Computer-aided training of Japanese learners of English to improve their performance of listening and speaking

(日本人英語学習者のための聴取・発音能力の向上を目的とした計算機援用型トレーニング)

張 昊宇

Zhang Haoyu

ID Number: 37-176877

Supervisor: Prof. Nobuaki Minematsu

Department of Electrical Engineering and Information Systems,
Graduate School of Engineering,
The University of Tokyo

Master Thesis

February 2019

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and acknowledgements.

Zhang Haoyu
February 2019

Acknowledgements

I would like to dedicate this thesis to all the people who have supported me both spiritually and materially. I am grateful to be the child of my parents, who never leave me and give a sense of safety to me. I want to thank my supervisor Prof. Nobuaki Minematsu and Lecturer Daisuke Saito, who gave me great help in my research, and all lab members who helped me a lot on my academical life and private life. Besides, I want to express my gratitude to all my friend and a particular person for appearing in my life. I am very grateful to all of them.

Abstract

The main aim of Computer-Assisted Language Learning (CALL), defined as ”the search for and study of applications of the computer in language teaching and learning.”, is to find methods for helping second language learners to learn language by using the computer. Reading, writing, speaking, and listening are the necessary language skills. Here, we found the possibility of computer use for helping Japanese English learners to practice listening skills and speaking skills. For CALL of listening, robustness in listening is investigated and compared between learners and native speakers, and then we tried to find methods of training to help learners to improve the robustness. For CALL of speaking, learners' pronunciation diversity or errors was modeled adequately with a G2P toolkit, phonetisaurus, to predict phonemic sequences that will be perceived by native listeners when they listen to any words pronounced by Japanese learners. Then, based on variations generated by phonetisaurus, phoneme error detection is carried out.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

Speaking, listening, writing and reading are the four skills that need to be acquired when learning a foreign language. Besides, if a person wants to have smooth communication, listening ability and speaking ability are especially important. Computer-assisted language learning (CALL) system is the system that is used for helping learners to study the second language by a technological method of the computer. Usually, there are four kinds of CALL system for assisting the four skills which needs to be acquired. In this thesis, we will focus on the listening and speaking. For the practice of listening, there are a large number of training materials which focusing on helping learners to enhance listening capabilities in the market. This training materials always consist of some questions in the textbooks and some audio CDs for clean speech, which ignores the fact that clean speech is rarely heard in real conversations. This is because extra-linguistic and environmental factors can easily degrade the clean speech and make it noisy. Although recordings of real and somewhat noisy conversations are also found in CDs, the quality degradation of the speech cannot be controlled and adjusted by users. If possible, quality degradation which is adjusted based on the listening capabilities of a target learner may be more helpful. For the practice of speaking, as pointed out in [2, 19], if the pronunciation cannot be understood by the other side of the conversation, no matter how much proper content can be constructed in words, nothing will be transmitted successfully at all. Thus, it is evident that pronunciation is extremely important when we start to learn a language. It is known that the English learners may be influenced by their native language in the following situation, 1) it is hard for them to pronounce certain sound in English which do not exist in their native language, 2) ways for combination of vowel and consonant is different from their native language, which means even though they may be able to pronounce the specific vowel or consonant separately, they may have trouble pronouncing some phoneme combination. It will help non-native speakers if we can train pronunciation for the English with an effective method.

This thesis is organized as two parts, the first half is about my experiment about the practice of listening and the second half is about my experiment about CALL system of speaking.

In Section 2, I will introduce some previous works on CALL system for listening and speaking. For listening, the sensation of listening in a different environment for English as a Second Language (ESL) is different. Then I will also introduce research on how to improve the sensation of listening with some training called High Variability Phonetic Training (HVPT). For speaking, works on non-native pronunciation error pattern generation will be introduced. Then model of g2p and how this model will be taken of use will also be introduced.

In Section 3, four kinds of speech modification by speech signal processing methods will be introduced.

In Section 4, I will introduce the detail of the listening experiment with the use of two kinds of deformation and the result of the analysis of that experiment. The detail and result of analysis of an additional experiment will be introduced. In the additional experiment, we made some listening materials with one kind of deformation technology and let students do the exercise with the listening material for a period of half a month. Then we did the same listening experiment again on the same students to see whether the listening materials make a difference.

In Section 5, information of corpus and technological implement and its principle of G2P model will be introduced. Based on the limitation of the dataset, we build a model for conversion from IPA to a set of phonemic transcription codes, which is called "AE-T2M". Then "JE-M2M" and "JE-G2M", which are the model of Japanese English Pattern Generation, will be introduced. The last part of this section is the introduction and analysis of the additional transcription

In Section 6, the result of our experiment on speaking will be introduced. Primary analysis of variations model will be done. Then results for the first trial of error detection and the final results of error detection by using additional dataset will be discussed.

In Section 7, a conclusion of both two works will be made. Besides, what needs to do to make an improvement will also ben introduced.

Chapter 2

Previous work

This chapter gives an overview of previous work on both listening practice and speaking practice.Then listening robustness and Japanese English error pattern generation will be discussed.

2.1 Previous works of listening robustness and HVPT

Acoustic variability in conversation has drawn the attention of linguistic researchers especially researchers of second language acquisition. And a method of training which is called as High Variability Phonetic Training (HVPT) has been verified experimentally to be effective in [12, 29, 8]. HVPT is a technology which uses multiple voices instead of single one voice with various phonetic contexts to train learners. In their research, they claim that stimulus variability is the core of HVPT, which has effectiveness. Besides, the differences of listening ability, which can also be called as listening robustness, also have been researched. With the use of acoustically degraded speech materials are found in L2 studies, they clarify differences in listening ability between learners and native speakers.[3] examined learners' listening abilities by using speech samples with babble noise. The experiment of comparing the learners' listening abilities in the two cases that the babble noise is L1 and L2 is carried out by [11]. The performance of phonological perception in the presence of reverberation, as well as background noise, was investigated in [15–18]. From the research before, we can know the importance of stimulus variability. However, studies prepared stimulus variability manually, e.g. collecting speech samples from multiple talkers, or using elementary signal processing techniques such as waveform addition or convolution.

In this thesis, to enhance the acoustic variability in stimuli, the core of HVPT, advanced speech technologies are introduced. Recently, voice conversion (VC) including voice morphing has been a very hot topic among speech engineers and voice conversion challenge was held [1]. GMM-based and DNN-based statistical conversion methods have been developed and

combined [7, 10]. Technically speaking, VC can be divided into two approaches, one is theory-based and the other is corpus-based. If the mechanism of voice changes is known and is divided into several steps, each step is technically implemented to generate new voices from original voices. If the mechanism seems complicated or unknown, a parallel corpus between original voices and new voices, e.g. a parallel corpus of two speakers, is used to estimate a mapping function between the original voice space and the new voice space. The latter is called *statistical* voice conversion, which draws attention of many speech engineers. However, the authors admit that the naturalness of converted voices is not so high as to be used in educational contexts. In this paper, considering the quality of converted voices and what kind of voice changes is needed for educational purposes, theory-based VC methods are adopted to prepare new acoustic variability for HVPT.

2.2 Previous works of Japanese English error pattern generation for phoneme error detection

For practicing pronunciation, especially English, traditional methods are to have a pronunciation teacher who is a native English speaker and has sufficient acoustic-phonetic knowledge to have a conversation and give correction about pronunciation to the learner at the same time, which is obviously costly. And in pronunciation guidance, it is proved to be useful to return appropriate teaching to the learning's utterance[26]. However, it is impossible to have a teacher to tell you where do you make a mistake and how to correct it at any time when you speak other language. So methods for teaching simulation by using speech signal processing technology has been tried[5]. In this case, it is necessary to model beforehand on what kind of pronunciation errors the learner will is expected to commit, which can also be called as English error pattern generation or pronunciation error prediction.

First of all, error pattern can be summarized based on the teacher's experience. Normally, three kinds of errors are considered, substitution error, insertion error, and deletion error. I. Thompson summarized some rules about substitution errors in Japanese English from vowels and consonants [28]. Besides, some rules about vowel epenthesis in Japanese speakers' L2 English is summarized by [31]. However, error pattern based on teacher's experience may have an insufficiency because those hand-crafted rules only focus on certain phoneme, which ignores the context information of the whole sentence and phonetic environment of the whole paragraph.

Because of the great development of machine learning, we can generate English error patterns which can consider context information automatically by data-driven methods. If we

can prepare a corpus that has pairs of canonical phoneme sequences and phoneme sequences generated by second language learners, we can easily model the non-native pronunciation error pattern. Meng transcribes pronunciation errors at the phoneme level of Hong Kong people's English utterance and model the pronunciation deviation using the transcribed corpus. By using the joint sequence model as G2P, modeling of pronunciation deviation as a phoneme sequence, which can be considered as sequential modeling, a model dependent on phoneme environment can be made.[22].

We can also view the error pattern generation as a machine translation task, which means we can view the canonical phoneme sequences and phoneme sequences for the pronunciation of second language learners as two different languages. So the machine translation framework can also be used if sufficient data is prepared. Stanley[25] applies statistical machine translation framework to model phonological errors for Japanese learners of English and Korean learners of English, with corpora called RS-JLE and RS-KLE. His system is divided into three part, translation model, language model, and decoder. For translation model, GIZA++ and Moses Trainer are used and for language model, IRSTLM is used. Then Moses Decoder combines language model, translation model, and native lexicon to generate Non-native lexicon.

Chapter 3

Technical implement of Acoustic Variabilities

In this chapter, the system that we made to generate acoustic variabilities will be introduced. In our system, attention is paid to extra-linguistic and environmental factors of speech variability. This is because they are generally uncontrollable to speakers and listeners. Four types of these speech modifications are technically implemented.

3.1 Vocal tract length manipulation

In animation movies, the voices of human actors and actresses are often manipulated acoustically to make the voices of insects and animals for example. These funny voices can become troublesome to learners. Here, vocal tract length manipulation, i.e., frequency warping, is introduced in the form of linear transform in the cepstrum domain[21]. Pitch manipulation is also done automatically according to the manipulated vocal tract length.

3.1.1 Basic knowledge used in this section

First of all, I want to introduce the basic knowledge of speech signal processing used for vocal tract length manipulation.

Short frame segmentation

Although speech signal is unstable and time-varying, in a "short" time, which ranges from 10ms to 30ms, it can be considered as a kind of stable time-invariant signal. When doing short-term analysis, the speech signal is segmented as a length of 10-30 ms to be analyzed, where the

small unit is called as frame. Thus, for the overall speech signal, each frame feature parameter constitutes a feature parameter time series. And because the rectangular window may lead the adjacent frames to be discontinuous, which may lead to bad inhibition of spectral leakage. We usually consider to use window function, and *Hamming window* is usually used.

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi}{N-1}) & 0 \leq n \leq N-1 \\ 0 & otherwise \end{cases} \tag{3.1}$$

Fourier transform

Then in order to enter the cepstrum domain. We need to do a series of processing. The core of this processing is Fourier transform. The Fourier transform, which is a "time domain to frequency domain" transformation, is actually equivalent to a decomposition or base operation.

For continuous signal, Fourier transform and its inverse operation as following equation.

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x)\, e^{-2\pi i x \xi}\, dx \tag{3.2}$$

$$f(x) = \int_{-\infty}^{\infty} \hat{f}(\xi)\, e^{2\pi i x \xi}\, d\xi \tag{3.3}$$

When we process signal in computer, we use discrete Fourier transform (DFT), which is defined as following equation.

$$\begin{aligned} X_r(k) &= \sum_{n=0}^{N-1} x_r(n)cos(2\pi nk/N) + x_i(n)sin(2\pi nk/N) \\ X_i(k) &= \sum_{n=0}^{N-1} -x_r(n)cos(2\pi nk/N) + x_i(n)sin(2\pi nk/N) \end{aligned} \tag{3.4}$$

And inverse DFT is defined as following

$$\begin{aligned} x_r(n) &= \frac{1}{N}\sum_{k=0}^{N-1} X_r(k)cos(2\pi nk/N) - X_i(k)sin(2\pi nk/N) \\ x_i(n) &= \frac{1}{N}\sum_{k=0}^{N-1} X_r(k)cos(2\pi nk/N) + X_i(k)sin(2\pi nk/N) \end{aligned} \tag{3.5}$$

When we process real time series by DFT, such as speech signal, we can simply let $x_i(n) = 0.0$. Because there are a "fast" version of DFT, which is implemented by an algorithm to calculate at high speed on the computer and called as fast Fourier transform(FFT). Actually we use the program for FFT instead of DFT to do this process.

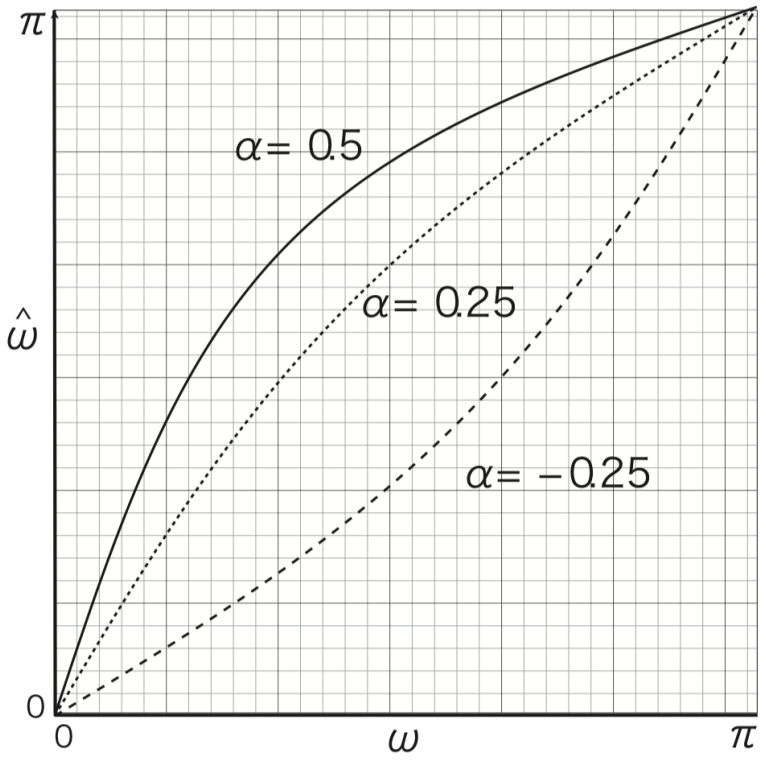


Fig. 3.1 Examples of frequency warping functions for different values of α [23]

Cepstrum

Then, when we get the $X_r(k)$ and $X_i(k)$ from the result of DFT/FFT, we can get logarithm power spectrum by following equation.

$$P(k) = \log 10 \left[\frac{1}{N} \{X_r(k)^2 + X_i(k)^2\} \right] \tag{3.6}$$

Then cepstrum will be get from doing inverse DFT on the $P(k)$.The logarithmic power spectrum is a real number sequence and it is necessary to perform inverse DFT with the imaginary term = 0. Since it is an inverse FFT of a logarithmic power spectrum rather than a spectrum, the result obtained is different from the original speech waveform.

3.1.2 Process of Vocal tract length manipulation

Vocal tract length manipulation, i.e., frequency-axis warping, can be represented in the form of linear transform in the cepstrum domain [21]. In this paper, generation of giant voices or fairy voices was realized by applying those linear transformations to original voices. By letting

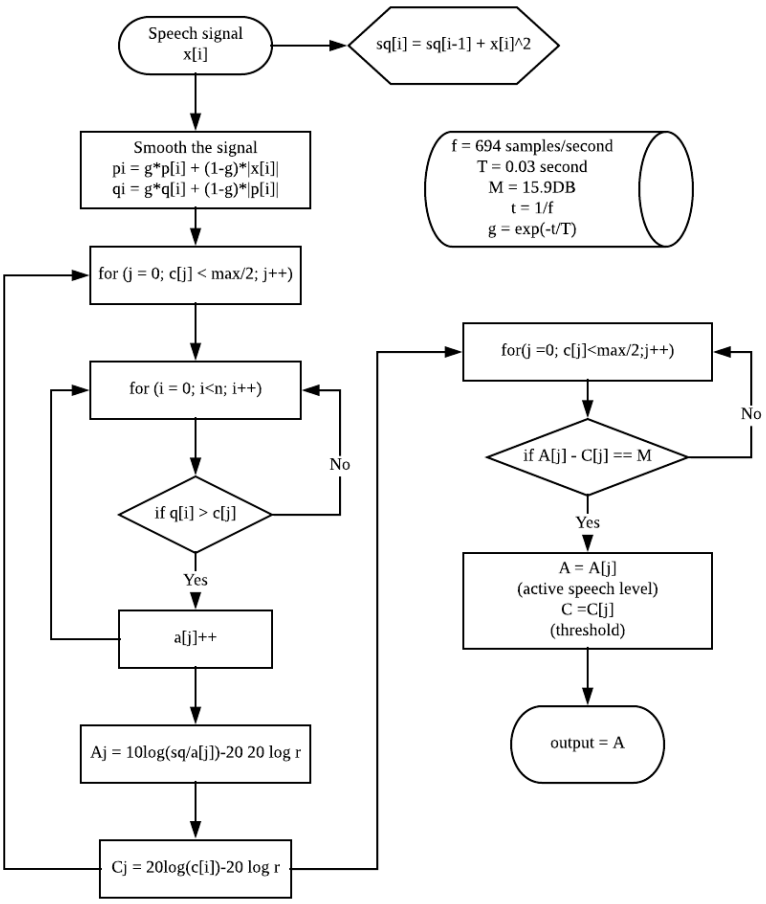


Fig. 3.2 Calculation flow of active level

ω and $\hat{\omega}(0 \leq \omega, \hat{\omega} \leq \pi)$ be the normalized angular frequency before and after warping, the warping is represented by the first order all pass function [23],

$$\hat{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \tag{3.7}$$

where $z = e^{j\omega}, \hat{z} = e^{j\hat{\omega}}$ and α is a real number where $|\alpha| < 1$. When $\alpha < 0$, the frequency axis is shrunk to simulate the voices of taller speakers. When $\alpha > 0$, the axis is extended to generate those of smaller ones. This transformation can be represented by linear transformation in the cepstrum domain [21]. Further, pitch manipulation is also done automatically according to the resulting vocal tract length.

3.2 Addition of environmental noises

By using publicly available noise databases, it is easy to add environmental noises to input clean speech with a desired signal-to-noise ratio. Various kinds of noises such as train noise, car noise, restaurant noise, babble noise, etc can be added. Here, instead of using SN ratio natively, we add the noise with a calculation of active level. Figure 3.2 shows how to calculate the active level of an input signal. Active level, also called as Active speech level, is a criterion measured as the presence of speech for the same proportion of time when a human listen to it. The addition of noise is based on *cleanGain*, which is defined as the following equation.

$$\begin{aligned} A_{clean} &= activelevel\,[cleanSpeech] \\ A_{noise} &= activelevel\,[noise] \\ cleanGain &= \sqrt{\frac{A_{noise}}{A_{clean}}} * 10^{\frac{SNR}{20}} \end{aligned} \tag{3.8}$$

Then we will add the noise based on following equation.

$$output = cleanSpeech * cleanGain + noise \tag{3.9}$$

3.3 Addition of reverberation

Every good researcher has experiences of troubles in listening to oral questions raised at the back of a big hall after he/she gives a research talk. A big hall generally has deep reverberation effects, which make oral questions difficult to understand. This is definitely the case with non-native researchers. The noise such as reverberation is multiplicative noise which is different from simply adds the noise to the input signal. The flow of multiplicative noise addition is 1) do DFT or FFT on both input signal and noise impulse, 2) multiply them in the frequency domain, 3) do inverse DFT or inverse FFT to get the mixed signal in time domain.

When we process different input of speech signal, overlap-add method (OA, OLA) is used to evaluate the discrete convolution of a very long input with a finite impulse response(FIR) filter, which is defined as the following equation.

$$\begin{aligned} y[n] = \left(\sum_k x_k[n-kL]\right) * h[n] &= \sum_k (x_k[n-kL] * h[n]) \\ &= \sum_k y_k[n-kL], \end{aligned} \tag{3.10}$$

Where, $y[n]$ is the output and can be written as a sum of short convolutions by OLA method.

By using publicly available databases of impulse responses, it is easy to simulate reverberation effects of a big hall, a big cathedral, etc. The size of halls or cathedrals can be controlled by changing the length (tail) of their impulse responses.

3.4 Simulation of channel distortion

The channel distortion of 3G mobile phones and 2G mobile phones can be simulated by using FFmpeg, where two codecs of AMR and GSM are simulated. Further, notorious channel distortions of air-traffic control, police radio, and taxi radio are simulated by doing waveform clipping and frequency modulation/demodulation, where transmission noises are artificially inserted just before the process of demodulation.

Distorted speech due to radio communication is simulated as follows. The amplitude of a given speech waveform is multiplied by integer n, and the samples above a fixed threshold are clipped. After applying low-pass filtering with 8kHz cut-off, the clipped signal $v(t)$ is frequency-modulated,

$$f(t) = A_c \cos\left(2\pi f_c t + \phi_c + K_{fm}\int_0^t v(t)dt\right), \tag{3.11}$$

where $f(t)$ is the signal after modulation, A_c is amplitude, f_c is the modulation frequency, and K_{fm} is constant. In this study, f_c is 100kHz. To implement distortions, three kinds of noises are introduced, one is $v(t)$ frequency-modulated with 110kHz , one is noise signals frequency-modulated with 100kHz, and the other one is oise signals frequency-modulated with 110kHz. After adding these three kinds of modulated signals, the resulting signals are frequency-demodulated with 100 kHz.

An example of original speech waveforms and its radio-distorted version is shown in Figures 3.3 and 3.4, respectively.

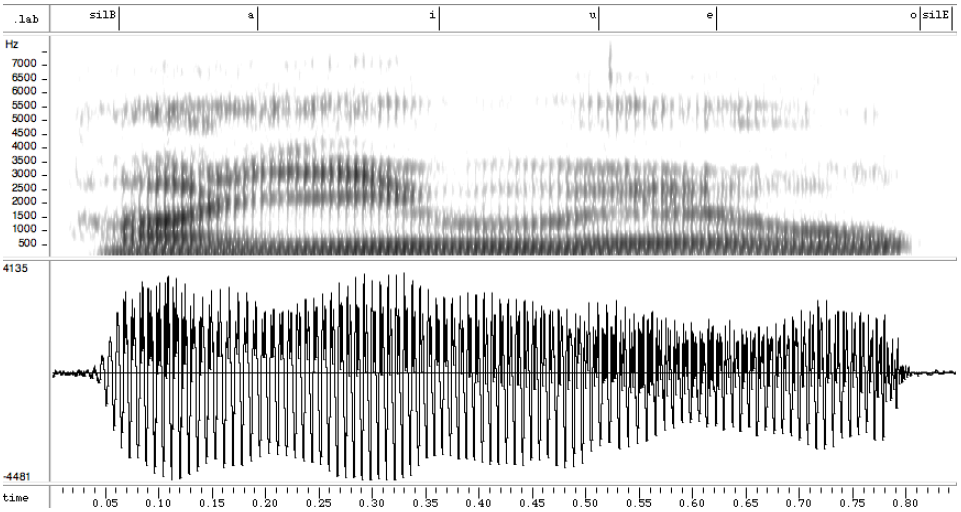


Fig. 3.3 An original speech sample

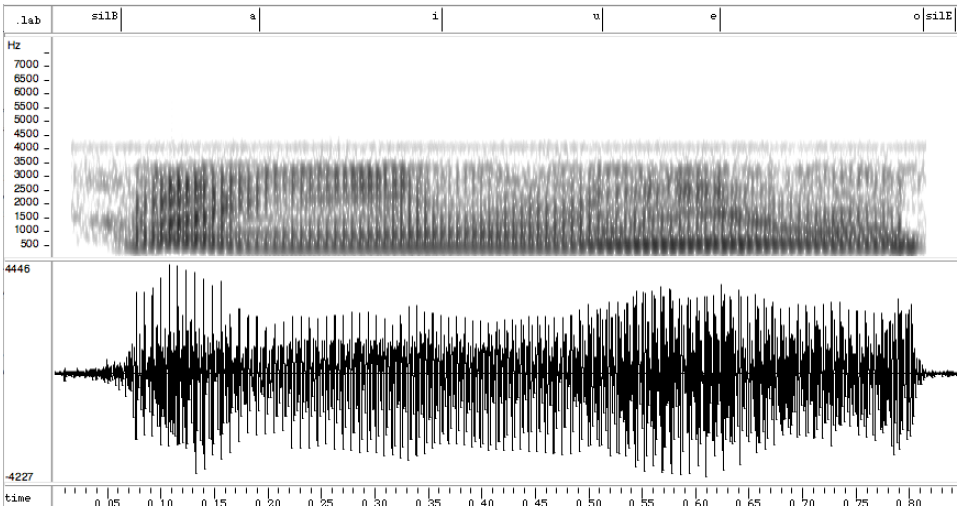


Fig. 3.4 Result of distortion based on radio communication

Chapter 4

Harsh Listening Comprehension Tests Using Distorted Speech

By using speech distorted by the toolkit developed in Chapter 3, listening tests were carried out with Japanese college learners. The learners firstly took part in the pre-test, then special listening drills were imposed on them. Finally, a post-test, where the same listening tests are used, was carried out after half a year.

As for the task imposed on learners, listening comprehension was adopted not identification of phonemes adopted in [15–18]. In Japan, as well as TOEFL and TOEIC, we have another well-known English proficiency test, called EIKEN [27]. EIKEN tests of Grade 2 are designed for high school students, listening questions of which are used as original questions in this paper. For our test, we only use the listening part of EIKEN G2, which consists of two part, questions using dialogues (Part A) and those using monologues (Part B). Participants are told to answer a four-choice question after they finish listening to a short dialogue between a male speaker and a female speaker (Part A) or a short monologue (Part B). Normally, the chance level of this test is 25% because participants can randomly select a choice even though they can't understand any information from the speech. Meanwhile, Part B is more difficult than Part A.

Our harsh listening tests are based on EIKEN G2 listening tests. A test is composed of 16 dialogue-based questions (Part A) and 16 monologue-based questions (Part B). The following three kinds of distortions are applied to the original clean utterances, when we prepared the distorted-version dialogues and monologues.

1) Male voices are converted to sounds with a male whose vocal tract is 1.3 times longer than his original tract and female voices are converted to to sounds with a female whose vocal tract is 0.7 times longer than her original tract. Because these sound like giant voices and fairy

voices, this kind of sounds are referred to as GF henceforth.
2) Any voice is converted into voice by "simulation of channel distortion". Because we adjust the parameter based on air traffic control sounds which the pilot listen to, this kind of sounds are referred to as ATC henceforth.
3) Combination of 1) and 2). Male/female voices are converted to voices of a giant/fairy pilot, called as GF+ATC.

Namely, four types of voices are prepared for our harsh tests, which are the original voices, GF sounds, ATC sounds and GF+ATC sounds. In order to keep the number of each distortion maintain same, 16 dialogue-based questions are composed of the four types of voices, each containing four questions. The 16 monologue-based questions are prepared similarly. We set the order of presenting the four types of voices is randomly, and the time length of a session of all the 32 questions is about half an hour.

4.1 The Pre-test

4.1.1 Subjects and experimental setup

Participants of the pre-test consist of 125 English learners who are Japanese college students and 2 native English speakers without hearing problems. The environment of the Japanese English learners is a normal classroom, where listening questions and material were presented by two loud speakers attached to the wall of the classroom. And the environment for native speakers are their private rooms and listening material are presented by the built-in speakers of their PCs.[32],[33]

We prepared four tests and each test is composed of 16 dialogue-based questions and 16 monologue-based questions, which will take about half an hour for a single test. The native speakers joined in all the four tests but each learner only took part in two, which means approximately 30 learners took part in each test. Besides, all the question and materials are obtained from the official EIKEN G2 tests, we can suppose that the difficulty level among the four tests can be ignorable.

4.1.2 Correct answer rates of learners and native speakers

Shown in table 4.1, the correct answer rates can be compared between the entire learners and the native speakers. For each part (A or B) and each type of voices, the correct answer rates are calculated independently. We also did the analysis of variance (ANOVA), where significant difference is observed ($p<0.001$) between any two distortion types in the learners' rate. That

result of correct answer rates in the learners' data are decreased in the order of Original, GF, ATC, and GF+ATC. When we compared the correct answer rate between GF and ATC, the ATC distorted utterances is more difficult to comprehend than the former one. The correct answer rates for GF+ATC are around 25%, which indicates that interpretation of both dialogues and monologues in GF+ATC is impossible to be done correctly at all. When we turned to the native speakers expect for GF+ATC , they can answer original, GF and ATC completely, and even for the most difficult GF+ATC, they also got a correct answer rate nearly 90%. The comment from two native speakers indicates that the difficulty order is Original, GF, ATC, and GF+ATC. But even the distortion of ATC and GF+ATC does not affect them so much when they are answering the four-choice question. A striking contrast between learners' weakness and native speakers' robustness is exhibited especially when we focus on the results in ATC and GF+ATC. For the task of phonological identification under the distortion of background noise addition and reverberation simulation, such huge difference was not observed [15–18]. This is probably because of differences of the experimental conditions that were adopted. In this study, listening comprehension was imposed on learners with radio communication distortions added to audio.

4.1.3 Analysis based on the learners' TOEIC scores

Before our tests, many participants took the TOEIC test. So we can divide them into three groups according to their TOEIC scores, by 400-600, 600-800, and 800-990, which can also correspond to the beginners' level, intermediate level and advanced level roughly. The correct answer rates which are separated for each level are shown in Table 4.2. From an overall view, the students who belong to the higher level group can have a higher correct answer rate. But no matter which group it is, even for advanced learners, a big drop in ATC can be found. And their correct answer rates are around 25%, which means they failed to understand monologues and dialogues in this condition.

We can know that the radio sounds are extremely important in an emergency, which will lead people to dangerous situations if they cannot comprehend this kind of sound, e.g. air traffic control, police radio, and taxi radio. In our experiment, the result suggests that native speakers have no trouble comprehending this type of acoustic distortion, while even advanced learners have huge trouble understanding it. Is there any significant difference between the listening strategy of native speakers and that of learners? Although in this paper, we do not discuss strategic differences between the two groups of listeners, it can be speculated that radio communication distortion may be a good tool to enhance the robustness of learners' listening comprehension. By introducing listening drills with radio communication distortions, learners may acquire a new strategy for listening.

Table 4.1 Correct answer rates of learners and native speakers for the four types of listening tests [%]

Part	Subject	N	Orig.	GF	ATC	GF+ATC
A	learners	125	68.9	59.2	34.2	25.9
	native	2	100	100	100	93.8
B	learners	125	55.6	46.1	31.4	25.4
	native	2	100	100	100	87.5

Table 4.2 Correct answer rates of each TOEIC-based group of learners for the four types of listening tests [%]

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	28	58.3	50.0	30.6	32.8
	600–800	52	78.2	62.0	35.1	23.4
	800–990	15	81.5	79.6	45.4	25.0
	native	2	100	100	100	93.8
B	400–600	28	42.2	41.1	23.9	25.0
	600–800	52	63.0	48.9	31.7	25.8
	800–990	15	74.1	67.6	41.7	24.1
	native	2	100	100	100	87.5

To verify this hypothesis, we did not stop at this step and planned to carry out the same test after the students have been trained with some distorted materials. The pre-test was held in July 2017 and we waited for four and a half months, during this period, we didn't inform the students that there is a post-test. We did this because we need to wait a long time to ensure that students did not remember anything of the pre-test. Besides, after the pre-test, the correct answer is not given to the students. Then after such a long time, an 18-day listening drill was carried out and then the same test was carried out again in December 2017. We will describe how we design this listening drill in the following section.

4.2 Listening Drill

4.2.1 Listening drills for 18 days

In this listening drill, five dialogue-based questions(Part A) and five monologue-based questions(Part B) were prepared for every day. Because we prepared this drill for 18-days so the

total is 180 questions. As was done in Section 4.1, the source of this drill materials are all from the official EIKEN G2 tests which are different from those used in 4.1.

Considering the fact that the difficulty level of materials with ATC distortion in pretest maybe too hard for the learners to practice. 4 levels of ATC distortion were prepared. The difficulty grows in the order of level 0, 1, 2, 3. Level 0 is equal to the speech quality of a classical and standalone telephone with no ATC distortion and level 3 is the difficulty of listening questions with ATC distortion in our harsh listening test(pre-test).

We prepared all the four kinds of ATC distortions of all the 180 listening questions, which means that 720 listening materials were generated in total. And we just used the ATC distortion and GF and GF+ATC were not used. The audio material and the questions are provided by a web site for learners to download for free. Besides, ranscriptions of all the materials were also available on that web.

4.2.2 Procedure of listening in the drill

If we did not give any prior instruction, the learners will use the prepared listening materials in their own ways, which will lead their behaviors uncontrollable. To avoid this, an experienced teacher of English who are at collaboration in this research gave the following instructions.

"You have five questions everyday. Start with level 3 of question 1. If you do not understand well what is said, then, repeat listening to level 3 of question 1 up to three times. If you still do not understand, then, use level 2 of question 1 and listen to it up to three times. After that, you may use level 1 and level 0. This is the end of question 1 and go to question 2."

4.3 The Post-test

4.3.1 Subjects of the post-test

Because we didn't inform the students that there will be a post-test in the pre-test, it is hard to let all the 125 students join the post-test. So 63 students from the 125 students who participated in the first listening test underwent the same test again in the same environment. In the pre-test, 4 sections of tests were carried out but due to time constraints imposed by the college curriculum, only the first two sections were carried out. The amount of each section is 16 dialogue-based questions and 16 monologue-based questions, which is the same as the pre-test. [34] [1].

[1]The official EIKEN G2 listening test is composed of 15 dialogue-based questions and 15 monologue-based questions.

4.3.2 Effectiveness of ATC-based HVPT

Not all the 63 students had their TOEIC scores, so only 55 students who satisfy the following condition, 1)joined both pre-test and post-test, 2)joined listening drill, 3)have TOEIC scores are considered. The result for 55 students in the pre-test is shown in Table 4.3. The result for 55 students in the post-test are shown in Table 4.4, where the post-test was held after a week after the 18-day listening drill. Table 4.5 shows the differences between Table 4.3 and Table 4.4, which was indicated in the form of incorrect answer reduction rate (IARR) to evaluate the effectiveness of ATC-based HVPT. And IARR is defined as

$$\text{IARR} = \frac{\text{IAR of the pre-test} - \text{IAR of the post-test}}{\text{IAR of the pre-test}},$$

where IAR means incorrect answer rate.

Since only the ATC-based distorted materials are used in the 18-day listening drill. So let's focus on the IARR of ATC first. We can indicate that IARR is always positive no matter of the proficiency level of students, which means the students have progress after the listening drill. Besides, for advanced learners, the effectiveness is more obvious, where nearly half of the error they made in the pretest are corrected.

Next, let us focus on the results of GF. the IARR is always positive and it is surprising that values of IARR in GF are generally higher than those in ATC. It can be indicated that the effectiveness of listening drill with ATC-based distortion can also make effects in other distortions. However the value of IARR become smaller in part B, so it is seen that robustness transfer does not occur always. So we can only say the for advanced learners, transfer of robust listening is stable.

When turning to the case with Original questions. Only the IARR with a large value is found in the case of advanced learners. The result of the experiment indicates that only the advanced learners can benefit effectively from our listening drill, while for other students, listening drill is not so effective as the advanced learners do. And the transfer of robust listening is also found only for advanced learners. The reason for this difference may be that the ability to comprehend the ATC sounds require well-integrated knowledge of English (phonology, syntax, semantics, pragmatics, etc). Or the advanced learners understand the the meaning and importance of this harsh test and train themselves more with the training materials. In either way, the authors think maybe the even we make four kinds of difficulty in the listening drill, these materials are still too hard for beginning learners and intermediate learners, so maybe the listening drill may have similar effects for them if the materials have easier materials, which is also a future work of this study.

Table 4.3 Results of the pre-test of the 55 learners [%]

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	66.7	48.3	25.0	41.7
	600–800	32	77.3	65.6	38.3	25.8
	800–990	8	84.4	84.4	43.8	21.9
B	400–600	15	50.0	43.3	28.3	23.3
	600–800	32	65.6	48.4	39.1	30.5
	800–990	8	78.1	62.5	37.5	28.1

Table 4.4 Results of the post-test of the 55 learners [%]

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	70.0	66.7	26.7	35.0
	600–800	32	73.4	73.4	40.6	32.8
	800–990	8	96.9	96.9	75.0	40.6
B	400–600	15	66.7	48.3	38.3	23.3
	600–800	32	61.7	51.6	42.2	35.2
	800–990	8	87.5	84.4	62.5	31.3

Table 4.5 Incorrect answer reduction rate (IARR) [%]

Part	TOEIC	N	Orig.	GF	ATC	GF+ATC
A	400–600	15	9.9	35.6	2.3	-11.5
	600–800	32	-17.2	22.7	3.7	9.4
	800–990	8	80.1	80.1	55.5	23.9
B	400–600	15	33.4	8.8	13.9	0
	600–800	32	-11.3	6.2	5.1	6.8
	800–990	8	42.9	58.4	40.0	4.5

Chapter 5

Japanese English error pattern generation for phoneme error detection

5.1 Error pattern generation and Phonetisaurus

5.1.1 Error pattern generation

Here, we set the phoneme sequence intended by the learners as /x_1 x_2 ... x_N/, and set phoneme sequence perceived by native speakers as /y_1 y_2 ... y_M/, the model of Japanese English error pattern generation is set up to find the correspondence between these two kinds of phoneme sequences, which is similar to the task of grapheme to phoneme conversion, that is G2P task. While doing speech synthesis, it will be troublesome if the word is rare and was not recorded in the dictionary. For example, It is necessary to guess phone sequence of the word of a person's name, even it can not be found in the dictionary. So here we use G2P model to deal with it. G2P model can learn the relationship between words spell and their pronunciation (phone sequence normally), and then it can generate the phone sequence which corresponds to the grapheme sequence of a new word. Generally, the task of G2P is to build a model that learns the relationship between paired corpus of sequences, which contains only the finite symbols. By using this model, prediction of the other symbol string for an arbitrary symbol string can be made. In the usual G2P task, a pronunciation dictionary is prepared, and the correspondence between the grapheme string and the phoneme string is automatically learned from the pronunciation dictionary, and a phoneme string of unknown words can be predicted.

5.1.2 Phonetisaurus

We used a G2P toolkit called as phonetisaurus [20], which is implementation of joint sequence model, and the corpus of Japanese English error pattern to build a G2P based pronunciation error pattern generation model. Maybe some Deep Learning methods may have a better performance if enough data are available. Because of the fact that in our experiment there are not so many data and it is very hard for us to gain new data, we consider that phonetisaurus is technically enough for our task. The phonetisaurus G2P, an open source toolkit, is a variation on joint sequence model in the WFST framework. This approach is in four steps. The first step is the data preparation, which is normally collecting pronunciation lexicon for training. However, in our task of pronunciation diversity modeling, "pronunciation lexicon" can be paired corpus between phonetic symbols, paired corpus between grapheme and phonemic symbols or paired corpus between phonemic symbols and phonemic symbols.

The second step is to make alignment between the "graphemes" and "phonemes" in the lexicon by EM algorithm. Based on pre-set parameters, word-pronunciation pairs can be constructed. Then expectation of word-pronunciation pairs can be calculated by an expectation function. By expectation maximization step, the parameters can be updated and this operation will repeat for several iterations.

The third step is to train a joint-sequence n-gram model if the alignment between the "graphemes" and "phonemes" is done. In the area of traditional G2P task, joint-sequence n-gram model is a considerably successful model so we think it is enough for our special G2P task. the joint-sequence n-gram model is the n-gram model which is slightly different from traditional ones used for training a language model [4]. In the language model, we need to calculate the probability of a sentence. And because we can't calculate $P(S)$ directly, we need to calculate based on prior probability.

$$p(S) = p(w_1, w_2, w_3, ..., w_n) \tag{5.1}$$

$$p(w_1, w_2, w_3, ..., w_n) = p(w_1) * p(w_2|w_1) * p(w_3|w_2, w_1).....p(w_n|w_1, ..., w_{n-1}) \tag{5.2}$$

And it is very costly to calculate the current word based on the all words before, we need to simplify it and it is very costly to calculate the current word based on all words before, we need to simplify it by Markov chain. N-gram is a model of Markov chain, which is a stochastic process that transfers from one state to another in state space. This process has a nature of "no memory", which means the probability distribution of the next step can only be determined by the current state and is independent of the steps before. This will greatly reduce the length of the above formula, which becomes the following equation.

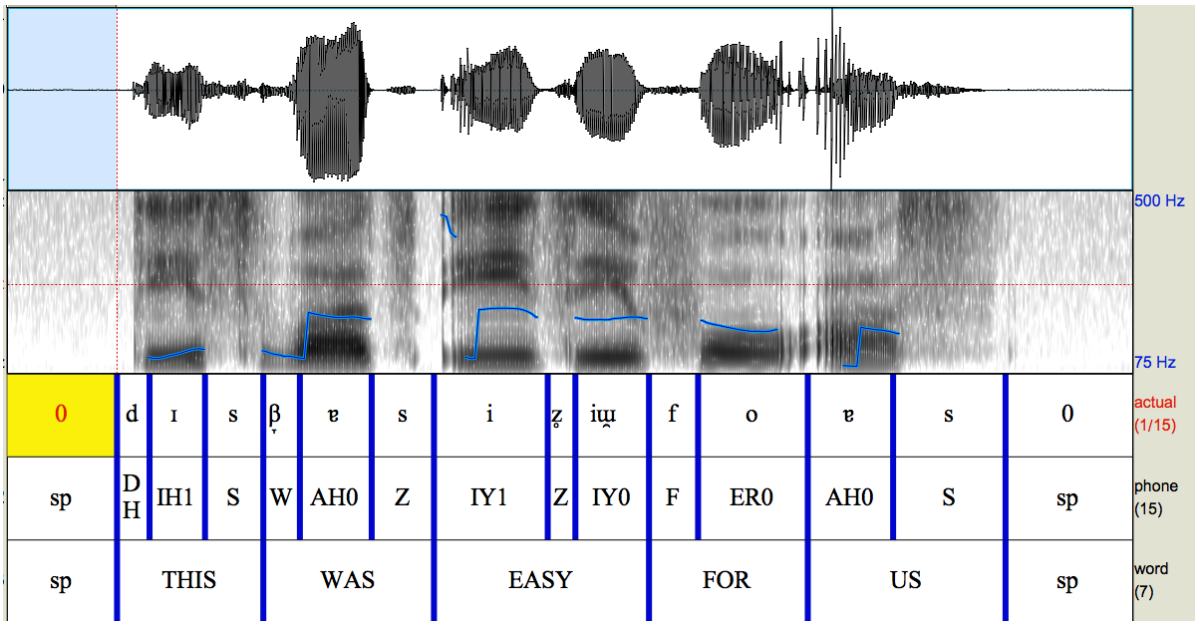


Fig. 5.1 An transcription sample

$$p(w_1, w_2, w_3, ..., w_n) \approx \prod_i p(w_i | w_{i-k} ... w_{i-1}) \tag{5.3}$$

In the joint-sequence n-gram model, it learns the relationship of joint (G,P) chunks instead of words in normal language model.

The fourth step is to transfer the joint-sequence n-gram model into WFST and pronunciations for the new word can be predicted by making the weighted combination in order to compute the intersection of the WFST representation of the target word and the joint n-gram model.

5.2 Corpus and pronunciation diversity modeling

5.2.1 Phonetic transcription

Here, we have two kinds of corpus from ERJ(English Read by Japanese) corpus. In ERJ corpus, 807 sentences and 1,009 word were read aloud by 100 male and 100 female native Japanese speakers in 20 places. First one is the corpus transcribed in detail with phonetic symbols with diacritics, second is the one contain no label. The speakers were 200 Japanese English learners (100 university students and 100 female university students) and 20 adult American native speakers who speak GA (General American) (8 men and 12 women). In the ERJ corpus, a subset of sentences are from the phonemically-balanced sentence set defined in TIMIT [6]

Table 5.1 Phonemic symbols

Consonants	P, T, K, B, D, G, CH, JH, F, TH, S, SH, HH, V, DH, , ZH, M, N, NG, L, R, W, Y
Vowels	IY, IH, EH, EY, AE, AA, AW, AY, AH, AO OY, OW, UH, UW, ER, AXR, AX

For the corpus contained with labels, there are a total of 800 utterances from learners and a total of 514 utterances from native speakers. And the speech symbol labeling is done using Praat and is provided as a TextGrid file. Phoneme symbols which are a subset of ARPABET, are displayed as a hint of pronunciation in the read-out text at the time of sound recording. The transcriber, who is an expert who has extensive experience in pronunciation guidance targeting Japanese people with familiarity with transcription with speech symbols, performed transcription of perceived phonetic symbols by IPA while referring to these phonemic symbols which is a subset of ARPABET, which is also listed in 5.1(In our ASR system, "AXR" is not used so "AXR" will be turned in to "ER") and sounds. For the corpus without label, there are 23944 utterances from learners which only contain wav files.

For transcription, shown in Fig 5.1, the layers "word" is the spells of words in this utterance. The layer "phone" is the phoneme sequence in phoneMic symbols from the standard English dictionary. The layer "actual" is what the transcriber perceived exactly in phonetic symbols. Because "actual" is in phonetic symbols which make it difficult to apply it to the ASR system, we have to think of how to turn it into phoneMic symbols or find someone to make transcriptions. It will be explained later.

5.2.2 Pronunciation diversity modeling

Unlike learning data (pronunciation dictionary) used in general G2P tasks, the size of paired corpus used in this study is small. So we will make a primary G2P, then use this G2P to generate the label for the 23944 unlabeled corpus in the future. We will denote the grapheme as G (Grapheme), the phoneme which is a subset of ARPABET, as M (phoneMic), and the phonetic symbols with IPA as T (phoneTic).

Rule from teachers' experiences

There are three kinds of pronunciation diversity model. First of all, let me introduce the hand-crafted rule. In this thesis, rules for phoneme deletion and substitution are based on [28],

and rules for phoneme insertion, which is also called vowel epenthesis are based on [31]. The following is the detailed rules for phoneme deletion and substitution. (" _ " means deletion)

1. /AO/ → /OW/
2. if /AX/ followed by /UH/: /AX/ → /OW/, else: /AX/ → /AA/
3. /AE/ → /AA/
4. /AH/ → /AA/
5. /ER/ → /AA/
6. /IH/ → _
7. /UH/ → _
8. /R/ → /L/
9. /L/ → /R/
10. if /HH/ followed by /UW/: /HH/ → /F/, else if /HH/ followed by /IY/: /HH/ → /SH/
11. if /F/ followed by /AO/, /F/ → /HH/
12. /TH/ → /S/ / /SH/
13. /DH/ → /Z/ / /ZH/
14. /V/ → /B/
15. /N/ → /M/ / /NG/ / _
16. if /T/ followed by /IH/ / /IY/: /T/ → /CH/, else if /T/ followed by /UH/ / /UW/: /T/ → /T S/
17. if /D/ followed by /IH/: /D/ → /DH/, else if /D/ followed by /IY/: /D/ →/CH/, else if /D/ followed by /UH/ / /UW/, /D/ → /DH/ / /Z/
18. if /S/ followed by /IH/ / /IY/: /S/ → /SH/
19. if /Z/ followed by /IH/ / /IY/: /Z/ → /ZH/

Because it is not rare for Japanese to inset a vowel after a consonant, if this consonant is not followed by a vowel. And rules of vowel epenthesis are defined by following rule:

1. /JH/ → /JH IH/, /CH/ → /CH IH/
2. /D/ → /D OW/, /T/ → /T OW/
3. /NG/ → /N G UW/
4. other consonants: insert /UW/

By the hand-crafted rules mentioned before, we can make phone variability by processing those rules based on the pronunciation dictionary.

AE-T2M, JE-M2M and JE-G2M640

Another method of pronunciation diversity modeling is to make use of the 800 labeled Japanese English corpus to generate diversity by machine learning method. However, because we want to use these diversity model to detect pronunciation error in the final step, which means this model has to generate the phoneMic(M) sequence but the transcription for how the sounds are sensed by transcriber is in phoneTic(T), and it is hard for people who are not experts to dear with IPA(International Phonetic Alphabet) symbol and it is necessary for our ASR system to use phoneme sequence, We can't use the label in paired corpus directly. First, for paired data of native speakers, we make a model by phonetisaurus, to learn the relationship between IPA labels and intended phonemes, which is also called as (AE-T2M). By applying this AE-T2M model to Japanese English phonetic transcriptions, we will get the phoneme sequences that are supposed to be perceived by native American listeners. By using the obtained sequence of phonemes, that are supposed to be perceived by native listeners, we can train two models. One is converting a sequence of phonemes intended by a Japanese learner to that of phonemes perceived by a native listener. The other is converting a sequence of graphemes that a Japanese learners is looking at to read to that of phonemes perceived by a native listener.

Because we want to build a pronunciation error detection system at last, which means we need to remain some transcription to use as test data, we divide the 800 labeled Japanese English corpus into 640 and 160. We will use 640 labeled Japanese English corpus to train the model and use the last 160 ones to evaluate. And JE-G2M model trained by 640 labeled data is called as JE-G2M640.

Pronunciation error detection and JE-G2M24k

Because we have got two diversity models, hand-crafted rules and JE-G2M640, we can do pronunciation error detection based on these two methods. By combination, we generated three kinds of G2M24k models, G2M640-based model, rule-based model and "G2M640+rule"-based

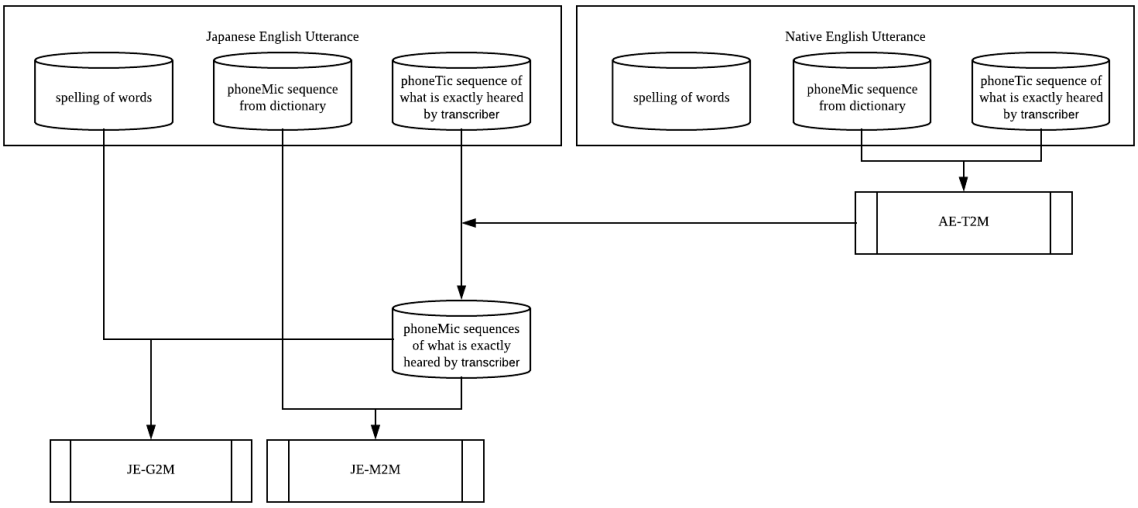


Fig. 5.2 Relationship between JE-G2M, JE-M2M, AE-T2M

model. Firstly, we have a Japanese English utterance, then we will do Automatic Speech Recognition (ASR) to identify words in the utterances which were intended by learners, where the ASR system here uses Japanese English acoustic model. Then we get the time stamps and spelling of words, by applying rules on phonemic sequence from dictionary or inputting spelling to the JE-G2M640, we can get a lot of candidate phonemic sequences of how Japanese would read this word and build grammar based on these candidates. By the word unit, we will do phoneme recognition based on grammar built before and native English acoustic model. Finally, we will get the phoneme sequences of each word, by comparing the obtained phoneme sequence with the canonical phoneme sequence from a dictionary, pronunciation error detection will be done.

As I have mentioned before, there are 23944 unlabelled Japanese English utterances. It is possible that we use the JE-G2M640 or hand-crafted rules to try pronunciation error detection on this unlabelled 23944 utterance first. Then we can use the result of pronunciation error detection to update the JE-G2M model, as shown in Fig. 5.3. Because these are the models learned from near 24k data, we call these as JE-G2M24K. And because the result of Japanese English ASR may not be so reliable, we will select a threshold here and generate a brunch of JE-G2M24k models, trying to find the best one. Of course, we can use these first generations JE-G2M24k to replace the position of JE-G2M640 in Fig. 5.3 and update JE-G2M24k iteratively.

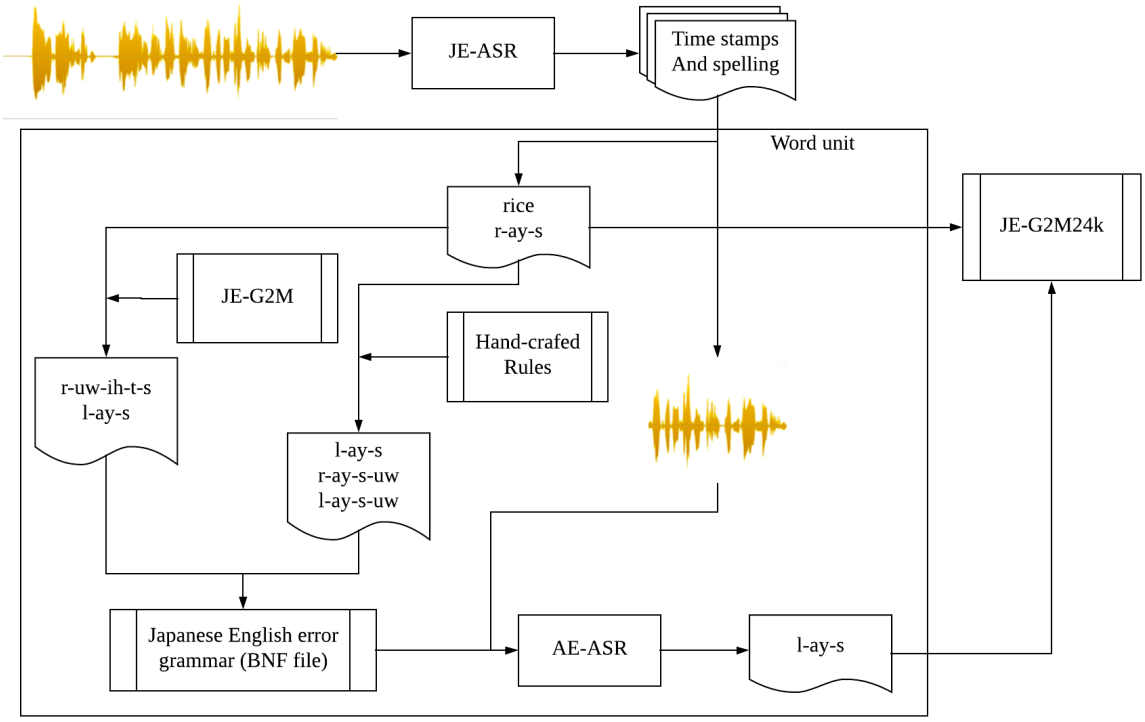


Fig. 5.3 JE-G2M24k

5.3 Additional phonemic transcription

5.3.1 Details on process of transcription

After the first period of experiment on pronunciation error detection, because we use the output of AE-T2M to evaluate our first version of pronunciation error detection, the output of AE-T2M maybe not 100% correct. So we have to find some supplement phonemic transcription for ERJ(English Read by Japanese) corpus. Two native speakers of American or Canadian English, K and D, finished the phonemic transcription of 800 utterances from learners which have already contained phonetic transcription. K and D are Native speakers of American English with fundamental knowledge of phonetics of English, fundamental knowledge of how to use Praat(the tool we use to transcribe), and both have little knowledge of Japanese phonetics. What we want to know is what phonemes are (will be) perceived by American listeners who do not know Japanese at all, when they listen to the utterances. So some Japanese sounds may be found in the utterances but transcription has to be done only with American English phonemes. The phonemic symbols used for the transcription are a slightly modified version of ARPABET, as listed in 5.1.

Table 5.2 Phone error rate compared with the canonical phoneme sequences in the dictionary

Transcriber	Accuracy	H	D	S	I	N
K	68.16%	17357	939	6565	411	24861
D	88.28%	22046	362	2449	122	24857

And after obtaining the phoneMic transcription, we can use it instead of phoneMic sequences generated by AE-T2M to build JE-G2M. So we will use it to build JE-G2M640 first and generate a new version of JE-G2M24k. We can build three kinds of JE-G2M640, one generated from K's transcription, one generated from D's transcription and the other one generated from combination of both transcription.

5.3.2 Comparison between different transcriptions

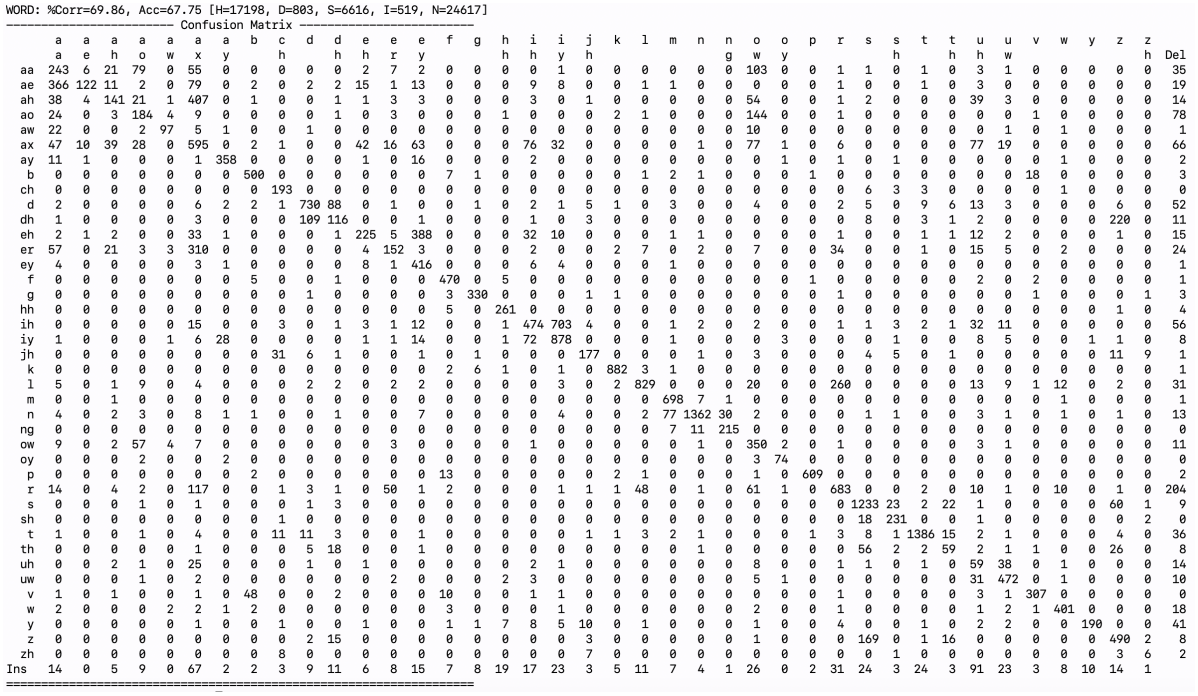
Some experiment results will be shown afterward and in this section, I will show the comparison between transcriptions from K and D. Although the K and D are both native English speaker, how they sense the Japanese English sound may be different.

In table 5.2, S means number of substitutions, D means the number of deletions, I means the number of insertions, H is the number of correct phonemes, N is the number of phonemes in the reference (N=S+D+C).

Totally, transcription from D has less difference from the standard English, which means that D is less strict than K. Although K's transcription has a lower correct rate and accuracy, we cannot say K's transcription is better but can only say K is more strict. When we focus on the error type(substitution, deletion, insertion), there is no difference between these two transcribers to have a tendency to choose a specific error.

Confusion Matrix are shown in 5.4. Now we compare K's and D's transcription directly, the Accuracy is 67.75 % and Correct rate is 69.86 %, which leads to the same result that their transcription is quite different. And when we focus on the error type(substitution, deletion, insertion), when they listened to the same Japanese English sounds, they may have an agreement that there is a substitution mistake but disagree on what phoneme is actually prounounced.

For certain symbols, there is not much disagreement on consonants. Most of their disagreement are in vowels, especially on "ae" and "aa", we can point out such obvious disagreements. For "ae", their transcription has a disagreement on whether it should be "ae" or "aa". For "er", their transcription have a disagreement on whether it should be "er" or "ax". Both two transcribers have the fundamental knowledge of phonetics of English, so these disagreements



```
WORD: %Corr=69.86, Acc=67.75 [H=17198, D=803, S=6616, I=519, N=24617]
------------------------ Confusion Matrix -------------------------
      a   a   a   a   a   a   a   b   c   d   d   e   e   e   f   g   h   i   i   j   k   l   m   n   n   o   o   p   r   s   s   t   t   u   u   v   w   y   z   z
      a   e   h   o   w   x   y       h       h   h   r   y               h   h   y   h                   g   w   y                   h       h   h   w                   h  Del
 aa 243   6  21  79   0  55   0   0   0   0   0   2   7   2   0   0   0   0   1   0   0   0   0   0   0 103   0   0   1   1   0   1   0   3   1   0   0   0   0   0  35
 ae 366 122  11   2   0  79   0   2   0   2   2  15   1  13   0   0   0   9   8   0   0   1   1   0   0   0   0   0   0   1   0   1   0   3   0   0   0   0   0   0  19
 ah  38   4 141  21   1 407   0   1   0   0   1   1   3   3   0   0   0   3   0   1   0   0   0   0   0  54   0   0   1   2   0   0   0  39   3   0   0   0   0   0  14
 ao  24   0   3 184   4   9   0   0   0   0   1   0   3   0   0   0   1   0   0   0   2   1   0   0   0 144   0   0   1   0   0   0   0   0   0   1   0   0   0   0  78
 aw  22   0   0   2  97   5   1   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  10   0   0   0   0   0   0   0   0   1   0   1   0   0   0   1
 ax  47  10  39  28   0 595   0   2   1   0   0  42  16  63   0   0   0  76  32   0   0   0   0   1   0  77   1   0   6   0   0   0   0  77  19   0   0   0   0   0  66
 ay  11   1   0   0   0   1 358   0   0   0   0   1   0  16   0   0   0   2   0   0   0   0   0   0   0   0   1   0   1   0   1   0   0   0   0   0   1   0   0   0   2
  b   0   0   0   0   0   0   0 500   0   0   0   0   0   0   7   1   0   0   0   0   0   1   2   1   0   0   0   1   0   0   0   0   0   0   0  18   0   0   0   0   3
 ch   0   0   0   0   0   0   0   0 193   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   6   3   3   0   0   0   0   1   0   0   0   0
  d   2   0   0   0   0   6   2   2   1 730  88   0   1   0   0   1   0   2   1   5   1   0   3   0   0   4   0   0   2   5   0   9   6  13   3   0   0   0   6   0  52
 dh   1   0   0   0   0   3   0   0   0 109 116   0   0   1   0   0   0   1   0   3   0   0   0   0   0   0   0   0   0   8   0   3   1   2   0   0   0   0 220   0  11
 eh   2   1   2   0   0  33   1   0   0   0   1 225   5 388   0   0   0  32  10   0   0   0   1   1   0   0   0   0   1   0   0   1   1  12   2   0   0   0   1   0  15
 er  57   0  21   3   3 310   0   0   0   0   0   4 152   3   0   0   0   2   0   0   2   7   0   2   0   7   0   0  34   0   0   1   0  15   5   0   2   0   0   0  24
 ey   4   0   0   0   0   3   1   0   0   0   0   8   1 416   0   0   0   6   4   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
  f   0   0   0   0   0   0   0   5   0   0   1   0   0   0 470   0   5   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   2   0   2   0   0   0   0   1
  g   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   3 330   0   0   0   1   1   0   0   0   0   0   0   0   1   0   0   0   0   0   0   1   0   0   0   1   3
 hh   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   5   0 261   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1   0   4
 ih   0   0   0   0   0  15   0   0   3   0   1   3   1  12   0   0   1 474 703   4   0   0   1   2   0   2   0   0   1   1   3   2   1  32  11   0   0   0   0   0  56
 iy   1   0   0   0   1   6  28   0   0   0   0   1   1  14   0   0   1  72 878   0   0   0   1   0   0   0   3   0   0   0   1   0   0   8   5   0   0   1   1   0   8
 jh   0   0   0   0   0   0   0   0  31   6   1   0   0   1   0   1   0   0   0 177   0   0   0   1   0   3   0   0   0   4   5   0   1   0   0   0   0   0  11   9   1
  k   0   0   0   0   0   0   0   0   0   0   0   0   0   0   2   6   1   0   1   0 882   3   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   1
  l   5   0   1   9   0   4   0   0   0   2   2   0   2   2   0   0   0   0   3   0   2 829   0   0   0  20   0   0 260   0   0   0   0  13   9   1  12   0   2   0  31
  m   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 698   7   1   0   0   0   0   0   0   0   0   0   0   1   0   0   0   1
  n   4   0   2   3   0   8   1   1   0   0   1   0   0   7   0   0   0   0   4   0   0   2  77 1362  30   2   0   0   0   1   1   0   0   3   1   0   1   0   1   0  13
 ng   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   7  11 215   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
 ow   9   0   2  57   4   7   0   0   0   0   0   0   0   3   0   0   0   1   0   0   0   0   0   1   0 350   2   0   1   0   0   0   0   3   1   0   0   0   0   0  11
 oy   0   0   0   2   0   0   2   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   3  74   0   0   0   0   0   0   0   0   0   0   0   0   0
  p   0   0   0   0   0   0   0   2   0   0   0   0   0   0  13   0   0   0   0   0   0   2   1   0   0   0   1   0 609   0   0   0   0   0   0   0   0   0   0   0   2
  r  14   0   4   2   0 117   0   0   1   3   1   0  50   1   2   0   0   0   1   1   1  48   0   1   0  61   1   0 683   0   0   2   0  10   1   0  10   0   1   0 204
  s   0   0   0   1   0   1   0   0   0   1   3   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0 1233  23   2  22   1   0   0   0   0  60   1   9
 sh   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0  18 231   0   0   1   0   0   0   0   0   2   0
  t   1   0   0   1   0   4   0   0  11  11   3   0   0   1   0   0   0   0   0   0   1   1   3   2   1   0   0   0   1   3   8   1 1386  15   2   1   0   0   0   4   0  36
 th   0   0   0   0   0   1   0   0   0   5  18   0   0   1   0   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0  56   2   2  59   2   1   1   0   0  26   0   8
 uh   0   0   2   1   0  25   0   0   0   1   0   1   0   0   0   0   0   2   1   0   0   0   0   0   0   8   0   0   1   1   0   1   0  59  38   0   1   0   0   0  14
 uw   0   0   0   1   0   2   0   0   0   0   0   0   2   0   0   0   2   3   0   0   0   0   0   0   0   5   1   0   0   0   0   0   0  31 472   0   1   0   0   0  10
  v   1   0   1   0   0   1   0  48   0   0   2   0   0   0  10   0   0   1   1   0   0   0   0   0   0   0   0   0   1   0   0   0   0   3   1 307   0   0   0   0   0
  w   2   0   0   0   2   2   1   2   0   0   0   0   0   0   3   0   0   0   1   0   0   0   0   0   0   2   0   0   1   0   0   0   0   1   2   1 401   0   0   0  18
  y   0   0   0   0   0   1   0   0   1   0   0   1   0   0   1   1   7   8   5  10   0   1   0   0   0   1   0   0   4   0   0   1   0   2   2   0   0 190   0   0  41
  z   0   0   0   0   0   0   0   0   0   0   2  15   0   0   0   0   0   0   0   0   3   0   0   0   0   1   0   0   0 169   0   1  16   0   0   0   0   0 490   2   8
 zh   0   0   0   0   0   0   0   0   0   8   0   0   0   0   0   0   0   0   0   0   7   0   0   0   0   0   0   0   0   0   1   0   0   0   0   0   0   0   3   6   2
Ins  14   0   5   9   0  67   2   2   3   9  11   6   8  15   7   8  19  17  23   3   5  11   7   4   1  26   0   2  31  24   3  24   3  91  23   3   8  10  14   1
=============================================================
```

Fig. 5.4 Confusion Matrix between K's and D's transcription

may come from how they understand a certain voice in phonetics and how they hear and sense it in their brain.

Chapter 6

Experiments on phoneme error detection with the pronunciation diversity models

6.1 Experiment on diversity modeling

6.1.1 AE-T2M

At first, we built the AE-T2M from the corpus of native speakers, which show the relationship between the phonetic symbols of IPA with phoneme which is a subset of ARPABET, which also can be understood as relationship between what the learners pretend to pronounce and what the native speakers exactly perceive. Because IPA is hard for non-phoneticians or non-experts to understand, let me first introduce some feature of IPA. In IPA, each phonetic symbol may be given a modifier called diacritics. Although diacritics is for expressing subtle differences in tones, the transcriber[14, 13] of the labeled corpus told us that diacritics is very important mark for expressing the sound. In our research, we firstly investigate the frequency of diacritics for different tones. For each tone, "tone + diacritics" and "tone only" are distinguished, and these are sorted in order of frequency. When cumulative frequency reaches N%, "tone + diacritics" are independently treated as a single mark. the diacritics for the rest "tone + diacritics" are excluded. Then, we did experiment on T2M to find the best N.[24]

For phonemic symbols in English, a diphthong, where although two vowels are uttered consecutively, the native speaker perceives it as one speech unit, is defined as one phoneme, but when it is described by a phonetic symbol, it becomes two vowel symbols. When we build AE-T2M, it is necessary for us to convert two consecutive vowel symbols into one phoneme symbol. If we have enough corpus, this conversion can be done automatically by model. In our case, because of the fact that we don't have enough labeled corpus, we need to treat a diphthong

Table 6.1 Accuracy of AE-T2M for different N

aligend	diacritics	diphthongs	symbols	accuracy
yes	0%	60%	88	88.2%
no	100%	100%	218	88.8%

as a special symbol. Like what we have done for diacritics, we investigate the frequency of different diphthong, and when cumulative frequency reaches N%, we treat it as a new symbol.

After preparing for the label, we can build AE-T2M by phonetisaurus. Phonetisaurus is implemented with a model called joint-sequence N-gram. In this model, the sequence of G and M are aligned firstly. Aligned sequence of G and M, which is also called a joint sequence, is used to train a language model of N-gram. In the labeled corpus, because T and M have been aligned by the transcriber, so we have two choices, use this alignment information to build a AE-T2M and train the model directly without using it. Regarding these two methods, we do experiment on checking the conversion accuracy while adjusting the value of N for the diacritics symbol and the diphthong symbol. Besides, 50 utterances selected randomly out of 514 sentences are evaluated data and the remaining 464 sentences are taken as learning data.

Experimental results showed that the value of N did not significantly change the conversion accuracy, but the condition showing the highest accuracy for the two methods is shown in 6.1. Based on these results, in the subsequent discussion, we decided to use the AE-T2M model built by using all the phonetic symbols with diacritics and the symbolization of diphthongs, without using manual alignment information. The number of types of phonetic symbols in this condition is 218.

6.1.2 JE-M2M

Because we have built AE-T2M before, we can generate multiple phoneme strings and calculate their confidence for a single IPA string. To make use of the confidence, we did following operation. For one word G_i, we had phonetic sequences T_i. We generate possible phoneme sequences $L_i^1, L_i^2, .., L_i^j, ..$ and the confidence p^j, by feeding T_i to AE-T2M. Then we round off $10p^j$ ($R = \text{round}(10p^j)$). Treating L_i^j as if L_i^j is observed R times in the corpus to built this paired corpus, which means sequence with probability lower than 0.1 will be thrown away.

Then we can build JE-M2M by using the generated sequences. If we feed a word to JE-G2M, we can get multiple phoneme strings and their confidence probabilities. If M kinds of

Table 6.2 Pronunciation perplexity by number of syllable

number of syllable	1	2	3	4	all
JE-M2M	4.9	5.6	6.2	6.8	5.3
Rule	3.5	9.4	49.8	103.6	6.7

Table 6.3 Output of JE-M2M with Input of LACK(/L AE K/)

	/L AA K/	/L AH K/	/L AE K/
probability	0.32	0.31	0.27

Table 6.4 Output of JE-M2M with Input of LUCK(/L AH K/)

	/L AH K/	/L AA K/	/R AA K/
probability	0.42	0.31	0.09

phoneme sequences are obtained and the probability value is written as $\{q_m\}$, the entropy can be defined as the following equation.

$$H = \sum_{m=1}^{M} -q_m \log_2(q_m) \tag{6.1}$$

Pronunciation perplexity is calculated by 2^H. This is a metric for entropy about how many different sequences one input will generate.

We compare pronunciation perplexity by JE-M2M and pronunciation diversity by [28], using the 850 Basic English Words in Table 6.2. Because when pronunciation transformation is considered in units of phonemes, the diversity increases as the word length becomes longer, we divided these words by their syllable and compared it. When the number of syllables is small, there are no difference, but when the number of the syllables grows, the diversity of [28] increase sharply and pronunciation perplexity of JE-M2M shows near a certain value. Although the influence of corpus size on the result of this experiment cannot be denied, it is considered that the effect of JE-M2M that models series correspondence between phoneme string and tone sequence is also shown.

In fact, several studies were conducted to find out what kind of phoneme sequence is available as the output of JE-M2M. In Table 6.3 and Table 6.4, the output of JE-M2M when the input is [LACK(/L AE K/)],[LUCK(/L AH K/)] are shown. The tendency of erroneous tendencies for learners on the rule:/AE/ to /AH/ was shown by this two words[28].

Table 6.5 Average pronunciation perplexity for different G2P models

JE-G2M	AE-G2M$_1$	AE-G2M$_2$	CMU-G2M
5.1	2.3	2.4	1.4

In our experiment, we use Basic English[30], which contain 850 basic english word. Average pronunciation perplexity when all 850 words are fed is indicated in the table.

Table 6.6 Ouput of JE-M2M with a input of THEREFORE(/DH EH R F AO R/)

	/D AA F OW/	/Z EH F OW/	/D AY F OW/
p	0.23	0.21	0.20

Table 6.7 Output of JE-G2M with input of THEREFORE(/DH EH R F AO R/)

	/S AH HH AO AX/	/S AH HH OW AX/
p	0.35	0.19

6.1.3 AE/JE-G2M

Besides, we can also build AE-G2M by using using labelled corpus of native speakers. In this case, there are two kinds of phoneme sequence, 1) Use the phoneme transcription directly (AE-G2M$_1$), 2) Use the phoneme sequences generated from phonetic sequences (AE-G2M$_2$). For 2), even for English of native speakers, I expect that diversity can be observed by transcribing at the symbol level. Besides, if we want to build G2P for standard English, there is no need to train the G2P model on a special small corpus. So we can train a CMU-G2M model by CMU dictionary[9]. Then we can do experiment on comparing the perplexity of different model.

Perplexity of AE-G2M is about 1.0 larger than CMU-G2M. We understand that in modeling low transparency transformation such as G2M, the size of corpus brings a great influence on accuracy. Besides, the pronunciation perplexity of JE-G2M is about twice as large as that of AE-G2M. Although the accuracy is low due to the influence of corpus size, if we consider that both are influenced, this double value may be somewhat reliable. This means that when a learner is given a grapheme sequence, it has about twice the pronunciation diversity compared with the native speaker.

We also compared the output of JE-M2M and JE-G2M, with the input of phoneme sequence and spelling sequence from the same word. Shown in Table 6.6 and Table 6.7, the error rule: /F/ and /HH/ often comes out before /AO/ can be found not just in M2M but also G2M. In G2M, we can also found a conversion: /DH/ to /S/, which is considered as lack of data.

6.2 Experiment on error detection

6.2.1 Levenshtein distance

Because of the fact that instead of speech recognition, our task is to automatically detect Japanese English pronunciation error, Phoneme Accuracy and Error Discrimination Accuracy are used as metric in this thesis. Both Phoneme Accuracy and Error Discrimination Accuracy are metrics which is similar as Word error rate (WER) and are derived from the Levenshtein distance at phoneme level.

The Levenshtein distance is a kind of edit distance. And this distance means if we want to transfer from one string to another one, the minimum number we need to do edit operation. Edit operation here allows replacing one character with another, inserting one character, and deleting one character. Dynamic programming is often used as one of the solutions to this problem.

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j)+1 \\ \mathrm{lev}_{a,b}(i,j-1)+1 \\ \mathrm{lev}_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases} \tag{6.2}$$

Where, Levenshtein distance between two strings a and b is $lev_{a,b}(|a|,|b|)$, and $1_{(a_i \neq b_j)}$ is the function that when $a_i = b_j$ it is 0 otherwise it is 1. In the equation before, we can also know which edit it will be. Then we can use it to implement our evaluation metric.

$$\begin{cases} lev_{a,b}(i-1,j)+1 & \text{deletion} \\ lev_{a,b}(i,j-1)+1 & \text{insertion} \\ lev_{a,b}(i-1,j-1)+1_{(a_i \neq b_j)} & \text{substitution} \end{cases} \tag{6.3}$$

6.2.2 Experiment without additional transcription

Evaluate method

For the first trial of Japanese English error detection, additional transcription is not utilized. Because what the transcriber exactly may feel are in phoneTic symbols, but the ASR system used in this thesis only use phoneMic symbols. That is why we trained AE-T2M before and

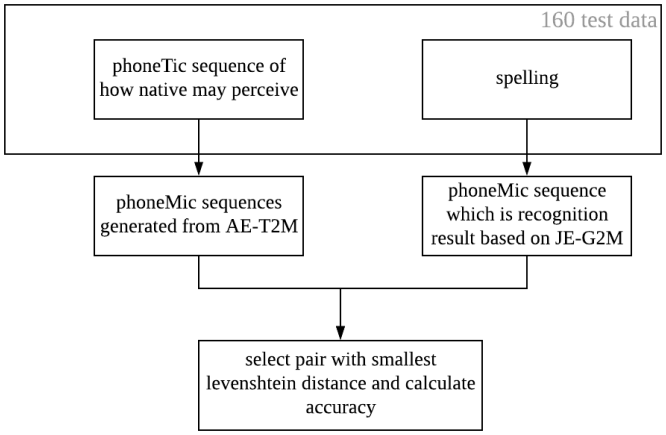


Fig. 6.1 How to calculate phoneme recognition accuracy(additional data not used)

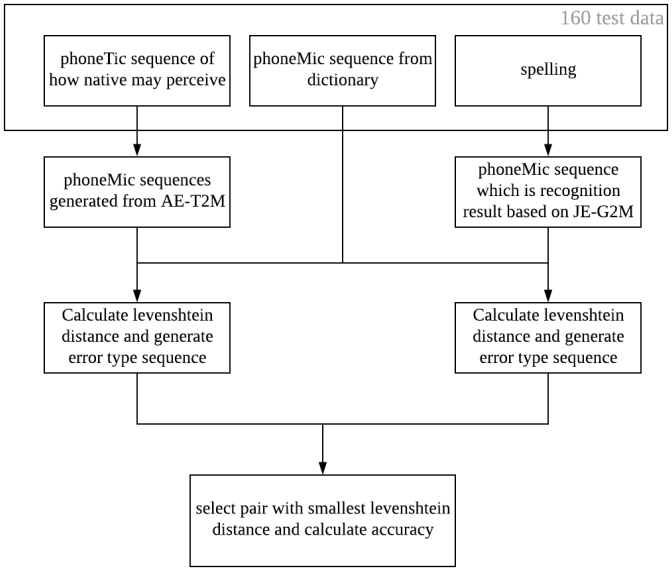


Fig. 6.2 How to calculate error type identification accuracy(additional data not used)

try to find a transformer from T(phoneTic) to M (phoneMic). When training diversity model, there is no problem to use the results of AE-T2M directly. However when evaluating, the multi-results will be generated from one label in phoneTic symbols, and we have to compare the difference between the "mutli results" and the error detection result. So what we do is putting IPA transcription, which is the testing data (160 JE utterances), into AE-T2M and only take phoneme string with a probability over 20% as a correct label(even it is a string of symbols appropriately phonemically written as a Japanese pronunciation of the Japanese, probably not a correct pronunciation as an American word.).

As a measure of evaluation, phoneme recognition accuracy and error type identification accuracy were calculated. Upon calculation of the phoneme recognition accuracy, shown in Fig. 6.1, Levenshtein distance is calculated between multiple label phoneme strings and a phoneme string of recognition result, a label phoneme string with the shortest distance is adopted as the final label phoneme string and accuracy were calculated based on it. Upon calculation of the error type identification accuracy, shown in Fig. 6.2, We first compare multiple label phoneme strings and dictionary string to generate "error type sequence". We can also generate "error type sequence" by comparing dictionary string and recognition result. Both accuracies are calculated based on the following equation.

$$accuracy = 1 - \frac{D+I+S}{N} \tag{6.4}$$

where C is the number of correct symbols, S is the number of substitutions, D is the number of deletions and I is the number of insertions.

Experiment result

Experimental results are shown in Table 6.8 and Table 6.9. When constructing JE-G2M24K, it is necessary to set a threshold value of reliability for speech recognition result of Japanese English, but as a result of the preliminary experiment, the accuracy doesn't change obviously if different threshold value are set. So the result of the model with the highest accuracy is shown. Besides, in both two tables, JE-G2M24KG1 means the first generation of JE-G2M24k, JE-G2M24KG2 means the second generation of JE-G2M24k and so on. This 640 utterance and 160 utterances of evaluation data partially overlap on speaker or text (it is completely open as on utterance), which may lead to a difference on the result. [35]

Shown in both error type identification accuracy and phoneme recognition correct accuracy, the rule-based method is less accurate than the corpus-based model. The reason is that the rule-based method excessively generates the number of error candidates, which leads to more recognition failures. Application of rules is inherently context-dependent, but there is a limit to

Table 6.8 comparison on phoneme recognition accuracy (PRA)(%)

JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3	Rule	Loop
58.19	57.66	57.92	57.88	51.64	-15.38

Table 6.9 comparison on error type identification accuracy(EIA)(%)

JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3	Rule	Loop
65.46	63.64	62.58	62.84	60.64	15.98

relying on the teacher 's experience on context dependency, and we should make effective use of analysis results based on corpus. In addition, continuous phoneme recognition performed without constraints(Loop) has a very low accuracy which is impossible to use.

Now let's move to see the result between JE-G2M640 and JE-G2M24K. No matter which generation it belongs to, JE-G2M24K has a little lower accuracy than JE-G2M640. JE-G2M24K is constructed by using the automatic recognition result obtained by applying the pronunciation diversity model according to the rule and JE-G2M640 from a large-scale corpus. So two reasons may lead to this deterioration. 1) the result of automatic recognition induces more incorrect error pattern learned by JE-G2M24k. 2) the test data itself contains some incorrect error pattern, which is because our test data is not correct for 100%.

As an example, Table 6.10 shows the results of an example (word "intelligible") in which many errors were observed. transcription in IPA and its result from AE-T2M, furthermore, phoneme strings prescribed by American pronunciation dictionary and recognition results obtained considering various kinds of pronunciation diversity are shown. The recognition result and phoneMic sequence from dictionary are the same. But they differ a lot from phoneTic sequence(IPA) and phoneMic sequence from AE-T2M. In the recognition result, some vowel phonemes are vowels that cannot be covered by pronunciation diversity and [h] at the beginning of the word is labeled with a state of breath leakage. It is a fact that transcription in IPA tends to be slightly too detailed, and it is also necessary to review this phonemicization method again. At the same time, if eventually, errors at the phoneme level are to be detected, it is considered that corpus maintenance should be considered regarding error transcription at the phoneme level. If we make transcription from Japanese English and American English by phoneTic symbols, some phoneTic symbols which occur in Japanese English will never occur in American English. So if we use AE-T2M to generate Japanese English phoneMic sequence, maybe it has out-of-domain problem. So using test data which use result form AE-T2M maybe

Table 6.10 example for "intelligible"

word	intelligible
phoneTic sequence(IPA)	h i n t e ɾ ɪ ʤ ɪ ɯ β ə l i̞
phoneMic sequence from AE-T2M	hh iy n t ey t ax d z ih uw b ih l iy
Dictionary	ih n t eh l ax jh ax b ax l
Recognition result	ih n t eh l ax jh ax b ax l

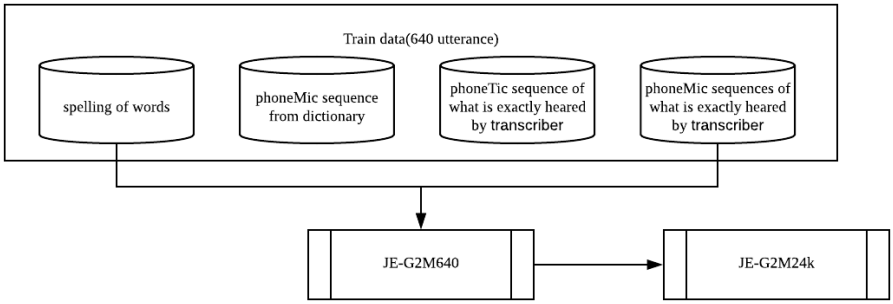


Fig. 6.3 JE-G2M with additional data

not a good idea. That is why we let two native speakers make transcription at phoneme level and the next section shows the results we get with additional transcription.

6.2.3 Experiment with additional transcription

Evaluate method

The second trial of Japanese English error detection uses additional transcription. In this trial, we do not use AE-T2M but we built JE-G2M models using additional transcriptions of JE data, which are manual phoneMic transcriptions and can train JE-G2M640 and update JE-G2M24k directly from additional transcription, which also shown in Fig. 6.3. For the measure of evaluation, we calculate phoneme recognition accuracy and error type identification accuracy either. The calculation is similar, but because we have phoneMic sequence in our test data now, so we can directly calculate Levenshtein distance directly without using the output from AE-T2M, which are also shown in Fig. 6.4 and Fig. 6.5.

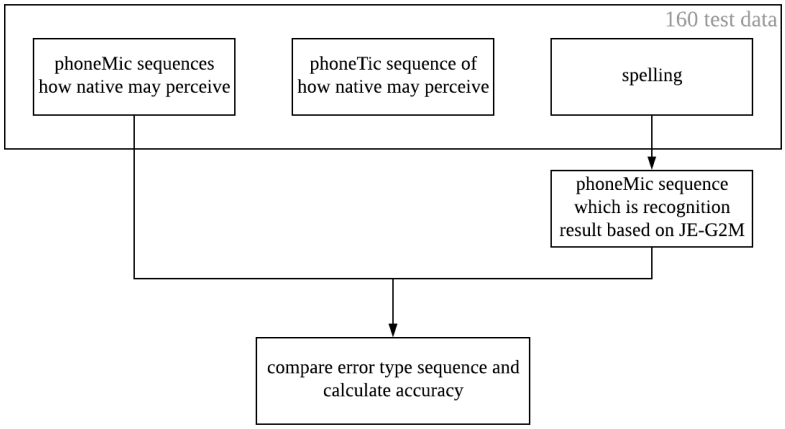


Fig. 6.4 How to calculate phoneme recognition accuracy(with additional data)

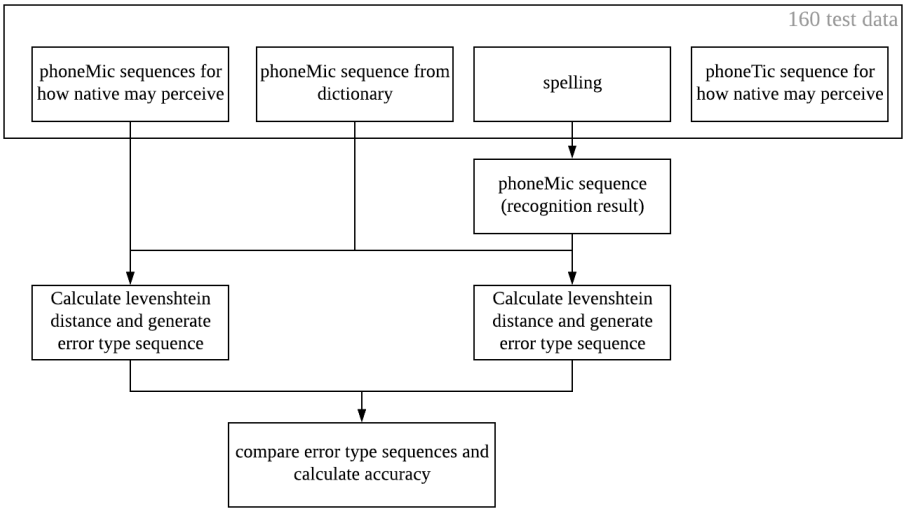


Fig. 6.5 How to calculate error type identification accuracy(with additional data)

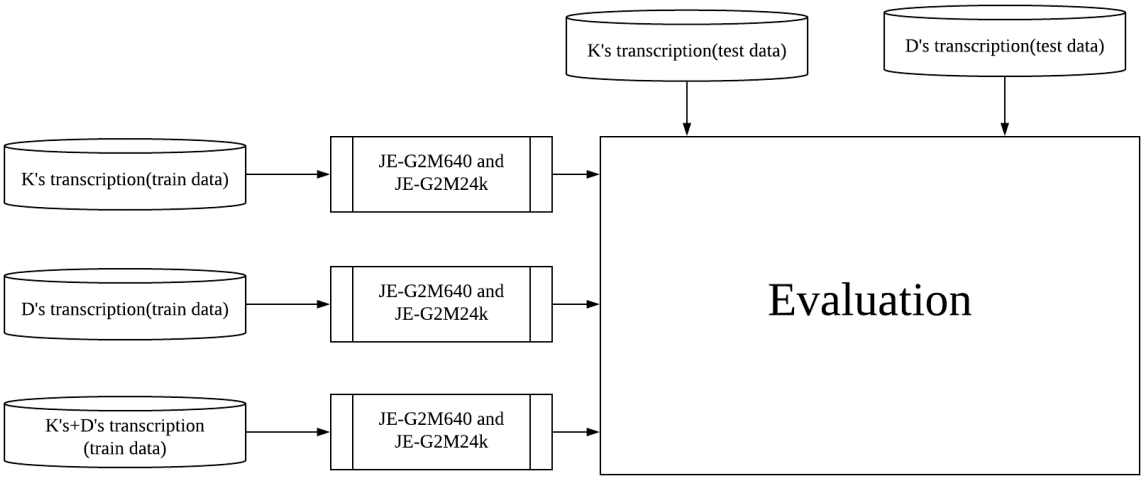


Fig. 6.6 Overview of evalutaion

Experiment result

Because we have additional transcription from two transcribers and their transcription are quite different, so we can do experiment in three methods, 1) only use D's transcription to train JE-G2M640, 2) only use K's transcription to train JE-G2M640, 3) combine both K's and D's transcription to train JE-G2M640. When evaluating, because of the existence of two kinds of transcription, we can compare their error detection result with two kinds of test data(transcriptions). The novelty of this experiment is not to find a model with better accuracy, but to use the existing model to compare different kinds of transcription to find out the difference of how different transcribers sense phonemic error and try our best to simulate it. which is also shown in Fig.6.6. And as same as the first trial of error detection, although we set a threshold on the result of speech recognition and generate series of G2M model, their accuracy doesn't change a lot, so the result shown in the following table are the highest accuracy from different models.

Let's firstly focus on the results between different models. No matter which transcription used, for phoneme recognition accuracy (PRA), JE-G2M24k have a slightly better result than JE-G2M640 and model of later generation will perform better. However, when metric is error type identification accuracy(EIA), although JE-G2M24k of a later generation may have higher accuracy than G1, EIA of all the JE-G2M24k don't have better results than JE-G2M640. And variations generated by Rule have lower accuracy in both PRA and EIA, maybe naive generation by hand-crafted rules generate too many variations and reduce accuracy.

Then let's compare the difference between different test data. Because we have transcriptions from 2 people, we also have two kinds of test data. No matter which transcription used

Table 6.11 Phoneme recognition accuracy(PRA) and error type identification accuracy(EIA) of model learned from D's transcriptions(%)

Metric	Test data	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3	Rule	Loop
PRA	D	82.80	84.18	84.56	84.38	67.16	-10.16
	K	64.57	64.59	65.32	65.26	56.95	-9.90
EIA	D	84.83	85.27	85.56	85.38	72.72	-5.36
	K	70.37	69.65	69.72	69.72	66.60	11.38

Table 6.12 Phoneme recognition accuracy(PRA) and error type identification accuracy(EIA) of model learned from K's transcriptions(%)

Metric	Test data	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3	Rule	Loop
PRA	D	68.92	74.06	76.22	76.42	67.16	-10.16
	K	68.06	68.72	68.58	68.00	56.95	-9.90
EIA	D	73.91	77.50	78.72	78.93	72.72	-5.36
	K	74.84	74.57	73.67	73.53	66.60	11.38

Table 6.13 Phoneme recognition accuracy(PRA) and error type identification accuracy(EIA) of model learned from K's and D's transcriptions(%)

Metric	Test data	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3	Rule	Loop
PRA	D	74.54	78.00	79.18	79.26	67.16	-10.16
	K	68.86	68.10	67.73	67.17	56.95	-9.90
EIA	D	78.13	80.79	81.66	81.74	72.72	-5.36
	K	74.68	73.41	72.59	72.59	66.60	11.38

to learn JE-G2M640, JE-G2M640 and JE-G2M24k will have a higher accuracy in Phoneme recognition accuracy(PRA) and error type identification accuracy(EIA) while we take D's transcription as Test data, which also shows that our JE-G2M model is more similar to D's transcription. As was compared before, D's transcription is more similar to dictionary sequence than K's, so our JE-G2M model can detect phoneme error not so strict as K.

Then we can compare models learned from three kinds of transcription, K's, D's and K's+D's. JE-G2M model will have a higher accuracy when compared with D's transcription than comparing with K's transcription, in all the three tables. And JE-G2M model learned from D's transcription has the highest accuracy when taking D's transcription as test data. And when we take K's transcription as test data, the model with highest accuracy is in the models

Table 6.14 Coverage of between results from JE-G2M(learned from D's transcription) and test data (%)

Test data	Dict added	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3
D	No	62.78	60.65	59.09	58.81
	Yes	67.76	65.34	64.35	64.35
K	No	28.12	25.85	25.14	25.00
	Yes	29.12	27.56	27.13	27.13

Table 6.15 Coverage between results from JE-G2M(learned from K's transcription) and test data (%)

Test data	Dict added	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3
D	No	36.36	51.14	50.85	50.43
	Yes	59.09	61.22	61.36	61.36
K	No	46.59	43.61	39.35	38.07
	Yes	49.29	45.03	41.05	40.20

Table 6.16 Coverage between results from JE-G2M(learned from K's+D's transcription) and test data (%)

Test data	Dict added	JE-G2M640	JE-G2M24KG1	JE-G2M24KG2	JE-G2M24KG3
D	No	62.22	58.66	55.68	56.11
	Yes	67.33	64.63	63.64	63.92
K	No	47.44	40.77	37.36	36.51
	Yes	48.30	42.05	38.78	37.93

trained from K's transcription but models trained from K's+D's transcription don't have so much difference and models trained from K's+D's transcription have a higher accuracy when taking D's transcription as test data. Altogether, models trained from D's transcriptions have a better performance in D's test data than K's test data and models trained from K's transcriptions have the highest accuracy in K's test data.

Throughout the whole experiment, what we want to build is a simulator that can detect phoneme error as a native speaker. Because how they can adapt to Japanese speech and how they exactly hear will affect strictness of error detection, so as introduced before, D is not so strict as K, while our JE-G2M model has a higher accuracy when taking D's transcription as test data. So we build a phoneme error detector not so strict as K but similar to D.

Furthermore, we also calculate the coverage of these models, which are shown in Table. 6.14 ,Table. 6.15 and Table. 6.16. In this three table, we calculate the coverage at the word level, which means for each word in test data, we generate a series of phoneme sequences. If the whole phoneme sequences contain exactly the same as the label in the test data, it will be counted as covered. Because we calculate it at "word level", so the number of coverage will lower than PRA or EIA before. And "Dict add" means that the coverage is calculated based on the variations with a combination of g2p results and dictionary sequence. Now let's focus on the number in each table, because we add canonical phoneme sequences in some case, so the coverage with the addition of dictionary will be higher than no-addition version. And in each table, most of the models are have a lower coverage in the following order: JE-G2M640, JE-G2M24KG1, JE-G2M24KG2 and JE-G2M24KG3, except for coverage between models learned from learned from K's transcription and D's transcription in test data. What we considered before is that we will get a better model if the coverage grows with more iteration, and it is different from what we thought before. Maybe this method can make an improvement by making the coverage higher. Besides, We can find that the JE-G2M640 trained from certain data will have more coverage in the same kind of test data, which means that model trained from K's data will have a higher coverage in K's test data and a lower coverage in D's test data. But JE-G2M24K will improve this dependence by creating labels for unlabelled utterances from the last generation and learns from it. That is maybe where JE-G2M24K has effects.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

In this thesis, we mainly present the experiment on applications of Computer-assisted language learning (CALL) system in two directions, listening and speaking.

First, we discuss the fact that listening robustness of language learners maybe not so good as native speakers. To make verification on whether this fact exist. We firstly design a pretest with two kinds of distortion, vocal tract length manipulation and radio simulation. The result of pretest shows that learners are weak in understanding distorted sounds. Then we design a listening drill and train learners with this distorted sounds. After the listening drill, the post-test was carried out and the result shows that high-level learners can have progress in post-test, which is also show the effectiveness of our listening drill.

Second, we investigate modeling learners' pronunciation variations and its application to phoneme error detection. First, we introduce corpus we have right now, and based on this 800 labeled data, we built a series of variation model, AE-T2M, JE-G2M and JE-M2M. Then we use the JE-G2M which is built from 640 labeled data to try phoneme error detection (the 160 ones remain to be the test data). And because we can use the result of 23944 unlabeled date to build a new variation model, called as JE-G2M24k, and of course, JE-G2M24k can be updated literately. Finally, we compare their performance on test data.

7.2 Future works

In future studies, we are going to:

- For robust listening, more kinds of noise will be made and we will consider the question of why native speakers can understand this extremely distorted speech from a scientific point of view.
- For phoneme error detection, more labeled data may be found and expect for joint-sequence n-gram model, some deep learning methods, such as sequence-to-sequence or LSTM, can also be used. Maybe these deep learning methods can have a better result.

References

[1] Voice Conversion Challenge 2016. http://vc-challenge.org/vcc2016/index.html.

[2] Jared Bernstein. Objective measurement of intelligibility. In *Proc. ICPhS*, pages 1581–1584, 2003.

[3] Anne Cutler, Andrea Weber, Roel Smits, and Nicole Cooper. Patterns of english phoneme confusions by native and non-native listeners. *The Journal of the Acoustical Society of America*, 116(6):3668–3678, 2004.

[4] Sabine Deligne and Frederic Bimbot. Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. In *Acoustics, Speech, and Signal Processing, 1995. ICASSP-95., 1995 International Conference on*, volume 1, pages 169–172. IEEE, 1995.

[5] Richeng Duan, Tatsuya Kawahara, Masatake Dantsuji, and Hiroaki Nanjo. Transfer learning based non-native acoustic modeling for pronunciation error detection. In *SLaTE*, pages 42–46, 2017.

[6] John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93, 1993.

[7] Tetsuya Hashimoto, Hidetsugu Uchida, Daisuke Saito, and Nobuaki Minematsu. Parallel-data-free many-to-many voice conversion based on dnn integrated with eigenspace using a non-parallel speech corpus. *Proc. Interspeech 2017*, pages 1278–1282, 2017.

[8] Hyosung Hwang and Ho-Young Lee. The effect of high variability phonetic training on the production of english vowels and consonants. In *Proceedings of the 18th International Congress of Phonetic Sciences edited by Maria Wolters, Judy Livingstone, Bernie Beattie, Rachel Smith, Mike MacMahon, Jane Stuart-Smith and Jim Scobbie. The Scottish Consortium for ICPhS*, 2015.

[9] K.Lenzo. http://www.speech.cs.cmu.edu/cgi-bin/cmudict.

[10] Gaku Kotani, Daisuke Saito, and Nobuaki Minematsu. Voice conversion based on deep neural networks for time-variant linear transformations. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2017*, pages 1259–1262. IEEE, 2017.

[11] ML Garcia Lecumberri and Martin Cooke. Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America*, 119(4): 2445–2454, 2006.

[12] Scott E Lively, John S Logan, and David B Pisoni. Training japanese listeners to identify english/r/and/l/. ii: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the acoustical society of America*, 94(3):1242–1255, 1993.

[13] Takehiko Makino. English read by japanese phonetic corpus.

[14] Takehiko Makino and Rika Aoki. English read by japanese phonetic corpus: an interim report. *Research in Language*, 10(1):79–95, 2012.

[15] Hinako Masuda and Takayuki Arai. Perception of english voiceless fricatives by japanese and english native listeners under various signal-to-noise ratios. In *Proceedings of the Spring Meeting of Acoustic Society of Japan*, pages 471–474, 2011.

[16] Hinako Masuda and Takayuki Arai. Perception of/r/and/l/in quiet and multi-speaker babble noise by japanese and english native listeners. In *Proc. Spring Meet. Acoust. Soc. Jpn*, pages 477–480, 2012.

[17] Hinako Masuda and Takayuki Arai. "perception of voiced english consonants in quiet and multi-speaker babble noise by japanese and english native listeners,". In *Proc. Autumn Meet. Acoust. Soc. Jpn*, pages 361–364, 2012.

[18] Hinako Masuda, Takayuki Arai, and Shigeto Kawahara. Preliminary analysis on the identification of english consonants in noise and/or reverberation by native japanese and english listeners. In *Proc. Autumn Meet. Acoust. Soc. Jpn*, pages 417–420, 2013.

[19] Nobuaki Minematsu, Koji Okabe, Keisuke Ogaki, and Keikichi Hirose. Measurement of objective intelligibility of japanese accented english using erj (english read by japanese) database. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[20] Josef Robert Novak, Nobuaki Minematsu, and Keikichi Hirose. Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the wfst framework. *Natural Language Engineering*, 22(6):907–938, 2016.

[21] Michael Pitz and Hermann Ney. Vocal tract normalization equals linear transformation in cepstral space. *IEEE Transactions on Speech and Audio Processing*, 13(5):930–944, 2005.

[22] Xiaojun Qian, Helen Meng, and Frank K Soong. On mispronunciation lexicon generation using joint-sequence multigrams in computer-aided pronunciation training (capt). In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[23] Daisuke Saito, Nobuaki Minematsu, and Keikichi Hirose. Rotational properties of vocal tract length difference in cepstral space. *Journal of Research Institute of Signal Processing*, 15(5):363–374, 2011.

[24] S.Kabashima, H.Zhang, D. Saito, N. Minematsu, S. Kobashikawa, and R. Masumura. Quantitative and corpus-based analysis of pronunciation diversity observed in japanese english. In *TECHNICAL REPORT OF IEICE*, 2018.

[25] Theban Stanley, Kadri Hacioglu, and Bryan Pellom. Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system. In *Speech and Language Technology in Education*, 2011.

[26] Y. Awaji S.Tetsuhito. If you really want to change the learner's pronunciation, it is limited to explicit correctional feedback. Language Education Expo 2017, 2017.

[27] EIKEN tests. http://www.eiken.or.jp/eiken/en/grades/.

[28] Ian Thompson. Japanese speakers. *Learner English: A teacher's guide to interference and other problems*, pages 296–309, 2001.

[29] Janice WS Wong. The effects of high and low variability phonetic training on the perception and production of english vowels/e/-/æ/by cantonese esl learners with high and low l2 proficiency levels. 2014.

[30] 850 Basic English Words. https://en.wiktionary.org/wiki/Appendix:Basic_English_word_list.

[31] Kakeru Yazawa, Takayuki Konishi, Keiko Hanzawa, Greg Short, and Mariko Kondo. Vowel epenthesis in japanese speakers' l2 english. In *ICPhS*, 2015.

[32] H. Zhang, Y. Inoue, D. Saito, N. Minematsu, H. Masuda, and Y. Yamauchi. Experimental study for enhancing listening capabilities by using speech modification technologies. In *Proc. Autumn Meet. Acoust. Soc. Jpn*, 2017.

[33] H. Zhang, Y. Inoue, Saito D., Minematsu N., Yamauchi Y., and Masuda H. Automatic speech quality control of english listening materials and examination of japanese learners' listening ability in terms of robustness. In *TECHNICAL REPORT OF IEICE*, 2018.

[34] H. Zhang, Y. Inoue, D. Saito, N. Minematsu, and Y. Yamauchi. Experimental study on enhancing listening comprehension with acoustically modified speech materials. In *Proc. Spring Meet. Acoust. Soc. Jpn*, 2018.

[35] H. Zhang, D. Saito, N. Minematsu, and S. Kobashikawa. Modeling of japanese english pronunciation diversity and its application to automatic phoneme error detection. In *Proc. Autumn Meet. Acoust. Soc. Jpn*, 2018.

Appendix A

Publications

National conferences and meetings

- Haoyu Zhang, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, Hinako Masuda and Yutaka Ymauchi “Experimental Study for Enhancing Listening Capabilities by Using Speech Modification Technologies’. In 日本音響学会講演論文集, pp. 371-372, 2017-9.

- Haoyu Zhang, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, Yutaka Ymauchi and Hinako Masuda “Automatic speech quality control of English listening materials and examination of Japanese learners’ listening ability in terms of robustness”. In 電子情報通信学会音声研究会資料, pp. 31-34, 2018-1. （学生ポスター発表賞受賞）

- Haoyu Zhang, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu, Yutaka Ymauchi “Experimental study on enhancing listening comprehension with acoustically modified speech materials”. In 日本音響学会講演論文集, pp. 1367-1370, 2018-3.

- Suguru Kabashima, Haoyu Zhang, Daisuke Saito, Nobuaki Minematsu, Satoshi Kobashikawa, and Ryo Masumura, “Quantitative and corpus-based analysis of pronunciation diversity observed in Japanese English”. In 電子情報通信学会音声研究会資料, pp. 69-74 (2018).

- Haoyu Zhang, Daisuke Saito, Nobuaki Minematsu and Satoshi Kobashikawa “Modeling of Japanese English pronunciation diversity and its application to automatic phoneme error detection”. In 日本音響学会講演論文集, pp. 1045-1048, 2018-9.

- Haoyu Zhang, Daisuke Saito, Nobuaki Minematsu and Satoshi Kobashikawa “Modeling learners’ pronunciation variations and its application to phoneme error detection". In 電子情報通信学会音声研究会資料, (2019).

International conferences and meetings

- Haoyu Zhang, Yusuke Inoue, Daisuke Saito, Nobuaki Minematsu and Yutaka Yamauchi “Computer-Aided High Variability Phonetic Training To Improve Robustness Of Learners’ Listening Comprehension". in ICPhS, 2018 (submitted)