普通话韵律短语的时长特性分析

李爱军

摘要 语句和篇章的韵律结构和信息结构的分析以及其模型化，是提高语音合成自然度的关键。我们认为汉语普通话的韵律结构是分层次的，由音步韵律词,韵律短语和语调短语组成。韵律短语可以按照感知分为两类: minor phrase (通常短语后感知到的停顿很短，使两个短语之间联结较紧密)和 major phrase (通常短语后感知到的停顿很明显)。本文报告对 145 个朗读语句的韵律结构进行标注后，minor phrase 和 major phrase 的时长特性及其音节数的统计分析结果。

1.引言

对句子进行句法分析，可以得到它的句法成分(constituent)，当我们念这句话的时候，却是按照其韵律结构来发音的。Selkirk(1984)提出了一种严格的韵律分层理论，她认为韵律结构从低到高的分层依次是音步 foot 、音节 syllable、音系词 phonology word (韵律词 prosodic word)、韵律音系短语 phonology phrase (韵律短语 prosodic phrase) 和语调短语 intonational phrase; 句子的韵律结构和句法结构之间存在着系统的影射关系。韵律结构有一些独立的完备的原则，这些原则在我们给一个句子分配其韵律结构时，是与句法到韵律的映射原则结合使用的(Selkirk & Shen 1988)。

韵律分段结构为语音合成的自然度和可懂度提供重要信息。例如韵律分段边界将一个语句分为几个更易处理的较小片段。对于句法上歧义的句，有些可以通过设置不同的韵律短语边界位置来排歧，如“下雨天留客天天留我不留”(Shen 1992)。

短语的边界在感知上也可以进行分类，与感知相关的物理量除了声调和语调特征以外，通常还与这样一些语音信号的物理特征相关(李爱军 1997)：停顿(pause)、短语的最末一个音节时长变长(final lengthening)等。

在文语转换中，如果能对韵律边界位置进行精确的预测的话，就能正确分配控制各级韵律分段的声学参数，如 F0，时长和幅度等。我们的最终目的就是要生成一个韵律分析器(prosodic parse)，对无限文本进行韵律结构和信息结构的分析，以产生最自然的语音。

国外对韵律短语边界的研究很多，如采用分级随机模型或统计模型，利用语言学知识自动预测韵律的边界位置(Ostendorf & Veileux 1994; Sanders & Taylor 1995)；采用 CART(Classification and Regression Tree)技术对手工标注的韵律边界和文本之间建立模型(Hirschberg 1991)；采用随机有限状态模型对汉语进行切词(Richard & Shih 1996)。

在汉语普通话韵律方面，有关于韵律词的研究(冯胜利 1997) 和关于连上变调的研究(Chilin Shih 1986，沈炯 1997)，Chilin Shih 提出了汉语普通话韵律的辖阈，她认为普通话的音步(foot)的组成分为三个步骤：先由两个直接成分(direct constituents)组成 disyllabic foot，在从左到右将单音节成分结合为 disyllabic foot，最后将剩下的单音节与邻近的 feet 结合。但对韵律短语边界的研究很少，尚不清楚普通话的韵律短语的边界的声学相关量。所以，目前很难象

英语那样从语调和声调上来定义韵律短语(Pierrehumbert 1980)，只能利用感知的方法来划分韵律短语的边界。

韵律短语按照感知分为两类: minor phrase 和 major phrase。本文报告对 145 个朗读语句的韵律结构进行标注后，minor phrase 和 major phrase 的时长特性及其音节数的分析结果。

2 . 韵律短语的标注和有关声学参数的测量

145 句普通话的语音材料选字 863 识别语音库中的朗读语句(祖漪清 李爱军 1997, Zu Yiqing 1997)，发音人为男性 M01。

在 Pierrehumbert 的语调理论中(Pierrehumbert 1980)，从音系上将短语分成两级，每种短语都是由 F0 曲线上高和低声调序列组成，这两级短语是：intermediate (或 minor) phrase 和 intonational (或 major) phrase。一个 intermediate (或 minor) phrase 包括一个或多个 pitch accent 加上一个 phrase accent。intonational (或 major) phrase 包括一个或多个 intermediate(或 minor) phrase 加上一个 boundary tone。因此一个 intonational 短语的边界一定是 intermediate 短语的边界，反之则不成立。

汉语普通话的韵律短语，不好从声调和语调上来进行定义，因此在标注韵律短语时，我们是依据感知结果来标注的。也将韵律短语按照感知分为两级：minor phrase 和 major phrase。如果短语后感知到的停顿较短，它与后面一个短语的连接较紧密，则将此短语标为 minor phrase(|)；如果短语后感知到的停顿较长，它与后面一个短语的连接较松散，则将此短语标为 major phrase(||)。 例如：

(1) 投资/数以/亿元/计的||“八五/计划/”工程||。
(2) 一个/又一个地||从连夜/熬出的/施工/蓝图||走向/工地||。
(3) 北约/欧洲/盟军||最高/司令|访问/俄罗斯||。
(4) 双方/达成|互设/代表/机构/协议||。
(5) 倒/确实令/法官们|有些/头疼||。
(6) 多么/值得/我们/深思啊||!
(7) 车臣/居民||至今/未能按|俄联邦的/法律||表达/意愿||。
(8) 决定/自己|共和国的/命运||。
(9) 国际/航空/公司|飞上海的/航班||因/大雾/取消了||。
(10) 远征/听后||急得/一下子/跳了起来||
注：其中“/”是韵律词边界的标志，本文不对韵律词进行研究。

除了标注韵律短语以外，还测量了短语的时长、短语最后一个音节的时长以及短语所含的音节数。如果短语起始音节是辅音，它的闭塞段时长不好测时，我们采用一个平均的闭塞段时长(陈肖霞 本刊)。另外，这里测得的 major phrase 的时长和音节数是 major phrase 边界到前一个边界之间的时长和音节数，但不包括停顿的时长。

3 . 韵律短语的时长和音节数的统计分析

3 . 1 韵律短语边界前一个音节时长的统计分析

所有的短语边界前一个音节, 按其不同的声调计算它们的平均时长，用 T0、T1、T2、 T3 和 T4 表示, (0 表示轻声或轻读, 1-4 表示四个声调), 有 T0<T4<T3<T1<T2(见表 1)，它们之间有明显的差异(F(4,348)=54.506, P<0.001，见表 2)。对它们两两进行 T 检验,发现一声和二声音节的时长,一声和三声音节的时长之间无显著性差异(P>0.1), 其它声调之间都有显著差异(见表

6)。

minor 短语边界前一个音节的平均时长关系也是 T0<T4<T3<T1<T2(见表 1)，它们之间有明显的差异(F(4,59)=15.461, P<0.001，见表 2)。短语边界前一个音节两两 T 检验的结果是一声和二声,一声和三声,一声和四声,二声和三声,二声和四声以及三声和四声音节的时长之间都没有显著的差异,其它声调之间都有显著差异(见表 6)。

major 短语边界前一个音节的平均时长关系也是 T0<T4<T3<T1<T2(见表 1)，它们之间有明显的差异(F(4,283)=43.073, P<0.001，见表 2)。前一个音节两两 T 检验的结果是一声和三声的时长无明显差异,其它声调之间都有显著差异。

3.2 韵律短语边界前一个音节的平均时长与所在短语的音节平均时长

边界前一个音节的时长与所在短语的音节的平均时长相比(见表 11), 轻声音节在 minor 和 major 短语中都短于短语的平均音节时长，并且有显著的差异(P<0.001)。其它声调的音节与短语的平均音节时长，除了 minor 短语边界前的去声以外,都有显著的差异(P<0.05)，但是 major 短语前的去声的时长却小于短语平均音节时长，其它情况下，边界前一音节的时长，均大于短语平均音节时长。minor phrase 和 major phrase 的末音节的平均时长则没有显著差异(P=0.305>0.1)。 总体来讲，除了轻声音节和 major phrase 末音节为去声音节以外，韵律短语边界前的音节有时长变长的特性。

3.3 短语边界前相同声调的音节时长的两两 T 检验

minor phrase 和 major phrase 边界前相同声调的音节平均时长的两两 T 检验结果表明(见表 7)，它们都没有显著性差异。 这说明，minor 和 major 短语不是靠边界前一音节的时长的不同来区分实现的，而是通过边界后的停顿和基频曲线等其它的韵律特征来区分实现的。

3.4 音节时长的统计分析

所有韵律短语（不区分 minor 和 major phrase）、minor phrase 和 major phrase 音节平均时长分别是 0.217S、0.212S 和 0.218S(见表 4)，并且它们之间没有显著性差异(见表 9)，T 值分别是 1.077、-0.355 和 -1.246，P 值都>0.1。

3.5 短语的平均时长

所有短语、minor phrase 和 major phrase 平均时长分别是 1.226S、1.069S 和 1.261S (见表 5)，并且 minor phrase 的平均时长小于 major phrase 平均时长，它们之间有显著性差异(见表 10,T=-3.409, P<0.01)。

3.6 短语的音节平均个数

在对语句进行韵律短语边界的自动切分时，音节数应该作为切分模型的非常重要的约束条件之一。

所有短语、minor phrase 和 major phrase 音节平均个数分别是 5.782、5.141 和 5.924(见表 3)。所有短语平均音节数大于 minor phrase 的平均音节数，且有显著性差异(见表 8，T=2.251, P<0.05)；所有短语平均音节数小于 major phrase 的平均音节数，但没有显著性差异(P>0.1)；minor phrase 的平均音节数大于 major phrase 的平均音节数，且有显著性差异(T=-2.706, P<0.01) 。

这说明朗读语句的韵律短语的音节个数(5-6 个)比对话和即兴话语的韵律短语的音节个数

(3-4 个)要多，在做韵律短语的自动切分时，应该根据不同的语体，选用不同的制约参数。
表格中符号说明：“*”：概率 P<0.05；“**”：概率 P<0.01；“***”：概率 P<0.001；“PP”、“P|”和“P||”分别表示所有短语、minor phrase 和 major phrase。

表 1. 不同短语边界前一个音节,按不同的声调统计的时长信息(时长单位为秒 S):

	声调	音节数	时长总和	音节平均时长 Ti	标准差
PP	不计调	353	76.508	0.217	0.049
	0	58	9.349	0.161	0.054
	1	47	11.387	0.242	0.050
	2	68	17.291	0.254	0.047
	3	45	10.488	0.233	0.039
	4	135	28.002	0.207	0.043
P|	不计调	64	13.524	0.2113	0.117
	0	24	3.874	0.161	0.069
	1	7	1.923	0.274	0.043
	2	11	2.624	0.239	0.059
	3	8	1.975	0.247	0.053
	4	14	3.128	0.223	0.044
P||	不计调	289	62.984	0.218	0.055
	0	34	5.475	0.161	0.041
	1	40	9.455	0.236	0.047
	2	57	14.667	0.257	0.044
	3	37	8.513	0.230	0.035
	4	121	24.874	0.205	0.043

表 2. 三组短语边界前不同声调音节时长的 F-test:

	F	P
PP	F(4,348)=54.5063	$4.44*10^{-11}$ ***
P|	F(4,59)=15.460671	$1.7158*10^{-6}$ ***
P||	F(4,283)=43.073078	$2.4112*10^{-10}$ ***

表 3. 不同短语中音节个数的统计

	音节总数	平均音节数	短语个数	标准差
PP	2041	5.782	353	2.113
P|	329	5.141	64	1.999
P||	1712	5.924	289	2.115

表 4. 不同短语中音节时长统计

	音节时长总和	音节平均时长	短语数	标准差
PP	76.523	0.217	353	0.035
P|	13.598	0.212	64	0.029

P‖	62.925	0.218	289	0.036

表 5. **不同短语时长统计**

	短语时长总和	**短语平均时长**	**短语数**	**标准差**
PP	432.815	1.226	353	0.414
P|	68.423	1.069	64	0.387
P‖	364.392	1.261	289	0.412

表 6. **三组短语边界前不同声调音节平均时长的两两间** t-test:

PP	**声调-声调**	T	**自由度** df	P
	0-1	-7.898686	103	$2.5848*10^{-7}$ ***
	0-2	-10.33626	124	$1.276*10^{-8}$ ***
	0-3	-7.543798	101	$4.274*10^{-7}$ ***
	0-4	-6.293442	191	$2.049*10^{-6}$ ***
	1-2	-1.311279	113	0.189
	1-3	0.9597988	90	0.342
	1-4	4.603319	180	$6.249*10^{-5}$ ***
	2-3	2.483459	111	$1.386*10^{-2}$ *
	2-4	7.122761	201	$5.021*10^{-7}$ ***
	3-4	3.59235	178	$7.286*10^{-4}$ ***
P|	0-1	-4.0792	29	$5.5246*10^{-4}$ ***
	0-2	-3.239416	33	$3.033*10^{-3}$ **
	0-3	-3.21041	30	$3.425*10^{-3}$ **
	0-4	-3.014282	36	$4.878*10^{-3}$ **
	1-2	1.351487	16	0.1929
	1-3	1.072536	13	0.3035
	1-4	2.52188	19	1.9717
	2-3	-0.30415	17	0.6882
	2-4	0.7776389	23	0.4506
	3-4	1.143759	20	0.2655
P‖	0-1	-7.249578	72	$9.21*10^{-7}$ ***
	0-2	-10.32399	89	$2.02*10^{-8}$ ***
	0-3	-7.645645	69	$5.69*10^{-7}$ ***
	0-4	-22.85744	153	$4.85*10^{-12}$ ***
	1-2	-2.249718	95	$2.5165*10^{-2}$ *
	1-3	0.631241	75	0.536
	1-4	3.861649	159	$3.764*10^{-4}$ ***
	2-3	3.141046	92	$2.634*10^{-3}$ **
	2-4	7.47185	176	$3.151*10^{-7}$ ***

	3-4	3.222871	156	$1.94*10^{-3}$ **

表 7. **不同边界前相同声调的音节的时长** T-test

	PP **与** P|			PP **与** P||			P| **与** P||		
tone	T	df	P	T	df	P	T	df	P
0	0	80	0.304	0	90	0.304	0	56	0.304
1	-1.604004	52	0.111	0.5733	85	0.572	1.995	45	0.326
2	0.9472636	77	0.3489	-0.365876	123	0.676	-1.1733	66	0.243
3	-0.88552	51	0.384	0.36287	80	0.677	1.1322	43	0.263
4	-1.322476	147	0.185	0.3715	254	0.674	1.47944	133	0.137

表 8. **不同短语中音节数** T-test

PP **与** P|			PP **与** P||			P| **与** P|| **短语边界**		
t	df	p *	t	df	p	t	df	p**
2.2509	415	$2.34*10^{-2}$	-0.84678	640	0.402	-2.705835	351	$7.18*10^{-3}$

表 9. **不同短语中音节时长的** T-test

PP **与** P|			PP **与** P||			P| **与** P||		
t	df	p	t	df	p	t	df	p
1.077453	415	0.2816	-0.3555	640	0.679	-1.24633	351	0.211

表 10. **不同短语的时长的** T-test

PP **与** P|			PP **与** P||			P| **与** P|| **短语边界**		
t	df	p	t	df	p	t	df	p **
2.818	415	5.279	-1.068	640	0.286	-3.4095	351	$1.09*10^{-3}$

表 11. **短语边界前一不同声调的音节的时长与短语中音节平均时长** T-test

音节的声调-短语边界	T	df	P
0-p	-10.34173	409	$5.323*10^{-9}$ ***
0-p|	-4.901914	86	$4.363*10^{-5}$ ***
0-p||	-8.602549	321	$4.83*10^{-8}$ ***
1-p	4.346225	389	$9.96*10^{-5}$ ***
1-p|	5.110637	69	$3.306*10^{-5}$ ***
1-p||	2.84664	327	$4.94*10^{-3}$ **
2-p	7.514402	419	$2.109*10^{-7}$ ***
2-p|	2.385344	73	$1.858*10^{-2}$ *
2-p||	7.191513	344	$3.693*10^{-7}$ ***
3-p	2.85004	396	$4.852*10^{-3}$ **
3-p|	2.897209	70	$5.208*10^{-3}$ **

3-p‖	1.9149	324	$5.319*10^{-2*}$
4-p	-2.643858	486	$8.373*10^{-3**}$
4-p|	1.162624	76	0.247
4-p‖	-3.143532	408	$2.198*10^{-3**}$
P 末音节与 P 中音节平均时长	0	704	0.3046
P 末音节与 P|中音节平均时长	$-4.6457*10^{-2}$	126	0.4685
P|末音节与 P‖中音节平均时长	0	576	0.3046

4.讨论

(1) 由表 1 可以知道，若不计声调，major phrase 最末一个音节的平均时长比 minor phrase 最末音节长，这一点与英语的结论相同。若计声调，则情况比较复杂，如 major phrase 最末一个阴平调的音节反而比 minor phrase 最末一个阴平调的音节短。除了轻声和轻读音节以外，去声在短语末最短，这一结论与以前有关汉语时长的研究的结果吻合(Nordenhake and Svantesson 1983)。

(2) 汉语的 final lengthening 比英语复杂,应根据短语最末音节的不同声调加以判别。

(3) 这里我们没有就短语最末音节是否是词重音、短语重音或句重音等情况分别考察。但是四声的时长确实要受不同语境的影响(Nordenhake and Svantesson 1983)，无论是在句末还是句中，焦点位置(focus position)的四个声调都会变长。

(4) 本文不讨论有关韵律短语的声调曲线(F0)问题，主要是由于目前笔者对它的研究还没有完全进行，但应该说，这是韵律短语最重要的特性，是合成语音是否自然的关键所在。在瑞典语的合成中(Bruce etc. 1997)，minor 和 major 短语的边界有不同的语调特性实现，minor phrase 边界后的短语的 register 无须 reset，major phrase 边界后的短语的 register 须 reset。Downstepping 出现在短语的焦点重音以后，minor phrase 的边界不会打断这种 downstepping，除非 major phrase 边界出现。我们这里的 major phrase 和 minor phrase 的 F0 的特性是否与英语一致，有待进一步研究。

有的学者认为汉语短语在 plain text 中有 declination 和 downsteping (Shih 1997)，也有的学者认为汉语没有(Xu 1997) 。汉语的语调比较复杂，受语用、语境、语义和认知等诸多因素的影响，不考虑这些因素而孤立地谈论语调，是毫无意义的。

(5) 对另外一位发音人的韵律结果的标注正在进行之中。初步结果表明，当发音人以中性语速朗读的时候，语句的韵律结构与 M01 惊人的相似。

参考文献

陈肖霞 (1998) 基于连续话语语料库的语音音段的初步统计分析，本刊。

冯胜利 (1997)《汉语的韵律、词法与句法》，北京：北京大学出版社。

李爱军 (1997) 普通话新闻广播话语中的停顿，中国声学学会 1997 年青年学术会议论文集，哈尔滨工程大学出版社。

沈炯 (1994) 北京话连读的调型组合和节奏形式，《中国语文》第 4 期。

祖漪清，李爱军 (1997) 语音识别和语音合成语料库的设计，《智能计算机接口与应用进展》，北京：电子工业出版社。

Bruce,G., Granstrom, B and House, D.(1992) Prosodic phrasing in Swedish speech synthesis,

Talking Machines: Theories, Models and Designs, edited by G.bailly, C.Benoit and T.R. Sawallis, published by Elsevier Science.

Hirschberg, J. (1991) Using Text Analysis to Predict Intonational Boundaries, In Proceedings of the 2nd European Conference on Speech Communication and Technology, Vol. 3: 1275-1278.

Horne,M. & Filipsson,M. (1995) Developing the prosodic component for Swedish speech synthesis,
In Proceedings of the 1st European Conference on Speech Communication and Technology, Vol.1: 611-615.

Horne,M. Filipsson,M. Ljungqvist,M. and Lindstrom,A. (1994) Computational modelling of contextual coreference: implication for Swedish text-to-speech, IBM working paper of the Institute for Logic & Linguistics, working paper 6: 103-112.

Horne,M. & Filipsson,M. (1996) Computational Extraction of Lexico-Grammatical Information for Generation of Swedish Intonation, Progress in speech synthesis, edited by Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, Julia Hirschberg, Springer.

Lindstrom, A., Horne, M. and Svensson, T. etc. (1995) Generating prosodic structure for restricted and "unrestricted" texts, ICPhS 95 Vol. 2: 330-333.

Nordenhake, Magnusand and Svantesson, Jan-Olof (1983) Duration of Standard Chinese Word Tones in Different Sentence Environments, Working Papers 25, Linguistics-Phonetics, Lund University, pp.105-111.

Ostendorf,M. Veilleux,N.(1994) A hierarchical Stochastic Model for Automatic Prediction of Prosodic Boundary Location, Computional Linguistics, Vol 20 (1): 27-53

Sanders,Eric and taylor,Paul (1995) Using Statistical Models to Predict Phrase Boundaries for Speech Synthesis, In Proceedings of the 4th European Conference on Speech Communication and Technology, pp.1811-1814.

Selkirk, E. (1984) Phonology and syntax: the relation between sound and structure. Cambridge, MA: MIT press.

Selkirk, E (1990) On the nature of prosodic constituency: comment on Beckman and Edwards' paper, Papers in laboratory phonology I: between the grammar and physics of speech, edited by
John Kingston and Mary Beckman. Cambridge: Cambridge University Press.

Selkirk, E. & Shen Tong (1988) Prosodic domains in Shanghai Chinese, The phonology - syntax connection, edited by Sharon Inkelas & Draga Zec. Chicago: The University of Chicago Press.

Shen,X-N.S. (1992) A pilot study on the relation between the temporal and syntactic structure in Mandarin, Journal of the International Phonetic Association 22.1/2: 35-43.

Shih, Chi-lin (1986) The Prosodic Domain of Tone Sandhi. University of California Ph.D. Thesis.

Sproat, Richard and Shih, Chilin, Gale, William and Nancy Chang (1996) A stochastic finite-state word-segmentation algorithm for Chinese. Computational Linguistics, Vol. 22(1): 27-53.

Xu Yi (1997) What can tone studies tell us about intonation? Intonation: Theory, Models and Applications. pp.337-340.

Zu, Yiqing (1997) Sentence Design for speech synthesis and speech recognition database by phonetic rules, In Proceedings of the 5th European Conference on Speech Communication and Technology, Vol. 2: 743-746.

Durational Characteristics of the Prosodic Phrase in Standard Chinese (SC)

Li Aijun

The prosodic phrase (PP) is divided into two levels by perception: minor phrase (MIP) and major phrase (MAP). The prosodic structures of 145 sentences were labeled and the duration of the last syllable in PP, the syllabic number of each PP and the duration of the PP were also measured. The statistical results on duration and syllable number of MIP and MAP are presented in this paper:

(1) The duration of the final syllable in PP, MIP and MAP with different tones. The duration of the neutral tone syllable is the shortest. If T1 T2 T3 and T4 stand for the duration of the four tones, then T4<T3<T1<T2. The F-test and T-test are shown in Table.1.

(2) The relation between the duration of the final syllable and the average syllable duration in PP, MIP and MAP are calculated. Final lengthening exists in MIP and MAP except the neutral tone syllable and tone 4 syllable in MAP.

(3) T-Test of the duration of the final syllable with the same tone in different phrases.

(4) The average syllabic duration of MAP and MIP is 0.212s and 0.218s respectively. No significant difference exists between them.

(5) The average phrasal duration of MAP and MIP is 1.069s and 1.261s respectively. Significant difference exists between them.

(6) The average syllabic numbers of MAP and MIP are 5.141 and 5.924. Significant difference exists between them.

Some control parameters such as the phrasal number and the final lengthening have been obtained by the above statistical analyses that can be used when we parse the prosodic structure in TTS.