Automatic Discrimination of Pronunciations of Chinese Retroflex and Dental Affricates

Akemi Hoshino[1], Akio Yasuda[2]

[1] Toyama National College of Technology, Ebie, Neriya, Imizu-city, Toyama, Japan
hoshino@nc-toyama.ac.jp
[2] Tokyo University of Marine Science and Technology, Etchujima, Koko-ku, Tokyo, Japan
yasuda@kaiyodai.ac.jp

Abstract. Retroflex aspirates in Chinese are generally difficult for Japanese students learning pronunciation. In particular, discriminating between utterances of aspirated dental and retroflex affricates is the most difficult to learn. We extracted the features of correctly pronouncing the aspirated dental affricates ca[ʦ‘a], ci[ʦ‘i], and ce[ʦ‘ɤ] and aspirated retroflex affricates cha[tʂ‘a], chi[tʂ‘i], and che[tʂ‘ɤ] by observing the spectrum evolution of breathing power during both voice onset time and voiced period of sounds uttered by nine Chinese native speakers. We developed a 35-channel filter bank on a personal computer to analyze the evolution of breathing power spectrum by using MATLAB. We then automatically evaluated the utterances of 20 students judged to be correct by native Chinese speakers and obtained a success rate of higher than 90% and 95% for aspirated retroflex and dental affricates, respectively.

Key words: Chinese aspirated retroflex and dental affricates, pronunciation training

1 Introduction

Retroflex aspirates in Chinese are generally difficult for Japanese students learning pronunciation, because the Japanese language has no such sounds. In particular, discriminating between utterances with aspirated dental and retroflex affricates is the most difficult to learn. We observed a classroom of Japanese students of Chinese uttering aspirated retroflex sounds modeled after examples uttered by a native Chinese instructor. However, the utterances sounded like a dental affricate to the instructor, and many students could not produce the correct sounds. They could not curl their tongues enough to articulate correctly, because there is no retroflex sounds in Japanese syllables.

adfa, p. 1, 2011.

We previously [1,2,3,4,5] showed that the breathing power during voice onset time (VOT) is a useful measure for evaluating the correct pronunciation of Chinese aspirates. We also developed an automatic evaluation system [6,7] for the students pronouncing Chinese aspirated affricates in accordance with the two parameters of VOT length and the breathing power during VOT.

However, since the system does not quite discriminate between aspirated retroflex and the dental affricates, we extracted the features of correctly pronouncing the aspirated dental affricates ca[ʦ‘a], ci[ʦ‘i], and ce[ʦ‘ɤ] and the aspirated retroflex affricates cha[tʂ‘a], chi[tʂ‘i], and che[tʂ‘ɤ] by analyzing the spectrum of breathed power during VOT of sounds uttered by Chinese native speakers. For this research, we developed a 35-channel frequency filter bank by using a personal computer. We found that the main difference between aspirated dental and retroflex affricates appeared in the spectrogram of the breathed power during VOT [8].

To improve the discrimination of these affricates, we extracted the features of correctly pronouncing aspirated dental affricates and aspirated retroflex affricates by analyzing the frequency spectrum of breathed power during both VOT and inside the voiced period of sounds and established improved evaluation criteria. We discuss the results of successfully discriminating between aspirated dental affricates and aspirated retroflex affricates by Japanese students.

We will continue to apply our system to other Chinese aspirated affricates to develop automatic training system.

2 Difference between Aspirated Dental and Aspirated Retroflex Affricates

The affricate is a complex sound generated by simultaneously articulating explosive and fricative sounds as one sound in the same point of articulation.

In this chapter, we define the distinctive features that discriminate between the dental affricate [ʦ‘] and retroflex one [tʂ‘] by examining the spectrogram of the pairs ca[ʦ‘a] - cha[tʂ‘a], ci[ʦ‘i] - chi[tʂ‘i], and ce[ʦ‘ɤ] - che[tʂ‘ɤ] uttered by a native Chinese speaker.

Figure 1 shows the temporal evolution of spectrograms of the aspirated retroflex sound cha[tʂ‘a] (left) and the aspirated dental sound ca[ʦ‘a] (right) uttered by a Chinese speaker. The lower part of the figure shows the waveform of the voltage evolution picked up by a microphone. The ordinate extended upward shows the frequency component and the shade of the stripes implies the approximate power level at the corresponding time and frequency. The aspirate appears in the brief interval in the right spectrogram of ca[ʦ‘a], indicated by light and thin vertical stripes during VOT, between the stop burst and the onset of vocal fold vibrations. This time interval is called the VOT [9], which is long, 160 ms. Although slightly darker stripes appear between 2500 and 5000 Hz in frequency and 70 and 150 ms in VOT, the temporal variation in the breathing power during VOT is not significant.

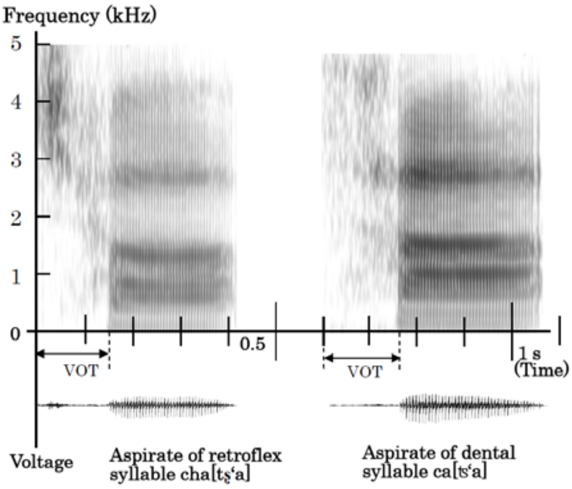


Fig. 1 Spectrograms of aspirated retroflex affricate cha[tʂʻa] (left) and aspirated dental affricate ca[ʦʻa] (right) pronounced by Chinese speaker

The left spectrogram is for the aspirated retroflex sound cha[tʂʻa] uttered by a Chinese speaker. The VOT was long, 150 ms. The dark vertical stripes in the upper left were observed between 2500 and 5000 Hz in frequency, during 0~70 ms of VOT. This is caused by friction of breath during breath release, which arises at a spot between the curled tongue and posterior alveolar. The large energy in the mouth dissipates at the early stage of VOT and generates high breathing power there. The thick horizontal bands in the voiced period in the right part of the spectrogram imply the formants that help to discriminate between the three dental affricates. The criteria are discussed later.

Figure 2 shows the temporal variation in spectrograms of the aspirated retroflex sound chi[tʂʻi] (left) and the aspirated dental sound ci[ʦʻi] (right) uttered by a Chinese speaker. The VOT of the aspirated dental sound ci[ʦʻi] was long, 225 ms, on the right hand side of the spectrogram. The unvarying darkness of the vertical bands shows that breathing power was rather steady during VOT. The left spectrogram is for the aspirated retroflex sound chi[tʂʻi]. The VOT was long, 250 ms.

During almost the entire VOT, the dark vertical stripes were observed in the frequencies between 2000~5000 Hz. This is due to the friction of breath at the breath release, which arises at a spot between the curled tongue and posterior alveolar.

Figure 3 shows the temporal variation in spectrograms of the aspirated retroflex sound che[tʂʻɤ] (left) and the aspirated dental sound ce[ʦʻɤ] (right) uttered by a Chinese speaker. The VOT of the aspirated dental sound ce[ʦʻɤ] was long, 180 ms. The stripes above 2000 Hz are darker and imply slightly stronger breathing power there.

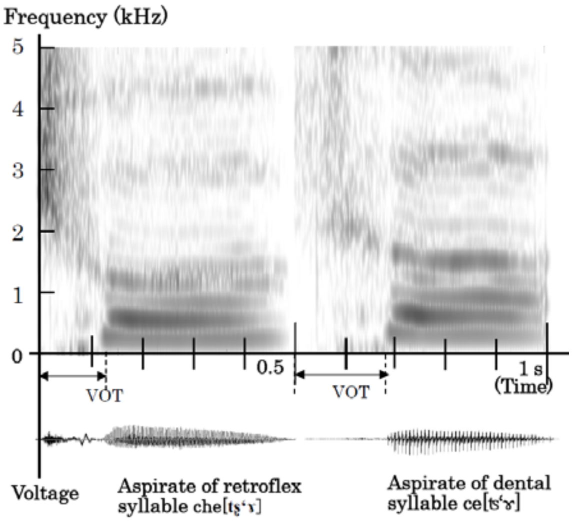


Fig. 2 Spectrograms of aspirated retroflex syllable chi[tʂʻi] (left) and aspirated dental syllable ci[ʦʻi] (right) pronounced by Chinese speaker

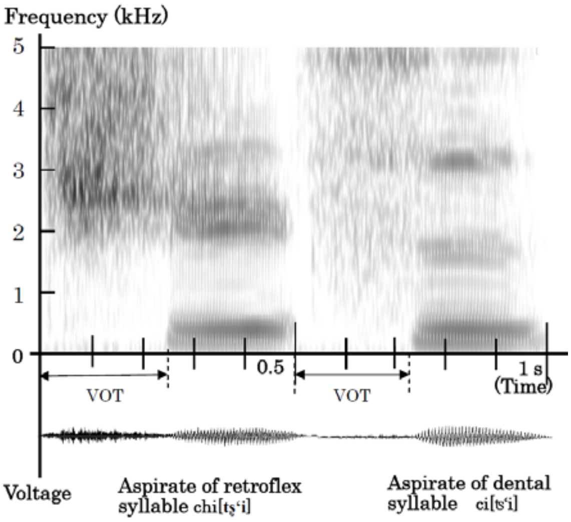


Fig. 3 Spectrograms of aspirated retroflex syllable che[tʂʻɤ] (left) and aspirated dental syllable ce[ʦʻɤ] (right) pronounced by Chinese speaker

For the frequency lower than 1200 Hz in VOT, the vertical stripes are light in accordance with weak breathing power. The distinctive feature of aspirated retroflex affricates is that they have a non-uniform spectrum in frequency and/or time during VOT, whereas aspirated dental ones have a rather uniform spectrum, as shown in the right spectrogram.

3 Automatic Measurement of VOT and Breathing Power

We showed that the correct utterance of aspirated retroflex and dental affricates is closely related to the frequency spectrum in VOT.

We previously developed an automatic measurement system of VOT and the breathing power by using a personal computer containing a 35-channel frequency filter bank, designed using MATLAB, in which the center frequency ranged from 50 to 6850 Hz with a bandwidth of 200 Hz [6,7]. We can extract the features of aspirated retroflex affricates and aspirated dental affricates of the frequency spectrum in both VOT and voiced periods.

3.1 VOT Measurement Algorithm

We automatically detected the onset of burst. Pronounced signals were introduced into the filter bank and split into the power at each center frequency every 5 ms. The start time of VOT, t1, was determined by comparing the powers for the adjacent time frames when the number of temporally increasing channels was maximum. The end of VOT, t2, was the start point of the formant. Thus, t2-t1 is defined as VOT.

We described the features of correct pronunciation of aspirated dental and retroflex affricates by observing the temporal variation of breathing power spectrum during VOT in Chapter 2. The powers at each frequency of the 35 channels every 5 ms with 11.025 kHz sampling were added in accordance with the frequency criteria defined in Chapter 2 during VOT.

3.2 Breathing Power Measurement Algorithm

The average power during VOT is defined as follows. The powers are deduced every 5ms and are referred to as Pi,j. which is the power at j×5ms of the i(1-35)-channel where Pi is the integration of the power at each time in VOT of the i-channel, as shown in Equation (1).

$$P_i = \sum_{j=1}^{J} P_{i,j}\,(t_j) \qquad (1)$$

Thus the energy Wi of the i-channel is defined as

$$W_{i,VOT} = P_i \times 5\mathrm{ms} \qquad (2).$$

The average power, Pi,av, of each frequency channel during VOT is defined as

$$P_{i,av}=W_{i,VOT}/VOT \qquad (3).$$

The average power at i-channel in voiced period, Tvs, Pvi,av can be defined similarly as

$$Pv_{i,av}=W_{i,vs}/T_{vs} \qquad (4).$$

4 Relationship between Breathing Power and Its Frequency Dependency during VOT and Quality of Pronunciations

Although several reports [9,10] on voiced retroflex have been published, there have been few reports on aspirated retroflex. We define the discrimination criteria of aspirated dental and aspirated retroflex affricates by examining the VOT and the breathing power spectrum during VOT of pronunciation of the pairs ca[ʦ‘a] - cha[tʂ‘a], ci[ʦ‘i] - chi[tʂ‘i], and ce[ʦ‘ɤ] - che[tʂ‘ɤ] uttered by 20 Japanese students. We used our automatic measuring system to define the parameters.

4.1 Scoring of Pronunciation Quality of Students

To investigate the correct pronunciation criteria of the aspirated retroflex affricates cha[tʂ‘a], chi[tʂ‘i], and che[tʂ‘ɤ] and the aspirated dental ones ca[ʦ‘a], ci[ʦ‘i], and ce[ʦ‘ɤ], the sounds uttered by 20 Japanese students were ranked using a listening test of the reproduced sounds conducted by nine native Chinese speakers [1-7]. The scores were as follows: 3 = correctly pronounced aspirated retroflex affricate or aspirated dental affricate; 2 = unclear sounds; and 1 = pronunciation in which the aspirated retroflex sounds were judged to be aspirated dental sounds and vice versa. We defined an average score of more than 2.6 as good. This score corresponds to the case in which six examiners give a score of ‘3’ and three give a score of ‘2’. 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.

4.2 Relationship between Scoring of Student Pronunciation and Evaluation Parameters

We now discuss the distribution of the student data with their scores are displayed on the surface of VOT and power respectively on abscissa and ordinate.

Figure 4 shows the data distributions on the surface of VOT and power with the scores of student pronunciations of aspirated retroflex cha[tʂ‘a] and aspirated dental ca[ʦ‘a]. The power of each utterance in this figure was automatically calculated at the frequencies between 2750 Hz (Channel-15) and 5750 Hz (Channel-29) averaged during the start time of VOT to 1/2VOT. The pronunciations of cha[tʂ‘a] with a good score gathered in the upper right from the center of the figure. The uttering power of aspirated retroflex affricate cha[tʂ‘a] increased by a continuous sequence of fricative articulations, and utterances with the power higher than 17 received scores higher than 2.6. In contrast, the utterances with insufficient curling of the tongue received a low score. The data with the power weaker than 16 received low scores. As for the utterance of aspirated dental ca[ʦ‘a], the data gathered downward a little from the middle of the figure. The data with the power of 8~12 scored higher than 2.8. The two data points of aspirated dental syllable ca[ʦ‘a], located at the top left, received a low score, presumably because unnecessary curling of the tongue resulted in high power utterance.

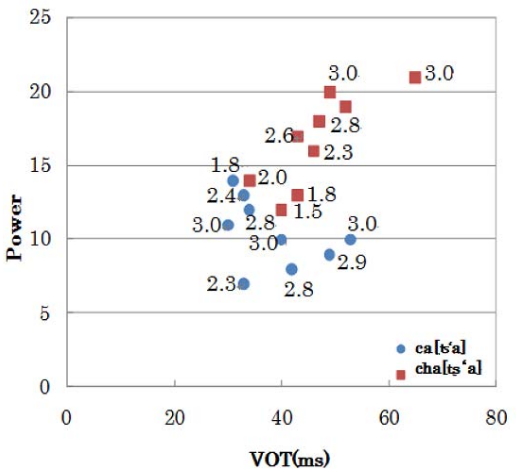


Fig. 4 Data distribution and scores for retroflex aspirated syllable cha[tʂ‘a],and dental aspirated syllable ca[ʦ‘a] with VOT on the abscissa and P_{av} at (2750-5750 Hz) on the ordinate.

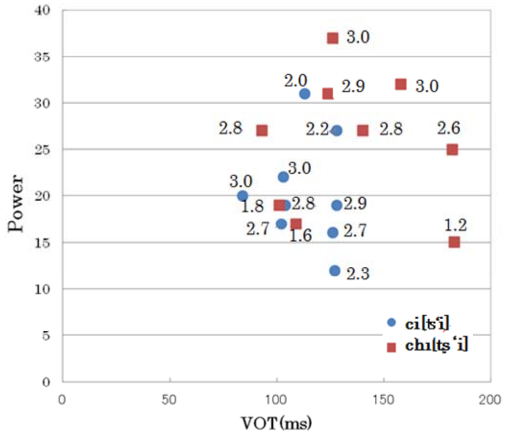


Fig. 5 Data distribution and scores for retroflex aspirated syllable cha[tʂ‘i] and dental aspirated syllable ca[ʦ‘i] with VOT on the abscissa and P_{av} at the frequencies between 1750 and 6350 Hz on the ordinate

Figure 5 shows the data distributions on the surface of VOT and power with the scores of the student pronunciations of aspirated retroflex affricate chi[tʂ‘i] and aspirated dental affricate ci[ʦ‘i]. The power of each utterance in this figure is summed one between the frequencies of 1750 Hz (Channel-10) and 6350 Hz (Channel-32) in VOT. The utterance with power higher than 25 of aspirated retroflex affricate chi[tʂ‘i] receives a good score. Three utterance data points in the lower part of the figure, of aspirated retroflex affricate chi[tʂ‘i] had utterance powers that are too low to pass the scoring test, i.e. 1.8, 1.6, and 1.2. As for utterances of aspirated dental syllable ci[ʦ‘i], the data with powers of 16~22 obtained higher scores.

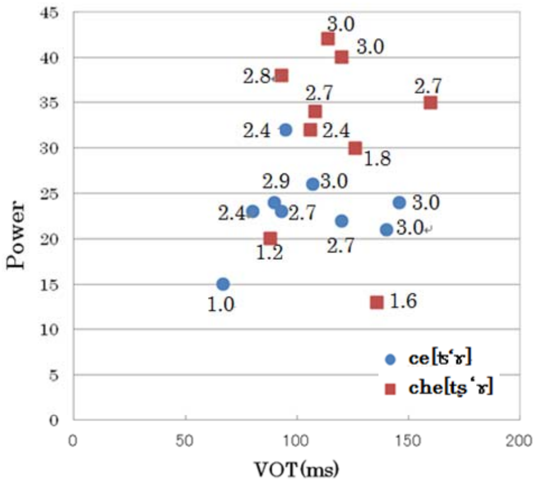


Fig. 6 Data distribution and scores for retroflex aspirated affricate che[tʂ‘ɤ] and dental aspirated affricate ca[ʦ‘ɤ] with VOT on the abscissa and P_{av} at the frequencies between 1150 and 5950 Hz on the ordinate.

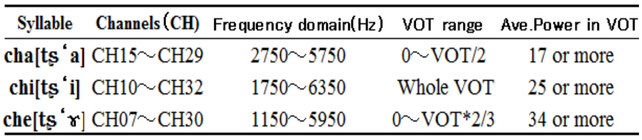
Figure 6 shows the data distributions on the surface of VOT and power with the scores of the student pronunciations of aspirated retroflex syllable che[tʂ‘ɤ] and aspirated dental affricate ce[ʦ‘ɤ]. The power of each utterance in this figure is summed one between the frequencies of 1150 Hz (Channel-7) and 5950 Hz (Channel-30) in VOT. Pronunciations of the aspirated retroflex affricate chi[tʂ‘ɤ] with the power higher than 34 scored higher than 2.7. Pronunciations with the power lower than 32 were not correct. For the pronunciations of the aspirated dental affricate ce[ʦ‘ɤ], the data with the power between 20~26 obtain successful scores.

5 Automatic Discrimination of Aspirated Retroflex Affricates and Dental Affricates

5.1 Parameters for Discrimination.

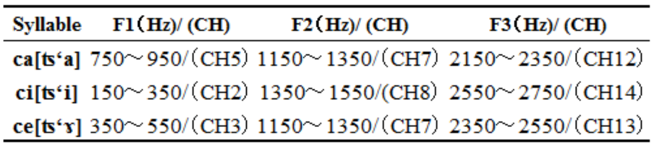
Table 1 lists the evaluation criteria on utterances of retroflex aspirated affricates. If the power was higher than 17 between 2750 Hz (CH15) and 5750 Hz (CH29) averaged during the onset of VOT to 1/2 of VOT, the utterances were judged to be aspirated retroflex affricate cha[tʂ‘a]. If the power was higher than 25 between 1750 Hz (CH10) and 6350 Hz (CH32) throughout VOT, the utterances were judged to be aspirated retroflex affricate chi[tʂ‘i]. If power was higher than 34 at the frequencies between 1150 Hz (CH7) and 5950 Hz (CH30) averaged during the onset of VOT to 2/3 of VOT, the utterances were judged to be aspirated retroflex affricate che[tʂ‘ɤ].

Table 1 Evaluation criteria on utterance of retroflex aspirated affricates



Syllable	Channels（CH）	Frequency domain(Hz)	VOT range	Ave.Power in VOT
cha[tʂ‘a]	CH15～CH29	2750～5750	0～VOT/2	17 or more
chi[tʂ‘i]	CH10～CH32	1750～6350	Whole VOT	25 or more
che[tʂ‘ɤ]	CH07～CH30	1150～5950	0～VOT*2/3	34 or more

Table 2 Evaluation criteria on utterance of dental aspirated affricates of formant frequencies



Syllable	F1（Hz)/ (CH)	F2（Hz)/ (CH)	F3（Hz)/ (CH)
ca[ʦ‘a]	750～950/（CH5）	1150～1350/（CH7）	2150～2350/（CH12）
ci[ʦ‘i]	150～350/（CH2）	1350～1550/(CH8）	2550～2750/（CH14）
ce[ʦ‘ɤ]	350～550/（CH3）	1150～1350/（CH7）	2350～2550/（CH13）

Table 2 lists the evaluation criteria on the utterances of dental aspirated affricates, which depend on the formant frequency values of F1, F2, and F3. If high power appears between 750 and 950Hz, 1150 and 1350 Hz, and 2150 and 2350 Hz, the utterances were judged to be aspirated dental syllable ca[ʦ‘a]. If high power appeared between 150 and 350 Hz, 1350 and 1550 Hz, and 2550 and 2750 Hz, the utterances were judged to be aspirated dental syllable ci[ʦ‘i]. If high power appeared at the frequency channels between 350 and 550 Hz, 1150 and 1350 Hz, and 2350 and 2550 Hz, the utterances were judged to be aspirated dental syllable ce[ʦ‘ɤ].

5.2 Experiment and Results

We tried to discriminate between the pronunciations of the pairs ca[ʦ‘a] - cha[tʂ‘a], ci[ʦ‘i] - chi[tʂ‘i], and ce[ʦ‘ɤ] - che[tʂ‘ɤ] uttered by 20 Japanese students. All utterances were evaluated to be correct by a listening test involving four native Chinese speakers.

Figure 7 illustrates the flow of our system for automatically discriminating aspirated retroflex and aspirated dental affricates. In step 1, the uttered sounds are input to the computer. In step 2, the sounds are automatically analyzed using our developed 32-channel filter bank to create a database of the temporal variation of power spectrum. In step 3, VOT is deduced using the algorithm described in Subsection 3.1.

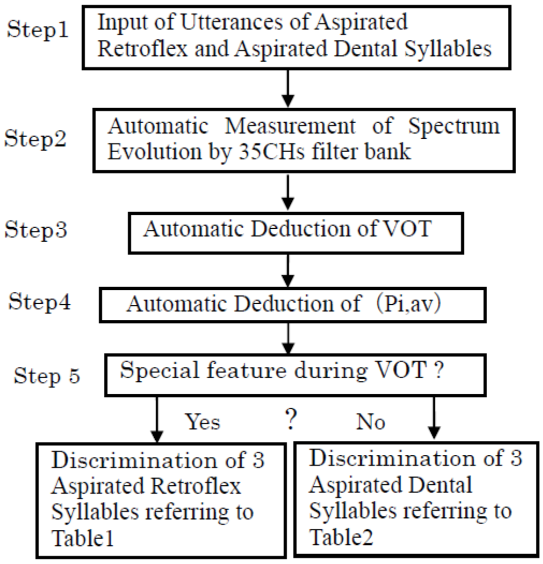
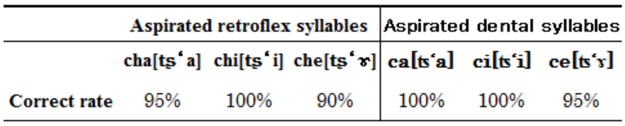

Step1
Input of Utterances of Aspirated Retroflex and Aspirated Dental Syllables
Step2
Automatic Measurement of Spectrum Evolution by 35CHs filter bank
Step3
Automatic Deduction of VOT
Step4
Automatic Deduction of（Pi,av)
Step 5
Special feature during VOT ?
Yes
?
No
Discrimination of 3 Aspirated Retroflex Syllables referring to Table1
Discrimination of 3 Aspirated Dental Syllables referring to Table2

Fig. 7 Discrimination diagram of aspirated retroflex and aspirated dental affricates

In step 4, the average power, Pi,av, is automatically calculated for each channel during VOT, as described in Subsection 3.2. In step 5, if any distinctive features are found during VOT, they are judged to be aspirated retroflex affricates and discriminated by referring to Table 1.

If there are no distinctive features during VOT, they are judged to be aspirated dental affricates and discriminated referring to Table 2. The Pvi,av is automatically calculated for each channel in voiced period, Tvs, as described in the Subsection 3.2.

Table 3 Correct judgment rate of aspirated retroflex and dental affricates



	Aspirated retroflex syllables			Aspirated dental syllables		
	cha[tʂ‘a]	chi[tʂ‘i]	che[tʂ‘ɤ]	ca[ʦ‘a]	ci[ʦ‘i]	ce[ʦ‘ɤ]
Correct rate	95%	100%	90%	100%	100%	95%

Table 3 lists the correct judgment rate of aspirated retroflex affricates cha[tʂ‘a], chi[tʂ‘i], che[tʂ‘ɤ] and aspirated dental affricates, ca[ʦ‘a], ci[ʦ‘i] and ce[ʦ‘ɤ] pronounced by 20 Japanese students. All utterances were evaluated to be correct by a listening test involving four native Chinese speakers.

The correct judgment rate of aspirated retroflex affricate cha[tʂ‘a] was 95%. One sample was too weak to be correctly detected. The correct judgment rate of retroflex affricate chi[tʂ‘i] was the perfect at 100%, and that of aspirated retroflex affricate che[tʂ‘ɤ] was the lowest at 90%. One utterance was too weak and another was too strong.

The correct judgment rates of aspirated dental affricates ca[ʦ‘a] and ci[ʦ‘i] were perfect at 100%, and that of aspirated dental affricate ce[ʦ‘ɤ] was 95%. One utterance had too little power.

6 Conclusion

We have been studying the instruction of pronunciation of Chinese aspirated sounds, which are generally difficult for Japanese students to perceive and reproduce. We closely examined the spectrograms of uttered sounds by native Chinese speakers and Japanese students and determined the criteria for correct pronunciations of various aspirated sounds [1-5]. We previously developed an automatic system for measuring and calculating the VOT and the power during VOT of student pronunciations [6,7].

In this paper, in order to develop an automatic training system for Chinese pronunciation, we aimed at automatic distinction of the three pairs of aspirated dental and aspirated retroflex affricates ca[ʦ‘a] - cha[tʂ‘a], ci[ʦ‘i] - chi[tʂ‘i], and ce[ʦ‘ɤ] - che[tʂ‘ɤ]. We automatically calculated the frequency spectrum of the utterance during VOT and voiced periods and extracted the distinctive feature of each utterance. Then we established criteria for automatically discriminating aspirated retroflex and aspirated dental sounds.

We conducted an experiment on automatic discrimination of 20 utterances of Japanese students using our automatic discriminating system. The results of the test showed that the system exhibited an average correct judgment rate for three aspirated retroflex affricates of 90% or more and aspirated dental affricates of 95% or more for the pronunciations evaluated to be correct by native speakers.

The authors appreciate for the financial support by Japan Society for the Promotion of Sciences (JSPS).

References

1. A. Hoshino, A. Yasuda, “Evaluation of Chinese aspiration sounds uttered by Japanese students using VOT and power (in Japanese), *Acoust. Soc. Jpn.*, **58**, No. 11, pp.689-695,(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, Hong Kong, pp. 472-475, (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, Hong Kong, pp. 121-124, (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, Hong Kong, pp. 125-128, (2004)
5. A. Hoshino and A. Yasuda, “Evaluation of aspiration sound of Chinese labial and alveolar diphthong uttered by Japanese students using voice onset time and breathing power,” Proceeding of ISCSLP 2006, Singapore, pp. 13-24,(2006)

6. A. Hoshino and A. Yasuda, “Pronunciation Training System for Japanese Students Learning Chinese Aspiration,” The 2nd International Conference on Society and Information Technologies(ICSIT), Orlando, Florida, USA,pp.288-293, (2011)
7. A. Hoshino and A. Yasuda, “Pronunciation Training System of Chinese Aspiration for Japanese Students,” Acoustical Science and Technology, Japan, Vol.32, No4, pp.154-157, July, (2011)
8. Hoshino, *etal,*Acoustics2012, April, 2012,Nantes, France. pp.339-344,(2012)
9. Ray. D. Kent, Charles Read, “The Acoustic Analysis of Speech,” Singular Publishing Group, Inc., San Diego and London, pp.105-109, (1992)
10. C. Zhu, “Studying Method of the Pronunciation of Chinese Speech for Foreign Students (in Chinese),” Yu Wu Publishing Co. China, pp. 63-71, (1997)
11. L. Zhou, H. Segi and K. Kido, “The investigation of Chinese retroflex sounds with time-frequency analysis,” The Acoustical Society of Japan, Vol54, No.8, pp.561-567, (1998)