递函数为 $1-0.97z^{-1}$ 进行高频提升；窗函数采用了 25.6ms 的 Hamming 窗，窗移为 12.8ms；每帧语音信号提取一组 16 阶的 LPC-CEP 参数.

该系统建立了 403 个无调音节的概率模型，模型是 5 状态的 SVQ-SPM，每个状态所对应的分段矢量量化器的码本大小分别是 64，64，64，32，32 个码字. 加上四声判别就形成了 1181 个有调音节的模型，这些模型对应了 6763 个国标一、二级汉字. 系统中已有 8940 个多字词. 由于系统对词识别采用了由音节的识别连接而成的策略，系统所能识别的词汇比系统中现有的词汇要多得多，系统添加新词也毫不困难（只需键入新词所对应的汉字即可，无需对新词进行训练）. 这样，此系统具有较好的实用性和较好的用户接口.

1991 年 2 月 4 日 我们对此“七五”科技攻关项目 75-68-05-22 中的大字表特定人连接词汉语语音识别系统进行了测试，2 月 6 日通过技术鉴定. 系统性能测试如下：

（1）识别率测试：

a. 从 1181 个单音节中任选 100 个进行测试.

第一名	前二名	前十名
80%	90%	96%

b. 从 7686 个二字词中任选 50 个进行测试.

第一名	前二名	前十名
82%	86%	100%

c. 从 542 个三字词中任选 50 个进行测试.

第一名	前二名	前十名
100%		

d. 从 725 个四字词中任选 50 个进行测试.

第一名	前二名	前十名
100%		

被测试的总字数为：

$$50\times4+50\times3+50\times2+100\times1=550$$

平均识别率：

第一名	前二名	前十名
93.1%	95.6%	99.3%

（2）词可以作为整体连呼输入，不需一字一断地呼入.

现在整体连呼还不能说是十分流利地进行，还得要求音节切割正确. 对于比较容易切割的词，如清华大学等，则可以正常呼入，而对于不易切割的词，如教务处等，则要求稍有停顿.

（3）响应时间基本实时.

(4) 词可以方便地任意添加，所添加的新词不需训练。

后来，我们对该系统的四字词词表进一步扩大，四字词词表由《汉语成语小词典》(商务印书馆出版，1990)中的所有成语，共 3170 个成语，四字词的测试结果仍是 100%。这一结果进一步说明了这种连接词汉语语音识别的系统结构是一个可行的、很有吸引力的方法。

三、小 结

在基于音节的词识别中，词识别的正确性除了受到音节识别率的影响，还取决于二个方面，即构成词的各音节之间的协同发音和词法(各音节之间的限制)信息的利用。音节之间的协同发音降低词识别率，而词法信息的利用则提高词识别率。从上面的测试结果中，我们可以看出：二字词的识别率比音节的识别率只是略高一些，而三字词、四字词的识别率则显著提高。这除了三字词、四字词的词表比较小之外，还因为在二字词的识别中协同发音的影响比较大，词法信息还不够；而在三字词、四字词识别中词法信息起主要作用。

我们对特定人大字表连接词汉语语音识别进行了比较深入的研究，研制成功了一个实用的、具有比较好的用户接口的汉语语音识别系统。通过对汉语语音的研究，我们认为 SVQ-SPM 方法因考虑了汉语音节的特点，更加适合汉语音节的识别。又从上述测试结果表明以汉语全音节识别为基础，以词识别为主的汉语语音识别是可行的，而且是很有吸引力的方法。

参 考 文 献

[1] Kai-Fu Lee, Large-Vocabulary Speaker Independent Continuous Speech Recognition: The SPHINX System, Ph. D Dissertation, Computer Science Dept., CMU, Apr. 1988.

[2] Rabiner, L. R., Juang, B. H. An introduction to hidden markov models, *IEEE ASSP Magazine*, 3: (1) (1986), 4—16.

[3] 李建民，基于音节的大字表语音识别系统的研究和实现，硕士论文，清华大学计算机系，1990.

[4] 蒋力，基于概率统计模型的非特定人语音识别方法与系统的研究，硕士论文，清华大学计算机系，1989.

[5] Hao Ying, An Introduction to a Speaker Adaptation Method in Speech Recognition System Based on SPM, Proc. of ICSP'90, 469—472.

[6] 李建民，方棣棠，语音端点检测中门限阈值的自动确定及音节切割的新判据，全国第一届语音识别学术报告与展示会议，1990.

[7] Jiang Li, Wu Wen-Hu *et al.*, A Real-time Speaker-Independent Speech Recognition System Based on SPM for 208 Chinese Words, ICSP'90, 473—476.

[8] Li Jian-min, Fang Di-tang, The SVQ-SPM for Large-Vocabulary Chinese Connected Speech Recognition, ICYCS'91 (International Conference for Young Computer Scientists, 1991 Beijing, China).

[9] 李建民，方棣棠，大字表语音识别系统的设计与研究，全国第三届汉语与语音识别会议，1989.

[10] Li Jian-min, Fang Di-tang, The Newts Speech Recognition System, International Conference on Computer Processing of Chinese and Oriental Languages, Aug., 1991, Taiwan.