# Evaluation of Aspiration Sounds of Chinese Labial Diphthong Uttered by Japanese Students Using VOT and Breathing Power

Akemi Hoshino ※

Abstract

A Chinese aspiration is generally considered to be very difficult to reproduce by Japanese students. We measured the voice on set time (VOT) and mean power during VOT, as the evaluation parameters, of diphthong sounds of pai[p'ai], pao[p'ao], pei[p'ei] and pie[p'iɛ] in Chinese aspirations uttered by nine Japanese students and nine native Chinese speakers. The sounds were evaluated by the listening test of eight native Chinese speakers. Then we proved that the VOT was not the sole measure for evaluating the pronunciations but also the mean power during VOT was the other measure for evaluating them. Thus, we conclude that power is also an important factor in evaluating the quality of pronunciation.

## 1. INTRODUCTION

The length of voice onset time (VOT) in uttering Chinese aspirated sounds, which are difficult for Japanese to pronounce, is an important factor in evaluating the quality of pronunciation. In the previous papers [1] [2][3][4], we measured the lengths of VOT and the powers during VOT for twenty one single-vowel syllables of six kinds of aspirates for nine Japanese students and nine native Chinese speakers. The quality of the students' pronunciation was evaluated using a hearing test judged by eight native Chinese. The results indicated that the correlation of the quality of the students' pronunciation to the power used in uttering a sound was greater than that to the VOT within a certain range of VOT which varied for different syllables. Thus, we concluded that power was also an important factor in evaluating the quality of pronunciation.

In the present paper, the same method with that proposed in the pervious papers was used to measure VOT and relative average power, Pav, during VOT. We examined the dependency of the quality upon these values for the Chinese labial aspiration sounds of diphthongs, pai[p'ai], pao[p'ao], pei[p'ei] and pie[p'iɛ] , as pronounced by nine Japanese students, who had studied Chinese 3 hours per week for one year and showed that the quality of pronunciation depended not only on VOT but also on the power during VOT.

## 2. DIFFERENCES BETWEEN DIPHTHONGS OF ASPIRATED AND UNASPIRATED SOUNDS

Figure 1 shows the spectrograms of labial diphthong syllable of unaspirated bai[pai], left, and the aspirated pai[p'ai], right, uttered by a Chinese speaker. The darker the horizontal bands, the higher the power of the frequency components. The aspirate appears in a brief interval in the right spectrogram, indicated by vertical light stripes, between the stop burst and the onset of vocal fold vibrations followed by a vowel. This time interval is called the voice onset time (VOT)[5].

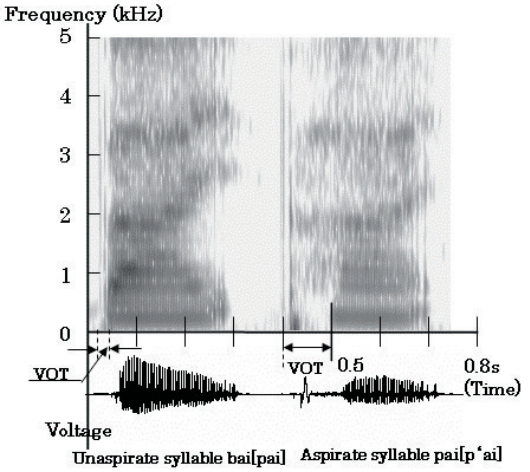


Figure 1 : Spectrograms of unaspirated syllable bai[pai] (left), and aspirated syllable pai[p'ai] (right) pronounced by a Chinese speaker.

※教養学科（准教授）

（平成21年３月31日受付）

The onset of the vocal fold vibration is so close to the burst in the left spectrogram that no aspiration interval appears. The VOT of aspirated syllable pai[p'ai] is very long and 100ms. The VOT of unaspirated syllable bai[pai] is very short and only 5 ms. These data were acquired and analyzed using a tool of Multi-Speech (Model 3700, Kay Electrics Corp, USA).

## 3. METHOD USED TO EVALUATE PRONUNCIATION

We measured the VOT and calculated the relative average power, Pav, during VOT from the spectrograms using the procedure reported previously [1][2][3][4]. The sounds uttered by nine Japanese students were ranked in a hearing test of the reproduced sounds conducted by eight native Chinese speakers. The grades were as follows: 3 = pronunciation which sounds aspirated; 2 = unclear sounds; and 1 = sounds unaspirated. The examiners checked with each other that their pronunciations were perfectly aspirated. Some data were excluded in cases of split evaluations and a standard deviation of larger than 0.64, broken sounds uttered very close to the microphone, and sounds with a low S/N uttered away from the microphone. An average grade of more than 2.6 was defined as a good pronunciation, five of the examiners awarded '3' and three examiners awarded '2'.

Figure 2 shows the spectrograms of the aspirated syllable pai[p'ai] of a Japanese student (left) and a native Chinese speaker (right). The VOT, on the left side for the Japanese student, is only 5 ms and so short compared with that of the native speaker. It is 103 ms on the right side for a native Chinese. This pronunciation received the worst grade of 1.0. The shortest VOTs of the sounds pao[p'ao], pei[p'ei] and pie[p'iɛ] of the Japanese students were 2 ms, 5 ms and 4 ms, respectively. These all pronunciations received the worst grade of 1.0.

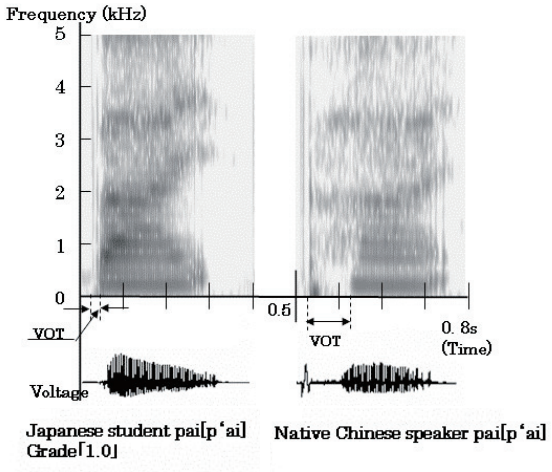


Figure 2 : Spectrograms of aspirate syllable pai[p'ai] pronounced by native Chinese speaker (right) and Japanese student (left).

The longest VOTs of the pronunciations pai[p'ai], pao[p'ao], pei[p'ei] and pie[p'iɛ] of the Japanese students were 45 ms, 70 ms, 65 ms and 92 ms, respectively. These pronunciations received the highest grade of 2.8, 3.0, 3.0 and 3.0, respectively.

These examples agreed well that when an aspirated syllable is pronounced with a brief VOT, it sounds unaspirated, and when it is pronounced with a long VOT, it sounds aspirated[6]. It is thought that VOT is one of the measures deciding the accuracy of the aspirated sounds.

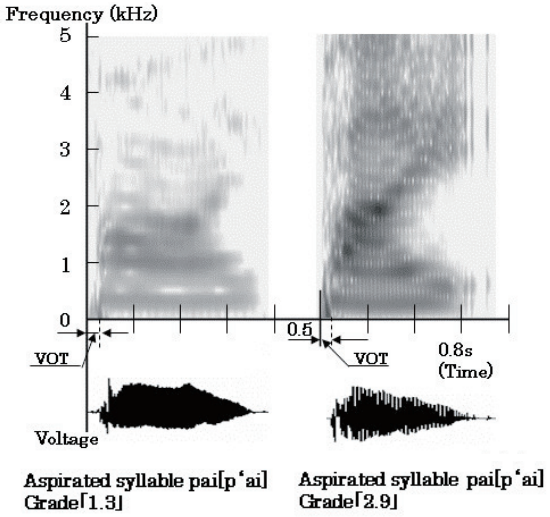


Figure 3 : Spectrograms of aspirated syllable pai[p'ai] pronounced by Japanese.

We can find, however, some exceptions as shown in figure 3 which shows the spectrograms of the aspirated syllable pai[p'ai] pronounced by Japanese students. The VOT on the right side is 25 ms, the color of the vertical stripes in VOT is dark, and the utterance received high grade of 2.9. The VOT on the left side is 28 ms, and it is slightly longer than that of right side. But, the evaluation is low and the grade is 1.3. Vertical stripes look light in the VOT of the pronunciation of the left side.

## 4. DEPENDENCY OF GRADE ON POWER

In the present study, we found some cases in which the student's pronunciation of aspirated sounds received a low grade even when the VOT was almost the same or longer than that of Chinese speakers. To find the reason for this, we use the relative average power, Pav, during the VOT[1][2][3][4], to have proposed before, of the labial sounds of diphthong aspirated syllables in Chinese uttered by nine Japanese students and nine native Chinese and examined the dependency of the pronunciation grade on them.

Table1 summarizes the VOT and the relative average power, Pav, during the VOT for diphthong of aspirated syllables, pai[p'ai], pao[p'ao], pei[p'ei] and pie[p'iε]. The average VOT and the average Pav for the three periods are shown where the VOT is short, in some specific range, and long enough. The average grade is just for the students.

The average VOT of the students' data for the syllable, pai[p'ai], which are shorter than 10 ms, is 4 ms. The average Pav is as low as 1.8. The grade is the worst and 1.0. In case of 20-30 ms, the average VOT is 25 ms, the average Pav is as low as 6. The grade is also low and 2.3. The average Pav of pronunciations with the VOT longer than 35 ms, has the highest value of 310. The average grade is the full mark of 3.0. The power is considered to be a measure of the quality of the pronunciation. VOTs of the pronunciations of all the native Chinese speakers are longer than 35 ms. They are about twice longer than that of the students.

Table 1 : Averaged VOT and Mean value of Relative Average Power (Pav) of aspirated syllable of pronunciation by nine Japanese students and nine Chinese native speakers

| | | Japanese Student | | | | Native Chinese speaker | | |
|---|---|---|---|---|---|---|---|---|
| Silyable | VOT Range | Average VOT | Average Pav | Average Grade | Data# | Average VOT | Average Pav | Data# |
| pai[p'ai] | 0-10ms | 4ms | 1.8 | 1.0 | 1 | — | — | 0 |
| | 20-30ms | 25ms | 6 | 2.3 | 4 | | | |
| | 35ms～ | 38ms | 310 | 3.0 | 4 | 72 | 106.5 | 6 |
| | Average | 28 | 142.3 | 2.5 | | 72 | 106.5 | |
| pao[p'ao] | 0-20ms | 8.3 | 0.81 | 1.1 | 3 | — | — | 0 |
| | 35-40ms | 38 | 130.0 | 2.7 | 3 | 33 | 130.7 | 3 |
| | 41ms～ | 60 | 57.0 | 2.9 | 3 | 91 | 13.2 | 4 |
| | Average | 36 | 62.7 | 2.2 | | 72 | 52.4 | |
| pei[p'ei] | 0-22ms | 16 | 30.5 | 1.5 | 3 | — | — | 0 |
| | 25-29ms | 25 | 4.3 | 2.5 | 2 | — | — | 0 |
| | 30ms～ | 43 | 146.3 | 2.9 | 4 | 65 | 220.7 | 5 |
| | Average | 30 | 76.3 | 2.4 | | 65 | 221.0 | |
| pie[p'iε] | 0-25ms | 18 | 13.3 | 1.3 | 3 | — | — | 0 |
| | 39-40ms | 40 | 37.6 | 2.9 | 2 | — | — | — |
| | 60ms～ | 80 | 60 | 3.0 | 3 | 106.3 | 4.7 | 4 |
| | Average | 47 | 36.8 | 2.4 | | 106 | 4.7 | |

The average VOT of the students' data for the syllable, pao[p'ao], which are shorter than 20 ms, is 8.3 ms, The average Pav is as low as 0.8. The grade is very low and 1.1. In case of 35-40 ms, the VOT is 38 ms, the power is as high as 130. The grade is higher and 2.7. The average VOT and average Pav of students' data are almost the same as those of Chinese speakers in this range. The power is considered to be a measure of the accuracy of the pronunciation. The power of pronunciations with the VOT longer than 41 ms, is higher and 57. They received a high average grade of 2.9. VOT of the pronunciations of Chinese speakers is very long, but average Pav is the lowest. It is understood that the accuracy of the pronunciations did not depend on the power when VOT is enough long.

The average VOT of the students' data for the syllable, pei[p'ei], which are shorter than 22 ms, is 16 ms. Although the average Pav is as high as 31, the grade is so low and only 1.5. In case of 25-29 ms, the power is as low as 4.3. The grade is low and 2.5. The average Pav of pronunciations with the VOT longer than 30 ms, has the highest value of 146 and the grade is the highest and 2.9. The average Pav of the pronunciation of Chinese speakers has also the highest value of 221.

The average VOT of the students' data for the syllable, pie[p'iε], which are shorter than 25 ms, is 18 ms. The average Pav is low and 13. The grade is also low and 1.3. In case of 39-40 ms, the average VOT of the pronunciation is 39 ms and the average Pav is high and 38. The grade is high and 2.9. The average Pav of pronunciations with the VOT longer than 60 ms, has the highest value of 60. The grade is the full mark of 3.0. The average Pav of the pronunciation of Chinese speaker has the highest value of 60. The average VOT for Chinese speakers is as long as 106 ms, the average power is as low as 4.7, It is understood that the accuracy of the pronunciation does not depend on the power in this period, either.

The upper examples show that pronunciations with a low grade have the short VOT. If VOT is long enough, the pronunciation even with a lower power gets a good grade. When the length of VOT is in a certain range, the pronunciation with a higher power gets a better grade and that with a lower power gets a worse grade.

## 4. 1The consideration on relation between the grade and the average power

Figures 4(a),(b),(c),(d) show the distributions of data for four Chinese aspirated sounds, pai[p'ai], pao[p'ao], pei[p'ei] and pie[p'iɛ], the abscissa represents the VOT and the ordinate represents the relative average power Pav. The average grade was added to the students' individual marks. Points were also plotted for the Chinese speakers to provide a reference.

### 4. 1. 1Case of the grade dependingon the power

Figure 4(a) shows the data for the aspirate syllable of pai[p'ai]. D1, located at the left, has a VOT of 20 ms and a relatively high power of 19. D2, located far below D1, had a longer VOT of 26 ms than that of D1. But the power is as low as 1.2, the grades difference of 2.6 and 1.3 is large. D3, located in the upper part, had a VOT of 35 ms and the grade of 3.0. Although the difference of VOT of D3 and D2 is only 9 ms, the power difference of 865 and 1.2 is very large. The pronunciation with a higher power gets a better grade. The figure shows that the data for the Chinese speakers had sufficient VOT or power even with a short VOT.

Figure 4(b) shows the data for the aspirate syllable, pao[p'ao]. D1, located at the top, has a VOT of 40 ms and grade of 3.0. D2, located just below D1, has VOT of 35 ms and grade of 2.4. Although the difference of VOT is just 5ms, the power difference of 340 and 50 is large. The pronunciation with a higher power gets a better grade. When we exclude four data with lower power, the VOT length and power of the other data are about the same as Chinese speaker's data.

Figure 4(c) shows the data for the aspirate syllable, pei[p'ei]. D1, located at the middle left, has a VOT of 25 ms and a relative power of 7.7. Although D2, located below D1, has almost the same VOT with D1, it has a low power of only 0.98. The grades received were 2.9 and 2.3, respectively. These data show that syllables pronounced with higher power received a higher grade. The data of Chinese speakers have the long VOT and higher power in the upper right of the figure.

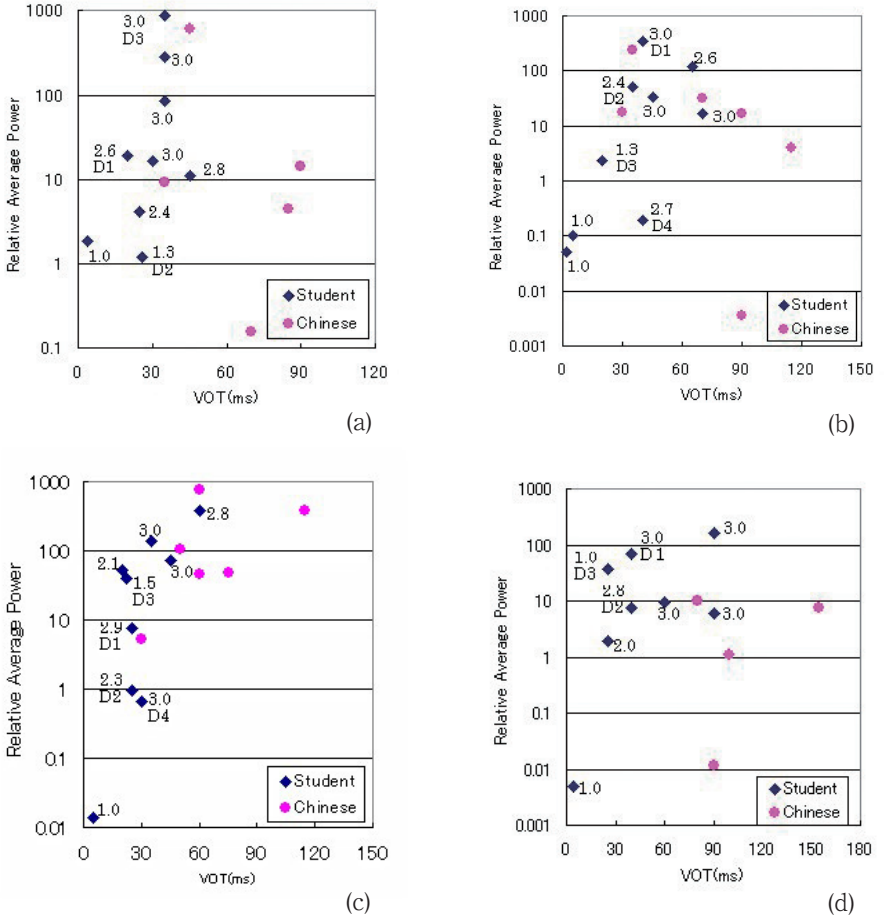


Figure 4: Data distribution of the aspirated syllables pai[p'ai] (a), pao[p'ao] (b), pei[p'ei] (c) and pie[p'iɛ] (d) on the surface of VOT in abscissa and relative average power in ordinate.

Figure 4(d) shows the data for the aspirate syllable, pie[p'iɛ]. D1, located at the second from the top, has a VOT of 40ms, a relative power of 68 and the grade of 3.0. Although D2, located below D1, had almost the same VOT with D1, the power is lower than D1, and the grade is a little lower than D1. As for the pronunciation of the students which had the longer VOT than 40 ms, it had a high power, and received higher grade than 2.8.

The upper examples show that the pronunciation with a higher power gets a better grade even though their VOTs are nearly equal or shorter than those with lower power

The grade of pronunciations with VOT between 20 and 30 ms for the aspirated of diphthong syllable pai[p'ai], that between 35 and 40 ms for pao[pa'o], that between 25 and 29 ms for pei[p'ei] and that between 30 and 40 ms for pie[p'iɛ] does not depend so much on the length of VOT rather depends on the average power used to breathe during VOT.

### 4. 1. 2 Case of very short VOT

The dependency of the grade on the power is not always true. Some examples show that dependency is not found, if the VOT is very short. Figure 4(b) shows the data for the aspirate syllable, pao[p'ao]. D3, located at the third from left, had a VOT of 20 ms and a power of 1.3. D4, located right below D3, had a lower power of 0.2. However, it received a higher grade of 2.7 than that of D3. Figure 4(c) shows the data for the aspirate syllable, pei[p'ei]. Although D3, located at the third from left, had a high power of 40, it received a low grade of 1.5. On the contrary, although D1, located below D3, had a lower power of 7.7 than that of D3, it received a higher grade of 2.9 than that of D3. Figure 4(d) shows the data for the aspirate syllable, pie[p'iɛ]. D3, located at the second from left, had a VOT of 25 ms and a high power of 38, however it received the worst grade of 1.0. On the contrary, although D2, located at right below D3, had a lower power of 7.4 than that of D3, it received the highest grade of 3.0. The same tendency is observed for the other syllables with a very short VOT.

### 4. 1. 3 Case of long enough VOT

The other examples show that grade does not depend so much on the breathing power.

As figure 4(a) for the aspirated syllable, pai[p'ai] shows, the datum of Chinese speaker at the bottom, has a VOT of 70 ms and the lowest power of 0.15. As figure 4(b) for aspirated syllable, pao[p'ao] shows, although D4, located at the third from the bottom, had a VOT of 40 ms and a low power of 0.2, it received a high grade of 2.7. The datum of Chinese speaker at the bottom, had the second longest VOT of 90 ms, but the lowest power of 0.005 in the same figure. As figure 4(c) for the aspirate syllable, pei[p'ei] shows, although D4, located at the second from the bottom, has a VOT of 30 ms and a low power of 0.7, it received the highest grade of 3.0.

All the examples shown above imply that the grades did not correlate closely with power if the length of the VOT was longer than a certain values. The same tendency is also observed in the other syllables.

## 5. CONCLUSIONS

We attempted to establish some useful evaluation measures to develop methods for teaching students how to pronounce Chinese aspirated syllables correctly. We examined the VOT and relative average power, Pav, during the VOT for four diphthong syllables of aspirated bilabial Chinese sounds uttered by nine native Chinese speakers and nine Japanese students. Grades for the pronunciation of each sound were determined by taking the average value of three grades given by eight native Chinese speakers in a hearing test. When the length of the VOT for the Chinese aspiration was within a certain range, pronunciation made using higher power received a higher grade, and vice versa. The results indicate that the quality of the pronunciation of aspirated sounds depends not only on the VOT but also on the power used during the VOT as we showed in the previous paper[1][2][3][4] for the case of the single-vowel aspirated syllables.

The average VOT of the pronunciation of the students is shorter than that of Chinese speakers, and about a half. If the VOT was very short, the pronunciation sounded poor. If the VOT was long enough, the pronunciation sounded good regardless with the power. Especially, the VOT of pronunciation of the Chinese speakers was enough long and the power was low. The results showed that there was no depend-

ency on the power in these VOT periods.

The results of these assessments suggest that, in general, when an aspirated syllable is pronounced with a long VOT it sounds as though it is correctly aspirated. However, our findings show that the quality of the pronunciation of aspirated sounds depends not only on the VOT but also on the power used during the VOT.

In order to develop the instruction device for the pronunciation of aspirated syllables, we will continue the effort to establish effective evaluation measures.

## 6 . REFERENCES

[1] A. Hoshino and A. Yasuda, 'Evaluation of Chinese aspiration sounds uttered by Japanese students using VOT and power (in Japanese),' The Journal of the Acoustical Society of Japan, Vol. 58, No. 11, pp. 689-695, Nov. 2002.

[2] A. Hoshino and A. Yasuda, "The Evaluation of Chinese Aspiration Sounds Uttered by Japanese Student Using VOT and power", 2003 International Conference on Acoustics, Speech, and Signal Processing IEEE Proceedings pp.472-475, Hong Kong,2003.

[3] A. Hoshino and A. Yasuda, "Dependence of Correct Pronunciation of Chinese Aspirated Sounds on Power During Voice onset Time", Proceeding of ISCSLP 2004, pp.121-124, Hong Kong, 2004.

[4] A. Hoshino and A. Yasuda, "Effect of Japanese Articulation of Stops on pronunciation of Chinese Aspirated Sounds by Japanese Students", Proceeding of ISCSLP 2004, pp.125-128, Hong Kong, 2004.

[5] Ray D. Kent and Charles Read, 'The Acoustic Analysis of Speech,' Singular publishing Group, Inc., San Diego and London, pp. 130-132, 1992.

[6] Zhu Chuan, 'Studying Method of the Pronunciation of Chinese Speech for Foreign Students (in Chinese)', Yu Wu Publishing Co., China, pp.63-71, 1997