第四届全国汉字与汉语语音识别学术会议论文集(NCCC&CSR)，pp.285-289, 1992年5月,杭州

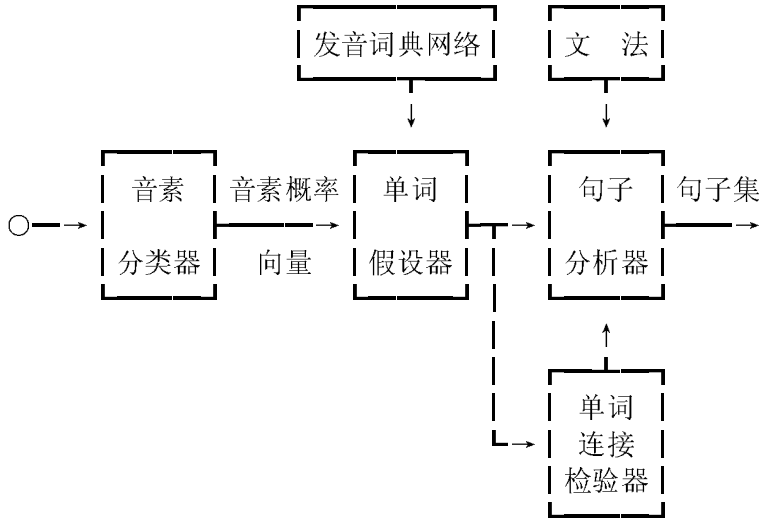
汉语连续语音识别中音节自动切分的研究

清华大学计算机系 郑　方 吴文虎

【摘要】本文介绍一种应用于汉语连续语音识别中的音节自动切分技术，它以平滑化了的帧能量轮廓为主要判据，利用相对阈值进行音节切分，并利用帧过零率和音长信息准确定位音节分隔点，收到很好的效果。

【Abstract】In this paper a new sort of technology for syllable-detection in Continuous Chinese Speech Recognition is proposed, which uses smoothed frame energy curve as its main criterion and the relative threshold, the frame zero-crossing rate, and the speech length information as its subsidiary criteria. As a result, this kind of method has got better performance.

连续语音的识别与理解的研究在各国都在抓紧进行，汉语连续语音识别的研究工作也已被提到日程上来，而音节切分的技术在其中占据相当关键的地位。先看一下美国卡内基梅隆大学计算机系研究的语音识别系统（CMU 系统）的结构[1]：

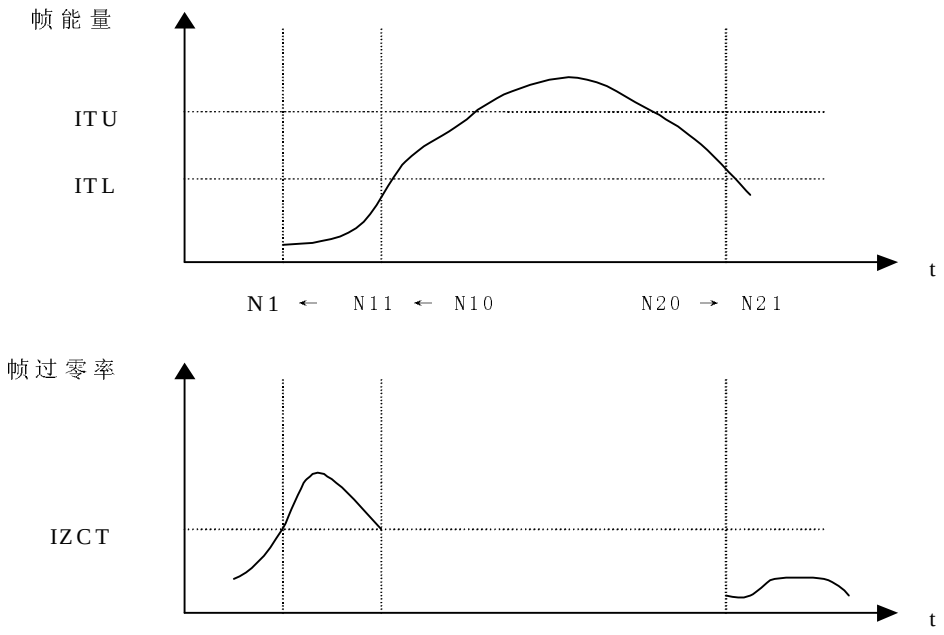


CMU 系统大致上分为三个模块：音素分类器（ Phoneme Classifier）、单词假设器（Word Hypothesizer）和句子分析器（Sentence Parser）。其中音素分类器完成信号处理和音素切分功能，它的正确与否直接影响到整个系统的识别率。

汉语和西语相比有比较简单的结构[2][3][4]。通常一个汉字就是一个音节。每个音节一般由声母、韵母及声调组成。因此汉语中组成所有汉字的 1286 个音节就可以由 21 个声母、37 个韵母和四个声调有机组合而成。声母主要由清辅音（主要是清擦音和塞音）和浊辅音（主要是鼻音和边音）组成。根据音长也可以分为长声母（ j、q、x、zh、ch、sh、z、c、s)、中长声母（p、t、k、h）和短声母（b、m、f、d、n、l、g、r)。韵母由元音、双元音和鼻辅音等组成。

根据汉语的特点，连续语音识别系统一般采用基于音节的识别方法（而不是基于音素）。它也有三级处理：信号→音节；音节→单词；单词→句子。按照这样的处理方式切分音节时，要求绝对正确。

许多语音工作者在音节切分方面都作了大量的研究工作。[5] 使用能量和过零率两个简单的时域度量为判据，用双门限进行音节切分。其典型的例子如下图：



（以强摩擦音开始的音节）

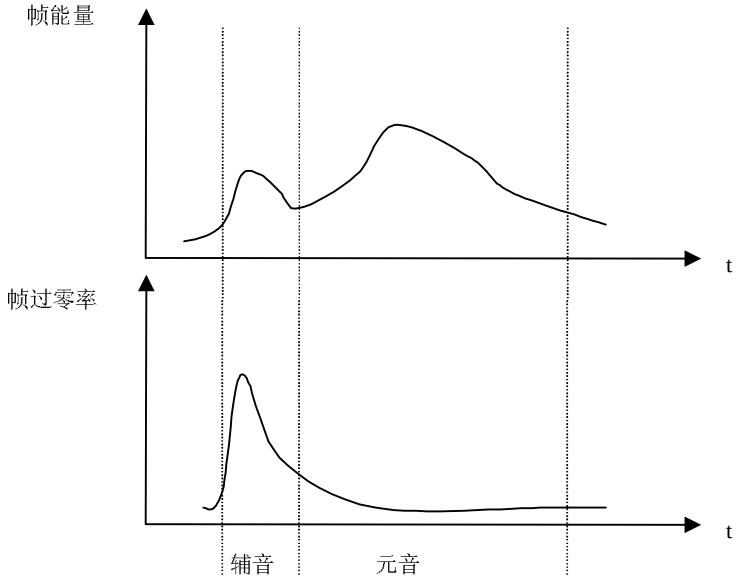
它在能量曲线轮廓中，先以一个很保守的门限 ITU 把音节定位在 N10 与 N20 之间；再从这两个分隔点开始，分别向左右扩张，以另一门限 ITL 把音节定位在 N11 与 N21 之间，最后在平均过零率曲线轮廓中，根据门限 IZCT 分别从 N11 向左或从 N21 向右找到真正的音节起止点 N1 和 N21。这种方法能较好地切分音节，但它有一个很致命的弱点，就是对语音信号的幅度（影响能量轮廓）很敏感，这是因为 ITU 和 ITL 是固定的值。

[6] 一改传统的方法，以能频值（帧能量与帧过零率的乘积）作为音节切分的判据。 一般情况下切分正确，但由于两种度量的相乘，一方面容易使音节分隔点模糊（谷点太低），另一方面容易使音节分隔点移位（左右错位）。

笔者在经过大量实验之后得出这样的结论：

• 音节分隔点几乎都在能量轮廓的谷点。

• 爆破音（b、d、g、p、t、k）和摩擦音（zh、ch、sh、z、c、s、j、q、x）音段处的过零率相对较大。一个典型的例子如下图：



• 经过平滑化处理后的能量曲线轮廓中，有一些谷点处于过渡音段处，而且谷点与左右峰点的幅度差不大。

• 不管语音信号的幅度如何，同一音的能量曲线轮廓形状大致相似；过零率则基本一样。

基于上述实验结果，本文提出一种音节切分方法：以归一化的短时能量曲线轮廓为主，以短时过零率和音长信息为辅，进行音节切分。由于对能量曲线轮廓进行归一化的过程要使用除法，因此具体实现时使用相对能量阈值来替代归一化过程。事实上，是用相对能量阈值和绝对过零率阈值（先验统计值）共同切分。

这里提出相对能量阈值（Relative Energy Threshold，简称 RET）的概念。RET 是一个后验值，它随不同的语音信号幅度而变，较好地跟随了信号的短时幅度和能量特性。它既作为能量曲线轮廓的阈值，又作为谷点和与其相邻峰点间落差的阈值。因此“相对”有两重含义：一方面阈值是随信号而变化的；另一方面该阈值又作为谷峰点间落差的度量。

这种音节切分方法的步骤如下：

1.使用一个绝对能量阈值(先验统计值)和(绝对)过零率阈值((Absolute)Zero-Crossing Threshold，简称 ZCT）进行有效语音段的端点检测。

2．对有效语音段进行能量曲线轮廓的平滑化处理。为简便起见，用一帧信号的幅度和代替帧能量。

3．在有效语音段内以平均帧能量的N倍或最大帧能量的M分之一作为本语音段的 RET。（M、N 都是经验整数值）

4．以 RET 和 ZCT 在有效语音段内进行音节切分。在实验的基础上，音节切分使用下面的一些规则：

①能量曲线轮廓中，谷点值为 RET 的一至三倍的地方为音节分隔点。

②能量曲线轮廓中，与左右峰值的落差为 RET 的四至六倍的谷点，一般为音节分隔点。

③能量曲线轮廓中，谷点处值较高但附近过零率超过 ZCT 的区域一般为辅音段。再根据该谷点离前一音节分隔点间的距离 — 音长，可以确定该谷点是辅音起点
（是音节分隔点）还是辅音终止点（非音节分隔点）。

④处于上升阶段的能量轮廓中出现的低落差谷点和处于下降阶段出现的低落差谷点，认为是过渡段分隔点（元辅音段，或双元音段的两元音之间）。

⑤音长已经达到一个音节的数量级，但依照前面规则尚不能判定音节分隔点的谷点，可按照一定的条件来判定它是否为音节分隔点。这条规则是一条补救规则，很少被采用，但实验证明这一补救思路是有效的。

大量实验表明，这种切分方法几乎完全正确地把一段以正常语速说出的语音的音节切分出来，只要说话者将每个音节都较清晰地发出即可。

在音节切分中一般会有三类错误发生：

1.音节分隔点检测不出。出现在为数不多的某些音节连接处（如 -i 和 y- 间，-u 和 w- 间等），比如“穷兵黩武”的“黩武”两字，如果读得快而含糊就不容易切分开。

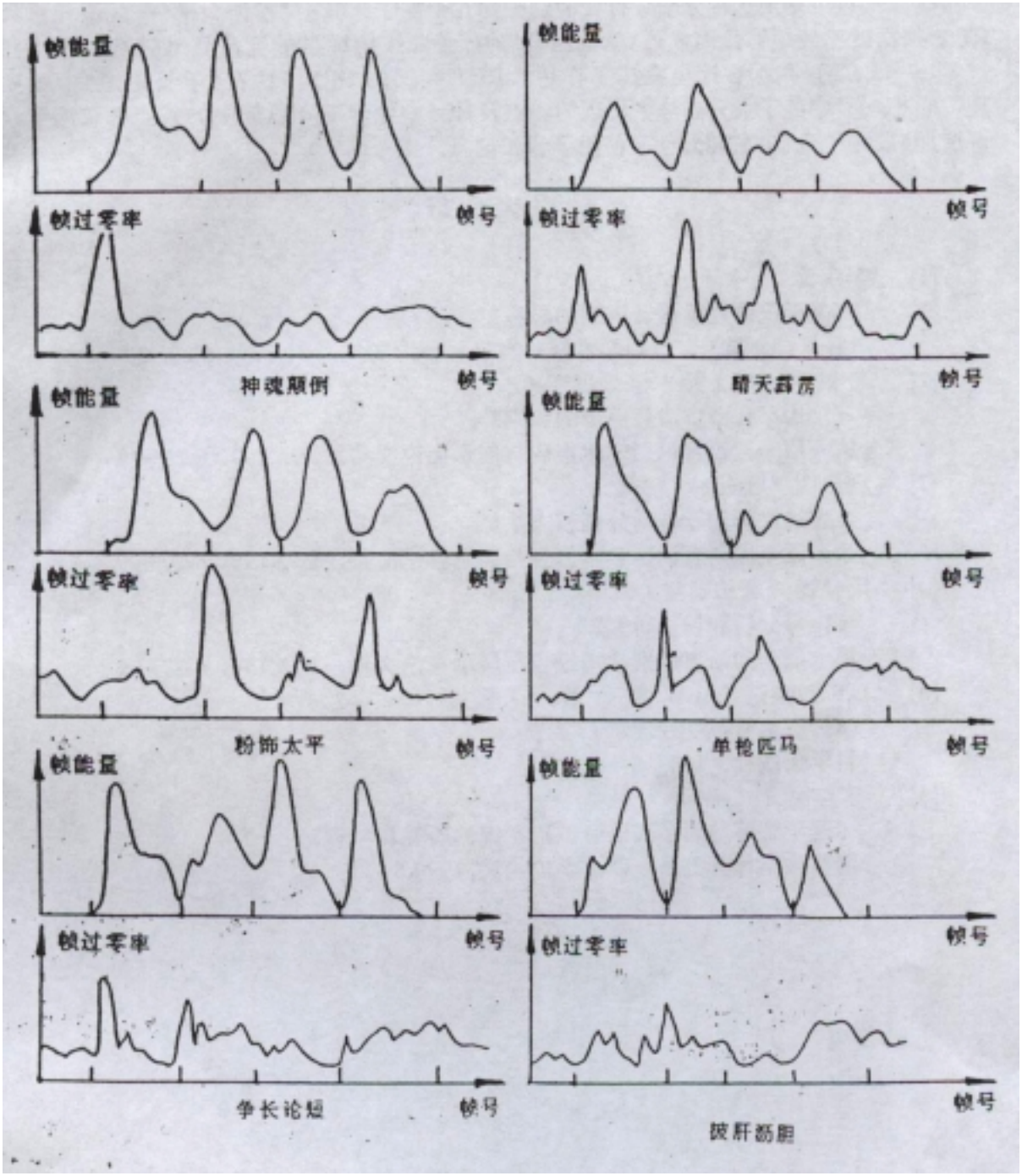
2.不该切分的地方误判为音节分隔点。这类错误多出现在发三声的音时，这时声调由高到低再由低到高的转折处能量曲线会有一个谷点，特别在强调这种声调的变化时容易被切成两个音节。

3.音节分隔点位置不准确（错位）。使用这种方法，这类错误几乎不发生。

通过大量实验，我们感到要做到完全正确地切分，只利用时域方法是有困难的。但如果再辅以其它一些判据（譬如利用 CEP 系数的变化、基音周期的有无等）是可以切得更为准确的。

音节切分技术是连续语音识别与理解的特别重要的一环，我们相信也期望使用时域度量再加上一些频域的判据能更准确地完成这一切分过程。

附录：以下是用这种方法进行音节切分的几个例子。其中“粉饰太平”一成语几乎只用 RET 判据就已经足够；而成语“单枪匹马”的能量曲线轮廓虽然无法将“匹”分割出来，但过零率（ZCT）判据在这里发挥了作用；RET 判据的“相对”性在“争长论短”中得到体现，它把“长”中处于上升趋势中出现的谷点及“论”中处于下降趋势中的谷点都忽略了；“披肝沥胆”既用了 RET 的相对性又使用了音长信息。



参考文献

[1] 李宗葛,《关于汉语连续语音识别的思考》,第一届全国语言识别学术报告与展示会论文集，P1,1990.6.27-29

[2] 王站东　叶士元　陈维民,《汉语单音节中清音声母的识别》，第一届全国语言识别学术报告与展示会论文集，P24,1990.6.27-29

[3] 程启明　王松林,《连续汉语语音自动分割新方法》，第一届全国语言识别学术报告与展示会论文集，P55,1990.6.27-29

[4] 周桑漪 渡边泰堂,《声母分析和识别研究》，第一届全国语言识别学术报告与展示会论文集，

P34,1990.6.27-29,
[5] L. R. 拉宾纳　R. W. 谢弗 著　朱雪龙等译,《语音信号数字处理》,科学出版社，1987
[6] 李建民,《基于音节的大字表语音识别系统的研究和实现》,清华大学计算机系，硕士学位论文, 1990.6