# SEGMENTAL DURATIONS OF A LABELED SPEECH DATABASE AND ITS RELATION TO PROSODIC BOUNDARIES

*ZU Yiqing & CHEN Xiaoxia*
Institute of Linguistics
Chinese Academy of Social Sciences
5 Jianguomennei Da Jie, 100732
Beijing,China
Tel: 086-010-656237408, E-mail: Phonlab@public.bta.net.cn

## ABSTRACT

The aim of this study is to obtain good understanding of segmental timing structures in a continuous speech database of standard Chines and find the relationship between segmental variability and prosody. Based on labeled speech data, this study examines the segmental durations which reflects cues of prosodic structures. By using normalized speaking rate, different levels of pause and segmental lengthening are found to distinguish prosodic groups within an utterance such as major phrase(MAP), minor phrase(MIP) and prosodic word(PW) in continuous utterances.

## 1. INTRODUCTION

A national continuous speech database suitable for training and testing Chinese speech recognizers has been built up. It contains 200 speakers each has about 500 utterances. In the procedure of designing reading text of speech database, phonetic knowledge has contributed to create speech units which can describe variabilities in continuous speech such as junctures between syllables [1]. By now the speech units used in collecting sentences are still limited in segments. We use syllables, inter-syllabic diphones, inter-syllabic triphones and final-initial structure as speech units to control the coverage. None of those four sets of speech units involve prosodic information including tones, stress and prosodic structures. We need a more elaborate acoustic model which involves both segmental and prosody. To solve this problem we must understand where is the boundaries of prosodic structures [2],[3] and the relationship between syntactic structures and prosodic structures [4]. Based on speech database, this study is restricted to the segmental duration which reflects one of important cues of prosodic structures.

## 2. OUTLINE OF SPEECH DATABASE

The reading text of the continuous speech database includes the most of following speech units inventories: (1) 401 syllables without tone; (2) 415 inter-syllabic diphones; (3) 3035 inter-syllabic triphones; (4) 781 inter-syllabic final-initial structures. We also give 17 sentence patterns to include the prosodic phenomena.

Among 200 speakers' data, 1560 utterances of 3 speakers has been hand-labeled [5]. To obtain the knowledge of the influence of prosodic structure on segments timing we use the labeled data to find the answer. The labeled utterances in the continuous speech database provide us the mean and standard deviation of duration in following segments [6] :

(1) 21 consonants:
b, c, ch, d, f, g, h, j, k, l, m, n, p, q, r, s, sh, t, x, z, zh;
(2) 17 voiced consonants:
bv, cv, chv, dv, fv, gv, hv, jv, kv, pv, qv, sv, shv, tv, xv, zv, zhv;
(3) 190 vowels (38 rhythms with five tones(1,2,3,4,0), where 0 stands for neutral tone):
a, e, ai, an, ang, ao, ei, en, eng, er, o, ou, ong, I, ia, iao, iu, ian, ie, in, ing, iong, iang, I1, I2, I3, ua, uan, uai, ui, uang, ueng, uo, un, v, ve, van, vn ;
(4) 374 c-v transitions:
Totally 602 are used to describe segmental durations. The data of a middle aged native male speaker M01 is used in this study.

## 3. NORMALIZING SPEAKING RATE

### 3.1 Intrinsic/inherent duration

The issue of duration timing is important not only for speech synthesis and speech recognition but also for phonetic research. There are numerous studies on this area [7],[8],[9]. An appropriate prosody will help speech synthesis become more nature. The prosodic information can also be used in speech recognition. The duration of each segments has to be expressed in term of how longer or shorter it is than expected, i.e. we should de-emphasis the variation caused by the intrinsic/inherent duration of segment [10],[11]. Tab.1 gives some examples of consonants segmental durations of speaker M01. Different articulatory manner of consonant have different durations: stop "b" is about 16 ms and voiceless fricative "f" is about 93 ms. Tab.2 shows that different vowels also have different durations. The vowels with lower vowel (e.g."a") are longer than higher one ("i,u"). Final durations also differ depending on both different tones associated with them and different numbers of segments. For example, the duration of tone 2 always has longest duration and that of tone 3 the second longest ( see Tab.3); In Tab. 4 "uang" is longer than "an", "an" is longer than "a".

**Tab. 1 The duration of some consonants**

| examples of consonants | occurrence number | mean (s) | deviation (s) |
|---|---|---|---|
| b | 240 | 0.016096 | 0.006659 |
| d | 505 | 0.015059 | 0.005259 |
| g | 325 | 0.025979 | 0.008007 |
| f | 202 | 0.092649 | 0.028997 |
| h | 289 | 0.092462 | 0.027856 |
| s | 136 | 0.120236 | 0.034682 |
| sh | 484 | 0.112097 | 0.033164 |

**Tab.2 The duration of vowels "a,i,u" with first tone**

| examples of vowels | occurrence number | mean (s) | deviation (s) |
|---|---|---|---|
| a1 | 122 | 0.119138 | 0.032192 |
| i1 | 96 | 0.116423 | 0.046816 |
| u1 | 52 | 0.103195 | 0.041503 |

**Tab.3 The duration of final "iang" with different tones**

| examples of segment tone | occurrence number | mean (s) | deviation (s) |
|---|---|---|---|
| iang1 | 41 | 0.180214 | 0.041482 |
| iang2 | 26 | 0.193486 | 0.044129 |
| iang3 | 26 | 0.180574 | 0.037586 |
| iang4 | 27 | 0.157241 | 0.032220 |
| iang0 | 1 | 0.123625 | 0.000000 |

**Tab.4 The durations of some vowel with different number of phone**

| examples of segments | number of occurrence | mean (s) | deviation (s) |
|---|---|---|---|
| a1 | 122 | 0.119138 | 0.032192 |
| an1 | 55 | 0.153333 | 0.048570 |
| uang1 | 18 | 0.161851 | 0.019325 |

### 3.2 Speaking rate

For the reason mentioned above we use normalized segment durations, instead of absolute durations [11],[12]:

$$\tau = (d-\mu)/\sigma \qquad (1)$$

where $\mu$ is mean and $\sigma$ is standard deviation. The speaking rate r of an utterance is defined as:

$$r = ( \sum\tau_i ) / N \quad i=1,2,..,N \qquad (2)$$

where N is the number of segments in that utterance.

The speaking rate of 520 utterances by speaker M01 shows a normal-like distribution. The center is 0. Fig.1 is the histogram of speaking rate which is within the range of what is typical for reading speech. The normalizing speaking rate provides a set of m01's inherent duration as standard to determine which segment is lengthened in every utterances read by speaker M01.

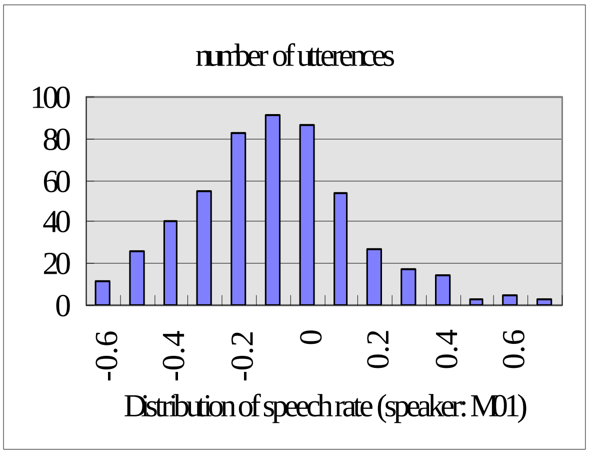

**Fig.1 The distribution of speaker M01's speech rate**

## 4. PAUSE, SEGMENTAL LENGTHENING AND ITS RELATION TO PROSODIC HIERARCHY

Native speakers tend to group words into phrase when speaking. A speaker use suprasegmental cues, such as pause, segmental lengthening and pitch, to implement in structures of utterances. Prosodic structures will help to get the meaning of an utterance and the speaker's intention to the listener. There are one more such groups or prosodic structures in an utterance which may have some relationship with syntactic and semantic constraints. The hierarchical prosodic structure [13] in continuous speech is assumed as follows (from large to small): intonational phrase, phonological phrase, prosodic word and foot. There are many evidences to demonstrate that pause and segmental lengthening are important acoustic cues of prosodic boundaries. The boundaries of the prosodic structure are breaks whose realization can be a silent pause, pre-boundary lengthening/final lengthening and pitch movement / F0 reset. Such a fact shows that pausing and segmental lengthening play a important role in prosodic structure. In general, pausing takes place at the major prosodic boundaries and lengthening may relate to miner prosodic boundaries. It can be inferred that perceived breaks are not necessarily signaled by silence, but probably by other acoustic cues, such as final lengthening. In some references silent pause and filled pause are introduced to refer silence gap and segmental lengthening.

Labeled speech data will provide all information of pause and segmental lengthening.

## 5. METHOD

### 5.1 syllable lengthening levels

This study concentrates the boundary cues on segmental timing structure. To determine segmental lengthening the inherent duration must be used. The normalized duration can not only deduce the difference of inherent duration, but also the different deviation range of durations. The speech rate should also be concerned. So we measure the duration by :

$$dur = \tau - r \qquad (3)$$

Where $\tau$ and r are defined in (1) and (2) respectively. In (1) $\mu$ (mean) and $\sigma$ (standard deviation) are calculated from labeled data. Based on the data we can easily find silence duration which is related to pause and segmental lengthening.

As mentioned above, 602 segments including initials, finals and c-v transitions are used in this study. Formula (2) is used to determine syllable lengthening. In this case $\tau 1$, $\tau 2$ and $\tau 3$ stand for consonant, transition and final respectively, N equals 3. In the case of zero initial, N equals 1. The syllable lengthening is contributed by sum of all segments within that syllable. Using formula (3) is to avoid the deviation raised by different speaking styles, where r is the speaking rate of the same utterance with lengthening segments.

The degrees of lengthening are classified into 4 levels at equal time intervals within 0 -- maximum.

By those degrees all silence gaps and segmental lengthening are automatically extracted from the labeling data on speaker M01.

### 5.2 The relationship between lengthening levels and prosodic boundaries

Finally we contrast our results to perception experiment done by Li Aijun [14] who selected 145 utterances in the data of speaker M01. In Li Aijun's experiment, prosodic boundaries are divided into three levels by perception: major phrase (MAP), minor phrase (MIP) and prosodic word (PW).

To simplify problem, at first we concern the syllable lengthening, which is the sum of lengthening of initial, transition and final. Two examples are shown as following. Each contains two lines. The first line is the text described by Pin Yin. The perception boundaries are marked on it. Where “ | ” , “ || ” and “ ||| ” are PW, MIP and MAP respectively. “%” represents the end of utterance. The second line shows the levels of syllable lengthening which is divided into four levels. Sign “gap” means silence duration.

(1) dang1nian2 | de0hai2tong2 ||| ru2jin1 |

0 0 0 1 1 gap 0 1

yi3cheng2zhang3wei2 | yi1ming2 |||

2 0 0 2 2 1

chu1se4de0 | jing3guan1 %

1 0 1 1 2

(2) shuang1fang1 | da2cheng2 || hu4she4 |

2 0 0 2 2 0

dai4biao3 | ji1gou4 | xie4yi4 %

0 0 0 1 1 4

## 6. RESULTS

157 utterances of speaker m01 are used to discuss and give some wonderful result for prosodic boundaries. The perception set and syllable duration set shown by two lines are matched and the result is shown in Tab.5 and Tab.6. Tab.5 gives the numbers of all lengthening levels and the numbers occurring of lengthening levels at perceptive boundaries. It illustrates that about 90% lengthening occur at MIP and PW, 100% silence gap correspond to MAP. The segmental duration of boundary cues (final lengthening, initial lengthening and pause) are shown in tab.6. The values in Tab. 6 are percentage of occurrence.

Tab. 6 shows the high agreement between lengthening levels and prosodic boundaries by perception.

**Tab.5 The no. of occurrence of all lengthening levels**

| | sum of occurrence | no. of currencies in perceptive boundaries | ratio |
|---|---|---|---|
| level1 | 569 | 512 | 90% |
| level2 | 227 | 208 | 92% |
| level3 | 55 | 48 | 87% |
| level4 | 69 | 62 | 90% |
| level5 | 47 | 47 | 100% |
| level6 | 41 | 41 | 100% |
| level7 | 14 | 14 | 100% |
| level8 | 10 | 10 | 100% |
| level9 | 8 | 8 | 100% |

**Tab. 6 The occurrence of different boundary levels and their relation to prosodic boundaries**

| (%) | S1 | S2 | Pw1 | Pw2 | MIP1 | MIP2 | MAP1 | MAP2 |
|---|---|---|---|---|---|---|---|---|
| level1 | 12 | 7 | 22 | 29 | 4 | 4 | 1 | 8 |
| level2 | 13 | 12 | 14 | 27 | 5 | 4 | 1 | 11 |
| level3 | 14 | 10 | 20 | 30 | 1 | 3 | 0 | 5 |
| level4 | 20 | 11 | 18 | 30 | 1 | 5 | 0 | 1 |
| level5 | 0 | 0 | 0 | 1 | 2 | 0 | 95 | 2 |
| level6 | 0 | 0 | 2 | 0 | 4 | 0 | 92 | 2 |
| level7 | 0 | 0 | 0 | 0 | 14 | 0 | 85 | 0 |
| level8 | 0 | 0 | 0 | 0 | 10 | 0 | 90 | 0 |
| level9 | 0 | 0 | 0 | 0 | 0 | 0 | 100 | 0 |

| | |
|---|---|
| S1: | first syllable of an utterance; |
| S2: | last syllable of an utterance; |
| PW1: | last syllable in PW; |
| PW2: | first syllable in PW; |
| MIP1: | last syllable in MIP final |
| MIP2: | first syllable in MIP initial |
| MAP1: | last syllable in MAP final |
| MAP2: | first syllable in MAP initial |

level1--level4: syllable lengthening levels;

level5: the syllable with zero-lengthening and followed with silence gap;

level6-9: syllable lengthening level1-4 and followed with silence gap.

## 7. CONCLUSIONS

Tab.6 shows that the levels of pause and segmental lengthening are very consistent with prosodic boundaries within an utterance. we can get following conclusion:

(1) Initial lengthening

Segmental lengthening occurs not only at final syllables in a prosodic unit, but also at the initial syllable. In other word, initial syllables in PW, MIP and MAP are always lengthened.

(2) MAP boundary and MIP boundary

PW boundaries are distinguished by segmental lengthening; MAP boundaries are distinguished by silence duration or silence accompanied by final lengthening; either pre-boundary lengthening or silence duration will occur at MIP boundaries. This fact indicates that silent pause is the cue of MAP boundaries, while the duration information is not unique cue for but MIP and PW. There is no significant difference in different levels of segmental lengthening .

(3) First and last syllables in an utterance

An utterance is a larger prosodic unit than MIP and MAP. So just like in MIP and MAP, initial and final syllables in an utterance are also marked by segmental lengthening.

(4) The definition of prosodic word a complex problem. In Tab. 6, different levels are not found to have distinct roles in the level of PW. Therefore the result of columns 3 and 4 which is related with PW in Tab.6 is not desirable.

The speech materials used by this study is isolated sentences which does not contain more complex phenomena of prosody. For example, there are fewer silence gaps in an utterance. The another limitation in this study is use of duration information as the boundary cue. Pitch movement is also a important boundary cue and emphasis will play a role in segmental timing.

## ACKNOWLEDGMENT

We are grateful for Ms. Li Aijun who provided us the result of perception experiment on prosodic boundary.

## REFERENCES

[1] Zu Yiqing (1997) “Sentence Design for speech synthesis and speech recognition”, Proceedings of 5th European Conference of Speech Communication and Technology, vol2., pp.743-746.

[2] Fant, Gunnar & Anita Kruckenberg (1989) , “Preliminaries to the Study of Swedish Prose Reading and Reading Style”, STL-QPSR 2.

[3] Blaauw, Eleonora (1994) “The Contribution of Prosodic Boundary Marks to the Perceptual Difference between Read and Spontaneous Speech”, Speech Communication 14, pp.359-375.

[4] Rossi, Mario (1997) “Is Syntactic Structure Prosodically Retrievable”, Proceedings of 5th European Conference of Speech Communication and Technology, vol.1, pp. KN 1-8.

[5] Chen, Xiaoxia (1997) “Segmenting and labeling on ontinuous speech database”, in printing.

[6] Chen, Xiaoxia and Zu, Yiqing (1998) “The factors affecting consonant durations in continuous speech”, this conference.

[7] Crystal, Thomas H. and Arthur S. House (1982), “Segmental durations in connected speech signals: Preliminary results”, J.Acoust.Soc.Am., 72(3), pp.705-716.

[8] Crystal, Thomas H. and Arthur S. House (1988), “Segmental durations in connected speech signals: Current results”, J.Acoust.Soc.Am., 83(4), pp.1553-1573

[9] Santen, J.P.H. van (1992), “Contextual effects on vowelduration”, Speech Communication 11, pp.513-546.

[10] Klatt, D. (1975),”Vowel Lengthening is Sytectically Determined in a Connected Discourse”, J. Phon.3, 129-140.

[11] Wightman, Colin W.& Stefanie Shattuck_Hufnagel(1992), “Segmental durations in the vicinity of prosodic phrase boundaries”, J.Acoust.Soc.Am., vol.91, No.3, pp.1707-1717.

[12] Wang Xue (1997), Incorporating Knowlege on Segmental Duration in HMM-based Continuous Speech Recognition, Foris Publications, The Nethlands.

[13] Selkirk, E (1990), Phonolog and syntaxt: the relation between sound and structure, Cambridge, MA: MIT Press.

[14] Li, Aijun (1998), “ Durational Characteristics of the Prosodic Phrase in Standard Chinese”, Report of Phonetic Research, pp.84-93.