ACOUSTICS 2012

Discrimination of Chinese pronunciations of aspirated dental and retroflex syllables according to breathing power and its frequency dependency during VOT

A. Hoshino and A. Yasuda

Toyama National College of Technology, 1-2 Ebie-neriya, Imizu city, Toyama, 933-0293, JAP,
933-0293 Toyama, Japan
hoshino@nc-toyama.ac.jp

Retroflexing of aspirates in Chinese is 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 syllables is the most difficult thing to learn. As we observed in our study, many students could not produce the correct sounds. Warping the tongue was not enough to produce the correct articulation, because there is no retroflex sounds amongst the Japanese syllables. In the present paper, the authors extracted the features of the correct pronunciation of the aspirated dental syllables ca[tsʻa], ci[tsʻi], and ce[tsʻɤ] and the aspirated retroflex ones cha[tʂʻa], chi[tʂʻi], and che[tʂʻɤ] by analyzing the spectrum of breathed power during voice onset time (VOT) of sounds uttered by nine Chinese native speakers. We found the main difference between aspirated dental and retroflex syllables appeared in the spectrogram of the breathed power during VOT. Then, we examined the relationship between the scores of 20 students' pronunciations and deduced evaluation parameters. We will develop an automatic discrimination system by using the deduced evaluation standards.

1 Introduction

Retroflexing of aspirates in Chinese is 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 syllables is the most difficult thing to learn. In our study, 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 dental syllables to the instructor, and many students could not produce the correct sounds. They cannot warp their tongues enough to articulate correctly, because there is no retroflex sounds amongst the Japanese syllables.

We previously [1,2,3,4,5] showed that the breathing power, Prel, 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' pronunciation of the Chinese aspirated syllables in accordance with the two parameters of length of VOT and the breathing power during VOT, which the present authors proposed.

In this study, the authors extracted the features of the correct pronunciation of the aspirated dental affricate syllables ca[tsʻa], ci[tsʻi], and ce[tsʻɤ] and the aspirated retroflex ones cha[tʂʻa], chi[tʂʻi], and che[tʂʻɤ] by analyzing the spectrum of breathed power during VOT of sounds uttered by nine Chinese native speakers. Then, we deduced these parameters for the pronunciations of 20 Japanese students and examined the relationship between the quality of the pronunciation and these parameters. For the next step, we are preparing to develop an automatic discrimination system by using the deduced evaluation standards.

2 Difference between aspirated dental affricate syllables and aspirated retroflex ones

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

The dental sound [tsʻ] of the aspirate of Chinese is called affricate of alveolar and is formed by articulating an explosive and a fricative sound at the point between the tip of the tongue and the tooth.

The Chinese aspirated retroflex sound [tʂʻ] is also called the sublamino-postalveolar affricate, and the point of articulation is the sublamino-postalveolar. In articulating it, one warps the tongue firmly and articulates explosive and fricative sounds simultaneously.

In this chapter, we define the feature that discriminates between the dental affricate [tsʻ] and retroflex one [tʂʻ] by examining the spectrogram of the pairs of ca[tsʻa] - cha[tʂʻa], ci[tsʻi] - chi[tʂʻi], and ce[tsʻɤ] - che[tʂʻɤ] uttered by a native Chinese speaker.

2.1 Difference between aspirated dental affricate syllable ca[tsʻa] and aspirated retroflex affricate syllable cha[tʂʻa]

Figure 1 shows temporal evolution of spectrograms of the aspirated retroflex sound cha[tʂʻa] (left) and the aspirated dental sound ca[tsʻa] (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 darkness 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[tsʻa], indicated by light and thin vertical stripes during VOT, between the stop burst and the onset of vocal fold vibrations followed by a vowel sound. This time interval is called the voice onset time, VOT [8]. It is long, 160 ms. Although the slightly darker stripes appear between 2500 and 5000 Hz in frequency and 70 and 150 ms in VOT, temporal variation of the breathing power during VOT is hardly seen.

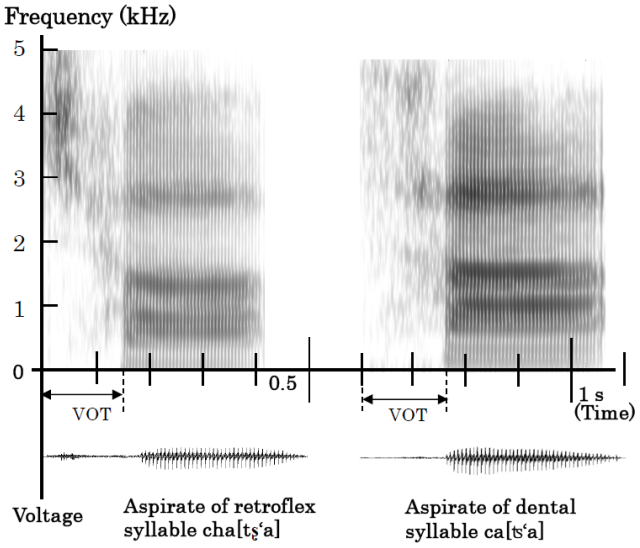

Figure 1: Spectrograms of retroflex aspirate syllable

cha[tʂʻa] (left) and dental aspirated syllable ca[tsʻa] (right) pronounced by a Chinese speaker.

The left spectrogram in figure 1 is for aspirated retroflex sound cha[tʂʻa] uttered by a Chinese speaker. The VOT was long, 150 ms. Dark vertical stripes are observed between 2500 and 5000 Hz in frequency, upper left during 0~70 ms in VOT. This is a friction of the breath caused by the breath explosion, which arises at a spot between the warped tongue and posterior alveolar. A large energy in the mouth dissipates at the early stage of VOT and generates the high breathing power there.

2.2 Difference between aspirated dental affricate syllable ci[tsʻi] and aspirated retroflex affricate syllable chi[tʂʻi]

Figure 2 shows temporal variation of spectrograms of the aspirated retroflex sound chi[tʂʻi] (left) and the aspirated dental sound aspirated ci[tsʻi] (right) uttered by a Chinese speaker. The lower part of the figure shows the voltage evolution picked up by a microphone. The VOT of the aspirated dental ci[tsʻi] was long, 225 ms, in the right hand side spectrogram. The unvarying darkness of the vertical stripes shows the breathing power is rather steady during VOT.

The left spectrogram in figure 2 is for the aspirated retroflex sound chi[tʂʻi]. The VOT was long, 250 ms. During almost the whole VOT, the dark vertical stripes are observed in the frequency band of 2000~5000 Hz. This is a friction of the breath caused by the breath explosion, which arises at a spot between the warped tongue and posterior alveolar. A large energy in the mouth dissipates during the whole VOT, and the strong breathing power is generated. Moreover, the weaker power was consumed at the frequency lower than 1500 Hz in the VOT, as the light vertical stripes imply.

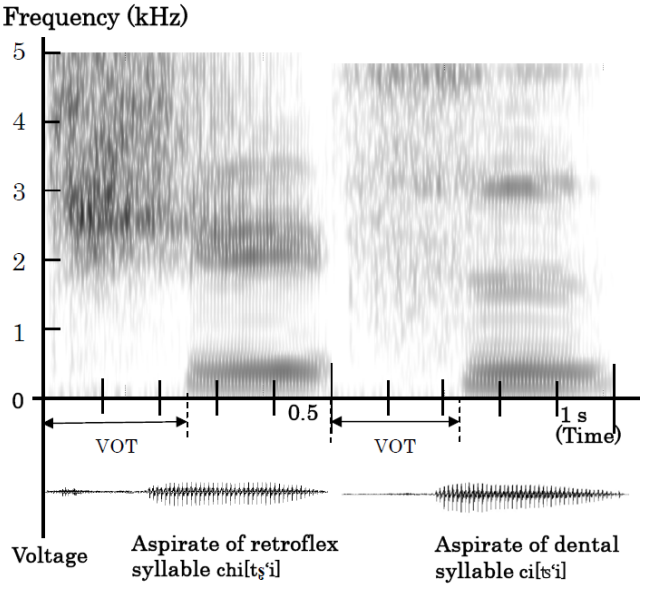

Figure 2: Spectrograms of retroflex aspirate syllable chi[tʂʻi] (left) and dental aspirated syllable ci[tsʻi] (right) pronounced by a Chinese speaker.

2.3 Difference between aspirated dental affricate syllable ce[tsʻɤ] and aspirated retroflex affricate syllable che[tʂʻɤ]

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

The left of figure 3 shows a spectrogram of the aspirated retroflex syllable che[tʂʻɤ]. The VOT was 125 ms. In the first half portion of VOT, dark vertical stripes in the frequency of 2000 Hz~5000 Hz are observed. This is also generated by a friction of the breath caused by the breath explosion, which arises at a spot between the warped tongue and posterior alveolar with a strong power, generated in this portion.

Moreover, for the frequency lower than 1200 Hz in VOT, the darkness of the vertical stripes is light in accordance with the weak power.

The distinctive feature of retroflex aspirated syllables is that they have non-uniform spectrum in frequency and/or time during VOT, whereas aspirated dental ones have rather uniform spectrum.

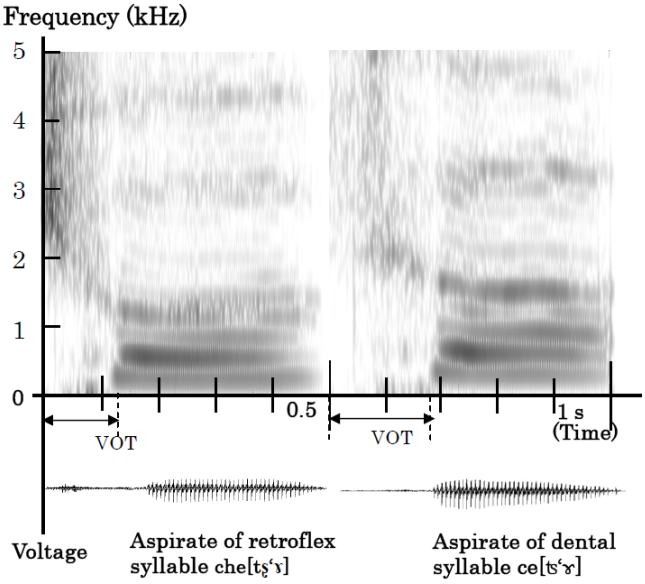

Figure 3: Spectrograms of retroflex aspirate syllable che[tʂʻɤ] (left) and dental aspirated syllable ce[tsʻɤ]] (right) pronounced by a Chinese speaker.

3 Relationship between power and its frequency dependency during VOT and the quality of pronunciations

Although several reports of research [9,10] on voiced retroflex have been published before, there are still few reports on research of aspirated retroflex. In this chapter, we define the discrimination standards of the aspirated dental affricate syllables and aspirated retroflex ones by examining the VOT and the power spectrum during VOT of the pronunciations of the pairs of ca[tsʻa] - cha[tʂʻa], ci[tsʻi] - chi[tʂʻi], and ce[tsʻɤ] - che[tʂʻɤ] uttered by 20 Japanese students.

3.1 Scoring of the pronunciation quality of students

To investigate the correct pronunciation criteria of the aspirated retroflex affricate syllables 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 in a hearing test of the reproduced sounds conducted by nine native Chinese speakers [1-7]. The grades were as follows: 3 = pronunciation that correctly sounded the aspirated retroflex affricate syllable or the aspirated dental one; 2 = unclear sounds; and 1 = pronunciation in which the aspirated retroflex sounds were judged to be aspirated dental sounds and vice versa. Data were excluded in cases of split evaluations and a standard deviation larger than 0.64, broken sounds uttered very closely to the microphone, and sounds with a low S/N uttered away from the microphone. Thus, more than half the utterances were abandoned. We defined a pronunciation with an average score of more than 2.6 as good. The score corresponds to the case where five examiners gave a score of ‘3’ and three gave a score of ‘2’.

3.2 Automatic measurement of VOT and Power

We automatically detected the onset of the burst by using a personal computer containing a 35-channel filter bank, designed using MATLAB, in which the center frequency ranged from 50 to 6850 Hz with a bandwidth of 200 Hz [6,7]. 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 the VOT, t2, was the start point of the formant. Thus, t2-t1 is defined as VOT.

We described features of correct pronunciation of aspirated dental and retroflex affricate syllables observing the temporal variation of breathed power spectrum during VOT in chapter 2. The powers at each frequency of the 35 channels every 5 ms were summed up in accordance with the frequency criteria defined in chapter 2 during VOT.

3.3 Relationship between the scoring of student pronunciation and evaluation parameters

The distribution of the students’ data with their scores is displayed on the surface of VOT and power respectively on the abscissa and ordinate.

Figure 4 shows the data distributions on the surface of VOT and power with the score of students’ pronunciation of aspirated retroflex cha[tʂ‘a] and aspirated dental ca[ʦ‘a]. The power of each utterance in this figure calculates automatically the power of the frequency between 2450 Hz (Channel-13) and 6850 Hz (Channel-35) averaged during the start time of VOT to the time of VOT/2. The pronunciations of cha[tʂ‘a] with a good score are gathered in the upper right from the middle of the figure. The ttering power of aspirated retroflex syllable cha[tʂ‘a] was made high by continuous sequence of fricative articulations, and utterances with the power higher than 17 received grades higher than 2.6. In contrast, the utterance with insufficient warping of the tongue received a low score. The data with power weaker than 16 received a low score. As for the utterance of aspirated dental ca[ʦ‘a], the data are 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 of aspirated dental syllable ca[ʦ‘a], located at the top left, received a low score, presumably because unnecessary warping of the tongue results in the high power utterance.

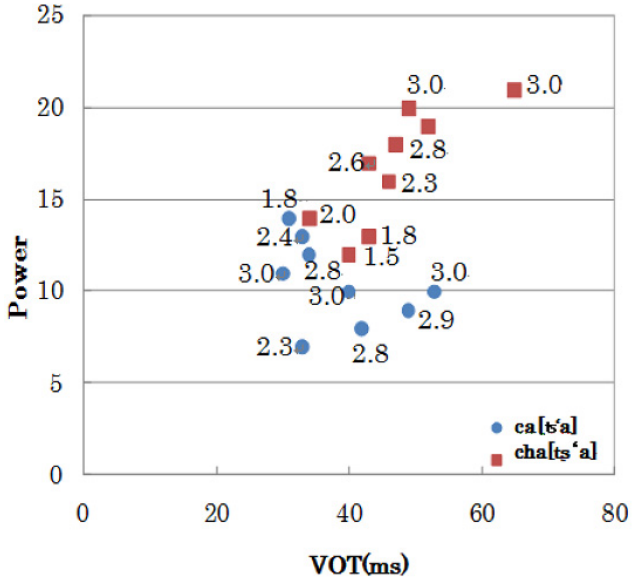

Figure 4: Data distributions on the surface of VOT and power with the score of students’ pronunciation of aspirated retroflex cha[tʂ‘a] and aspirated dental ca[ʦ‘a].

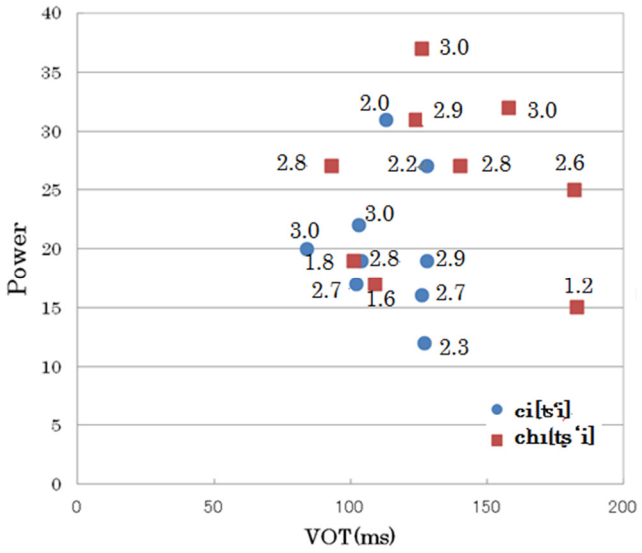

Figure 5: Data distributions on the surface of VOT and power with the score of students’ pronunciation of aspirated retroflex chi[tʂ‘i] and aspirated dental ci[ʦ‘i].

Figure 5 shows the data distributions on the surface of VOT and power with the score of the student’s pronunciation of aspirated retroflex chi[tʂ‘i] and aspirated dental ci[ʦ‘i]. Power of each utterance in this figure is summed one between the frequencies of 1450Hz (Channel-7) and 6850 Hz (Channel-35) in VOT. The utterance with power higher than 25 of the aspirated retroflex chi[tʂ‘i] receives a good score. Three utterance data, at the lower part of the figure, of the aspirated retroflex chi[tʂ‘i] have powers that are too low to pass the scoring test. They are

1.8, 1.6, and 1.2. As for utterances of aspirated dental ci[tsʻi], the data that have powers of 16~22 obtain higher scores. For the two data located at the top, the power was too high to obtain a high score. The tongue was unnecessarily warped too much in the utterance and high breathed power was produced as if it was for the retroflex sounds.

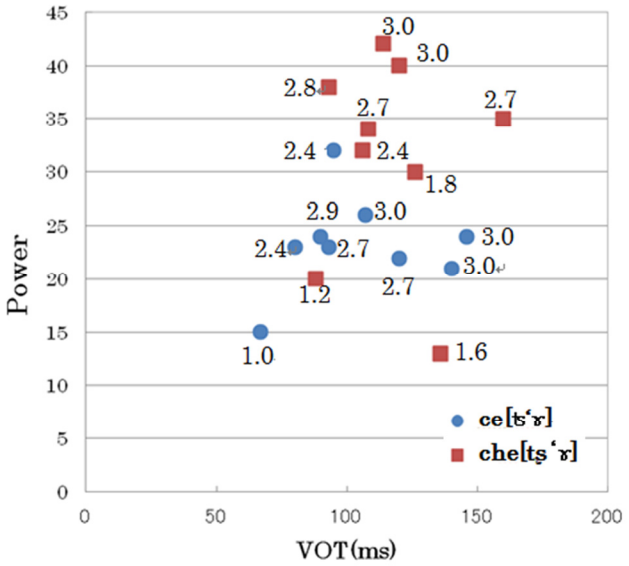

Figure 6: Data distributions on the surface of VOT and power with the score of students' pronunciation of aspirated retroflex che[tʂʻɤ] and aspirated dental ce[tsʻɤ].

Figure 6 shows the data distributions on the surface of VOT and power with the score of the student's pronunciations of aspirated retroflex che[tʂʻɤ] and aspirated dental ce[tsʻɤ]. Power of each utterance of this figure is summed one between the frequencies of 1250Hz (Channel-6) and 6850 Hz (Channel-35) in VOT. Pronunciations of the aspirated retroflex chi[tʂʻi] with power higher than 34 scored higher than 2.7. Pronunciations with power lower than 32 are not successful.

As for the utterance of the aspirated dental ce[tsʻɤ], the data with the power between 20~26 obtain successful scores.

4 Conclusion

The authors have been studying the instruction of pronunciation of Chinese aspirated sounds, which are generally difficult for Japanese students to perceive and reproduce as there is no such sound in Japanese pronunciation. We have examined the spectrograms of uttered sounds by native Chinese speakers and Japanese students closely and determined the standards for correct pronunciations in the various aspirated sounds [1-5]. Moreover, we developed an automatic system for measuring and calculating the VOT and the power during VOT of students' pronunciations and scored them [6,7]. The system automatically measured the voice onset time (VOT), and the breathed power by using a personal computer with a 35-channel filter bank in it, designed using MATLAB, in which the center frequency ranged from 50 to 6850 Hz with a bandwidth of 200 Hz.

In this paper, we deduce the distinctive features of three pairs of aspirated dental affricates syllables and aspirated retroflex ones of ca[tsʻa] - cha[tʂʻa], ci[tsʻi] - chi[tʂʻi], and ce[tsʻɤ] - che[tʂʻɤ], which are considered to be even more difficult for Japanese students to perceive and pronounce correctly, by using the correct pronunciation of the sounds uttered by nine Chinese native speakers. Then we applied the features to scored samples of 20 Japanese students.

Generally, the students' pronunciations that have long enough VOT and higher power than each threshold receive a higher score in the aspirated retroflex syllables, whereas those with power that is too high receive a low score in the aspirated dental affricates syllables, presumably because they are uttered with unnecessary warped tongues.

We are going to establish the evaluation standards for correct pronunciation using VOT and power spectrum of breathed power during VOT and are preparing to develop the automatic discriminating system of aspirated dental affricates syllables and aspirated retroflex ones using a personal computer.

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] Ray. D. Kent, Charles Read, "The Acoustic Analysis of Speech," Singular Publishing Group, Inc., San Diego and London, pp.105-109, 1992

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

[10] 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.