《软件学报》, Vol. 10, No. 11, pp.1212-1215， 1999 年 11 月

# 连续汉语语音识别中基于归并的音节切分自动机

**张继勇 郑 方 杜 术 宋战江 徐明星**
清华大学计算机科学与技术系语音实验室 北京 100084
*zjy@sp.cs.tsinghua.edu.cn, (010)62772001*

**摘要：**本文研究并实现了汉语连续语音中的音节自动切分算法——基于归并的音节切分自动机(MBSDA, Merging-Based Syllable Detection Automaton)算法。MBSDA 算法利用了包括语音的短时能量、过零率和基音周期在内的多种特征参数，把特征参数高度相似的相邻帧(一帧或若干帧)的语音信号进行“归并(Merging)”，形成“归并类似段(Merged Similar Segment，简称 MSS)”，它们被认定属于同一音节的相同状态。这些 MSS 经过一个包含若干状态的“音节切分自动机(Syllable Detection Automaton，简称 SDA)”后，输出音节的切分点。每个确定的切分段中所包含的音节的个数的范围(Range of Syllable Number，简称 RSN)也由 MBSDA 算法给出。
**关键词：**音节切分，归并，音节切分自动机，韵母特征类段，音节个数范围估计

## 1. 引言

在非特定人连续语音的识别中，目前最广泛使用的声学模型是 HMM 模型及其改进模型，其中传统的帧同步算法[1]或 Viterbi 解码算法[2]给出了状态解码序列。借助于一定的词法信息和语言模型，就可以由声学搜索结果给出最大似然句子输出，完成句子识别[3]。

在这样的方案中，声学的搜索算法有两个问题不容忽视。(1) 搜索路径的组合爆炸问题；(2) 解码出的状态序列错位问题。它们在很大程度上影响了整个识别系统的性能。我们的实验表明，即使在连续语流中，在音节切分点已知的情况下孤立音节的识别率也能达到了理想的水平[4]，因此理想的做法是把一段连续语音信号准确地切分至单音节然后再加以识别。但是由于目前研究条件和水平的局限性，不加任何限制要达到这一点是比较困难的。对有些音节，我们可以给出准确的边界，而有些音节之间的边界却非常难以区分。本文采取的方案是把完全确定的地方切开，切不开的地方则给出音节的个数范围(RSN, Range of Syllable Number)，从而较好地解决上述的问题。为此，本文提出了一个基于相似语音帧合并的音节切分自动机(MBSDA, Merging-Based Syllable Detection Automaton)算法。汉语音节性很强的特点为这种算法提供了很好的理论依据。

## 2. 算法基本原理

### 2.1 特征参数的提取

MBSDA 算法使用的主要特征参数是短时帧能量、过零率和基音周期。特征参数的提取和分析以帧为单位。具体的计算请参看有关文献[5]，在此不作介绍。

### 2.2 相近语音帧的归并

在同一个音素的发音过程中，声道会在一定的时间间隔内保持稳定；而当从一个音素到下一个音素过渡时，声道会发生变化。因此，如果连续几帧语音特征没有发生比较大的变化，我们有理由认为它们是属于同一个音素的。基于此，我们提出了对这种特征相近的帧进行“归并(Merging)”的概念。

在进行归并之前，我们首先检查语音特征是否发生“转折(Transition)”。转折有 I 类转折和 II 类转折两种，描述如下：

I 类转折：特征发生突然变化。即当前帧的能量(或过零率)大于前一帧能量(或过零率)的$\alpha$倍；或当前帧能量(或过零率)的$\alpha$倍小于前一帧的能量(或过零率)。(这里取$\alpha = 2$)

II 类转折：特征发生缓慢变化。即当前帧的前$T$帧语音能量(或过零率)的均值与后$T$帧语音的能量(或过零率)的均值之间存在类似于 I 类转折中的变化关系。(这里取$T = 3$)

如果当前的语音帧发生了上述的 I 类或 II 类转折，则给该帧语音作上“转折标记(TT, Transition Tag)”。连续的一帧或几帧没有 TT 标记的语音被归并到同一个 MSS 类段中。类段的一个很重要的性质是它反映了这段语音中各个音素中最稳定的部分。

### 2.3 音节切分自动机的实现

通过类段并不能直接给出音节的切分边界，我们构造一个音节切分自动机(SDA, Syllable Detection Automaton)，由自动机根据其内部状态的转移来确定音节的切分点。SDA 的状态划分为以下几类：静音、噪声、一类声母、二类声母、伪静音、韵母和韵尾， SDA 中状态的含义分别如下：

**静音**(SIL)**和噪声(NOI)：**静音指能量和过零率都是很低的信号，它不包含语音信息。而当环境噪音比较强时，静音转变为噪声。

**一类声母(SM1)和二类声母(SM2)：**对声母特征的研究和统计表明，l,m,n,r 具有类似于韵母的特点，我们将它归入二类声母的状态；剩下的声母的特征与韵母有教明显的区分，我们将它们归入一类声母。

**伪静音(Psudo-SIL)：**有些音节在发音时，声母和韵母中间有一个能量的“低谷”。比如音节“kai”，声母“k”和韵母“ai”的能量都比较高，但它们之间的过渡段的能量却很低，类似于静音。如果把它归到静音状态，则显然会发生错误。为了对这种情况加以区分，我们在这里引入了伪静音状态。

**韵母(YM)和韵尾(YW)：**韵母一个很明显的特点是它具有准周期、较高的能量和适中的过零率。而当发音从一个音节转变到下一个音节时，韵母部分能量和过零率都有比较明显的下降，这就是韵母中的韵尾。

通过大量的实验，我们得到了各个状态和特征参数之间的大致规律如表 1 所示。自动机各个状态间的转换图如图 1 所示。

表 1 自动机状态和特征参数的关系

| MSS 性质 / 帧参数 | 静音 | 噪声 | 一类声母 | 二类声母 | 伪静音 | 韵母 | 韵尾 |
|---|---|---|---|---|---|---|---|
| 规整能量 | < 100 | <500 | 100~1000 | 100~1000 | 0~500 | > 500 | 100~1000 |
| 过零率 | < 3 | 10~ 40 | > 50 | < 30 | <10 | 5~ 80 | < 30 |
| 基音周期 | 无 | 无 | 无 | 有 | 无 | 有 | 无 |

### 2.4 SDA 的输出和音节切分

如果我们把静音和噪音都当作一个“伪音节”或“广义音节”来看待，那么根据汉语音节的特点，当表 2 中的任何一个状态转移条件满足时，都表明语音流发生从一个广义音节到另一个广义音节的转变。这也是 SDA 需要输出音节切分标记的地方。

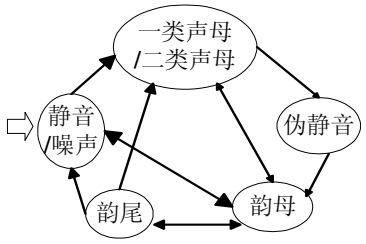


图 1 音节切分自动机的状态转换图

表 2 作音节切分标记的条件

| 状态转移条件 | 自动机前一状态 | 自动机的当前状态 |
|---|---|---|
| 条件 1 | 静音、噪声 | 声母、韵尾 |
| 条件 2 | 韵母 | 静音、噪声、声母 |
| 条件 3 | 韵尾 | 声母、静音、噪声、韵母 |

### 2.5 切分段所含单音节个数范围估计

切分段音节个数范围(RSN, Range of Syllable Number)的确定有下面几个步骤：

对于第 $n$ 个切分段，统计出其中所包含的韵母特征类段(VFS，Vowel Feature Segment)的个数，记为 $C_{VFS(n)}$。当前切分段的平均音节长度(ASL, Average Syllable Length)值，如果不考虑初始的情况，可按下式来计算(以帧为单位)：

$$ASL_n = \sum_{i=-1}^{-M} l_n^{(1)}(i) / M \tag{1}$$

其中 $l_n^{(1)}(-i) = l^{(1)}(n-i)$ 表示第 $n-i$ 个“音节估计个数为 1”的切分段长度(以帧为单位)，也即当前位置以前第 $i$ 个“音节估计个数为 1”的切分段长度。上式用当前位置以前的 $M$ 个“音节估计个数为 1”的切分段长度来估计当前段的 ASL 值。

这里 $M$ 的选择必须合适：如果太小，则估计对音节长度的局部变化太敏感，缺乏抗干扰能力；如果太大，音节个数的估计缺少对语速变化的跟随特性。通常我们选 $M = 10$。

在给出 RSN 之前，我们先利用 ASL 值定义一个上限参考值

$$C_{R(n)} = \lfloor l_n / ASL_n \rfloor \tag{2}$$

其中 $l_n$ 表示待估计的切分段长度。

在求得待估计切分段的 $C_{VFS}$ 值和 $C_R$ 值之后，则该切分段所包含的音节个数的上限确定为

$$C_{\max(n)}=\begin{cases}C_{R(n)}-1, & C_{VFS(n)}<C_{R(n)}-1\\ C_{R(n)}+1, & C_{VFS(n)}>C_{R(n)}+1\\ C_{VFS(n)}, & \text{其他}\end{cases} \tag{3}$$

利用本文给出的切分算法，一般$C_{\max}$值不超过 3，大部分情况下为 1。

## 3. 实验及评价

### 3.1 实验数据

在实验中我们使用了两批测试数据。第一批测试数据采用了 863 语音数据库。这批测试数据是在安静的办公室环境下采录的，基本上没有噪声的干扰。我们从中抽取了 5 男 5 女的语音数据，每个录音者各取 20 句。这样组成了一个男声和女声各 100 句的测试集。为了对比噪声对切分结果的影响，我们在环境噪声较强的房间里采录了第二批测试数据。另外，还故意加入了说话者的“吹气”、“咳嗽”等干扰信号。这批数据共有 100 句，都是男声。两批数据的采样频率均为 16KHz，量化精度为 16bit。录音者的语速为每分钟 150~180 个字。帧长为 16ms(256 个样本点)。

### 3.2 音节切分算法的评价方案

切点正确率$p_d$和个数范围估计正确率$p_n$采用如下的公式来计算：

$$p_d=\frac{SDA\text{给出正确切点个数}}{SDA\text{给出总的切点个数}}\times 100\%,\quad p_n=\frac{RSN\text{估计正确的切分段数}}{SDA\text{给出的总切分段数}}\times 100\% \tag{4}$$

总的切分正确率$p_c$由$p_d$和$p_n$共同决定。为了反映在切分算法中能切出的音节占总音节的比例，我们引入了切出率$p_{out}$的概念。定义分别如下：

$$p_c=(p_d+p_n)/2,\qquad p_{out}=\frac{SDA\text{切出的音节个数}}{\text{实际的音节总个数}}\times 100\% \tag{5}$$

### 3.3 实验结果

根据上面的统计公式，适当调整切分算法的参数，可以得到切分算法在不同操作点下的 ROC(Receiver Operating Characteristics)曲线，如图 2 所示，该曲线中，MBSDA 算法的$p_c$是$p_{out}$的函数。我们可以根据不同的切出率需求调整 MBSDA 参数，从而得到 ROC 曲线上的不同操作点。但一般情况下，一个实际使用的切分算法要求$p_c$接近 100％。

从 ROC 曲线上我们可以看到，如果刻意追求切出率或者正确率都会导致算法整体的性能的下降。在允许发生一定错误的情况下，当切出率大致为 90%时，系统整体的性能最理想。表 3 中列出了其中一次实验统计的结果。

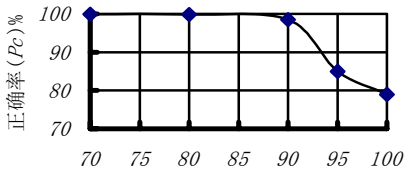

图2 MBSDA切分算法的ROC曲线

表 3 实验统计结果

| 测试数据库 \ 结果(%) | | 切点正确率($p_d$) | RSN 正确率($p_n$) | 总正确率($p_c$) | 切出率($p_{out}$) |
|---|---|---|---|---|---|
| 第一批数据 | 男声 | 99.2 | 98.3 | 98.8 | 90.3 |
| | 女声 | 99.7 | 98.5 | 99.1 | 92.1 |
| 第二批数据 | 男声 | 98.9 | 96.2 | 97.6 | 89.8 |
| 平　　均 | | 99.3 | 97.7 | 98.5 | 90.7 |

从实验的结果可以看出，总体来讲切分的效果是很不错的。在男声和女生的对比中可以发现，女生的切分效果要比男生好，原因主要是由于数据库中的女声发音比男声发音要清晰。此外，通过实验结果还可以看出，噪声对切分的结果有一定的影响，但算法的整体性能没有很大幅度的下降，这反映了本算法对噪声有较好的鲁棒性。

## 4、结论

无论是从算法的可行性还是算法的复杂度来看，基于语音切分的汉语语音识别是一种行之有效的方案。如何能保持本算法现有的切分正确率，提高算法的音节切出率，将是我们今后努力的方向。可以看出，在此算法的基础上，如果再加入语音信号的一些其他特征参数，如倒谱参数，将会提高算法的切出率。当切出率达到很高的值时，实际上也就是达到了切分至单音节的目标。此外，由于本切分算法的自动机已经给出了每一帧的状态，因此对该算法加以适当的修改即可实现连续语音中音节的声/韵切分。

【参考文献】

[1] C.H. Lee and L.R. Rabiner. “A Frame Synchronous network search algorithm for connected word recognition.” *IEEE Trans. On ASSP*, 37(11): 1649-1658, Nov. 1989

[2] Zheng Fang, Chai Haixin, Shi Zhijie, Wu Wenhu and Fang Ditang, “A Real-World Speech Recognition System Based on CDCPMs”, in Int’l Conf. On Computer Processing of Oriental Languages(LCCPOL’97), 1:204~207, Apr.2,97, Hong Kong

[3] 郑 方，牟晓隆，徐明星，武 健，宋战江等. “一个语文转换文本编辑器的实现.”. 见：王承发、张凯等编，第五届全国人机语音通讯学术会议(NCMMSC’98)论文集. 哈尔滨：哈尔滨工业大学出版社，1998 年 7 月，280-285

(Zheng Fang, Mou Xiao-long, Xu Mingxing, *et al*. The Implementation of a speech-to-text editor. In: Wang Chengfa, Zhang Kai, eds. Proceedings of the ‘1998 National Conference on Man-Machine Speech Communication. Harbin: Harbin Institute of Technology Press, 1998. 280~285.)

[4] 郑 方，吴文虎，方棣棠. “CDCPM 及其在语音识别中的应用.” 软件学报，863 高技术项目智能主题专刊, 7:69-75, 1996 年 10 月

(Zheng Fang, Wu Wenhu, Fang Ditang. CDCPM with its application to speech recognition. J. Of Software, 1996, 7(863 Special Issue): 69~75)

[5] 杨行峻，迟惠生. 语音信号数字处理. 北京：电子工业出版社，1995 年.

(Yang Xingjun, Chi Huisheng. Speech Signal Processing. Beijing: Publishing House of Electronics Industry, 1995)

## Merging-based Syllable Detection Automaton in Continuous Speech Recognition

Zhang JiYong,Zheng Fang,Du Shu,Song ZhanJiang and Xu MingXing

Speech Lab., Dept. of Computer Science and Technology, Tsinghua Univ., Beijing, 100084, P.R.China

*zjy@sp.cs.tsinghua.edu.cn,(010)62772001*

**Abstract:** In this paper an automatic syllable detection method namely merging-based syllable detection automaton (MBSDA) is studied and implemented. The MBSDA uses a variety of features including the frame energy, the zero crossing rate and the fundamental frequency to merge similar frames (one or several frames) into one merged similar segment (MSS). The frames in the same MSS are treated as frames of the same state of a phonetic. These MSSs are passed into a syllable detection automaton (SDA) to give the syllable detection results. In addition, the MBSDA gives the range of syllable number (RNS) of each definite detection segment.

**Keyword:** syllable detection, syllable detection automaton, vowel feature segment, range of syllable number