自然口语对话中的韵律特征分析

刘亚斌　李爱军

中国社会科学院语言研究所

lyabin@sina.com　liaj@linguistics.cass.net.cn

摘要

本研究以自然口语对话语音语料库 CADCC(Chinese Annotated Dialogue and Conversation Corpus)为基础，对自然口语语料中各级韵律边界和各级重音的声学特征进行了统计分析，包括韵律边界处的停顿、边界前后各级韵律单元的时长和音高(F0)表现、各级重读音节的时长和音高(F0)表现等，并在此基础上初步得出了自动识别韵律边界和重音等级的一些规则。

1．引言

近年来，语音技术有了日新月异的发展，在语音合成的可懂度和语音识别的准确性方面都有了很大的提高，也出现了一些以语音技术为核心的产品。但是在对连续话语的处理上，无论是识别的准确性还是合成的自然度，都还是不能让人十分满意。从处理孤立话语到处理连续话语，许多新的问题涌现出来，其中最重要的就是韵律问题。

鉴于韵律问题的重要性，越来越多的人也开始对韵律进行研究，并得到了许多有用的结论，但大多局限于朗读语料，且多为小语料库，不能真实全面的反映人们实际语言的特点。研究表明，自然话语与朗读语料之间存在着很大的差异[1]，因此，我们认为，要想真正弄清自然话语的韵律特征和规律，就必须采用较大的自然口语语料库作为研究材料，然后在其基础上展开研究，才能最终找到破解韵律难题的钥匙。本文正是这样来做的，以近一个小时的真实的自然口语语料为基础，对其韵律特征进行逐一考察，归纳出其中的规律，并进而提出了用于韵律自动识别的一些规则。

2．语料与标注

2．1 语料库简介

自然口语对话语料库 CADCC（详见[2]）包括两个子库，其中 SET 1 是电话对话库，SET 2 是正常通道对话库。SET 2 中共有 13 对发音人，对话双方是同事或同学，有共同的爱好或话题，谈话内容不限，也就是语篇话题可以自由转换。其中有 8 位发音人曾参加过朗读语篇 ASCCD 的录音，这样以便详细对比朗读和自然口语的各种差异。录音在普通办公室或宿舍进行，对话者身带无线话筒，无线录音设备放置在另外的房间，这就保证了对话双方完全进入自然谈话状态。每一对发音人的谈话时间在 1 个小时左右。

本文中用的语料是从 SET2 中随机取的一部分语料，总时长为 41 分钟，包括各种音节共 11131 个，发音人为两位女发音人。语料首先进行了文本转写，用 xWaves+进行音段和韵律标注，并用 praat 软件提取出基频曲线，对其进行手工修改，以去掉某些错误点。

2．2 语料的转写和标注

我们对选出的语料进行了文字转写和人工标注，标注包括音段标注和韵律标注，音段标注采用 SAMPA-C 音段标注系统，韵律标注采用 C-ToBI 韵律标注系统，详见[3]。

标注共 7 层：正则的音节和声调标注、声韵母标注、韵律结构标注、重音结构标注、语句功能类型标注、杂类标注和话轮标注。

在声韵母标注层，用 SAMPA-C 音段标注系统来标注实际发音，包括标注超音段特征（声调的变化、上上相连的变调和轻声变化）和音段特征（增音、减音、浊化、清化、喉化、送气化、成音节化、音素音变等等）；杂类层主要标注背景噪音、口语现象等非语言学和副语言学现象，由于口语语气词和叹词在口语功能研究中的重要性，我们也在这层中标出。另外，我们还对 CADCC 的所有语音都进行了汉字转写，并将口语的非语言学和副语言学现象也按照特定的符号（参见[1]）进行了转写，在文字转写中还标记了语篇话题（非句子话题）转换的位置。

3．韵律边界的声学表现

3．1 韵律边界处的停顿统计

图 3.1 是自然口语语料中边界处的停顿个数统计，图 3.3、图 3.4 分别是对朗读语料中两位发音人（一男一女）

fig3.1 自然口语中边界处停顿统计

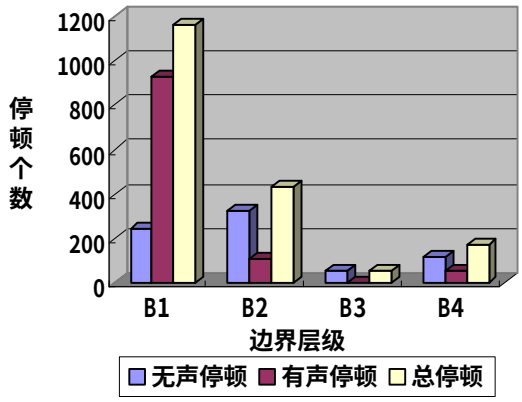


fig3.2 边界处的停顿时长

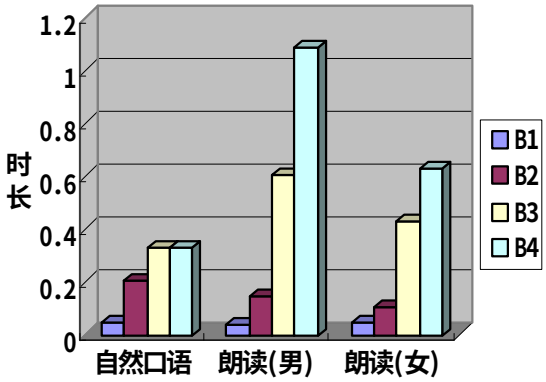


fig3.3 朗读语料边界停顿统计（男）

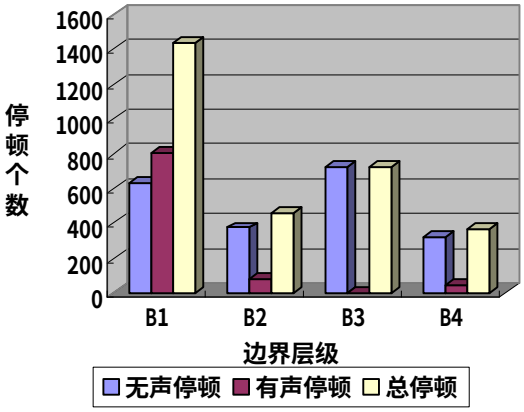
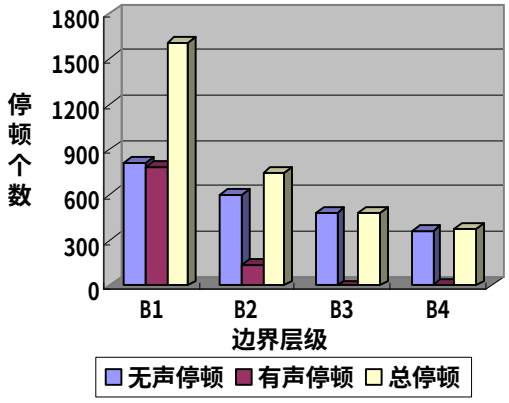


fig3.4 朗读语料边界停顿统计(女)



的语料中的边界停顿个数的统计。图 3.2 是对两种语体中的边界停顿时长的统计。其中，B1、B2、B3、B4 分别代表韵律词边界、韵律小短语边界、韵律大短语边界、语调组边界。

由图可得以下结论：

(1) 各种韵律边界后的停顿出现率在朗读和自然口语中不同。朗读语篇的停顿出现次数从大到小是 B1>B3>B2>B4 (男发音人 m002)；B1>B2>B3>B4 (女发音人 f002)；自然口语的停顿出现次数从大到小是 B1>B2>B4>B3。特别是自然口语中 B3 出现数目极小，仅占 3%。B1 在两种语体中出现的次数都很多，口语中占 64%，朗读中占 50%。

(2) 间断对应无声停顿(SP: Silent Pause)和有声停顿(FP: Filled Pause)两种[4]。随着边界级别增加（B1,B2,B3），边界后面出现 FP 的数目或可能性依次递减，就是说出现 SP 的数目或可能性不断增加。对于自然口语，B1 边界后面出现 FP 的概率约为 80%，B2 边界后面出现 FP 的概率约为 25%，B3 边界后面出现 FP 的概率约为 4%。对于朗读话语来说，B1 边界后面出现 FP 的概率约为 60%，B2 边界后面出现 FP 的概率约为 20%，B3 边界后面出现 FP 的概率几乎为 0。

(3) 对于 B4，可能对应语段尾或话轮转换边界，在这两种情况下，边界后面可能没有静音，但是也不能称之为有声停顿。

(4) 边界后的无声停顿时长，随着边界级别的增加而递增。

3．2 边界前后各韵律单元时长统计

3．2．1 边界前后的音节时长统计

本文分别用三种归一方法对边界前后的音节时长进行了统计，三种方法分别是按所有音节的时长均值归一、对每个音节分别归一和按照声韵母单元进行归一[5]。结果发现，前两种方法相比，方法一得到的边界前音节时长方差较小，方法二得到的边界后音节时长方差较小（边界 4 除外），而方法三则在所有情况下方差均为最小，因此，方法三为最优方法。本文以下所进行的时长统计均是采用方法三即按声韵母单元进行归一的方法来做的。

除此之外，本文还对所有非边界音节的时长进行了统

计，并结合边界前后音节时长的数据作了图 3.5。由图可见，除边界 1 外，边界前后音节的时长均值均大于非边界音节时长均值。边界 1 前音节的时长均值较小，怀疑是受轻声音节影响所致，从音节统计结果发现，各边界前的轻声音节所占

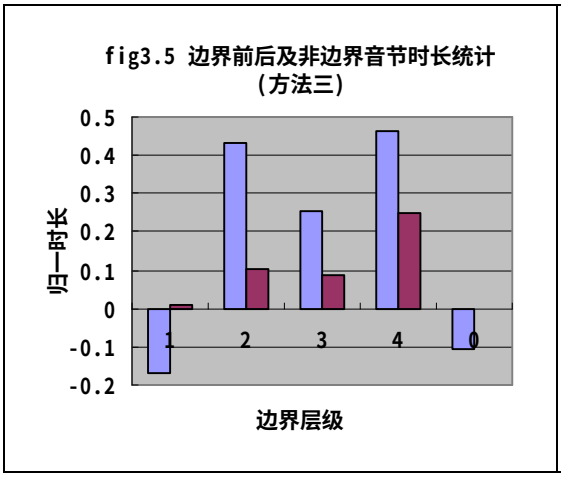


fig3.5 边界前后及非边界音节时长统计（方法三）

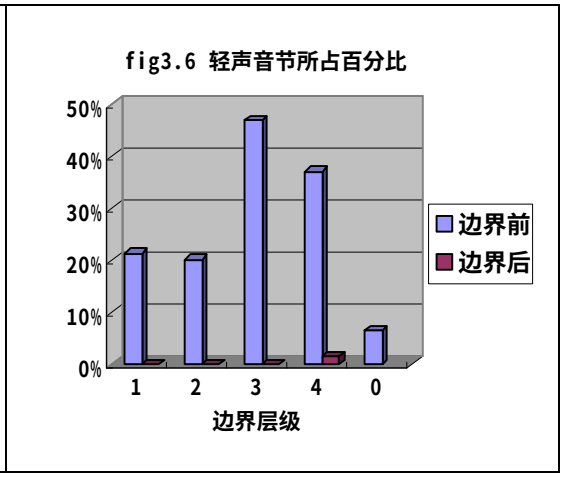


fig3.6 轻声音节所占百分比

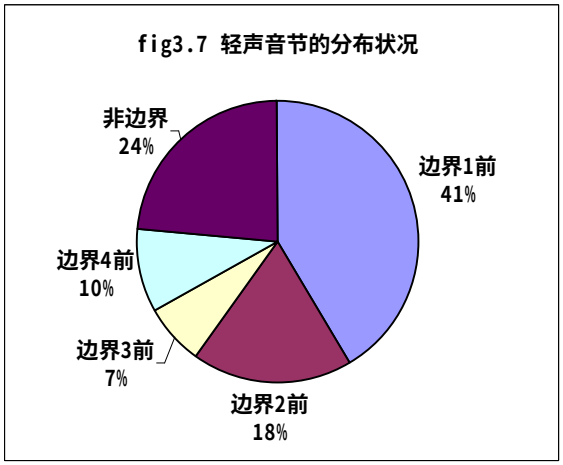


fig3.7 轻声音节的分布状况

百分比都较大，均在 20%以上（见图 3.6），而由轻声音节的分布情况（图 3.7）可知，分布在边界 1 前的轻声音节占到了所有轻声音节的 41%，而其他边界前后的轻声音节所占比例均低于此数字，这应该是边界 1 前音节时长均值较小的一个重要原因。另外还可能是因为在边界 2、3、4 前的音节大都有不同程度的拉长，故其时长均值仍较大，而边界 1 前的音节则一般不会有明显的拉长，所以时长均值较小。

3．2．2 边界前后不同声调的音节时长统计

为了研究边界对不同声调的影响情况，本文对边界前后以及非边界音节的时长按不同声调进行了分类统计，并据此绘制出图 3.8～3.13，其中 A 图为发音人 A 的语料归一后的数据，B 图为发音人 B 的语料归一后的数据（下同）。由于统计语料有限，边界 3、4 出现较少，尤其是在边界 3 和边界 4 后，因为大多为无声停顿，数据非常稀少，故不予考虑。

由图可知，阴平、阳平、去声、轻声在边界前后和非边界处的时长表现基本一致，在边界 2、3、4 前均比非边界处有较大拉长，在边界 2 后也有拉长（轻声除外），但程度较小，在边界 1 前后则有轻微缩短；上声则略有不同，上声在各边界前与非边界处相比均有不同程度的拉长，在边界 2 后也有一定拉长，边界 1 后则没有明显变化。

另外，本文还对每个边界前后和非边界时各种声调的时长进行了比较，发现并非像以前所认为的“上声比其他声调都长”，而仅在边界 1 前后、边界 2 后和非边界处上声最长；边界 2 前各声调差别不大；边界 3 前阳平最长，上声次之；边界 4 前则是去声和阳平最长，上声最短。

3．2．3 边界前后不同声韵母时长统计

传统语音学认为，语句中主要是韵母时长发生变化，声母时长无明显变化。为验证这一点，本文对不同边界前后和非边界处的声韵母时长进行了分类统计，见图 3.14～3.15[①]。

由图可知，与非边界处相比，单元音和二合元音在边界前后均有拉长，边界前拉长更显著，三合元音则在边界前有较大拉长，边界后没有明显变化。声母的变化较为复杂，由图 3.14 可见，除鼻声母在边界前略有拉长外，其他各类声母在边界前均有不同程度的缩短，塞音缩短程度最大；而边

[①]这是根据两位发音人的语料得出的数据，没有进行发音人归一，事实上，作者按发音人归一也得出了类似的结论，限于篇幅，此处不再给出。

界后的所有类别声母均有不同程度的拉长，其中塞音、塞擦音、边音拉长程度较大，浊擦音最小。由此看来，在自然口语中，塞音和塞擦音声母时长也会发生变化，这与传统的语音学观点不同。

fig3.8A 阴平在边界前后的时长统计
(发音人A)

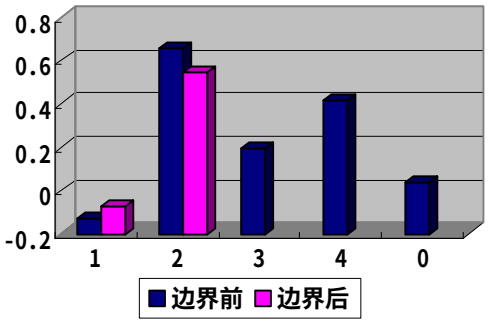


fig3.8B 阴平在边界前后的时长统计
(发音人B)

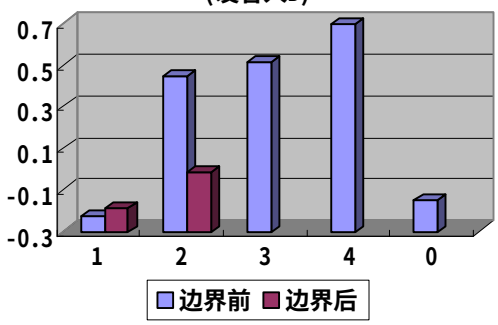


fig3.9A 阳平在边界前后的时长统计
(发音人A)

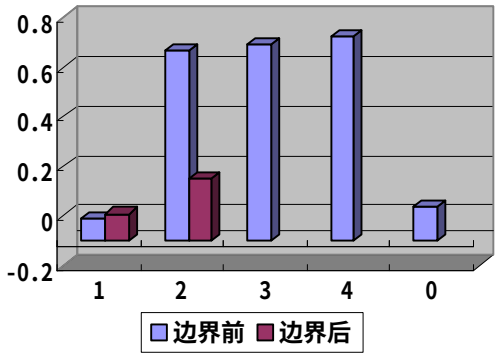


fig3.9B 阳平在边界前后的时长统计
(发音人B)

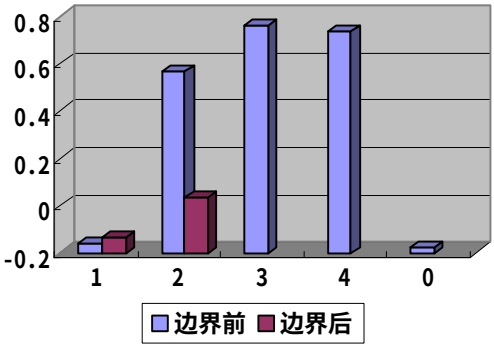


fig3.10A 上声在边界前后的时长统计
(发音人A)

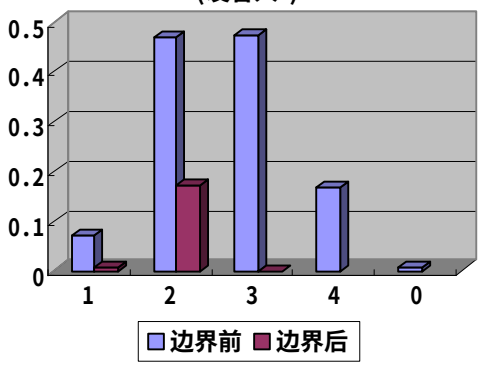
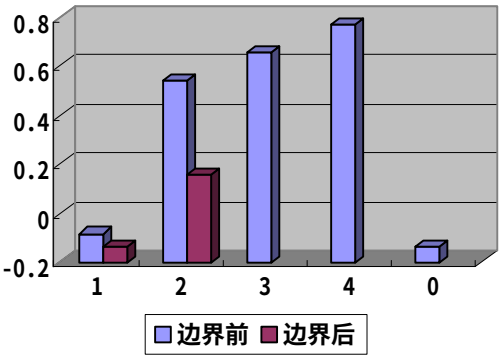
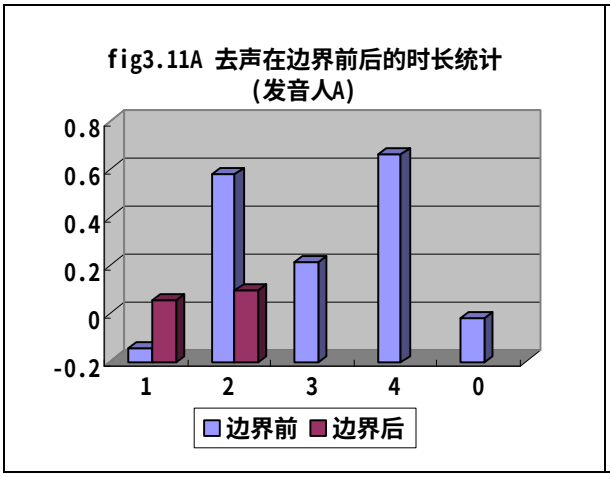
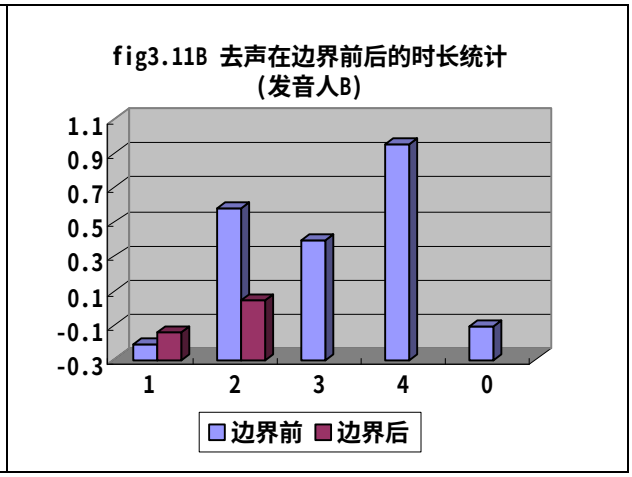


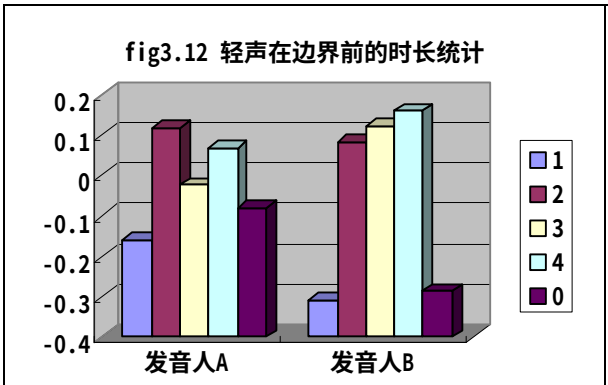
fig3.10B 上声在边界前后的时长统计
(发音人B)

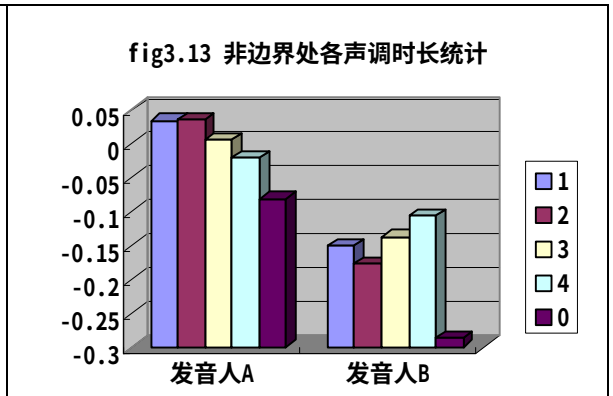


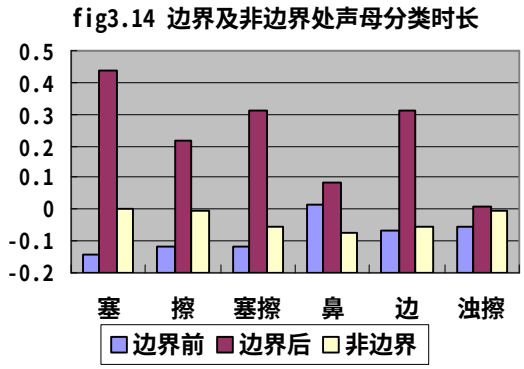


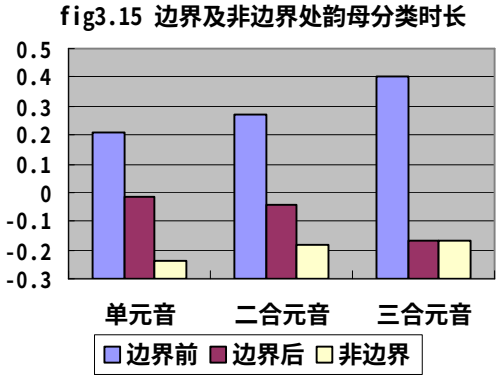


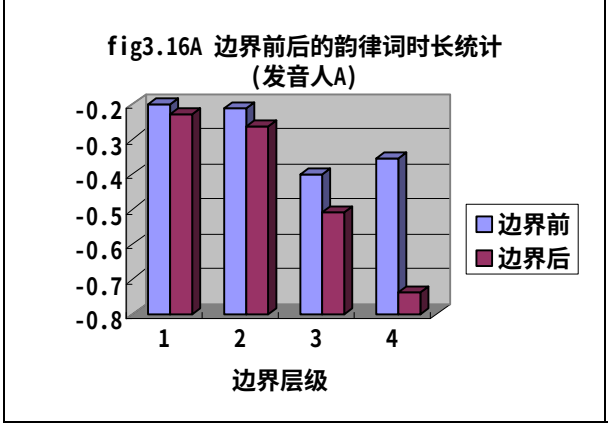
(注：图中自左向右依次代表边界 1、2、3、4 前后的音节时长，由于边界 3、4 后数据过于稀疏，故未画出)

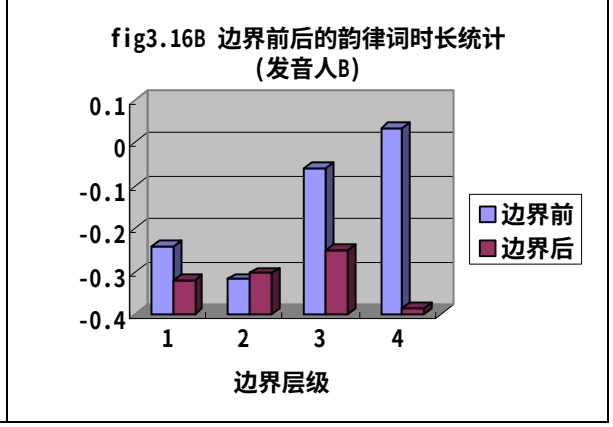


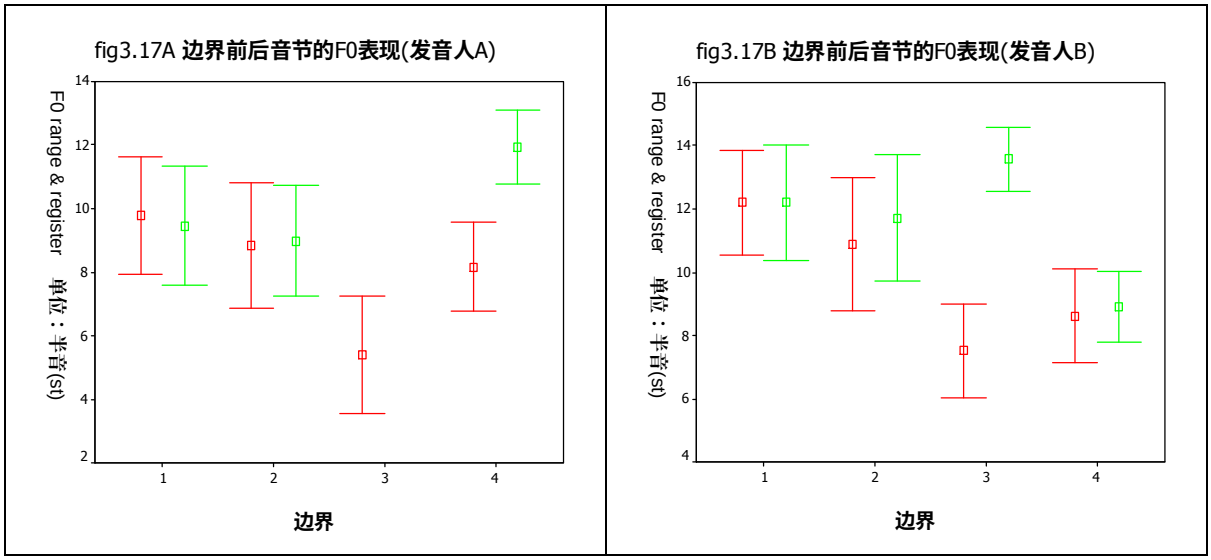




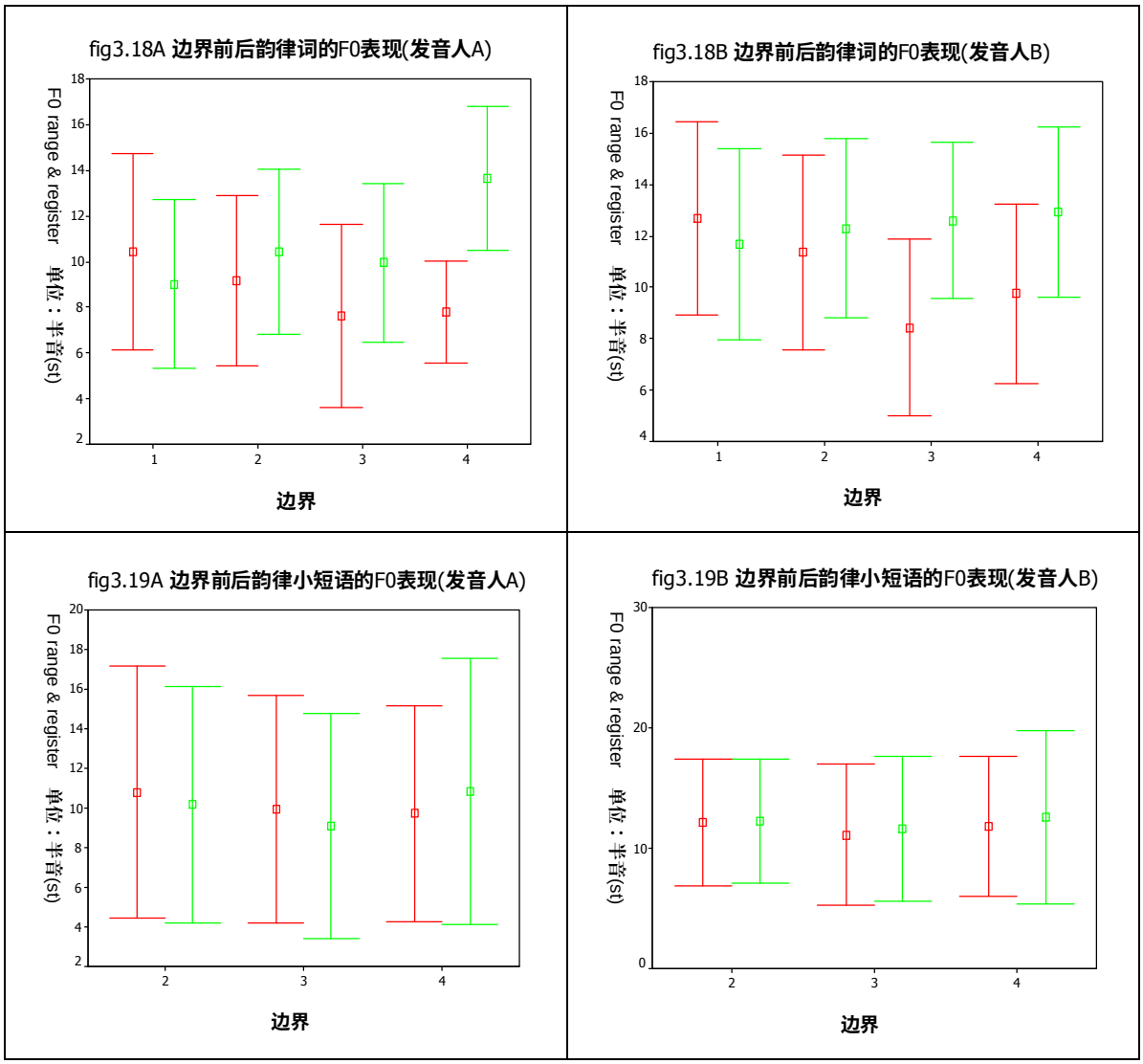








（注：图中自左向右依次代表边界 1、2、3、4 前后的 F0 表现，由于边界 3 后数据过于稀疏，故此处未画出其 F0 情况）



（注：图中自左向右依次代表边界 1、2、3、4 前后的 F0 表现）

3．2．4 边界前后的韵律词时长统计

考察过音节和半音在边界前后和非边界处的时长表现以后，本文又对边界(2、3、4)前后的韵律词时长进行了统计，见图 3.16。

观察图 3.16A 和 3.16B 发现，发音人 A 和发音人 B 也有差别，发音人 A 的韵律词在边界前均比边界后时长要长，且边界 1、2 前后的韵律词时长明显大于边界 3、4 前后的时长；而发音人 B 尽管也是在各边界前都比边界后要长（除边界 2 处略有缩短），但边界 1、2 前后的韵律词时长却明显小于边界 3、4 前后的时长，这一点和发音人 A 恰好相反；而有一点两人是相同的，即边界 1 前比边界 2 前长，边界 4 前比边界 3 前长。

3．3 边界前后各韵律单元的 F0 表现

边界对各个韵律单元的影响不仅体现在时长的拉长与缩短，还会体现在基频的变化上，包括音域的扩大与缩小，音阶的升高与下降等等。本文分别统计了边界前后各种韵律单元的音阶与音域，并将其转化为半音，考察每种边界对韵律单元的影响。

3．3．1 边界前后音节的 F0 表现

由图 3.17 可知，边界前后音节的 F0 范围没有太明显变化，边界 2、4 后略有缩小；而边界 2、4 后音节的 F0 中值均有升高，边界 1 后则为降低。可见，边界 2、4 前后音节有轻微的 F0 重设，但不明显，若数据充足的话，边界 3 的情况应该也与边界 2、4 类似。

3．3．2 边界前后韵律词的 F0 表现

由图 3.18 可以看出，音阶方面，两位发音人在边界 2、3、4 前后音阶均有不同程度的升高，尤其是边界 3 和 4，而在边界 1 处则为下降；音域方面，两人在边界 2、3 后音域均缩小，在边界 4 后发音人 A 音域加大，而 B 则有轻微缩小，边界 1 处变化不明显。

3．3．3 边界前后韵律小短语的 F0 表现

观察图 3.19A 和图 3.19B 可知，与韵律词相比，韵律小短语在各边界前后的 F0 变化不明显，仅在边界 4 处有不大的音阶升高和音域增大。

由上述结果可见，边界 2、3、4 后均会发生基频重设，且级别越高，重设程度也越大，基频重设仅限于边界前后的韵律词之间。这同时也表明，比韵律词大的韵律单位内部都有 F0 下倾(declination)现象发生。

3．4 小结

本节主要考察了韵律边界处的停顿和边界前后各级韵律单位的时长情况，得出以下结论：

a. 边界后出现无声停顿的可能性及其时长与边界级别有关；
b. 除韵律词边界前音节时长缩短外，其余各级边界前后的音节时长均有不同程度的拉长；
c. 边界前后各种声调的时长表现基本一致（轻声、上声在韵律词边界前略有不同）；上声并非在所有情况下都比其他声调长；
d. 各种声母（包括不送气塞音、塞擦音）在边界前均有缩短，在边界后均有拉长，韵母在边界前后都有显著拉长；
e. 韵律词在韵律短语以上级别的边界前有拉长；
f. 除韵律词边界外其余各级边界前后均有 F0 重设现象，且级别越高，重设程度越大；F0 重设仅限于边界前后紧邻的韵律词之间；比韵律词大的韵律单位内都存在 F0 下倾现象。

本节结论对韵律识别的启示：

A. 首先可以根据无声停顿、前后音节有无拉长、有无基频重设、是否轻声音节以及是否语段尾或话轮转换处来确定出边界位置；
B. 一般说来，可以根据无声停顿最长来确定该边界是否 B4，但当对应语段尾或话轮转换处时后面可能没有静音，这时可以根据有无基频重设来判断是否属于话轮转换，若是则可以定为边界 B4；
C. 根据较大的停顿、拉长和基频重设且不是语段尾或话轮转换处来确定 B3；
D. B1 的判断依据可以定为：没有或只有极小的无声停顿，前音节时长较短，没有明显的 F0 变化；
E. 其余边界位置若无明显不同则为 B2。

4．重音的声学表现

研究表明，时长和基频（F0）对重音起主要作用[6][7][8]，尤其是时长，被认为是影响重音的最主要因素，本节就从这两个方面对各级重音（1、2、3）进行了统计分析。为处理方便，这里所统计的时长和基频表现均是指该级重音所在的音节。

4．1 各级重音的时长统计

图 4.1A 和图 4.1B 分别是对两位发音人的语料归一得到的各级重读音节时长的统计数据。由图可知，发音人 A 和发音人 B 在 2 级重音和 3 级重音的时长表现基本一致，都比非重音音节有明显拉长，1 级重音则有不同表现，发音人 B 的 1 级重音也比非重音音节有较大拉长，而发音人 A 的 1 级重音反而比非重音音节有明显缩短，这表明词重音的声学表现受发音人说话习惯影响较大，有的发音人是靠时长拉长来实现重音，有的发音人则是利用其它特征的变化（如基频）来实现重感。

4．2 各级重音的 F0 表现

图 4.2A 和图 4.2B 分别是对两位发音人的语料归一得到的各级重读音节 F0 表现（音阶和音域）的统计数据，图

中自左向右依次表示词重音、韵律小短语重音、韵律大短语重音和非重读音节的 F0 数据（分别以 1、2、3、0 表示）。

很容易看出，两位发音人在各级重音处的 F0 表现模式基本一致，各级重音之间的音阶和音域的变化趋势非常接近：各级重音的音域都比非重读音节的音域要大，其中 3 级重音尤为显著；1 级重音和 3 级重音的音阶都明显高于非重读音节的音阶，而 2 级重音的音阶却略低于非重读音节。

这里，结合上面的时长统计结果，我们发现，3 级重音则是时长拉长音域加大音阶也升高，2 级重音的时长拉长音域加大而音阶降低， 1 级重音除了音域加大和音阶升高外，时长的变化因发音人而异。由此看来，不同级别的重音有不同的实现方法，同一级别的重音不同的发音人也有不同的实现策略，要想弄清楚各级重音的产生机制到底是怎样的，还需要更多的语料和更进一步的研究，这也是我们下一步努力的方向。

4．3 小结

本节分别对各级重音的时长和 F0 表现进行了统计分析，总结出了各级重音的声学特征，并发现不同级别的重音有不同的实现方法，同一级别的重音不同的发音人也有不同的实现策略，这一点应当引起言语工程学家的注意，在合成和识别重音时对各个级别要区别对待。

通过分析，这里给出几条重音识别的参考规则：

a. 3 级重音在时长、音阶、音域上都明显高于其他级别，可以作为识别的依据；

b. 2 级重音时长和音域都较大，仅次于 3 级重音，但音阶下降，这里不能肯定这是普遍现象，但可以作为判断标准之一，当然也可以把相反的情况即“ 音域减小而音阶升高 ”作为可能的标准之一进行考虑；

c. 1 级重音即词重音较难分辨，必须综合考虑时长、音域、音阶各方面的表现。

fig4.1A 各级重音的时长统计(发音人A)

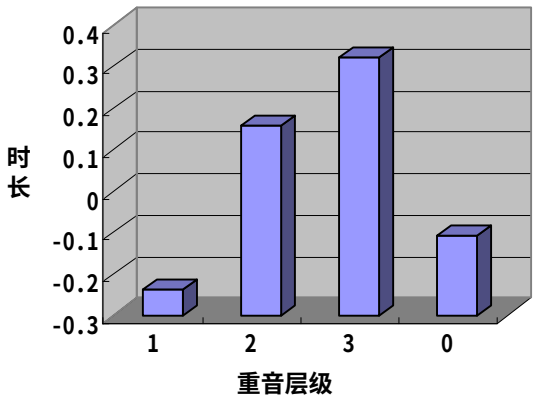


fig4.1B 各级重音的时长统计(发音人B)

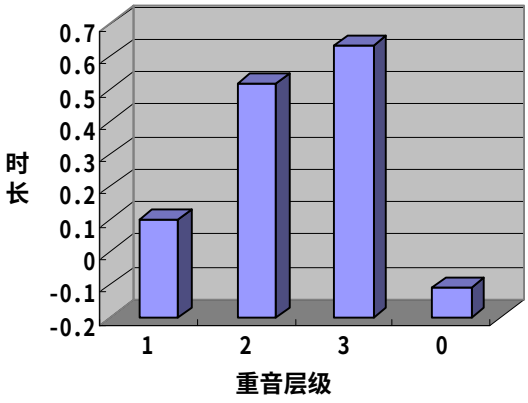


fig4.2A 各级重音的F0表现(发音人A)

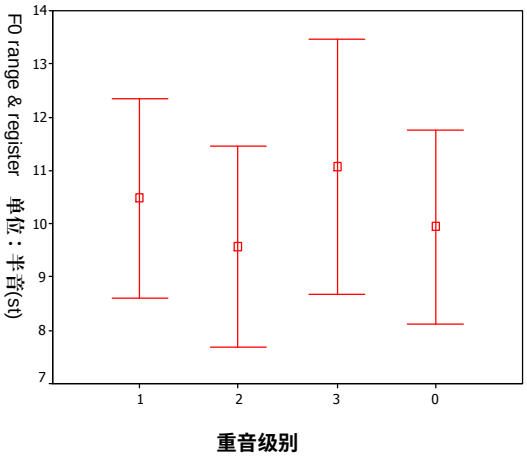
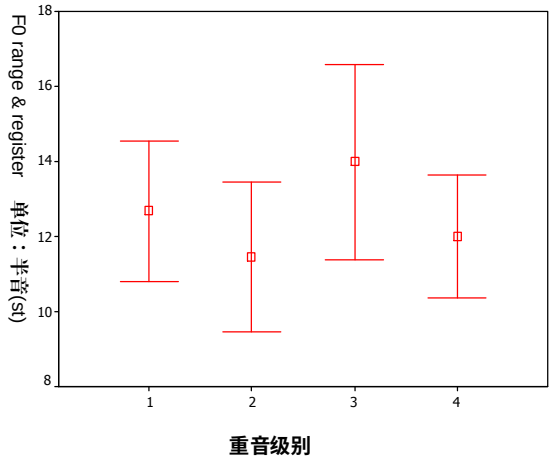


fig4.2B 各级重音的F0表现(发音人B)



5．结论和讨论

本文以自然口语对话语音语料库 CADCC 为基础，对自然口语语料中各级韵律边界和各级重音的声学特征进行了统计分析，包括韵律边界处的停顿、边界前后各级韵律单元的时长和音高(F0)表现、各级重读音节的时长和音高(F0)表现等，并在此基础上初步得出了自动识别韵律边界和重音等

级的一些规则。同时，从分析结果看出，韵律词边界的判定和 1 级重音的识别是韵律识别中的两个难点，还需要更多学者的进一步努力才能解决。

参考文献

[1] 刘亚斌，李爱军，《朗读语料与自然口语的差异分析》，《中文信息学报》，2002 年第 1 期。(另见本书)

[2] 李爱军，徐波等，《口语对话语音语料库 CADCC 和其语音标注》，第五届全国现代语音学学术会议，《新世纪的现代语音学》，蔡莲红，周同春，陶建华编，清华大学出版社，2001。

[3] 陈肖霞，祖漪清，李爱军，《对汉语普通话正则标音系统的探索》，第四届全国现代语音学学术会议，《现代语音学论文集》，吕士楠等编，金城出版社，1999。

[4] 林茂灿，《普通话语句中间断和语句韵律短语》，《当代语言学》，2000 年第 4 期。

[5] 殷治纲，《基于语料库的时长规整化研究》，第五届全国现代语音学学术会议，《新世纪的现代语音学》，蔡莲红，周同春，陶建华编，清华大学出版社，2001。

[6] CAO Jianfen, LV Shinan, YANG Yufang, 《Chinese prosody and a proposed phonetic model》, *Report of Phonetic Research*, 2000.

[7] 颜景助，林茂灿，《北京话三字组重音的声学表现》，《方言》，1988 年第 3 期。

[8] 沈炯，J.H.v.d.Hoek，《汉语语势重音的音理》，《语文研究》，1994 年第 4 期。

[9] 李爱军，《普通话对话的韵律短语和语句重音的声学分析》，第四届全国现代语音学学术会议，《现代语音学论文集》，吕士楠等编，金城出版社，1999。

An Analysis of the Prosodic Features of Chinese Spontaneous Speech

LIU Yabin, LI Aijun

Institute of Linguistics, CASS

lyabin@sina.com *liaj@linguistics.cass.net.cn*

This paper analyzes the acoustic performance features of prosodic boundaries and stresses at all levels in Chinese spontaneous speech using statistical method, and manages to find some rules so as to help achieve automatic detection of the prosodic boundaries and stresses in Chinese spontaneous speech.

This research is based on the Chinese Annotated Dialogue and Conversation Corpus (CADCC), which is an authentic corpus consisting of recordings and transcripts of more than 15 hours of natural, spontaneous conversations produced by 13 pairs of speakers. We have selected and annotated manually about 1 hour of the transcript of the recorded conversations, using 8 tiers of segmental and prosodic labels. The hierarchies of prosodic labeling include the boundaries of prosodic words, minor prosodic phrases, major prosodic groups and intonation groups whereas the hierarchies of stress labeling include the stress of prosodic words, minor prosodic phrases, major prosodic groups and intonation groups. All statistical data are based on the annotated part of the corpus.

With an array of statistical analysis of the acoustic performance preceding or following the boundaries of all levels, this study investigates the segmental and suprasegmental characteristics, which include the duration of all segments and prosodic units and the F0 resetting. The preliminary results are:

a. The occurrence frequency and the length of the silent pause following boundaries are both relevant to the boundary levels;

b. The duration of the syllables preceding/following boundaries are lengthened except preceding the word boundary;

c. The syllable duration of different tones has the same changing tendency preceding/following boundaries; the third low level tone is not always longer than all the other three tones;

d. Initials including stops and affricates are shortened/lengthened when they're preceding/following boundaries while finals are lengthened in both positions;

e. The duration of the prosodic words preceding boundaries is lengthened;

f. F0 resetting is observed at boundaries for all prosodic units except the prosodic word.

An analysis on acoustic features of stresses is also proposed and some results are listed here: a. the prosodic phrase stress is usually accompanied by the changing of both duration and F0; b. the prosodic word stress is more complicated and seems to be speaker-dependent.

Finally, some rules of automatic prediction of prosodic boundaries and stresses are presented.