

MASTER THESIS

**A Learner Adaptive
Chinese Pronunciation Education System
for Japanese**

日本人学習者を対象とした適応型
中国語発音教育システム

by

Ma Minyi 48-076439

Supervisor

Professor Keikichi HIROSE

February 4th, 2009



Department of Information and
Communication Engineering
GRADUATE SCHOOL OF
INFORMATION SCIENCE AND
TECHNOLOGY
UNIVERSITY OF TOKYO

Abstract

This document deals with a Learner Level Adaptive Chinese Pronunciation Education System for Japanese.

As China continues to be one of the world's fastest growing economies, many foreigners are eager to learn Mandarin Chinese to enhance their international business. There is an increased urgency to find ways to help foreign speakers to learn Chinese.

In language learning, one of the most difficult parts is pronunciation. It is widely recognized that one of the best ways to learn natural pronunciation is through practice with a native teacher. However, this is not a practical method due to the lack of native teachers and the high expenses. A potential solution to this problem is to rely on Computer Assisted Language Learning systems for pronunciation education.

Through analyzing Japanese speakers' Mandarin speech data, we found out features of Japanese learners. We developed a Computer Assisted Language Learning system specifically for helping Japanese people to improve their Chinese pronunciation.

Our Learner Level Adaptive system can judge learners' pronunciation level by speech recognition technology, suggest practice emphases and provide corresponding pronunciation practice courses. Also, the system can provide corrective audio feedback in learners' own voice using speech modification technology.

Up to now, we had finished simple system evaluation. According the evaluation results, all testers felt the instructions and judgment results offered by the system are proper and helpful, all testers are making progress.

Acknowledgments

My very special thanks to Prof. Hirose. This work could not have been finished without your kind advice and superb guidance.

I would also like to sincerely thank all the members of my laboratory for their warmhearted helps and encouragement during my study in the laboratory. Without those discussions and suggestions, especially while writing this document, my everyday life would have been rather different. It's you all made my stay at the University of Tokyo so fruitful and enjoyable.

Finally, I am grateful to my parents, for their encouragement, love and understanding ...

Contents

Abstract	i
Acknowledgments	ii
Contents	iii
List of Figures	v
List of Tables	vii
1 Introduction	1
1.1 Background	1
1.2 Research Objective	2
2 Overview of CALL systems	3
2.1 What is CALL	3
2.2 CALL history	3
2.3 Classification	6
2.4 Advantages and Disadvantages	8
2.5 Two Selected CALL systems	9
2.5.1 A Chinese Learning System	9
2.5.2 PSC test Evaluation System	12
3 Mandarin Introduction	18
3.1 Phonemes	18
3.2 Tones	19

3.3	Difficulties for Japanese	21
4	Learners' Speech Analysis	23
4.1	Speech Data	23
4.2	Method	24
4.3	Experimental Results	25
4.4	Speech Modification Experiment	33
4.5	Conclusion	38
5	Adaptive CALL System	39
5.1	System Overview	39
5.2	Level Judgment Module	41
5.2.1	Phoneme Scoring	42
5.2.2	Tone Scoring	43
5.2.3	Judgment Report	44
5.3	Word-level Practice Course	47
5.3.1	Speech Modification using TD-PSOLA	48
5.4	Sentence-level Practice Course	52
5.4.1	Duration and Tonal Modification	52
5.5	Conclusion	53
6	System Evaluation	54
6.1	Recognition Rate	54
6.2	Judgment	55
6.3	Feedback	55
6.3.1	Instructions	55
6.3.2	Audio Feedback	55
6.4	Conclusion	56
7	Future Work	57
A	Reading Script for Speech Analysis	59
	Bibliography	63

CONTENTS

v

Publications

67

List of Figures

2.1	<i>A brief History of CALL</i>	4
2.2	<i>CALL History comparing with Computer Technology</i>	5
2.3	<i>NHK System Structure</i>	10
2.4	<i>Smoothing of pitch contour</i>	11
2.5	<i>Normalization in Time and Averaging of Pitch</i>	11
2.6	<i>Procedure for converting user's F0</i>	12
2.7	<i>Structure of PSC System</i>	13
2.8	<i>F0 distribution of different speakers</i>	15
2.9	<i>F0 distribution after normalization</i>	16
2.10	<i>Pronunciation error detection process</i>	17
3.1	<i>Relative pitch changes of the four tones</i>	20
4.1	<i>Correct and Incorrect Pitch contours</i>	28
4.2	<i>Difference of Utterance Rhythms</i>	33
4.3	<i>Vowel, Consonant, Short Pause Segmentation</i>	34
4.4	<i>Learners' Segmentation TXT File</i>	35
4.5	<i>Expanding or Contracting Time Axis of Learners' Data</i>	36
4.6	<i>Process of Pitch Modification</i>	36
4.7	<i>Listening Test Results of Modified Speech</i>	38
5.1	<i>Learner Adaptive CALL System Structure</i>	40
5.2	<i>Learner Adaptive CALL System Form Structure</i>	40
5.3	<i>Block Diagram of GOP Scoring</i>	46
5.4	<i>Learner Level Judgment Flow Chart</i>	47

5.5	<i>Pronunciation Instructions of 'r' and 'l'</i>	48
5.6	<i>Word-level Course Form Interface</i>	49
5.7	<i>Word-level Course Structure</i>	49
5.8	<i>Overall Principle of the PSOLA Method</i>	50
5.9	<i>Overview of the Teacher Mapping Technique</i>	51
5.10	<i>Sentence-Level Course Structure</i>	52
6.1	<i>Listening Test Results of Modified Speech</i>	56

List of Tables

3.1	<i>Unites of Chinese syllables</i>	19
4.1	<i>Speech Data Collected</i>	23
4.2	<i>Possible Vowel Mispronunciation List</i>	26
4.3	<i>Possible Consonant Mispronunciation List</i>	27
4.4	<i>Evaluation of Phonemes Pronunciation</i>	27
4.5	<i>Evaluation of Tones Pronunciation</i>	28
4.6	<i>Comparison of Character Duration between Non-native and Native</i>	30
4.7	<i>Arithmetic mean and SD of sentence durations with/without SP</i>	30
4.8	<i>Arithmetic mean and SD of sentence duratons with/without SP</i>	31
4.9	<i>Average Durations of per Vowel and Consonant</i>	31
4.10	<i>Average durations of per vowel and double-vowel</i>	32
4.11	<i>Comparison of Durations between Nouns and Verbs</i>	32
4.12	<i>Comparison of Consonant and Vowel Durations between Nouns and Verbs</i>	33
4.13	<i>Subjective Scoring Criteria</i>	37
4.14	<i>Listening Test Results of Modified Speech</i>	37
4.15	<i>Difficulties per Level</i>	38
5.1	<i>Recording Sentences of Level Judgment Module</i>	41
5.2	<i>HMM Monophone List</i>	42

Chapter 1

Introduction

1.1 Background

The world is changing rapidly, and learning a foreign language is more necessary nowadays than it ever was. Second language acquisition has been widely researched throughout the world, and the recent improvements in computers performances make it possible to use them as a tool for second language acquisition. Such systems are referred as CALL¹ systems.

With the development of China, Chinese is becoming important for one's career. Speaking even a little Chinese can greatly enhance your international business relations. Thus, nowadays in Japan, there are many Japanese people that are learning Chinese to acquire proficiency in communicating with Chinese.

Most people believe Chinese pronunciation is difficult. There are 37 vowels and 21 consonants in Chinese. The number of vowels and consonants implies the difficulties of Chinese pronunciation. There are four patterns of tones in a Chinese (Mandarin) syllable, and therefore, F0 movements are much more complicated than non-tonal languages like Japanese and English, etc.

In order to help Japanese people to improve their Chinese pronunciation, we collected Mandarin speech data from Japanese speakers and analyzed features of their utterances. We found that problems of different level – beginners or advanced learners – are different. Beginners are usually poor at phonemes and tones, which makes their utterances hard to be understood; while, advanced learn-

¹Computer Assisted Language Learning, will be introduced in details in further chapters.

ers may manage to pronounce phonemes well, but when they speak sentences, there are problems in tones and durations that make their utterances sounds unnatural. Thus, we proposed our Learner Level Adaptive CALL system to provide different pronunciation practice emphases to satisfy different levels of learners.

1.2 Research Objective

The primary goal of our project is to develop a computer application which is specialized to help Japanese people to improve their Chinese pronunciation.

The system is:

- Using Speech recognition technology to automatically evaluate learner 's speech.
- Producing learner level judgment report and recommending practice instructions.
- Providing corresponding effective pronunciation practice.
- Giving visual directions and corrective speech feedback in learner 's own voice.

This document is organized as follows. First, it gives an introduction of CALL and CALL systems. Second, it explains the main difficulties of Chinese pronunciation and introduces our speech analysis results of features of Japanese learners' utterances. Then, it presents detail of our learner level adaptive CALL system. Lastly, system evaluation results are summarized, leading to suggestions about further aspects that could be investigated if one would consider continuing research.

Chapter 2

Overview of CALL systems

2.1 What is CALL

Computer Assisted Language Learning ¹ is an approach to language teaching and learning in which computer technology is used as an aid to the presentation, reinforcement and assessment of material to be learned, usually including a substantial interactive element ².

For many years, foreign language teachers have used computers to provide supplemental exercises. In recent years, advances in computer technology have motivated teachers to reassess the computer and consider it a valuable part of daily foreign language learning. Innovative software programs, authoring capabilities, storage technology, and networks are providing teachers with new methods of incorporating pronunciation, grammar, culture, and real language use while students gain access to audio, visual, and textual information about the language and the culture of its speakers. Computers increasingly play an important role in education, particularly language learning.

2.2 CALL history

CALL is a field tied closely to other areas such as the development of linguistics, approaches of foreign language teaching and the development of computer science. Figure 2.1 shows brief history of CALL.

¹From now on referred as **CALL**

²Ken Beatty. "Teaching & Researching Computer-Assisted Language Learning "

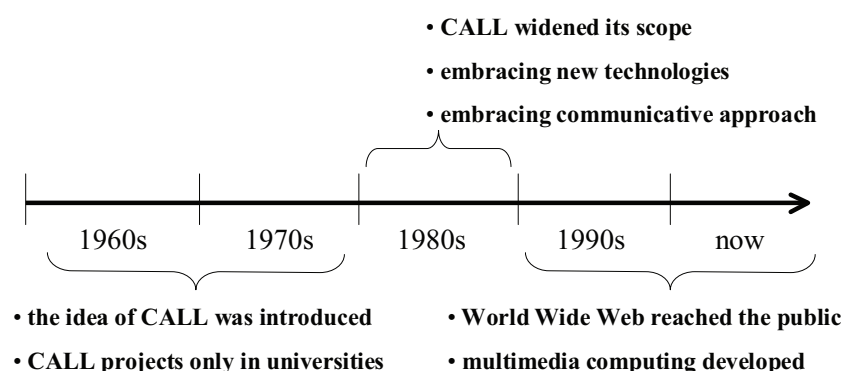


Figure 2.1: A *brief History of CALL*

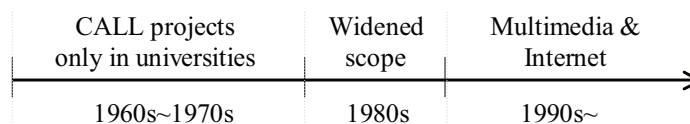
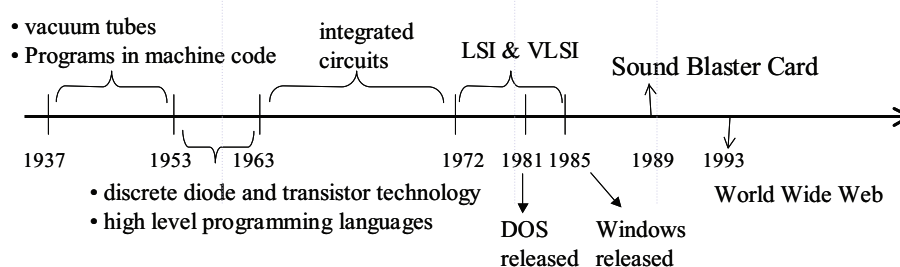
CALL's origins can be traced back to the 1960s. Up until the late 1970s, CALL projects were limited mainly to universities, where computer programs were developed on large mainframe computers. The PLATO project ³ is an important landmark in the early development of CALL. The early CALL favoured an approach that drew heavily on practices associated with programmed instruction.

The 1980s may be regarded as the most important decade in development of computing and in development of computers as a tool for learning. Throughout the 1980s, CALL widened its scope, embracing the communicative approach and a range of new technologies, especially multimedia and communication technology. It was also in the 1980s that major progress occurred in related areas of machine intelligence such as speech recognition, machine-assisted translation, artificial intelligence and generally natural language processing, including in CALL.

In the early 1990s, the World Wide Web reached the general public. The web is changing the world, and also had a great impact on CALL. Vocabulary practice, grammar lessons, reading and writing tasks, and even pronunciation exercises can be put on the web and made interactive in a variety of ways. Multimedia computing, the Internet, and the World Wide Web have provided an incredible boost to CALL applications. CALL is finally receiving the recognition it deserves thanks in large part to developing technologies.

Figure 2.2 shows brief history of CALL comparing with that of computer technology.

³Graham Davies. "Computer Assisted Language Learning"

History of CALL:**History of Computer Technology:**Figure 2.2: *CALL History comparing with Computer Technology*

I concluded the development of CALL systems into three stages:

Classical CALL system

In early CALL programs the stimulus was in the form of text presented on screen, and the only way in which the learner could respond was by entering an answer on the keyboard. In Classical CALL programs, people made sufficient use of the computer advantages, such as doing grammar tests and vocabulary tests that can be practiced again and again. Also, many reviews of grammar points, reading comprehension materials and dictionaries have been developed which are electronic versions of normal textbooks, providing students with drills and practices but with no judgmental feedback. Typical example is the TICCIT project⁴.

Communicative CALL system

Error analysis and feedback is a common feature of Communicative CALL. Based on the communicative approach, communicative CALL is favouring a learner centered, explorative approach, rather than a teacher centered, drill based approach. Communicative CALL programs provide skill practice through language

⁴1976-77 TICCIT Project. Final Report.

games, reading and text reconstruction, etc. It also uses the computer as a tool that enable the learner to understand and use the language, such as word processors, desk-top publishing, spelling and grammar check programs. One example is the CLEF package ⁵. The approach adopted by the authors of CLEF was to anticipate common errors and build in appropriate feedback.

Integrative CALL system

The most recent approach is Integrative CALL, which is based on multimedia computers and the Internet. These technological developments have brought text, graphics, sound, animation and video to be accessed on a single computer, enabling learners to navigate through CD-ROMS and the Internet at their own pace and path. For example, one feature of many multimedia CALL programs is the role-play activity, in which the learner can record his/her own voice and play it back as part of a continuous dialogue with a native speaker. Other multimedia programs make use of Automatic Speech Recognition (ASR) software to diagnose learners' errors. Most CALL programs under development today fall into the category of Integrative CALL.

2.3 Classification

Classification of CALL is difficult. From different viewpoints, there are different ways of classification. About education content, there are CALL for grammar, for pronunciation and for conversations; about education purpose, there are CALL for test preparation, for children's formative education; about technology being used, there are CALL using HTML, Weblogs, JavaScript, RealAudio, and CGI scripts.

I would like to introduce the form styles of CALL systems as follows:

1. E-materials

Large and ever growing collections of materials, taking the form of textbooks, which is also the form of earliest CALL, provide large amount of reference data. Since collections of resource have been perhaps the fastest growing area on the Web, metasites of CALL are also being created ⁶.

⁵CLEF French Grammar Package is sold at: <http://www.camsoftpartners.co.uk/clef.htm>

⁶An example of metasites: <http://www.word2word.com/coursead.html>

2. Drills and Practices

There is a multiplicity of exercises, most of them using fill-ins, (usually in the context of a whole sentence), or multiple-choice questions. Some are an integral part of a structured language class. There exist web-based courses, with exercises linked to pages that explain the structures. But the majority of the exercises supple an off-line course, and may be linked directly to specific textbooks.

3. Quizzes, games

There is a large variety of quizzes and games developed to help students learn languages in an interesting way. Games, such as 'Trivial Pursuit' from Gessler publishers ⁷ in foreign language versions, provide an entertaining environment for students to learn about foreign countries culture and the target language through problem solving and competition.

4. Simulation programs

Simulation programs present students with real-life situations in which they learn about the culture of a country and the protocol for various situations. For example, in 'Quick Start French' ⁸, Spoken and written words are associated with simulations, allowing students to improve listening, reading, spelling, and speaking. Encyclopedia-type programs are information programs that allow students to conduct research in the target foreign language.

5. Virtual classrooms

These tend to be fee-paying courses, often offering free trial materials open to anyone. They employ considerable staff and offer other extensive services. Such as Rosetta Stone ⁹ which proclaims itself as the World's No.1 Language Learning Software.

6. Virtual Connections

Through Internet, computers can communicate across thousands of miles. Communication abroad allows for direct interaction with native speakers. Such virtual chats provide solid opportunities for authentic language use among native and non-native speakers on an unprecedented scale in terms

⁷'Trivial Pursuit' is available at: <http://www.trivialpursuit.com/>

⁸'Quick start French' is available at: <http://www.selectsoft.com/repository/LQIMMFRENJ/view>

⁹Homepage: <http://www.rosettastone.com/>

of the numbers of users and the geographical distances involved. Discussion groups, such as BBC learning English world discussion group, are becoming very popular and progressed tremendously.

2.4 Advantages and Disadvantages

Advantages

The range of uses to which computers have been put in service of language teaching and learning is remarkable. Students can do repeating exercises for grammar and vocabulary with simple programs; they can do interactive exercises of all types on desktop machines or on the internet; they can record their voices, get feedback of their pronunciation and compare it with that of a native speaker; they can consult digital dictionaries and look up nearly anything. Web-based CALL is naturally suited to distance and flexible learning. It allows students to send and receive those assignments to and from instructors in faraway places; it provides excellent examples of ways to motivate students and keep them interested in their work; it lets individual practitioners combine different approaches to their own suiting.

Disadvantages

However, as with all IT solution, CALL is not without problems. In the case of web-based CALL, access to the web is still often unreliable and slow. There are too many distractions to study. Properly trained staff, which do not only understand the content, but also be highly trained in the use of the computer and the Internet, must also be hired to work with students on-line.

But CALL is constantly undergoing change because of technological advances that create opportunities to conduct new research and to challenge established beliefs. The goal remains to use the web for meaningful, realistic activities, to rethink the teaching approach, and to exploit the various communication resources available in the most motivating way possible.

2.5 Two Selected CALL systems

In this section, two selected CALL systems for pronunciation education will be introduced. One with focus on pronunciation correction and the other with focus on pronunciation evaluation. Both of them inspired us to propose and develop our Learner Level Adaptive Pronunciation Education System.

2.5.1 A Chinese Learning System

NHK developed a Chinese Language Learning System with visualization and speech correction for prosody, and used it in its television language study program [1].

The system construction is shown in Figure 2.3. It incorporates a speech database, when Japanese learners record (input) their speech, the system will find tone discriminations between model speech and user speech and display the difference. And, in order to help learners to improve their tone, user speech is converted with corrective tone as audio feedback.

Model Speech

The model voice of Chinese native was collected beforehand and was aggregated into the database including the position of pitch extraction contours, the Chinese characters and the PINYIN information.

First, by using an autocorrelation function and the data of power and zero crossing, judgment of voiced/unvoiced/silent was made and pitch extraction is done automatically. The extracted pitch contour is shown in the middle part of Figure 2.4.

In Figure 2.4, it is easily visible that the beginning and the end part fall into disorder and there are small irregular changes in the voiced part. This is because sometimes either the vocal chord vibrates irregularly or the vibration frequency changes in the connecting part between consonants and vowels. Because they are not distinguishable to a listener, the system is designed to omit these 'mistakes' and only display a smoothed pitch contour. The cutoff frequency of LPF used for smoothing is 8HZ. In addition, to make the tone of each character easy to understand, even in case of continued voice and consecutive pitch contour, the system

omits the boundary parts between different tones. The result of pitch contours finally obtained is shown in bottom part of Figure 2.4.

User Speech

User speech is processed with the following manner:

1. Judgment of voiced/unvoiced/silent data and pitch extraction by the same way of Model Speech.
2. DP matching between the model speech and the user speech, then the part where the same sound is uttered is associated. But because users occasionally insert wrong pauses in the places where model speech does not, or speak without necessary pauses, which will cause wrong result of DP Matching. To avoid this, all of the silent parts of both the model speech and the user speech are omitted beforehand, and only the voiced/unvoiced parts are placed in a line and DP Matching is used on them. After the voiced and unvoiced parts are with respect to each other, the omitted silent parts are put

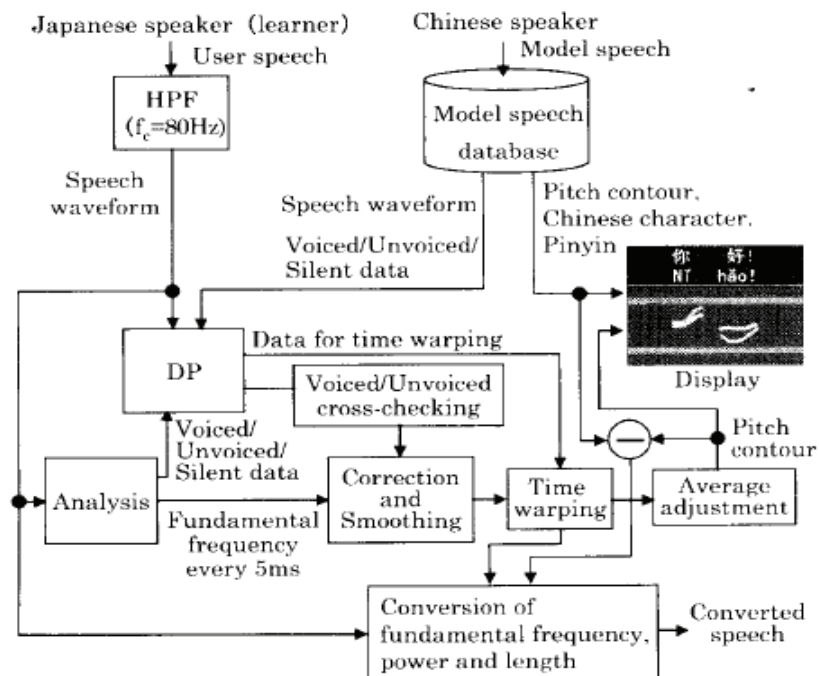
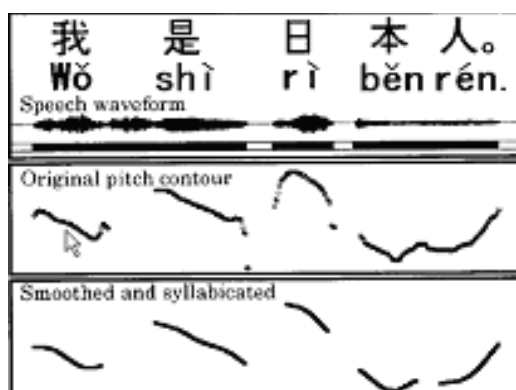
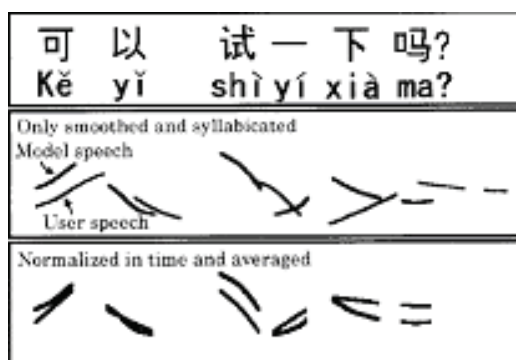


Figure 2.3: NHK System Structure

Figure 2.4: *Smoothing of pitch contour*

back into places at former time positions.

3. The same smoothing and cutting of the boundary parts of each character as model speech. The pitch contour obtained by now is shown in the middle part of Figure 2.5. But it seems hard to understand which part correspond with model speech.

Figure 2.5: *Normalization in Time and Averaging of Pitch*

4. In this case, time axis of the pitch contour of the user voice is to be expanded or contracted according to the model voice based on the result of step 2. The mean value of pitch in all voiced parts of the user speech is made to match that of the model speech.

5. Not only the time direction, but also the direction of height is normalized, and it becomes easy to compare the tone of each character between model speech and user speech, as shown in the bottom part of Figure 2.5.

Converted Speech

To reach efficient training, the user will be given a corrective audio feedback in his own voice converted by the system. The user's speech is modified in its fundamental frequency (F0), power, duration and pause. The procedure of F0 modification is shown in Figure 2.6.

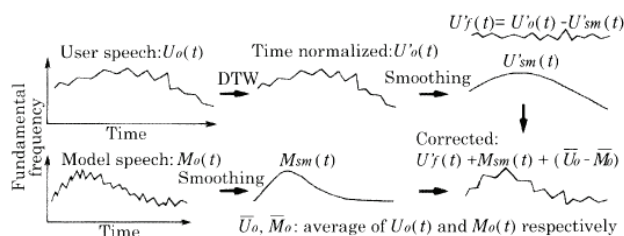


Figure 2.6: Procedure for converting user's F0

To explain simply, the first step of the procedure is replacing the user's smoothed pitch contour with model's smoothed pitch contours and preserving the user's average height of pitch and the detail changes in frequency. Power, duration and pause are converted into the same as those of the model speech. As a result, preserving the average height of pitch guarantees the characteristics of individual users' voices, and maintaining the detailed changes in frequency prevents quality deterioration. Then, the converted speech with correct tone is given back to user.

2.5.2 PSC test Evaluation System

The Chinese Proficiency Test, abbreviated as PSC, is the People's Republic of China's only standardized test of Standard Mandarin Chinese proficiency for non-native speakers, namely foreign students, overseas Chinese, and members of ethnic minority groups in China. It is also known as the "Hanyu Shuiping Kaoshi (abbreviated as HSK)" and the "Chinese TOEFL".

Currently evaluation of the PSC test is conducted entirely by human testers, which leads to subjectivity. Furthermore, large-scale testing is practically impossible due to the low efficiency and high expenses. To resolve this urgent need, the PSC test Evaluation System is developed [2].

PSC test Evaluation System is a kind of CALL system, which mainly works on pronunciation evaluation. Many researchers have studied the CALL system's pronunciation evaluation, including members of the SRI speech group [3], who mainly focused on evaluating the overall pronunciation quality of learners. They take word posterior probability, timing and duration scores as methods of evaluation; the joint research by the speech group of Cambridge University and the AI lab of MIT [4] focused on pronunciation error detection and phone-level pronunciation evaluation.

Although there are various pronunciation scoring methods, PSC is a little different because other CALL systems only need an approximate evaluation of the learners with rough scores, while the PSC test system needs to evaluate the speakers precisely.

Structure of System

PSC test system is made up of 3 main modules, which are the pre-processing module, the evaluation module and the mapping module. The overall structure is shown in Figure 2.7.

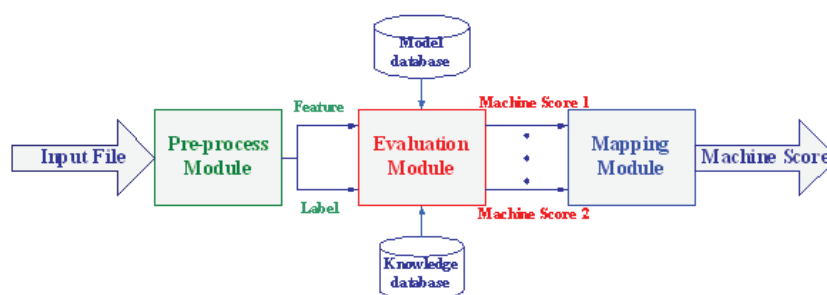


Figure 2.7: Structure of PSC System

The pre-processing module is used to process the input WAV file and text file in order to get the feature and label files for the evaluation module. The evaluation module completes the machine evaluation and generates the machine scores. The

mapping module maps the various machine scores to obtain the ultimate evaluation score.

The model database includes a set of 4 Gaussian mono-phoneme HMM models, which are used to generate coarse segmentation and select adaptation data for the evaluation module. The model database also includes a set of 16 Gaussian mono-phoneme HMM models, which are used to carry out the evaluation after adaptation. The knowledge database includes typical error patterns of the different dialect accents of Chinese. The outputs of the evaluation module are the various machine scores, which are used as inputs to the mapping module. In mapping module, linear regression is used as the mapping method to get the final score.

Pronunciation Evaluation Algorithm

A universal language-independent evaluation algorithm is based on ASR. The features of the HMM-based ASR are 13-dimensional MFCCs¹⁰ and its first and second order derivatives. Optimized algorithm is used to calculate the output probability $P(O|T)$ of given text T to the observation vector O , which is shown as follows:

$$\begin{aligned}
 P(T|O) &= \sum_{i=1}^N (|\log(P(T_i|O^{(T_i)}))|/NF(T_i))/N \\
 &= \sum_{i=1}^N (|\log(\frac{P(O^{(T_i)}|T_i)P(T_i)}{\sum_{q \in Q_{error}^{T_i}} P(O^{(T_i)}|q)P(q)})/NF(T_i))/N \\
 &= \sum_{i=1}^N (|\log(\frac{P(O^{(T_i)}|T_i)}{\max_{q \in Q_{error}^{T_i}} P(O^{(T_i)}|q)})/NF(T_i))/N
 \end{aligned}$$

Q is the model set, q is the phoneme that T_i could be misread as. $NF(T_i)$ is the total frame of phoneme T_i . $P(O^{(T_i)}|T_i)$ is the likelihood of given HMM model T_i to the observation vector $O^{(T_i)}$.

Tone Evaluation

¹⁰Mel-Frequency Cepstral Coefficients

Tone evaluation is a critical part of PSC test system. Tones in Chinese can be represented by F0 contour. The identifying property of the F0 contours for tone is dramatically lessened by variations across speakers and manners of pronunciation. Normalization is thus necessary for compensating these variations. Recognizing tone via F0 contour is based on the assumption that a tone can be represented by its F0 contour consistently, which in turn means that different speaker's F0 distribution should be somewhat similar.

From Figure 2.8 it can be found that difference between the two sets of F0 distributions is so large that the models could easily fail if F0 is directly used as a tone classification feature without any normalization.

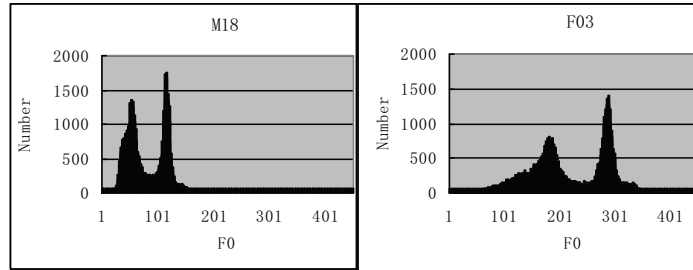


Figure 2.8: *F0 distribution of different speakers*

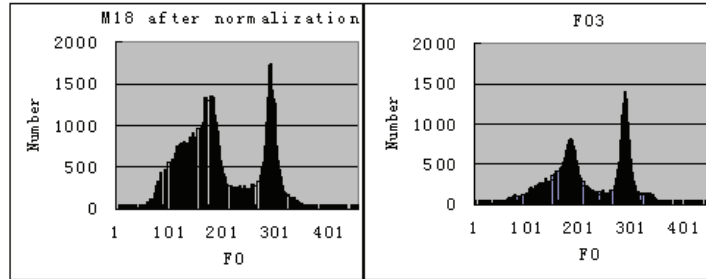
Here using CDF-Matching for F0 normalization, one standard person's (F03) F0 distribution is selected as the objective CDF, and all other persons' F0 distributions are mapped to speaker F03's F0 distribution via CDF-matching. The CDF-matching is implemented using histogram equalization.

Figure 2.9 displays the normalization result of CDF-matching. In contrast to Figure 2.8, it can be seen that the F0 distributions of different speakers become more similar after CDF-matching.

Posterior probability is used for tone evaluation as well. The posterior probability of tone is calculated as follows:

$$P(T|O) = \frac{P(O|T)P(T)}{\sum_{T' \in T_{set}} P(O|T')P(T')} \quad (2.1)$$

O is the F0 contour of one syllable. T is the tone label of text. T_{set} is the set

Figure 2.9: *F0 distribution after normalization*

of tone models.

Next is to use calculated tone posterior probability to evaluate tone. After posterior probability is derived, tone error detection is done as in equation:

$$\begin{cases} \text{Right} & \text{if } P(T|O) \geq \text{Thresh}_T \\ \text{Error} & \text{if } P(T|O) < \text{Thresh}_T \end{cases}$$

Thresh_T is the tone error detection threshold for tone T . The equation indicates that the pronounced tone T is an error if its posterior probability is less than the threshold. T is judged as accurate, if it matches or exceeds the threshold.

Pronunciation Error Detection

Pronunciation error detection is essential for pronunciation evaluation systems because it can give users information about the kinds of errors made as well as give them the corrective advice.

Error detection is also based on posterior probability. A segment is marked as an error if its posterior probability is below a predefined threshold. The process is shown in Figure 2.10.

Because the posterior probability varies from phone to phone, the threshold should be phone-specific. This is implemented via Equation as follows:

$$T_p = \mu_p + \alpha\sigma_p + \beta \quad (2.2)$$

where T_p is a phone-specific threshold, μ_p and σ_p are the mean and variance

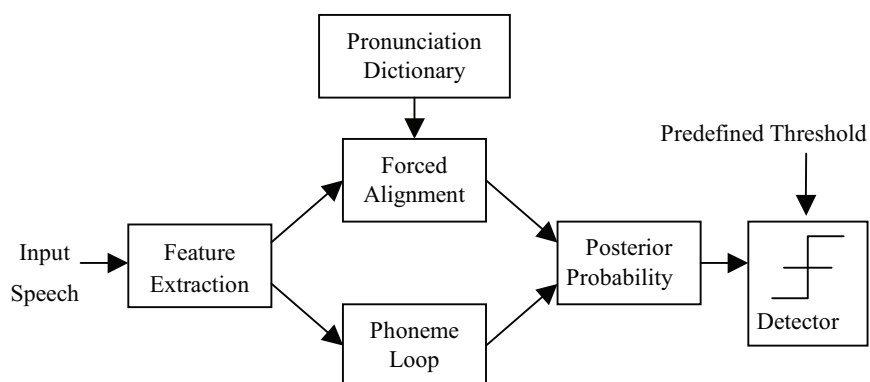


Figure 2.10: *Pronunciation error detection process*

of the posterior probability for phone p in the standard training database, and α and β are empirically-determined scaling constants.

Chapter 3

Mandarin Introduction

Nowadays, Chinese has become the number one spoken language in the world, and Chinese is one of the six official languages in United Nations. Speakers of Chinese live not only in China, Taiwan, and Singapore, but spread throughout the world.

With the rapid recent emergence of China as a major player in the global economy, people get more and more interested in learning the Chinese language in preparations for stronger business link and greater prospects in China's emerging market.

Chinese language is quite different from many western languages in various features. Chinese language is not alphabetic. All the characters are ideographic symbols and there are more than 10000 commonly used Chinese characters [5] in China. Almost every character is a morpheme with its own meaning and can play some linguistic role independently. Meanwhile, one or several characters can also compose a word with its meaning sometimes completely different from the meaning of the component characters. Then several words can compose a sentence, sometimes in very flexible ordering. These characteristics make Chinese considered quite special and challenging.

3.1 Phonemes

The characters of Chinese language are pronounced as monosyllables which are usually consisted of one consonant or nasal in the front and one vowel at the

end. So by traditional Chinese phonology, the syllables are divided into two kinds of phoneme sets. The front consonant or nasal is called the initial while the ending vowel is called the final. To Mandarin, it has a total of 22 initials (including null initials) and 38 finals (including null finals) [6] which compose the whole 408 Mandarin syllable set.

Romanization is the process of transcribing a language in the Latin alphabet. There are many systems of romanization for the Chinese languages due to the Chinese's own lack of phonetic transcription until modern times. Chinese is first known to have been written in Latin characters by Western Christian missionaries in the 16th century. Today the most common romanization standard for Standard Mandarin is Hanyu Pinyin, often known simply as pinyin. Pinyin is almost universally employed now for teaching standard spoken Chinese in schools and universities. Table 3.1 shows "Mandarin Chinese Pinyin Table". By combining all those phonemes and tones listed, there are a total number of 1302 different tonal syllables in Mandarin.

<i>Initial consonants</i>						
b	p	m	f	d	t	n
l	g	k	h	j	q	x
z	c	s	zh	ch	sh	r
<i>Final vowels</i>						
a	i	u	e	o	v	ao
ai	an	ang	ou	ong	ei	en
er	eng	ia	iao	ie	iu	in
ian	ing	iang	iong	ua	uo	ui
uai	un	uan	uang	ve	vn	van
<i>Tones</i>						
1	2	3	4	0		

Table 3.1: *Unites of Chinese syllables*

3.2 Tones

Another characteristic of Chinese spoken language is the tonal nature. There are four lexical tones and one neutral tone in Chinese language. The same syllables with different tones have different lexical meanings. Official modern Man-

darin has only 400 spoken monosyllables but over 10,000 written characters, so there are many homophones only distinguishable by the five tones. Even this is often not enough unless the context and exact phrase is identified.

A very common example used to illustrate the use of tones in Chinese are the four main tones of Standard Mandarin applied to the syllable "ma". The tones correspond to these five characters:

媽 (ma + tone1)	"mother"	high level tone
麻 (ma + tone2)	"hemp" or "torpid"	high rising tone
馬 (ma + tone3)	"horse"	low falling-rising tone
罵 (ma + tone4)	"scold"	high falling tone
嘛 (ma + tone5)	"question particle"	neutral

In human languages, the most important perceptual attribute of these tones is pitch. The pitch variation used in short stretches of syllable length, such as in small grammatical units like words and morphemes, is called tone [7]. By the phonetic studies on the acoustic characteristics, pitch is the primary cue in tone perception [8] and the F0 contours are used to represent the pitch variation in the study of speech. Usually pitch range is divided into 5 levels, naming 1 to 5 and level 5 corresponding to the highest. When the tones are pronounced in isolation, the fundamental frequency (F0) contours produced are relatively stable and their shapes well described phonetically: a flat high F0 contour for the tone 1, a rising contour for tone 2, a curved falling and rising F0 shape for the Tone 3 and a sharply falling contour for tone 4 (tone 5 has no canonical shape, and is not shown). Figure 3.1 shows traditional description of pitch changes of Mandarin tones.

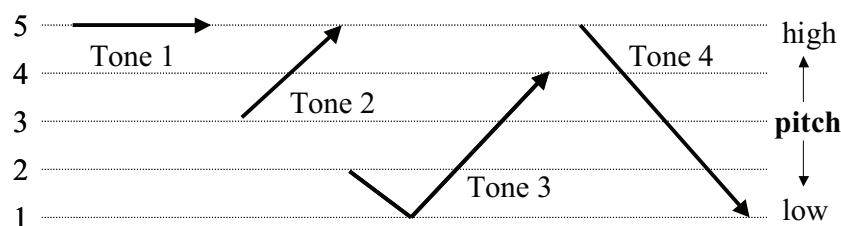


Figure 3.1: *Relative pitch changes of the four tones*

But when the tones are in a continuous sentence context, the F0 contour differs greatly with the speaker and the sentence environment, such a tonal change

process is commonly referred as tone sandhi. For example, a tone 3 is often realized with a rising F0 contour before another tone 3, which is known as the tone 3 sandhi. Furthermore, emotion or intonation of the speaker may also affect the pitch variation in long stretches of the pitch contour speech. So it isn't easy to use the relations between the tones and pitch variation. Thus far, there is no conclusive evidence that speakers are able to exploit the subtle acoustic differences in distinguishing the two tones. That is also why the study of how to use the tone information is still a big challenge.

3.3 Difficulties for Japanese

For Japanese people, since there are Kanjis in Japanese, Chinese characters are not a big problem for them.

But as a tonal language, in Chinese, F0 movements are much more complicated than non-tonal languages like Japanese and English. In Japanese, the pitch of each syllable could be either "High" or "Low". The Japanese language does not have tone, but does have pitch accent.

The role of pitch in Japanese is quite different from Chinese. In a pitch accent language, there is one accented syllable in a word, and the position of that accented syllable determines the tonal pattern of the whole word. This is unlike the situation in tonal languages, where the tones of each syllable can be independent of the other syllables in the word.

In the case of Japanese, there are 2 patterns of accents, for example, hashi(HL) and hashi(LH), which mean chopsticks and bridges. The pitch accent can make the difference between different words.

$$e.g. \begin{cases} \text{"chopsticks"}, [hashi], \text{High-Low} \\ \text{"bridge"}, [hashi], \text{Low-High} \end{cases}$$

Although there are also pitch controls in Japanese pronunciation, the pitch movements of Chinese are far more dynamic. And not only the approach of changing pitch, but also the degree of rising and dropping, the changing speed and the power controlling is difficult to learn for Japanese people.

In Chinese, there are 38 vowels, 22 consonants, while in Japanese there are 5 vowels, 14 consonants. The gap of vowels and consonants' quantity implies the pronunciation difficulties too. Furthermore, there are double-vowels and tri-vowels in Chinese, such as 'ou', 'uang', while no similar ones in Japanese. In Japanese, /r/ and /l/ hardly distinguishes, while in Chinese, we raise tongue to pronounce /r/. Native speakers of Japanese have difficulty pronouncing Chinese consonant clusters because the syllabic structure of Japanese is more restricted than that of Chinese. Japanese has a maximum of two syllable-initial consonants and one syllable-final consonant, while Chinese allows more consonant combinations. Inserting vowels within consonants clusters or after syllable-final consonants is a prevalent error in Chinese spoken by Japanese. With these problems in mind, our system can automatically detect inserted vowels and instruct learners how to pronounce the target utterance correctly.

In next chapter, we will introduce our speech analysis results of Japanese people's Chinese utterances.

Chapter 4

Learners' Speech Analysis

4.1 Speech Data

In order to get an idea of the feature characters of Chinese pronunciation uttered by Japanese people, we designed a reading script (see Appendix), which includes subsets of 12 Japanese sentences, 24 Chinese words and 28 Chinese sentences.

We went to several Chinese language classes and collected Mandarin speech data from a total of 12 native Japanese speakers and a total of 6 native Chinese speakers.

Learner	Male	6
	Female	6
Native	Male	4
	Female	2

Table 4.1: *Speech Data Collected*

All the words and sentences are labeled with PINYIN (with phonetic and tonal information). In order to let the participants make better understanding of the sentence-structure and word meaning, all the Chinese are labeled with Japanese translations. It takes about 45 minutes for each one to read on average. Most of the recordings were done in quiet classrooms or libraries (no soundproof) using a microphone (Sony Electret condenser microphone ECM-360).

All the participants filled in questionnaires to tell their mother language, their

birthplace, their years of learning Chinese/Japanese, their test levels and so on. We defined who has studied Chinese for less than 1 year as beginners and those who has studied more than 1 year as advanced learners. Among the 12 Japanese participants, 7 of them are beginners and the other 5 are advanced learners.

4.2 Method

A native Chinese trained in phonetic labeling labeled all of the utterances at the phone level.

We used the toolkit Praat [9] to analyze the speech wave, the pitches, the power, the pauses, etc. We cut students' utterances into syllables and asked natives to listen to these syllables and write down what did they hear. Some statistical analysis is also done to find out how duration influence non-natives' utterance. Several subjective listening tests were held. 4 native speakers gave their evaluation of the differences between native's speech and learners' speech.

For an incorrect pronunciation of a word, there may be two aspects of reasons: one is the phonetic aspect and the other is the prosodic aspect. From the phonetic aspects, we can evaluate pronunciations of vowels and consonants; while from prosodic aspects we can evaluate tones by watching pitch contours. And there are also other parameters like duration and power and so on. We put our analysis emphases on the aspects shown as follows:

1. The reading script covered all the vowels and consonants of Chinese. And some difficult consonants and vowels appeared both in sentences and words. We would like to check the performance of a same syllable in word level and sentence level. Also, when developing CALL system, we need to give a dictionary of all the possible error pronunciations of learners. For example, for the syllable 'sh+i', possible errors may be 's+i', 'sh+e', and 's+e'. Thus, we need to conclude a list of easy-to-make-mistake vowels and consonants and all possible error candidates through our speech analysis.
2. Duration is also one of the main analysis points. In the recording, participants are made to read a same sentence several times when with/without supply of sentence segmental information. We chose words that begin with voiced sound that make it easier to make accurate labeling and subjective

listening tests are to be held to analyze how is the influence of duration.

3. To analyze the prosodic features, we found many words that are made up of same phonemes but with different tones. For example, "知識 (zhi1 shi5)" and "指示 (zhi3 shi4)", "聯系 (lian2 xi4)" and "練習 (lian4 xi2)". We intend to find out what kind of combination of tones is difficult and how do participants perform same tones at word level and sentence level.
4. In Chinese, there are many words that have similar pronounce to Japanese. Such as "開始 (kai1 shi3/カイシ)", "管理 (guan3 li3/カンリ)". Although the pronunciation may be similar, they are not the same. In order to check whether Japanese learners are leaving the influence of their mother language or not, we are going to compare their Japanese and Chinese pronunciation.

4.3 Experimental Results

The features of Japanese learners' utterances we found are shown as follows:

1. About vowels

Vowels came out to be a big problem. It seems difficult for Japanese learners to expand their five Japanese vowels to 37 Chinese vowels. Learners perform vowels unstable. the same vowel uttered differently when pronounced in different words. Furthermore, lip movements for Japanese utterances are generally smaller than those for Chinese. This makes, for instance, pronunciation of vowel 'a' in Japanese and Chinese is totally different, but Japanese students tend to use Japanese version of 'a'. Also, when saying double-vowels and tri-vowels. The duration of learners is obviously longer than natives'. And as for Chinese people, double-vowels and tri-vowels, such as 'ou', 'uang', 'ao', and 'uo', is only one syllable that cannot be pronounced separately. But when listening to learners data, we heard that vowels are divided into several syllables.

From natives' listening test results, and combining with some open analysis data of Chinese Proficiency Test (HSK), we concluded Table 4.2. In the table, all Mandarin vowels are listed out, and for each vowel, a list of possible mispronunciation vowels by Japanese is followed. For example, for the vowel 'a', possible errors may be 'e', 'ai', 'ao', 'ua', 'ia'. This table will be used in our CALL system to help HTK to build N-best networks.

a	e; ai; ao; ua; ia;	o	u; ou;
e	a; o; u; ei;ie;	er	e; o;
ai	ei; uai; ao;	ei	ai; ui; ve; ie;
ao	ai; uai; ou; o;	ou	uo; uao; o;
an	ang; en; eng; uan;	en	eng; an; ang; un;
ang	an; eng; en; uang;	eng	en; ang; an;
ong	un; ang; eng; ung;	i	v; ue; ie;
ia	ie; i; iu; a;	ie	ia; iu; e;
iao	ie; ia; ian; ao;	iu	ie; iao; ou;
ian	iang; ia; ie; ien;	in	ing; iong; ian;
iang	ian; ing; iong;	ing	in; eng; en;
iong	un; ing; ian;	u	e; i; o; uo;
ua	a; uai; ao;	uo	u; ou; ao; o;
uai	ui; uai; ai;	ui	ei; uai; ai;
uan	un; an; uang;	un	uan; uan; an;
uang	uan; ueng; ang; an;	ueng	eng; uang; un;
v	ue; iu;	ve	ie; u; ue;
ue	ie; u; ve;	uan	uang; un; ang;
un	uan; eng; ong;		

Table 4.2: Possible Vowel Mispronunciation List

2. About consonants

Several consonants are especially be easy to make mistakes. For example, 'r', which is a raised tongue consonant in Chinese, is often pronounced with the tongue being raised too much or not enough by Japanese students. 'b/p', 'd/t', 'g/k', and 'f/h' are not distinguished clearly.

We concluded Table 4.3 in the same way as vowels. In the table, all Mandarin consonants are listed out, and for each consonant, a list of possible mispronunciation consonants by Japanese is followed.

We also held a evaluation experiment that estimate phonemes pronunciation. We randomly chose utterances of 48 characters from every learners' speech, and 1 native speaker gave her judgment of the corrective percentage of phoneme pronunciation. The result is shown in Table 4.4.

b	p; m; f;	j	q; x; zh; z;
p	b; h; m;	q	j; x; c; ch;
m	b; p; f;	x	j; q; sh; s;
f	h; p; b; m;	zh	z; ch; sh; j;
d	t; n; l;	ch	c; zh; sh; q;
t	d; n; l;	sh	s; ch; zh; x;
n	l; d; t;	r	ch; zh; sh; r;
l	r; n; d; t;	z	zh; c; s; j;
g	k; h; j;	c	ch; s; z; q;
k	g; h; q;	s	z; c; sh; x;
h	f; k; g; q;	y	w; r;

Table 4.3: Possible Consonant Mispronunciation List

	Corrective Percentage
Beginner	67.92%
Advanced learner	84.17%

Table 4.4: Evaluation of Phonemes Pronunciation

3. About tones

Japanese people have the concept of high pitch and low pitch, but it is still hard for them to manage tones. Changing pitch too slowly (not in one syllable), or not enough changing range (when rising or dropping) is the reason why Tone 2, 3 and 4 can not be done very well. For example, according to our data, when saying tone 2 and tone 4 in sentences, natives change pitch about 100 Hz, while learners change about 50 to 60 Hz. Also, we found that the Chinese speakers produced higher average pitch, larger pitch range than the Japanese speakers both on phoneme and utterance level.

We randomly chose 24 word-utterances and another 24 word-utterances cut from sentences of every learner. 1 native speaker gave her judgment of the corrective percentage of tone pronunciation. The result is shown in Table 4.5. From the table, we can see that advanced learners pronounce word-level tone better than sentence-level tone.

When focusing on the error situations, our test results indicated that most errors made by learners of various levels happened between tone 2 and tone 3. The

	Word-level Tone	Sentence-level Tone
Beginner	68.06%	69.43%
Advanced learner	89.17%	78.33%

Table 4.5: *Evaluation of Tones Pronunciation*

error rate is higher than that of tone 1 and tone 4. Typical examples with tone mispronunciations are plotted in Figure 4.1.













	tone1	tone2	tone3	tone4
Correct contour				
Incorrect contour				
				

Figure 4.1: *Correct and Incorrect Pitch contours*

- The first tone (55)¹: it is sometimes pronounced too short or not flat as shown in Figure 4.1. Another representative mistake is that pitch values are not high enough, such as 44 or even lower value.
- The second tone (35): usually not enough rising of pitch in learners' data. Also, it is easily confused with the third tone when the descending part of the contour lasts long enough to be perceived.
- The third tone (214): a rising trend is required at the end of the pitch contour, however some learners ignore such requirement and pronounce as 21. Some learners pronounce it as 35 that becomes the second tone. Another mistake is beginning value is so high to be comparable with or even higher than the ending's, such as 313 or 413 and so on. The performance of tone 3 is especially easy to influenced by the tones before and after.
- The fourth tone (51): the beginning is sometimes pronounced not high enough. Learners whose pitch changing ranges are not wide enough are more possible to make such mistakes.

¹As explained in Figure 3.1

Another point is that, to native Chinese, tone and phoneme fuse together. But for Japanese learners, since there is no similarity in their mother tongue, it is extremely hard for them to do tonal change and phonetic change in one syllable. This is showed as when the tone is pronounced right but phonemes are not done clearly or when the phoneme is clearly said and the tone is right but the duration is so long that one tri-vowel is divided into several syllables. These all make reading data sounded unnatural.

Our reading script was prepared so that all tone combinations were included. Through a listening test focused on tone pronunciation, the combination of [tone3 + tone2] was found especially difficult for Japanese students. In addition, they failed to pronounce tone3+tone3 sequence as tone2 and tone3. Utterances with this type of error are hard to be understood by native.

We also compared pronunciations of the same word in word level and in sentence level. Beginners pronounce neither word level nor sentence level good. But it was obvious that, for advanced learners, the performance in word level was better than that in sentence level.

4. About duration

We randomly selected 2 learners and 1 native, and 7 sentences of these 3 people. We did some statistical analysis and tried to find out how duration influence non-natives' utterances and what is the main difference of duration between non-natives and natives.

Character Duration Analysis

Table 4.6 is the comparison of character durations between non-native and native. It shows that, on average, non-native speaks characters slower. The average time length of each character is about 1.31 times as long as the time length of native's. Also, the standard deviation of non-native is bigger which implies that non-native changes their speech speed more when reading different sentences.

Short Pauses Analysis

SP means short pause between phrases. Table 4.7 summarizes the durations of sentences with/without short pause. When getting rid of short pauses, the average

	non-native	native
average number of characters per sentence	14.71	
standard deviation	3.62	
average duration of per character	0.305	0.232
standard deviation	0.038	0.018
average number of characters per second	3.33	4.322
standard deviation	0.397	0.335

Table 4.6: *Comparison of Character Duration between Non-native and Native*

sentence durations of non-native are about 1.3 times as fast as the native's. While in the case of concluding short pauses, the ratio rose to 1.34, which implies that, the short pause that non-native takes between phrases is too long.

with SP	MEAN	SD	without SP	MEAN	SD
non-native	4.859	1.159	non-native	4.436	1.185
native	3.638	0.789	native	3.374	0.674
RATIO	1.34	1.47	RATIO	1.31	1.76

Table 4.7: *Arithmetic mean and SD of sentence durations with/without SP*

Vowels and Consonants Analysis

In Table 4.8, there shows the percentage of vowel durations and consonant durations in sentences. In native speech, the ration of durations between vowels and consonants is about 61:39; while in non-native speech, the ration is about 55:45. As we know there are 37 vowels in Chinese, so perhaps in order to speak clearly and not to confuse these vowels, Chinese people are used to spend more time on vowels.

Moreover, from Table 4.7 we have noticed that the arithmetic mean of duration of non-native is about 1.3 times as fast as the native's. Thus, in Table 4.8, we have learned the further detail that when speaking vowels, non-native is only 1.2 times as fast as native, while speaking consonants, non-native is 1.55 times faster! According to this data, we can conclude that in Japanese-Chinese, it is the short duration of vowels and long duration of consonances that makes the duration gap between non-native and native.

Vowels	MEAN	PERCENTAGE
non-native	2.548	54.67%
native	2.117	61.44%
RATIO	1.20	

Consonants	MEAN	PERCENTAGE
non-native	2.011	45.33%
native	1.301	38.56%
RATIO	1.55	

Table 4.8: *Arithmetic mean and SD of sentence durations with/without SP*

In Table 4.9, it concludes the average duration of consonants and vowels. For non-native, the ratio of the time length of a single consonant and vowel is 1:1.27, and for native, the ratio is 1:1.62. This also proves that native pronounce longer vowels and shorter consonants.

		MEAN	SD	RATIO
Consonants	non-native	0.139	0.017	1.55
	native	0.091	0.011	
Vowels	non-native	0.176	0.020	1.20
	native	0.147	0.013	

Table 4.9: *Average Durations of per Vowel and Consonant*

Vowels and Double-vowels Analysis

In Table 4.10, it concludes the average duration of vowels and double-vowels. For non-native, the ratio of the time length of per single vowel and double-vowel is 1 to 1.11, and for native, the ratio is 1 to 1.09. Therefore, 1) double-vowels cost more time than vowels to pronounce; 2) Non-native spends more time to pronounce double-vowel than native; 3) According to SD, the duration of native is more stable.

Nouns and Verbs Analysis

Table 4.11 is the comparison of durations between nouns and verbs. In Chi-

		MEAN	SD	RATIO
Vowels	non-native	0.176	0.030	1.20
	native	0.147	0.013	
Double-vowels	non-native	0.195	0.029	1.21
	native	0.161	0.018	

Table 4.10: *Average durations of per vowel and double-vowel*

nese, verbs and nouns generally consist of one, two, or more characters. In Table 4.11, first, there gives average durations of per character in nouns and verbs. The value of mean implies that both non-native and native pronounce nouns shortly and pronounce verbs longer. But the ratio indicates that the duration gap between nouns and verbs is greater in native speech.

Table 4.12 is the comparison of consonant and vowel durations between nouns and verbs. What causes the duration gap between nouns and verbs? To solve this question, We separated nouns and verbs into vowels and consonants, and calculated the durations again. Now from Table 4.12 it is clearly that both in non-native speech and native speech, consonants are shortened in nouns and lengthened in verbs. However, in the case of vowels, non-native still shorten the vowels in nouns and lengthen the vowels in verbs, while native do the opposite.

Nouns & Verbs		MEAN	RATIO
non-native	noun	0.140	0.826
	verb	0.170	
native	noun	0.129	0.952
	verb	0.135	

Table 4.11: *Comparison of Durations between Nouns and Verbs*

5. About utterance rhythm

Natives speak long sentences at a stable speed, while learners tend to speak some parts very fast and some parts slower.

Vowels		MEAN	RATIO
non-native	noun	0.162	0.911
	verb	0.177	
	MEAN	0.169	-
native	noun	0.171	1.122
	verb	0.152	
	MEAN	0.161	-

Consonants		MEAN	RATIO
non-native	noun	0.119	0.733
	verb	0.162	
	MEAN	0.140	-
native	noun	0.086	0.732
	verb	0.118	
	MEAN	0.102	-

Table 4.12: Comparison of Consonant and Vowel Durations between Nouns and Verbs

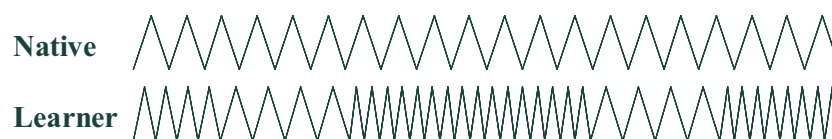


Figure 4.2: Difference of Utterance Rhythms

4.4 Speech Modification Experiment

From our speech analysis results, we became anxious about how pitch and duration influence learners' utterance. Thus, we arranged a speech modification experiment.

We wrote Perl program to replace learners' pitch contours by native's contours and modified learners' duration according to native's. Then, 4 native speakers did subjective listening test to evaluate the naturalness of the new generated learners' speech. We randomly chose 2 learners, and 7 sentence utterances of each one. And chose corresponding native utterances of the same 7 sentences.

Duration Modification

Our general approach to modify duration of learners' speech is:

- First, doing manual segmentation. We marked vowels and consonants and short pauses of learners' speech and native's speech manually (See Figure 4.3). Both learners' and native's results of segmentation is saved as TXT files.

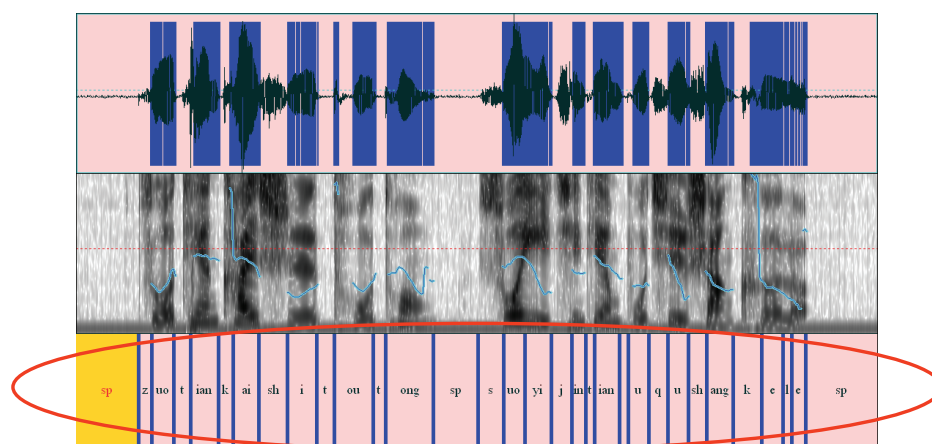


Figure 4.3: *Vowel, Consonant, Short Pause Segmentation*

- Second, run our perl program to change learners' segmentation TXT file according to native's one, see Figure 4.4. Then, duration of learners' speech is changed exactly the same with the native's.
- Third, using toolkit Praat to apply the new TXT duration file to learners' WAV file. After this step, by expanding or contracting time axis of learners' data, we can generate new WAV file of learners' speech with duration modified (See Figure 4.5).

Pitch Modification

In the case of pitch modification, first, we use the same process of duration modification. But this time, we modify native's speech according to learners' duration data. Then the pitch contours of both learners and native are extracted at

```

File type = "ooTextFile"
Object class = "TextGrid"
xmin = 0
xmax = 10
tiers? <exists>
size = 1
item []:
  item [1]:
    class = "IntervalTier"
    name = "yoshihara-s_2-01-01"
    xmin = 0
    xmax = 10
    intervals: size = 32
    intervals [1]:
      xmin = 0
      xmax = 1.2933017125765804
      text = "sp"
    intervals [2]:
      xmin = 1.2933017125765804
      xmax = 1.3352710586565624
      text = "z"
    intervals [3]:
      xmin = 1.3352710586565624
      xmax = 1.453309844506512
      text = "uo"
    intervals [4]:
      xmin = 1.453309844506512
      xmax = 1.5359369946014767
      text = "sp"
    .....

```

Figure 4.4: *Learners' Segmentation TXT File*

10ms intervals from the original speech using toolkit Praat. By replacing learners' pitch tiers by native's, new pitch modified learners' speech is generated. The whole process can be seen in Figure 4.6.

Similarly, if we modifies learners' duration according to native ones, and changes learners' pitch tiers, we can obtain new both duration and pitch modified learners' speech.

Subjective Listening Test

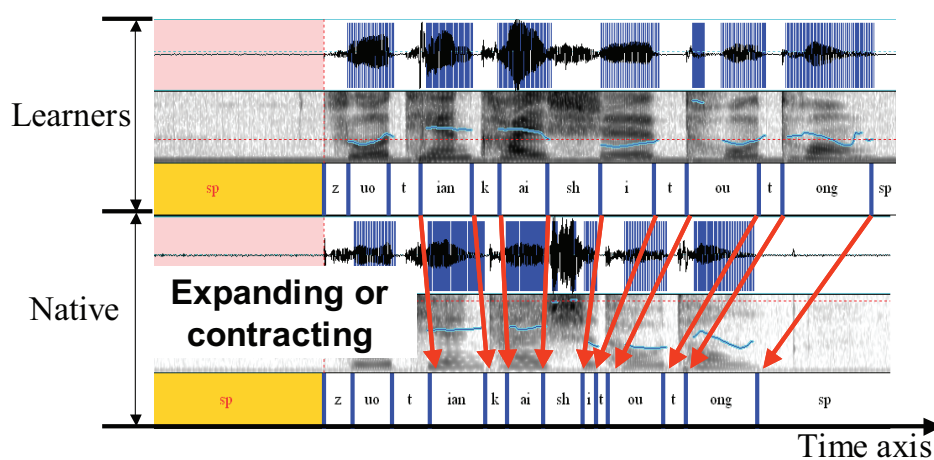


Figure 4.5: *Expanding or Contracting Time Axis of Learners' Data*

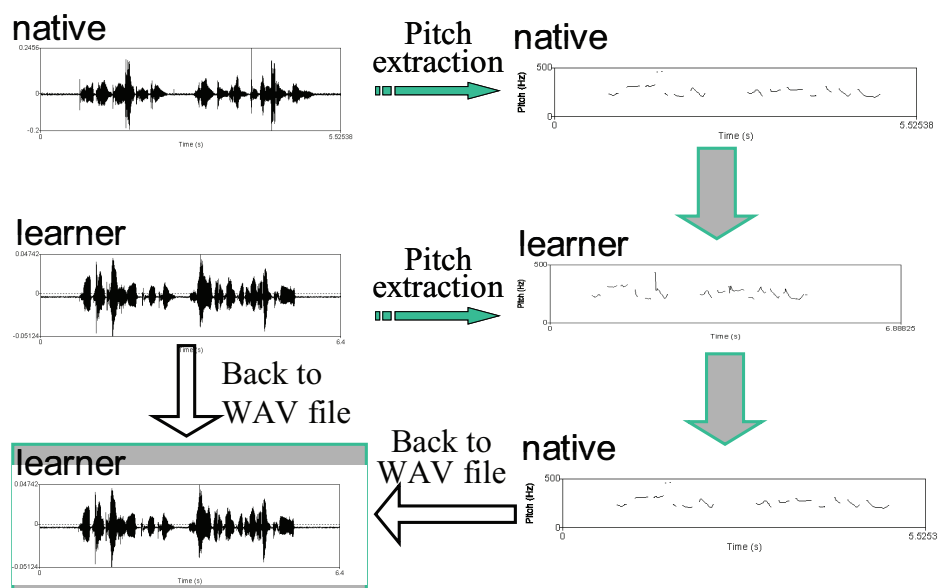


Figure 4.6: *Process of Pitch Modification*

In order to evaluate the naturalness of the generated speech, we carried out a subjective listening test. Now, for each sentences, there are four kinds of data: the original ones, duration changed ones, pitch changed ones and both duration & pitch changed ones. Four native people listened to the four kinds of WAV file and gave their evaluation scores on a five-point scale (Scoring Criteria shown in Table 4.13).

Score	Characteristics
5	Sounds natural, and can't be distinguished as a generated utterance
4	Only a few of the syllables sounds inconsistent, but doesn't spoil the whole utterance
3	Some prosodic phrase sounds unnatural, and causes damage to the utterance, a little annoying
2	Inconsistent throughout the whole utterance, sounds annoying, hard to accept
1	Unacceptable

Table 4.13: *Subjective Scoring Criteria*

Sentence No.	Original	Duration-modified	Pitch-modified	Dura.&Pitch modified
1	3.33	4	3.5	4
2	3.33	4	3.5	4.25
3	3	4	3.5	3.75
4	3.5	3.75	3.75	4
5	3.16	3.5	3.5	3.75
6	3.25	4.25	3.5	3.5
7	3.	3.75	3.5	3.5
AVERAGE	3.296	3.893	3.536	3.857

Table 4.14: *Listening Test Results of Modified Speech*

The evaluation result is shown in Table 4.14. From Figure 6.1, we can see whatever 'pitch' or 'duration' or both of them are changed, the naturalness is improved! The results indicate that generated speech are closer to native quality than the original learner speech. And now we do regard duration as an important factor to influence learners' utterance. We will add duration practice module to our CALL system. Someone may also interested in why duration modified speech even get higher score than pitch modified ones. Here we gave our simple explanation. In our idea, it is because there are some weird sounds in pitch modified speech. At first, we thought maybe this is because the automatic pitch extraction is not very accurate, so, we added the step of smoothing the pitch contour. But,

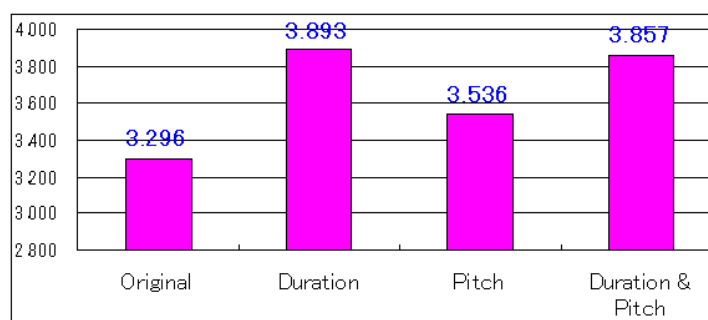


Figure 4.7: *Listening Test Results of Modified Speech*

still, weird sounds appears. We can do nothing about the speech quality deterioration.

4.5 Conclusion

In this section, we concluded many characteristics of Japanese learners' utterances. Besides, a speech modification experiment is also done and validated that duration and pitch practice is necessary and important.

From our speech analysis, we found that the problems of beginners and advanced learners are different. Beginners are usually poor at phones and tones, which make their utterances hard to be understood. On the other hand, advanced learners may manage to pronounce phones well, but when speak sentences, there are problems in tones and durations that make their utterances sounds unnatural. This finding inspired us to develop a learner level adaptive system.

Subject	Difficulties	Practice Emphases
Beginners	on phonetics	Phoneme, Word level Tone
Advanced	on prosody	Duration, Sentence level Tone

Table 4.15: *Difficulties per Level*

Chapter 5

Adaptive CALL System

5.1 System Overview

According to the speech analysis results of Japanese learners, we proposed our learner level adaptive CALL system to provide different pronunciation practice emphases to satisfy different levels of learners.

Figure 5.1 illustrates the basic architecture of the system. First, learner is asked to record 10 sentences to have level judgment. After a process to give level judgment result, learner would be recommend to a certain course, word level practice course or sentence level practice course. Then learner will be trained in either course with different emphases.

The system's form structure is shown in Figure 5.2. There are three main modules in our system: Learner Level Judgment Module, Word-level Practice Module and Sentence-level Practice Module:

- Judgment module is to judge learner's pronunciation level and recommend learner to enter proper practice course.
- Word-level Practice Module puts emphases on phoneme training and word-level tone training. Its aim is to help beginner level learner to pronounce clearly and correctly in order to make their pronunciation easy to be understood.
- Sentence-level Practice Module puts emphases on tone training and duration training on sentences level. Its aim is to help advanced learner to pronounce naturally and thus gradually reach a native level.

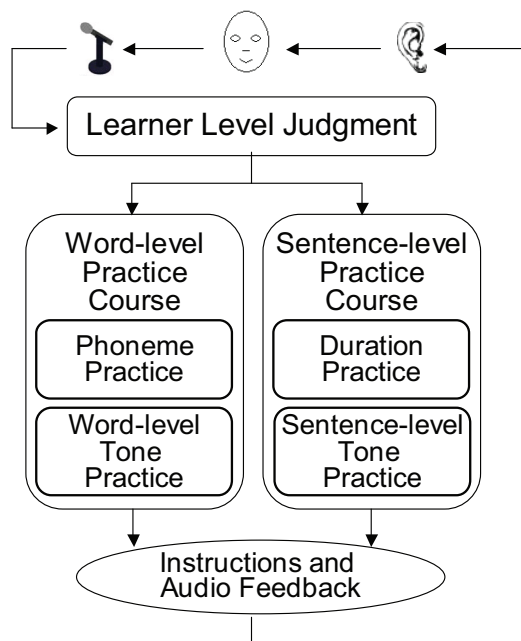


Figure 5.1: Learner Adaptive CALL System Structure

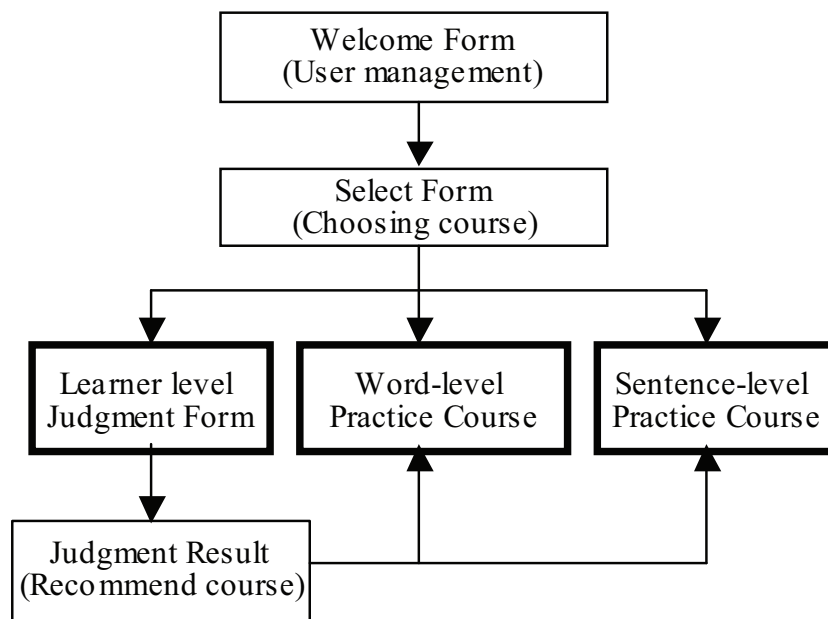


Figure 5.2: Learner Adaptive CALL System Form Structure

5.2 Level Judgment Module

Learner level judgment is one of our core modules, and it represents the main innovation of our system. In this module, our system will first ask learner to record 10 short sentences and then through using speech recognizer to analyze these recording, the system will give a judgment report of learner's pronunciation.

The 10 sentences are shown in Table 5.1. They are carefully designed that:

- a) they include all Chinese 36 vowels and 21 consonants;
- b) they include most of tone combinations;
- c) they are all short sentences with no punctuation mark (uttered without pause);
- d) they are all easy to be understood sentences.

01	星期四下午的音樂課最有趣。 xing1 qi1 si4 xia4 u3 de1 in1 ue4 ke3 zui3 ou3 qv4
02	他感覺頭痛得利害。 ta1 gan3 jue2 tou2 tong4 de1 li4 hai4
03	那是我第二次去北京旅行。 na2 shi2 wo3 di4 er4 ci4 qv4 bei3 jing1 lv3 xing2
04	原來姐姐也會游泳。 yuan2 lai2 jie3 jie1 ie3 hui4 ou2 ong3
05	弟弟每天都要被老師批評。 di4 di1 mei3 tian1 dou1 ao4 bei4 lao3 shi1 pi1 ping2
06	克服困難勇敢前進。 ke4 fu2 kun4 nan1 ong3 gan3 qian2 jin4
07	小恒是一頭快樂的熊。 xiao3 heng2 shi4 yi4 tou2 kuai4 le4 de1 xiong2
08	大家一起來創造美好的未來。 da4 jia1 yi4 qi3 lai2 chuang4 zao4 mei3 hao3 de1 wei4 lai2
09	在花的海洋里自由地飛翔。 zai4 hua1 de1 hai3 yang2 li3 zi4 ou2 de1 fei1 xiang4
10	非常熱愛中國古典文學。 fei1 chang2 re4 ai4 zhong1 guo2 gu3 dian3 wen2 xue2

Table 5.1: Recording Sentences of Level Judgment Module

5.2.1 Phoneme Scoring

Mandarin is a tonal language which means that tone quality and phonetic quality can be considered independently.

In order to score phoneme quality of learners, we used a method based on Hidden Markov Models (HMMs). The HMM is imagined to generate observation sequences by jumping from state to state and emitting an observation with each jump. In this method, a set of context-independent models along with the HMM phone is used to compute an average posterior probability for each phone. The posterior probability is selected as the evaluation feature for scoring and the context-independent model is regarded as the correct pronunciation of each phone.

We used the speech recognition toolkit HTK [19]). The speech recognizer uses Chinese monophone HMMs. Monophone HMMs for Chinese are trained on native speech data as in regular monolingual speech recognition. Observation feature is extracted from input speech and fed into HMM model net to do decoding. The standard HTK decoder HVite was used for the recognition experiments.

The hmmlist is shown in Table 5.2.

b	p	m	f	d	t
n	l	g	k	h	j
q	x	z	c	s	zh
ch	sh	r	a	ai	an
ang	ao	e	ei	en	eng
er	i	ia	ian	iang	iao
ie	ii	iii	in	ing	iong
iou	o	ong	ou	u	ua
uai	uan	uang	uei	uen	uo
v	van	ve	vn	sp	sil

Table 5.2: *HMM Monophone List*

Because pronunciation errors occur when the correct phonemes are not the most probable ones, systems generally use subsequent processing on more than one syllable hypothesis to obtain high phoneme recognition accuracy. This is referred to as an "N-best" strategy, since the selected syllables are the N most probable phoneme sequences found by the recognizer. We developed all syllables'

network according to Table 4.2 and Table 4.3.

Then the post-processing of the N-best list, known as "N-best rescoring", re-orders the list using a linear combination of scores from acoustic likelihood for each of the N-best hypotheses, resulting in a new top choice. The phoneme sequence with the highest probability score is selected finally as the output recognition result.

Compare the phoneme list of the recognition result with the correct script, we can calculate the percentage of correct pronounced phonemes.

5.2.2 Tone Scoring

Tone recognition plays an important role and provides very strong discriminative information for Chinese speech recognition. Tone recognition for fluent Mandarin speech has always been a very difficult problem, because of the complicated tone behavior. Traditionally, detailed tone features such as the entire F0 curve are used for tone recognition. In paper [14], the observation sequence ($F0, \Delta F0$) of the whole syllable are used with a combination of VQ (Vector Quantization) and HMM for tone recognition. Tone models of different size, ranging from very simple one-tone-one-model tone models to complex phoneme dependent tone models, have different ability to characterize tone.

When preparing for system development, we have read many papers about Chinese tone recognition. So far, the studied techniques for standard Chinese tone recognition mainly depend on the feature-based or HMM-based methods. Tone refers to the movements of the F0 contour within a syllable. Many of tone recognition algorithms are based on features extracted from the F0 contours. Although, as shown in a number of documents, each F0 contour of four tones can be stylized to a simple pattern distinctive from others. Although there are only five tones in continuous speech, to discriminate them well is a very difficult task. Due to complex F0 variations, isolated tone recognition methods cannot be directly applied to continuous speech. Tone patterns of syllables in continuous speech are subject to various modifications. Tone pattern of a syllable may be seriously affected by the tones of neighboring syllables. This effect is commonly known as tone sandhi. Coarticulation effect makes the F0 contour of a syllable be affected by the F0 contours of neighboring syllables. There are also some other factors, which can affect the realization of tone patterns of syllables.

A simple measure for tone scoring is just based on tone recognition results. If the tone is recognized correctly, its pronunciation is judged as correct and otherwise it will be judged as one mistakes. Such kind of measure will be highly dependent on the recognition accuracy of the HMM models.

At first, we had thought about using independent tone recognizer. But some methods are too complicated to be put into practice in our system or some are not suitable for our system objective. Thus, we tried another method. In Chinese, tones are usually marked on vowels, so we were thinking of extend the number vowels in hmmlist. That means, for example, extending the vowel /a/ to /a1/, /a2/, /a3/, /a4/. The number labeled behind the vowel indicates the tone type. Then, the speech recognizer can do phoneme recognition and tone recognition at the same time.

Though theoretically speaking, our idea is feasible, when using our new trained HMM model, it is a pity that it seems not working well. The number of vowels in hmmlist increases 4 times, but the volume of training corpus is limited. Effective HMMs can not be obtained. Furthermore, spectral features such as MFCC can improve tone recognition accuracy in ASR¹ systems, however, the most discriminative feature for tone is F0. In order to use features be beneficial to the detection of tone mispronunciation, we had better used F0 related features and its combination with MFCC_E_D_Z. But when considering phoneme features, we are using MFCC_D_A_0, it is hard to make the best of both worlds. Besides, when deal with tone issues, female voice and male voice should be processed separately. Currently, because of the time limit and in order to hold most of the system's function, we are ignoring tone scoring part, and only be concern about phoneme scoring. We know this is not good, but we could only leave tone scoring part as a regret.

5.2.3 Judgment Report

In the judgment report, we are not only intended to provide a result of level judgment like "Beginner" or "Advanced Learner", but also to tell the learner what are the problems of his/her pronunciation and how to overcome his/her weak points. Therefore, the system will provide the following aspects of information in the report:

¹Automatic Speech Recognition

- a) Display HMM phoneme recognition results
- b) Give evaluation score
- c) Recommend practice course
- d) Tell instructions for practice

We tried to design the pronunciation evaluation algorithm to reach a judgment result as closer as possible to human scoring. Finally, we decided to calculate following indices to get machine scores:

1. Strictness

According to the recognition result of HMM, we calculate the percentage of wrong phonemes, in order to estimate how necessary should the learner have phoneme practice through word level course.

$$S = \frac{\text{Count of Rejected Phonemes}}{\text{Total Count of Phonemes}} \quad (5.1)$$

2. Segment Duration Scores

Insertions, deletions or substitutions of phonemes will result in duration differences, so we use the simplest approach – a measure of Rate of Speech (ROS) to compute the average number of phonemes per unit of time. From native speech data, we can calculate average number of phonemes per seconds of every sentences. The problem is that a learner probably pronounces the sentence slower than a native speaker. Therefore, we used the result of our speech analysis. According to our analysis results in Section 4.3, add a coefficient "1.3", then, calculate learners' rate of speech. By scaling the difference between learner speech and native speech (multiply coefficient 1.3), the ROS shows learner's performance of phoneme durations and indicates how necessary should the learner have duration practice through sentence level course.

3. Phone-level Likelihood-based GOP ²

The posterior probability is an absolute measure of how the pronunciation is close to the acoustic model. The models are trained by standard Mandarin speech corpus, consequently, Posterior probability can be directly used for phonemic pronunciation assessment. The aim of the GOP measure is to provide a score for each phoneme of an utterance.

²Goodness of Pronunciation

$$GOP = \left| \text{Log} \left(\frac{P(O^{(p)}|p)}{\max_{q \in Q} P(O^{(p)}|q)} \right) \right| / NF(p) \quad (5.2)$$

where Q is the set of all phone models and $NF(p)$ the number of frames in the acoustic segment $Q^{(p)}$.

A block diagram of the resulting scoring mechanism is shown in Figure 5.3. For each sentence, we first extract MFCC features from waveform, then use continuous speech recognizer to do forced alignment and unconstrained phoneme loop. Based on the results of forced alignment and phoneme recognition, the individual GOP scores are calculated for each phoneme. Finally, a threshold is applied to each GOP score to reject badly pronounced phonemes. The choice of threshold can decide the level of strictness. That means, the system's sensitivity of detecting phoneme error can be adjusted. For instance, pruning thresholds can be set so that the system detects more phoneme errors than human judgments or vice versa. By using threshold as a filter, it becomes possible to generate a seriously mispronounced phoneme list of the learner. And then by calling a guidance dictionary, the system can provide practice instructions to let learner focus on practices of more noteworthy phonemes.

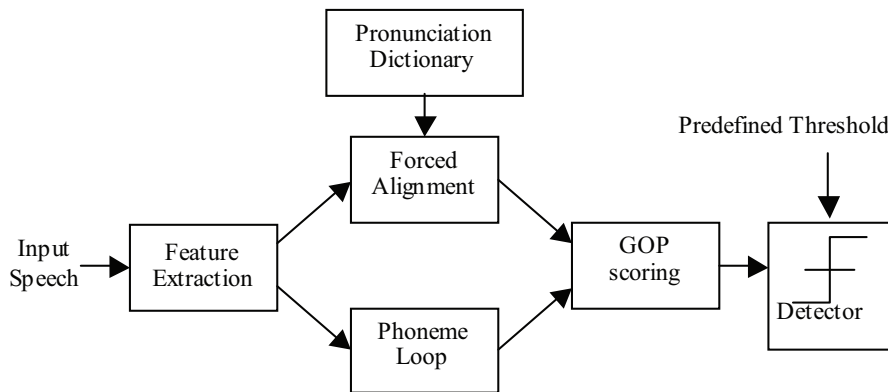


Figure 5.3: Block Diagram of GOP Scoring

By using phoneme scoring result, and combining with some other parameters, system will give a level judgment result to the learner. The flow chart of level judgment is shown in Figure 5.4. If the learner pronounced more than 30% wrong

phonemes, he/she would be recommend to Word-level Course; or, if the learner's average phoneme duration is $\pm 20\%$ bigger than normalized native one, he/she would be recommend to Word-level Course; or, from GOP score, if there are more than 3 fatal phoneme errors, he/she would be recommend to Word-level Course.

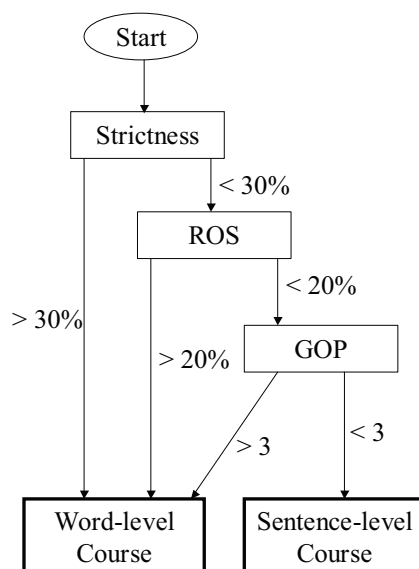


Figure 5.4: *Learner Level Judgment Flow Chart*

5.3 Word-level Practice Course

In word-level practice module, the system provides word-level phoneme and tone practices. Form interface is plotted in Figure 5.6. We selected words that Japanese learners easily make mistakes. At first there are initial instructions to remind the learner, such as suggesting which phoneme to take care or how to pronounce some particular tone correctly.

Learner are allowed to listen to teacher's reference utterance repeatedly and can record his own pronunciation. After learner records his pronunciation, By pressing "Evaluate" button, system will detect phoneme errors and tone errors. According to detected errors, the text of instruction will be changed. For example, if user mispronounced 'r' as 'l', then text of how to pronounce raised tongue phoneme will be shown:

中国語 R と L の発音に関して:
R は濁ります。「り」って感じです。
L は濁りません。「リ」と同じです。
L の発音時は前歯付近まで下を巻いて下を前に送り出す感じで話すと上手く出ます。
R の発音時は舌を喉チンコに届け! っていうくらい意識して舌を前に出さないで話すと好いかも知れません。

Figure 5.5: Pronunciation Instructions of 'r' and 'l'

Feedback has some great importance in second language acquisition, as it is the way the learner gets to understand his mistake and the way correct them. Thus, system will also modify user's utterance and generate corrective audio feedback in learners' own voice using speech modification to help them understand the problems of their pronunciations [21].

Word-level practice course structure is shown in Figure 5.7.

5.3.1 Speech Modification using TD-PSOLA

TD-PSOLA stands for Time Domain Pitch Synchronous Overlap Add. It was introduced in the speech processing community in the early nineties [22], and is now widely known to be simple and efficient in producing high quality speech modification.

The PSOLA scheme involves the three following steps: an analysis of the original signal, modifications brought to this intermediate representation, and finally the re-synthesis of the modified intermediate representation. The overall principle of the method can be observed in Figure 5.8.

The efficiency of the TD-PSOLA technique relies on the determination of the pitch-marks. In our case, we used some open-source toolkit of **Galatea Project**³.

Using TD-PSOLA scheme and pitch-marking process, we can do speech modifications like:

- Modify the delays between the synthesis signals to change the pitch of new signal

³see <http://hil.t.u-tokyo.ac.jp/galatea/index.html>

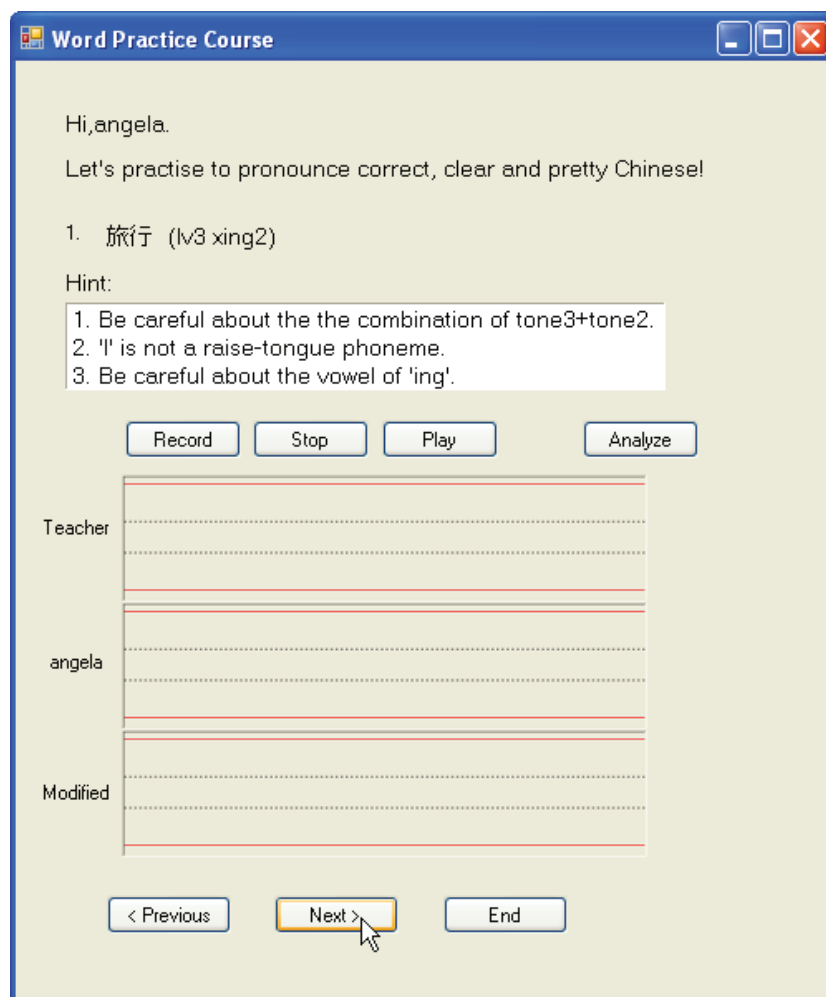


Figure 5.6: *Word-level Course Form Interface*

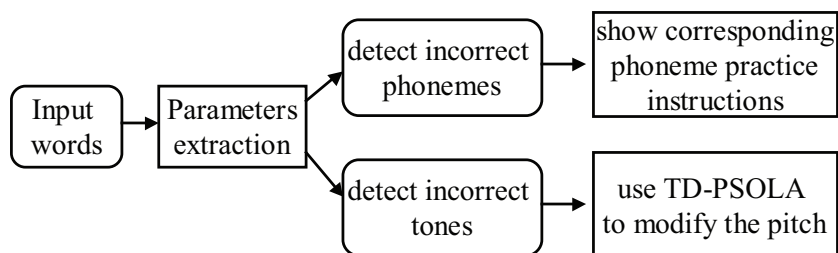


Figure 5.7: *Word-level Course Structure*

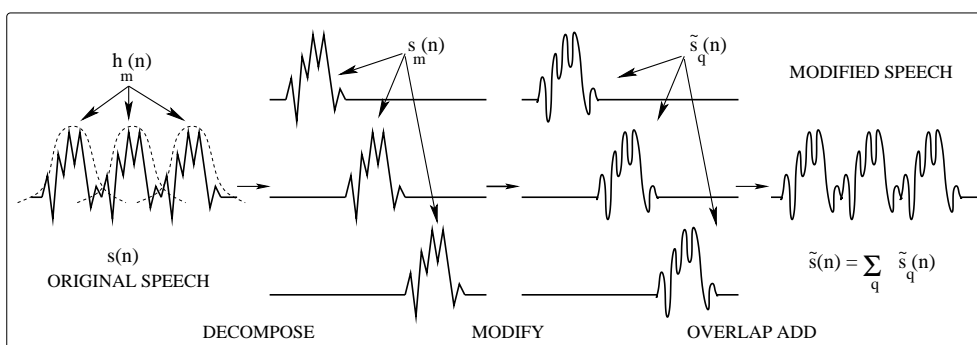


Figure 5.8: Overall Principle of the PSOLA Method

- Modify the number of synthetic pitch-marks to change the duration of each phoneme

In order to modify learners' utterances according native's speech, we applied another method called "Teacher Mapping Technique", and that consists in mapping some features extracted from a native speaker's speech, into the learner's one. Our modification method supposes that the contexts of the words are fixed and that that same words uttered by a native speaker is available. In this situation, the comparison between the native speaker's and learner's utterances is one of the promising ways to automatically 'correct' the learner's utterance. See Figure 5.9 for an overview of the method.

Pitch mapping is adapting the native's pitch marks onto the learners' signal. First we obtain the intervals between two consecutive pitch marks for the native's speech, then apply this pattern of intervals to the learners' signal after having multiplied it with a pitch ratio in order to respect the voice tone difference. This pitch ratio was calculated by:

$$P_{ratio} = \frac{PmD_{ut.}(L)}{PmD_{ut.}(N)} \quad (5.3)$$

PmD standing for pitch-marks delays, $_{ut.}$ indicates that the calculation was realized on the whole utterance, and L , N standing for "Learner", "Native" respectively.

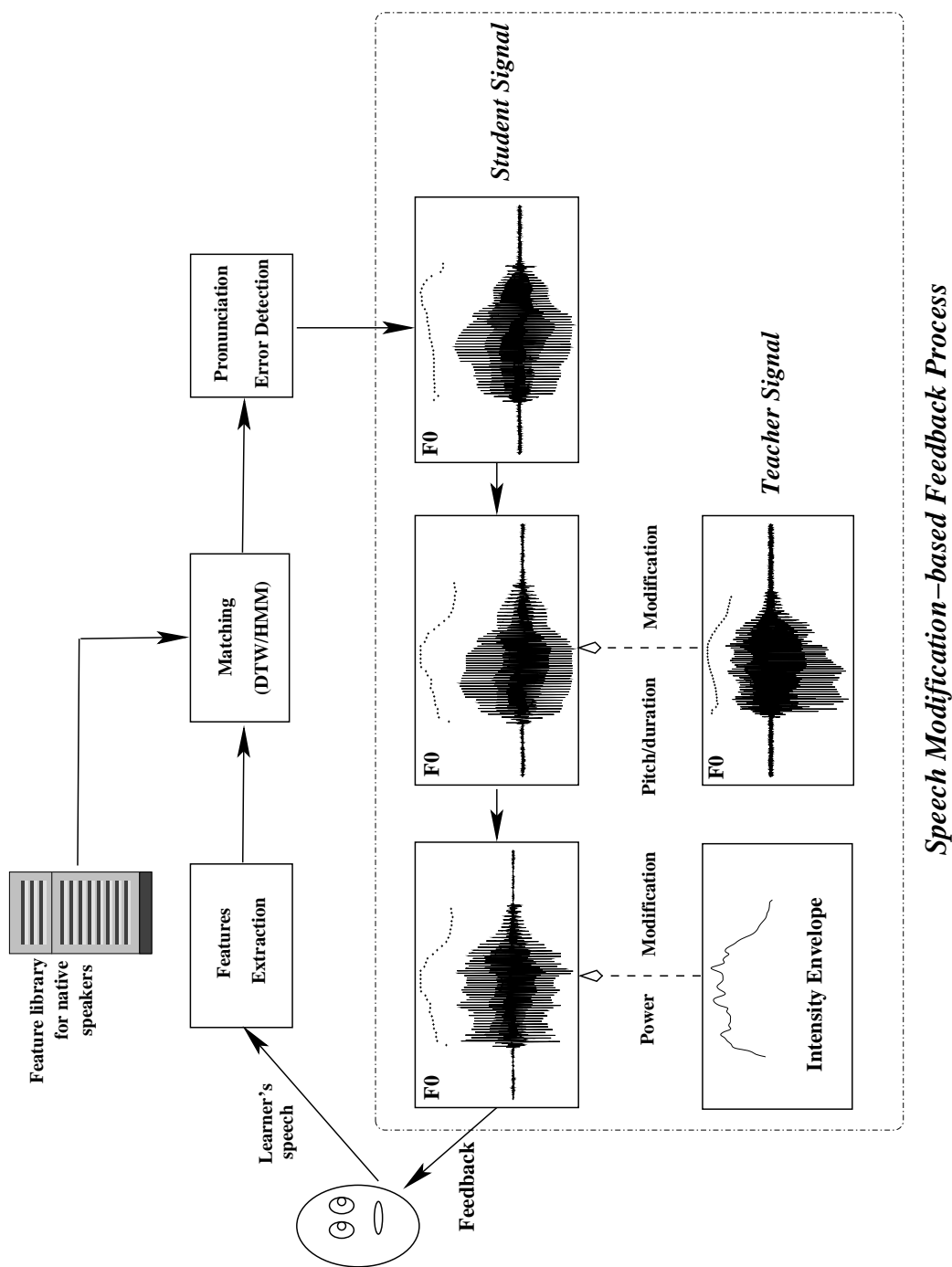


Figure 5.9: Overview of the Teacher Mapping Technique

5.4 Sentence-level Practice Course

In sentence level practice module, the general practice process is similar to word-level practice module. That is, The system displays Chinese sentences on the computer screen and instructs the learner. Since the practice emphases now are durations and sentence-level tones, the initial instructions would be how to segment sentence into small words correctly, and tone changing information. System will enable learner to listen to model utterances by native speakers of Chinese and to record their own pronunciation of those utterances. System will generate corrective audio feedback in learner's own voice. Learners are able to compare their own utterances with generated new voices and those of the native model speaker. The convenient interface allows learners to repeat their recording many times, each time attempting to approximate the native speech more closely. Hence the purpose is to enable learner to practise the correct intonation of Chinese entire sentences by direct comparison of their original utterance with generated new corrective ones. Sentence-level practice course structure is shown in Figure 5.10.

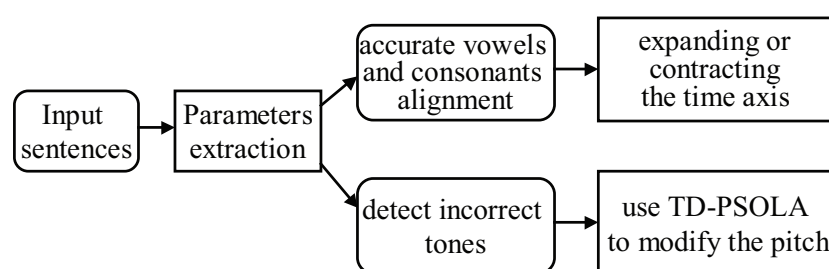


Figure 5.10: *Sentence-Level Course Structure*

5.4.1 Duration and Tonal Modification

Prosody, perceived as stress, intonation and rhythm, is extremely important to our perception of natural speech, and timing is an essential part of prosody.

Our general approach to provide duration and tonal corrections in sentences is to modify a waveform of the learner's speech. This is done in a two stage process. The first stage generates duration tier and a pitch contour from native speech. The second stage alters the duration and pitch in the learner waveform to match the generated duration tier and a pitch contour. For detail, see Section 4.4.

5.5 Conclusion

In this chapter, we gave a detail introduction of the realization of a learner level adaptive CALL system which is based on the speech recognition technology and speech modification technology.

- We have studied typical Chinese pronunciation errors and the characteristics of mistakes of Japanese learners, then designed the system which gives effective instruction utilizing speech recognition technology.
- Our system enables learners of different levels to learn by themselves according to their own proficiency. It makes an independent learning possible by provide a target-oriented opportunity to practice spontaneously by oneself as many times as possible, in their own time and in addition to the normal face-to-face contact with the language teacher.
- Feedback is an important part of foreign language learning. Our system provides feedback similar to human teachers. For phoneme, duration and tone in particular, the differences between native and learner pronunciations are highlighted, and the learners of Chinese are told pronunciation instructions.
- For pronunciation tutoring, one method to provide feedback is to provide examples of correct speech for the student to imitate. However, this may be frustrating if a student is unable to completely match the example speech. Our system provides audio feedback using a student's own voice. The system uses speech modification techniques to alter student pronunciations while maintaining other voice characteristics.

Chapter 6

System Evaluation

We evaluated our system on following aspects:

a) Phoneme Recognition Rate

Since learner level judgment is mainly based on the phoneme recognition result of system, recognition accuracy is a very important point of system evaluation.

b) Judgment Validity

In order to verify the reliability of system's learner level judgment, judgment result done by the system is compared to the mean of pronunciation scores evaluated by native speakers. The correlation coefficient shows the judgment validity.

c) Audio Feedback and Instructions

The system provides various kinds of feedbacks, like text-based instructions and corrective audio feedback. Evaluating whether these instructions and feedback work or not is necessary.

6.1 Recognition Rate

We randomly chose learner utterances of 5 sentences (including 89 phonemes). The system's recognition result is totally 18 phonemes marked as phoneme errors. On the other side, natives also mark out phoneme errors by listening test. Comparing to human result, the system judgment of phoneme correction is about 90%.

6.2 Judgment

The process of how system obtains judgment result is described in Section 5.2.3. The thresholds used are all carefully designed in order to be close to human estimation.

Since the borderline of beginner and advanced learner is not clear, it is hard for natives to evaluate a learner's level only by listening to 10 utterances. So, we cannot simply compare system judgment result and human listening test result to finish the judgment function evaluation. We collected 4 learners utterances and asked native speaker to decide "phoneme & tone course" or "duration & intonation course" to recommend. Then, comparing with system's judgment, correlation is 100%. The perfect correlation indicates good judgment function.

6.3 Feedback

In an ideal way, feedback is adapted to the learner's errors and capacity to integrate corrections. The system should detect where errors are, inform the learner and give he/she necessary information to correct these errors.

6.3.1 Instructions

Automatic feedback is important for language learning systems in order to make the system interactive for students. All testers felt the instructions offered system are proper and helpful. After practicing several times of each words or sentences, phoneme errors detected by system decreases, which indicates all testers are making progress.

6.3.2 Audio Feedback

Comparison-based speech evaluation/modification method involve a risk of giving a low score for a natural utterance if the pattern of the utterance happens to be different from the reference utterance. However, it is very efficient for training learners (especially second language learners) to mimic a teacher's utterance.

Audio feedback prompts learners to pronounce words and sentences and returns immediate feedback. According to a subjective listening test (details shown in Section 4.4), the audio feedback generated by system shows good improvement of pronunciation. Our system is providing useful feedback information to help learners to practice.

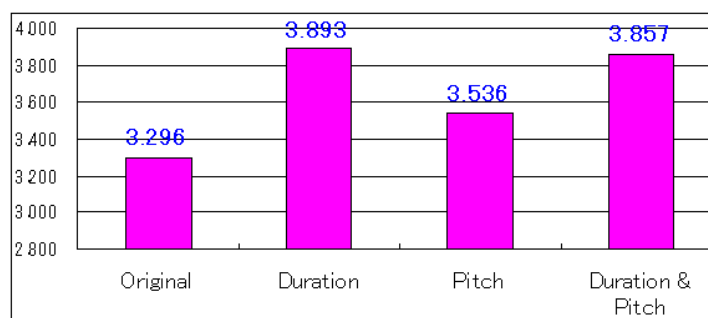


Figure 6.1: *Listening Test Results of Modified Speech*

6.4 Conclusion

Experimental results showed that level judgment result the system evaluated is reliable and all the instructions the system offered could bring improvements to learner pronunciation performance. Furthermore, the modified speech of our system has high sound quality, relatively higher naturalness and intelligibility.

Chapter 7

Future Work

There are several unsettled issues of our learner level adaptive CALL system which need doing in the future:

- Up to now, only simply evaluation is finished. After we manage to make our system more stable, we are going to contact class of learners to try our system and get more tryout reports.
- Difficulty of pronunciation is different depending on phonemes. So when calculating GOP, we are planning to set a different threshold for each phoneme. This may help enhance evaluation precision.
- Tone is an acoustic characteristic of Chinese language, which is also a prominent difficulty in Chinese learning. How to guide the learners to correctly distinguish pronunciations of four tones is a key point of Chinese language education. However, in the judgment module, we did not do tone evaluation at the moment. How to improve tone recognition accuracy and add it to our system is also one of our future works.

In future versions, we also have some other ideas described as follows:

- An registry system that provides user log, enables every learner to check history learning record.

- To make the system give a more direct perception visually through illustrations of pitch patterns, durations on time axis, etc; give an effective guidance to system users who are not familiar with speech processing technology, have difficulties to understand the meaning of the different features displayed on the screen and are sometimes confused as how to understand the information they conduct.
- E-learning language courses offer a natural framework of collecting corpora. A large volumes of language data may be collected in a short time. An online exercise can collect dozens of answers with mistake patterns. These can be minded automatically and fed back into assessment tools. To make the system as a web-based application, and it can record all the results from different users, to provide scientific materials for Chinese education through detailed error analysis. It can provide teachers more complete and detailed statistics for making effective education plans. Students can also be benefit from the same information offered by the system.

Appendix A

Reading Script for Speech Analysis

Part 1

1. 昨日から頭が痛いので今日の授業は休みました。
2. 木曜日の午後にある音楽の授業が一番の楽しみだ。
3. ビールを飲みながら皆で語り合い にぎやかな夜だった。
4. 鍋に水を入れて湯を沸かしてください。
5. ラーメンが食べたいと思つて近所の店まで歩いた。
6. テレビを見ながら家で一人でワインを飲んだ。
7. 音楽のリズムに合わせて子ども達が元気に踊っている。
8. 何度も繰り返し練習しなければうまくなることはできません。
9. この仕事が終わるまで帰らないでください。
10. 5時になったら帰つてもいいです。
11. 読みたい本は借りられているので今日は借りられない。
12. 覚えることがたくさんあつて全部覚えられない。

Part 2

1. zuótiānkāishǐtóutòng suǒyǐjīntiānbùqùshàngkèle
昨天开始头痛，所以今天不去上课了。
2. xīngqīsìxiàwǔdeyīnyuèkèzúiyǒuqù
星期四下午的音乐课最有趣。
3. nàshìyígedàjiāyìbiānhēpíjiǔyìbiānliáotiānderènaodeyèwǎn
那是一个大家一边喝啤酒一边聊天的热闹的夜晚。
4. qǐngjiāngshuǐdàorùguōlǐshāokāi
请将水倒入锅里烧开。
5. xiǎngchīlāmiànle yúshìbùxíngqùfùjìndediàn
想吃拉面了，于是步行去附近的店。
6. dúzìyígerénzài jiālǐyìbiānkàndiànshìyìbiānhējiǔ
独自一个人在家里一边看电视一边喝酒。
7. hái zǐmen bàn zhe yīnyuè de jiézòu kuài lè dì tiào wǔ
孩子们伴着音乐的节奏快乐地跳舞。
8. bù fǎn fù liànxí de huà jiù bù néng zuò de hěn hǎo
不反复练习的话就不能做的很好。
9. qǐng gàn wán zhè jiàn gōng zuò hòu zài huí jiā
请干完这件工作后再回家。
10. dào diǎn zhōng jiù kě yǐ huí jiā le
到5点钟就可以回家了。
11. xiǎng kàn de shū yǐ jīng bèi rén jiè le jīn tiān jiè bù liǎo
想看的书已经被别人借了，今天借不了。
12. yào jì de dōng xī tài duō bù néng quán bù jì zhù
要记的东西太多，不能全部记住。

Part 3

- | | | |
|-----------------|-------------------|----------------------|
| 1. zhīshì
知识 | 9. gōngzuò
工作 | 17. huāngzhāng
慌张 |
| 2. hèlǐ
贺礼 | 10. yǐlái
以来 | 18. zhǐshì
指示 |
| 3. rěnrǎn
忍受 | 11. tiàowǔ
跳舞 | 19. lǚxíng
旅行 |
| 4. jiézòu
节奏 | 12. hélǐ
合理 | 20. guànlì
惯例 |
| 5. yīnyuè
音乐 | 13. tónghuà
童话 | 21. dōngxī
东西 |
| 6. liánxì
联系 | 14. kāishǐ
开始 | 22. yīlài
依赖 |
| 7. fùjìn
附近 | 15. rénshǒu
人手 | 23. yīnyuē
隐约 |
| 8. guǎnlǐ
管理 | 16. liànxí
练习 | 24. gǎnmào
感冒 |

Part 4

- tā yīn wéi gǎn mào fā shāo le, suǒ yǐ qǐng le yī tiān jiǎ.
他因为感冒发烧了, 所以请了一天假。
(彼は風邪を引いて熱が出たので1日休みをもらった。)
- dà jiā dōu shuō zhè bù diàn yǐng hǎo kàn, yú shì wǒ yě qù kàn le.
大家都说这部电影好看, 于是我也去看了。
(みんなこの映画は良いと言っていたので私も見に行つた。)
- chú le tā yǐ wài, hái yǒu méi yǒu rén zhī dào?
除了她以外, 还有没有人知道?
(彼女以外に知っている人はいますか?)
- suí zhe wǒ guó jīng jì de fā zhǎn, rén men de shēng huó shuǐ píng yě tí gāo le.
随着我国经济的发展, 人们的生活水平也提高了。
(わが国の経済発展にともない、人々の生活レベルも向上した。)
- dǎ gè diàn huà wèn wèn bù jiù zhī dào ma?
打个电话问问不就知道吗?
(電話して聞いてみれば分かる。)
- jīn tiān shì xīng qī liù ma?
今天是星期六吗?
(今日は土曜日ですか?)
- wèi le yǒu yì yǔ hé zuò, xū yào jì chéng hé fā yáng zhōng rì yǒu hǎo yuǎn liú cháng de lì shǐ chuán tǒng.
为了友谊与合作, 需要继承和发扬中日友好源远流长的历史传统。
(友情と協力のために、中日友好の悠久な歴史と伝統を受け継ぎ、発展させる必要があります。)

Part 5 強調

1. zhèshìwǒdìyīcìdào zhōngguó lǚxíng
这是我第一次到中 国旅行
(今回は、私にとって一回目の中国旅行です。)
2. zhèshìwǒdìyīcìdào zhōngguó lǚxíng
这是我第一次到中 国旅行
(今回は、私にとって一回目の中国旅行です。)
3. dàxuébìyèhòuzài běijīng gōngzuò le liǎng nián ránhòu lái shànghǎi de
大学毕业后在北京工 作了两 年。然后来上 海的。
(大学卒業後、北京で2年間仕事をしました。その後、上海へ来ました。)
4. dàxuébìyèhòuzài běijīng gōngzuò le liǎng nián ránhòu lái shànghǎi de
大学毕业后在北京工 作了两 年。然后来上 海的。
(大学卒業後、北京で2年間仕事をしました。その後、上海へ来ました。)

Part 6 Segmentation

1. wèile yǒuyì yǔhézuò xūyào jìchéng hé fāyáng zhōng rì yǒu hǎo yuányuǎn liúcháng de lìshǐ chuántǒng
为了|友谊与合作，需要|继承和发扬|中 日友好|源 远 流 长 的|历史传统。
(友情と協力のために、中日友好の悠久な歴史と伝統を受け継ぎ、発展させる必要があります。)
2. nàshíyí gè dàjiā yíbiān hē pǐjiǔ yíbiān liáo tiān de rènao de yèwǎn
那是一个|大家|一边喝啤酒|一边聊天的|热闹的|夜晚。
(ビールを飲みながら皆で語り合い にぎやかな夜だった。)
3. hái zǐ mēn bàn zhe yīnyuè de jiézòu kuàilè de tiàowǔ
孩子们|伴着|音乐的节奏|快乐地|跳舞。
(音楽のリズムに合わせて子ども達が元気に踊っている。)

Part 7

1. zhèshìwǒdìyīcìdào zhōngguó lǚxíng
这是我第一次到中 国旅行
(今回は、私にとって一回目の中国旅行です。)
2. dàxuébìyèhòuzài běijīng gōngzuò le liǎng nián ránhòu lái shànghǎi de
大学毕业后在北京工 作了两 年。然后来上 海的。
(大学卒業後、北京で2年間仕事をしました。その後、上海へ来ました。)

Bibliography

- [1] Tohru TAKAGI, Akiko HATTORI, Megumi KOMIYA, Atsushi IMAI, Kenshi KISHI and Takayuki ITO.
A Chinese Language Learning System with Visualization and Speech Correction for Prosody.
Institute of Electronics, Information, and Communication Engineers
Vol.J88-D-I No.2 pp.478-487.
- [2] RenHua WANG, Qingfeng LIU, and Si WEI.
PUTONGHUA PROFICIENCY TEST AND EVALUATION.
- [3] L.Neumeyer, H.Franco, V.Digalakis, and M.Weintraub.
Automatic Scoring of Pronunciation Quality.
Speech Communication, 83 (2000).
- [4] S.M.Witt, and S.J.Young.
Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning.
Speech Communication, 95 (2000).
- [5] Aijun LI, and Xia WANG.
A contrastive Investigation of Standard Mandarin and Accented Mandarin.
EUROSPEECH 2003, 2345-2348, GENEVA.
- [6] Lin-shan LEE.
Structural Features of Chinese Language – Why Chinese Spoken Language Processing is Special and Where We Are.
keynote Speech, 1998 International Symposium on Chinese Spoken Language Processing, Singapore, 1998, pp1-15.
- [7] L. Wang, E. Ambikairajah, and E.H.C. Choi.
Automatic Tonal and Non-Tonal Language Classification and Language Identification Using Prosodic Information.
2006 International Symposium on Chinese Spoken Language Processing, Singapore, 2006.
- [8] Jongman, A., Y.Wang, C. B. Moore, and J. A. Sereno.

- Perception and production of Mandarin Chinese tones.*
Handbook of Chinese Psycholinguistics, Cambridge University Press, 2006.
- [9] Boersma, P., and Weenink, D.
Praat.
<http://www.praat.org/>.
- [10] Goh KAWAI.
Spoken Language Processing Applied to non-native Language Pronunciation Learning.
Doctor thesis, The University of Tokyo, 1999.
- [11] Frederic GENDRIN.
Accent Pattern CALL System using Speech Modification based Corrective Feedback.
Master thesis, The University of Tokyo, 2003.
- [12] Yiochi YAMASHITA, Keisuke KATO, and Kazunori NOZAWA.
Automatic Scoring for Prosodic Proficiency of English Sentences Spoken by Japanese Based on Utterance Comparison.
IEICE TRANS.INF.&SYST., Vol.E88-D, No.3(20050301) pp. 496-501.
- [13] Fric GENIN, Kei HSE, and Noki MISU.
The Evaluation of the Japanese students' Chinese Four Tones Discrimination ability By Computer-Assisted Instruction System for Self-Teaching.
2nd International Workshop on Language and Speech Science, Sep.4-5, 2008.
- [14] W.J.Yang, J.C.Lee, Y.C.Chang, and H.C.Wang.
Hidden Markov model for Mandarin lexical tone recognition.
IEEE Tran.Acoustics, Speech, and Signal Processing, v36, n7, 1988, pp.988-992.
- [15] Oh-pyo KWEON, Motoyuki SUZUKI, Akinori ITO, and Shozo MAKINO.
A Study on Japanese Pronunciation Learning System for Korean Using Speech Recognition.
TECHNICAL REPORT OF IEICE. SP2002-164 (2003-01).
- [16] Keiichi TOKUDA.
FUNDAMENTALS OF SPEECH SYNTHESIS BASED ON HMM.
IEICE technical report. Speech, Vol.100, No.392(20001019) pp. 43-50, SP2000-74.
- [17] Zhang, L., Huang, C., Chu, M., Soong, F., Zhang, X. D. and Chen, Y. D.
Automatic detection of tone mispronunciation in Mandarin.

- In Proceedings of the 5th International Symposium on Chinese Spoken Language Processing (ISCSLP '06). 590–601.
- [18] Keikichi HIROSE.
Speech Prosody and CALL.
Journal of the Phonetic Society of Japan, Vol.9 No.2, August 2005, pp.38-46.
- [19] Steve YOUNG et al.
Speech Recognition toolkit HTK.
<http://htk.eng.cam.ac.uk/>.
- [20] Witt, S. and Young, S.J.
Phone-level pronunciation scoring and assessment for interactive language learning.
Speech Communication, 30 (2-3). pp. 95-108. ISSN 0167-6393.
- [21] Frederic GENDRIN, Keikichi HIROSE and Nobuaki MINEMATSU.
Corrective feedback for accent pattern CALL systems using speech modification.
IEICE Technical Report, SP2002-161, pp.1-6 (2003-1).
- [22] H. VALBRET, E. MOULINES, and J.P. TUBACH.
Voice transformation using PSOLA technique.
Speech Communication, 11:175–187, 1992.
- [23] Leonardo NEUMEYER, Horacio FRANCO, Mitchel WEINTRAUB, and Patti PRICE.
Automatic Text-independent Pronunciation Scoring of Foreign Language Student Speech.
Proc. of ICSLP 96, pp.1457-1460, Philadelphia, Pennsylvania, 1996.
- [24] Frank K.SOONG, Wai-Kit LO, and Satoshi NAKAMURA.
General-ized word posterior probability (GWPP) for measuring reliability of recognized words.
Proc.SWIM, 2004.
- [25] Ye Tian, Jian-Lai Zhou, Min Chu, Chang, E.
Tone recognition with fractionized models and outlined features.
Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP '04).
- [26] Si WEI, Hai-Kun WANG, Qing-Sheng LIU, and Ren-Hua WANG.
CDF-MATCHING FOR AUTOMATIC TONE ERROR DETECTION IN MANDARIN CALL SYSTEM.
ICASSP2007, pp.205-208, 2007.

- [27] Oh-pyo KWEON, Motoyuki SUZUKI, Akinori ITO, and Shozo MAKINO.
*A Study on Japanese Pronunciation Learning System for Korean Using
Speech Recognition.*
IEICE Technical Report, VOL.102, NO.618(SP2002 161-168), PAGE.19-
24(2003).
- [28] Valbret H, Moulines E, Tubach J.P.
Voice transformation using PSOLA technique.
Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE .

Publications

- [1] Minyi MA, Hiroko HIRANO, Keikichi HIROSE and Nobuaki MINEMATSU.
Proposal of Adaptive CALL system for Japanese Learners of Mandarin
Proc. Autumn Meeting of the *Acoustical Society of Japan*, 1-Q-6, pp.333-334 (2008-9).
- [2] Minyi MA, Keikichi HIROSE and Nobuaki MINEMATSU.
A Learner Adaptive Chinese Pronunciation Education System for Japanese
Proc. Spring Meeting of the *Acoustical Society of Japan*.
(to be published)