A Quantitative Parameter of Pronunciation, TVVF

Yin Zhigang

Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
yinzhg@cass.org.cn

Abstract. The process of pronunciation is accompanied by energy planning and consumption. Energy in a broad sense is a 'resource'. Many phonetic phenomena are related to the quantity of the ‘resource’. For example, stress is often considered to occupy a larger amount of ‘resource’, which is why it appears particularly prominent. But what parameters are used to describe the quantity of pronunciation? This is a question worth studying.

This article starts with the issue of stress, studies various acoustic factors that affect stress perception, and then finds an effective parameter that reflects the quantity of pronunciation - Times of vibration of vocal folds (TVVF).

Previous studies in phonetics have suggested that pitch and duration are important acoustic parameters that affect stress perception. However, for tonal languages like Mandarin, these two parameters may not be the best parameters for representing stress. Because different tones have different pitch and duration characteristics, these parameters cannot reflect the stress levels well.

This study found that TVVF better reflects stress levels than pitch and duration. TVVF represents the times of vibration of vocal folds during pronunciation and the number of pulses of a syllable in acoustics. Since the number of pulses is the integral of pitch and duration, TVVF is a fusion parameter of pitch and duration. Research has shown that TVVF is a more effective parameter for indicating tonal language stress level than other acoustic parameters.

In addition, TVVF can effectively reflect the quantity pattern of pronunciation in large prosodic units. We found that the number of syllables in a big prosodic unit is inversely related to the average TVVF of syllables. That is to say, there is a conservation phenomenon in the TVVF of prosodic units. We refer to it as the principle of ‘Quantitative conservation of pronunciation’.

In summary, TVVF is an effective quantitative parameter for describing pronunciation and can be used in various quantity-related studies.

Keywords: Quantitative Parameter, Times of vibration of vocal folds (TVVF), stress level, Quantitative conservation of pronunciation.

1. Introduction

The process of pronunciation requires energy consumption. Energy can be seen as a kind of 'resource'. So the pronunciation process can be seen as a process of allocating and consuming pronunciation ‘resource’.

Many phonetic phenomena are related to the quantity of the ‘resource’. For example, stress is often considered to occupy a larger amount of ‘resource’, which is why it appears particularly prominent. Previous studies have suggested that the stress level of syllables may be related to parameters such as pitch, duration, and intensity [1] [2]. However, for Mandarin, a monosyllabic tonal language, the relationship between stress levels and acoustic factors is more complex.

The first aim of the study is to identify the parameters in Mandarin that have the strongest correlation with stress levels. This parameter can also be seen as a quantitative characteristic and applied to all studies related to the quantity of pronunciation.

2. Some acoustic parameters that affect stress levels

2.1 Pitch

Previous studies have suggested that pitch is the most important acoustic parameter that affects stress perception [3] [4]. Stressed syllables typically increase the maximum pitch value (PitchMax) and broaden the pitch range (PitchRange). As shown in Figure 1, stress enhances the [+high] features of tone and lowers its [+low] features [6][7].

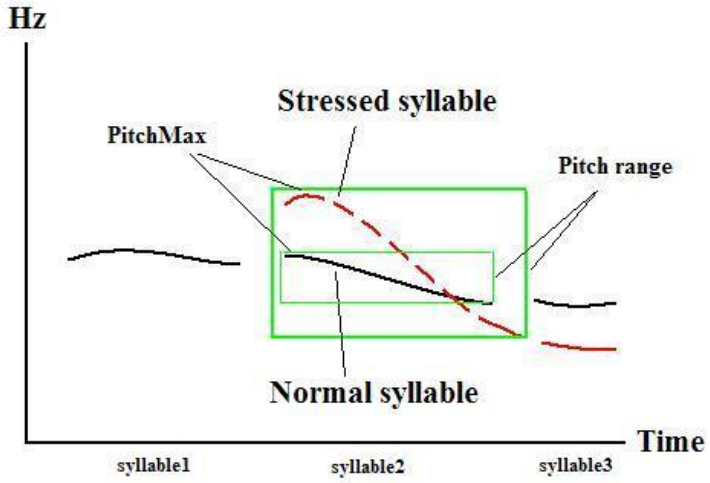

Fig. 1. The pitch contours of a normal syllable and a stressed syllable

For tonal languages, different types of tones have different pitch and duration characteristics. Mandarin is a language with four tones. Its four vocabulary tones have their pitch characteristics and pitch values (5-degree values) - Tone 1 (HH, 55), Tone 2 (LH, 35), Tone 3 (LL, 214), and Tone 4 (HL, 51). Figure 2 shows their pitch patterns. Tone 1 (HH), Tone 2 (LH), and Tone 4 (HL) have high features [+high]. The only tone type that does not have a high feature [+high] is Tone 3 (LL). Therefore,

stressed syllables with Tone 3 will not increase PitchMax, but will expand the Pitch range by decreasing PitchMin.

Due to the complexity of Mandarin tone types, any pitch feature (PitchMax, PitchMin, PitchRange) is difficult to use alone as a parameter to indicate stress levels.

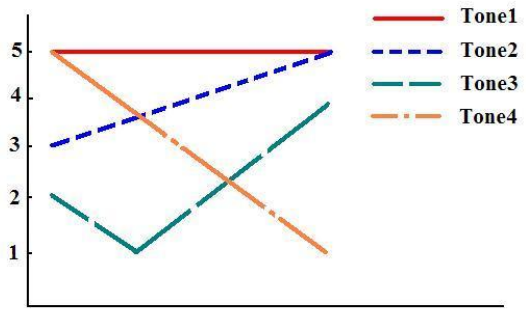

Fig. 2. Diagrammatic sketch of Mandarin tones

- The expanded PitchRange can detect the stress of Tone2/Tone3/Tone4, but it is difficult to detect the stressed syllables of Tone1 (HH).
- The raised PitchMax can reflect the stress of Tone 1/Tone 2/Tone 4, but it is difficult to reflect Tone 3. In addition, the phenomenon of pitch drop that occurs in large rhythmic units can also weaken the effectiveness of pitchMax in detecting stress levels (as shown in Figure 3).

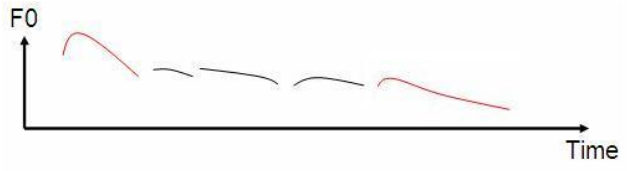

Fig. 3. The pitch curve of a sentence

2.2 Duration

The extended duration is often considered an important acoustic parameter for detecting stress levels. However, as a monosyllabic tonal language, Mandarin has similar syllable durations, and there are systematic minor differences in the duration of different tones. Generally speaking, Tone 3 is the longest, followed by Tone 1 and Tone 2, and Tone 4 is the shortest. In addition, the duration of syllables at the boundaries of rhythm may vary in length [2]. Due to these phenomena, duration is not the best parameter for detecting stress levels.

For the two parameters of pitch and duration, except for a few studies [7], most studies believe that pitch has a greater impact on stress than duration [8] [9] [10] [11].

2.3 Times of vibration of vocal folds (TVVF)

Due to the advantages and disadvantages of pitch and duration in representing stress levels, we believe that integrating them into a new parameter may be a better idea.

If the pitch curve is viewed as a function curve, then integrating pitch and duration can yield a new parameter - the number of pulses, also known as the Times of vibration of vocal folds (TVVF). The relationship among TVVF, pitch, and time (duration) can be represented by formula 1 and Figure 4:

$$TVVF = \int f(pitch)dt \tag{1}$$

In formula 1, *t* means time and its maximum is equal to the duration value of the voiced part of a syllable. *Pitch* means F0 values. The figure below shows the relationship among TVVF, pitch, and duration on syllable-final.

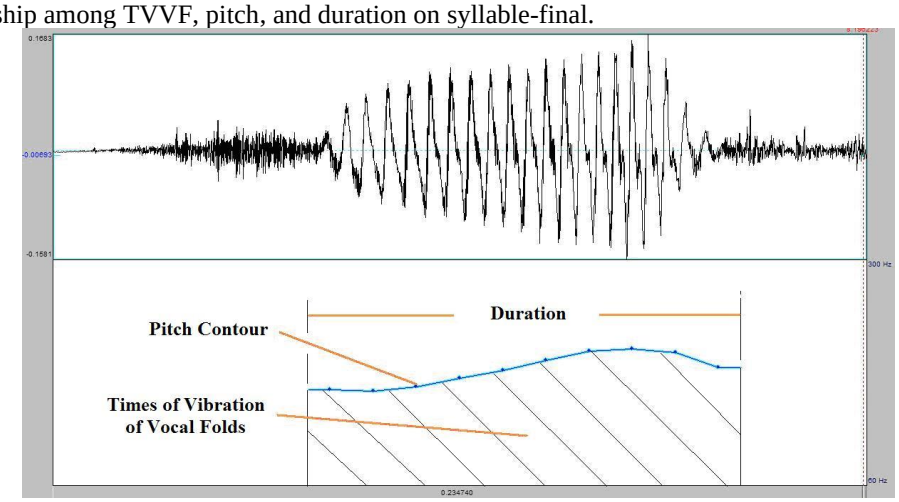

Fig. 4. The relationship among TVVF, pitch, and duration

In Figure 4, the blue curve means the pitch contour, and the area (signified by diagonal lines) under the blue curve (pitch contour) means TVVF. It is related to both pitch and duration: the higher pitch will increase TVVF, and the lengthening duration will also increase it.

TVVF combines the advantages of pitch and duration while reducing their disadvantages when indicating the syllable stress levels of tonal language, theoretically achieving the best results.

Due to the fact that a vibration of vocal folds generates a pulse, the TVVF of a syllable is also equal to its number of pulses in acoustics. Therefore, the value of TVVF can be easily obtained through the function of extracting pulse numbers in PRAAT software, without the need for complex mathematical calculations.

3. Experiment 1: the relationship between TVVF and syllable stress Levels

3.1 Corpus

To verify whether TVVF can better reflect stress levels, we conducted a test using the ASCCD (Annotated Speech Corpus of Chinese Discourse) database.

The corpus consists of 18 text recordings in stereo, 16 kHz, 16 bit format. We selected data from 2 female speakers (F001, F002) and 2 male speakers (M001, M002) among 10 speakers. All records are manually annotated with orthography, syllable initials, syllable finals, prosodic boundaries, and stress levels (0, 1, 2, 3). The level 0 represents weak, level 1 represents normal, level 2 represents strong, and level 3 represents very strong. The consistency rate of the six annotators is approximately 87.5%.

3.2 The correlation between various parameters and stress levels

We analyzed the correlation coefficients between various parameters and stress levels, and the results are shown in Table 1. In the table, the correlation between TVVF and stress level is the highest. Although the correlation coefficient is less than 0.6, considering that all data have not been normalized, this result still proves that TVVF can be a good parameter for reflecting stress levels.

Table 1. The ranks of the acoustic parameters in accordance with the correlation coefficients

Correlation ranks (stress level)	F001	F002	M001	M002	Mean
TVVF	0.50	0.56	0.55	0.58	0.55
Duartion	0.46	0.53	0.54	0.57	0.53
PitchMax (excluding Tone3)	0.43	0.52	0.55	0.56	0.52
PitchMax (all tones)	0.37	0.45	0.48	0.51	0.45
Pitch range (excluding Tone1)	0.12	0.35	0.50	0.55	0.38
Pitch range (ST, all tones)	0.15	0.30	0.30	0.43	0.30
Pitch range (Hz, all tones)	0.09	0.25	0.39	0.41	0.29
Intensity	0.06	0.05	0.05	0.08	0.06
Energy	0.05	0.04	0.02	0.06	0.04

4. Experiment 2: the role of TVVF in large language units

4.1. The law of quantitive conservation of pronunciation in large prosodic units

In addition to reflecting stress levels, TVVF can also be used to reflect the allocation of pronunciation resources in large prosodic units.

Just as physics has the law of conservation of energy, we believe that there is also some form of ‘conservation’ phenomenon in the process of pronunciation. This conjecture is based on the following:

(1) Breathing is the energy foundation of pronunciation. Everyone has relatively stable values for their natural breathing volume and maximum lung capacity.

(2) When people pronounce, breathing energy forms a series of pronunciation movements. The basic pronunciation action at the sound source is glottal vibration. If the vocal energy is a stable value, it means that the vocal actions (such as glottal vibrations) accompanying each breath are also stable.

Based on the above conjecture, we studied the law of ‘quantitative conservation’ in large rhythmic units using TVVF as the acoustic feature.

4.2. Quantitative conservation of pronunciation in intonation phrases

Many studies [12] in prosodic phonology suggest the existence of a prosodic hierarchy in language. This prosodic hierarchy system is composed of a series of prosodic units from small to large. Each major prosodic unit is composed of smaller prosodic units that are one level lower than it. These prosodic units can be divided from small to large into syllable, prosodic word (PW), prosodic phrase (PP), intonation phrase (IP), and utterance.

Among these prosodic units, intonation phrases are relatively opposite large prosodic units. Its typical acoustic feature is the presence of pauses before and after, which are usually accompanied by breathing and ventilation processes.

Taking intonation phrases as an example, we calculated the relationship between the number of syllables and the mean value of TVVF, and found that there is an inverse relationship between them. This means that the more syllables a intonation phrase contains, the less energy is allocated to each syllable. For example, the following figure and table shows the correspondence between the number of syllables within an intonation phrase and the TVVF mean (F001).

Table 2. Relationship between the number of syllables and the TVVF mean in intonation phrases

The number of syllables (N)	The mean value of TVVF
1	45.83
2	39.04
3	36.28
4	33.41
5	32.82
…	…

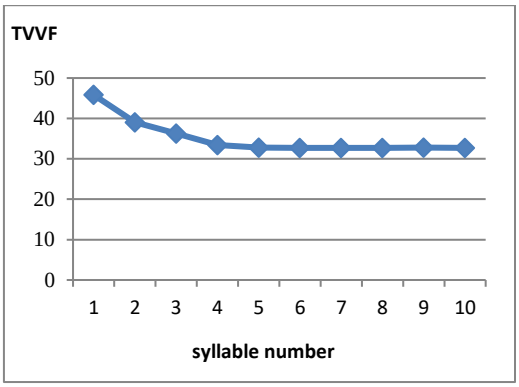

Fig. 5. The relationship between the syllable number and the mean TVVF of syllables in intonation phrases

Through mathematical modeling, we found that the relationship between the number of syllables and the mean value of TVVF can be expressed as:

$$（TVVF-K）* N=T \quad (2)$$

In Formula 2, K is the base (constant) of the TVVF of the speaker, which is the lower limit of the average TVVF of syllables. N is the number of syllables contained in the intonation phrase. T is the slope and also a conservation constant.

We call the above rule the 'Quantitative conservation of pronunciation' rule. The above formula indicates that the product of the number of syllables in an intonation phrase and its average TVVF tends towards a constant T. It should be pointed out that each speaker's K and T are different and need to be observed separately. The deeper linguistic significance behind it also needs to be explored in our subsequent research.

4.3. Quantitative conservation of pronunciation in other prosodic units

In addition to intonation phrases, we investigated whether the above rule also exist in other prosodic units, such as prosodic words and prosodic phrases. The data in Tables 3 and 4 respectively show the relationship between the number of syllables and the average TVVF of syllables within prosodic words and prosodic phrases (F001).

Table 3. Relationship between the number of syllables and the TVVF mean in prosodic words

The number of syllables (N)	The mean value of TVVF
1	36.18
2	33.79
3	30.40
4	30.28
5	28.50

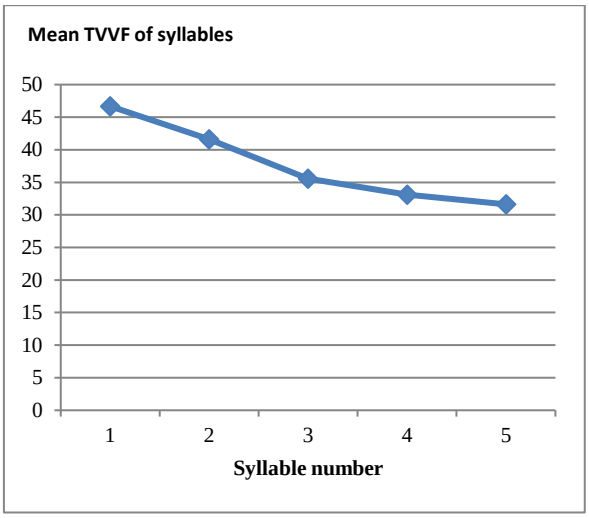

Fig. 6. The relationship between the syllable number and the mean TVVF of syllables in prosodic words

Table 4. Relationship between the number of syllables and the TVVF mean in prosodic phrases

The number of syllables (N)	The mean value of TVVF
1	46.67
2	41.63
3	35.56
4	33.10
5	31.62
…	…

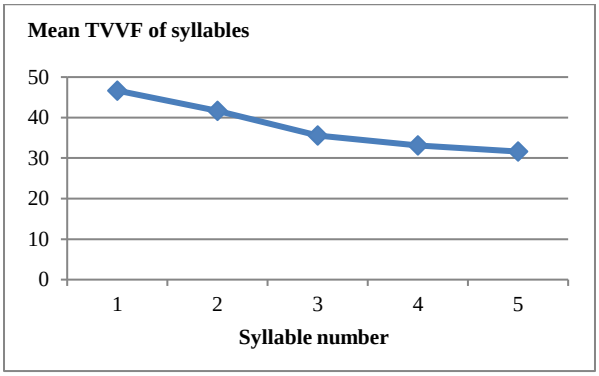

Fig. 7. The relationship between the syllable number and the mean TVVF of syllables in prosodic phrases

The above data shows that there is a negative correlation between the number of syllables and the average TVVF of syllables in both prosodic words and prosodic phrases. This experiment can demonstrate that the rule of 'quantitative conservation of pronunciation' is widely present in various prosodic units, and TVVF is a reliable quantitative indicator of pronunciation.

5. Conclusion

The results of experiment 1 indicate that TVVF is an effective parameter reflecting the quantity of pronunciation and it can effectively predict the stress level of syllables.

The results of Experiment 2 indicate that TVVF can analyze the quantity patterns in large prosodic units. And based on TVVF, we obtained *the law of quantificative conservation of pronunciation*.

It should be pointed out that the study is based on a Mandarin corpus but the same conclusion applies to other languages. In the future, more work will be carried out with other language corpora to check the validity of the conclusion.

6. References

1. Ye, J.: The Research of Rhythm of Modern Chinese, Shanghai Century Press, Shanghai, 40-49 (2008, in Chinese).
2. Yin, Z.: The Research of Rhythm of Standard Chinese, PhD Academic Dissertation of CASS, 61-71 (2011, in Chinese).
3. Fry, D.B.: Experiments in the Perception of Stress, Language and Speech 2(1), (1958).
4. Xu, Y.: Effects of tone and focus on the formation and alignment of f0 contours, Journal of Phonetics 27, 55-105 (1999).
5. Cao, J.: The Relationship of Tone and Intonation, China Language (3), 195-202 (2002, in Chinese).
6. Lin, M., Yan, J., Sun G.: The Stress Pattern of Foot in Beijing Dialect, Dialect (1), (1984, in Chinese).
7. Zhong, X., Wang, B., Yang, Y., Lv, S.: The Stress Perception of Standard Chinese, Psychological Sinica 31, (2001, in Chinese).
8. Wang, Y., Chu M., He L.: A preliminary Study of Focus Stress and Semantic Stress in Chinese, Chinese Courses for Foreigners (2), 86-98 (2006, in Chinese).
9. Cao, W.: The Prosody of Chinese Focus, pp. 5-10, Beijing Language Press, Beijing (2010, in Chinese).
10. Cai, L., Wu Z., Tao J.: The Computability Research of Chinese Prosody Characters, The Modern Phonetics of New Century, Tsinghua University Press, Beijing (2001, in Chinese).
11. Xu, J., Chu, M., He L., Lv S.: The Infection of Utterance Stress to Pitch and Duration, Acoustics Journal 25(4), (2000, in Chinese).
12. Tseng, C.Y., Pin S.H., Lee Y.L.,. Speech prosody: issues, approaches and implications. From Traditional Phonology to Modern Speech Processing, pp.417-437. Foreign Language Teaching and Research Press, Beijing (2004).