Ph.D. Thesis

博士論文

Improvements in Pronunciation Evaluation for Reading-Aloud and Shadowing Speech Based on Speech Technology

（音声情報処理に基づく音読・シャドーイング音声の自動評価の改良手法）

2010 年 6 月 15 日

指導教員　峯松 信明 准教授

東京大学大学院工学系研究科
電子工学専攻　**37-077090**

羅　徳安
Dean Luo

Abstract

The main goal of this research is to improve automatic pronunciation evaluation of reading-aloud and shadowing based on speech technology for Computer-Assisted Language Learning (CALL) systems. One of the biggest challenges in CALL development based on speech processing is the mismatches between learners' speech and the native speech data that is used to train acoustic model. In Automatic Speech Recognition (ASR), speech adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) have been used to reduce these mismatches by using small amount of the target speaker's speech as adaptation data. However, in the case of CALL, learners' pronunciations often contain errors. Conventional speaker adaptation techniques that use learners' imperfect pronunciations as adaptation data can cause the over-adaptation problem, in which case errors can be transformed into good pronunciations after adaptation. Although some studies use MLLR adaptation (with only one transformation for all pronunciations) to keep the main characteristic of speaker while ignoring the pronunciation details, to the best of the authors' knowledge, no quantitative analysis has been reported to investigate the adverse effects of conventional speaker adaptation techniques.

To address the over-adaptation problems, we first analyze the effects and side effects of conventional MLLR adaptation for pronunciation evaluation in terms of automatic scoring and error detection. Evaluation experiments show that: a) although global adaption with only one transformation for all pronunciations indeed improves performances, when more transformations are used for different pronunciations, over-adaption occurs. b) In automatic scoring, when the number of regression tree is larger than 4, the correlation between automatic scores and manual scores is worse than the original models. c) In error detection, the performance of recall rate decreases due to over-adaptation but the performance of precision rate increases even with over-adaptation.

In order to better benefit from speaker adaption and prevent over-adaption at the same time, this thesis presents a novel idea that uses a group of teachers' perfect pronunciations to regularize learners' transformation so that over-adaptation problems can be prevented. We name this method Regularized Maximum Likelihood Linear Regression (Regularized-MLLR) and implement it in two ways: one is using the average of the teachers' transformations as constraints adding to conventional MLLR to prevent radical pronunciation transformation, and the other is using linear combinations of teachers' transformation matrices to represent learners' transformations. We refer to the formal implementation as R-MLLR1 and the latter as R-MLLR2. We compare R-MLLR1 and R-MLLR2 with conventional MLLR by conducting experiments on the same conditions as we investigate the adverse effects of MLLR. Automatic scoring and error detection experiments show that the proposed methods outperform conventional MLLR. By adding constraints to MLLR, R-MLLR1 indeed reduces the adverse effects of MLLR, yet performances still drop due to over-adaptation. R-MLLR2 not only out-performs MLLR global adaption, which is widely use for CALL, but also prevents over-adaptation by using linear combinations of teachers' matrices instead of using learners' directly. The proposed methods can better utilize speaker adaptation and prevent adverse effects, and thus more suitable for CALL systems.

Automatic evaluation methods for shadowing are also proposed. Shadowing is a kind of “repeat-after-me” type exercise, but rather than waiting until the end of the phrase heard, learners are required to reproduce nearly at the same time. Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages for its effects of improving both listening and speaking skills. Since learners have to follow the speaking rate of the presented utterance, their pronunciation often becomes very inarticulate and unintelligible. These features of shadowing make it very difficult to build a reliable scoring system for shadowing speech.

Three techniques are proposed for evaluating shadowing speech. One is using Goodness of Pronunciation (GOP) scores calculated through HMM-based forced alignment. In this method, for automatic scoring, the transcription of the presented utterance and the acoustic models of the target language are required. Another is based on continuous phoneme recognition, in which the acoustic models are also needed, but no transcription is required. The third method is using a time-constrained bottom-up clustering technique. Here, only the presented utterance and the shadowed response are required. The transcription and the acoustic models are not needed. Correlations between automatic scores and manual scores, and correlations between automatic scores and learners' TOEIC scores have been investigated and very good results have been obtained.

We also compare the evaluation performances of shadowing and reading-aloud with different cognitive loads posed on learners. Experimental results prove that shadowing can better reflect learners’ true proficiency than reading-aloud by posing an adequate level of cognitive load on learners. Therefore, our proposed shadowing evaluation methods can be used to predict learners’ over-all language proficiency. A shadowing scoring system has been developed based on these methods. The system is being used for English classes in several universities in Japan and has received very positive feedbacks from teachers and students.

Finally, automatic prosodic evaluation has also been proposed for learners’ personal-best shadowing. Experimental results show that rather high correlation with manual prosodic scores has been found. Automatic prosodic scores and segmental ineligibility scores are combined together by using a multiple regression model and the combined scores further improve the performance of automatic scoring that predicts learners’ over-all language proficiency.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Current conditions of English education in Japan and the technologies that support it

Amid the advance in the internationalization of Japan's society, English-speaking Japanese are in greater demand than ever. However, it is well known that English is very different from Japanese phonetically (including rhythm and intonation), and logistically [1, 2, 3]. On top of that, the differences of perception process of speech [4,5] and the cultural differences make it very difficult for Japanese to master English [6-9], especially in terms of oral competence.

In order to improve the current situation of English education in Japan, Ministry of Education, Culture, Sports, Science and Technology (MEXT) set up a strategic plan in 2002 [10], which is titled "Developing A Strategic Plan to cultivate 'Japanese with English Abilities' – a Plan to Improve English and Japanese Abilities". It also set up an action plan, titled "Regarding the Establishment of an Action Plan to Cultivate Japanese with English Abilities" in 2003 [11]. Robert Hughes points out that the most important aspect of the MEXT plans is the upper echelon recognition of a serious educational problem in Japan [12]. This problem is that throughout junior and senior high school, students battle with English grammar and vocabulary to succeed at entrance examinations, and yet with successful entry into a university, for most students, the reason for studying English is gone. As a result, most high school graduates enter university with minimal communicative competency in English and with low level of motivation, additional English classes may not improve student oral communicative competency. Although MEXT has also decided that from 2011, English education will be compulsory at elementary schools, a severe shortage of English teachers is deeply concerned for Japan's English education. With the limitations of large classes and few hours of instruction, it is very difficult to optimize learner motivation and to improve their oral communicative competency [12].

One possible solution to the problems mentioned above is using technology to supporting language learning and teaching. By utilizing the power of computers and the internet, language teachers can collaborate with engineers to develop systems that enable students to learn even without the present of a language teacher. With rich multimedia contents and communication with teachers or other leaner's through the internet, these systems can be more appealing and boost learner motivation than conventional teaching methods.

Recent advance in spoken language processing has made it possible for computer to evaluate learners' pronunciation automatically and thus assist teaching and learning a foreign language, especially spoken language. Systems based on these technologies are often referred to as Computer Assisted Language Learning (CALL) systems [13]. These systems usually compare learners' pronunciations with speaker-independent acoustic models train on native speech of target language. Base on the comparison results, CALL systems can detect errors in learners' pronunciations or give scores for learners' pronunciations to indicate how good they are pronounced. To improve learners' listening, reading, speaking and writing skills, many CALL systems incorporate multimedia (video, audio and text) with speech and natural language processing technologies.

CALL has many advantages over traditional language education in many ways such as enabling self-evaluation for students, reducing time and costs that are required in human-to-human education, efficiency of computer and internet, etc. However, the pronunciation techniques of most of the CALL systems are based on automatic speech recognition (ASR) that performance much better with native speech than foreign-accented speech. Low performance of ASR on learners' speech results with many false alarms of pronunciation error detection or proficiency prediction, which would frustrate and even misguide students in their pronunciation acquisition. For this reason, some experts are skeptical about the usefulness of CALL.

1.2 Research Objectives

This study addresses the problems of conventional ASR technique for CALL and aims at proposing methods for developing reliable CALL systems that can help with pronunciation education. A detailed analysis of conventional speaker adaptation, which is often used in ASR to improve recognition performances, will be conducted and their effects and limitation will be closely examined. Based on the analysis, novel methods will be proposed to improve performance of pronunciation evaluation.

We also collaborate with language teachers to provide technology that are suitable for the needs of pronunciation education. In this research, we focus on automatic evaluation of a popular pronunciation practice, shadowing. We propose several methods for automatic scoring of shadowing and develop a shadowing evaluation system for pronunciation education classes. We also compare shadowing with convention pronunciation practice and provide proof of the volatility of the advantage of shadowing over conventional practices from speech engineering point of views.

1.3 Outline of This Thesis

In the following chapters, first, the background knowledge will be introduced in Chapter 2 and a detailed overview of various kinds of CALL systems and the technologies behind them will be explained in Chapter 3. In Chapter 4, a quantitative analysis of Maximum Likelihood Linear Regression (MLLR) adaption, which is often used in Automatic Speech Recognition (ASR) to improve recognition performance, is conducted on CALL with publicly available databases and over-adaption problem is closely examined. Based on the investigation results in Chapter 4, Chapter 5 presents a novel idea, Regularized-MLLR, which uses a group of teachers' speech data to regularize learners' transformations so that erroneous pronunciations will not be transformed into good pronunciations. We implements this idea in two ways: one is using the average of the teachers' transformation added to learners' transform and the other is representing each learner’s transform as linear combination of the teachers’. Evaluation experiment is conducted to prove validity of the proposed method over conventional MLLR adaptation. In Chapter 6, automatic scoring methods for shadowing is proposed and compared with manual scores and TOEIC over-all proficiency scores. Comparison of shadowing and reading-aloud is described in Chapter 7 and in Chapter 8, we propose prosodic evaluation for shadowing. Finally, conclusions and future works will be presented in Chapter 9.

Chapter 2

Research Background

2.1 Introduction

As explained in the previous chapter, pronunciation education is very important to achieve MEXT's goals. This chapter presents background knowledge of Japanese and English in the context of pronunciation acquisition and the position of the research in improving students' oral communicative competency.

2.2 Phonetic differences between Japanese and English

2.2.1 Vowels

In phonetics, a vowel is a sound in spoken language pronounced with an open vocal tract so that there is no build-up of air pressure at any point above the glottis. Japanese vowels are shown in Figure 2.1 [14]. This figure shows the position of tongue, where vertical scale means the height of tongue and horizontal scale means the part of tongue (front or back). As shown in Figure 2.1, in Japanese language, there are only 5 vowels, /a, i, u, e, o/.

Figure 2.2 shows the English vowels (monophthongs only) [15]. Although diffinition of English vowels differs from dialect to dialect, here, we only consider the most common dialect spoken in the United States, i.e. General American English. In General American, there are 5 short vowels (/ɪ, ʊ, ɛ, ʌ, æ/), 4 long vowels (/i, u, ɔ, ɑ/), 5 diphthongs (/eɪ, ɔɪ, aɪ, aʊ, oʊ/), schwa (/ə/) and /ɚ/, a central vowel before /r/. All together, there are 22 vowels in General American English. Therefore, for Japanese learners of English, due to the lack of vowels in their mother tongue, replacement errors of English vowels by the 5 vowels of Japanese are very common.

2.2.2 Consonants

English and Japanese consonants are shown in Table 1 and 2 as pairs of voiced / unvoiced phones with IPA (International Phonetic Alphabet) pronunciation symbols. In articulatory phonetics, a consonant is a speech sound that is articulated with complete or partial closure of the vocal tract. Table 1 and 2 shows the features of consonants in English and Japanese.

/l/ and /r/ are particularly difficult for Japanese learners to pronounce correctly and are often replaced by a Japanese corresponding consonant /ɸ/. Since in Japanese pronunciations there are no such phones as /f , v/ and /θ, ð/, substitution errors (substitution between /v/ and /b/, /f/ and /h/, /θ/ and /s/, /ð/ and /d/) are among most common errors of Japanese learners of English.

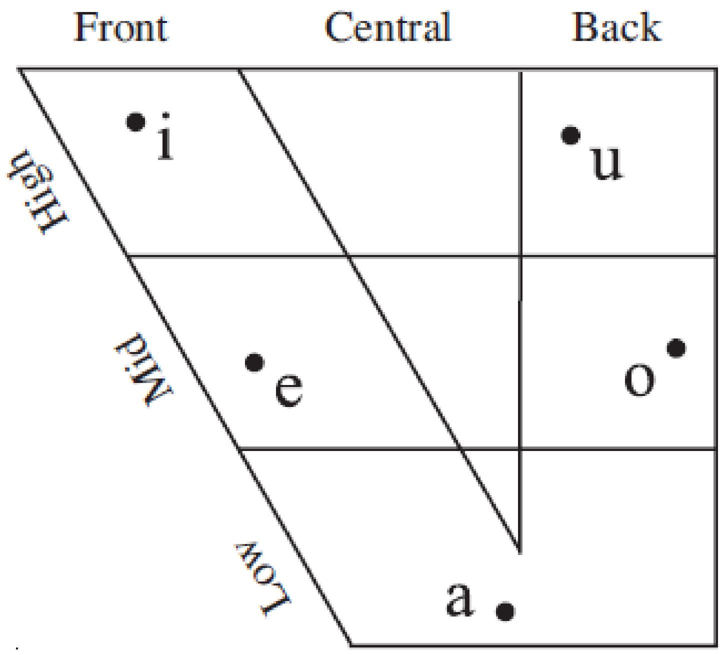


Figure 2.1: Japanese vowels

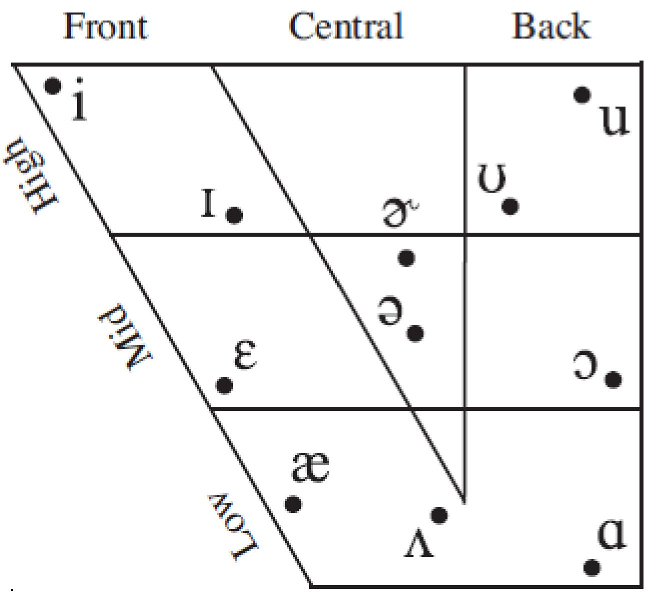


Figure 2.2: English vowels

Table 1: Japanese and English consonants #1

		Bilabial	Labiodental	Interdental	Alveolar	Postalveolar
Stops	J	p/b			t/d	
	E	p/b			t/d	
Fricatives	J	ɸ			s/z	
	E		f/v	θ/ð	s/z	ʃ/ʒ
Affricates	J				ʦ/ʣ	ʧ/ʤ
	E					ʧ/ʤ
Nasals	J	m			n	
	E	m			n	
Liquids	J				ɾ	
	E				l/ɹ	
Glides	J	w				
	E	w				

Table 2: Japanese and English consonants #2

		Retroflex	Palatal	Velar	Uvular	Glottal
Stops	J			k/g		ʔ
	E			k/g		ʔ
Fricatives	J		ç			h
	E					h
Nasals	J			ŋ	N	
	E			ŋ		
Liquids	J	ɻ				
	E					
Glides	J		j	w		
	E		j	w		

2.2.3 Syllables

In English, the pronunciation unit is syllable. However, the Japanese pronunciation unit is mora. Table 3 shows the differences between syllable and mora. As mention in the previous section, there are only 5 vowels in Japanese and yet there are 20 in English. Due to the difference of Mora and syllable, altogether, there are proximally 100 kinds of mora in Japanese and there are more than 10,000 kinds of syllable in English.

2.2.4 Accent

In Japanese, accent is defined by a higher of lower pitch and is often called pitch accent and in terms of speech processing, a type of accent can be decided by using pitch pattern (F0) lonely. However, English is stress-accent, which varies with the change of pitch, power, duration, etc. Stress is the relative emphasis that may be given to certain syllables in a word, or to certain words in a phrase or sentence.

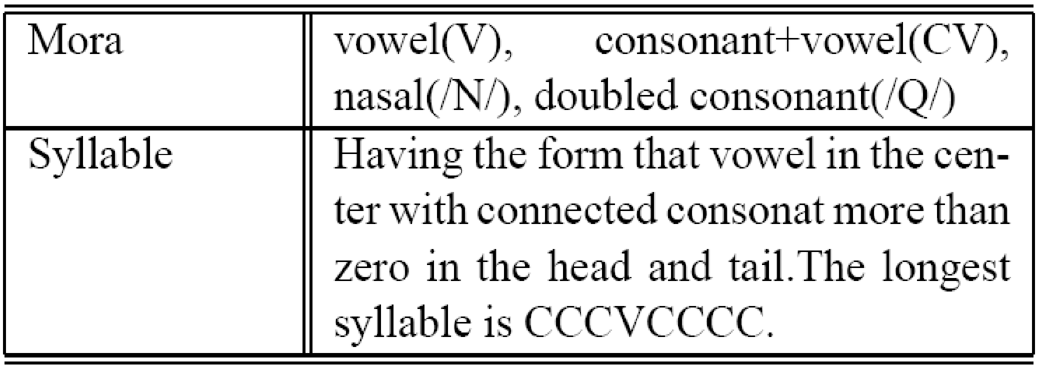
2.2.5 Rhythm

The rhythm of English and Japanese is different in terms of isochrony. Isochrony is the idea that a language rhythmically divides time into equal portions. Three types of divisions are postulated:

1. The temporal duration between two stressed syllables is equal (stress-timed);
2. The duration of every syllable is equal (syllable-timed);
3. The duration of every mora is equal (mora-timed).

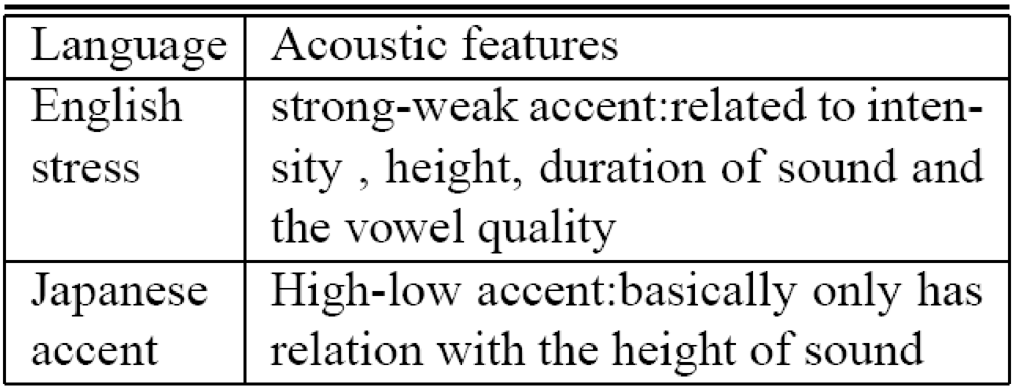
Japanese is mora-timed and English is stress-timed rhythm. In Japanese mora-timed rhythm, a V or CV syllable takes up one timing unit. Japanese does not have long vowels or diphthongs but double vowels, so that CVV takes twice the time as CV. A final /N/ also takes as much time as a CV syllable. In a stress-timed language such as English, syllables may last different amounts of time, but there is perceived to be a fairly constant amount of time (on average) between consecutive stressed syllables. Stress-timing is sometimes called Morse-code rhythm. Stress-timing is strongly related to vowel reduction processes. The summary of differences of Japanese and English in rhythm is shown in Table 4.

Table 3: The difference between mora and syllable



Mora	vowel(V), consonant+vowel(CV), nasal(/N/), doubled consonant(/Q/)
Syllable	Having the form that vowel in the center with connected consonat more than zero in the head and tail.The longest syllable is CCCVCCCC.

Table 4: The difference of English stress and Japanese accent



Language	Acoustic features
English stress	strong-weak accent:related to intensity , height, duration of sound and the vowel quality
Japanese accent	High-low accent:basically only has relation with the height of sound

2.3 Typical errors of Japanese learners of English

2.3.1 Segmental errors

Influenced by Japanese mora-structure, vowel insertion changes the structure and amount of syllables are one of the most common error patterns among Japanese learners of English. For example, one syllable word “strike” is often pronounced by Japanese learners as [su/to/ra/i/ku], which has 5 moras or syllables. Such pronunciations totally destroy the original syllable structure of the words and thus cause misunderstanding when communicating with people who are not familiar with Japanese-style pronunciations of English. Table 5 shows some typical errors of vowel insertion.

2.3.2 Intonation and stress related errors

Intonation and stress are key prosodic factors in spoken language in terms of effective communication. Stress in the wrong positions and with wrong patterns can cause confusion or misunderstanding. Since Japanese accent are pitch accent, Japanese learners tend to emphasize important words by changing the pitch instead of the whole set of acoustic features that characterize stress in English. This often causes perceived errors. Incorrect phrasing can also cause stressing at wrong positions, which also results in misunderstanding.

2.4 Position of pronunciation education in English Learning

The goals of learning a foreign language are usually to acquire skills in reading, writing, listening and speaking. Recently, on top of acquisition of basic language skills, more and more efforts have been focused on improving learners’ communication skills in real world. This approach requires especially higher abilities in listening and speaking. CALL systems based on speech processing mainly support improving these two skills.

To evaluate learners’ communication abilities, there can be many factors used as measures. For example, Bachman has proposed a model of language competence to categorize different measures, which is shown in Figure 2.3 [16]. In Bachman’s model, pronunciation is one of the key factors of language competence

According to [17], a learner’s communication ability score can be defined as the following equation.

Table 5: Examples of vowel insertion errors

[i] is attached			
after	[tς]	-catch	[kætςi]
[u] is attached			
after	[l]	-pool	[pulu]
	[k]	-book	[buku]
	[g]	-egg	[Egu]
	[p]	-top	[tapu]
	[b]	-cab	[kæbu]
	[f]	-knife	[naifu]
	[v]	-have	[hævu]
	[z]	-prize	[praizu]
	[ς]	-mash	[mæςu]
	[θ]	-teeth	[tiθu]
[o] is attached			
after	[d]	-bread	[brEdo]
	[t]	-note	[nouto]

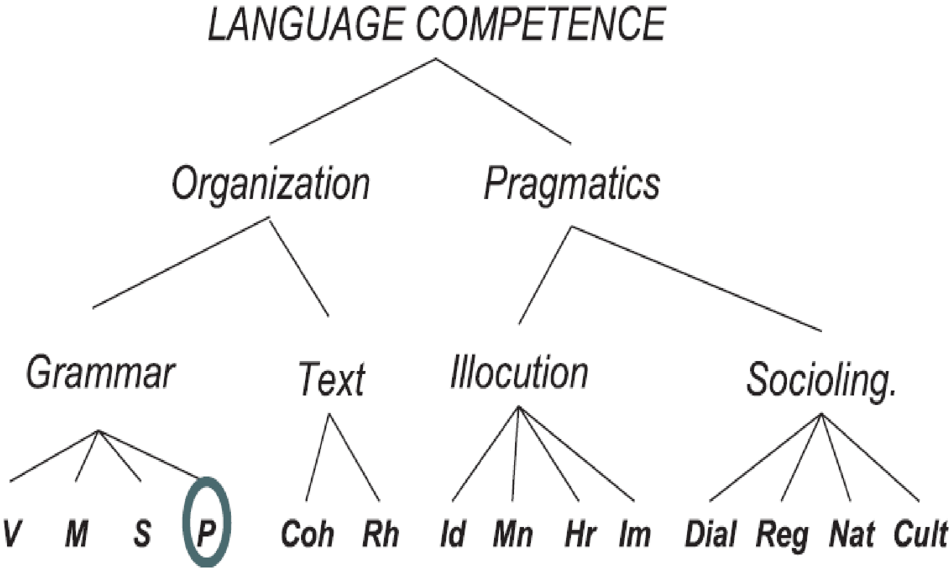


Figure 2.3: Bachman’s model

$$comm \approx pron. \bullet lex. \bullet (1 + syn. + rhet. + illoc. + soc.) \tag{2.1}$$

While indicates that pronunciation (pron.) and vocabulary (lex.) are the most important parts of communication. In terms of communication skill acquisition, pronunciation education and the technology that assist it is very importance in language learning.

2.5 Position of this study in pronunciation education

As mentioned in the previous section, new trends in language education focus more on communication and more and more language teachers and learners are beginning to use CALL for teaching and learning. However, most of the

conventional CALL systems are based on speaker-independent automatic speech recognition (ASR) which is much more reliable in recognizing native speech than foreign accented speech. On top of speaker characteristics, the diverseness of pronunciations causes mismatches between acoustic models and evaluation speech data and thus makes conventional pronunciation evaluation system unstable. In automatic speech recognition, speaker adaptation is widely used to reduce such mismatches by using a small amount of the target user's speech data. This study aims at finding a way to deal with these problems and propose methods for developing reliable CALL systems that can help with pronunciation education.

Since the purpose of CALL system development is to improve learners' communication abilities with target foreign language, we collaborate closely with language teachers to keep abreast with latest trend of language education and provide technological supports. Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages. Shadowing is a kind of "repeat-after-me" type exercise, but rather than waiting until the end of the phrase heard, learners are required to reproduce nearly at the same time. Although shadowing was originally designed to train simultaneous interpreters, its effects on foreign language learning have been widely recognized and being used in classrooms [18, 19, 20]. Studies show that in shadowing, speakers can hardly imitate the presented speech only, but use their own speech habits and language knowledge of their mother tongue unconsciously as well [21]. The adequate measurement of shadowed utterances can be a good indicator of the speaker's overall language proficiency.

Most existing works on automatic pronunciation scoring have been built on HMM-based speech recognition technologies. The HMMs were trained with native and/or non-native "read" speech samples. However, in shadowing, since learners have to follow the speaking rate of the input native utterance, the speaking style of the learners is very different from "read" speech. Especially in the case of beginners, the text content of the utterances generated through shadowing can be completely different from that of the presented ones. To the authors' knowledge, no automatic pronunciation scoring method has been proposed or investigated for shadowing.

This study proposes several methods for automatic scoring of shadowing and develops a shadowing evaluation system for pronunciation education classes. We compare shadowing with convention pronunciation practice and provide proof of the volatility of the advantage of shadowing over conventional practices from speech engineering point of views.

Chapter 3

Overview of CALL systems

The previous chapter introduces phonetic knowledge of Japanese and English and the importance of pronunciation education and Computer-Assisted Language Learning (CALL) systems. This chapter reviews the existing CALL systems and the technologies that support them.

3.1 Introduction

Improvements in information technology make it possible to turn computers into virtual tutors that allow learners to be able to receive training virtually anytime anywhere without a teacher at present. It can also saves teachers a lot of time and efforts in teaching a foreign language. Therefore, Computer-Assisted Language Learning (CALL) systems are becoming more and more popular in language teaching and learning. As a result, many research works have been done to improve these technologies for CALL.

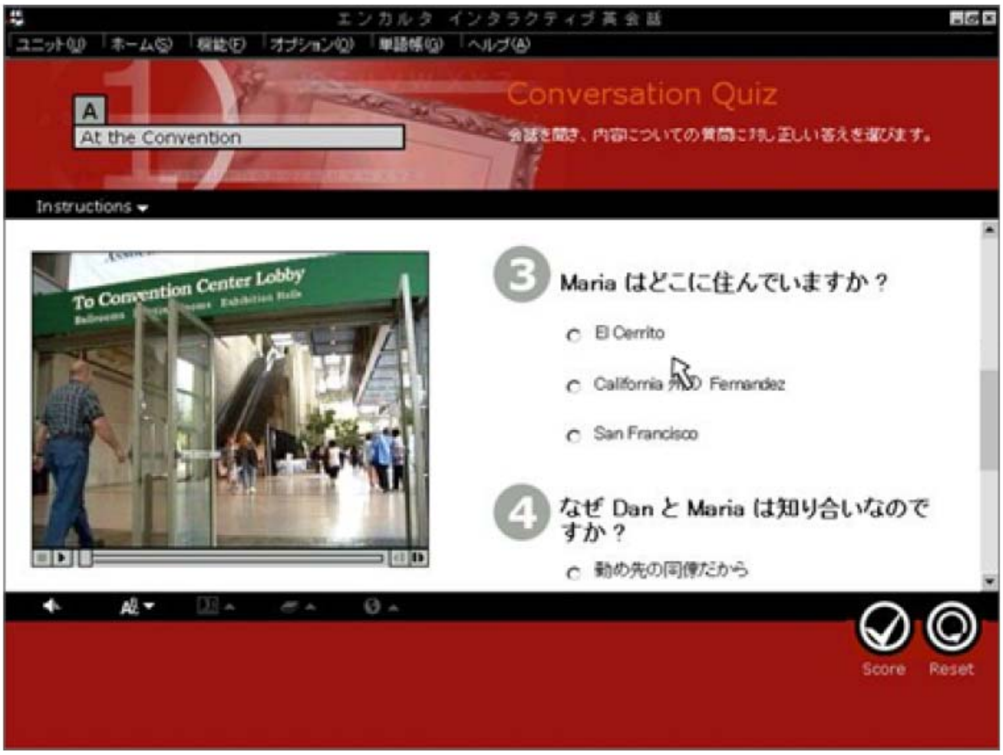
Basically, there are two kinds of CALL systems in the current market: one is purely utilizes multimedia as learning materials and provides courseware designed by human teachers for learners to access via computer; the other is utilizing automatic speech processing to assess learners' pronunciations, detect errors and provide diagnosis feedbacks. The formal one is straightforward and its implementation does not require advance techniques in speech processing. The performance and reliability of latter one highly depend on automatic speech processing techniques.

In order to provide technologies to build a reliable CALL system that can improve learners' communication abilities, we will examine more technical details of CALL systems that are based on speech processing, typically on automatic speech recognition (ASR).

3.2 CALL systems based on multimedia

Many of CALL software programs on the market utilize multimedia (text, audio, video) to provide predesigned exercises for students to learn a foreign language on computers. One example is Microsoft ENCARTA [22]. As shown in Figure 3.1, this software shows learners a short video clip of English conversion and then presents questions for learners to answer. Although it has some useful functions such as turning on or off the transcriptions while watching videos, playing audios of the questions, learners' pronunciations are not evaluated and thus no feedback of pronunciations are given.

As mentioned the Chapter 2, pronunciation skill is one of the key part of communication ability, so it is desirable for a CALL system to be able to diagnose learners' pronunciations and give proper feedbacks. In the following sections, such CALL systems will be reviewed and the technologies behind them will be closely examined.



At the Convention

Conversation Quiz

会話を聞き、内容についての質問に対し正しい答えを選びます。

3 Maria はどこに住んでいますか？

○ El Cerrito

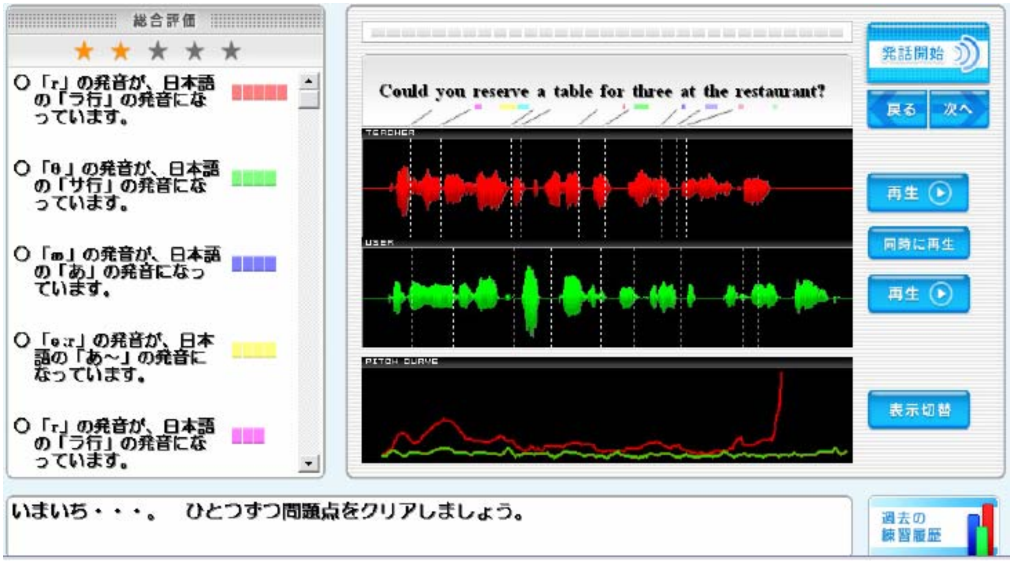
○ California 州の Fernandez

○ San Francisco

4 なぜ Dan と Maria は知り合いなのですか？

○ 勤め先の同僚だから

Figure 3.1: CALL based on multimedia: Microsoft ENCARTA



総合評価

★★★★★

○「r」の発音が、日本語の「ラ行」の発音になっています。

○「θ」の発音が、日本語の「サ行」の発音になっています。

○「æ」の発音が、日本語の「あ」の発音になっています。

○「ə:r」の発音が、日本語の「あ～」の発音になっています。

○「r」の発音が、日本語の「ラ行」の発音になっています。

Could you reserve a table for three at the restaurant?

いまいち・・・。　ひとつずつ問題点をクリアしましょう。

Figure 3.2: CALL based on speech processing: AmiVoice

3.3 CALL systems based on Automatic Speech Recognition (ASR)

Speaker-independent automatic speech recognition (ASR) was emerging in the early 1980s and improved significantly during the following decade. By the end of 1990s, ASR became the main language technology used for language learning systems. Here, after introducing the basics of ASR, I will explain how to apply techniques in ASR to computer-assisted language learning (CALL).

3.3.1 Basics of Automatic Speech Recognition (ASR)

Figure 3.3 shows mechanism of automatic speech recognition (ASR). From speech S, speech vector X of acoustic features that are extracted through acoustic analysis. In the statistical framework, the process of automatic speech recognition can be considered as a probability distribution function as $p(w|\mathbf{x})$. This probability function defined the probability distribution of the word sequence w ginve the input speech x. According the Bayes rule, the goal of ASR can be defined as,

$$\mathop{\arg mx}_{w} p(w|x) = \mathop{\arg mx}_{w} \frac{p(x|w)p(w)}{p(x)} = \mathop{\arg mx}_{w} p(x|w)p(w) \quad (3.1)$$

where $p(x|w)$ is called the acoustic model and $p(w)$ is called language model. In the case of CALL, acoustic models are used to judge how close acoustically a learner’s pronunciations are to native speakers’, and language models are used to

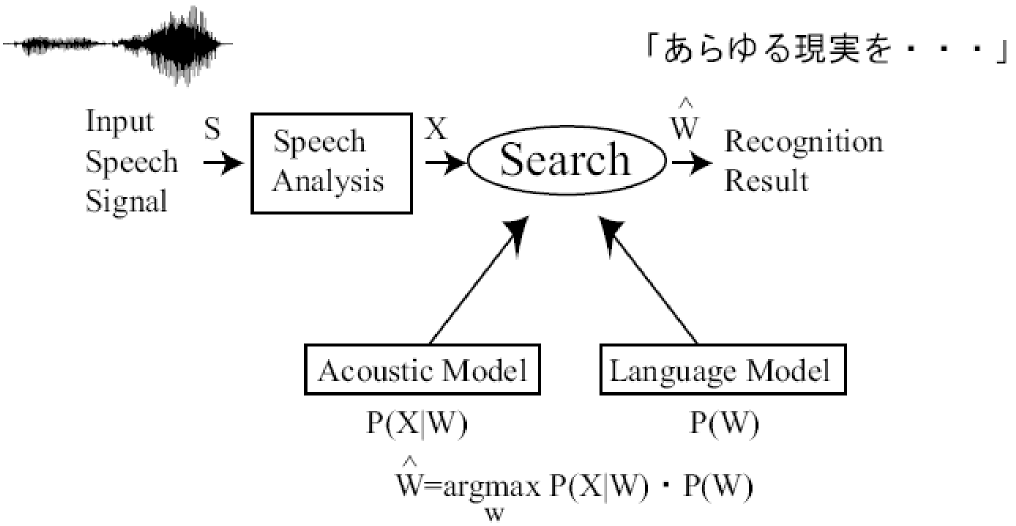


Figure 3.3:　Automatic speech recognition mechanism

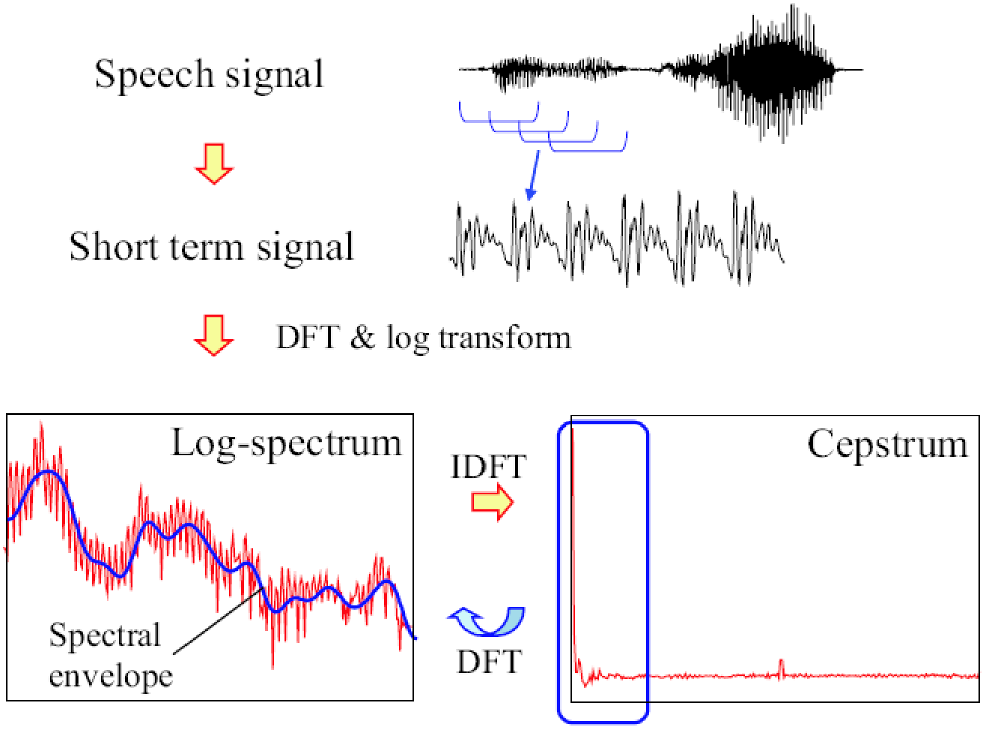


Figure 3.4:　Acoustic analysis of speech

predict the possibility of error patterns that might occur with that learner.

3.3.2 Acoustic model

As mentioned in the previous section, acoustic model, is represented by $P(\mathbf{X} \mid \mathbf{W})$, i.e, the probability of speech observance being **X** when the word **W** is spoken. Hidden Markov Model (HMM) as shown in Figure 3.5 is de facto standard for speech recognition. Here S_i represents the *i*-th state, a_i is the transition profanity from S_i to S_{i+1}, and $b_i(x)$ is the probability that output speech is generated from S_i. Probability density of $b_i(x)$ is often represented by Gaussian Mixture densities.

For CALL systems, speaker independent acoustic models of the target language are often used. These models are trained on vast amount of speech data from a large number of speakers for every phoneme of the target language. However the mismatches between native speakers and the learners often cause low speech recognition rate by using models trained in this way. Speaker adaptation that uses a relative small amount of data from a learner to reduce mismatches is often adopted.

3.3.3 Viterbi algorithm

The algorithm that calculates $p(\mathbf{X} \mid \mathbf{W})$ with HMM is called Viterbi algorithm. Here, considering each word correspond to each set of HMM (word HMM), Figure 3.6 depicts the possible state transition paths of outputting speech vector $\mathbf{X} = \{x(1), x(2), ..., x(7)\}$. The possibility of each path is calculated by the multiplication of a_i and every $b_i(x)$ in the path. Adding up all these probabilities yields the probability of output speech **X** coming from the HMM that represents the word **W**, i.e. $P(\mathbf{X} \mid \mathbf{W})$. However, in practice, instead of summing up all the probabilities, the probability of the path with maximum likelihood is calculated, which is called Viterbi algorithm.

In the case of continues speech recognition, each phone instead of word is corresponding to one set of HMM and be realized by connecting all the phone-HMM models together to represents $P(\mathbf{X} \mid \mathbf{W})$. The isolated phone-hmm is called monophone and the phone-hmm that takes into account of the effects of adjacent phones in the context of phonemes sequence is called triphone.

In CALL system implementation, Viterbi algorithm is often used to perform forced-alignment that identifies the location of each phoneme given the transcript and pronunciation dictionary of a given utterance. Therefore, a segmental local errors can detected at phoneme-level or a proficiency score can be given for each phoneme.

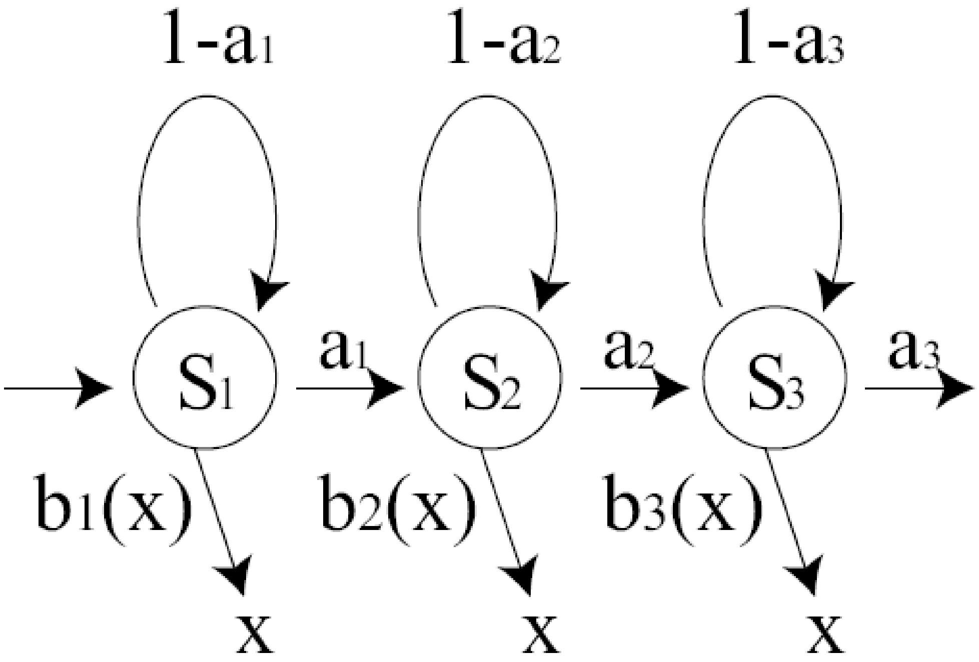


Figure 3.5: Hidden Markov Model (HMM)

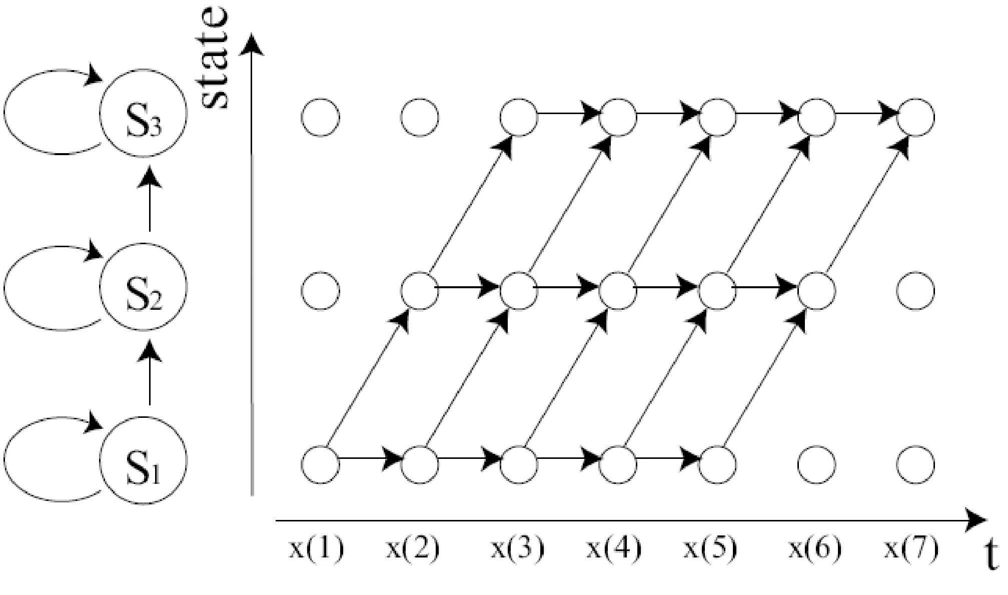


Figure 3.6: The Viterbi algorithm

3.3.4 Examples of CALL based on ASR

ASR is usually used to calculate phone-based likelihood as intelligibility measures. An automatic scoring system proposed by [23] is depicted in Figure 3.7. Likelihoods calculated with HMM acoustic models of native speech according to Vitabi algorithm and other speech features are combined by using a regression models to provide evaluation score for each learners. Although the combined scores show higher correlation with manual scores, the correlation by using the likelihood alone is only 0.36.

Silke Witt and Steve Young proposed a posterior probability score based scheme to detect phone-level mispronunciation [24]. The posterior probability score, or so-called Goodness of Pronunciation (GOP) score, is calculated by conducting forced-alignment and continuous phoneme recognition with unconstraint phone loop grammar as shown in Figure 3.8. Figure 3.9 shows that by presetting a threshold for phoneme-level GOP score, any phoneme that has a GOP score lower than the threshold will be judge as mispronunciation.

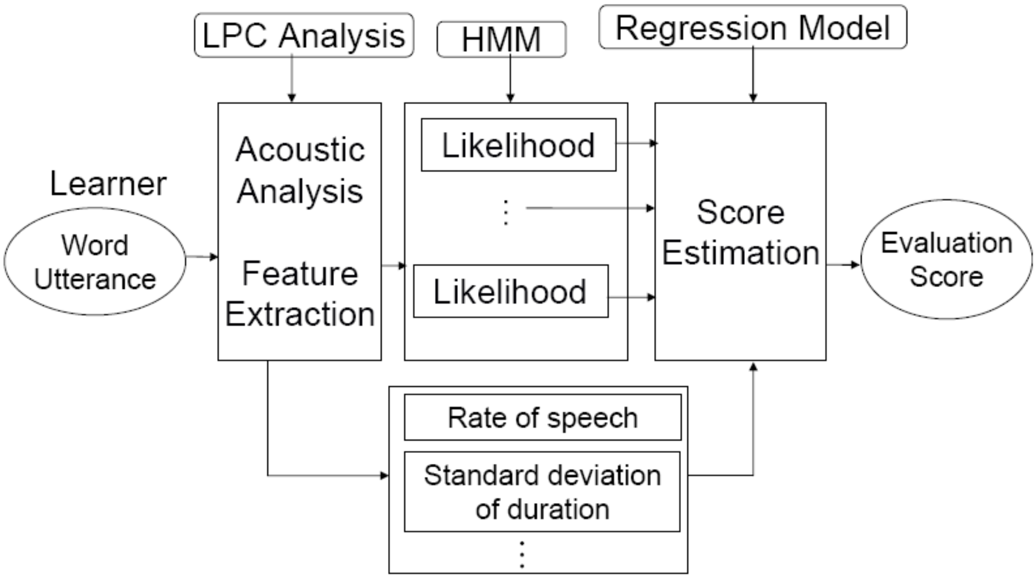


Figure 3.7: CALL based on ASR: automatic scoring

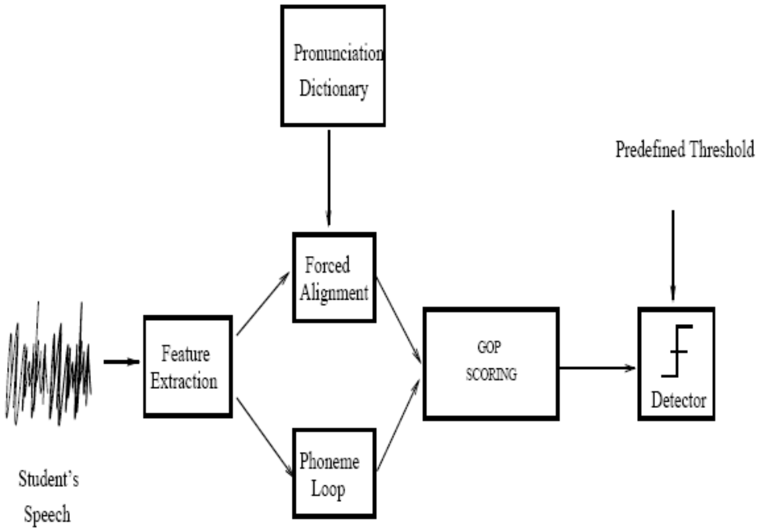


Figure 3.8: GOP scoring system

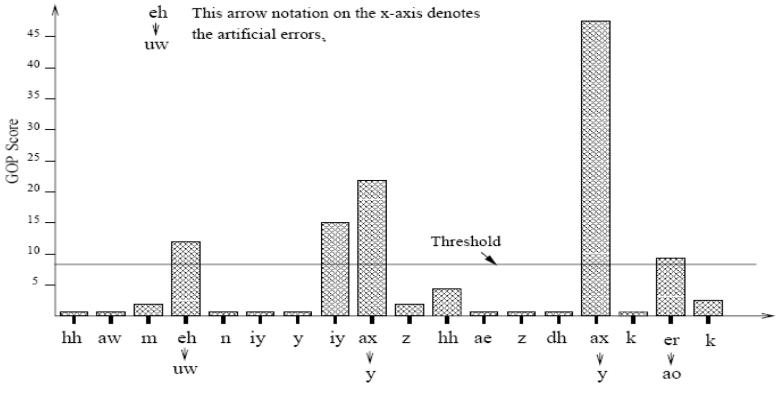


Figure 3.9: CALL based on ASR: error detection

3.4 Prosody evaluation

[30] proposed an automatic evaluation system of English prosody for Japanese learners. As shown in Figure 3.10, the system extracts rhythm and intonation features, then calculate prosodic scores by comparing with teachers' speech. In this case, forced-alignment based on ASR is conducted to detect word boundaries.

3.5 Conclusions

This chapter reviews various CALL systems the technologies that support it. Most of the CALL related researches utilize automatic speech recognition (ASR) for pronunciation evaluation. Although some prosodic features that are not directly used by HMM-based ASR are intergraded with scores derived from ASR to yield better results, ASR is still the core techniques for CALL system. Even for prosody evaluation, Viterbi algorithm in ASR is still used to locate the position of phonemes. Therefore, in order to improve reliability of CALL systems, improving the pronunciation evaluation techniques based on ASR is very important.

From the following chapters, I will focus on how to improve the ASR-based pronunciation evaluation techniques for the purpose of CALL.

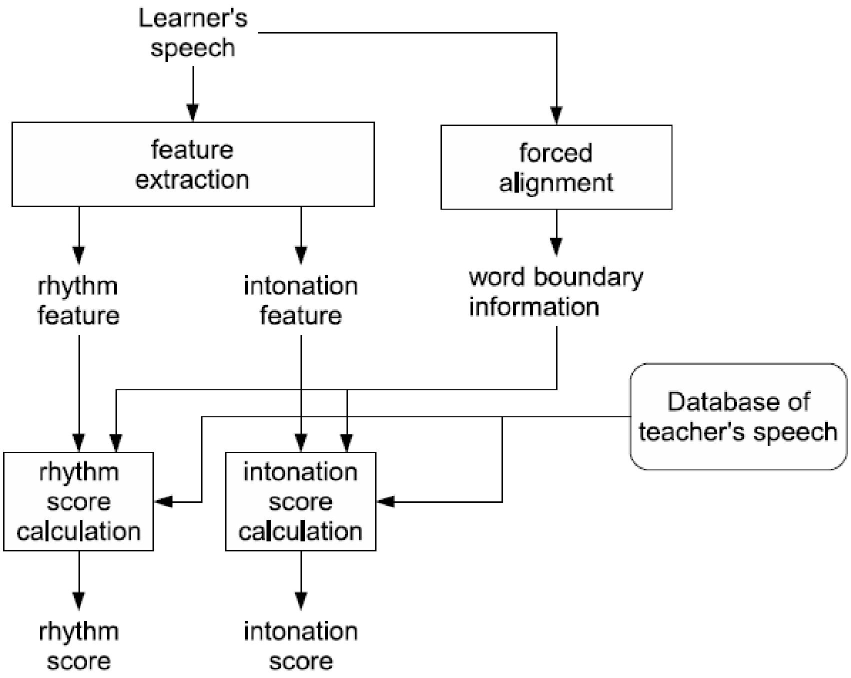


Figure 3.10: Prosody evaluation system

Chapter 4

Analysis of MLLR Adaptation for CALL

4.1 Introduction

As mentioned in the previous chapter, CALL systems usually use speaker-independent HMM acoustic models of target language to evaluate learners' pronunciations. However, each learner has his or her own speaker and linguistic characteristics in their speech. The differences of these characteristics between a specific learner and the native speakers whose speech data is used for training acoustic models cause mismatches that reduce the performance of phoneme recognition. In Automatic Speech Recognition (ASR), speaker adaptation is widely used to reduce such mismatches. These adaptation techniques often use a relatively small amount of a speaker's speech data to calculate transformation by yielding maximum effect of reducing the mismatches between the original acoustic models and adaptation data.

Although speaker adaptation has been proved very effective for ASR, problems occur when speaker adaption is directly applied to CALL for pronunciation evaluation. Instead of recognizing intended contents of the speech, the purposes of CALL are to evaluate goodness of pronunciations and detect errors. Since learners' pronunciations are not necessarily correct, there can be many errors in their speech. If a specific learner's adaptation data contains many errors, it can cause the over adaptation problem, in which case, erroneous pronunciations are adapted as correct. Although there are some studies use global Maximum Likelihood Linear Regression (MLLR) instead of local adaptation to avoid looking into too many details of pronunciation, to the author's best knowledge, no quantitative analysis has been reported to investigate the adverse effects of speaker adaptation.

In the following sections, I will first introduce a widely used adaptation technique, MLLR adaptation, and investigate how over-adaptation problem occurs by conducting two kinds of experiments: one is automatic scoring and the other is error detection. For automatic scoring, conventional Goodness of Pronunciation (GOP) scores and proposed forced-aligned GOP scores are used. The correlations between automatic scores and human scores are used to measure the performance of automatic scoring. For error detection, networkgramma-based and GOP-based schemes are adopted.

4.2 Maximum Likelihood Linear Regression (MLLR) Adaptation

4.2.1 Basic procedure of MLLR adaptation

Maximum Likelihood Linear Maximum (MLLR) estimates a set of transformations that reduce the mismatches between speaker-independent models and learners' data [31]. Usually, MLLR computes a set of transformations for the mean or variance parameters of a Gausssian mixture

HMM models. By shifting the component means or changing the variances in original models with these transformations, the adapted models are closer to speaker-dependent models. The transformation matrix that are applied to mean is given by

$$\hat{\boldsymbol{\mu}} = \mathbf{W}\boldsymbol{\xi} \quad , \qquad (4.1)$$

where **W** is the $n \times (n+1)$ transformation matrix, n is the dimensionality of the data, and ξ is the extended mean vector given by,

$$\xi = [\, w\,\mu_1\ \mu_2 \ldots\ \mu_n \,]^T \quad , \qquad (4.2)$$

where *w* represents a bias offset whose value is fixed at 1 with speech recognition tool kit HTK we used for evaluation.

Hence, transformation matrix ***W*** can be decomposed into

$$\mathbf{W} = [\,\mathbf{b}\,\mathbf{A}\,] \quad , \qquad (4.3)$$

where **A** represent an $n \times n$ matrix and **b** represents a bias vector.

4.2.2 Regression Classes

Since comparing with the model parameters, the amount of adaptation data is relatively little, these parameters are often clustered into regression classes. MLLR makes use of a regression class tree to group the Gaussian parameters so that the set of transformations to be estimated can be chosen according to the amount and type of adaptation data is available. The tying of the each transformation across a number of mixture components makes it possible to adapt distributions for which there were no observations at all.

Regression class tree construction that we used for evaluation experiment is implemented in the HTK tool kit [32]. This implementation is to cluster together components that are close in acoustic space, so that similar components can be transformed in a similar way. The tree is built with a centroid splitting algorithm, which uses a Euclidean distance measure.

Figure 4.2 shows a simple of a binary regression tree with four base classes $\{C_4, C_5, C_6, C_7\}$ as implemented with HRest tool in HTK [32]. The diagram shows a solid arrow and circle (or node), indicating that there is sufficient data for a transformation matrix to be generated using the data associated with that class. A dotted line and circle indicates that there is insufficient data. HTK uses a top-down approach to traverse the regression class tree. Here the search starts at the root node and progresses down the tree generating transforms only for those nodes which

1. have sufficient data and

2. are either terminal nodes (i.e. base classes) or have any children without sufficient data.

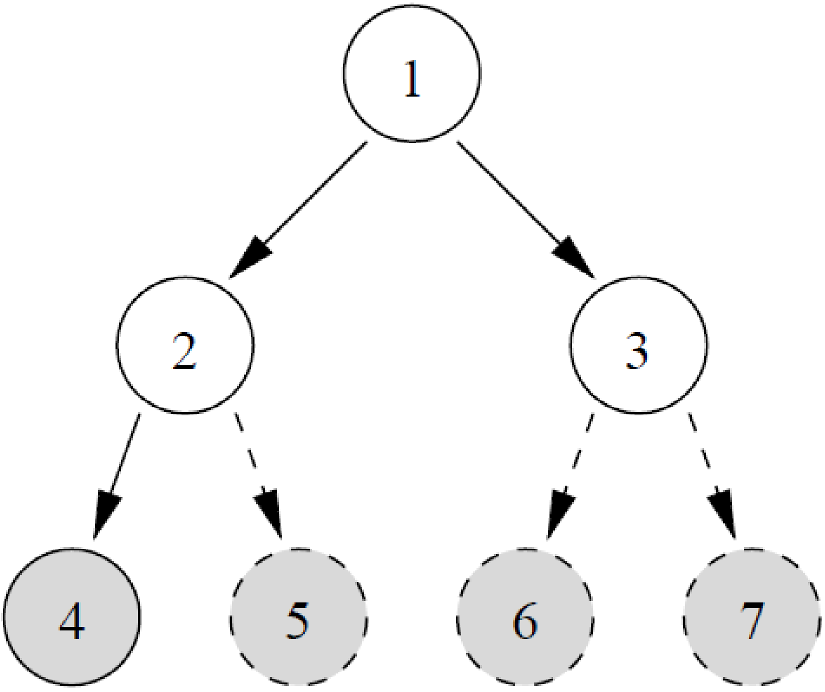


Figure 4.1: A binary regression tree implemented with HTK

$$\left\{ \begin{array}{ccc} \mathbf{W}_2 & \longrightarrow & \{C_5\} \\ \mathbf{W}_3 & \longrightarrow & \{C_6, C_7\} \\ \mathbf{W}_4 & \longrightarrow & \{C_4\} \end{array} \right\}$$

Figure 4.2: Transformation for base-class clusters

In the example shown in Figure 4.1, transforms are constructed only for regression nodes 2, 3 and 4, which can be denoted as W2, W3 and W4. Hence when the transformed model set is required, the transformation matrices (mean and variance) are applied in the fashion shown in Figure 4.2 to the Gaussian components in each base class.

4.2.3 Definition of MLLR

Consider M_r Gaussian components $\{m_1, m_2, ..., m_{M_r}\}$ that are tied together as decided by the regression class tree. The standard auxiliary function used to estimate the transforms is given by,

$$Q(M, \hat{M}) = \frac{1}{2}\sum_{r=1}^{R}\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t) \times$$
$$[K^{(m)} + \log\left|\hat{\Sigma}_{m_r}\right| + (o(t) - \hat{\mu}_{m_r})^T \hat{\Sigma}_{m_r}^{-1}(o(t) - \hat{\mu}_{m_r})]$$

(4.4)

where M is the HMM model set, $\hat{M}$ is the adapted model set, R is the number of the nodes of regression class tree, M_r is the number of Gaussian components that is to be tied together, $K^{(m)}$ subsumes all constants, $\hat{\mu}_{m_r}$ and $\hat{\Sigma}_{m_r}$ are the adapted mean vector and covariance matrix for the mixture component m_r respectively, and $L_{m_r}(t)$ is the occupation likelihood defined as

$$L_{m_r}(t) = p(q_{m_r}(t) \mid M, O_T) \quad , \qquad (4.5)$$

where $q_{m_r}(t)$ is the Gaussian component at time *t*, and O_T is the adaption data.

Here, we assume diagonal covariance matrices and the adaptation is only applied to the mean vector for each Gaussian component,

$$diag(\Sigma_{m_r}) = [\sigma^2_{m_r,1}, \sigma^2_{m_r,2}, \ldots, \sigma^2_{m_r,n}] \qquad (4.6)$$

$$\hat{\mu}_{m_r} = W_r \xi_{m_r} \qquad (4.7)$$

$$\hat{\Sigma}_{m_r} = \Sigma_{m_r} \qquad (4.8)$$

Substituting them into auxiliary function yields

$$\mathrm{Q}(\mathrm{M},\hat{\mathrm{M}}) = \mathrm{K} - \frac{1}{2}\sum_{r=1}^{R}\sum_{j=1}^{d}(w_{rj}G_r^{(j)}w_{rj}^T - 2w_{rj}k_r^{(j)T}) \qquad (4.9)$$

where w_{rj} is the j-th row of W_r,

$$\mathrm{G}_{\mathrm{r}}^{(\mathrm{i})} = \sum_{m_r=1}^{M_r}\frac{1}{\sigma_{m_r i}^2}\xi_{m_r}\xi_{m_r}^T\sum_{t=1}^{T}L_{m_r}(t) \qquad (4.10)$$

and

$$K_r^{(i)} = \sum_{m_r=1}^{M_r}\sum_{t=1}^{T}L_{m_r}(t)\frac{1}{\sigma_{m_r i}^2}o_i(t)\xi_{m_r}^T \qquad (4.11)$$

Differentiating $\mathrm{Q}(\mathrm{M},\hat{\mathrm{M}})$ with respect to the transform W_r, optimal transformation can be obtained by

$$w_{ri} = K_r^{(i)}G_r^{(i)-1} \qquad (4.12)$$

4.3 Pronunciation evaluation experiments with MLLR

To investigate the effects and side effects of conventional MLLR adaptation technique, automatic scoring and error detection were conducted on two public available databases. We examine the changes of performance while increasing the number of regression classes.

4.3.1 Acoustic models

The acoustic models we use for evaluation experiment are triphone HMM models train on TIMIT [33] and WSJ databases [34] with CMU pronunciation dictionary that includes a phoneme set of 39 phonemes. As acoustic features, 39-dimensional feature vectors, consisting of 12-dimensional MFCC, log-energy, and their first and second derivatives, were extracted from utterances using a 25 ms-length window shifted every 10 ms. The CMS (cepstral mean subtraction) was applied to each utterance unit. Each HMM has three output states with a left-to-right topology with self-loops and no transitions which skip over states.

For MLLR adaption, we use regression class trees described in Section 4.2.2 to cluster Gaussian components. The number of the nodes of regression tree increases from 1 to a certain number according to the amount of the adaptation data that are available.

4.3.2 Databases

For automatic scoring, we use English Read by Japanese (ERJ) database [35]. This database is constructed under the guidelines below:

1) The target language is General American

2) Speakers are Japanese students of universities or colleges.

3) Students read given words or sentences with phonemic/prosodic symbols. In addition to orthographical information, phonemic/prosodic information is give to subjects as text.

The text contents of the database are shown in Table 6. These sentences were divided into 8 groups and each subject read approximately 120 sentences.

This database contains proficiency scores manually rated by 4 experts who are native speakers of General American English and are familiar with Japanese pronunciation. 10 sentences of each subject were randomly chosen and each sentence was given a score by each expert and the average of the scores for the 10 sentences uttered by each subject is used as his or her score of indelibility score. We will use ERJ database for our automatic scoring evaluation Experiment. The inter-rater correlation of manual scores for 42 chosen learners are shown in Table 7.

Because the ERJ database does not contain phoneme labels with erroneous pronunciations, we use another corpus of English words spoken by Japanese students for our evaluation experiments of error detection. The database [36] consists of 5950 utterances of 850 basic English words read by seven Japanese speakers.

This database contains manually annotated phonemic labels that were faithfully transcribed and include erroneous phonemes. This database has been used to evaluate the performances of acoustic models for CALL [37].

We used the utterances of 4 speakers (2 males and 2 females) with many typical errors of Japanese learners. For each learner, 450 word utterances are used as adaptation data, and the remaining 400 utterances are used as test data.

Table 6: Sentence sets in ERJ in terms of segmental aspect of English Pronunciation

Set	Number of sentences
TIMIT-based phonemically-balanced sentences	460
Sentences including phoneme sequences that are difficult for Japanese to pronounce correctly	32
Sentences designed for test set	100

Table 7: Inter-ratter correlation of manual scores

	A	B	C	D
A	1	0.90	0.87	0.79
B		1	0.77	0.80
C			1	0.84
D				1

4.3.3 Automatic scoring with GOP scores

The confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for assessing speakers' articulation and shows good results. In this study, we use HMM acoustic models trained on WSJ and TIMIT corpus to calculate GOP scores defined as follows. For each acoustic segment $O^{(p)}$ of phoneme p, GOP($O^{(p)}$) is defined as posterior probability by the following log-likelihood ratio.

$$GOP(O^{(p)}) = \frac{1}{D_p}\log(P(p \mid O^{(p)})) \tag{4.13}$$

$$= \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)P(p)}{\sum_{q\in Q} P(O^{(p)} \mid q)P(q)}\right) \tag{4.14}$$

$$\approx \frac{1}{D_p}\log\left(\frac{P(O^{(p)} \mid p)}{\max_{q\in Q} P(O^{(p)} \mid q)}\right) , \tag{4.15}$$

where $P(p \mid O^{(p)})$ is the posterior probability that the speaker uttered phoneme p given $O^{(p)}$, Q is the full set of phonemes, and D_p is the duration of segment $O^{(p)}$. The numerator of equation (4.15) can be calculated by scores generated during the forced Viterbi alignment, and the denominator can be approximately attained by continuous phoneme recognition with an unconstrained phone loop grammar.

Since the boundaries of phoneme p yielded from forced alignment do not necessarily coincide with the boundaries of phoneme q resulted from continuous phoneme recognition, the frame average log likelihoods of the same speech segment are often used in traditional GOP calculation.

42 learners (21 males and 21 females) with higher agreement among raters and a variety of proficiency were selected. Average phoneme GOP score over 30 sentences read by each learner are calculated as automatic score for the learner. 60 sentence utterances of each leaner were used as adaptation data.

We investigate the correlations between GOP scores and human scores while increasing the number of the nodes of regression tree. Here the number 0 means without adaption, and 1 represents global adaption. As shown in Figure 1, global adaptation yielded the best correction of 0.65, yet while the number of nodes of regression class tree increases from 2, the performance drops. When the number is larger than 4, the correlation is even worse than the original models.

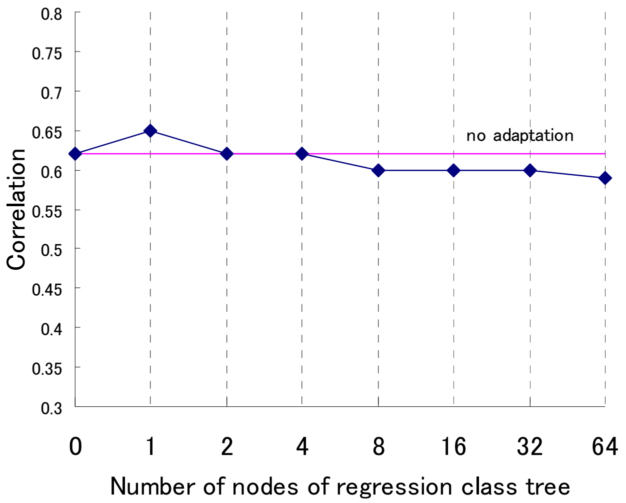


Figure 4.3:　Correlations between GOP scores and manual scores as the number of classes in MLLR increases

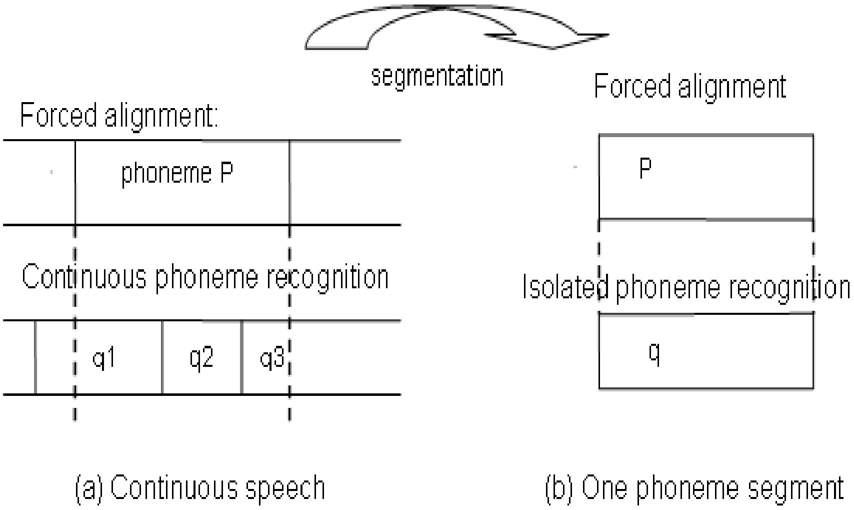


Figure 4.4:　Forced-aligned GOP method

4.3.4 Automatic scoring with forced-aligned GOP

Conventional GOP calculation refers to the results of both forced alignment and continuous phoneme recognition. This causes a problem as depicted in (a) of Figure 4.2, that there might be 3 phonemes resulting from continuous phoneme recognition, which correspond to one forced aligned phoneme p. In this case, GOP score for p is calculated using the log likelihood of p and average log likelihood of q1, q2 and q3 within the segment of p.

As an alternative way of calculating GOP score, we can first obtain the phoneme boundaries for phoneme p based on the result of forced alignment, and then calculate the posterior probability of that segment using equation (3) directly. We call this method Forced-aligned GOP (F-GOP). This method always refers to the boundaries of forced alignment and actually separates the calculation of GOP score into two processes, one is detecting the phoneme boundaries and the other is calculating the posterior probability for that segment. We can use different models for the two processes. We used the same data set as mentioned in Section 4.2.2 to evaluate the performance of F-GOP. We tested two different combinations of acoustic models for detecting phoneme boundaries and calculating posterior probabilities. Figure 4.5 shows the results of three conditions: F-GOP1, which used the same set of models for both phoneme boundary detection and posterior probability calculation, F-GOP2, which used the adapted models (the number of classes ≥ 1) to detect phoneme forced alignment boundaries, and the original models to calculate posterior probabilities, and the conventional GOP scores.

As shown in Figure 4.5, two kinds of F-GOP outperformed the conventional GOP. We consider this is because F-GOP did not refer to the results of continuous phoneme recognition that is often unreliable. Figure 4.6 shows an example of phoneme segmentation results of A) forced alignment, B) unsupervised bottom-up clustering and C) continuous phoneme recognition. In this example, the result of continuous phoneme recognition is even worse than segmentation based on unsupervised clustering, which uses no prior knowledge at all.

F-GOP2 shows better performance than F-GOP1, especially when the number of the nodes of regression trees is larger than 2. The only difference between F-GOP1 and F-GOP2 is that while F-GOP1 used the adapted acoustic models to calculate posterior probabilities, F-GOP2 used the original models to evaluate the same phoneme segment. This indicates that with more transforms used for adaption, the “judgment” of the acoustic model becomes worse. By utilizing the better phoneme alignment results, F-GOP can better benefit from speaker adaptation.

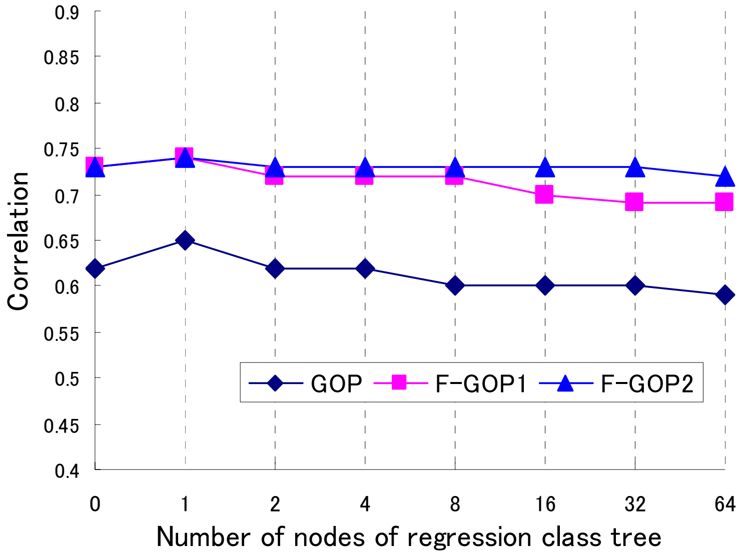


Figure 4.5:　Correlations between human scores and Forced-aligned GOP, comparing with conventional GOP

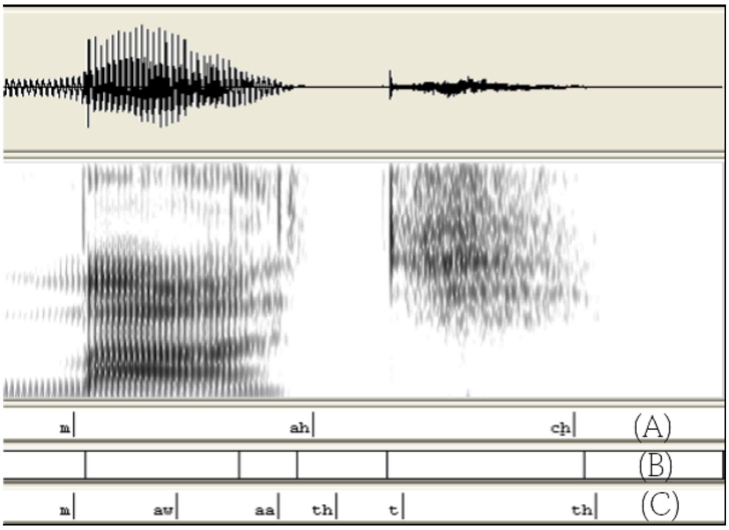


Figure 4.6:　Phoneme segmentation results, A) forced alignment, B) unsupervised bottom-up clustering, C) continuous phoneme recognition

4.3.5Error detection based on network grammar

The first method we use to detect pronunciation errors is using pronunciation networks that include correct pronunciations and various error patterns to predict learners' possible mispronunciations. These pronunciation networks are often called network grammars. An Example of network grammar is shown in Figure 5.4. The network grammar predicts 4 possible errors that might occur when a Japanese learner utter English word "grid": inserting /uh/ after /g/, substituting /r/ with /l/, substituting /ih/ with /iy/, and inserting /uh/ after /d/. Any combination of these 4 possible errors can be detected according to the acoustic scores calculated with HMM models. By referring to [38] and [39], 12 major error patterns shown in Table 7 were defined and any irregular errors in the labels were added to the prediction networks. Although the error detection performance highly depends on pronunciation networks and a larger network often results in lower detection precision, when the same network is used, the relative increase or decrease of detection accuracy can be used to measure the performances of the acoustic models with MLLR adaptation. In actually implementation, pronunciation network can be fine-tuned according to the proficiency levels of the learners and their error tendencies so that an optimal network can be constructed to yield best error detection results.

We used precision and recall rates defined as below to measure the performance of acoustic models with MLLR.

$$\text{Precision} = \frac{N_{hit}}{N_{total}} = \frac{N_{hit}}{N_{hit} + N_{FR}} \qquad (4)$$

$$\text{Recall} = \frac{N_{hit}}{N_{labeled}} \quad , \qquad (5)$$

where N_{hit} represents the number of the errors that were correctly detected , N_{total} is the total number of detected errors, N_{FR} is the number of false rejections (i.e. correct pronunciation falsely recognized as errors) and $N_{labeled}$ is the number of all the errors that were detected by phoneticians, and F-measure defined as below is also calculated to combine the two measures.

$$\text{F} - \text{measure} = \frac{2\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \qquad (6)$$

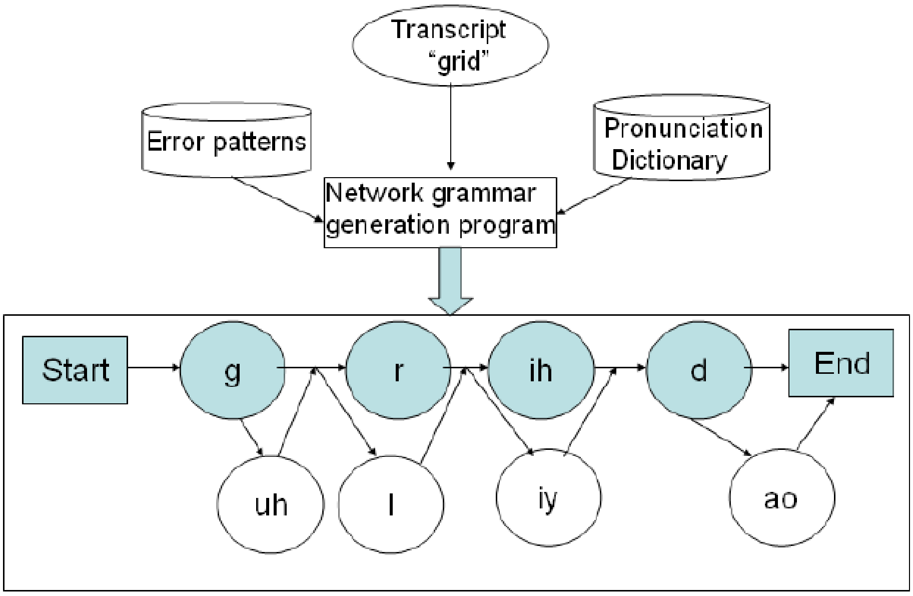
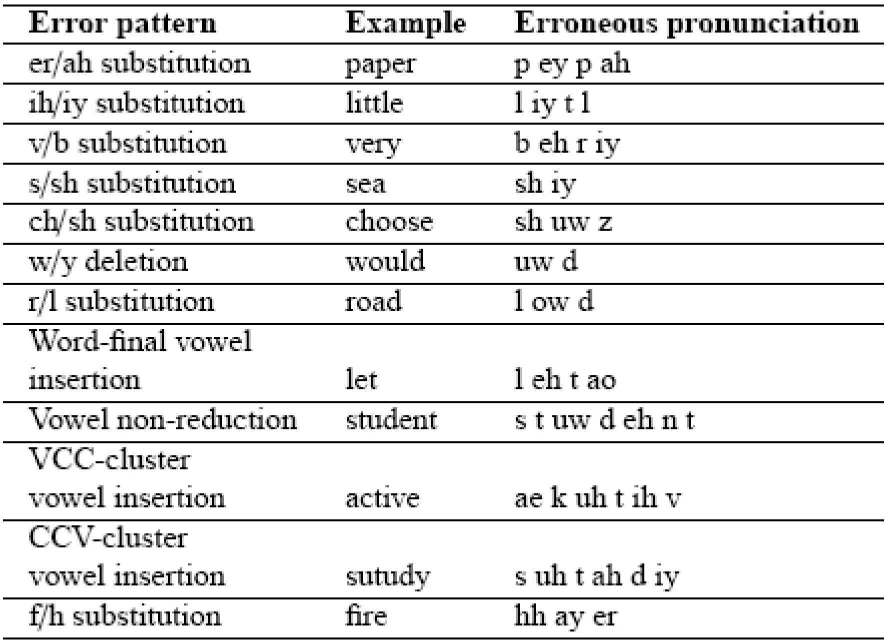


Figure 4.7: An example of network grammar

Table 8: 12 basic error patterns for constructing network grammars.



Error pattern	Example	Erroneous pronunciation
er/ah substitution	paper	p ey p ah
ih/iy substitution	little	l iy t l
v/b substitution	very	b eh r iy
s/sh substitution	sea	sh iy
ch/sh substitution	choose	sh uw z
w/y deletion	would	uw d
r/l substitution	road	l ow d
Word-final vowel insertion	let	l eh t ao
Vowel non-reduction	student	s t uw d eh n t
VCC-cluster vowel insertion	active	ae k uh t ih v
CCV-cluster vowel insertion	sutudy	s uh t ah d iy
f/h substitution	fire	hh ay er

Figure 4.6 shows the performances of error detection with MLLR adaption. Although the precision rates keep increasing when more transforms were used for adaptation, the recall rates drop when the number of nodes is larger than 2. This indicates that with adaptation to reduce model mismatches, the number of false rejections N_{FR} drops significantly, therefore the precision rates increase. However, since the number of $N_{labeled}$ is only decided by the label, the decrease of recall means the decrease of the number of correctly detected errors. This result shows that over adaption can cause more errors to be recognized as correct pronunciation (i.e. N_{hit} decreases), yet at the same time, even with over-adaptation, more false rejections can be reduced. How to benefit from reducing N_{FR} and preventing decreasing N_{hit} is goal of our research and we will provide a novel solution to achieve such goal in the next chapter.

4.3.6 Error detection based on GOP scores

Two most popular methods of error detection are employed for our phoneme error detection experiments: one is based on pronunciation networks or so-called network grammar and the other is based on GOP scores. The former method predicts possible error patterns and thus is able to detect specified types of errors such as phoneme-level substitution, deletion or insertion. However, the detection performance is largely depending on the size of the pronunciation networks. The latter method often uses a pre-set threshold to determine whether a phoneme is correctly pronounced or not. Although this method cannot specify the type of an error that occurs, by choosing the optimal threshold for each phoneme, much better detection performance can be obtained.

For the error detection method based on GOP scores, the recall and precision can be adjusted by changing the values of the thresholds. According to [40], erroneously rejecting correct pronunciations would be more detrimental for learners than erroneously accepting mispronunciations. Therefore, we need to keep the false rejection rate at relatively low level, which means to keep the precision relatively high, and find the optimal thresholds that maximize the recall. Here, we investigate the change of recalls at precision level of 70% while increasing the number of regression classes for MLLR and Regularized-MLLR. Here, the number 0 means no adaptation, i.e. using the original acoustic models.

As shown in Figure 4.9, in the case of MLLR adaptation, only global adaption shows slight improvement over original models and when the number of regression classes is larger than 2, the performance drops significantly. This clearly indicates that over-adaptation occurs with MLLR.

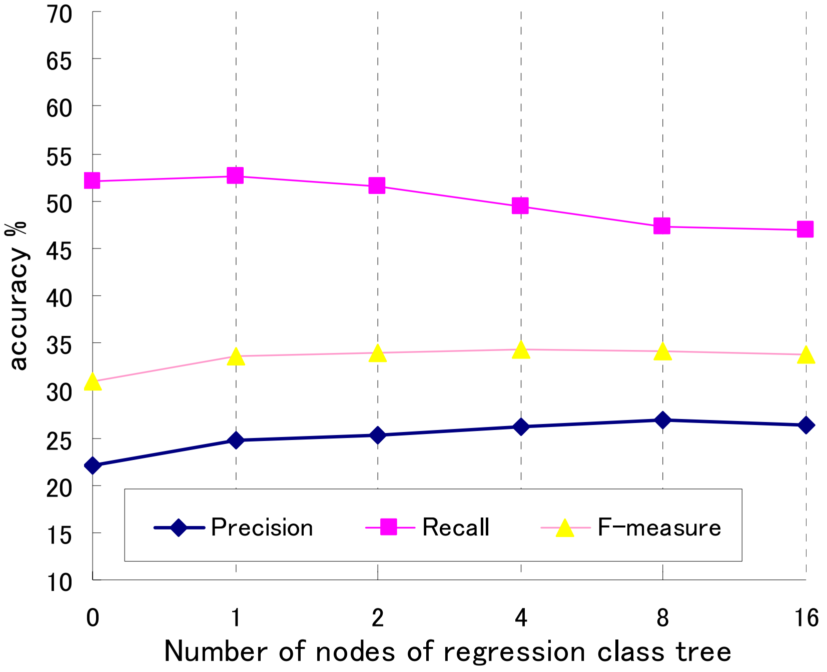


Figure 4.8:　The performances of error detection based on pronunciation networks

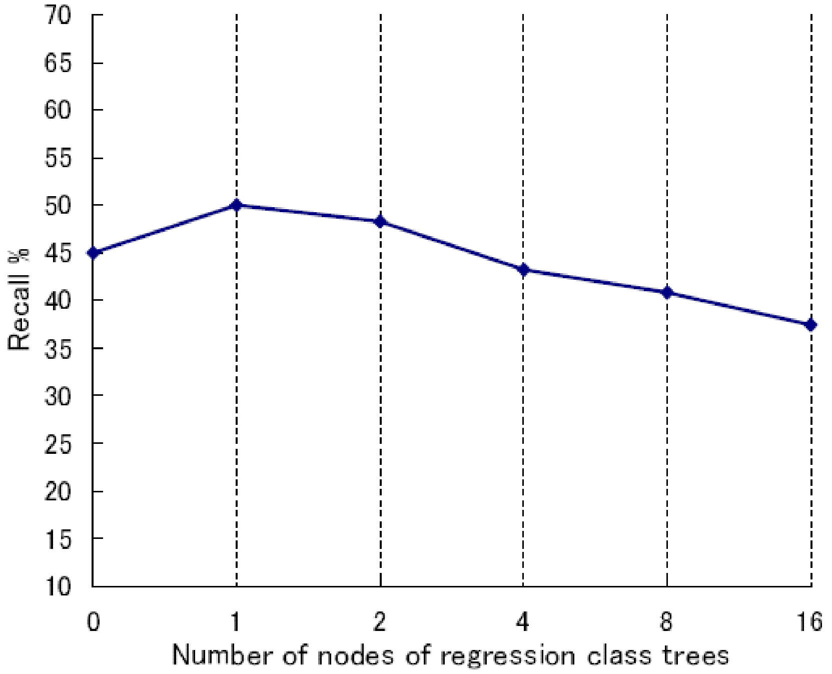


Figure 4.9:　Recall at the precision level of 70% (based on GOP)

4.4 Conclusions

In this chapter, we investigated the effects and side effects of conventional MLLR speaker adaptation technique on pronunciation evaluation for CALL systems. Automatic scoring and error detection experiments were conducted while increasing the number of regression classes of MLLR. For automatic scoring, GOP scores and Forced-aligned GOP scores were used as automatic scores, and correlations between automatic scores and manually rated scores were investigated. For error detection, network grammar and GOP-base schemes were adopted for evaluation experiments.

We first introduced the basic concept of speaker adaptation and the definition of conventional Maximum Likelihood Linear Regression (MLLR) adaptation. We then conducted pronunciation evaluation experiments on publicly available databases in two ways: one is automatic scoring and the other is error detection. We investigated the performances of automatic scoring and error detection with MLLR adaption by increasing the number of regression classes.

Experimental results shows that global MLLR adaption (the number of regression classes is one) slightly improves performances comparing with the original models. However, when the number of regression classes is larger than 2, over-adaption occurs. These results indicate that when too many details of the pronunciations are being looking into during adaptation, learners' erroneous pronunciations can be adapted as good ones. In order to fully utilized the benefits of speaker adaptation and solve the over-adaption problem, some kind of constraint needs to be added to the conventional adaption, so that only mismatches that are caused by speaker characteristics are adapted while at the same time, the transformation of wrongly pronounced pronunciations into good ones will be prevented. In other words, the transformation of adaption needs to be regularized to yield best performances for the purposes of CALL. In the following chapter, we will introduce such regularization of MLLR adaption as solution to over-adaption problem.

Chapter 5

Regularized-MLLR Adaptation for CALL

5.1 Introduction

In the previous chapter, we have analyzed the effects and side effects of conventional MLLR adaptation when applied to pronunciation evaluation. It is very clear that over-adaption can cause erroneous pronunciations being adapted as good ones.

Based on the analysis results, we provide solutions to the over-adaption problem. Since the reason that causes the over-adaptation problem is that conventional MLLR adaptation using learners' imperfect pronunciation as adaptation data and if there are too many errors in the adaptation data, those errors would be transformed as good pronunciations. Therefore, if we can prevent such transformations that erroneous pronunciation being transforming into good ones, we can prevent over-adaption.

To regularized pronunciation transformation during speaker adaptation, we use a group of teachers' data to calculate each teacher's transformation matrix with MLLR, and then use the teachers' matrices to regularized learners' transformation. We refer to this method as Regularized-MLLR and implement it in two ways: one is use the average of the teachers' matrices as a constraint to the conventional MLLR objective function, and the other is using a linear combination of the teachers' matrices to represent each target learners' matrices. Experimental results show the high validity of the proposed methods.

5.2 The first implementation of Regularized-MLLR adaption

In order to regularize MLLR transformation so that the erroneous pronunciation will not be "transformed" to good pronunciation, we add constraints to conventional MLLR.

The standard auxiliary function for MLLR is defined as below to estimate the transform W_r for each regression class r.

$$Q(M, \hat{M}) = \frac{1}{2}\sum_{r=1}^{R} \sum_{m_r=1}^{M_r} \sum_{t=1}^{T} L_{m_r}(t) \times \left[K^{(m)} + \log\left|\hat{\Sigma}_{m_r}\right| + (o(t) - \hat{\mu}_{m_r})^T \hat{\Sigma}_{m_r}^{-1}(o(t) - \hat{\mu}_{m_r})\right] \tag{5.1}$$

where M is the HMM model set, $\hat{M}$ is the adapted model set, and R is the number of the nodes of regression class tree, M_r is the number of Gaussian components that is to be tied together, $K^{(m)}$ subsumes all constants, and $L_{m_r}(t)$ is the occupation likelihood defined as

$$L_{m_r}(t) = p(q_{m_r}(t) \mid M, O_T) \quad , \tag{5.2}$$

where $q_{m_r}(t)$ is the Gaussian component at time t, and O_T is the adaption data.

Here we obtained a set of transforms estimated from a group of teachers who are native speakers of General English. Teachers' transforms are used to constrain the transforms for the learners to avoid bad pronunciation being transformed into good pronunciation.

Let $\{W_r^{C_1}, \ldots, W_r^{C_k}\}$ denote a set of transformation matrices estimated from a group of K teachers, and $W_r^C = \frac{1}{K}\sum_k W_r^{C_k}$ represents the mean of these matrices. The objective function for Regularized-MLLR is defined as

$$\max_{W_r}\{Q(M,\hat{M}) - \lambda\sum_{r=1}^{R} \left\|W_r - W_r^C\right\|_F^2\} \quad , \tag{5.3}$$

where λ is a parameter depending on the acoustic characteristics of the speaker.

We assume diagonal covariance matrices and the adaptation is only applied to the mean vector for each Gaussian component,

$$\hat{\mu}_{m_r} = W_r \xi_{m_r} \quad , \tag{5.4}$$

where ξ_{m_r} is the extended mean vector for the Gaussian component m_r.

Considering the row decomposition $W_r = [w_{r,1}; w_{r,2}; \ldots; w_{r,d}]$, the cost function for each row vector becomes,

$$f(w_{r,j}) = K_j + w_{r,j} H_r^j w_{r,j}^T - 2 w_{r,j} N_r^{(j)T} \quad , \tag{5.5}$$

where

$$H_i^j = \sum_{m_r=1}^{M_r} \frac{1}{\sigma_{m_{r,j}}^2} \xi_{m_r} \xi_{m_r}^T \sum_{t=1}^{T} L_{m_r}(t) - \lambda I \tag{5.6}$$

$$N_r^j = \sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)\frac{1}{\sigma_{m_{r,j}}^2} o_j(t)\xi_{m_r}^T - \lambda w_{r,j}^C \tag{5.7}$$

The optimal $w_{r,1}$ is given by solving

$$\frac{\partial f(w_{r,j})}{\partial w_{r,j}} = 0 \quad , \tag{5.8}$$

which yields,

$$w_{r,j} = N_r^j (H_r^j)^{-1} \tag{5.9}$$

We will refer to this method as R-MLLR1 here after.

5.3 The second implementation of Regularized-MLLR

R-MLLR1 uses the average of a group of teachers' transformations as a constraint adding to convention MLLR. The scale of that constraint which is decided by the parameter λ, needs to be manually adjusted for each learner according to his or her acoustic characteristics. When the number of learners is very large, it can be very time-consuming to find an optimal parameter for each learner. Therefore, we try another approach that can automatically estimate optimal parameters for different learners.

In the second regularization, we assume a learner's transformation matrix W_r can be represented as the linear combination of a group of N teachers' transformation matrices $\{W_r^{C_1},...,W_r^{C_N}\}$,

$$W_r = \sum_n \alpha_n W_r^{C_n} \tag{5.10}$$

Then the objective function becomes,

$$\max_{\{\alpha_k\}} g(\alpha_1,\alpha_2,...,\alpha_N) = \sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)(o(t)-\sum_n \alpha_n W_r^{C_n}\xi_{m_r})^T \Sigma_{m_r}^{-1} \times \\ (o(t)-\sum_n \alpha_n W_r^{C_n}\xi_{m_r}) \tag{5.11}$$

By calculating the derivative,

$$\frac{\partial g}{\partial \alpha_k} = -2\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)\Sigma_{m_r}{}^{-1}(o(t)-\sum_n \alpha_n W_r^{C_n}\xi_{m_r})(W_r^{C_n}\xi_{m_r})^T$$

$$= 0 \quad , \qquad (5.12)$$

and changing $k = 1,2,...,N$, we have N linear equations on $\{\alpha_n\}$. For simplicity, if we set

$$\xi'_{m_r,n} = W_r^{C_n}\xi_{m_r} , \qquad (5.13)$$

then the linear equations become,

$$\sum_{m_r=1}^{M_r}\sum_{t=1}^{T} L_{m_r}(t)\Sigma_{m_r}{}^{-1}(o(t)-\sum_n \alpha_k \xi'_{m_r,n})\xi'^{T}_{m_r,n} = 0 \quad . \qquad (5.14)$$

By solving these linear equations, we obtain the optimal $\{\alpha_n\}$. Then we can use equation (5.10) to calculate the target learner’s transformation matrix.

We will refer to this implementation of Regularized-MLLR as R-MLLR2 hereafter.

5.4 Evaluation experiments

In order to prove the validity of our proposed methods, we directly compare the effects of R-MLLR1, R-MLLR2 and MLLR on automatic scoring and error detection. The databases and experiment conditions are the same as investigation of adverse effects of MLLR mentioned in the previous chapter. To regularize MLLR transformation, we use 20 teachers' utterances from ERJ database.

These teachers are native speakers of General American English. 60 sentence utterances of each teacher are used to calculate his or her transformation matrices.

5.4.1 Automatic scoring results

We apply transformations estimated with R-MLLR1 and R-MLL2 to original HMM models by increasing the nodes of regression trees from 1 to 64. Then we use the adapted models to calculate average phoneme GOP score for each learner. The correlations between GOP scores and manual scores with adapted models are shown in Figure 5.1. The learners and the amount of evaluation data and adaptation data are the same as the previous chapter.

As shown in Figure 5.1, R-MLLR1 and R-MLLR2 show better performance than conventional MLLR. When the number of regression classes increases after 1 (global adaptation), the effect of regularization becomes rather obvious. Although by adding some amount of constraints, R-MLLR1 reduces the adverse effects of over-adaptation, the performance still drops when the number of classes is larger than 1. In the case of R-MLLR2, it not only always shows the best results, the performance never drops. This can be explained that in the case of R-MLLR2, the direct use of learners' transformations estimated by their imperfect pronunciations with MLLR is avoided. However, in the case of R-MLLR1, these transformations are still used and since there is not sufficient labeled data for each learner, the constraint scale parameter λ manually chosen for each of the 42 learners might not be optimal to yield best results.

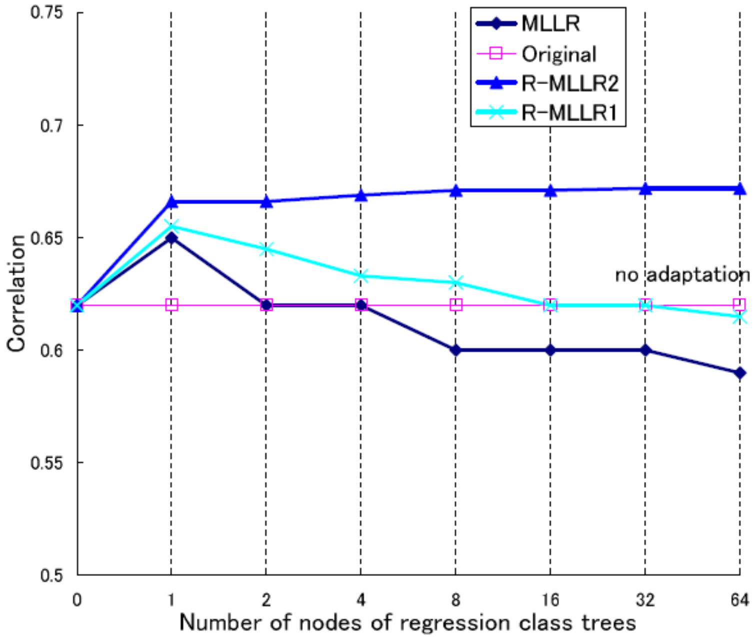


Figure 5.1: Correlations between GOP scores and manual scores as the number of classes increases

5.4.2 Results of Error detection based on network grammar

We apply transformations estimated with R-MLLR1 and R-MLLR2 to original HMM models by increasing the nodes of regression trees from 1 to 16. Then we use the adapted models to perform error detection experiments with network grammar and also with GOP scores. Experimental setups are the same as error detection experiments with MLLR adaptation mention in the previous chapter. For regularization, the same 20 teachers' utterances from the ERJ database are used to estimate their transformation in the same way as automatic scoring experiments with R-MLLR1 and R-MLLR2.

The results of error detection based on network grammar with comparison of any two of the three adaptation methods are shown in Figure 5.2, Figure 5.3, and Figure 5.4, respectively.

As shown in Figure 5.2 and Figure 5.3, R-MLLR1 and R-MLLR2 improve recall significantly and also keep very high precision rates. Especially in the case of R-MLLR2, recall keeps high level when the number of classes increases. This indicates the proposed method not only benefits from reduction of mismatches (increase of precision) but also prevents over-adaptation.

The comparison of R-MLLR1 and R-MLL2 is shown in Figure 5.4. When the number of regression classes is larger than 4, in the case of R-MLLR1, the performance of recall drops, which indicates the over-adaptation problem still occurs. This problem is solved by R-MLL2, and the performances of precision of R-MLLR1 and R-MLLR2 are almost the same.

5.4.3 Results of Error detection based on GOP scores

Figure 5.5 shows the performances of recall at the precision level of 70%. R-MLLR1 improves the performances comparing with conventional MLLR, however, the performance drops when the number of regession classes is larger than 2, i.e. over-adaptation problem remains. R-MLLR2 outperforms MLLR global adaptation or R-MLLR1, especially when the number of regression classes becomes larger. In the case of R-MLLR2, recall rate never drops, which again shows that this method can avoid the over-adaption problems by using linear combination of teachers' MLLR transformations instead of their own transformation matrices.

5.5 Conclusions

In this chapter, we implement two forms of Regularized-MLLR, R-MLLR1 and R-MLLR2, by using teachers' perfect pronunciations to regularize learners' transformations. R-MLLR1 uses the average of a group of teachers' transformation matrices as a constraint adding to the conventional MLLR transformations. This constraint prevents radical transformations when there are too many errors in the adaptation data. R-MLL2 uses linear combination of the

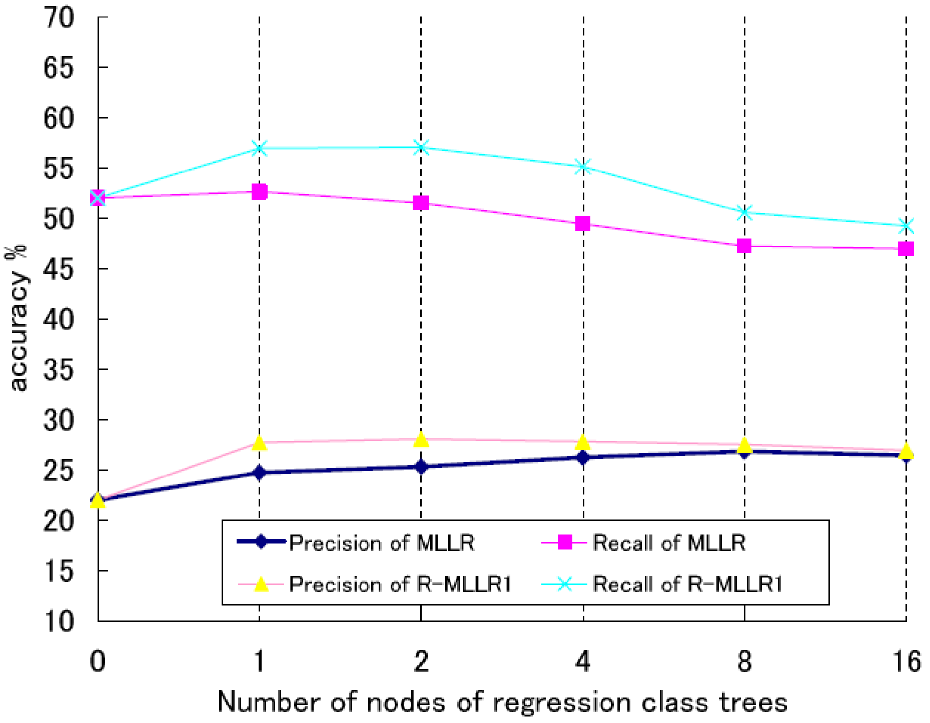


Figure 5.2: Correlations between GOP scores and manual scores as the number of classes increases

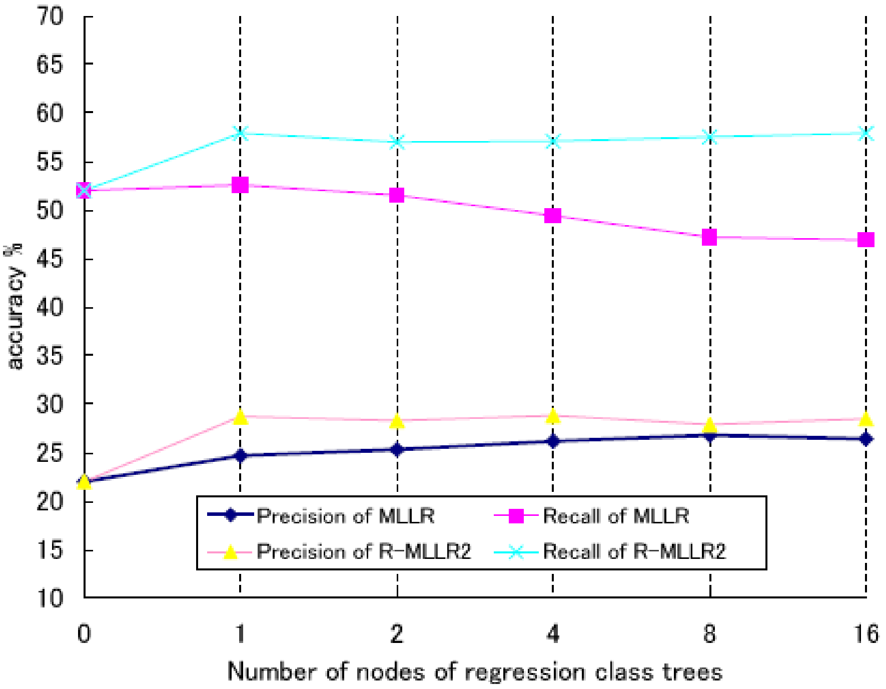


Figure 5.3: Correlations between GOP scores and manual scores as the number of classes increases

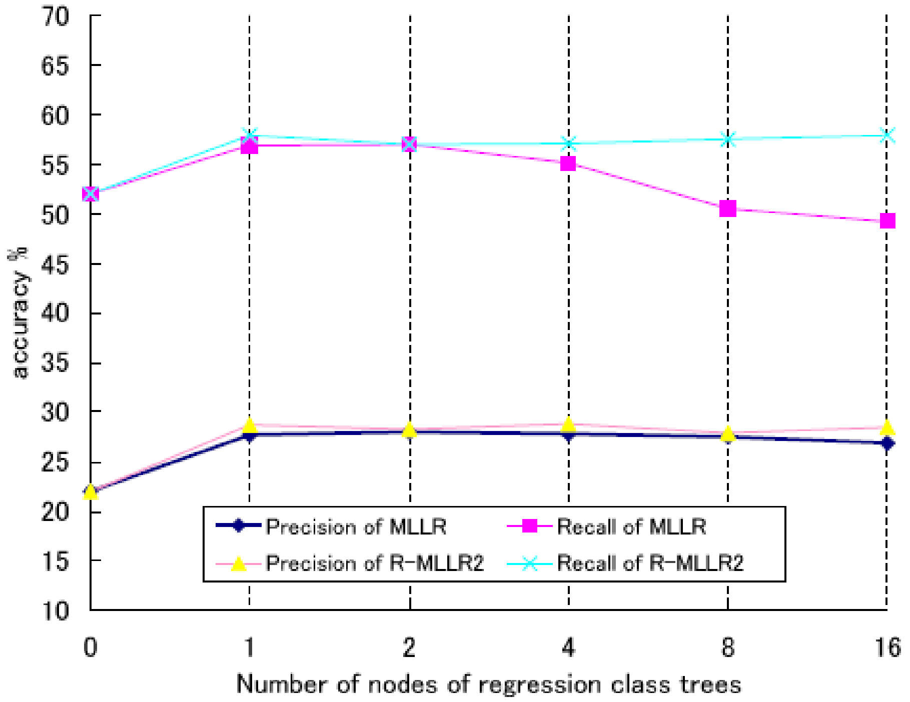


Figure 5.4: Correlations between GOP scores and manual scores as the number of classes increases

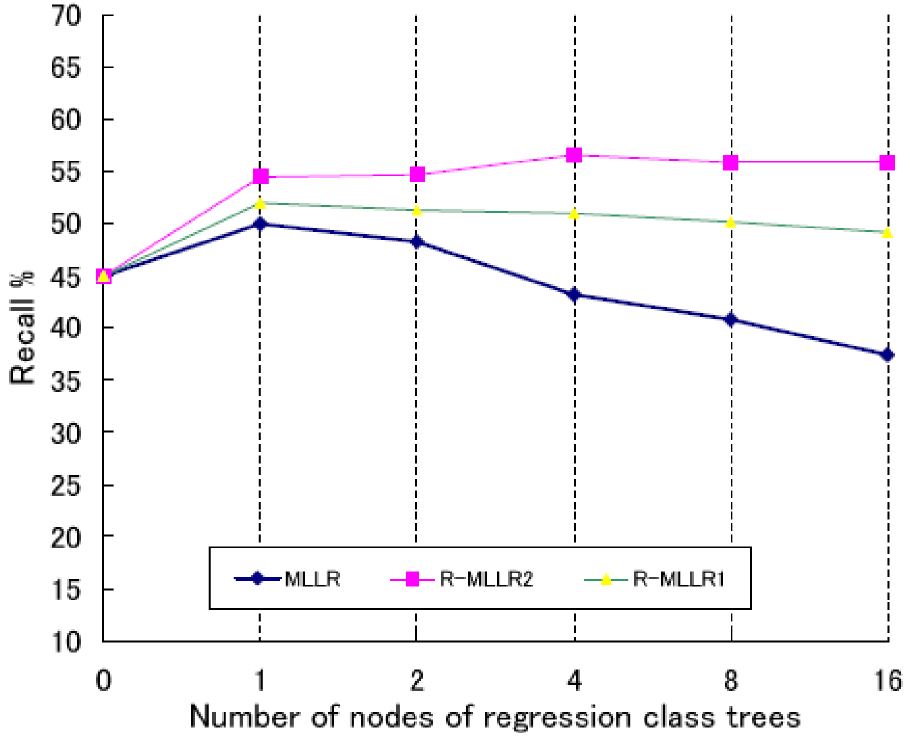


Figure 5.5: Correlations between GOP scores and manual scores as the number of classes increases

teachers' MLLR transformation matrices to represent each learner's transformation. This approach does not directly use learners' MLLR transformations that are estimated from their imperfect pronunciations, therefore prevents over-adaption. We compare R-MLLR1 and R-MLLR2 with conventional MLLR by conducting experiments on the same conditions as we investigate the adverse effects of MLLR on pronunciation evaluation of automatic scoring and error detection. Experimental results show that the proposed methods outperform conventional MLLR.

By adding constraints to MLLR, R-MLLR1 indeed reduces the adverse effects of MLLR, yet performances still drop due to over-adaptation. R-MLLR2 not only out-performs MLLR global adaption, which is widely use for CALL, but also prevents over-adaptation by using linear combinations of teachers' matrices instead of using learners' directly. The proposed methods can better utilize speaker adaptation and prevent adverse effects, thus more suitable for CALL systems.

Chapter 6

Automatic Assessment of Shadowing

6.1 Introduction

Recently, shadowing has attracted much attention in the field of teaching and learning foreign languages. Shadowing is a kind of "repeat-after-me" type exercise but rather than waiting until the end of the sentence heard, learners are required to repeat nearly at the same time. Although shadowing was originally designed to train simultaneous interpreters, its effects on foreign language learning have been widely recognized and being used in classrooms. Studies show that in shadowing, speakers do not just imitate the presented speech, but use their own speech habits and language knowledge as well. The measurement of shadowed utterances can be an indicator of the speaker's overall language proficiency.

Existing works on automatic pronunciation scoring have mainly been focused on "read" speech, mostly using Hidden Markov Models (HMM) which have been trained with native "read speech". However, in shadowing, since learners have to follow the speaking rate of the input native utterance, the speaking style of the learners is very different from "read" speech. Especially in the case of beginners, the text content of the utterances generated through shadowing can be completely different from the presented ones. To the authors' knowledge, no automatic pronunciation scoring method has been proposed or investigated for shadowing.

In this study, we propose a supervised technique by using HMM likelihood-based Goodness of Pronunciation, and an unsupervised technique based on time-strained bottom-up clustering to measure shadowed utterances by Japanese learners of English and language teachers. Correlations between automatic scores and manual-rated scores or speakers' TOEIC overall proficiency scores have been investigated and the results are promising.

6.2 Shadowing as a method for language training

Shadowing is originally introduced for training simultaneous interpreters. It requires subjects to repeat a presented native speech as quickly and closely as possible. Recently, shadowing has been widely used in langue training for its effects on improving students' speaking and listening abilities. Many language teachers have reported that students' learning is greatly enhanced by shadowing [41,42].

According to Brian McMillan [43], in shadowing, students should think about what they are repeating and can be encouraged to focus on meaning, grammar, pronunciation, or a combination of these as they shadow. Therefore, shadowing poses a cognitive load on students and can help them to improve their overall language proficiency.

6.3 Unsupervised scoring techniques

Since the learners need to immediately repeat whatever they hear, the speaking style in shadowing is very different from that of "read" speech. Especially in the case of beginners, their pronunciation often becomes corrupt and inarticulate. Considering two facts:

1) the supervised scoring technique based on HMMs, which are trained on "read" speech, can inevitably cause segmentation errors in evaluating utterances generated through shadowing,

2) it is desirable to build a scoring system that requires only an utterance pair: a native utterances presented to learners and a learner's utterance generated in response to the native utterance

A new unsupervised method is proposed here for automatic scoring of utterances in shadowing. The new method does not use any acoustic models such as HMMs at all, and just compares the two utterances based on time-constrained bottom-up clustering. Details of the time-constrained bottom-up clustering will be explained in the next section.

6.3.1 Unsupervised phoneme segmentation based on time-constrained bottom-up clustering algorithm

Most of the previous approaches to unsupervised phoneme segmentation have been focused on detecting the change points of speech signals and considering them as the boundaries of phonemes. Different from these approaches, we have proposed a bottom-up segmentation algorithm that starts with each frame as segments and merge acoustically similar adjacent segments into lager segments in a greedy way until the optimal segmentation is found. A class of statistical measures has been used to decide the 2 segments (clusters) to be merged and shows better results than other published methods. In this study, we used a fast implementation of the proposed algorithm by using Ward's method.

Ward's method is hierarchical agglomerative clustering method, which searches the similarity matrix for the similar pair of clusters and reduces the number of clusters by one through merging the most similar pair of clusters until all clusters are merged. The Word objective is to find at each stage those two clusters whose merger gives the minimum increase in the total within group error sum of squares (or distances between the centroids of the merge clusters). Suppose that adjacent speech segments p and p+1 are to be merged into new cluster r (=p ∪ q). If the segments are m-dimensional vectors $(x_1, x_2, ..., x_m)$, within group error sum of square E(p) is defined as

$$E(p) = \sum_{i=1}^{n_p} \sum_{j=1}^{m} (x_{ij}^{p} - \overline{x}_{j}^{p})^2 \qquad (4)$$

where n_p is the number of samples, and $\bar{x}_j^p$ is the j-element of the centroid of *p*. The increase of within group error sum of square when merging segments p and p+1 into r thus can be calculated as

$$\Delta E(p, p+1) = E(r) - \{E(p) + E(p+1)\}$$

By merging adjacent segments p and p+1 with minimum $\Delta E(p, p+1)$, the number of cluster would be reduced by one. We can realize bottom-up clustering of speech segments with iteration of the process.

6.3.2 Stopping condition of clustering

Considering the stage at which each segment approximately corresponds to each phoneme, the next step to merge two segments would be merging 2 clusters that belong to different phonemes. In this case, we assume that the minimum distances between the centroids of each two phonemes are approximately speakers-invariant. Therefore, regardless of speakers, when the proposed segmentation is conducted on an utterance, the merging step after the optimal stage should yield larger ΔE, i.e. $E(p \cup p+1) >> E(p) + E(p+1)$. Then we can set a predetermined threshold K for *ΔE(p,p+1)*, which can be used as stopping condition of clustering.

Figure 6.1 shows an example of the proposed phoneme segmentation based on time-constrained bottom-up clustering, comparing with manual label. Although some phoneme segment boundaries are not correctly detected, the results are rather good.

Figure 6.2 shows the segmentation results on presented read speech and the showed utterances of two learners with TOEIC scores of 421 and 202 in response to the presented utterance. Vertical axis is the increase of within group error sum of square $\Delta E(p, p+1)$, and horizontal axis is the number of clusters. The threshold K was set to be 0.23, which has been tested on various databases and proved optimal for English and Japanese.

By examining the results of segmentation on these utterances, it is clear that even with the same linguistic content, the more distinctly the utterance is spoken, the more segments can be found when clustering stops. Therefore, the number of the clusters or the segments in shadowing speech can be considered as an indicator of the learner's proficiency.

6.3.3 Distances between speech evens and articulatory efforts

If a sound (pronunciation) is not acoustically intelligible or distinct, we can say it is not articulatory distinct. Recently, a structural representation of speech, which consists of every even-to-even distance to form a geometrical structure (distance metric), has been proposed. Previous work has showed that the size of this structure of speech can be interpreted as magnitude of articulatory efforts made in speech production. For example, schwa is located in the center of the structure of vowels, which indicates that schwa is produced with the least articulatory effort and other vowels need more articulatory efforts to control the shapes of vocal tract to generate a distinct sound.

Considering these facts, we can use the distances between speech evens as an indicator of articulatory efforts. In our proposed automatic segmentation, when the clustering stops, by calculating the number of segments, we can estimate the articulatory efforts to generate the utterance. In other words, we can evaluate how intelligible a given utterance is with this method. In the case of shadowing, we can say the higher proficiency the learners have, the more intelligible their shadowed utterances would be, vice versa. Therefore, our proposed unsupervised technique is suitable for evaluating utterances in shadowing.

6.4 Supervised scoring techniques

6.4.1 GOP measurement

Various supervised techniques using HMM have been tried in many works to evaluate pronunciation. As mentioned in the previous chapters, confidence-based pronunciation assessment, which is defined as the Goodness of Pronunciation (GOP), is often used for accessing speakers' articulation and shows good results on read speech. In this study, we used HMM acoustic models trained on TIMIT [33] and WSJ [34] corpus to calculate GOP scores defined in Chapter 5. We calculate average phoneme GOP score for each learner as his or her proficiency scores.

6.4.2 Continuous Phoneme Recognition Scores

In case of transcription not being available, we can use HMM acoustic models to conduct continuous phoneme recognition. We consider for each utterance, the less intelligible the pronunciation is, the less distinct the individual segments are in the utterance. The number of recognized phonemes per utterance can be used as an index to measure the intelligibility. Here the number of phonemes normalized by the number in the presented utterance thus can be defined as continuous phoneme recognition (CPR) score. CPR score is very similar with the scores based on unsupervised clustering and do not require transcription of the utterances. The only different between the two scores is that, CPR scores are calculated by using acoustic models and clustering based scores calculation does not need any acoustic models.

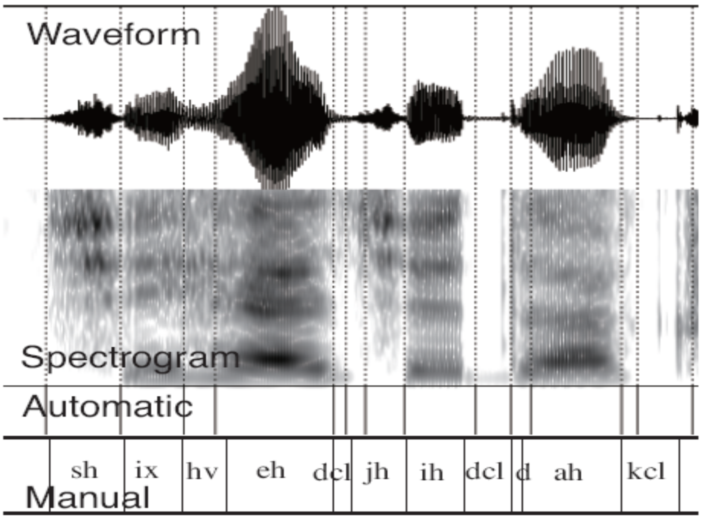


Figure 6.1: An example of unsupervised phoneme segmentation

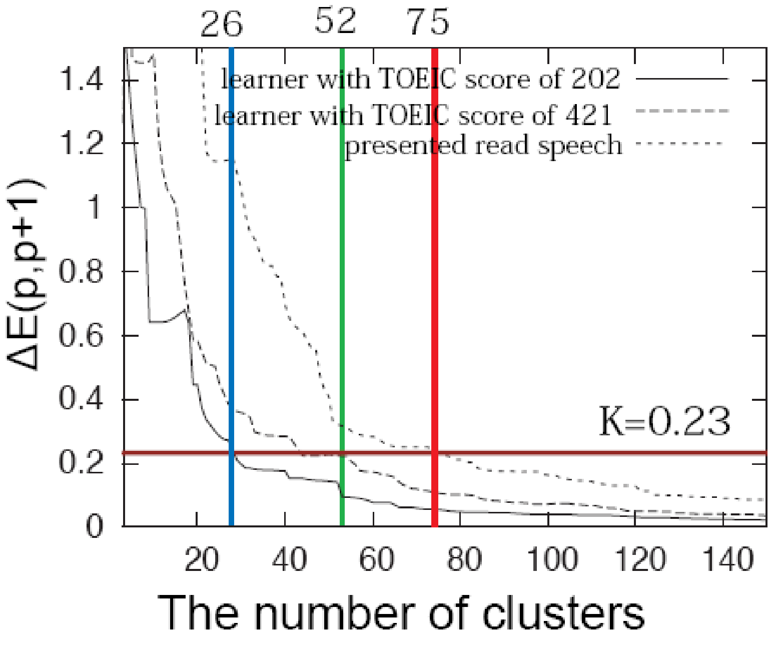


Figure 6.2: Unsupervised phoneme segmentation on shadowed utterances and presented read speech.

Table 9: Subjects' TOEIC scores

Proficiency	TOEIC scores	Average
Advanced	990, 990, 968, 955, 940, 895, 825	938
Intermediate	625, 601, 592, 581, 512, 436, 432, 427, 421	514
Beginners	395, 367, 308, 301, 289, 278, 275, 252, 202, 197, 158	275

6.5 Experiments

6.5.1 Shadowing database and manual assessment

In order to evaluate the proposed techniques, we collected a database of shadowing productions from 27 speakers, in which there are 7 advanced learners, 9 intermediate learners and 11 beginners. The subjects' overall proficiency scores measured by TOEIC (Test of English as International Communication) are shown in Table 9.

The presented utterances recorded by a native speaker of English contain 21 sentences and its topic was carefully chosen to be familiar to Japanese learners. However, the utterances themselves had never been presented to any of the subjects before. All the sentences were presented to the subjects sequentially at the rate of 140 wpm (words per minute), and the subjects were instructed to repeat as closely and as quickly as possible. The subjects' shadowing productions in response to the presented utterances were recorded in the environment of classroom.

Manual assessment was conducted by an expert in language education. Utterances of 10 sentences shadowed by 11 learners were chosen. The rater examined each utterance word by word. For each correctly pronounced word, the score would be 1. For any inserted word, the score of the word would be -1. For each partially correct word, the score would be 0.5. Thus by summing up the score of every word and normalized by the number of the words in the presented utterance, the result can be used as manual score for each shadowed utterance. Note that manual assessment of shadowing speech is very time-consuming. It took about 1 hour for the expert to evaluation per learner.

6.5.2 Acoustic conditions for analysis

The acoustic conditions for analysis for HMM-based evaluation are shown in Table 10. The acoustic models we use are triphone HMM models trained on

Table 10: Acoustic conditions in HMM-based method

Sampling	16bit / 16kHz
Window	Hamming / 25 ms length/10 ms shift
parameters	MFCC, log-energy, and their $\Delta, \Delta\Delta$

Table 11: Acoustic conditions in clustering-based method

Sampling	16bit / 16kHz
Window	Hamming / 16 ms length /10 ms shift
Parameters	MCEP (1～12)
Threshold	K = 0.23

TIMIT [33] and WSJ [34] databases, which are the same as Chapter 4 and 5.

The acoustic conditions for analysis in clustering-based automatic segmentation are shown in Table 11. The threshold is set to be 0.23 as stopping condition. This threshold has been proved valid on different databases of English and Japanese.

6.5.3 Comparison of automatic assessments

GOP scores， CPR scores and clustering scores are supposed to play an equal role in pronunciation evaluation. To demonstrate this, we compared these 3 methods quantitatively. The correlations at utterance level and speaker level are shown in Figure 6.3 and 6.4 respectively.

Very high correlations have been found between any two of the three scores. At utterance level, the correlation between clustering scores and CPR scores is 0.83, correlation between GOP scores and CPR scores is 0.80 and correlation is 0.75. At speaker level, the correlations are even higher. As shown in Figure 6.4, the correlations between any two of the three measures at speaker level are higher than 0.90.

6.5.4 Correlations between automatic scores and manually-rated scores

The correlations between automatic scores and manually-rated scores at utterance-level and speaker-level are shown in figure 6.5 and 6.6 respectively. Again, very high correlations have been found. At utterance level, GOP scores and CPR scores shadow highest correlation of 0.85 and the correlation between clustering is 75. At speaker level, CPR scores shows highest correlation of 0.97 and the other two scores also show high correlation of 0.94 and 0.92.

6.5.5 Correlations between automatic scores and TOEIC scores

The correlations between automatic scores and TOEIC scores are shown in Figure 6.7. Since TOEIC scores are at speaker-level and all 27 subjects have their TOEIC scores, we calculate GOP scores, CPR scores and clustering scores at speaker-level for all subjects. As shown in Figure 6.7, GOP score shows the best correlation of 0.82 and language-independent clustering score also shows a good result of 0.72.

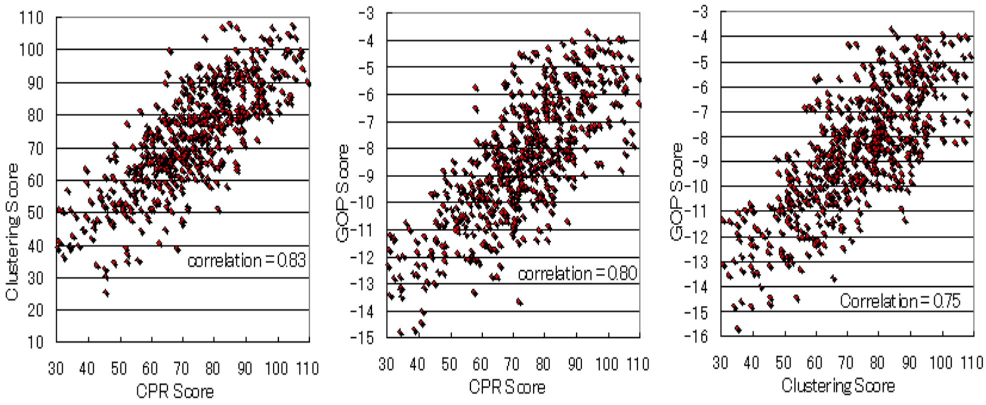


Figure 6.3: Comparison of every two of the three automatic scores at utterance level

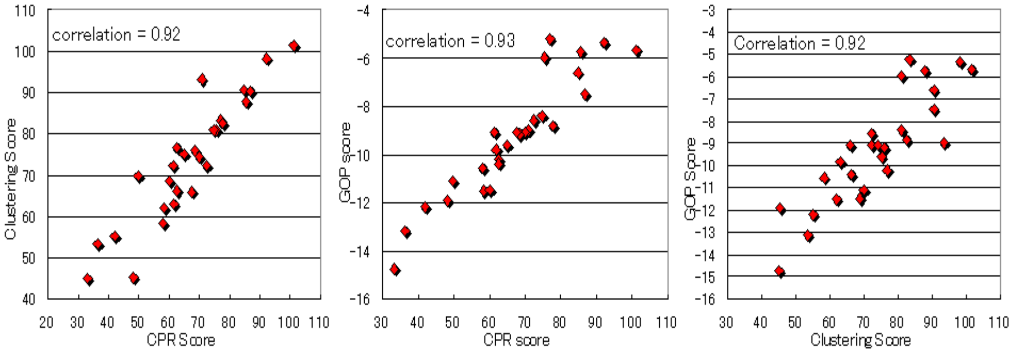


Figure 6.4: Comparison of every three automatic scores at speaker level

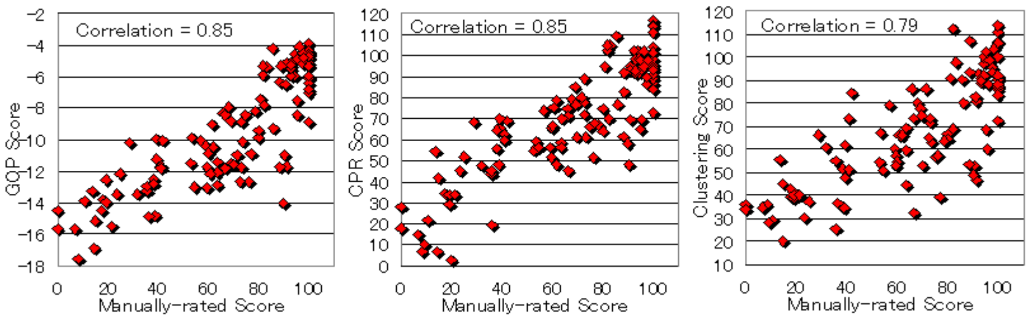


Figure 6.5: Correlation between automatic scores and manual scores at utterance level

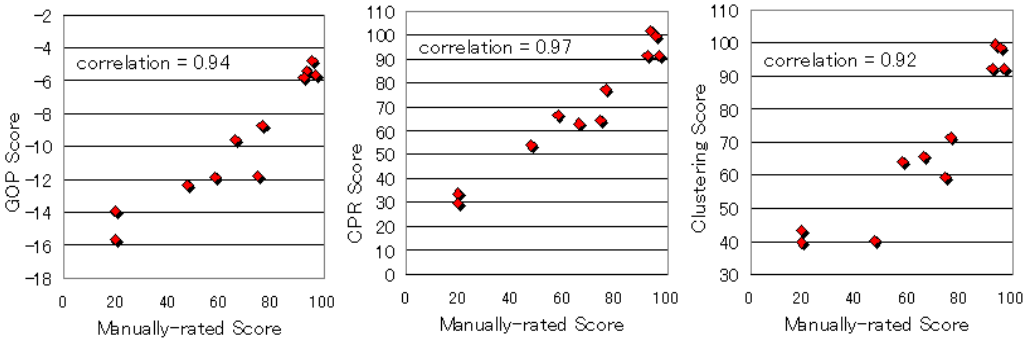


Figure 6.6: Correlation between automatic scores and manual scores at speaker level

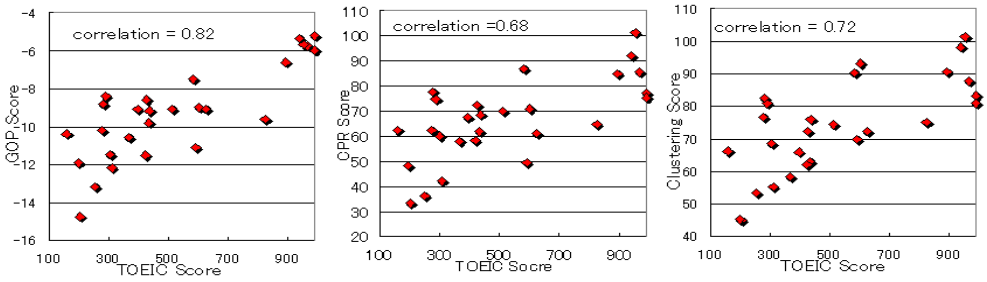


Figure 6.7: Correlation between automatic scores and TOEIC

6.6 Discussion

In read speech evaluation, even by using similar HMM-based GOP techniques, much lower correlations between machine and human scores were reported in recently published studies [29, 52]. This might be because shadowing poses a cognitive load on learners adequately and, therefore, the shadowing productions may reflect the learners' "true" proficiency level rather precisely. We will conduct comparison experiment of shadowing and read speech or so-called reading-aloud.

6.7 Conclusions

In this chapter, we have proposed three scoring methods for utterances generated through shadowing: GOP scores, CPR scores and clustering scores. For GOP scores, both acoustic models and transcripts are required. For CPR scores, only acoustic models are required and for clustering scores, neither acoustic models nor transcripts are required.

We described how to implement these techniques and compared them with each other. Evaluation experiment results show that automatic scores have strong correlation with manual scores or learners' overall language proficiency. Comparison of scores derived from different techniques shows that the proposed language-independent clustering-based scoring technique is still available for evaluation of shadowing productions.

Chapter 7

Comparison of Shadowing and Reading-aloud

7.1 Introduction

As mentioned in the previous chapter, shadowing is becoming more and more popular in English education in Japan and learners' shadowed speech can be good indicators of their true language proficiency.

Reading-aloud has always been a popular practice to improve speaking skill in language learning. Unlike shadowing, utterances generated through reading aloud, or so-called read speech, are more stable and closer to the speaking style of the speech corpuses on which acoustic models (HMMs) are often trained. Therefore, read speech is often used for automatic pronunciation evaluation. Improving the evaluation performance on read speech is also one of the goals of our research.

In this chapter, we compare shadowing to the conventional practice of reading-aloud and in order to examine how cognitive loads affect learners' speech, we also consider two situations of shadowing with and without text presented. With text, the difficulty of shadowing is reduced. We use Goodness of Pronunciation (GOP) based scores calculated through HMMs as automatic scores. Correlations between automatic scores and speakers' TOEIC overall proficiency scores are investigated to analyze the results based on the tasks posed on learners with various cognitive loads.

7.2 Automatic Scores

Since supervised automatic scoring methods show better results on shadowing and is widely used for evaluating read-speech (reading-aloud speech), supervised automatic scores will be used for the comparison experiments between shadowing and reading-aloud. We use HMM-based GOP and F-GOP scores described in Chapter 5 as measurements for intelligibility of learners' shadowing and reading-aloud speech.

For acoustic models, we use HMM triphone models trained on TIMIT and WSJ databases as basic acoustic models. We then apply MLLR global adaptation to the basic acoustic models to examine effects of speaker adaptation on shadowing and reading-aloud.

7.3 Data collection

In order to compare shadowing with reading aloud, we have designed a program to record learners' utterances in three modes with different levels of phonation difficulty: shadowing (only native model utterances are presented), reading aloud (only texts are presented), and shadowing with texts (both native model utterances and text contents are presented). In shadowing and

Table 12: Subjects' TOEIC scores

Proficiency	TOEIC scores	Average
Advanced	955, 926, 855, 832, 825, 792, 773, 752	838
Intermediate	687, 686, 668, 563, 524,	625
Beginners	496, 425, 399, 378, 252	392

shadowing-with-text modes, learners were required to repeat at the same speed as that of the presented native utterances, but in reading-aloud mode, learners were allowed to read the presented text at his/her own pace. For each mode, the contents of presented utterances or texts were carefully selected by experts so that they contain three levels of semantic difficulty: easy, intermediate, and difficult. The subjects were instructed to first record their shadowing productions, then shadowing with text and finally reading aloud of each task with different level of semantic difficulty. Utterances under these conditions were collected from 18 Japanese learners (8 advanced learners, 5 intermediate learners and 5 beginners) with a variety of proficiency.

We use TOEIC (Test of English as International Communication) scores as the references of learners' overall language proficiency. The subjects' TOEIC scores are shown in Table 12.

7.4 Evaluation Experiments

7.4.1 Comparison of shadowing, shadowing with text and reading aloud by using GOP scores

The correlations between GOP scores and TOEIC scores are shown in Table 13. In all tasks with three different levels of difficulty, GOP scores calculated from shadowing showed the highest correlations. The results from shadowing with text are lower than shadowing but better than reading aloud. Shadowing with the intermediate level of semantic difficulty shows the highest correlation of 0.81. This indicates that the contents of shadowing need to be carefully chosen to better measure learners' proficiency.

We then applied MLLR adaption by using a part of each learner's utterances from reading aloud to the native acoustic models. The results are shown in Table 14. Although the improvement of reading aloud utterances are more significant than shadowing, automatic scores calculated from shadowing utterances still show better performances. This further confirms the advantage of shadowing over reading aloud in overall language proficiency assessment.

Table 13: Correlations between GOP scores and TOEIC scores without adaptation

Level of difficulty	Shadowing	Shadowing with text	Reading aloud
Easy	0.74	0.65	0.48
Intermediate	0.81	0.68	0.59
Difficult	0.71	0.67	0.61

Table 14: Correlations between GOP scores and TOEIC scores with MLLR adaptation

Level of difficulty	Shadowing	Shadowing with text	Reading aloud
Easy	0.74	0.68	0.60
Intermediate	0.82	0.71	0.68
Difficult	0.70	0.69	0.67

7.4.1 Comparison of shadowing, shadowing with text and reading aloud by using F-GOP scores

We calculate F-GOP scores with acoustic models and then conducted MLLR global adaptation by using part of the utterances for reading-aloud as adaption data. Correlation between F-GOP scores and TOEIC scores is used as automatic scoring performance measurement.

Figure 7.1 shows the results of correlations of F-GOP scores and TOEIC by using original HMM acoustic models (without adaptation) with three different levels of difficulty: easy, intermediate and difficult, compared with GOP. Figure 7.2 shows the performance of F-GOP scores with or without MLLR global adaptation.

As shown in Figure 7.1, although F-GOP without adaptation did not improve the scoring performances on shadowing, the improvement on read speech (reading aloud) is rather significant. We consider this might because the forced aligned boundary information F-GOP refers to is not as accurate in the case of shadowing as that of read speech.

As shown in Figure 7.2, with MLLR adaptation, the performance of F-GOP can be further improved. In GOP or F-GOP scoring, with or without adaptation, shadowing always out perform shadowing-with-text or reading-aloud. By comparing the results in terms of different levels of text difficulty, shadowing with intermediate level of text difficulty show better results than “easy” or “difficult” levels.

7.5 Discussion

In every different task, shadowing has shown better results than reading aloud. This indicates that shadowing, which poses a certain amount of cognitive load on learners, can better reflect the true language proficiency of the learners.

However, MLLR adaptation, which improved the results of reading-aloud significantly, did not improve the performances of shadowing evaluation as much. We considered that it is because the difference of the speaking style between shadowing and reading aloud, even by the same speaker, causes much of the mismatches between utterances generated through shadowing and the original acoustic models. The use of read speech as adaptation data can not reduce the mismatches caused by the difference of speaking style. In order to further improve the performance of shadowing evaluation, we need to address the problems caused by the speaking style of shadowing in the future.

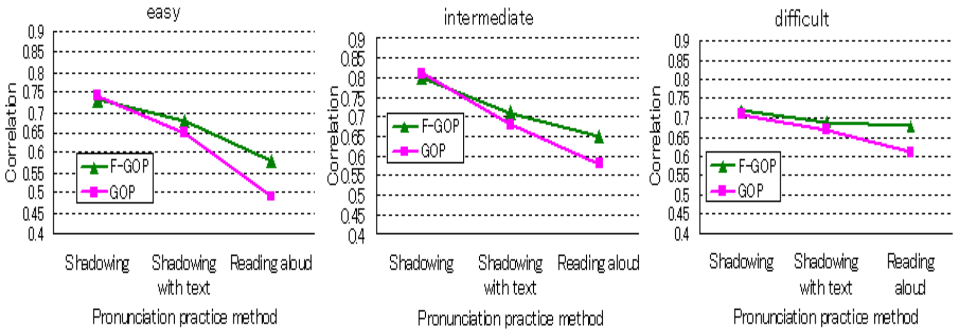


Figure 7.1: Comparison of F-GOP and GOP

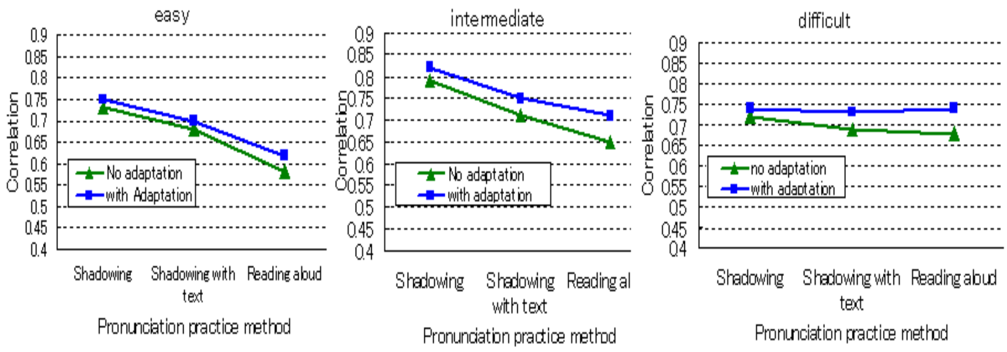


Figure 7.2: Performance of F-GOP with adaptation

7.6 Conclusions

In this chapter, we compare automatic proficiency assessment results on utterances generated through three different ways of pronunciation practices: shadowing, shadowing with text, and reading aloud. Three different degrees of difficulty of the presented text or native utterances are employed to examine the effects of cognitive loads posed on learners.

Experimental results show that shadowing with a proper degree of difficulty, or cognitive load, can be used to assess language learners' proficiency with the best accuracy. We also analyze the effect of MLLR adaptation on automatic scores and find out that MLLR improves the performances on reading out significantly but little improvement is found on shadowing. We are planning to investigate the change of learner's proficiency after routinely shadowing practices over a period.

Chapter 8

Prosodic Evaluation of Shadowing

8.1 Introduction

In the previous chapters, we have proposed several automatic scoring methods for first-time shadowing, in which case the presented speech has not been seen or heard by the subjects before shadowing. High correlations between automatic scores of first-time shadowing and TOEIC overall proficiency scores have been found.

However, we found that learners used different strategies to shadow a given native utterance. For example, some learners might focus on the contents of the presented utterance and repeat individual words with their own style of speaking. Some might focus on segmental phoneme pronunciations and others might only focus on the prosodic features yet ignoring the intelligibility of pronunciations.

In order to further analyze segmental and prosodic features of shadowing speech, instead of first-time shadowing, more stable personal-best shadowing utterances, which are recorded after sufficient practices without the transcription, are used for our analysis. Figure 8.1 shows a procedure of recording learners' utterances of shadowing and reading-aloud. This study focus on how learners' degree of understanding the contents during shadowing affects their pronunciations in shadowed utterances in terms of phoneme intelligibility and prosodic fluency. To measure learners' degree of understanding the contents, we introduce two types of scores. One is a comprehension test that contains 7 questions. Each question asks learners to choose the best answers out of 4 candidates according to the presented native speech they heard during shadowing. The other is learners' self-check of words that they do not recognize during shadowing. In this case, the transcription of the native speech is shown to the learners and, by referring to it, they are required to mark out any words they did not follow up during the personal-best shadowing. We prepare other two types of scores. We ask a language expert to rate the shadowing utterances in term of prosodic features, intonation and rhythm, and an overall prosodic score is assigned to each subject. TOEIC score is also provided from the learners.

For automatic analysis, we use Goodness of Pronunciation scores as the measure for phoneme intelligibility. As for prosodic features, we focus on F0-based and power-based DTW distances between shadowed utterances and the presented native speech, utterance-level variance of F0, length of pauses and rate of speech. The relations between reference scores and automatic scores are examined.

8.2 Data collection

32 subjects participated in our shadowing data collection. These subjects are Japanese learners of English from two universities in Japan and their TOEIC scores are shown in Table 15.

The contents of presented speech were carefully chosen by a language expert that contains 14 sentences of an intermediate level of difficulty. The presented native speech was provided by an English teacher of native General American

Table 15: Subjects' TOEIC scores

TOEIC scores	Number of subjects
600-800	13
400-600	11
100-400	8

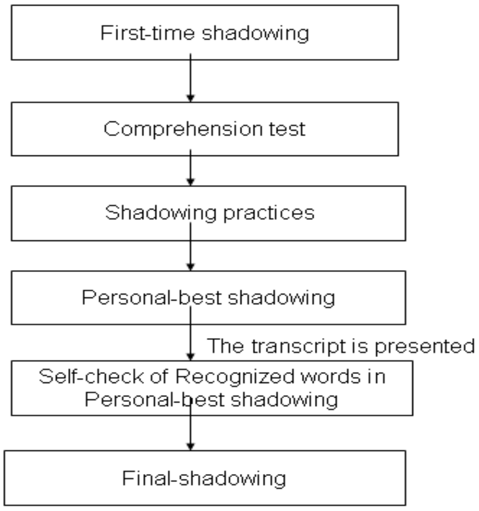


Figure 8.1: Recording procedure of shadowing.

English speaker with normal speed but with a variety of changes in intonation.

The contents of the presented speech are shown in Appendix D. The transcription or speech was never presented to the subjects before recordings. The subjects were asked not only to pay attention to segmental pronunciations, but also to the prosodic features of the presented speech and to mimic them as closely as possible instead of speaking in their own ways.

After recording the first-time shadowing, the subjects were asked to take a comprehension test. The test is written in Japanese with seven questions related to the contents of the presented speech. For each question, the subjects need to choose the best answer out of four candidates. After the comprehension test, the subjects practiced shadowing for several times until they became familiar with the native pronunciations. Then the subjects' personal-best shadowing was recorded. After personal-best shadowing recording, the transcript was presented to the subjects and while listening to their own recorded personal-best shadowing utterances, they were asked to mark out any words that they did not recognize during shadowing.

Now that the transcript has been shown to the subjects, we record their shadowing speech one more time for comparison with their personal-best shadowing. We will refer to this final shadowing recording as final-shadowing hereafter. Figure 8.1 shows the total procedure of a sequence of recordings including a comprehension test.

8.3 Reference scores

For reference scores, first, we calculate the number of words that the subjects recognized correctly during shadowing and define recognized word scores (RWS) based on the subjects' self-check results as below.

$$\mathrm{RWS} = \frac{\text{number of recognized words}}{\text{total number of words}} \times 100\% \tag{8.1}$$

And comprehension test scores (CTS) is defined as,

$$\mathrm{CTS} = \frac{\text{number of correct answers}}{\text{total number of questions}} \times 100\% \tag{8.2}$$

These two scores measure learners' degree of understanding the contents of the native utterances in different ways. RWS and CTS correspond to word-level comprehension and overall comprehension, respectively.

When learners are asked to shadow presented native utterances, they sometimes pay more attention to the prosodic aspects, intonation and rhythm, of the presented utterances. In order to measure prosodic proficiency, we ask an English education expert to rate a score for each subject based on the expert's subjective impression of that learner's prosodic fluency. We refer to this score as manually-rated prosodic score (MPS).

Table 16 shows the correlations of any two of the referenced scores including TOEIC scores. RWS shows very high correlation with TOEIC overall proficiency scores and manual prosodic scores (MPS). This indicates that the level of word recognition during shadowing not only reflect learners' overall language proficiency but also affects prosodic fluency of shadowed utterances. The relatively low correlation between CTS and MPS might indicate that it is possible to mimic prosodic features of the presented speech without comprehending the whole contents.

8.4 Scores based on prosodic measures

8.4.1 Fundamental frequency (F0)

In our experiment, F0 is extracted by using Praat [50], which analyzes F0 every 5 ms with 20ms frames of each utterance. The log scale values of F0 are normalized to cancel the differences due to gender. In addition, F0 pattern is smoothed with regression fitting.

[30] uses the DTW distances between native utterances and learners' read speech as measure for intonation proficiency. In the case of shadowing, the presented native speech is the only source that learners refer to during shadowing. The distances of presented native utterances and learners' shadowed ones are reasonable measure for intonation fluency.

Table 16: Correlations between any two of the referenced scores

	RWS	CTS	TOEIC	MPS
RWS	1	0.53	0.70	0.72
CTS		1	0.73	0.54
TOEIC			1	0.72
MPS				1

Word-level DTW distances we use is defined as below,

$$g(i,j) = \min \begin{cases} g(i-1,j) + d(i,j) \\ g(i-1,j-1) + 2d(i,j) \\ g(i,j-1) + d(i,j) \end{cases} , \qquad (8.6)$$

where *d(i,j)* is a local difference between normalized F0 values of the *i*-th frame of shadowed utterance and the j-th frame of the presented speech and *g(1,1)=d(1,1)*. If the speech segment of a word has *I* frames in native speech, and its corresponding segment has *J* frames in learners' shadowed speech, the DTW distance of this word is calculated by,

$$D(native, learner) = \frac{g(I,J)}{I+J} . \qquad (8.7)$$

We refer to scores calculated by Equation (8.7) as F0_DTW. The smaller F0_DTW is, the shorter the distance between the 2 utterances, i.e. the closer the learner's pitch pattern is to the presented native speech.

According to [56], at utterance level, Japanese learners' pitch contours are more flat than those of native English speakers' are. Thus, the variance of normalized F0 at utterance level can be used as an indicator to judge if the learners' shadowed utterances are Japanese-like or native-like.

In the case of shadowing, differences of F0 variance among different levels of learners are rather clear. For example, Figure 8.2 shows the F0 pattern of presented native speech, Figure 8.3 shows the F0 pattern of an advanced learners' shadowed utterance and Figure 8.4 shows the F0 pattern of an intermediate learner's shadowed utterance. As shown in these figures, intermediate learner's pitch contour is much smoother than that of advance learner or the native speaker.

8.4.2 Power

Power (or intensity) parameters are also extracted by Praat. Power contours of learners' utterances have strong relation with the rhythm of their speech.

DTW distances between intensity contours of learners' shadowed speech and the presented native speech are calculated in the same way as mentioned in previous section. We refer to these scores as Power_DTW scores.

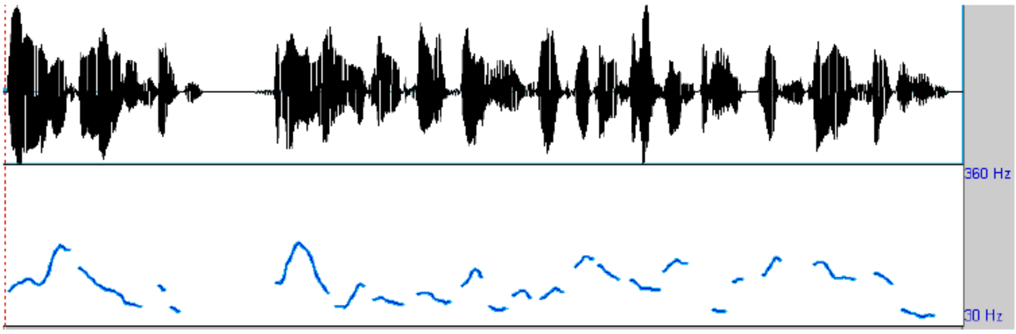


Figure 8.2: Pitch contour of presented native speech.

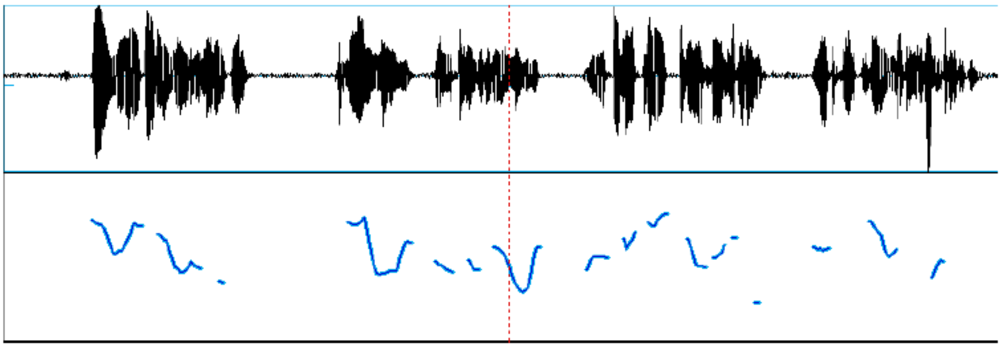


Figure 8.3: Pitch contour of an advanced learner’s shadowing speech.

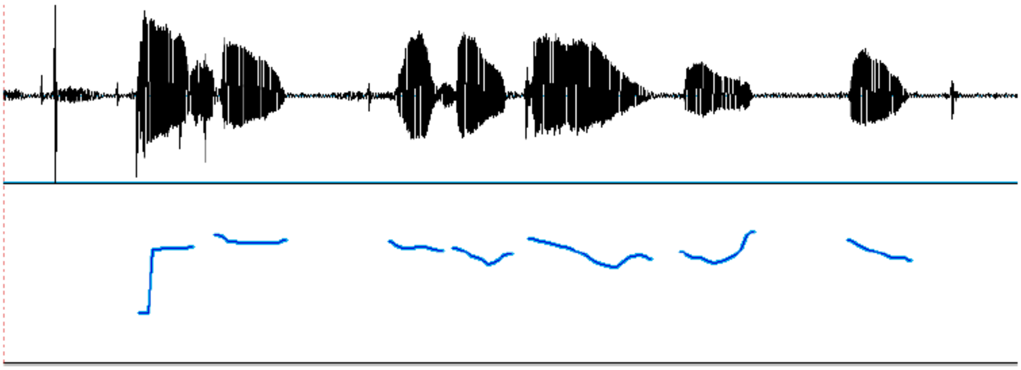


Figure 8.4: Pitch contour of a intermediate learner’s shadowing speech

8.4.3 Length of pauses

Pauses are automatically detected by using a threshold-based scheme for the values of power. Durations of silence segments between words are calculated and normalized by the length of the presented utterance. We consider that there should be more pauses in a learner's shadow speech if he or her cannot follow the presented speech.

8.4.4 Rate of speech

Rate of speech (ROS) is calculated as,

$$ROS = \frac{N_{phonemes}}{D_{utterance} - D_{Silence}} \quad , \qquad (8)$$

where $N_{phonemes}$ is the number of phonemes and $D_{utteranc}$ is the duration of the utterance and $D_{silence}$ is the length of silence.

ROS can be used as an indicator of fluency of learners' shadowed speech or how well a learner can repeat the present speech at the same speed.

8.5 Evaluation Experiments

8.5.1 Correlations between automatic scores and reference scores

For personal-best shadowing, we investigate correlations between every automatic scores described in Section 8.4 and referenced scores mentioned in Section 8.3. Correlations between automatic scores and recognized word scores (RWS) are shown in Table 17. GOP scores, F0-based scores and ROS show better results than scores based on power or pauses.

Correlations between automatic prosodic scores and manual prosodic scores (MPS) are shown in Table 18. Again, F0-based scores perform better than Power-based scores and ROS shows better result than Pauses.

Table 17: Correlations between automatic scores and RWS (Recognized word scores)

Measures	Correlation
GOP	0.63
F0_DTW	-0.45
F0_variance	0.55
Power_DTW	-0.30
Pauses	-0.20
ROS	0.58

Table 18: Correlation between automatic prosodic scores and MPS (manual prosodic scores)

Measures	Correlation
F0_DTW	-0.55
F0_variance	0.49
Power_DTW	-0.30
Pauses	-0.37
ROS	0.59

8.5.2 Multiple regression models

We use a set of multiple regression models to combine different measures. The combined scores are given by the following equation,

$$S = \sum_{k=1}^{K} \alpha_k F_k \tag{9}$$

where F_k is the k-th feature score of K scores and α_k is obtained by using training data.

Here we adopted leave-one-out cross verification to estimate target scores with different features. First, we use the 6 measures shown in Table 17 to estimate RWS. The correlation between estimated scores and RWS is 0.68 which higher than any one of the features. Although the result is lower than the correlation between RWS and TOEIC or MPS (shown in Table 16), the differences are not significant. We then use the five automatic scores based on prosodic features to estimate MPS. The correlation between the estimated scores and MPS is 0.6, which is again higher than any single measure.

8.5.3 Comparison of personal-best shadowing and final shadowing

The difference between personal-best shadowing and final shadowing is that final shadowing is done after checking the individual words in the presented native speech. We have expected that learners' pronunciation might improve significantly by checking the transcript. However, by closely examining the MPS of both types of shadowing speech, we find that they are very similar.

Considering the fact that there are no advanced learners whose TOEIC scores are higher than 800 in the subjects, the correlations we obtain are rather high.

Correlations between RWS and automatic scores calculated by using the data of personal-best shadowing and final shadowing are shown in Table 19. Although correlations between GOP and RWS change significantly, in the case of prosodic measures, the correlations are almost the same. This indicates that knowing the contents of showing might not help learners with their prosodic fluency in shadowing. Correlations between MPS and automatic scores calculated by using the data of personal-best shadowing and final shadowing are shown in Table 20. Similar conclusion can be draw as in the case of RWS that except GOP, other feature scores show similar conrrelations.

Table 19: Correlations between automatic scores and RWS, comparing personal-best shadowing with final shadowing

Measures	Personal-best shadowing	Final shadowing
GOP	0.63	0.55
F0_DTW	-0.45	-0.43
F0_variance	0.55	0.56
Power_DTW	-0.3	-0.2
Pauses	-0.2	-0.25
ROS	0.58	0.56

Table 20: Correlations between automatic scores and MPS, comparing personal-best shadowing with final shadowing

Measures	Personal-best shadowing	Final shadowing
GOP	0.65	0.58
F0_DTW	-0.55	-0.50
F0_variance	0.49	0.47
Power_DTW	-0.30	-0.29
Pauses	-0.37	-0.39
ROS	0.59	0.60

8.6 Conclusions

In this chapter, we analyze shadowing with automatic measures related to phoneme indelibility and prosodic fluency. We compare these automatic measures with several reference scores and propose several methods for shadowing evaluation. Experimental results show that the proposed automatic scoring methods are suitable for shadowing evaluation. Comparison of personal-best shadowing and final shadowing shows that knowing the contents before shadowing does not necessarily affect the prosodic aspects of learners' shadowed utterances. Future works include detailed comparison of shadowing with other conventional training methods, especially on prosodic aspects.

Chapter 9

Conclusions

9.1 Summary

In this thesis, several novel methods have been presented for improvement in pronunciation evaluation of reading-aloud and shadowing speech based on speech processing technology.

A detailed introduction of background knowledge is presented in Chapter 2 and an overview of various exiting CALL systems and the technology behind them are closely examined in Chapter 3.

From Chapter 4 to 5, we address the over-adaption problem that occurs when using conventional MLLR adaption for pronunciation evaluation and propose a novel adaptation technique, Regularized Maximum Likelihood Linear Regression (Regularized-MLLR), for CALL systems. The idea is to use a group of teachers' transformations to regularize learners' transformations so that erroneous pronunciations will not be transformed into good pronunciations.

First, we investigate the effects of MLLR on pronunciation evaluation in two ways: automatic scoring and error detection. Experimental results show that although the MLLR global adaptation (number of regression classes is one) can indeed improve evaluation performances, when the number of regression classes increases and more details of learners' pronunciations are adapted, over-adaptation occurs so that erroneous pronunciations are recognized as correct ones. However, even with over-adaptation, conventional adaption can still improve precision rate of error detection performance, which indicates that false rejections can be reduced by conventional MLLR.

Based on these results, we implement two forms of Regularized-MLLR, R-MLLR1 and R-MLLR2 by using teachers' perfect pronunciations to regularize learners' transformations. R-MLLR1 uses the average of a group of teachers' transformation matrices as a constraint adding to the conventional MLLR transformations. This constraint prevents radical transformations when there are too many errors in the adaptation data. R-MLL2 uses linear combination of the teachers' MLLR transformation matrices to represent each learner's transformation. This approach does not directly use learners' MLLR transformations that are estimated from their imperfect pronunciations, therefore prevents over-adaption.

We compare R-MLLR1 and R-MLLR2 with conventional MLLR by conducting experiments on the same conditions as we investigate the adverse effects of MLLR. Automatic scoring and error detection experiments show that the proposed methods outperform conventional MLLR. By adding constraints to MLLR, R-MLLR1 indeed reduces the adverse effects of MLLR, yet performances still drop due to over-adaptation. R-MLLR2 not only out-performs MLLR global adaption, which is widely use for CALL, but also prevents over-adaptation by using linear combinations of teachers' matrices instead of using learners' directly. The proposed methods can better utilize speaker adaptation and prevent adverse effects, thus more suitable for CALL systems.

From Chapter 5 to 8, we proposed method for segmental and prosodic evaluation of shadowing speech and compare shadowing with reading aloud. We first propose supervised and un-supervised methods for automatic scoring of shadowing. Correlations between automatic scores and manual scores or TOEIC overall proficiency score are investigated. Experimental results show that very high correlations between automatic scores and manual scores or TOEIC scores have been found. The language independent unsupervised method is also available for shadowing evaluation.

We then compare shadowing with reading aloud with different cognitive load posed on the subjects. Experimental results show that with adequate amount of cognitive load, shadowing can better reflect learners' true proficiency than conventional reading aloud. We conduct speaker adaption for shadowing and reading aloud and find that speaker adaption have much more effects on reading-aloud than shadowing.

Finally, we propose prosodic evaluation for personal-best shadowing and final shadowing. Personal-best shadowing is the shadowing speech recorded after extensive practices and final shadowing is the shadowing speech recorded after the transcriptions are presented. By combining different prosodic aspects such as F0, power, pauses and rate of speech, we can obtain a reliable score for each speaker. The performance of TOEIC score prediction can be further improved by combing prosodic scores with GOP scores. The comparison of personal-best shadowing and final shadowing shows that showing the text contents to the learners does not necessarily improve their prosodic proficiency.

9.2 Future work

To regularize learners' transformation, we only use the 20 teachers' speech data from ERJ database. Increasing the number of teachers will increase the variety of speaker characteristics of the teachers' data we use for regulation. We need to investigate if increasing the number of teachers would improve the effectiveness of adaptation.

The method we use to cluster model parameters into regression classes for MLLR and Regularized-MLLR is according to how close they are in acoustic space. By using some phonetic expertise in deciding which components should be clustered together, we might obtain better recognition and error detection results.

Other adaptation techniques such as MAP and Eigenvoices need to be examined and compared with MLLR-based methods. By looking into the details of each method, we can combine them to further improve evaluation performances for CALL systems.

Since the proposed speaker adaption techniques are language-independent for pronunciation evaluation, we would also like to test them on different databases of different languages such as Japanese, Chinese etc.

For shadowing evaluation, we need to examine the long-term effect on learners' proficiency of shadowing practice. Error detection should be conducted on shadowing when the learners are more familiar with shadowing and their shadowed utterances do not become so broken and recognizable with automatic speech recognition (ASR). We also need to compare evaluation results of shadowing on different conditions such as changing the speech of presented speech, the dialect of accent of the presented speech, etc.

Since manual segmental or prosodic scores are rated by one expert, we need to increase the number of experts and compare the manual scores by different raters. For prosodic scores, an overall score of each learner’s prosodic proficiency is given by the expert. We need to examine different aspects of prosodic such as intonation and rhythm separately and find out if there is any relationship between them.

Appendix A

The phoneme set with 39 phonemes of CMU pronunciation dictionary used for acoustic models are shown as below.

Phoneme	Example	Translation
aa	odd	aa d
ae	at	ae t
ah	hut	hh ah t
ao	ought	ao t
aw	cow	k aw
ay	hide	hh ay d
b	be	b iy
ch	cheese	ch iy z
d	dee	d iy
dh	thee	dh iy
eh	ed	eh d
er	hurt	hh er t
ey	ate	ey t
f	fee	f iy
g	green	g r iy n
hh	he	hh iy
ih	it	ih t
iy	eat	iy t
jh	gee	jh iy
k	key	k iy
l	lee	l iy
m	me	m iy
n	knee n	iy
ng	ping	p ih ng
ow	oat	ow t
oy	toy	t oy
p	pee	p iy
r	read	r iy d
s	sea	s iy
sh	she	sh iy
t	tea	t iy
th	theta	th ey t ah
uh	hood	hh uh d
uw	two	t uw
v	vee	v iy
w	we	w iy
y	yieldy	iy l d
z	zee	z iy
zh	seizure	s iy zh er

Appendix B

Text content for first-time shadowing

In 1996, three men in California were taken to a hospital with strange symptoms. They felt dizzy, tired, and weak. They couldn’t speak, and they had trouble breathing. The hospital doctors thought the men had been poisoned, but couldn’t work out what was wrong with them. Then they found out the three men were all chefs, and they had just shared a dish of fugu. Fugu, the Japanese name for the puffer fish, is one of the strangest fish in the ocean. The puffer fish gets its name from the way the fish protects itself from enemies. Whenever it is attacked, the fish puffs up (blows up) its body to over twice its normal size! The reason the three men were taken to the hospital is because the puffer fish is also very poisonous. As a rule, if you eat a whole puffer fish, you will probably die. The three men had a close call, but they all survived. The symptoms of fugu poisoning are a strange feeling around the mouth and throat, and difficulty breathing. You can’t breathe and your body can’t get any air. Your brain still works perfectly, however, so you know you are dying, but you can’t speak or do anything about it. Despite the danger of fugu poisoning, this strange, ugly, and very poisonous fish is actually a very expensive, and very popular, kind of food in Japan. Customers pay up to$200 per person to eat a fugu meal. Because of the danger, fugu can only be prepared by chefs with a special license from the government. These chefs are trained to identify and remove the poisonous parts of the fish. Most people who die from eating fugu these days are people who have tried their hand at preparing the fish themselves. Fugu is said to be so delicious that it has even started to be imported into Hong Kong and the United States. Several tons of fugu are now exported from Japan every year.

Appendix C

Text contents with 3 levels of difficulty for comparing shadowing and reading-aloud.

Level 1: Easy

February 14 is a day for people who have fallen in love. On this day, these men and women give gifts and cards to each other to celebrate Valentine's Day. At first, February 14 was the old Roman festival, Lupercalia. Then, on February 14, 270 A.D., a man named Valentine was killed by the Romans because of his Christian beliefs. Before Valentine was killed, he fell in love with the daughter of his jailer and would pass notes to her. His final note read, "From your Valentine." Later, February 14 became known as Saint Valentine's Day. Since then, people in love around the world have given gifts and cards to each other on Saint Valentine's Day. Gloves, chocolates, and even underwear have all been popular as gifts. Valentine cards did not become popular until the 1750s. The first Valentine cards were made by hand. People wrote their own words on the cards, usually a kind or funny message. Cards made by machines became more popular around 1850. All of a sudden, Valentine's Day became a big holiday for people who made and sold cards. Now, every year around February 14, cards and chocolates fill stores around the world, for all the people who have fallen in love.

Level 2: Intermediate

In 1996, three men in California were taken to a hospital with strange symptoms. They felt dizzy, tired, and weak. They couldn't speak, and they had trouble breathing. The hospital doctors thought the men had been poisoned, but couldn't work out what was wrong with them. Then they found out the three men were all chefs, and they had just shared a dish of fugu. Fugu, the Japanese name for the puffer fish, is one of the strangest fish in the ocean. The puffer fish gets its name from the way the fish protects itself from enemies. Whenever it is attacked, the fish puffs up (blows up) its body to over twice its normal size! The reason the three men were taken to the hospital is because the puffer fish is also very poisonous. As a rule, if you eat a whole puffer fish, you will probably die. The three men had a close call, but they all survived. The symptoms of fugu poisoning are a strange feeling around the mouth and throat, and difficulty breathing. You can't breathe and your body can't get any air. Your brain still works perfectly, however, so you know you are dying, but you can't speak or do anything about it. Despite the danger of fugu poisoning, this strange, ugly, and very poisonous fish is actually a very expensive, and very popular, kind of food in Japan. Customers pay up to$200 per person to eat a fugu meal. Because of the danger, fugu can only be prepared by chefs with a special license from the government. These chefs are trained to identify and remove the poisonous parts of the fish. Most people who die from eating fugu these days are people who have tried their hand at preparing the fish themselves. Fugu is said to be so delicious that it has even started to be imported into Hong Kong and the United States. Several tons of fugu are now exported from Japan every year.

Level 3: Difficult

Otaku is now a popular word used to refer to young people -mainly males- whose "lives "center around their hobbies, usually computers, computer games, comic books and animated films Akihabara has long been famous as a major electronic appliance shopping district, but these days it is also well known as a mecca for otaku.",FALSE The word otaku literally means"your house," and is also a polite form of "you."",FALSE Sometime in the 1980s, the word came into general use as a term for these young enthusiasts, because such people tend to confine themselves to their own rooms where they can indulge in their hobbies to their hearts' content without interference from others .",FALSE Such people also used the term otaku to refer to each other, a way of showing each other sympathy and respect, so to speak. ",FALSE At first the term had a largely negative connotation, suggesting an isolated, fanatical person with poor social skills (especially when it came to girls and dating) and no common sense. Since otaku tended to prefer each other's company and dressed oddly as well, the otaku subculture was often made fun of. The growing number of such people can be attributed to several factors. For one, there is young people's greater economic power, which has enabled them to buy the expensive electronic goods that are the backbone of their hobbies. Declining birthrates mean that most couples have on average only one or two children, which means that parents have more money to lavish on their children (grandparents can afford to be more generous, too). Another factor for the rise of the otaku is the rapid development of computer technology, which is particularly attractive to boys and young men. Modern parents often believe that computers are the key to success in today's information technology world, so they tend to buy expensive gadgets for their children. A final factor might be the growing emphasis on individuality in Japanese society. People today think being different is cool. Otaku is usually translated into English as "nerd." The word "geek" is used to refer to young people who are specifically interested in computer technology. Otaku, by the way, are no longer looked down upon as they once were. In fact, they are just as often admired for their expertise in one specific field.

Appndix D

Text content of personal-best shadowing for prosodic evaluation.

The MacDonald’s house has been broken into. A policeman has come to check it out. He finds a boy standing nearby. The policeman is now talking to the boy. He wants to know how the door of the MacDonald’s house was broken open. The boy said that it had already been broken before he and his friend went to the house. He said that they simply walked into the house. The police officer asked, “why were your fingerprints found all over the door? And why were your boots scratched? It was you who kicked the door open, wasn’t it? Why did you steal the stereo and the CDs? Did you just want to have a bit of fun, or were you trying to get some money? Now then, tell me the truth.” I don’t want to hear any more of your lies.

Acknowledgement

I would first like to thank my supervisor, Professor Nobuaki Minematsu, and Professor Keikichi Hirose for their supervision, advice, encouragement and support during my Ph.D. course.

I would also like thank Dr. Yu Qiao for many discussions, advice and cooperation that realize Regularzied-MLLR implementation.

I would also like to thank Professor Yutaka Yamauchi at Tokyo International University for educational insights and cooperation with Shadowing database collection, Dr. Yasushi Tsubota and Prof. Masatake Dantsuji at Kyoto University for providing the database of Basic English Words Read by Japanese Students.

I would like to thank the technician Mr. Noboru Takahashi for his help of providing computer hardware parts and networking environment of the laboratory. Many thanks to the secretary of Minematsu lab, Ms. Ikegami for help with various kinds of school paper work. I also would like to thank all the other members of Hirose/Minematsu lab for their friendship and creating a wonderful research environment together.

Bibliography

[1] F. Ramus, “Acoustic correlates of linguistic rhythm: Perspectives,” Proc. Speech Prosody, pp.115–120 (2002)

[2] 小川直樹，理屈で分かる英語の発音，ノヴァ・エンタープライズ（1999）

[3] 深澤俊昭，英語の発音パーフェクト辞典，アルク（2000）

[4] 田嶋圭一，山田玲子，山田恒夫，“Perceptual learning of English syllable rhythm by elderly Japanese listeners,” 第16回日本音声学会全国大会予稿集，pp.103−108（2002）

[5] A. Cutler, “Listening to a second language through the ears of a first,” Interpreting, vol.5, no.1, pp.1–23 (2001)

[6] 松本亨，英語で考える本，英友社（1968）

[7] R. B. Kaplan, “Cultural thought patterns in inter-cultural education,” Language Learning, vol.16, pp.1–20 (1996)

[8] 淺間正通，“英語教育における「異文化コミュニケーション」の普遍的視点をめぐって”，静岡大学情報学研究，vol.3，pp.1−10（1997）

[9] D. Crystal, English as a global language, Cambridge University Press (1995)

[10] Ministry of Education, Culture, Sports, Science, and Technology, Developing a Strategic Plan to Cultivate “Japanese with English Abilities”,

http:www.mext.go.jp/English/news/2002/07/020901.htm (2002)

[11] Ministry of Education, Culture, Sports, Science, and Technology, Regarding the Establishment of an Action Plan to Cultivate “Japanese with English Abilities”,

http:www.mext.go.jp/English/news/2002/07/020901.htm

[12] R. Hughes, “The MEXT English Education Reform Objectives and Student Motivation”, Journal of Regional Development Studies, pp.353-359, 2008

[13] A. Neri, C. Cucchiarini, and H. Strik, “Automatic speech recognition for second language learning: How and why it actually works,” Proc. Int. Congress of Phonetic Sciences (ICPhS’2003), pp.1157–1160 (2003)

[14] 竹林滋，斎藤弘子，英語音声学入門，大修館書店（1998）

[15] International Phonetic Association, Handbook of the International Phonetic Association, Cambridge University Press (1999)

[16] L. Bachman, Fundamental considerations in language testing, Cambridge University Press (1990)

[17] J. Bernstein, M. Lipson, G. Halleck, and J. Martinez, “Comparison of oral interviews and automatic tests of spoken language,” Language Testing Research Colloquium (LTRC’99) (1999)

[18] T. Hori, “Exploring Shadowing as a Method of English Pronun-ciation Training,” A Doctoral Dissertation Presented to the Graduate School of Language Communication and Culture, Kwansei Gakuin University. 2008

[19] S. Miyake, “Cognitive processes in phrase shadowing and EFL listening,” JACET Bulletin Tokyo: Japan Association of College English Teachers. Forthcoming

[20] H. Mochizuki, “Shadowing and English language learning,” Unpublished MA thesis, Kwansei Gakuin University, 2004

[21] P.W. Nye et al., “Shadowing latency and imitation: the effect of familiarity with the phonetic patterning of English,” Journal of Phonetics, pp.63–69, 2003

[22] Microsoft Corporation, ENCARTA インタラクティブ英会話（2000）

[23] Ohta et al, “A statistical method of evaluating pronunciation proficiency for Japanese words,” Proc Interspeech pp.2233-2236, 2005

[24] S.M. Witt and S.J. Young, “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning,” Speech Communications, 30 (2–3): pp.95-108, 2000

[25] A. C. Gimson, An introduction to the pronunciation of English, Edward Arnold Ltd.(1980)

[26] 渡辺和幸，英語のリズム・イントネーションの指導，大修館書店（1994）

[27] 河合剛，石田朗，“日本人の英語の発音評価の信頼性に関する実験的評価”，電子情報通信学会教育工学研究会，ET95-44（1995）

[28] N. Minematsu, C. Guo, and K. Hirose, “CART-based factor analysis of intelligibility reduction in Japanese English,” Proc. European Conf. Speech Communication and Technology (EUROSPEECH’2003), pp.2069–2072 (2003)

[29] Abhishek Chandel et al, ”Sensei: Spoken Language Assessment for CALL Center Agents,”Proc. ASRU, pp.711-716, 2007

[30] Motoyuki Suzuki, Tatsuki Konno, Akinori Ito, Shozo Makino,"Automatic Evaluation System of English Prosody Based on Word Importance Factor," Journal of Systemics, Cybernetics and Informatics, vol. 6, no.4, 2008

[31] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Comput. Speech Lang., vol.9, pp.171-185, 1995

[32] "The Hidden Markov Model Toolkit (HTK)"

http://htk.eng.cam.ac.uk/.

[33] J.S. Garofolo, "The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM", Technical Report NISTIR 4930, NationalInstitute of Standards and Technology, 1986

[34] John Garofalo, "Wall Street Journal-based Continuous Speech Recognition (CSR) Corpus", Linguistic Data Consortium, Philadelphia, 1994

[35] Minematsu et al, "English Speech Database Read by Japanese Learners for CALL System Development," Proceedings of Inter-national Conference on Language Resources and Evaluation, pp896-903, 2002

[36] Tanaka et al, "Acoustic models of language-indigent phonetic code systems for speech processing," Spring meeting of the Acoustical Society of Japan, pp191-192, 2001

[37] Y. Tsubota et al, "Practical Use of English Pronunciation System for Japanese Students in the CALL Classroom," Proc. ICSLP2004, pp1689-1692 , 2004

[38] Y. Tsubota et al, "An English pronunciation learning system for Japanese students based on diagnosis of critical pronunciation errors", ReCALL 16(1), pp173-188, 2004

[39] S. Kohmoto, "Applied English Phonology: Teaching of English pronunciation to the Native Japanese Speaker," Tokyo Tanaka Press, 1965.

[40] Kanters et al., "The goodness of pronunciation algorithm: a detailed performance study," Proc. SLaTE, CD-ROM, 2009

[41] E. Stevick, "Success with Foreign Languages: Seven Who Achieved It and What Worked for Them", Prentice Hall, 1989

[42] T. Murphey et al., "Meaningful communicative repetition", English Teaching Forum, 33, 4, pp. 37-38, 1995

[43] B. McMillan, "Shadowing: Powerful teachniques for developing listening and speaking skills", Interactive Vol.24. 10-12, 2008

[44] O. Engwall et al, ”Design strategies for a virtual language tutor,” Proc. ICSLP, pp.1085-1088,2004

[45] O. Engwall et al, ”Audio-visual phoneme classification for pronunciation training application, ” Proc. INTERSPEECH 2007, pp.702-705, 2007

[46] A. Samir et al, ”Enhancing usability of CAPL system for Qur ’an recitation learning,” Proc. INTERSPEECH 2007, pp.214-217, 2007

[47] Sharmisha Gray and John H.L. Hansen, ”An Intergraded Approach To The Detection And Classification of Accents/Dialects For A Spoken Document Retrieval System,” Proc. ASRU, pp.35-40 (2005)

[48] Stephane Dupont et al, ” Feature extraction and acoustic modeling: an approach for improved generalization across languages and accents,” proc. ASRU, pp.29-34 (2005)

[49] Kevin Lenzo, “The Carnegie Mellon University Pronouncing Dictionary”, http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[50] “Praat: doing phonetics by computer” www.praat.org

[51] Y. Okawa et al, “A speaker adaptation method for non-native speech using learners’ native utterances for computer-assisted language learning systems”, Speech Communication, Vol. 51, Is-sue 10, pp875-882, 2009

[52] Zhang, et al, “Generalized segment posterior probability for automatic Mandarin pronunciation evaluation,” Proc. ICASSP, pp.201-204, 2007

[53] C.Huang et al, “Improving automatic evaluation of mandarin pronunciation with speaker adaptive training (SAT) and MLLR speaker adaptation”, Proc. ISCSLP2008, pp37-40, 2008

[54] D.Luo, et al, “Quantitative analysis of the adverse effect of speaker adaptation on pronunciation evaluation” , Proc. ASJ spring meeting, 1-P-22, pp.173-176 , 2009

[55] D.Luo, et al, “Analysis and utilization of MLLR adaptation technique for learners’ pronunciation evaluation,” Proc. INTERSPEECH, pp.608-611, 2009

[56] Miwa, et al,"Analysis and comparison of the prosodic features for Japanese English and native English," IEICE technical report. Speech 101(744), 51-58, 2002-03-2

Publication

Journal Papers

[1] D. Luo, Y. Qiao, N. Minematsu and K. Hirsei, “Regularized Maximum Likelihood Linear Regression Adaptation for Computer-Assisted Language Learning Systems”, IEICE Trans. INF.&SYST., (2010, submitted)

[2] Y. Qiao, D. Luo, and N. Minematsu, "A study on optimal unsupervised phoneme segmentation and its invariant properties," IEEE Trans. on Signal Processing, (2010, submitted).

Peer-Reviewed International Conferences

[3] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, K. Hirose, "Regularized-MLLR Speaker Adaptation for Computer-Assisted Language Learning System, " Proc. INTERSPEECH2010, 2010-9 accepted.

[4] D. Luo, N, Minematsu, Y. Yamauchi, K. Hirose, " Speech Analysis for Automatic Evaluation of Shadowing," Proc. SLaTE, 2010-9, accepted

[5] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, K. Hirose, "Analysis and utilization of MLLR speaker adaptation technique for learners' pronunciation evaluation, " Proc. INTERSPEECH2009, pp.608-611 2009-9.

[6] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, “Automatic Assessment of Language Proficiency through Shadowing," Proc. ISCSLP2008, pp.41-44, 2008-12.

[7] D. Luo, N, Minematsu, Y. Yamauchi, K. Hirose, "Analysis and Comparison of Automatic Language Proficiency Assessment between Shadowed Sentences and Read Sentences," Proc. SLaTE, CD-ROM 2009-9.

[8] D. Luo, N. Minematsu, Y. Yamauchi, "Development of a CALL system to enhance ESL/EFL learners' skills of shadowing and reading aloud," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), Demo Session 2009-9.

[9] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, “Automatic Pronunciation Evaluation of Language

Learners’ Utterances Generated through Shadowing," Proc. INTERPEECH2008, pp.2807-2810, 2008-9.

[10] M. Suzuki, D. Luo, N. Minematsu, K. Hirose, "Improved structure-based automatic estimation of pronunciation proficiency," Proc. ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE), CD-ROM 2009-9.

[11] M. Suzuki, N. Minematsu, D. Luo, and K. Hirose,"Sub-structure-based estimation of pronunciation proficiency and classification of learners,"Proc. Int. Workshop on Automatic Speech Recognition and Understanding (ASRU'2009), pp.574-579 2009-12.

[12] C. Tsurutani, Y. Yamauchi, N. Minematsu, D. Luo, K. Maruyama, and K. Hirose, “ Development of a program for self assessment of Japanese pronunciation by English learners,” Proc. ICSLP’2006, pp.841-844,2006-9.

Other International Conferences

[13] Dean Luo, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose: "Automatic Assessment of Utterances Shadowed by Learners Using Speech Technologies", Language and Speech Science Workshop on L2, Waseda University, 2008-9.

Domestic (Japanese) Conferences

[14] 羅徳安，峯松信明，山内豊，牧野武彦，広瀬啓吉，"シャドーイング・音読発音評価を目的とした話者適応の分析と応用", 電子情報通信学会音声研究会，SP2009-32, p.51-56 (2009-6)

[15] 羅徳安，峯松信明，山内豊，牧野武彦，広瀬啓吉，"話者適応処理が発音評定精度に及ぼす影響に関する定量的分析", 日本音響学会春季講演論文集，1-P-22，pp.173-176 (2009-3)

[16] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, "A Study on Automatic Scoring Methods for Language Learners' Shadowing Productions", IEICE Technical Report, SP2008-29, pp. 55-60, 2008-9.

[17] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, "A Study on Automatic Scoring Methods for Language Learners' Shadowing Productions", IEICE Technical Report, SP2008-29, pp. 55-60, 2008-9.

[18] Dean Luo, Naoya Shimomura, Nobuaki Minematsu, Yutaka Yamauchi and Keikichi Hirose, "Pronunciation Analysis and Evaluation of Shadowed Utterances for Language Education," Proc. ASJ (Acoustical Society of Japan) 2008 Autumn Conference，pp.485-488, 2008-9.

[19] 羅徳安, 峯松信明, 鶴谷千春，山内豊，広瀬啓吉，"英語話者を対象とした日本語 CALL システムにおける発音評価"，日本音響学会秋季講演論文集，3-P-14，pp.363-364, 2006-9.

[20] 羅徳安，峯松信明，鶴谷千春，山内豊，広瀬啓吉，"日本語 CALL システムのための学習者発音分析とその自動評価"，電子情報通信学会音声研究会，SP2006-78，pp.13-18 , 2006-11

[21] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，"音声の構造的表象を用いた自動発音評定法の改善"，情報処理学会音声言語情報処理研究会，2009-SLP-77-17，pp.1-6 (2009-7)

[22] 鈴木雅之，羅徳安，峯松信明，広瀬啓吉，"発音構造を用いた話者の違いに頑健な発音評定と学習者分類", 日本音響学会秋季講演論文集，1-2-5，pp.243-246 (2009-9)

[23] 山内豊，峯松信明，羅徳安，川村明美，"音声情報処理技術を活用したシャドーイング自動評価システムの開発", 外国語教育メディア学会全国研究大会講演集, (accepted) 2009-8

[24] 鶴谷千春，山内豊，峯松信明，羅徳安，"日本語音声の難易度　～英語を母語とする学習者の場合～"，日本音声学会全国大会予稿集，pp.51-56 (2006-10).

Media Coverage

Our work has been reported by English learning magazine “Tadoku Tacho”.

特別取材

【ただいま開発中】
TOEIC® テスト
スコアが予測可能?!

シャドーイング自動評定システム

コンピュータでシャドーイングのパフォーマンスを自動評定し、しかもTOEIC スコアまで予想してくれるというシステムが開発されている。今までリスニング力を中心として、英語力を伸ばす学習法として注目されてきたシャドーイングだが、今度は言語運用能力を測る手段としての新たな可能性が出てきた。

取材・編集部

言語運用能力測定とシャドーイング

「コンピュータでシャドーイングを自動評定して、その結果でTOEICのスコアも予測できる？！」……2009年8月4日から6日にかけて神戸市・流通科学大学で開催された、第49回外国語メディア学会で驚きの発表が行われた。東京国際大学の山内豊先生と東京大学の峯松信明先生とを中心とするグループの発表だ。

最近の英語教育に関わる学会の大会では、タイトルに「シャドーイング」という文字が入った発表が増えてきている。しかし、学習・訓練方法としてのシャドーイングの効果や活用法ではなく、言語運用能力を測定する方法としてシャドーイングを取り入れ、成果を出しつつある研究開発を知ったのははじめてだ。

いったい、どういうシステムだろうか。音声認識ソフトを使って音声を書き起こす場合、明瞭な音声が必要となる。発声がくずれやすいシャドーイングで「コンピュータによる自動評定」はどのくらいの精度でできるのだろうか。まだ被験者の数は30人ほどと少ないとはいえ、そのシステムによるシャドーイング・パフォーマンスの評定とTOEICスコアの相関係数は高い数値を示しているという。

（1）シャドーイングの自動評定、（2）その結果を用いたTOEICスコアの予測、というふたつを結びつけた野心的なシステム開発を行っている東京大学大学院の情報理工学系研究科・電子情報学専攻にある峯松先生の研究室を訪問して、直接、お話を伺うことにした。

シャドーイングの評価

カラオケで歌うと、『あなたは何点です』と点数を出すシステムがあるでしょう。あれと同じように、ある量の英文をシャドーイングすると、『あなたは何点です』と

右から、英語教育の面からシステム開発に携わる山内豊先生（東京国際大学）、主にプログラミングを担当している羅徳安氏（東京大学：博士課程在籍中）、システム開発の全体を統括している峯松信明先生（東京大学）。東京大学・峯松研究室で。

30 2009 DECEMBER

Ph.D. Thesis

博士論文

東京大学大学院
工学系研究科
電子工学専攻

077090　羅　徳安 (Dean Luo)