第 26 卷 第 4 期
2000 年 7 月

自 动 化 学 报
ACTA AUTOMATICA SINICA

Vol. 26, No. 4
Jul., 2000

汉语连续语音识别系统与知识导引的搜索策略研究[1)]

宋战江　郑 方　徐明星　武 健　吴文虎

（清华大学计算机科学与技术系语音实验室　北京　100084）

（E-mail：szj@sp.cs.tsinghua.edu.cn）

摘　要　从整体上介绍了汉语连续语音识别系统的基本原理，并重点对声学和语言两个层面的建模与搜索策略进行了分析. 在对传统帧同步搜索算法进行研究的基础上，提出了基于统计知识的帧同步搜索算法 SKB-FSS. 它包含了三个主要的功能层次：基于归并的音节切分自动机产生确定的搜索边界点，由统计得到的差分状态驻留信息控制搜索过程中的状态转移，利用词搜索树控制音节候选的扩展规模并根据动态前向预测的方法进行合理而及时的路径剪枝. 实验结果验证了该搜索策略的有效性.

关键词　连续语音识别，基于统计知识的帧同步搜索，差分状态驻留分布.

RESEARCH ON CHINESE CONTINUOUS SPEECH RECOGNITION SYSTEM AND KNOWLEDGE BASED SEARCH STRATEGIES

SONG Zhanjiang　ZHENG Fang　XU Mingxing　WU Jian　WU Wenhu

(*Speech Lab, Dept. of Computer Science & Technology, Tsinghua University, Beijing* 100084)

Abstract　In this paper, the principle of a Chinese continuous speech recognition system is introduced, the modeling and search strategies of its acoustic layer and language layer are also discussed in detail. On the basis of the research on the traditional frame synchronous search algorithm, the statistical knowledge based frame synchronous search (SKB-FSS) algorithm is proposed. It contains three principal functional modules, generating definite search boundaries by a merging based syllable detection automaton, controlling state transitions by the statistical differential state dwell information, and restricting the syllable expansions by a word search tree and pruning unpromising paths by the dynamic forward prediction. The experimental results show the validity of the novel search strategies.

Key words　Continuous speech recognition, statistical knowledge based frame synchronous search, differential state dwell distribution.

---

1) 本文的部分内容曾在第五届(1998 年)全国人机语音通讯学术会议上宣读.

收稿日期　1998-11-09　　收修改稿日期　1999-12-30

1　引言

本文基于清华大学计算机系语音实验室开发的大词汇量、非特定人、连续汉语语音识别系统(又称听写机)EasyTalk[1],研究了连续语音识别的两个最基本部分——声学模型和语言模型.对于输入的连续语音流,EasyTalk 的声学识别层给出由多候选音节串组成的音节网络,然后由语言处理层进行组句分析,得到最终的识别结果.

在声学层面上,用来进行时间对准的搜索策略是一个非常重要的组成部分.采用 HMM[2,3]拓扑结构时,搜索算法基本上可以分为两大类:一类是时间同步的,如著名的 Viterbi 解码算法[4]和帧同步搜索算法[5]等;另一类是非时间同步的,如堆栈解码算法和 A* 搜索算法[6,7]等.此外还有一些更加复杂的搜索算法,如双向图搜索[8]算法等.

堆栈解码算法一般需要一个快速匹配过程来产生候选基元列表,然后再进行路径扩展;A* 搜索算法从理论上可以给出最佳的识别结果,但是其计算量和存储空间的消耗都非常大,而且往往需要一次额外的预搜索来确定 A* 的启发函数.

传统的 Viterbi 解码算法利用动态规划的原理,通过搜索识别基元的最佳状态序列来得到近似最优的目标基元序列.它一般只能给出一个最佳的候选状态序列,当需要多个候选时,就需要进行修改以存储多个回溯指针,而这会使其存储空间和搜索时间成倍增加.

帧同步搜索算法不但简洁高效,而且可以给出多个声学候选,因此在 EasyTalk 中,就采用基于帧同步的搜索算法来产生声学层面的候选音节网络.为进一步提高搜索效率和识别性能,我们提出了基于统计知识的帧同步搜索(SKB-FSS, Statistical Knowledge Based Frame Synchronous Search)算法.它通过基于归并的音节切分自动机产生若干可靠的音节边界点供搜索过程使用,利用状态驻留长度的统计分布和词搜索树来约束搜索过程,通过动态前向预测来进行路径剪枝.本文将重点介绍该搜索策略.

本文各个部分的组织情况如下:第二部分介绍 EasyTalk 的整体结构,包括声学层面的建模和语言层面的处理方法,以及基于归并的切分自动机的实现;第三部分介绍 SKB-FSS 的原理,即如何将状态驻留的统计信息应用于搜索过程中,以及对扩展路径的约束和剪枝策略等;第四部分给出实验结果和分析;第五部分进行了总结.

2　声学模型、语言模型及切分预处理

EasyTalk 核心识别引擎的总体功能结构如图 1 所示.

2.1　声学模型与识别基元

对模型之间距离度量的研究和实验表明,HMM 中状态转移概率矩阵的作用远不如观测概率矩阵重要[9].为此,我们对标准 HMM 模型进行修改[1],去掉了状态转移概率矩阵,仅保留其观测概率矩阵.各状态内部的特征空间采用混合高斯分布进行描述[3],其协方差矩阵采用对角形式.系统以汉语的 418 个无调单音节作为识别基元进行建模.

训练数据库采用 863 汉语普通话连续语音数据库,它是在安静环境下以 16 位精度和 16kHz 采样率进行采样的.语音特征选为 16 阶 Mel 倒谱系数[10]及其自回归分析系数[11],帧长 32ms,帧移 16ms,回归分析宽度为 5 帧.数据库已进行了单音节的手工初始标注.

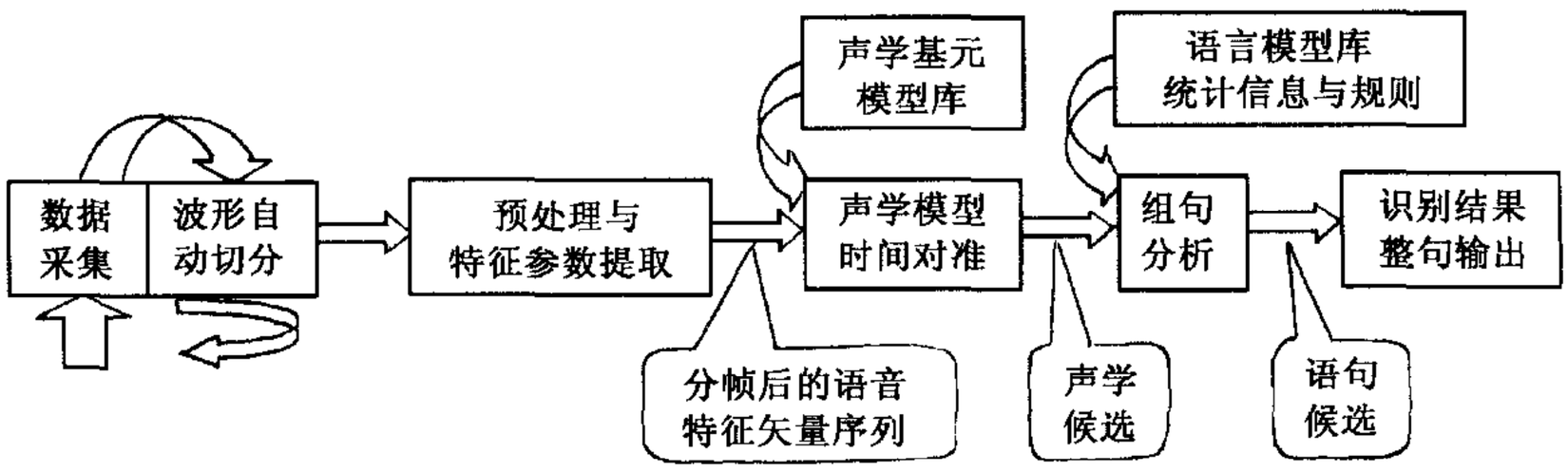

图 1 EasyTalk 核心识别引擎的总体功能结构

在模型的训练阶段，首先采用鲁棒性较高的非线性分段算法[12]给出每个单音节内部各个状态的初始分点，计算出各个状态内特征的初始观测概率密度函数，然后进行迭代，直到各个模型的状态分点达到稳定；而在识别阶段，采用基于统计知识的帧同步搜索算法 SKB-FSS(将在第三部分介绍)来产生声学候选音节网络.

2.2 连续语音流的切分预处理

在声学搜索中有两个问题不容忽视，一是搜索路径的组合爆炸问题，二是解码出的状态序列错位问题. 事实上，由于汉语的连续语音是以词的边界为瞬间间歇的，若能在声学搜索中加入词边界判决信息，就可以在一定程度上缓解上面两个问题，同时部分地解决语言层的分词问题. 因此我们设计了对动态语音数据流进行预处理的切分引擎[13]，它充分利用了声学、语言等方面的统计知识和规则，不断地从语音流中分离出一些相对独立和完整的语音段供后续搜索过程进一步处理.

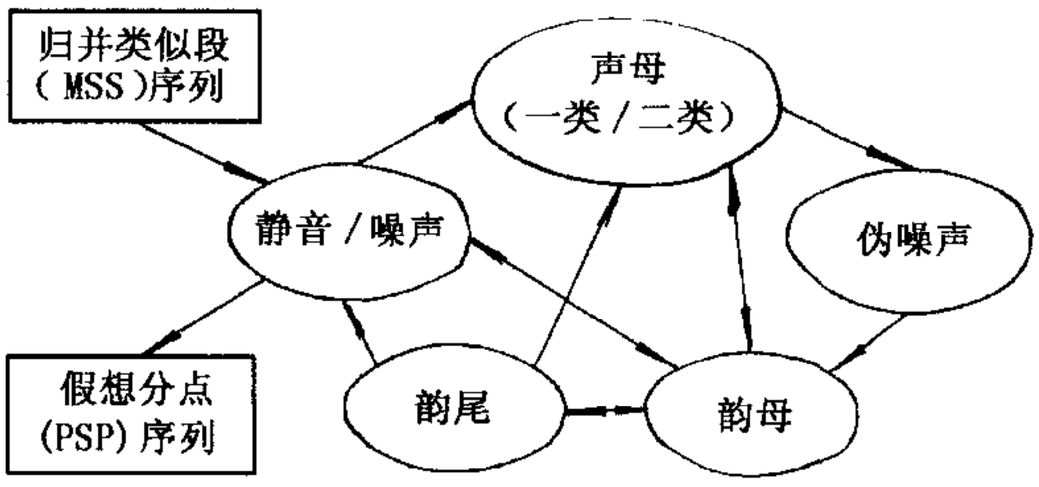

图 2 音节切分自动机的状态转移示意图

切分引擎采用“基于归并的音节切分自动机”思想. 它充分利用语音的短时能量、过零率、基音周期和傅里叶频谱等多种特征参数及其差分信息，把特征参数高度相似的相邻多帧语音(它们被认为属于相同的发音状态)进行归并，形成归并类似段(MSS, Merged Similar Segment). 这些 MSS 经过一个包含静音(噪声)、声母、伪噪声、韵母、韵尾等状态的“音节切分自动机”后，输出候选的音节边界点. 图 2 是它的状态转移示意图.

音节切分自动机充分利用 MSS 中的多种特征和汉语特有的声韵音节结构，以及依据声韵及噪声段的统计特性而制定的若干规则，给出一系列候选音节边界点，它们被称为假想切分点(PSP, Putative Separation Point). 每个 PSP 同时又具有一个信任度值，表明对其正确性的把握程度. 信任度值高于某个预设的“接受阈值”的 PSP 被认为是正确的音节分点，定义为真实切分点(TSP, True Separation Point)；否则被认为是错误的或不确定的.

为了评价切分引擎的性能，定义其切出率为 TSP 个数占实际音节分点个数的百分比，定义其分点正确率为 TSP 中实际正确的音节分点个数的百分比. 图 3 给出了在不同的信任度阈值情况下，切出率与分点正确率的关系，可见分点正确率是切出率的单调不增函数.

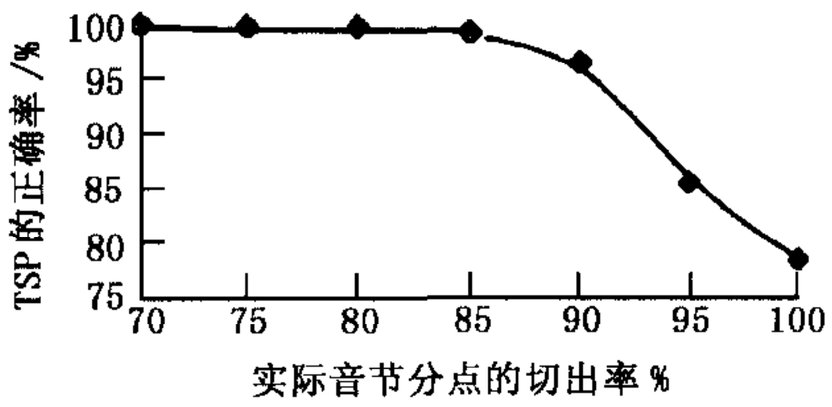

图 3　切分引擎的切出率与分点正确率的关系

根据图 3 的规律,可以把“接受阈值”选得比较苛刻,从而保证认定的 TSP 有接近 100%的正确率.对于那些没有确定为 TSP 的实际音节分点,将由后续的声学搜索过程来发现.另一方面,根据音节切分自动机经历的状态转移过程,并通过一系列规则,可以在每个语音段内准确地估计出它所包含的音节个数范围.切分引擎给出的 TSP 以及有效语音段内的音节个数范围,可以作为提供给后续声学搜索过程的一种导引知识,降低状态解码边界的不确定性,显著地提高搜索效率,使识别速度达到或接近实时.

相邻两个 TSP 之间的有效语音段定义为确定段,每个确定段内包含一个或多个音节,根据汉语语音的特点和发音速度,确定段内不会含有太多的音节数目.随后的 SKB-FSS 将在每个确定段内部进行,以产生多候选音节串组成的网络,供语言层面进行组句分析.

2.3　语言模型和组句分析

EasyTalk 的基本词汇有 5 万多个词,其中单音节词占 20.81%,双音节词占 65.50%,三音节词占 7.36%,四音节词占 6.33%.此外用户还可任意添加不长于 10 个音节的自定义词汇.语言层采用基于 Trigram 的统计语言模型,训练语料库包括 1993 和 1994 年《人民日报》的全文,以及《市场报》、《新华社文稿》、《经济日报》的摘编等约 2 亿字的文本,并预先进行了分词和词号标注.

但无论训练语料库的规模有多么大,也不可能包含所有有意义的 Trigram 词串,而识别时若出现训练文本中从未出现过的 Trigram 词串,则其最大似然估计为 0,这显然是不合理的.在 EasyTalk 中,我们使用一种基于 Turing 概率估计的平滑算法[14]来解决这种由于数据稀疏而造成的 0 概率 Trigram 问题,从而降低了语言模型的困惑度.

与声学层面的识别过程相似,语言层面的组句分析也是通过搜索实现的.在 EasyTalk 中,根据声学识别层给出的候选音节网络,以及语言模型提供的候选词串的 Trigram 概率,采用“音节同步网络搜索”算法[15],来得到最终的候选语句(目标词串)识别结果.

3　基于统计知识的帧同步搜索算法 SKB-FSS

3.1　传统的帧同步搜索算法

在传统的帧同步搜索算法的实现中,搜索过程是逐帧进行的.对于截止在时刻 t 的每一条部分路径,时刻 $t+1$ 的特征矢量被假设为可属于这些路径的任何可能的后续状态.于是搜索进行到 $t+1$ 时刻时,t 时刻的每条部分路径都根据其后续的可能状态列表,用 $t+1$ 时刻的特征矢量扩展出若干条新的部分路径.经过一些必要的状态合并动作并扔掉一些低竞争力的候选路径后,搜索过程再向前推进一帧.重复这一过程直到语音结尾.

帧同步搜索算法简洁而且高效.然而对于大词汇量、连续语音的识别任务来说,随着时刻 t 的增加,扩展出来的搜索路径会急剧增加.因此需要根据一定的准则随时剪除一些

低竞争力的路径.但是,较严的剪枝阈值可能会造成一些正确的路径被过早扔掉(这将无法在后续过程中得到恢复);而较宽的剪枝阈值又会大大增加存储空间和搜索过程的负担.

针对上述算法的不足,我们认为可以将一些与帧驻留长度有关的统计知识应用于搜索过程的状态转移中,以达到降低搜索复杂度的目的.有两类统计知识可以利用:

一种是基于纯统计知识的概率描述.比较典型的是对状态驻留长度进行建模,用概率密度来刻划状态驻留长度的分布情况.在搜索时,把系统处在当前状态、当前驻留长度下的条件概率作为惩罚分数加进路径的似然得分中,以此控制搜索路径的取舍.

另一种则是基于统计知识的规则.比如,根据类似的方法统计得到状态驻留长度分布的直方图,搜索时只有驻留长度落在允许范围内的路径才可以进行相应的状态转移或驻留,这可以看成是把第一种方法中的概率分布近似为均匀分布.本文讨论的重点就是这种基于规则的统计知识在帧同步搜索中的应用.

3.2 利用状态驻留分布信息的 SKB-FSS

状态驻留分布(SDD, State Dwell Distribution),或被称作统计直方图,是被广泛使用的用以进行搜索剪枝的一种信息.按照 32ms 的帧宽、16ms 的帧移对正常语速的 863 数据库进行统计(采用 6 状态 HMM),得到了表 1 所示的状态内驻留帧数的统计结果.

表 1 用 863 数据库训练的模型中各状态驻留帧数统计(%)

驻留帧数	0	1	2	3	4	5	6	7	8
第零状态	0.00	6.30	40.89	36.82	12.88	2.48	0.41	0.10	0.04
第一状态	1.06	25.14	44.55	21.76	5.82	1.20	0.25	0.09	0.04
第二状态	0.62	17.74	40.10	27.01	10.30	3.03	0.81	0.20	0.07
第三状态	0.59	13.68	32.96	29.32	15.18	5.81	1.79	0.47	0.11
第四状态	1.08	19.17	34.47	25.73	12.69	4.70	1.46	0.45	0.14
第五状态	0.00	7.29	44.64	32.24	11.64	3.11	0.72	0.18	0.07

从表中可以看出,每个状态内驻留的帧数比例最大的是 2 帧,平均覆盖了 39.6%;而比例最大的驻留帧数区间为 0～5 帧,覆盖了 98.7%以上,或 1～4 帧,覆盖了 95.0%以上.

我们摒弃利用驻留帧数的分布概率来作为惩罚分数的做法,而是确定一个允许的驻留帧数区间$[D_{\min},D_{\max}]$,该区间确定了在搜索时哪些驻留长度是允许的.很显然,若仅使用 SDD,对语速的变化不会有很好的鲁棒性,也不能保证很好的识别率.因为绝对的区间宽度限制了语速的变化范围,如果发音过长或过短,就不可能得到正确的识别结果.

3.3 利用差分状态驻留分布信息的 SKB-FSS

考虑到状态驻留分布 SDD 对语速变化的低鲁棒性,我们不再以绝对的驻留区间作为控制状态转移的参考因素,而是采用差分状态驻留分布(DSDD, Differential State Dwell Distribution)作为参考.表 2 是统计得到的相邻状态间差分驻留帧数的统计结果.

从表 2 可以看出,差分驻留帧数主要集中于 0 帧,平均覆盖了 32.7%;而比例最大的差分驻留帧数区间为－4～4 帧,覆盖了 99.8%以上,或－2～2 帧,覆盖了 95.4%以上.

因此,对于第一个状态,仍然给它分配一个较宽的允许驻留帧数的范围,以保证后续各个状态驻留范围的灵活性;而对于其它的状态,分别为之确定一个允许的差分驻留帧数区间$[d_{\min}^{(s)},d_{\max}^{(s)}](s>0)$,于是其有效状态驻留帧数定义为$[D^{(s)}+d_{\min}^{(s)},D^{(s)}+d_{\max}^{(s)}]$.其中,

$D^{(s)}$可以定义为第 $s-1$ 状态的驻留帧数(这被称为 DSDD-LST),也可以定义为前 $s-1$ 个状态的平均驻留帧数(这被称为 DSDD-AVG). 显然,采用这种搜索过程中的状态转移控制策略,就很容易与语速的变化相匹配了.

表 2　用 863 数据库训练的模型中相邻状态间的差分驻留帧数统计(%)

差分驻留帧数	−4	−3	−2	−1	0	1	2	3	4
零一状态间	0.18	2.38	14.67	35.02	32.98	12.31	2.14	0.27	0.04
一二状态间	0.06	0.65	4.89	18.87	33.00	27.59	11.39	2.87	0.57
二三状态间	0.20	1.27	6.25	19.06	30.41	25.53	12.28	3.80	0.92
三四状态间	0.61	2.87	11.26	26.17	30.95	18.44	6.97	1.98	0.48
四五状态间	0.27	1.36	5.96	20.21	35.99	27.25	7.57	1.13	0.16

3.4　SKB-FSS 的路径剪枝策略

虽然采用 SDD 或 DSDD 信息后,可以大大降低搜索时部分路径的膨胀速度,但是其空间消耗仍然很大. 因此在 EasyTalk 中,又采取了如下措施对搜索进行限制和剪枝:

第一,采用词(音节)搜索树作为路径扩展过程中音节转移时的词法限制. 词搜索树中每个叶结点都对应于词表中的一个词,从根结点到叶结点的树枝上的每个非叶结点依次代表了这个词的一个音节读音,且具有起始于根结点的相同"部分发音序列"的词共享相同的"部分非叶结点序列". 当某个部分路径发生音节转移时,仅仅根据词搜索树扩展出可能的后续音节,而不是所有的 418 个音节候选. 这就可以大大压缩搜索空间的规模.

第二,每当完成当前帧的路径扩展时,都要对具有相同候选音节序列和相同当前状态的那些路径进行筛选,仅保留其中具有最高累积声学得分的一条或多条路径. 根据 HMM 状态转移的无后效性原理,这种路径剪枝策略是合理的. 它可以及时扔掉一些低竞争力的部分路径,以保证后续扩展时较小的空间开销,并且减少多余路径的干扰.

第三,采用动态前向预测路径有效性的方法进行路径剪枝. 根据当前确定段的 TSP 和音节个数范围信息,SKB-FSS 必须保证在搜索结束时,有效路径的结尾候选音节的末状态恰好落在确定段的右边界上,否则这条路径就是不合理的. 因此,在每帧完成扩展之后,根据 SDD 或 DSDD 中每个后续状态的允许驻留帧数范围,确定每条路径从当前状态出发,总共可以到达的最小和最大的帧数范围$[F_{\min}, F_{\max}]$,然后与当前确定段的总帧数 F_{DS}相比,若 $F_{\min} > F_{DS}$或 $F_{\max} < F_{DS}$,则说明这条路径将要扩展的"任何一条路径"中的某音节的末状态都不可能恰好落在确定段的右边界上,因此它被作为无效路径立即删除.

通过上述的路径扩展限制和剪枝策略,即保证了较小的时空开销、提高了搜索效率,也减少了无效路径的干扰,提高了声学层面的整体识别率.

4　实验结果

我们通过实验对 SKB-FSS 的各种策略进行了比较. 为了更好地测试这些策略在纯声学层面上的有效性,首先禁止了 EasyTalk 中的语言模型处理选项,挑选了 204 个汉语词组,每个词组包含 2 到 4 个音节,分别由 10 个人进行发音. 测试了采用 SDD 和 DSDD(包括 DSDD-LST 和 DSDD-AVG)方案时 SKB-FSS 对它们的声学识别结果. 对每种方案,分别统计识别结果中各个词组前 5 个候选的累积识别率,见表 3.

表 3 控制状态转移的搜索策略性能比较(词识别率%)

策略 \ 候选数	1	2	3	4	5
SDD	63.2	67.1	67.6	69.1	69.6
DSDD-LST	77.9	82.8	84.8	86.8	88.2
DSDD-AVG	86.3	89.7	92.6	94.1	96.1

从表中可以看到,对于上述词组的首选词识别率,利用差分状态驻留信息的 DSDD-AVG 的性能比仅利用静态状态驻留信息的 SDD 的性能相对提高了 36.6%. DSDD 规则之所以有较好的效果,其主要原因是它能够较好地与语速相匹配,同时把一些不符合实际的搜索路径及时剪除,排除了"干扰",使得搜索算法能够把有限的搜索空间限制在最有意义的路径上. 同时,与基本的帧同步搜索算法相比,SKB-FSS 的搜索速度也有较大的提高.

另一方面,在采用 SKB-FSS(选择 DSDD-AVG 策略)的当前版本的 EasyTalk 上,允许了语言模型处理的功能后,从 863 数据库的 10 个人的样本中随机抽取出 200 个语句进行了连续语音的整句识别测试,其总体的字正确率达到了 87.6%.

5 结论

本文从整体上介绍了连续汉语语音识别系统 EasyTalk 的实现原理,并重点介绍了声学层面采用的基于统计知识的帧同步搜索算法 SKB-FSS. SKB-FSS 包含了三个基本层次:

1) 基于归并的音节切分自动机产生待搜索的语音确定段,减少基元边界的不确定性.

2) 采用基于统计知识的(差分)状态驻留信息,来控制搜索过程中的状态转移过程.

3) 利用词搜索树控制基元的扩展规模,并根据动态前向预测进行合理而及时的剪枝.

通过纯声学层面的词组识别,以及听写机系统上的连续语音整句识别的实验结果,验证了 SKB-FSS 算法在汉语连续语音识别中的有效性.

参 考 文 献

1 Zheng F, Song Z J, Xu M X *et al*. EasyTalk: A large-vocabulary speaker-independent Chinese dictation machine. In: Proceedings EUROSPEECH. Budapest, Hungary, 1999, **2**: 819～822

2 Rabiner L R, Juang B H. Introduction to hidden Markov models. *IEEE ASSP Magazine* (*Acoustics, Speech, and Signal Processing*), **3**(1): 4～16,1986

3 Huang X D, Jack M A. Semi-continuous hidden Markov models for speech signals. *Computer Speech and Language*, 1989, **3**: 239～251

4 Viterbi A J. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Information Theory*. 1967, IT-**13**: 260～267

5 Lee C H, Rabiner L R. A Frame Synchronous network search algorithm for connected word recognition. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 1989, **37**(11): 1649～1658

6 Kenny P, Hollan R, Gupta V *et al*. A*-Admissible heuristics for rapid lexical access. In: Proceedings ICASSP. Toronto, Canada, 1991, **1**: 689～692

7 Paul D B. Algorithms for an optimal A* search and linearizing the search in the stack decoder. In: Proceedings ICASSP. Toronto, Canada, 1991, **1**: 693～696

8 Li Z, Boulianne G, Labute P *et al*. Bi-directional graph search strategies for speech recognition. *Computer Speech and Language*, 1996, **10**: 295～321

9 Juang B H, Rabiner L R. A probabilistic distance measure for hidden Markov models. *AT&T Technical Journal*, 1985, **64**(2): 391～408

10 Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 1980, ASSP-**28**(4): 357～366

11 Furui S. Speaker-Independent isolated word recognition using dynamic features of speech spectrum. *IEEE Trans. Acoustics, Speech, and Signal Processing*, 1986, **34**(1): 52～59

12 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[硕士学位论文].北京:清华大学计算机科学与技术系,1989

13 张继勇,郑方等.连续汉语语音识别中基于归并的音节切分自动机.软件学报,1999,**10**(11): 1212～1215

14 牟晓隆,詹津明等.基于修正退化频度估计算法的 N-gram 语言模型.第五届全国人机语音通讯学术会议论文集.哈尔滨:哈尔滨工业大学,1998.206～209

15 Zheng F. A syllable-synchronous network search algorithm for word decoding in Chinese speech recognition. In: Proceedings ICASSP. Phoenix, USA, 1999, **2**: 601～604

宋战江　1972 年 9 月生.分别于 1994 年和 1997 年在南开大学获得计算机软件专业学士学位和计算机应用专业硕士学位.现在清华大学计算机系攻读博士学位,研究方向为语音识别和理解.

郑　方　1967 年 3 月生.分别于 1990 年和 1992 年获清华大学计算机科学与技术专业学士学位和硕士学位,于 1997 年获清华大学计算机应用专业博士学位.现为清华大学计算机系副教授、语音实验室主任、清华-ADI DSP 技术研究中心主任、IEEE 会员、《中文信息学报》编委.专业兴趣包括信号处理,语音识别和理解等.