

# Study and Auto-Detection of Stress Based on Tonal Pitch Range in Mandarin

*Xipeng Shen, Bo Xu*

National Laboratory of Pattern Recognition, Institute of Automation  
Chinese Academy of Sciences, Beijing, P.R. China  
{xpshen, xubo}@nlpr.ia.ac.cn

## Abstract

In Mandarin, there is a special acoustic feature—tonal pitch range, which is relative to stress. In this paper, we present a novel concept—tonal range ratio (TRR), which is based on tonal pitch range, and make a study on the correlation between TRR and stress in Mandarin. And we developed a system to automatically detect stresses in words and sentences based on TRR in Mandarin. We obtained high success rate (92.67% in words and 82.0% in sentences). The results show that TRR has strong correlation with stress and is powerful in detecting stresses in Mandarin.

## 1. Introduction

Prosodic stress is important for people to communicate with each other as to make clear focal points of topics and emphasize one's intention. Speakers typically convey sentence focus by an increase in the fundamental pitch, power and duration. The study on stress plays an important role in text-to-speech (TTS), computer-assisted language learning (CALL), and studies on emotional speech. Recently, more and more efforts have been made on studying stress and auto-detection of stress. In [1], the rhythm instruction was realized by judging the stress with duration and vowel quality. Pitch and power also become key factors of sentence stress in [2]. In [3], an HMM-based method, a DTW-based method, and a human strategy only with visual inspection are compared in terms of their performance in judging whether two utterances of a word have the same stress pattern, e.g. r'ecord and rec'ord.

As a tonal language, Chinese has its particular features on stress. Tonal range of pitch is one of these features. The widening of pitch range of tones, especially the shifting up of the top-line of the pitch range of tones, correlates to the stressed word [4] greatly. In [5], the authors show that high point of the pitch range is the main acoustical correlative with word stress perception among low point, pitch range and average pitch. In the experiment of [5], the authors are chiefly based on analysis of one type of tones in Mandarin—the falling tone. In our work, we proposed a novel term—TRR (tonal range ratio), which is normalized for each syllable with any tone and makes it possible to make a reasonable comparison of the pitch ranges of syllables with

different tones. We made some experiments to study the correlation between the TRR and stress. Then, according to results of these experiments, we constructed a system to auto-detect stresses in words and sentences. We obtained high success rate (92.67% in words and 82.0% in sentences). From the results of these primary experiments, we concluded that TRR is a good feature for stress detection. It has strong correlation with stress and not only can eliminate the effect of tones on pitch range of syllables, but also can combine speeches of different speakers together for model training.

In the following parts, Section 2 describes the meaning of tonal range parameter and TRR. Meanwhile, the experiments on the correlation between TRR and stresses are showed in this section. In Section 3, we present experiments and results of stress auto-detection. Finally, we discuss the experimental results and draw some conclusions in section 4. Additionally, those aspects needing to be improved in our works are presented in this section.

## 2. Correlation between TRR and stress

### 2.1 Tonal pitch range

In Chinese, there are five types of tones—tone1, tone2, ..., tone5. Each character is pronounced as a monosyllable with a tone association. The tonal pitch range (TPR) is the difference between the highest point and the lowest point of the fundamental pitch curve (F0 curve) of a syllable. Each tone has its own type of F0 curve. The pitch range of each syllable's pronunciation is relative with its tone. Thus, we call the pitch range of one syllable's pronunciation TPR.

### 2.2 Tone-normalized pitch range and TRR

TPR is relative to the tone and quality of the syllable. We present tone-normalized pitch range (TNPR) to eliminate the effect of tones. And then, based on TNPR, we present tonal range ratio (TRR) to eliminate the effect of quality of syllables. Thus, TRR can eliminate the effect on TPR by both tones and quality of syllables, so that syllables with different tones can be reasonably compared with each other. The TRR of one pronunciation of a syllable is defined as following:

$$Tr_i = \gamma_i / \bar{\gamma}_{si} \quad (1)$$

Where,  $i$  denotes the syllable,  $Tr_i$  denotes the TRR of a pronunciation of syllable  $i$ ,  $\bar{\gamma}_{si}$  denotes the standard TNPR of syllable  $i$ ,  $\gamma_i$  denotes the TNPR of a pronunciation of syllable  $i$ .  $\bar{\gamma}_{si}$  is obtained according to the following formula:

$$\begin{aligned} \bar{\gamma}_{si} &= \frac{1}{M} \sum_{j=1}^M \bar{\gamma}_{sij} \\ &= \frac{1}{M} \sum_{j=1}^M \frac{\bar{r}_{sij}}{\bar{r}_{tj}} \\ &= \frac{1}{M} \sum_{j=1}^M \left( \frac{1}{N_j} \sum_{k=1}^{N_j} \frac{r_{ijk}}{\bar{r}_{tj}} \right) \\ &= \frac{1}{M} \sum_{j=1}^M \sum_{k=1}^{N_j} \frac{r_{ijk}}{N_j \bar{r}_{tj}} \end{aligned} \quad (2)$$

where,  $\bar{\gamma}_{sij}$  denotes the standard TNPR of syllable  $i$  in the speech of speaker  $j$ ,  $M$  denotes the total number of speakers,  $\bar{r}_{sij}$  denotes the average TPR of the syllables with the same pronunciation as syllable  $i$  in the speech of speaker  $j$ ,  $t$  denotes the tone of syllable  $i$ ,  $\bar{r}_{tj}$  denotes the average TPR of those pronunciations of syllables with tone  $t$  in the speech of speaker  $j$ ,  $r_{ijk}$  denotes the TPR of a pronunciation of syllable  $i$  in the speech of speaker  $j$ ,  $N_j$  is the total number of the pronunciations of syllable  $i$  in the speech of speaker  $j$ .

While,  $\gamma_i$  is obtained by the following formula:

$$\gamma_i = \frac{r_i \cdot \bar{r}_{as}^m}{\bar{r}_t^m \cdot \bar{r}_{at}} \quad (3)$$

Where,  $r_i$  denotes the TPR of one pronunciation of syllable  $i$  in the test corpus,  $t$  denotes the tone of syllable  $i$ ,  $\bar{r}_t^m$  denotes the average TPR of those syllables with tone  $t$  in the speech of speaker  $m$ ,  $m$  is selected from 1 to  $M$  randomly,  $\bar{r}_{as}^m$  denotes the average TPR of all the syllables in the speech of

speaker  $m$ ,  $\bar{r}_{at}$  denotes the average TPR of all the syllables in the test corpus. Thus,  $Tr_i$  can be obtained by the following formula:

$$\begin{aligned} Tr_i &= \gamma_i / \bar{\gamma}_{si} \\ &= \frac{r_i \cdot \bar{r}_{as}^m}{\bar{r}_t^m \cdot \bar{r}_{at}} \bigg/ \frac{1}{M} \sum_{j=1}^M \frac{\bar{r}_{sij}}{\bar{r}_{tj}} \\ &= \frac{r_i \cdot \bar{r}_{as}^m}{\bar{r}_t^m \cdot \bar{r}_{at} \cdot \bar{\gamma}_{si}} \end{aligned} \quad (4)$$

From this formula, we can see that the TRR not only can eliminate the effect of tones on TPR of syllables with different tones, but also can eliminate the effect of quality of syllables. Additionally, TRR can combine speeches of different speakers together so that all speeches can be used to generate the standard parameters. It made it possible to reasonably compare the pitch ranges of any two syllables with different tones. Thus, if there's a strong correlation between TRR and stress, we can compare the degree of stresses of syllables in a sentence by comparing the TRR of these syllables. In the next part, we present the experiment and results of testing this correlation.

## 2.3 Experiment I and results

In this part, we present experiment I to test the correlation between TRR and stress. We didn't consider the syllables of tone-1 and tone-5 because there is not direct correlation between TPR and stresses on those syllables.

### 2.3.1 Corpus

In the experiment, the training corpus includes 20 speeches by 20 male speakers. The text corpora for these speakers consist of 3 collections, each of which includes 520 sentences. One speaker only read sentences in one collection. Totally, there are 1560 sentences. Except for the syllables of tone-1 and tone-5, there are 12558 characters in the text corpora. The test set includes 80 sentences, 936 characters. Except for the syllables of tone-1 and tone-5, there are 699 characters in the test set. It is of a male's speech. We manually labeled each of these characters as unstressed or stressed syllables. Three listeners labeled the corpus. For a syllable, if 2 or 3 listeners label it as stressed, the syllable is considered as stressed. Unstressed syllables are in the same case. Table 1 is the final labels of the syllables in this test set.

Table 1. Labels of the test corpus in experiment I

	Unstressed	Stressed	Total
Tone-2	212	35	247
Tone-3	191	32	223
Tone-4	184	45	229
Total	587	112	699

### 2.3.2 Experimental results

In our experiment, we first calculated the data of standard syllables, which include  $\bar{r}_{as}^m$ ,  $\bar{r}_t^m$  ( $t = 2, 3, 4$ ) and  $\bar{y}_{sj}$  ( $i = 1, 2, \dots, V$ ).  $V$  is the total number of syllables in training corpus. Then, we calculate the average TPR of all the test syllables  $\bar{r}_{at}$  and TRR of each syllable in test set.

The TRRs of the test syllables are showed in figure 1. It is showed that the TRRs of the stressed syllables are significantly higher than most of the unstressed syllables.

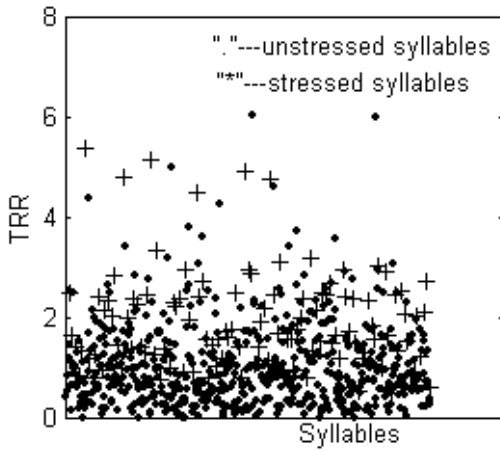


Figure 1. The TRRs of syllables in test corpus

The average TRRs of the test syllables are showed in table 2. The average TRRs of stressed syllables of each tone are much higher than that of unstressed ones. And the average TRR of all of the stressed syllables is about 1.78 times of that of unstressed syllables.

Table 2. Average TRRs of syllables in the test corpus

	Tone-2	Tone-3	Tone-4	Total
Unstressed	1.168	1.081	1.013	1.091
Stressed	1.986	1.903	2.070	1.937

We sorted the TRRs of syllables in each sentence from maximum to minimum. Then, for each stressed syllable, we divided its sequence number by the total number of syllables in the sentence to obtain a value called stressed syllable's TRR rank. We show the stressed syllables' TRR rank values in figure 2. From the figure, we can see that most of the values are 0-0.2, which means the TRRs of almost all of the

stressed syllables are among the biggest ones in a sentence.

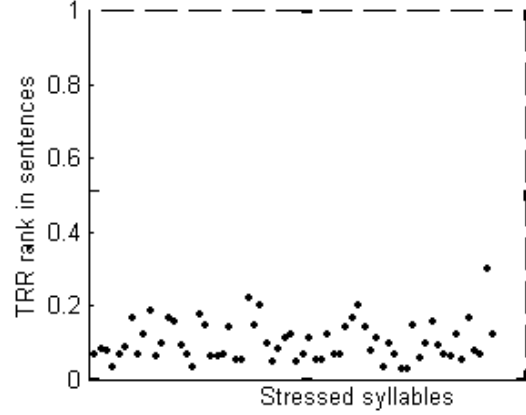


Figure 2. Stressed syllables' TRR ranks in test corpus

Thus, we can see that there are strong correlation between TRR and stress in Mandarin. This correlation may be used in stress auto-detection, just as the experiments showed in the next section.

### 3. Stress auto-detection

In this section, we present two experiments of stress auto-detection based on TRR. Experiment II is to detect the prosodic stresses within some words, which is composed of 2 or 3 characters. Experiment III is to auto-label the prosodic stress for syllables in some sentences.

#### 3.1 Experiment II--detect stresses within words

The test set in this experiment includes 150 words pronounced naturally by a male speaker. In the ways showed in Section 2.3, each word is labeled a stress mark on one syllable. All the syllables in the test set are of tone-2, tone-3 or tone-4. This test corpus is showed in table 3, in which, Fs denotes the stressed syllable is the first syllable, Ms denotes the stressed syllable is the middle syllable, Ls denotes the stressed syllable is the last syllable, 2-ch denotes the word is composed of 2 characters, 3-ch denotes the word is composed of 3 characters.

Table 3. Test corpus in experiment II

	2-ch	3-ch	Total
Fs	95	16	111
Ms		6	6
Ls	25	8	33
Total	120	30	150

The auto-detect system includes 2 chief modules: the first one is to calculate the TRRs of each syllable of

the test set, the other one is to judge which syllable in a word should be labeled as stressed one by comparing the TRRs of all the syllables in the word. The criterion is simple: the one with the biggest TRR is labeled as a stressed syllable. The confusion matrix is showed in table 4. Overall, the correct rate is 92.67%. From the results, we can see that the auto-detection system based on TRR is effective in detecting the stress within a word.

Table 4. Confusion matrix for stress detection within words

	2-ch Fs	2-ch Ls	3-ch Fs	3-ch Ms	3-ch Ls
2-ch Fs	91	3			
2-ch Ls	4	22			
3-ch Fs			15	2	1
3-ch Ms			0	4	0
3-ch Ls			1	0	7

### 3.2 Experiment III--Auto-label stresses in sentences

In this experiment, the test set includes 80 sentences read by a male and was labeled with stress marks manually. Totally, 115 syllables were labeled as stressed.

In this experiment, we adopted 3 methods to auto-label the stress. One is to label the syllable with highest TRR as stressed and the others unstressed. The second one is to label the two syllables with highest TRRs as stressed and the others unstressed. The last one is to calculate the similarity between the  $M$  syllables with the highest TRRs and the maximum TRR of a sentence. The similarity is calculated as following:

$$Sim_j = 1 - \frac{T_m - T_j}{T_m - \bar{T}} \quad (j = 1, 2, \dots, M - 1) \quad (5)$$

Where,  $\bar{T}$  denotes the average TRR of all the syllables in the sentence;  $T_m$  denotes the maximum

TRR in the sentence;  $T_j$  denotes a TRR value of one of the  $M$  syllables. In our experiment, we set  $M$  the half of the total number of syllables in the sentence.

In the last method, we set a thresh value  $K$ . When  $Sim_j$  is larger than  $K$ , syllable  $j$  are labeled as stressed. In our experiment, we set  $K$  equal to 0.75. Although these three methods are all simple, they are

suitable to test the capability of TRR in detecting stresses in sentences. Table 5 shows the results. The meaning of precision and recall can be seen in [6].

Table 5. Results of detecting stresses in sentences

	Precision	Recall
Method I	0.83	0.58
Method II	0.57	0.79
Method III	0.71	0.82

## 4. Discussion and conclusions

From all of the experiments, we can see that TRR is a good feature to detect stress. It has strong correlation with stress and not only can eliminate the effect of tones and quality of syllables on pitch range of syllables, but also can combine speeches of different speakers together.

Although we obtained fairly good results in stress detection, there is still big space for improvement. One idea is to take use of other features, such as duration and power. The other idea is to take consider of the vowel and consonant around each syllable and take the corresponding bi-phone or tri-phone as the unit to calculate the TRR, which should make the TRR more powerful for stress detection.

In our experiments, we didn't consider the syllables of tone-1 and tone-5 because there is not direct correlation between TRR and stresses on those syllables. However, we are trying to take use of duration and average F0 value to detect stresses on these syllables now.

## 5. References

- [1] S. Hiller, E. Rooney, J. Laver and M. Jack, "An Automated System for Computer-Aided Pronunciation Teaching", *Speech Communication*, 1993.
- [2] M. Sugito, "English spoken by Japanese", published by *Izumi shoin* (1996, in Japanese).
- [3] N. Minematsu, Y. Fujisawa, and S. Nakagawa, "Performance Comparison among HMM, DTW, and Human Abilities in Terms of Identifying Stress Patterns of Word Utterances", *ICSLP2000*.
- [4] Jiong Shen, "The pitch range of tone and intonation of Mandarin", *The Collection of Acoustic Study of Mandarin*, 1985 (Chinese).
- [5] W. Bei, Z. Bo, L. Shinan, C. Jianfen and Y. Yufang, "The Pitch Movement of Word Stress in Chinese", *ICSLP2000*.
- [6] Taylor, P. and Black, A.W. 1998. "Assigning phrase breaks from part-of-speech sequences", *Computer speech and language*, vol. 12, 1998, 99-117.