INTERNET TRAINING SYSTEM FOR LISTENING AND PRONUNCIATION OF CHINESE STOP CONSONANTS

Jiapeng Tian and Jouji Miwa

Department of Computer and Information Science, Iwate University
4-3-5 Ueda Morioka, Iwate, 020-8851, Japan

ABSTRACT

This paper describes a system for Chinese pronunciation and listening training on the Internet. The target Chinese stop consonants /b/, /p/, /d/, /t/ were chosen for this experiment.

In the pronunciation component, a dynamic low-pass filtered power (DLP) was used as an acoustic feature to discriminate between aspirated sounds and unaspirated sounds. With the sounds uttered by Chinese native speakers, we obtained a discrimination accuracy of 98.37% and 96.71% for aspirated consonants and unaspirated consonants respectively. Then we applied the pronunciation system to the sounds uttered by Japanese speakers who are learning Chinese as a second language and obtained an accuracy of discrimination. At last, a score is fed back to trainee according to the accuracy of his pronunciation and at the same time, the system instructs the trainee to improve his/her pronunciation.

In the listening component, in contrast to traditional methods, our system can not only let the trainee listen to the sound, but can also give him a feedback which will show whether or not his judgment about what he has heard is correct. The listening training system was constructed in Java language so that it can be used on any platform.

We will introduce the background and basic principles of the system, and give some examples of its use in language education.

1. INTRODUCTION

With the popularization of computers, the number of people using the Internet increases year by year. It has been possible to study foreign language through the Internet because of many online pronunciation instruction systems[1][2]. In order to instruct trainees just as a real human teacher does, many researchers have applied speech processing technology to their systems, by which the trainee’s sound can be analyzed or recognized and the result will be fed back to trainee. But little research has been reported for Chinese. Constructing such a system for Chinese is our purpose.

As target stimuli, we chose Chinese stop consonants /b/ ([p]), /p/ ([pʰ]), /d/ ([t]), /t/ ([tʰ]). Although these stop consonants are also included in many languages (such as Japanese), Chinese stop consonants are divided into aspirated consonants (called "AC" below) and unaspirated consonants (called "UC" below), unlike those of Japanese which are divided into voiced and voiceless sounds. For many foreign learners, especially for Japanese learners, stop consonants are some of the most difficult phonations to utter in Chinese.

2. DESCRIPTION OF THE WHOLE SYSTEM

The pronunciation component is the main part of the system, but in considering the correlation between hearing ability and phonation ability, we also constructed a listening component, expecting that it can bring a good effect to the phonation ability. Therefore the whole system consists of two components: the pronunciation component and the listening component. When using this system, the trainee will be instructed to do listening training exercises and pronunciation tests alternatively. Figure 1 shows a block diagram of the whole system.

![Figure 1: Block diagram of the training system.](image-url)
3. PRONUNCIATION TRAINING COMPONENT

The aim of this component is to analyze the trainee’s phonation, then give the trainee a feedback of score, and lastly instruct the trainee to improve his/her pronunciation by the score.

3.1. ACOUSTIC FEATURE

In the case of Chinese stop consonants, in contrast to those of Japanese, when the closure is opened, the vibration of the vocal cored is generally not accompanied. For AC, there is a considerable breath flow after the closure is opened, and for UC there is not. As a result, the consonant duration of AC is longer than that of UC. Using the features appeared in this duration, we can discriminate between AC and UC.

By observation of spectrum, we found that the time variation of power is clearly different during the transition region for AC and UC, especially in a low frequency band. In this paper, the time variation of power, which we called DLP (Dynamic Low-pass Power), is used as the acoustic feature. Figure 2 shows the procedure of extracting DLP.

![Figure 2: Procedure of extracting DLP.](image)

AC’s and UC’s low-pass filtered average power (shown on Figures 3 and 4) were obtained from the spectrum envelope through a low-pass filter (refer to equation (1)).

\[
p[i] = \frac{1}{n_1 - n_0 + 1} \sum_{i=n_0}^{n_1} G_i \Omega_n
\]

(\(\Omega_n = \frac{j \omega_n}{2 \pi}, N = 512\))

\(\Omega_0\) : low frequency (0Hz)  
\(\Omega_1\) : high frequency (515Hz)  
\(i\) : frame number  
\(G_i\) : log power spectrum envelope

Then DLP was calculated by equation (2). Equation (2) is a weighted filter which can calculate differential and remove small variations at the same time using moving average.

\[
dp[i] = \sum_{m=1}^{M} (m \cdot p[i + m])
\]

(2)

\(2M + 1\) is filter’s order, \(m\) is the coefficient of weight. \(M\) is decided considering that the small variation of DLP is made as small as possible while the meaningful peak of DLP can be determined. Figures 5 and 6 show DLP of AC and UC respectively.

![Figure 3: low-pass filtered power for AC.](image)

![Figure 4: low-pass filtered power for UC.](image)

![Figure 5: An example of DLP for AC.](image)
unspirated sound by Chinese speaker

![Diagram](image)

**Figure 6:** An example of DLP of UC.

In Figures 5 and 6, peak1 is the first meaningful peak at the explosive section and peak2 is the peak whose reliability (the height from the peak to the valleys on both sides) is the greatest at the transition section as shown in Figures 5 and 6 with range. AC and UC are mainly discriminated by the value of peak2.

### 3.2. DISCRIMINATION EXPERIMENT

Transition section range is determined by the distribution of consonant duration of AC. In addition, small variations are still appear because of the effect of noise. So we also use convexity reliability (the height from the peak to the valleys on both sides) as a parameter of discrimination. As a result, the discrimination algorithm is:

1. The first convex, which is greater than thpeak1 as well as the convexity reliability is greater than theon f, is peak1, meaning the explosive point.

2. In range of "range", if peak2 is convex which reliability is the greatest is greater than thpeak2, the consonant is AC. If no peak2 was found, set peak2 to zero. If peak2 isn’t greater than thpeak2, the consonant is UC.

The speakers involved in this experiment were 6 Japanese speakers (JS) who were all learning Chinese as a second language and 10 native speakers of Chinese (CS). The stimuli are about 200 real bisyllabic Chinese words and were chosen to have all possible combinations of initial stop consonants and final vowels, and four tones. In considering the practical training surroundings, the recordings did not take place in soundproof booths. Table 1 shows the analysis condition.

<table>
<thead>
<tr>
<th>Table 1: Analysis condition.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>sampling condition</strong></td>
</tr>
<tr>
<td><strong>Analysis interval</strong></td>
</tr>
<tr>
<td><strong>frame period</strong></td>
</tr>
<tr>
<td><strong>FFT</strong></td>
</tr>
<tr>
<td><strong>filter degree</strong></td>
</tr>
<tr>
<td><strong>frequency range</strong></td>
</tr>
</tbody>
</table>

By analyzing 10 CS’s speech data, distributions of feature parameters were obtained. From these distributions, threshold (such as thpeak1, thpeak2, theonf, range) for discrimination were determined. For example, thpeak1 and thpeak2 were determined from the scatter diagram of peak1 and peak2 shown on Figure 7.

![Diagram](image)

**Figure 7:** Scatter diagram of peak1 and peak2.

The threshold values are shown in Table 2.

<table>
<thead>
<tr>
<th>Table 2: Threshold value for discrimination.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>threshold parameter</strong></td>
</tr>
<tr>
<td>thpeak1</td>
</tr>
<tr>
<td>thpeak2</td>
</tr>
<tr>
<td>theonf</td>
</tr>
<tr>
<td>range</td>
</tr>
</tbody>
</table>

To evaluate the discrimination system, two experiments were done for 6 JS native speaker’s perception and discrimination by computer. We selected a phonetician for perception and adapted the threshold values obtained from CS to discrimination for JS. Table 3 shows the discrimination result for CS. Table 4 shows the rate that the discrimination result corresponded to the perception result.

<table>
<thead>
<tr>
<th>Table 3: Discrimination result for Chinese.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
</tr>
<tr>
<td>AC</td>
</tr>
<tr>
<td>UC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 4: Discrimination accuracy rate for JS (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>JS1</strong></td>
</tr>
<tr>
<td>AC</td>
</tr>
<tr>
<td>UC</td>
</tr>
</tbody>
</table>

For JS who haven’t mastered Chinese stop consonants, the possibility of vague utterances’ appearance is strong. On the other hand, stop consonants have too short a duration to be heard clearly and judged correctly even for a phonetician. This is one of the reasons why discrimination accuracy is low for some trainees. In this sense, the main purpose of this paper is to provide a quantitative assessment method rather than making discrimination accuracy as close as possible to perception accuracy.
3.3. SCORING AND INSTRUCTION

Many training systems return a feedback, like speech waveforms, spectrum, etc. to trainees. However, these kinds of feedback are too difficult to be understood for the general trainee. The most intuitive feedback is score. In this paper, we use a posterior probability (PP) of CS's feature variable peak2 to score JS's utterance. Figure 8 shows an example of scoring function for AC. The value of peak2 is larger, the score is higher.

![Figure 8: Scoring function for AC.](image)

In Chinese classroom education, we found the difference of phonation method between Chinese and Japanese. In Chinese, AC is uttered by so-called lung style phonation method which can not be found in Japanese. For UC, lung style phonation method is not needed, but when uttering UC, the closure is closed more firmly than Japanese voiceless consonants /b/ and /d/. For these reasons, if trainee's AC utterance's score is low, we instruct him to send out more breath from the lungs as well as close the closure more firmly before utterance; if trainee's UC utterance's score is low, we instruct him to close the closure more firmly before utterance and utter the next vowel as soon as possible after the closure is opened without time of sending out breath.

4. LISTENING TRAINING COMPONENT

Doing listening training by computer has many advantages such as (1) using anywhere at anytime (2) avoidance of self-conscious if mistakes occur (3) the possibility of feedback. In order to be used by all over the world on any platform such as Windows, Macintosh and UNIX, this component was constructed in Java language. Figure 9 shows an example for the listening component.

As target stimuli, we chose minimal pair bisyllable words (20 pairs 40 words for /b/ /p/ /d/ /t/) which means that the final vowel and tone in minimal pair words are the same except the initial stop consonant. Each time, two words are presented on screen in random order. At the same time one utterance of the two words are played. The trainee selects one word as the answer, then the system will judge if his selection was correct, display a circle mark for a correct answer or a cross mark for a mistake. In the textfield at the bottom, the system calculate the result of the whole training after the trainee began. If "hint" is clicked, a Japanese hint (the word's meaning) appears.

This listening training system will be opened on the Internet. Please refer to the web address below:

http://sp.cis.iwate-u.ac.jp/sp/lesson/c/

5. CONCLUSION

In this paper, we described a training system by computer for Chinese stop consonants /b/, /p/, /d/ and /t/. It consists of a pronunciation training component and a listening training component. We recommend trainees to do the two kinds of training alternately so that both their pronation ability and hearing ability will improve.

In the pronunciation component, dynamic low-pass filtered power was used as feature variable. Using the thresholds gained from Chinese native speakers data, we realized a quantitative measurement for Japanese speakers' data that make it possible to instruct the trainee adjust to their utterance method quantitatively.

In the listening component, Java language was adopted for the usage of the system on different platforms.

6. REFERENCE

