Automatic Proficiency Evaluation of Spoken English by Japanese Learners for Dialogue-Based Language Learning System Based on Deep Learning

A DISSERTATION

SUBMITTED TO THE DEPARTMENT OF COMMUNICATIONS ENGINEERING,

GRADUATE SCHOOL OF ENGINEERING

AND THE COMMITTEE ON GRADUATE STUDIES

OF TOHOKU UNIVERSITY,

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN ENGINEERING

JIANG FU

FEBRUARY 2020

Copyright © 2020 by Jiang Fu

ABSTRACT

AUTOMATIC PROFICIENCY EVALUATION OF SPOKEN ENGLISH BY JAPANESE LEARNERS FOR DIALOGUE-BASED LANGUAGE LEARNING SYSTEM BASED ON DEEP LEARNING

Jiang Fu

The rapid development of globalization leads to an increase in the demand for the second language (L2) learners to study foreign languages. English is the most universal language for speech communication along the world. Traditional classroom English teaching indoors contributes to timely interaction between the teachers and the students. However, such a language teaching method tends to be costly and time-consuming. With the widely used electrical devices and the Internet, it is possible to provide a variety of materials for the needs of English self-study. This method is still hard for them to check the language proficiency as there is no teacher to give feedback directly. In most cases, this boring way of learning alone can make learners lose interest in learning a new language. Due to the increasing computing power and modern information technology in the past two decades, computer-assisted language learning (CALL) system has attracted much attention in the field of language teaching, as an efficient tool for the L2 learners.

Speech-based computer-assisted language learning (CALL) systems should recognize the utterances of the learner with high accuracy and evaluate the language proficiency of the specific speaker with appropriate methods. In this thesis, we discuss the automatic proficiency assessment of the second language (L2) for non-native speakers. For the language proficiency evaluation, pronunciation evaluation and grammatical error detection (GED) are both important components in CALL systems.

Regarding the automatic assessment on language pronunciation, there are many existing works on pronunciation evaluation by applying the goodness of pronunciation (GOP) method. Based on the automatic speech recognition (ASR) technology, we introduce an automatic pronunciation proficiency evaluation system that combines various kinds of non-native acoustic models and native ones, such as Gaussian mixture model (GMM)-hidden Markov model (HMM) and deep neural network (DNN)-HMM. Most of the existing works assume that we know the transcription of an utterance (the reference sentence) when evaluating the utterance, especially in reading and repeating tasks. To realize a reference-free pronunciation evaluation, we propose a novel machine score named as the reference-free error rate (RER) to evaluate English proficiency. In our experiments, the DNN-based non-native acoustic models outperformed the traditional acoustic models on non-native speech recognition. Thus, we calculated the RER by regarding the recognition result from the DNN-based non-native acoustic model as “reference” and the result from the native acoustic model as “recognition result”. The proposed RER has high correlation with human proficiency scores, which indicates the effectiveness of RER for automatically estimating the pronunciation proficiency. By combining the RER with other machine scores such as the log-likelihood scores, we obtained high correlation (reading aloud task: $r = 0.826, p < 0.001, N = 190$; constrained interactive dialogue task: $r = 0.803, p < 0.001, N = 26$; spontaneous English conversation task: $r = 0.799, p < 0.001, N = 28$) to the human scores.

Automatic speech recognition (ASR) and grammatical error detection (GED) are both vital tasks in the modern speech-based CALL system. Compared with a traditional GED system on typing in sentences from the learners, the GED task on speech is different. We designed the oral translation task for learners’ English grammatical usage as the sub-module in our CALL system. To realize a reliable speech-based CALL system with adequate feedback provided for the L2 learners, it is necessary to recognize the learner’s speech as accurately as possible including the grammatical errors at the first stage. To correct oral translation in our system, the recognized sentences by ASR contain ASR errors and grammatical errors. Without the errors in the training data, to some extent, the system would recognize the wrong sentence into the right one. Previous works established handcrafted grammatical error rules and implemented these rules on the prompted right source sentences for generating new sentences with errors, which might cause a problem that it is unclear whether the proposed rules are "optimum" in some senses, as the rules were heuristically established, the generated sentences do not necessarily reflect the comprehensive situation of English spoken by L2 learners. For the GED task in our CALL system, we describe methods for improving the recognition accuracy of speech with grammatical errors. We introduce the process of sentence generation with grammatical errors by adopting the state-of-the-art neural machine translation (NMT) for training a new language model. Then we compare five language models with different training data: 1) baseline LM (language model), 2) rule-based LM, 3) NMT-based LM, 4) rule & NMT-based LM, 5) gold LM (closed-condition). As a result, by comparing the five different language models, the proposed method gives more accuracy on speech recognition rate, which indicates that if the language model contains the human transcribed text of the recorded speech (sentences with realistic errors), the ASR could get a high recognition rate for non-native speech. Therefore, the NMT models generated related errors, which would promote the performance of ASR in the spoken GED module of our CALL system. To a certain extent, it is worth to use the NMT for generating sentences with errors, which will increase the efficiency of the entire CALL system and save labor costs.

Keywords:

Automatic proficiency assessment, Non-native speech, Speech recognition, Computer assisted language learning (CALL), Grammatical Error Detection (GED), Neural Machine Translation (NMT) English study, Deep neural network (DNN)

ACKNOWLEDGMENT

About this thesis, it reveals my last three years' study in the field of Communications Engineering in Tohoku University. Along the way, this research would not have gone so smoothly without the help and the support of many people, who deserve to be thanked and appreciated.

First and foremost, I would like to express my deep gratitude to Prof. Akinori Ito, from the department of Communications Engineering in Tohoku University, who gave me the opportunity for this research and afforded valuable advice in summarizing this thesis, and also for his selfless support, constant guidance and his encouragement in the research direction.

I would like to thank Prof. Hiroki Nishiyama, from the department of Communications Engineering in Tohoku University and Prof. Shuichi Sakamoto, from the department of Research Institute of Electrical Communication in Tohoku University, for their appropriate advice and recommendations in the process of review to this research.

In carrying out this research, I also would like to express my heartfelt gratitude to Associate Professor Takashi Nose, from the department of Communications Engineering in Tohoku University, who gave me conscientious advice and guidance from time to time.

I would like to appreciate assistant Professor Yuya Chiba, from the department of Communications Engineering in Tohoku University for his instructive comments and suggestion on my research.

I would also like to express my appreciation to all my colleagues from Ito-Nose laboratory. With their kind help could I finish this research smoothly.

Finally, I sincerely thank my family for showing a generous understanding of my study aboard and continuing encouragement, along with their economic support.

This thesis was supported by financial aids below.

- JSPS KAKENHI Grant-in-Aid for Scientific Research (A) Grant Number JP17H00823

- Support foundations for improvement of prominent graduate school

These warm supports help to propel my research forward. I would like to express my gratitude to them for their financial supports.

Jiang Fu
Sendai, Japan
February 2020

Contents

List of Tables

List of Figures

1 . Introduction

Based on the English learning of non-native speakers, this thesis explores the development of a dialogue-based computer-assisted language learning (CALL) system centered on modern deep learning (DL) technology. In this initial chapter, we first review the importance of English learning, then describe the methods of learning English and give the introduction of the CALL system with new features. At the end of this chapter, we describe our research objective and give the structure of the entire thesis.

(1) English among L2 learners

With the rapid development of globalization entering into the modern world, humans have more opportunities to communicate with each other. There is an increasing need for English study to the second language (L2) learner whose mother tongue is not English. As one of the most important information carriers, English has become the most widely used language in all fields of human life, yielding the English ability an essential skill. According to some incomplete statistics[3),4),5)] [1], we list some points of English usage scenarios:

1. English is the official language of 60 countries.

2. One out of five people in the world can speak or at least understand English.

3. English is the language occupying the 75% media shows.

Figure 1.1 shows the distribution of English speakers in each region[2]. As the native English

[1] https://en.wikipedia.org/wiki/English-speaking_world#cite_note-Two_thousand_million-2

[2] https://www.babbel.com/en/magazine/how-many-people-speak-english-and-where-is-it-spoken

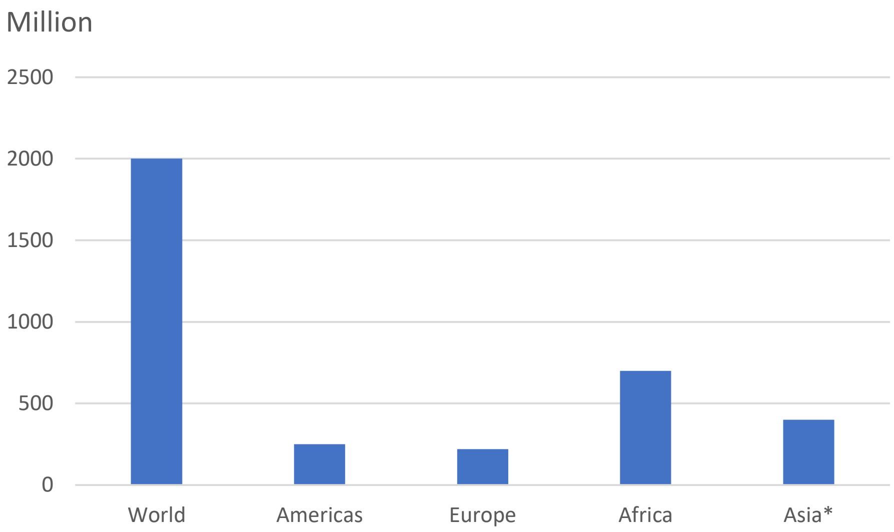

Fig. 1.1 Distribution of people able to speak English in each region.

speakers in Asia are relatively less, to some extent, it is difficult to calculate the number in east Asia. In Figure 1.1, the number of English speakers in Asia is rough and under discussion.

The birth of the Internet has given us more access to information acquisition and knowledge learning. According to a recent web server survey[3], as of January 2020, there were approximately 1.3 billion websites on the Internet. Based on the historical trends in the usage of content languages for websites from January 2019 to January 2020[4], English increases from 54% to 57.4%. By comparing the occupation of different languages used for websites, English obviously ranks the top one among all languages, which reaches 57.4%. The usage of Japanese and Chinese occupies 2.8% and 1.4%, which has a little decrease compared with that of one year ago. Therefore, it is possible to calculate approximately 730 million English content websites. Through learning English well, we would have the ability to view more than 700 million English-language based websites.

Besides, in terms of research, the number of English documents is also very large. Learning English well helps us with academic research.

Take the google scholar search as an example, since this thesis is also a research in the field of

[3] https://news.netcraft.com/archives/2020/01/21/january-2020-web-server-survey.html

[4] https://w3techs.com/technologies/history_overview/content_language

communication, search Japanese word “Tsushin” will obtain about 3.7 million results (as the word in Japanese and Chinese are the same, the number in actual situation will be much less) and on the other hand, searching English word “communication” gets about 6.65 million results.

As the most widely spoken language in the world, English covers all aspects of human life. Whether it is a simple information exchange or rigorous literature, the usage of English accounts for a large proportion. To the non-native English speakers, learning English well is like opening a door to a new world. We can no longer be banned from the original channels of access to information and knowledge, and finally, achieve the goal of diversified learning and value.

Speak English in Japan As discussed in the above, no official data shows the exact number of English speakers in Japan. The real situation relies on the education of the school and the English proficiency level of the Japanese learners. According to a survey in 2013[5], among the 1200 Japanese respondents aged from 20 to 49, 41.6% of them can’t speak English and 30.4% of them can barely emerge few of English words. It is estimated that 20% of the respondents might be able to communicate using English. Regarding the studying willingness, 57% of them expressed the confirmation of learning English.

The “English education reform implementation plan in response to globalization”[6] was proposed by the Ministry of Education, Culture, Sports, Science and Technology in Japan since 2013, along with the decision to host the 2020 Tokyo Olympic Games, which have made rapid progress in accelerating English education for Japanese learners.

(2) Previous methods of learning English

Difficulty in English study for Japanese learners First and foremost, before digging into the English learning methods, we should discuss the difficulties in English study for Japanese learners.

Here are three major reasons why Japanese learners suffer from learning English:

[5] http://economic.jp/?p=29982

[6] http://www.mext.go.jp/b_menu/houdou/24/09/attach/1326084.htm

1. Japanese and English are very different languages, due to the fundamental differences in grammatical and expression aspects, as well as the differences in pronunciation.

2. English study time for Japanese is not sufficient.

3. Japanese students rarely speak English as there are seldom chances for communication in daily life.

a) Classroom English teaching

Traditional language classroom teaching with human teachers contributes to timely interaction between the teachers and the students, which might be the most ideal way for language learners. It is possible to practice English conversation with the teacher and obtain the appropriate feedback to various errors, therefore, this seems to be the most effective way of promoting the listening and speaking skills.

Regarding the rapid development of the Internet and communication technology, it is possible for the learner in his or her home studying English with a human teacher on the other side via computer's communication software. This Internet teaching can be regarded as one way of one-on-one language teaching. However, since one-on-one language teaching is extremely expensive, the practical English study classroom is always a form of collective large class education, involving only one teacher with dozens of students. Hence, it is common that all of the students in one classroom receive the same English education which ignores the different language proficiency level and the learning ability among the students. The teaching method might be suitable for knowing the level of the corresponding student to a certain extent, but the effectiveness of normal education is still under consideration. Moreover, due to the nature of such a classroom, there might be constrained conditions on the time and the place which seems to be a not perfect way for English study.

b) Self-learning for English

Self-learning is a way that uses the available commercially teaching materials, textbooks and video lessons, which are studied by the student himself or herself. Especially the nowadays Internet age makes it possible to provide a variety of materials which meets the needs of self-study. The student can study at his or her own pace. The biggest advantage of self-study is to cultivate self-discipline. If you choose the right materials, plus strong self-study and self-discipline, such learning is largely effective.

However, the self-study method only trains on the language skills of “reading”, “writing” and “listening”, to which, the “speaking” ability is not developed which can be the key skill for the English communication to Japanese learners. When conducting the self-study, especially in the process of “reading” or “self-speaking”, it is not easy for the language learners to check where the pronunciation or grammatical errors are. For the pronunciation aspect, they may record the speech from themselves and compare it with the right pronunciation. However, it is still hard for them to check as there is no teacher to give feedback directly, and that will cost much of time with low efficiency. In most cases, this boring way of learning alone can make people lose interest in learning a new language.

(3) Computer-assisted language learning system

Due to the increasing computing power and modern information technology in the past two decades, the educational field and learning style have been changed, the language learners can use applications on the computers or the smartphones for a new language learning. Such a foreign language learning program is called computer-assisted language learning (CALL) system. Regarding the definition of a CALL system, any method or process in the system that a learner can use, as a result, improves his or her language proficiency[6)]. The CALL system links various kinds of research aspects, such as psychology, artificial intelligence, human-computer interaction and linguistics.

The CALL system has attracted much attention in the field of language teaching, as an efficient tool for L2 learners. The CALL systems have easier access and more flexibility with lower cost and saving time than the traditional classroom based-language learning. The state-of-the-art speech-based CALL system can recognize the words or sentences uttered by the L2 speakers via automatic speech recognition (ASR) technology.

Table 1.1 The features between the reading aloud CALL and the dialogue-based CALL.

Reading aloud CALL	**Dialogue-based CALL**
Prepared material to read; recorded speech to repeat	Scenario; free-conversation
Boring, tiresome	Friendly, emotional
Difficult to system maintain	Self-study (Chat bot; AI assistant)

a) Traditional language learning system

The origins of CALL can data back to the 1960s[7]. Traditional CALL systems always presented a stimulus on the screen of the computer to the learner who had to respond with the answer by type in with the keyboard or the mouse. Further in the traditional CALL system is that audio content and dictionaries are embedded there as a simple feature machine[8].

b) Dialogue-based CALL system

The dialogue-based CALL system is derived from the speech-based CALL system. The early speech-based CALL are systems only equipped with repeating and reading tasks, in which some reference-dependent automatic scoring methods were developed[9],[10].

However, a CALL system only giving practice of reading and repeating is always restricted and constrained for language study. To improve the learner's oral communication skills, a few interactive conversational CALL systems were researched[11],[12],[13]. Table 1.1 shows the different features between reading aloud CALL and dialogue-based CALL system.

The Speech Rater[12] sponsored by ETS is not only for research development, but also taken as a commercial application on the English test. The system aimed at automatically evaluates spontaneous speech based on a large vocabulary continuous speech recognition (LVCSR) methodology.

POSTECH's interactive English CALL system[13] is intended for Koreans, with which the learners can practice English conversation in constrained situations such as shopping and asking directions. This system prepared the example for each dialogue in advance, the learner's pronunciation would be scored with feedback after compared with that prepared example.

Table 1.2 The scenario in shopping.

Speaker	Speech
System:	Can I help you?
Learner:	I want two lemons，three peaches，and two packs of cherries，please．
System:	Is that all?
Learner:	Yes．
System:	OK．That would be nine hundred and eighty yen．
Learner:	Here you are．
System:	Thank you very much．Here’s your change．

Table 1.3 The scenario in breakfast.

Speaker	Speech
System:	What do you have for breakfast?
Learner:	I often have toast and milk．How about you?
System:	I have rice and miso soup．
Learner:	Anything else?
System:	I also have an egg，natto，and tofu．
Learner:	Oh，you eat a lot!
System:	I’m hungry in the morning．

What we want to do in this study is to build a dialogue-based CALL, which can be a conversational partner to L2 learners for practicing English, anytime and anywhere. Our dialogue-based CALL system also prepared the sub-module by adopting the prompt samples for conversation[2)], where more details and evaluation will be shown in Chapter 3. Table 1.2 and Table 1.3 describe the content of the samples in our CALL system.

c) Automatic language proficiency evaluation

Pronunciation evaluation A large number of researchers have investigated pronunciation evaluation in terms of various scoring methods ranging from phonemic domain[9),14),15),16)] to prosodic domain[17),18),19)].

There are three main types of automatic pronunciation evaluation on the phonemic side:

1. Evaluating the overall goodness of one sentence or a pronounced specific word. Franco *et al.*[9)] indicated that the combination of machine evaluation scores from ASR system, e.g., phone log-posterior probability scores based on Gaussian mixture model (GMM)-hidden Markov model (HMM), segment duration scores and timing scores, achieved a high correlation with human rating scores in sentence-level. In addition, the correlation between the mentioned machine scores and human scores had a comparable result to human-to-human correlation when an adequate amount of evaluated speech data is available[20)]. Speech recognition technology was also used in the pronunciation evaluation system to increase human-machine score correlation[21)].

2. Evaluating goodness of a specific phoneme. L2 learners can use this system to derive a score for one arbitrary L2 phoneme they selected in their non-native speech evaluation. In this branch, Kim *et al.*[22)] set a series of particular phonemes to examine how well the machine scores correlate with the corresponding human scores rating of single phone utterances. To validate the automatic pronunciation evaluation results, Witt and Young[23)] developed a system for scoring pronunciation of each phoneme in an utterance and detecting pronunciation errors. The most well-known approach might be the Goodness of Pronunciation (GOP)[24)], which is an approximation of the probability of the target phoneme.

3. Detecting and diagnosing mispronounced phonemes. Compared with evaluating goodness of a specific phoneme, such kind of system could give a direct and accurate output on whether the pronounced phoneme is acceptable and unacceptable, or the comparison between correct phoneme and wrong one. Besides the general approach, the discriminative approach is often applied in these systems for pronunciation modeling. Franco *et al.*[25)] proposed a method to conduct automatic mispronunciation detection with phone segment score and a log-likelihood ratio

score calculated from two dissimilar acoustic models trained by non-native speech, one from acceptable correct native-like speech and the other one from extremely non-native speech. Regarding the performance improvement in mispronunciation detection, Ito *et al.*[26] developed a decision tree-based clustering method for phoneme-level classification. In addition, approaches using the deep neural network (DNN)-based acoustic models significantly improved the results of mispronunciation detection and diagnosis[27),28),29),30)].

We aim at improving the performance of pronunciation evaluation in type 1.

Evaluation of language pronunciation on L2 learners, to some extent, is an important objective as well as a difficult mission. Even in the case of traditional human evaluation, the inevitable variability in the evaluation standard of multiple teachers leads to less objective and consistent assessments. As an approach of pronunciation evaluation, the GOP is widely used in phone level pronunciation scoring of non-native speech[31] and utterance verification assessment[32]. GOP is originally calculated through the GMM, and then a different GOP estimation method based on DNN has been proposed[33].

To improve the evaluation accuracy, the native and non-native acoustic models are used for measuring whether a specific utterance sounds native-like or non-native like[31),34),14)]. The acoustic model in an ASR system trained from native speech database is regarded as a baseline to measure how close the given non-native speech is to the target native pronunciation, as the native acoustic model includes nearly whole phonemic features of the native language. The phoneme-level log-likelihood score and posterior probabilities generated from the native acoustic model achieved a high correlation to human score by using linear regression for non-native speech[35], and the log-posterior probabilities of aligned phonemes in the GOP are also obtained from an acoustic model trained with native speech. However, the speech from non-native speakers is more or less influenced by the accent, rhythm, or other phonemic characteristics of their mother tongue (L1). The DNN-based acoustic models trained from non-native English speech uttered by Japanese native speakers gave very high accuracy for recognizing English speech by Japanese speakers[36]. There exists a plenty of studies that use both native and non-native acoustic models to conduct automatic pronunciation evaluation[37),14),38),16)].

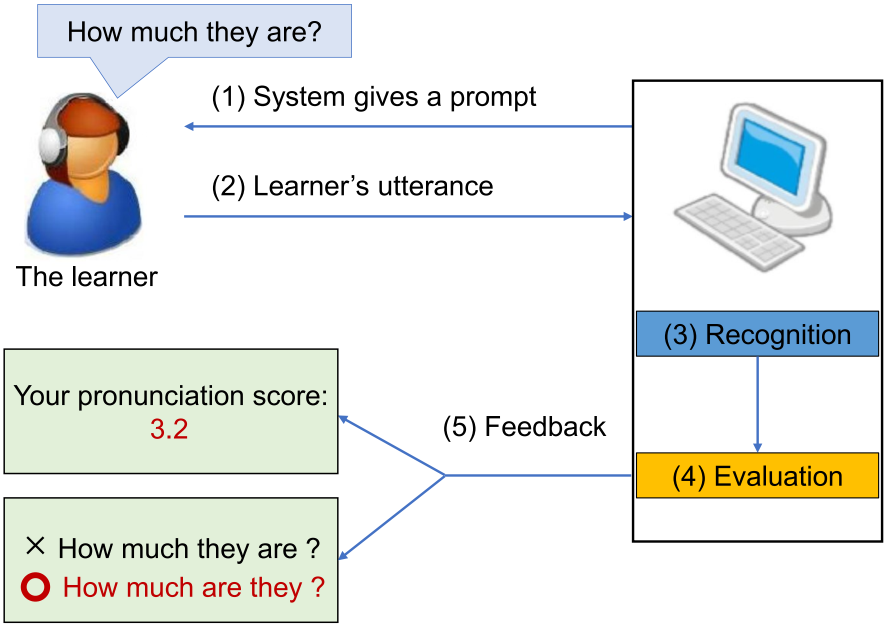

Fig. 1.2 A example of our dialogue-based CALL system with adequate feedback on pronunciation and grammatical errors.

Grammatical error detection To realize a speech-based CALL system, language proficiency evaluation in such a CALL system is always a challenging task. Regarding the English speech from L2 learners, in most cases, it contains mispronunciations and grammatical errors largely due to the influence of the learners' mother tongues.

The evaluation on pronunciation above means how to pronounce the word correctly at the phonemic-level, however, which has no correlation with the content of the speech. As a complete automatic advanced CALL system, it not only judges the pronunciation, but also further understands what the learner is saying to judge whether the learner has errors on the content or grammatical aspect. That is the speech-based grammatical error detection (GED) task in the dialogue-based CALL system.

For most of the GED system, they are often applied to the written texts with errors. Spoken speech has much difference to the written text, such as the special action features happened in the spoken utterance, no punctuation there regarded as no grammatical error, and no spelling mistakes in the word.

For the GED task in the spoken dialogue-based system, the quality of ASR is always depended on its two main components: acoustic models (AMs) and language models (LMs). The AMs describe how the sounds are pronounced in a language and the LMs cover the probabilities of words or sequences of words. Not only the AMs with extremely high performance is needed, but also language modeling is very important as the language model contains a lot of verbal information (correct or wrong grammar and expression). A good language model helps to improve the speech recognition rate, which further makes the GED system better.

With respect to the above two aspects of the English proficiency evaluation, an example of our dialogue-based CALL system is showed in Figure 1.2. This system has the function of talking based on ASR technology. By received what the learner has said, it is normal for the system to translate the speech into text and then make an analysis on the speech. Finally, appropriate feedback and score are proposed to the learner with the content on the screen.

(4) Purpose

a) Problem in previous work

For pronunciation evaluation, most systems assume that the transcription of a learner ’ s utterance can be exploited at the time of evaluation. It can be realized by letting a system specify the sentence to be read. However, the text-dependent evaluation cannot be used for evaluation on any spontaneous speech, which is essential to combine a spoken-dialog-based CALL system with pronunciation evaluation. Moustroufas and Digalakis[14)] proposed a system that used acoustic models in both L1 and L2 and compared the sentence-level likelihood scores from these acoustic models for any utterance. Still, there is nearly no other advanced method for pronunciation evaluation on non-native speech.

Regarding the grammatical error detection on spoken language, since an ordinary English corpus does not contain many grammatical mistakes, a language model trained using such as corpus does not model grammatical mistakes. Thus, Anzai[39)] developed a rule-based method to artificially generate sentences with grammatical errors that Japanese learners tend to make. They developed error rules according to the corpus that contains mistakes in spoken English sentences[40)], and applied the rules to correct English sentences to generate training data for the language model. A problem of their method is that it is unclear whether the developed rule is “optimum” in some sense. As the rules were heuristically developed, the generated sentences do not necessarily reflect the tendency of actual sentences spoken by Japanese learners.

b) Purpose in this research

The first objective of our research is to develop a highly accurate evaluation system for sentence-level pronunciation evaluation that can be used for any utterance, assuming both L1 and L2. For developing the reference-text-independent pronunciation scoring in a dialogue-based CALL system, we propose a novel machine score (Reference-free Error Rate, RER) that is based on speech recognition results using native and non-native acoustic models. The DNN-HMM acoustic models are utilized for improving the accuracy of evaluation. We investigate pronunciation evaluation by combining multiple machine scores including log-likelihood scores and the proposed RER.

The second target is to exploit the neural machine translation (NMT), which is able to convert correct English sentences into those with grammatical errors, to obtain the training data for language modeling. In that way, we can promote the speech recognition rate unless the generated sentences from NMT have realistic grammatical errors. Furthermore, the output of the ASR system could be directly utilized for a normal grammar checker. To this end, we compare the proposed language model with the conventional ones and analyze the performance of the proposed language model in grammatical error detection task.

(5) Thesis outline

In the second chapter, a more detailed description of previous work and the key technology used in the modern dialogue-based CALL systems are described. Next, the two main proficiency evaluation outlined above will be tackled in the third and fourth chapters. In the third chapter, an automatic pronunciation proficiency evaluation system for English utterances from Japanese native speakers is built by utilizing both native and non-native acoustic models. To this end, the RER is proposed as a new machine score for the regression-based prediction of proficiency. Meanwhile, in the fourth chapter, language modeling with different training data would be discussed and investigated. Neural machine translation (NMT) would be applied for the generation of the training data, which is used as the proposed language model. Finally, in the fifth chapter, the conclusions of the thesis will be described.

2 . Previous work

In this chapter, the five key points in the dialogue-based CALL system are explained along with the previous work: automatic speech recognition (ASR) module, grammatical error detection (GED) task, machine translation (MT), modeling toolkits and required data sets in the system. These main components are related with the fundamental speech technologies needed in a dialogue-based CALL system, or more or less required to establish the system. Towards to an initial comprehensive understanding of the specific CALL system, this chapter just meets the demand.

(1) Automatic speech recognition

Automatic speech recognition (ASR) has attracted the human interest for almost 60 years. The goal of this technology can be served as a bridge for smooth communication between human and human, or human and machine. Since the 1960s, the University of Tokyo, NEC Lab in Japan, Carnegie Mellon University in the United States, and scientists from the former Soviet Union have successively proposed several basic concepts for speech recognition, laying a solid foundation for the future development of ASR. The most brilliant achievement of ASR in 1980s is that the center of technology has shifted from pattern matching to statistical model methods, and in particular, Hidden Markov Model (HMM) have made great progress[41),42),43)].

However, ASR has not been a vital form of human-machine communication in the past, largely due to the backward information technology and very small amount of available speech data. Speech technology substantially changes the way we live and work. Through some tiny digital devices, voice interaction became the main type of human-machine communication. Here are the reasons for this trend:

1. The continuous improvement of computing power and modern information technology, making it possible to train a more complex and powerful acoustic model and language model.
2. With the ubiquitous Internet, we can collect a lot of speech resources.

For developing a friendly human-machine dialogue-based CALL system, utilizing the modern ASR technology is necessary and important.

The purpose of ASR is to convert speech into text. Specifically, it is to input a speech signal and find out a sequence of words (consisting of words or characters), so that it matches the speech signal to the highest degree. This degree of matching is generally expressed in terms of probability. Let x be the speech signal and w represent the sequence of words, then the question can be solved as follows:

$$\hat{w} = \arg\max_{w} P(w|x) \tag{2.1}$$

It is generally believed that calculating $P(w|x)$ is difficult, and normally the speech is generated after having words, namely first comes the words, and then the sounds of them are sent out. Therefore, from Bayesian formula, the above equation is transformed as follows:

$$\hat{w} = \arg\max_{w} \frac{P(x|w)P(w)}{P(x)} \tag{2.2}$$

Here, the denominator $P(x)$ can be omitted as it is a constant to w. Therefore, the equation can be equaled as:

$$\hat{w} \approx \arg\max_{w} P(x|w)P(w) \tag{2.3}$$

The above equation is the core formula in speech recognition. The probability of $P(w)$ expresses how likely a sequence of words itself; $P(x|w)$ is the probability of a speech after a given text. Calculating the values of these two items is the task of the acoustic model ($P(x|w)$) and the language model ($P(w)$). Therefore, how to train an acoustic model and a language model is a big issue in order to improve the accuracy of speech recognition. The description of acoustic model and language model will be explained in the next.

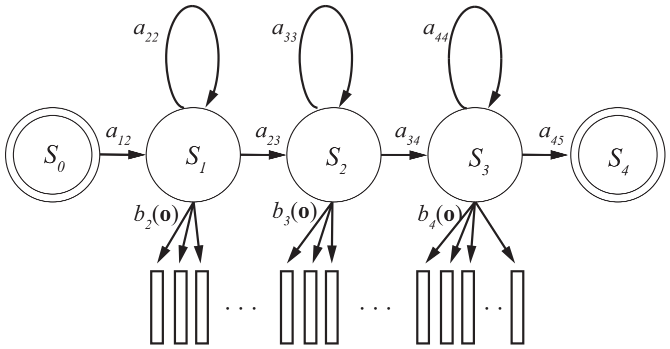

Fig. 2.1 Standard 3-state phone Hidden Markov Model.

a) Acoustic model

The task of an acoustic model (AM) is to compute the observation likelihood $P(x|w)$, which is the probability of uttering the speech after a given text. Here, the primary task is to know what kinds of sounds or symbols should be assigned to each word. The lexicon is another dictionary module in ASR, which converts the word-level string to a phoneme-level string. For the process in ASR, the lexicon is considered to be a component alongside the acoustic model and language model. With the help of the lexicon, the AM knows which sound uttered successively with the given string of text. In order to compute the likelihood of matching between speech and phoneme-level string, it is also necessary to know the start and the end of each phoneme (typically in a real ASR, the linguistic units are always sub-parts of phonemes). This can be done by a dynamic programming (DP) algorithm, which is called Viterbi algorithm. By solving the above work, Hidden Markov model (HMM) is widely applied to the acoustic model. The hidden Markov model is based on the Markov chain and adds observed events. Wherein, the hidden layer is mapped to the observed layer by an emission probability or an observation probability, and the transition between the hidden states is obtained by a transition probability. Figure 2.1 shows a standard 3-state phone HMM. Here, S_0 and S_4 represent the non-emitting start and end states for the specific phoneme, respectively. S_1, S_2 and S_3 are the left-to-right three states. $a_{ij}(1 \leq i, j \leq 5)$ is the probability for each sub-phone of taking a self-loop or going to the next sub-phone. $b_i(\mathbf{o})(2 \leq i \leq 4)$ means the emission probability, expressing the probability of a feature vector being generated from the corresponding sub-phone state $S_i(1 \leq i \leq 3)$.

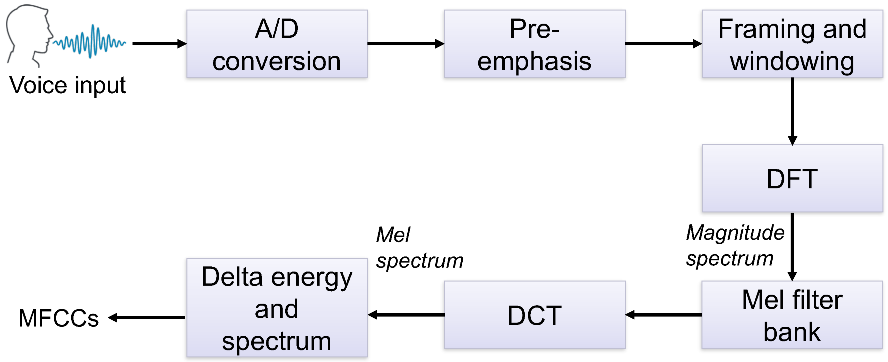

Fig. 2.2 The process of extracting Mel-Frequency Cepstral Coefficient (MFCC) feature.

Through the decoding in HMM by Viterbi algorithm, not only considers the matching degree between each segment of speech and phoneme, but also considers the transition probability between individual phonemes.

In the process of pointing out the phoneme boundary and calculating the matching, the acoustic model should have the original speech signal represented in a typical form, in which the stage is feature extraction. Mainstream speech recognition systems[44)] usually use Mel-Frequency Cepstral Coefficient (MFCCs), filter banks (FBANK) or Perceptual Linear Prediction (PLP) as features. The extraction of MFCCs is showed in Figure 2.2. The speech signal goes through a pre-emphasis filter and is sliced into (overlapping) frames. Then a window function is applied to each frame. Afterward, a Fourier transform is conducted on each frame and the power spectrum is calculated which leads to filter banks. Subsequently, compute the filter banks by applying triangular filters. To obtain MFCCs, the Discrete Cosine Transform (DCT) is applied to the filter banks. Finally, after adding energy and delta and double-delta features, 39 MFCCs are obtained. MFCCs are useful in many cases due to the cepstral coefficients are nearly uncorrelated which will make the acoustic model much simpler.

GMM-HMM From the speech training data, a large number of feature vectors and their corresponding phonemes can be extracted. From this data, we can train a classifier from feature to phoneme. The most commonly used classifier in previous years is the Gaussian Mixture Model (GMM). The general principle of GMM is to estimate the distribution of the feature vector to each

Table 2.1 Example of monophone training and triphone training. Here, “speech” is an example word.

Training type	speech
Monophone	S P IY CH
Triphone	SIL-S+P S-P+IY P-IY+CH IY-CH+SIL

phoneme. Then, in the recognition phase, calculate the probability that the feature vector of each frame generated by the corresponding phoneme, and multiply the probability of each frame. GMM is expressed as follows:

$$b_j(\mathbf{o}) = \sum_{m=1}^{M} c_{jm}\mathcal{N}(\mathbf{o}|\mu_{jm}, \Sigma_{jm}) \tag{2.4}$$

$$\mathcal{N}(\mathbf{o}|\mu_{jm}, \Sigma_{jm}) = \frac{1}{\sqrt{(2\pi)^n|\Sigma_{jm}|}} \exp\left(-\frac{1}{2}(\mathbf{o}-\mu_{jm})^\top \Sigma_{jm}^{-1}(\mathbf{o}-\mu_{jm})\right) \tag{2.5}$$

Here, **o** represents the feature parameter corresponding to one frame of speech. c_{jm}, μ_{jm} and Σ_{jm} are respectively the mixing parameter, mean, and covariance matrix of the m-th Gaussian component of the j-th state. $b_j(\mathbf{o})$ is the observation likelihood that the speech feature parameter **o** belongs to the j-th state, and n is the dimension of the observation vector. By multiplying the probability $P(\mathbf{o}|S_i)$ of each frame, $P(x|w)$ can be achieved.

The HMM model assumes that a phoneme contains 3 to 5 states, the pronunciation of the same state is relatively stable, and one state can transit to a different one with a certain probability. The feature distribution of a state can be described by a probability model, where the model can be GMM. Therefore, in the GMM-HMM framework, HMM describes the short-term steady dynamics of speech, and GMM is used to describe the internal pronunciation features of each state of the HMM.

In the pronunciation process, due to the influence of collaborative pronunciation, the pronunciation of the same phoneme varies in different positions. The monophone is context-independent. In order to be able to represent this difference, the concept of triphone is proposed. Table 2.1 explains the training classification of monophone and triphone. The triphone is determined by the Left and right phonemes, therefore, it is context-dependent.

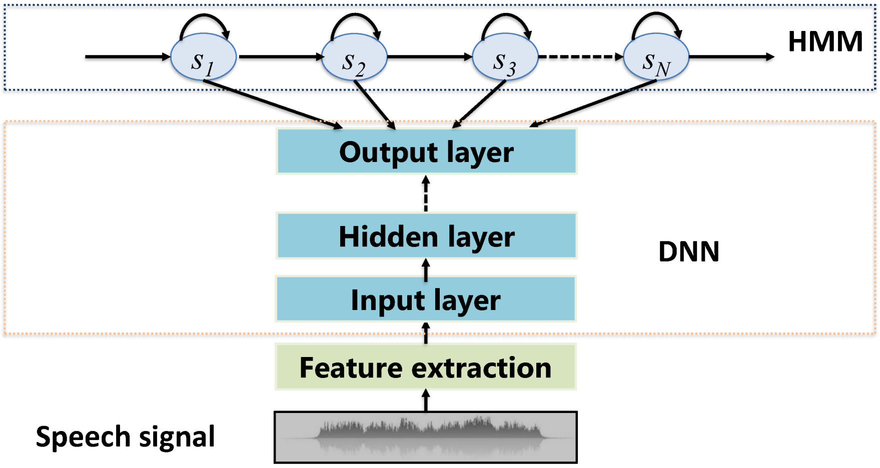

Fig. 2.3 The DNN-HMM model for speech recognition.

DNN-HMM Although GMM-HMM has achieved a lot of success in the past, with the development of deep learning, the deep neural network (DNN) model shows a significant performance over the GMM model[45)], instead of GMM for HMM state modeling. As the speech signal is continuous, there are no obvious boundaries between individual phonemes, syllables, and words. Each phoneme pronunciation is also affected by the context. The advantage of DNN is that it no longer needs to make assumptions about the distribution of speech data, and stitches adjacent speech frames including the time series structure information, which makes the classification on probability of the state significantly improved. At the same time, DNN also has a strong environment learning ability, which can improve the robustness to noise. In simple terms, DNN is the state probability corresponding to the input string of features. Figure 2.3 shows the explanation of DNN-HMM model for speech recognition. Compared with traditional GMM, DNN-based acoustic model is able to directly calculate the value of $P(S_i|\mathbf{o})$. Through Bayesian formula, the $P(x|w)$ is obtained by multiplying the $P(\mathbf{o}|S_i)$ of each frame.

b) Language model

When it comes to the language model (LM), the most familiar usage scenario nowadays is the intelligent keyboard input. By inputting several advanced characters with the smart input method, it can directly give the fully right words or appropriate next word in a sentence. That is the credit of the language model.

Language model is a model that assigns probabilities into sentences or sequences of words. In a word, the purpose of the language model in automatic speech recognition is to give the text sequence with the highest probability according to the output of the acoustic model. The language model generally uses the chain rule to disassemble the probability of a sentence into the product of the probability of each word. Let string w be composed of words: w_1, w_2, $\cdots$, w_n, then $P(w)$ can be broken into:

$$P(w) = P(w_1, w_2, \cdots, w_n) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\cdots P(w_n|w_1, w_2, \cdots, w_{n-1}) \tag{2.6}$$

In the above formula, each individual item is the conditional probability of the current word w_n given previous string of words. However, when the strings are too long, the probability is not easy to estimate. Therefore, based on the Markov assumption, the most common practice is to think that the probability distribution of each word depends only on the last few words in history. Such a language model is called a n-gram language model. The n-gram language model is represented as:

$$P(w) = \prod_{i=1}^{n} P(w_i|w_{i-N+1}, \cdots, w_{i-1}) \tag{2.7}$$

Here, N is the number of dependent words. When N is 1, 2, and 3, the n-gram models are called unigram, bigram, and trigram language models, respectively.

- Small N: If words frequently occurred in the training corpus, that would have higher reliability.
- Large N: More constriction on the next appeared word, with greater discrimination, but needs a lot of training data.

The larger the N in the n-gram language model, the more training data is needed. In theory, the bigger n shows the better; from the experience, the general speech recognition system use trigram models.

Usually, the language model is constructed by calculating the Maximum Likelihood Estimate (MLE), which is the best estimate of the training data. The formula is as follows:

$$P(w_i|w_{i-2}, w_{i-1}) = \frac{N(w_{i-2}, w_{i-1}, w_i)}{N(w_{i-2}, w_{i-1})} \tag{2.8}$$

$N(\mathbf{w})$ means the frequency of a word sequence **w** appeared in the training corpus. The larger the size of the training corpus is, the more reliable the results of the parameter estimation would be. But even if the training data is very large, such as a few GB, there will still be many linguistic strings that have not appeared in the training corpus, which will result in many parameters being 0. This problem is also known as data sparseness. Data sparseness can be solved by Data Smoothing method.

Here are some open source language model tools:

1. The SRI Language Modeling Toolkit (SRILM) [1]
2. The IRST Language Modeling Toolkit (IRSTLM) [2]
3. The MIT Language Modeling Toolkit (MITLM) [3]
4. An *n*-gram Language Model Library from UC Berkeley (BerkeleyLM) [4]

(2) Grammatical error detection

Learning a new language is not an easy task. Especially in the text written by learners taking English as a second language (L2), there are a large number of grammatical errors. Considering the machine has a high performance to automatically detect these errors and assist the L2 learners for English study attracts the attention of researchers.

In general, grammatical error detection (GED) is the task of automatically detects grammatical errors in the written text. To a certain extent, another similar parallel task to GED is the error correction. Both of these two tasks take the text from language learner to be compared with the reference text from the human annotator or the grammar checker. Regarding the error detection, it is only concerned with pointing out parts of text in which are grammatically incorrect. This can be an alternative

[1] http://www.speech.sri.com/projects/srilm/

[2] https://hlt-mt.fbk.eu/technologies/irstlm

[3] https://github.com/mitlm/mitlm

[4] https://code.google.com/archive/p/berkeleylm/

choice of reporting the result to the learners or not. As the GED systems are not always accurately correct to the input sentences, it is better not to report unclear result to the leaner. On the other hand, the task of error correction should give direct comments and suggestions to identify an error in the sentence. For the vast majority of cases in CALL system, there is no actual difference between these two tasks, both need to detect the errors and give the suitable feedback to the language learners.

Specifically, a GED system identifies and corrects any grammatical errors that exist in the sentence by analyzing the context of the input sentence, and finally outputs a sentence without grammatical errors. The output sentence retains the semantics of the original sentence. Of course, if there is no error in the input sentence, the GED system should return the original sentence directly. The following paragraph gives an intuitive example.

Incorrect: Book of my class intresed to me.

Correct_human: A book in my class interested me.

Correct_GED1: Books for my class interested me.

Correct_GED2: The books of my class were interested to me.

From this example, the L2 learner wrote a sentence with errors that including spell error, article error or preposition usage error. Compared with the feedback from a human teacher, the GED systems performed not perfect.

a) Grammatical errors

Generally speaking, the common errors in L2 learners' study include spelling errors and grammatical errors[46)]. The spelling errors open occur in the written text from the language learner, which are easily handled with the application of word processors (e.g., Microsoft Word[5]). On the aspect of grammatical errors, there are a variety of types of classification, such as word missing, word order errors, unnecessary use of word, and misuse of articles, prepositions, noun number and verb form. Here are some detailed descriptions on three main typical grammatical errors:

[5] https://products.office.com/en-us/word

1. Article errors: The concept of article is the word come before a noun which is general or specific, including a, an, the, this, that and so on. It is a common issue to misuse the article word for a new language study.

2. Preposition errors: Almost refer to the relationship between one word to another, especially after the verb. The difficulty here is that there are not many rules for the preposition, therefore, it needs the L2 learners to remember those usages.

3. Verb errors: In this category, subject-verb agreement and the verb tense shift should be noticed.

Spoken speech has much difference to written text, such as the influences in the spoken utterance, no punctuation there regarded as no grammatical error, and no spelling mistakes in the word. Each of these requires some modifications to the standard written text GED setup.

b) Approaches of GED system

In recent years, the GED systems have changed the methods from rule-based to statistical-based, which can be roughly divided into three stages. Here is a brief introduction.

Rule-based approach This type of method corrects errors by handing rules in coding source of the system. The world's first GED system, Writer's Workbench[47)], is based entirely on matching and replacement of strings. For some specific categories of errors, rule-based methods are easy to implement and very effective. Until now, that kind of method is still widely used. However, considering the complexity of the languages, the rule-based approach is not suitable as a general method of GED system.

Traditional machine learning method With the accumulation of corpus resources, data-driven methods have become the mainstream of GED since the 1990s. People used machine learning techniques (naive Bayes classifier, support vector machine (SVM)) to design separate classifiers for different error categories[48)]. The effect of such methods on errors such as articles and prepositions is obvious, but there are some other problems. It only uses the local context information in the sentence and considers different error categories independently, thus can't do anything for some interactive errors.

Machine translation (MT) based approach The idea of regarding GED task as a kind of translation from the wrong sentence into the right one was first derived from Brockett *et al.*[49]. However, this outstanding method gradually became the mainstream of GED until 2014, and then developed rapidly. This method is initially applied with the traditional statistical machine translation (SMT)-based GED system[50]. After that, with the replacement of SMT by Neural Machine Translation (NMT), the MT-based method becomes a core component of the state-of-the-art GED system.

c) Evaluation metrics on GED system

As we have seen the grammatical error detection example in the above paragraph, a GED system is able to detect one or a few kinds of errors in the given text. However, the output sentence from the GED system is not always exactly grammatically correct. Therefore, evaluating the performance of a GED system is necessary. A GED system is generally evaluated by a comparison between the output from the system (to be seen as a hypothesis) and a gold standard reference from the human annotator. Before ready to do the comparison, each position of the system edited string can be divided into a true positive (TP), a true negative (TN), a false positive (FP) and a false negative (FN).

TP: This item means the perfect match in the position where both the GED system's output and the gold standard reference. Here, both of the system and the human annotator indicated that part in the text was not right.

TN: This item also means the perfect match in the position where both the GED system's output and the gold standard reference. Different from TP, both of the systems and the human annotator didn't do any change to the part in the text.

FP: This item means the GED system predicted a change in a part of the text, while the gold standard reference gave nothing changed.

FN: This item means the GED system didn't show a change in a part of the text, while the gold standard reference showed that was false.

Table 2.2 An example for evaluation on GED systems.

Learner	he	like	speak	English	on	*	classroom
Annotator	he	likes	speaking	English	in	the	classroom
GED system	he	liked	speaking	English	on	the	classrooms
Evaluation	TN	TP	TP	TN	FN	TP	FP

This definition referred to Chodorow's work in their evaluation for GED systems[51)]. Table 2.2 shows an example of this evaluation standard.

By counting the four different items, these counts are used for evaluation metrics such as precision (P), recall (R) and F-measure (F). The definition of these metrics are showed as follows:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{2.9}$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \tag{2.10}$$

$$F = \frac{2PR}{P + R} \tag{2.11}$$

P is the proportion of actual errors' number among all the number of the errors predicted by the GED system. It shows the quality of the GED system. R is the proportion of the actual errors from the GED system among all the errors in the text from the learner. F reacts the intermediate ratio of P and R.

(3) Machine translation

In the last section of grammatical error detection, machine translation (MT) based GED system is mentioned. This section will show more details on the MT as not only the machine translation is an efficient method applied in the GED system, but also it can be adopted into the language modeling of ASR for the dialogue-based CALL system (especially presented in the fourth chapter).

Machine translation (MT) automatically translates a natural language text (source language) into another natural language text (target language) by the power of machine. Meanwhile, machine translation is a relatively vague concept. In general, all technologies and services that use machines to help humans with translate activities can be classified as machine translation, such as an online dictionary, translation assistant, etc. In the vast majority of cases, machine translation is a system that can be utilized for a complete sentence or paragraph translation, excluding the only online dictionary function. Besides the input length of the text to the MT, the MT can be specified to a new sentence pair. Regarding the GED system, the input-output sentence pair is not a normal natural language to another natural language, just as a wrong to right sentence pair.

a) Traditional methods

Machine translation can be traced back to the 1950s, the United States began to develop translation machines from Russian to English. At that time, the machine translation was very simple, that is, the corresponding words were automatically translated one by one referred from the dictionary. That time was also an early exploration period in which machine translation was based on the bilingual dictionary and grammar. This is called rule-based machine translation (RBMT).

With the development of statistics, the researchers began to apply statistical models to machine translation, which is based on the analysis of bilingual text corpus to generate translation, known as statical-based machine translation (SMT)[52]. The SMT method performed better than the RBMT and dominated the field for nearly 20 years. Here, in the SMT, there are some similar points between the SMT and the speech recognition. Image a translation task: suppose the source language is x and the target language is y. The goal of machine translation is to find y, so that $P(y|x)$ is the largest, which is showed below:

$$\hat{y} = \arg\max_{y} P(y|x) \tag{2.12}$$

Further, from Bayesian formula, this equation can be split into the product of two probabilities: where $P(y)$ is the language model (LM) introduced earlier, in which the simplest method can be n-gram; $P(x|y)$ is the translation model from the target language y to the source language x.

$$\hat{y} = \arg\max_{y} P(x|y)P(y) \tag{2.13}$$

Where the problem of translation is changed into how to calculate for the translation model $P(x|y)$ and the language model $P(y)$. In general, for an SMT, to train these models, two kinds of training data is required: parallel corpus, in which language x and language y are mutually corresponding, also knows as bilingual corpus; monolingual corpus, for example, only for language y. The translation model $P(x|y)$ can be trained as a phrase table similar to a dictionary, followed by a probability to measure the likelihood of two phrases in source and target languages. In this way, the phrase table establishes a bridge between the two languages. Alignment in this process can be thought of as the mapping of words and words or phrases and phrases between two different languages. Alignment itself is very complicated as there are one to one, one to many, many to one and many to many. The monolingual corpus is used to train the language model $P(y)$. The language model measures whether a sentence is fluent and authentic in the target language y. The combination of these two models, together with other features, constitutes a statistical machine translation. The summary of SMT is as follows:

1. Requires a large number of feature engineering to construct the SMT system.
2. Takes a lot of time and effort for manual maintenance.
3. The main problem of SMT is that it is not only a complex system, it also needs to have multiple machine learning models (translation model ($P(x|y)$), language model ($P(y)$), etc.) to be established.

b) Neural machine translation

Due to the diversity and complexity of natural language, it is very challenging to translate one language into another. At present, under the conditions of large-scale corpus and modern computing power, neural machine translation (NMT) has showed great potential and has been developed into a new machine translation method. This method only needs bilingual parallel corpus, which is convenient for training large-scale translation models. NMT has a tremendous research value, which has become a hot spot in current machine translation research. From the academic view, neural machine translation is a method of using deep neural networks (DNN) to obtain mapping relationships between natural languages. In 2013, Kalchbrenner and Blunsom[53)] firstly proposed a new end-to-end encoder-decoder architecture for machine translation, which can be the birth of NMT. The nonlinear mapping of NMT is different from the linear SMT model, and the state vector between the encoder module and the decoder module is used to describe the equivalence relation of semantics.

The encoder-decoder architecture in NMT can be called a sequence-to-sequence (seq2seq) model, which typically involves two recurrent neural networks (RNNs).

The purpose of RNNs is to process sequence data[54)]. In the traditional neural networks, from the input layer to the hidden layer, and then to the output layer, the layers are fully connected and the nodes between each layer are disconnected. But this common neural networks is powerless for many problems. For example, if we want to predict what the next word of a sentence is, the previous word or several words are needed as the words in a sentence are not independent. RNNs are called recurrent neural networks, where the current output of a sequence is also related to the previous output. The specific form of expression is that the networks memorize the previous information and apply it to the calculation of the current output, namely the nodes between the hidden layers are no longer unconnected but connected, and the input of the hidden layer includes not only the output of the input layer but also includes the output of the hidden layer at the previous moment. In theory, RNNs can process sequence data of any length. However, in practice, in order to reduce complexity which is a little similar with n-gram language model, it is often assumed that the current state is only related to a few of previous states. Figure 2.4 shows a typical RNNs.

Firstly, look at the picture on the left. x is a vector that represents the value of the input layer (a series of words representation in NMT), and h is a vector that represents the value of the hidden layer. U is the weight matrix of the input layer to the hidden layer, and V is the weight matrix of the hidden layer to the output layer. y is also a vector, which represents the value of the output layer. Here, the weight matrix W is the weight of the value in the hidden layer last time, which is as well regarded as the input weight at current. Expanding the graph by the timeline, it equals the right side one. RNNs contain the input set which is labeled with $\{x_0, x_1, \cdots, x_{t-1}, x_t, x_{t+1}, \cdots\}$, the output set which is labeled with $\{y_0, y_1, \cdots, y_{t-1}, y_t, y_{t+1}, \cdots\}$, and hidden units set which is marketed

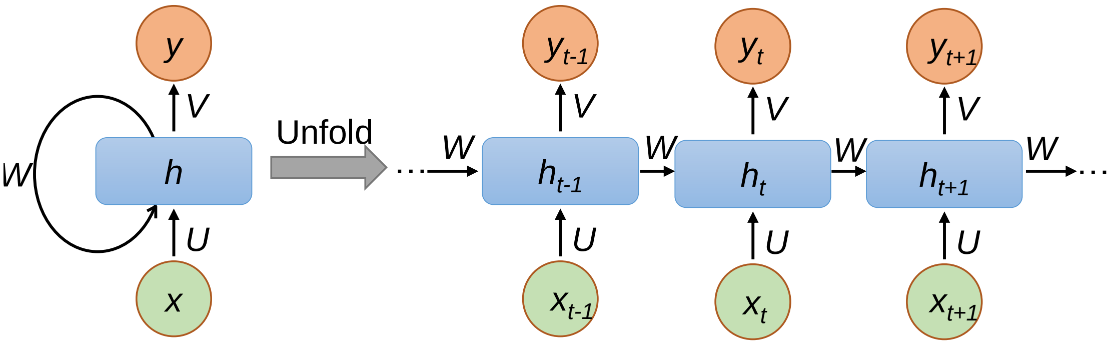

Fig. 2.4 Structure of a simple RNNs.

as $\{h_0, h_1, \cdots, h_{t-1}, h_t, h_{t+1}, \cdots\}$. These hidden units have done the most important work in RNNs. The following formulas represent the calculation method of the RNNs.

$$h_t = f(Ux_t + Wh_{t-1}) \tag{2.14}$$

$$y_t = softmax(Vh_t) \tag{2.15}$$

Where f is generally a nonlinear activation function, such as *tanh* or *ReLU*. When calculating h_0, the hidden layer state of the first word, h_1 is assumed to be needed, but it does not exist, therefore it is generally set to be 0 in the implementation.

The NMT of seq2seq is showed in Figure 2.5. It consists of two RNNs. The blue part on the left is called the encoder RNNs, which is responsible for encoding the source language (Encoding). The red part on the right is called the decoder RNNs, which is responsible for the target language (Decoding). The encoder RNNs learns the implicit features of the source language. The hidden state of the last neuron of the encoder RNNs is the initial implicit state of the decoder RNN. The decoder RNNs is called as a conditional language model because its initial implicit state is based on the output of the encoder RNNs. The decoder RNNs is used to generate the target language. The decoder RNNs continuously predicts the next word until the prediction output word ends tag with $\langle EOS \rangle$. The input to the encoder RNNs is the word embedding of the source language, and the input to the decoder RNN is the word embedding of the target language.

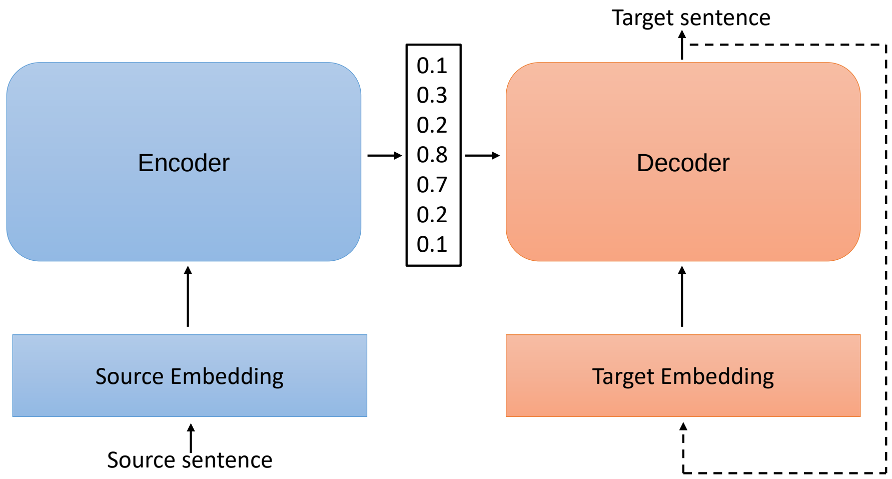

Fig. 2.5 Pipeline illustration of a encoder-decoder neural machine translation (NMT).

The training process of seq2seq is end-to-end, that is, the encoder RNN and the decoder RNN are trained as a whole, and there are not many sub-modules that are trained separately like SMT.

Advantages of NMT over SMT:

1. Better performance and better use of the context.
2. Simpler model which only requires a neural network with end-to-end training.
3. Less human maintenance and feature engineering. Regarding a different language, the network structure can remain the same, just change the training parallel bilingual corpus.

Insufficiency in NMT:

1. The neural network is hard to explain and difficult to debug.
2. Difficult to control with adding some artificial rules, which may cause some unexpected effects and safety issues.

(4) Modeling toolkit

a) Kaldi for automatic speech recognition

Like many great technology applications, there are many combinations of modules behind ASR. As introduced in the explanation of ASR, the acoustic model (AM) and language model (LM) are mostly based on the hidden Markov model (HMM) and *n*-gram model. Improvements to its implementation process (training the models) often save development costs to a certain extent and speed up the iteration of technology. Choosing the appropriate modeling toolkit in ASR would be a contribution to our dialogue-based CALL system. The emergence of Kaldi[55)] is based on the above motivation.

The goal of Kaldi is to have a modern and flexible code that is easy to understand, modify and extend.

The four open-source speech recognition toolkits listed here are all derived from academic research.

1. As the name suggests, CMU Sphinx is a product from Carnegie Mellon University. There are two versions of C and Java on the GitHub platform. Its research and development (R&D) history dates back approximately 20 years and is currently being updated on the GitHub[6].

2. Hidden Markov Toolkit (HTK) began at Cambridge University in 1989 and was once commercialized but is now back at Cambridge University. HTK is not a strictly open-source tool, and it is slow to update[7].

3. Kaldi is an open-source toolkit based on a 2011 workshop, which is written in C++ and licensed under the Apache License v2.0. The source code of Kaldi is hosted by GitHub[8]. There are currently about 7700 stars and 3400 forks. The code is continuously maintained and updated, which is a very popular algorithm toolkit in the field of speech recognition.

[6] https://github.com/cmusphinx

[7] https://github.com/ibillxia/htk_3_4_1

[8] https://github.com/kaldi-asr/kaldi

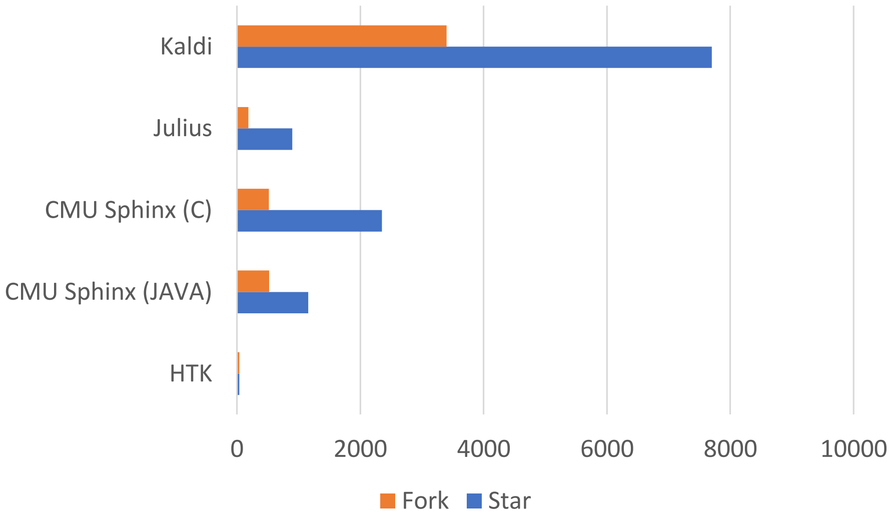

Fig. 2.6 Comparison of developer interest among different speech recognition toolkits.

4. Julius began in 1997 with high-performance, large vocabulary continuous speech recognition (LVCSR) decoder software for speech-related researchers and developers[9].

Figure 2.6 shows the developers' interest among the four ASR toolkits. Kaldi attracts a lot of attention from the developers. On the other hand, Kaldi users have more ways to interact, including mail, Google forums[10], and GitHub repository.

b) Keras: Deep Learning framework

Keras is a well-known high-level neural network API that can run on top of TensorFlow, CNTK, or Theano for researchers[11]. Keras was first released in March 2015 and was supported by its ease of use and succinct simplicity. Keras is a framework supported by Google. Keras is easy to learn and conduct experiments with standard layers, eliminating the need for developers to consider the complexity of deep learning.

[9] https://github.com/julius-speech/julius

[10] https://groups.google.com/forum/#!forum/kaldi-help

[11] https://keras.io/

Table 2.3 Assessment criteria for the sentences in ERJ database[1].

1 (Very poor)	Inaccurate, and apt to be misunderstood
2 (Poor)	Inaccurate, and considerable practice needed
3 (Fair)	Fair and common
4 (Good)	Accurate, but some practice needed
5 (Excellent)	Good and near-native speaker level

(5) Data sets

As data-driven approaches are applied in this research, here, we introduce the data sets used in our system. The data sets can be roughly divided into two categories: the speech data trained/evaluated in ASR and the text data used for training NMT.

a) ERJ database

In this dialogue-based CALL system, the automatic proficiency evaluation in the phonemic category of non-native speech would be conducted. Therefore, regarding the human score, the segmental score of sentence utterances pronunciation was selected in ERJ (English Read by Japanese) database[1],[56]. The ERJ database is widely used for many research studies on the development of the CALL systems[15],[57],[58], non-native speech recognition[59] and synthesis[60]. Every Japanese student was asked to be fully prepared with sufficient practice before recording. They were informed to read each given sentence as accurately as possible, and they had three chances to re-record or skip those sentences failed to be recorded during the main recording time. They can choose to challenge to re-record those failed sentences after the main recording time. English speech utterance samples spoken by 100 male and 102 female Japanese students are included in the database. Five phonetically trained American English teachers were recruited as rating experts. They were informed to listen ten sentences read by each Japanese student and give a 1-5 scale segmental score to every sentence. The 5-grade scales are showed in Table 2.3.

b) TIMIT database

As a native speech database, the TIMIT database[61] is widely used for research in the field of speech technology[62),63]. The TIMIT database collected by Texas Instruments includes 630 speakers and 2343 different sentences. Each of the 630 individuals from the eight major dialect regions in the United States spoke out the given ten sentences. All the sentences were manually segmented and marked at the phoneme level.

c) TMNED corpus

Another non-native speech database as an evaluation set in our system is Tohoku multi-modal non-native English dialogue (TMNED) corpus recorded in Tohoku university[2]. For constructing this corpus, Wu[2] let the learners take English conversation with the CALL system based on scenarios learned in advance. Since university students in Japan have few opportunities to use English for conversation and communication, to a certain extent, most of them don't have much confidence to speak English out directly. To construct the corpus with high quality for research use, the learners took lots of effort to remember the content of the conversations before video recording. The original conversation text and the human transcription from the learners' speech are nearly the same. Therefore, this corpus contains restricted English speech of two different dialogue topics spoken by 13 undergraduate students (9 males and 4 females). Each dialogue covers three sentences from the system and three responding sentences from the learner. 26 dialogues were collected from this corpus. The human evaluation step of the collected speech follows the design of the ERJ database's 1-5 scale segmental score: 3 American English language teachers were employed to annotate all dialogue data with assessment criteria presented in Table 2.4.

d) TEC corpus

We constructed a new speech database named as Tohoku English conversation (TEC) corpus for the spontaneous speech evaluation. 14 (12 males and 2 females) Tohoku University students participated in the recording of the TEC corpus. Each student was asked to participate in two different conversations. Each conversation is a three-minute free English conversation with every two students. To eliminate the tension of the students during the recording process, and to better understand the

Table 2.4 Assessment criteria of segmental score for the dialogues in Tohoku multi-modal non-native English dialogue (TMNED) corpus[2].

1 (Very poor)	Really non-native speech
2 (Poor)	Some non-native speech
3 (Fair)	Fair and common
4 (Good)	Native-alike prounciation
5 (Excellent)	Very native-alike pronunciation

dialogue, we asked participants to speak at a relatively slower pace, and in the pauses of context, try to reserve enough time for understanding, thinking and responding. The recorded conversations cover a variety of topics, such as food, movies or sports. We collected 195 different sentences spoken by those 14 students in total. 3 American English native speakers who have some teaching experience were recruited to annotate all dialogue data with 1-5 scale phonetic segmental score based on the ERJ database assessment criteria presented in Table 2.3.

e) Movie-Dialogs corpus

Regarding the development of NMT aiming to generate sentences with errors, since most of the components in input and output sentences are the same, we first imagine an initial NMT with the ability to output the same sentence as the input sentence. To train such an initial NMT, the Cornell Movie-Dialogs Corpus is selected [64], which is a large collection of English conversations extracted from the movie scripts. The raw data in this corpus is represented in dialogue turns labeled with movie metadata. As this initial NMT concentrates on generating identical sentences, we trimmed the data in this dialogue corpus by removing the sentences containing low-frequency words from the original 304713 utterances, which reproduces 201935 identical sentence pairs for the initial NMT.

f) NICT JLE corpus

The NICT Japanese Learner English (JLE) Corpus[65] mainly includes the transcribed text of the audio-recorded speech samples in an English oral proficiency interview test (Standard Speaking Test, SST) for Japanese students. As a non-native English spoken corpus, it contains the world's largest transcribed text data (no audio speech data)[12]. 167 files of the interview in this corpus were annotated with grammatical error tags. The error tag set includes 47 tags of lexical and grammatical errors of learners. In total, 8476 right-to-wrong sentence pairs were managed out for training the NMT.

g) TJTEOT corpus

In this research, for the grammatical detection in the speech-based CALL system, we developed an oral translation corpus called Tohoku Japanese to English oral translation (TJTEOT) corpus.

For the motivation of constructing this corpus, we assume every topic in the dialog-based CALL system to be a preset one. The learners study words, expressions and grammars from lessons as pre-exercise, which would make it easier for learners to speak out when talking with the system. Therefore, we call the reference correct sentence expected to be uttered by the learners "a target sentence". However, in a real session, as the language proficiency levels of the learners are different, not all of the utterances from the learners match the target sentences. we regard the sentence actually uttered by the learner as "an uttered sentence". An uttered sentence often contains grammatical errors. And we assume that this CALL system has the ability to recognize the speech uttered by the learners as "a recognized sentence", which have recognition errors in addition to the grammatical errors.

To develop the speech database for GED system's use, we prepared grammar exercise in the form of Japanese-to-English translation existed in the system, which has the advantage of only focusing on the grammatical use in English study. In this task, we designed the recording part as the flexible use of English expression to show the actual grammatical proficiency level of the learner. More or less grammatical errors would happen in the recording part as it is difficult to translate all the target sentences correctly.

The translation content defined for TJTEOT corpus follows the previous work[39)]. The procedure of recording the utterances was as follows:

1. Practice. We prepared the Japanese sentences and the corresponding target sentences. 14 target sentences were used as a material of grammar exercise. The learners were asked to read and practice these sentences. There was no time limitation for practice.

[12] https://alaginrc.nict.go.jp/nict_jle/index.html

2. Recording. After the learners thought they were ready, the recording began. The Japanese text was presented to them, and they were asked to speak out the English translation sentence by sentence.

We recorded the speech data from five students in Tohoku University, including 2 males and 3 females. By counting the different sentences that occurred in the recording set, we obtained 144 different uttered sentences from the 42 original target sentences.

(6) Summary

In order to establish a dialogue-based CALL system, there would be many sub-tasks, such as automatically recognizing the speech from the learners, evaluating the pronunciation proficiency of each learner and giving the appropriate feedback on the grammatical errors in their speech. Behind the above tasks, the technology in spoken language processing field is important and required.

This chapter has showed a variety of speech technology along with the previous work and recent deep learning. It should be noted that in order to maximize the application of these technologies, sufficient data is needed. After all, most of the approaches are data-driven. Therefore, we also descried the databases utilized in the system, including some famous databases established in the previous years and the corpus recorded in Tohoku University.

After reading this chapter, we believe that the audiences will have a fundamental understanding of the entire dialogue-based CALL system. Next, the following two chapters will enter the design and implementation of the real CALL system, as well as the experimental verification of this system.

3 . Automatic pronunciation evaluation on English speech by Japanese learners based on acoustic models

At the first stage of building the dialogue-based CALL system, it is necessary to recognize the utterances of the learners with high accuracy and evaluate the language pronunciation proficiency of the specific speakers with appropriate methods. In this chapter, we discuss the automatic assessment of the second language (L2) for non-native speakers.

(1) Previous metrics

a) Traditional automatic pronunciation evaluation system

The automatic proficiency assessment systems for L2 learners include a variety of tasks ranging from restricted speech to spontaneous speech[66),67),12)]. Figure 3.1 describes one of the conventional proficiency evaluation systems for reading aloud and repeating tasks[10)]. At first, the learner's speech is processed with feature extraction, and sent to the speech recognizer for decoding. Then, the output of the recognizer, reference sentence and extracted features are calculated as several ASR-based scores, including the rate of speech (ROS), GOP and accuracy. Finally, these scores are sent to the regression model which produces the prediction score for the speech response. SpeechRater is a typical automatic proficiency assessment system for scoring non-native English speaker's spontaneous speech[12)], which is showed in Figure 3.2. In this scoring process, the input speech is decoded into word sequence with the non-native acoustic model and forced-aligned to the native acoustic model for calculating the log-likelihood and the durations of phonemes. The linear regression model is used to predict proficiency score for the utterance of non-native speakers. The conventional systems provide

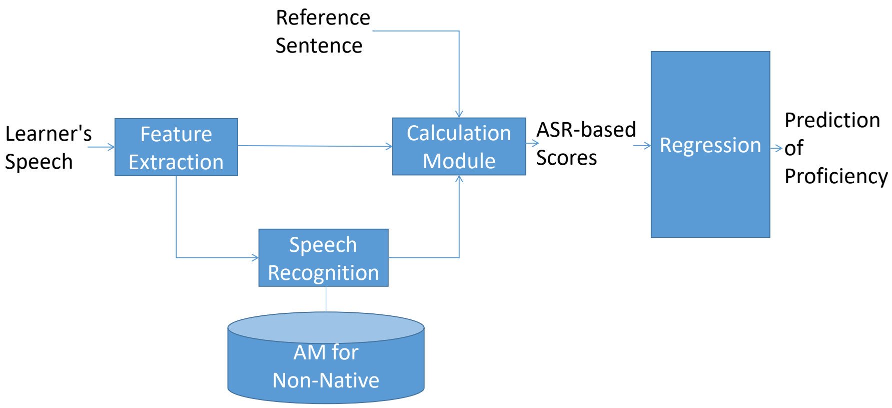

Fig. 3.1 Automatic proficiency evaluation systems for reading aloud and repeating tasks. Here, the reference sentence is required during the calculation step for ASR-based scores.

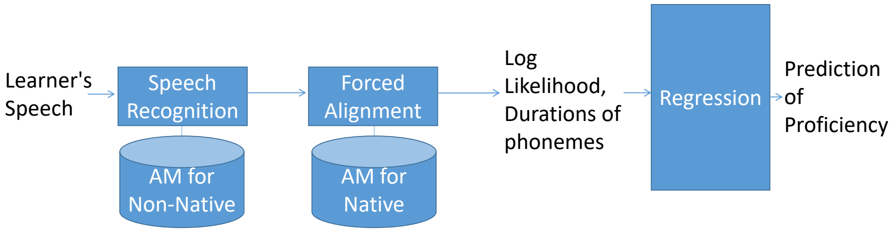

Fig. 3.2 Schematic diagram of a standard automatic scoring system for spontaneous speech. The input speech is decoded into word sequence with the non-native acoustic model and forced-aligned to the native acoustic model for calculating the log-likelihood and the durations of phonemes. The linear regression model is used to predict proficiency score.

log-likelihood scores or other ASR-based scores as machine scores in regression-based proficiency prediction.

Log-likelihood scores are widely used for automatic pronunciation scoring as a traditional machine score[35),68),14)]. WER could be a machine score for assessing proficiency based on its high relevance with the function of hearing and distinguishing in the ASR system. Tao[69)] systematically demonstrated that WER obtained from the ASR module in an automatic assessment system played an important role in scoring. Moreover, the WER is the relatively straightforward manifestation and easily obtained among the output results from the ASR system. One disadvantage of the WER is that it requires the prepared reference transcription of the utterance. It is not always possible to prepare the reference sentences, especially in a dialogue-based CALL system.

b) HMM-based phone log-likelihood scores

HMM-based acoustic models, trained from the speech database of native speakers, can be used to generate phonetic time alignments of the non-native speaker's speech during the decoding step by the Viterbi algorithm. For each phone segment in each sentence, the normalized log-likelihood score $\hat{l}_i$[35] is defined as

$$\hat{l}_i = \frac{1}{d_i} \sum_{t=t_0}^{t_0+d_i-1} \log p(y_t|q_i) \tag{3.1}$$

where $p(y_t|q_i)$ is the likelihood of the current observation vector y_t to the i-th phone q_i, d_i is the phone's duration in frames, and t_0 is the staring frame index of the phone segment.

The phone-based log-likelihood score in one sentence L[35] is defined as

$$L = \frac{1}{N} \sum_{i=1}^{N} \hat{l}_i \tag{3.2}$$

where the whole phone's likelihood score is summed in the sentence over the number of phones N to obtain the average phone-based log-likelihood score.

c) Word error rate

To evaluate the accuracy of an ASR system, the word error rate (WER) is the standard measurement. Generally, we obtain the WER in the case of knowing the text spoken by the language learner as the reference transcription. The output word sequence from the ASR system is forced to align with the reference text. Assume we have I utterances in the database. $T_i^{(\mathrm{ref})}$ and $T_i^{(\mathrm{asr})}$ are the corresponding reference sentences and recognized sentences to the i-th utterance, respectively. According to the Levenshtein distance[70], the WER is calculated as

$$WER = D(\mathbf{T}^{(\mathrm{ref})}, \mathbf{T}^{(\mathrm{asr})}) = \frac{\sum_{i=1}^{I} d_L(T_i^{(\mathrm{ref})}, T_i^{(\mathrm{asr})})}{\sum_{i=1}^{I} |T_i^{(\mathrm{ref})}|} \tag{3.3}$$

where $\mathbf{T}^{(\mathrm{ref})} = (T_1^{(\mathrm{ref})}, ..., T_I^{(\mathrm{ref})})$ and $\mathbf{T}^{(\mathrm{asr})} = (T_1^{(\mathrm{asr})}, ..., T_I^{(\mathrm{asr})})$ are the reference sentence-set and recognized sentence-set, respectively. $D(\mathbf{X}, \mathbf{Y})$ is the WER between the sentence-sets $\mathbf{X}$ and $\mathbf{Y}$, $|T_i^{(\mathrm{ref})}|$

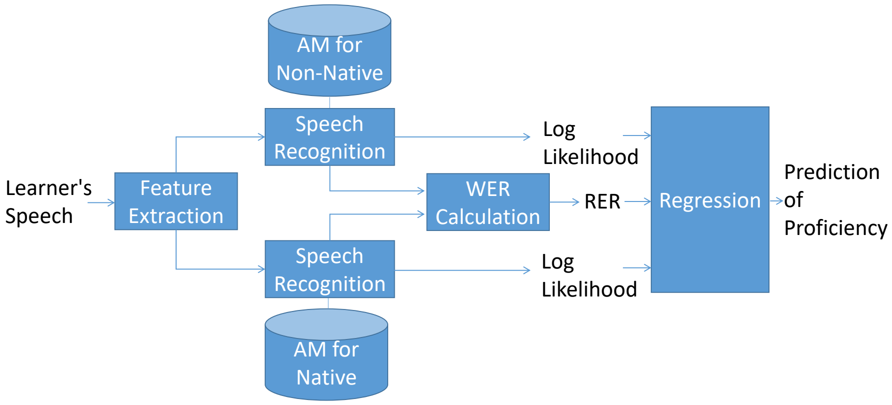

Fig. 3.3 Overview of the proposed proficiency evaluation system. We propose the RER as a new machine score in the regression-based prediction of proficiency, which based on the two assumed conditions: a) high recognition accuracy from non-native ASR and b) the effectiveness of evaluating language proficiency with WER from native ASR.

is the number of words in the i-th sentence, and $d_L(X,Y)$ is the Levenshtein distance between sentences X and Y.

(2) A novel pronunciation evaluation method: RER

a) Our proposed pronunciation evaluation system

To recognize and assess the non-native speech in a dialogue-based CALL system, we propose to utilize two acoustic models as shown in Figure 3.3: one is trained with a non-native speech database and the other one with a native database. Here, we expect that the non-native ASR system recognizes non-native utterances with high accuracy regardless of the speaker's proficiency; besides, the native ASR system should give less accurate results for utterances with low proficiency. If this is true, we can distinguish the proficiency of the non-native speaker by comparing the two different recognition results from non-native and native ASR systems with the input speech. Based on this idea, we propose a new method for automatic assessment: reference-free error rate (RER).

Table 3.1 The components calculated in WER and RER.

	WER	**RER**
ASR output	Acoustic model output	Native AM output
Reference transcription	Human transcription	Non-native AM output

b) Reference-free error rate

The proposed reference-free error rate (RER) is calculated by comparing the recognized text obtained from native ASR to that obtained from non-native ASR based on the Levenshtein distance. If we have I utterances of a speaker, let $\mathbf{T}^{(\mathrm{non})} = (T_1^{(\mathrm{non})}, ..., T_I^{(\mathrm{non})})$ and $\mathbf{T}^{(\mathrm{nat})} = (T_1^{(\mathrm{nat})}, ..., T_I^{(\mathrm{nat})})$ be the recognized sentences by the non-native and native ASR systems, respectively. Then the RER is calculated as

$$RER = D(\mathbf{T}^{(\mathrm{non})}, \mathbf{T}^{(\mathrm{nat})}). \tag{3.4}$$

If the RER is high, it means that the two recognition results differ much, which happens when the proficiency of the utterance is low.

Table 3.1 shows the components utilized in WER and RER. From this comparison, it is obvious that the results from non-native acoustic model are regarded as “human transcription”.

(3) Experimental set-up

a) Non-native ASR system based on the ERJ database

Data preparation in the ERJ database

All the sentences in the ERJ database were divided into 8 groups (S1 to S8) and the required amount of the recording in each group is about 120 sentences. Therefore, sentences in one group were read by nearly 12 male speakers and 13 female speakers, respectively. Table 3.2 shows the details of the sentence subsets in the ERJ database.

In order to build this non-native ASR system, only the ERJ database was used to train the acoustic models in this part. At first, all the sentences in the ERJ database were selected out and analyzed by

Table 3.2 Sentence subsets in ERJ database[1]. Totally contains eight groups (S1 to S8) and the required amount of the recording in each group is about 120 sentences.

	S1	S2	S3	S4	S5	S6	S7	S8
Number of sentences	120	123	122	123	124	123	123	122
Number of words	742	838	823	823	814	899	882	879
Speakers (male)	13	12	11	11	13	14	12	14
Speakers (female)	13	13	11	12	13	13	13	14

considering the frequency of phones. We used the default phonemic symbols in the CMU pronouncing dictionary. After the repeated sentences were selected out from each sentence subset in ERJ database, each of 39 phones was counted in every modified sentence subset. Then, we checked the number of each phone in these revised sentence subsets and ensured that all the 39 phones appeared in training, development and test sets. Finally, the sets S1, S2, S4, S5, S6, and S7 were determined as the training set, S3 as the development set and S8 as the test set. Furthermore, regarding the WER of every student in ERJ database, we applied leave-one-out cross-validation (8 folds) to redefine the training set and test set for decoding based on the initial non-native ASR system. For the other two tasks of evaluating the constrained conversation speech in TMNED corpus and the free conversation speech in TEC corpus, the non-native acoustic models were trained with the whole speech of sentences in ERJ database.

Lexicon and language model

We modified the CMU pronouncing dictionary as the baseline pronunciation lexicon by changing all the stressed phonemes to non-stressed phonemes. Table 3.3 shows examples of this change in CMU pronouncing dictionary. One reason for this modification is that, in our CALL system, phonemic features are intended to assess non-native English speech for Japanese learners. The other reason is that Japanese speakers' English pronunciation is generally different from native speakers. Therefore, if we have more data for a specific phoneme modeling, it could improve the performance of the acoustic model.

Regarding the language models in this non-native ASR system for the proficiency assessment

Table 3.3 Examples for changing stressed phonemes to non-stressed phonemes in CMU pronunciation dictionary.

Word	Original stressed phonemes	Non-stressed phonemes
English	IH1 NG G L IH0 SH	IH NG G L IH SH
Speech	S P IY1 CH	S P IY CH
Assess	AH0 S EH1 S	AH S EH S

of students in ERJ database, we selected all the sentences subsets of ERJ database and removed all repeated sentences. We developed bigram and trigram language models from these 980 different sentences. These language models are database-closed on the decision that this system is able to figure out any sentence in ERJ database as much as possible. However, the design of this language model is exclusive for ERJ database while leading to poor performance for the unknown text out of this database. Therefore, in the tasks of evaluating TMNED and TEC, we established an open trigram language model for the dialogue-based speech proficiency assessment. The open language model was trained with the content of ERJ sentences, TIMIT sentences and a variety of American English conversations from the United States Department of State[1] and the book of Conversational American English[71)]. We searched the language model scale among the values between 10 and 50 with a beam width of 13 to obtain the WER.

b) Native ASR system based on the TIMIT database

Data preparation in the TIMIT database

We converted all the transcriptions in TIMIT from phoneme level to word level, and meanwhile, set phonemic symbols in data preparation by using the non-stressed phonemic symbols in the CMU pronouncing dictionary. In contrast to the original separate data set as training, development and test, all of the utterances in TIMIT were used for training acoustic models in the native ASR system.

[1] https://americanenglish.state.gov

Language model

The TIMIT based native ASR system is the English proficiency assessment component in the CALL system. Considering the language model in this system, we decided the content including 980 sentences from ERJ and 2343 different sentences from TIMIT and established a trigram language model for evaluating each student's proficiency in ERJ database. The open language model in non-native ASR system was used for evaluation on TMNED and TEC.

(4) Experiments and results

a) Non-native speech recognition within ERJ database

HMM-based monophone and triphone acoustic models were primarily trained as the basis of DNN-based acoustic models. The details of settings in four different DNN training methods are listed in Table 3.4. In which, NN11 stands for the NNET1 method and the other three NNs stand for the NNET2 methods in Kaldi toolkit. Here, p-norm is the non-linearity $y = ||x||_p$, where the vector x represents a small group of inputs[72]. There is no hidden layer dimension in the p-norm networks, instead, there are two parameters: p-norm input dimension/p-norm output dimension. The p-norm E.L. stands for the ensemble p-norm non-linearity, where the ensemble size is 4. We regarded NNET1 method as a trial, experiment in NNET2 methods as the main focus. Four hidden-layer neural networks is enough in most of the databases[73]. MFCC feature without dimension reduction was applied to NNET2, which was slightly different with FBANK feature in NNET1[74],[75]. As this paper concentrates on the emphasis of acoustic models, we present all the results only by using the same trigram language model trained from ERJ sentences.

The recognition results within ERJ database are presented in Table 3.5. Compared with the results of traditional GMM-HMM acoustic models, DNN-based acoustic models significantly improve the accuracy of recognition. In addition, the acoustic model with non-stressed phonemes performed better than that with stressed phonemes, which was mentioned in Section 3.3. Six kinds of acoustic models are used for decoding. The acoustic model with non-stressed phonemes performs better than that with stressed phonemes. The third method with non-stressed phonemes in NNET2 series models has the

Table 3.4 Settings in DNN training. Here, p-norm is the non-linearity $y = ||x||_p$, where the vector x represents a small group of inputs.

	NN11	NN21	NN22	NN23
Model Type	NNET1	NNET2	NNET2	NNET2
Input Feature	40FBANK	40MFCCs	40MFCCs	40MFCCs
Hidden Type	sigmoid	tanh	p-norm	p-norm E.L.
Input Nodes	440	360	360	360
Hidden Layer	4	4	4	4
Hidden Dim	1024	1024	1000/200	1000/200
Output Nodes	3168	1551	1551	1551

Table 3.5 Experimental results in the development set and test set. Six kinds of acoustic models are used for decoding. The acoustic model with non-stressed phonemes performs better than that with stressed phonemes.

	Monophone	Triphone	NN11	NN21	NN22	NN23
(non-stressed)						
DEV-WER[%]	7.09	7.83	2.99	4.99	3.13	2.81
TEST-WER[%]	6.90	5.11	2.05	3.42	2.31	1.97
(stressed)						
DEV-WER[%]	8.39	8.77	*	5.38	3.39	2.93
TEST-WER[%]	7.59	5.34	*	3.55	2.29	2.08

lowest WER in the test set as 1.97%, which shows a proper potential to use the output recognized sentences as reference transcriptions.

b) Conditions in automatic pronunciation assessment

The goal in this part is to develop a method to predict the English proficiency of a Japanese native speaker's English utterance without reference transcription. To this end, for the human scoring on the evaluation data set, we first investigated the inter-rater agreement calculated by the Pearson correlation coefficient (r). In ERJ database, one of the raters scored only male students when the other four (R1,R2,R3 and R4) scored both males and females. Therefore, we calculated the human-human agreement among the four full scoring teachers. Figure 3.4 shows the inter-rater correlation matrix in ERJ database. The distribution of each human-related score is displayed on the diagonal line. The two variables (one human score with the other) scatter plots are displayed with a regression line at the lower triangle, while the correlation coefficient with the level of significance as stars on the upper triangle. Each level of significance is associated with a symbol: p value (0: "$* * *$", 0.001: "$**$", 0.01: "$*$", 0.05: ".", > 0.05: " "). The average correlation r among the raters is around 0.8 (N = 190). The total number of Japanese students scored by the teachers is 190 (95 males and 95 females). Additionally, we calculated the average segmental score from all of the 5 raters as the proficiency score for every Japanese student.

Figure 3.5 shows the histogram of proficiency scores in the ERJ database. The average value of proficiency is 2.97, and the standard deviation is 0.55. In order to explicit the relationship between the proficiency and WER obtained from native ASR system, we divided all speakers of ERJ database into three classes: LOW, MID and HIGH, where the speaker having a lower score than 2.5 were classified into LOW, those having higher score than 3.2 were HIGH, and the others were MID. The ensemble p-norm activation function in NNET2 was only selected in DNN based acoustic model, which has the same parameter with NN23 showed in a). Figure 3.6 shows the relationship between the WER from our established native ASR system and proficiency score for both GMM-HMM and DNN-HMM. These results indicate that with the proficiency score getting higher, WER of the corresponding student becomes lower.

Next, to investigate the correlation coefficients between the proficiency score and other ASR-based machine scores, we prepared various combinations of conditions. Table 3.6 shows the symbols to describe the condition of machine scores. A machine score is described as a combination of the training data, score type, and the acoustic model. For example, the log-likelihood of GMM-HMM triphone trained from the ERJ database is denoted as "E_LL_TRI". As the calculation of RER uses acoustic models trained from both TIMIT and ERJ, in which output text from the DNN-based acoustic model in native ASR is regarded as the reference, RER-based scores are described without the training

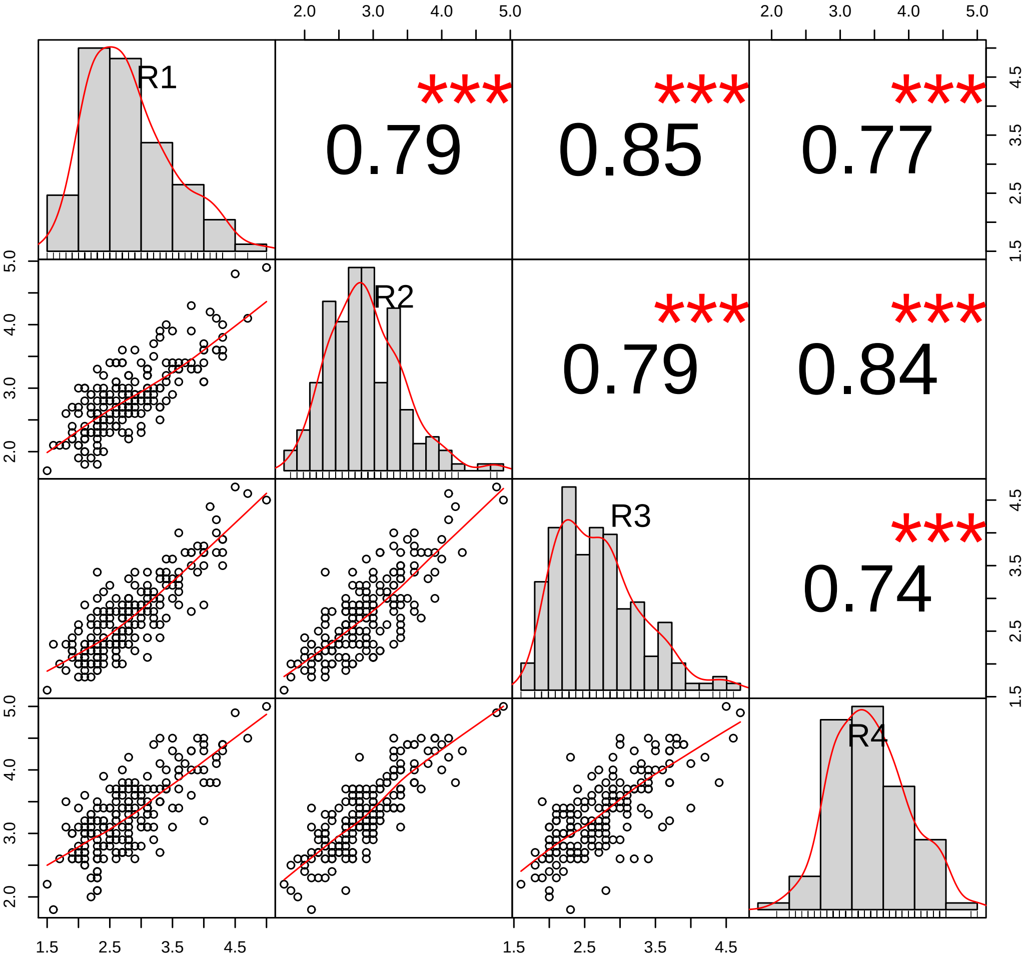

Fig. 3.4 The scatter plot matrix of inter-rater correlation in ERJ database. The distribution of each human-related score is displayed on the diagonal line. The two variables (one human score with the other) scatter plots are displayed with a regression line at the lower triangle, while the correlation coefficient with the level of significance as stars on the upper triangle. Each level of significance is associated with a symbol: p value (0: "$* * *$", 0.001: "$**$", 0.01: "$*$", 0.05: ".", > 0.05: " "). N = 190.

data part, such as "RER_MONO" or "RER_NN".

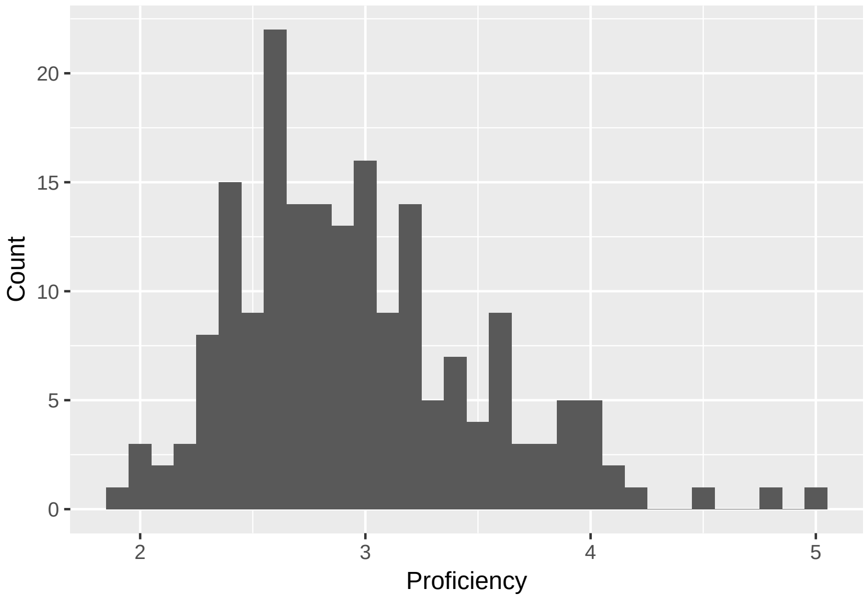

Fig. 3.5 Histogram of proficiency scores in ERJ database. The average value of proficiency is 2.97, and the standard deviation is 0.55. N = 190.

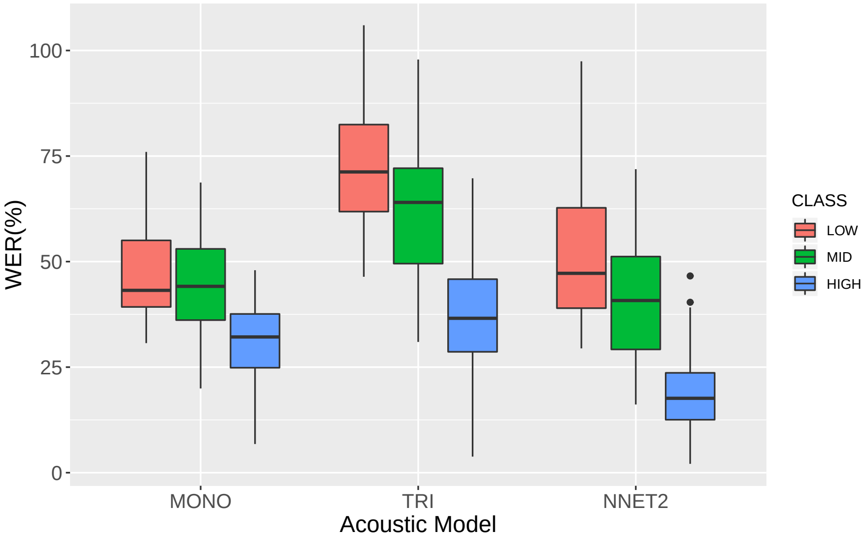

Fig. 3.6 Box-plots of WER from three acoustic models in the native ASR system, dependent to the proficiency score of the speakers. N = 190.

Table 3.6 Components of machine score combination. A machine score is described as a combination of the training data, score type, and the acoustic model.

Training	Score Type	Acoustic Model
TIMIT (T_)	Log-likelihood (LL_)	Monophone (MONO)
ERJ (E_)	WER (WER_)	Triphone (TRI)
	RER (RER_)	NNET2 (NN)

c) The relationship between the proficiency and ASR-based machine scores

Figure 3.7 shows the absolute correlation for all combination of machine scores and the proficiency score. This figure is a matrix of absolute correlation coefficients, and the vertical and horizontal order of scores are the same. We also show the dendrogram obtained by the hierarchical clustering conducted based on the correlation. In this figure, white color means high absolute correlation and black color means low correlation. From this figure, it is obvious that all scores are classified into three groups: the first one includes WER of ERJ-based acoustic models (E_WER_MONO, E_WER_TRI, E_WER_NN) and the log-likelihood of ERJ-based NN acoustic model (E_LL_NN). The second one includes log-likelihood of GMM-HMM trained from both TIMIT and ERJ (T_LL_TRI, T_LL_MONO, E_LL_TRI, E_LL_MONO). The third group includes all other scores and the proficiency score.

As presented in Figure 3.7, it is obvious that the log-likelihood scores of GMM-HMM have a strong correlation with each other regardless of its training data. However, the log-likelihood of DNN-HMM shows a different tendency. The log-likelihood of E_LL_NN and T_LL_NN have a very low correlation ($r = -0.041, p = n.s$), which suggests that the DNN-HMM trained from different training data captures the distribution of speech differently.

Figure 3.8 shows the absolute correlation between the proficiency score and all machine scores. T_LL_NN has the highest correlation ($r = 0.736$, $p < 0.001$), and T_WER_NN, T_WER_TRI, RER_TRI, RER_NN have almost the same values (0.721 to 0.726). The difference between T_WER_NN and RER_NN is really small, which means the reference transcriptions are nearly the same with the output recognized text from the non-native ASR by using DNN-based acoustic model. Note that WER or RER has a negative correlation to the proficiency score because low proficiency utterance leads to

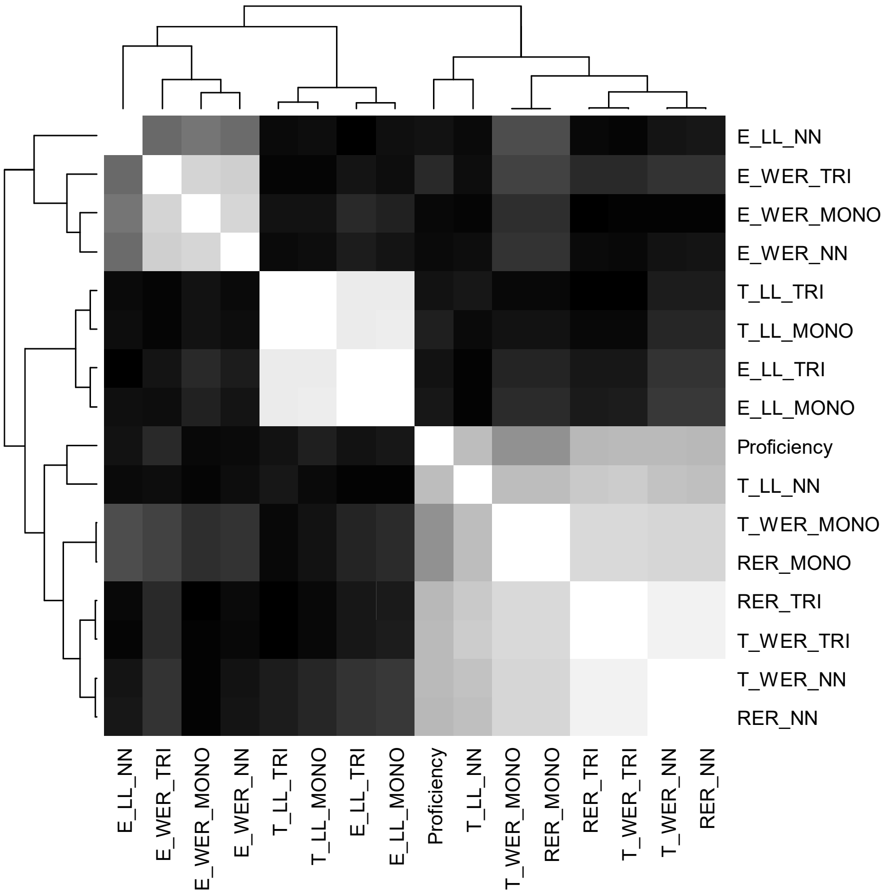

Fig. 3.7 Absolute correlation between all ASR-based machine scores and the pronunciation proficiency score. white color means high absolute correlation and black color means low correlation.

larger WER.

Next, we analyzed the effect of combining multiple scores. Considering the purpose that aims to develop a method to predict a learner ’ s proficiency without a reference transcription, we excluded the WER-based scores. As a result, we used nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN). Then we examined all combination of these scores (511 combinations in total). Linear regression was used for combining those scores. Two-fold cross-validation was used for evaluation, where the data was split into even-

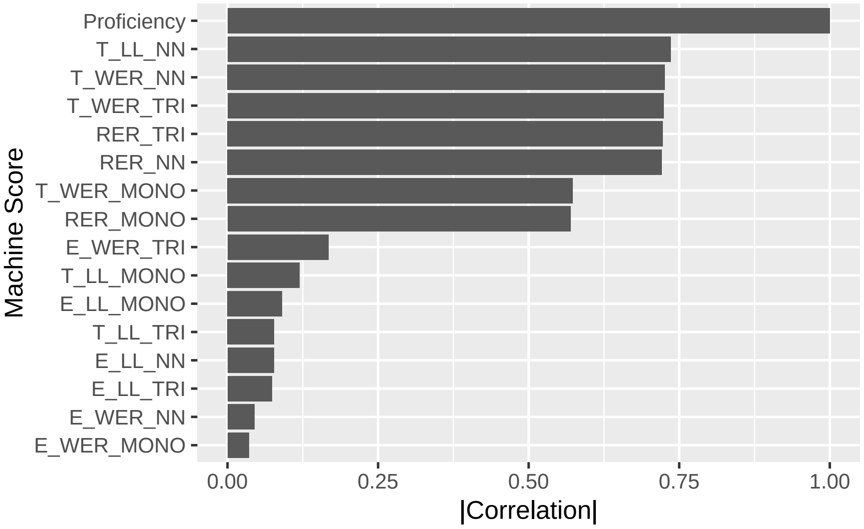

Fig. 3.8 Absolute correlation between the proficiency score and machine scores on ERJ database.

numbered and odd-numbered samples, and then the regression coefficients were calculated from one sample set, and the correlation coefficient was calculated using the other set. Table 3.7 shows the result. To summarize the result, we only show the result with the highest correlation when one to nine machine scores were combined.

We conducted the test of the difference between two dependent correlations with one variable in common[76)], this test was calculated based on the two-fold cross-validation basis. The summary of this test is presented in Table 3.8. Here, we selected three machine scores for comparison. This test shows there are significant differences among them. As shown in Figure 3.8, T_LL_NN has the highest correlation as a single score. The correlation coefficient exceeds 0.8 when three scores were combined (T_LL_NN, RER_MONO, and RER_NN). The highest correlation was obtained when all machine scores were used, where the correlation coefficient was 0.826, $p < 0.001$. When we did not use RER-based scores, the best combination was T_LL_MONO + T_LL_TRI + T_LL_NN + E_LL_TRI, where the correlation coefficient was 0.786, $p < 0.001$. These results show the importance of RER for estimating proficiency. Figure 3.9 shows a scatter plot of the proficiency score and the prediction score when all scores are used. This result shows that the prediction score has a good

Table 3.7 Correlation coefficients by combining multiple machine scores. This table only shows the nine high correlation when one to nine scores were combined.

T_LL			E_LL			RER			Correlation
MONO	TRI	NN	MONO	TRI	NN	MONO	TRI	NN	
		✓							0.736, $p < 0.001$
		✓						✓	0.778, $p < 0.001$
		✓				✓		✓	0.791, $p < 0.001$
✓		✓		✓			✓		0.808, $p < 0.001$
✓	✓	✓	✓					✓	0.815, $p < 0.001$
✓	✓	✓	✓			✓		✓	0.819, $p < 0.001$
✓	✓	✓	✓		✓	✓	✓		0.820, $p < 0.001$
✓	✓	✓	✓		✓	✓	✓	✓	0.823, $p < 0.001$
✓	✓	✓	✓	✓	✓	✓	✓	✓	0.826, $p < 0.001$

Table 3.8 Summary of the test of the difference between two dependent correlations with one variable in common. S1 is T_LL_NN, S2 is the combination of four scores (T_LL_MONO, T_LL_TRI, T_LL_NN and E_LL_TRI) and S3 is the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). P stands for the proficiency score. $N = 190$.

Scores	Correlation	Test for difference in two correlations
S1	r(S1-P) = 0.73	r(S1-P) *vs*. r(S2-P): $z = -1.96$, $p = 0.049$
S2	r(S2-P) = 0.78	r(S2-P) *vs*. r(S3-P): $z = -2.93$, $p = 0.003$
S3	r(S3-P) = 0.82	r(S3-P) *vs*. r(S1-P): $z = -2.03$, $p = 0.042$

correlation with the proficiency score. The standard error was 0.31.

d) Automatic assessment on TMNED corpus

In order to examine the effectiveness of our proposed method on real constrained dialogue task, we conducted the experiment on the evaluation of TMNED. By using the same way in ERJ database,

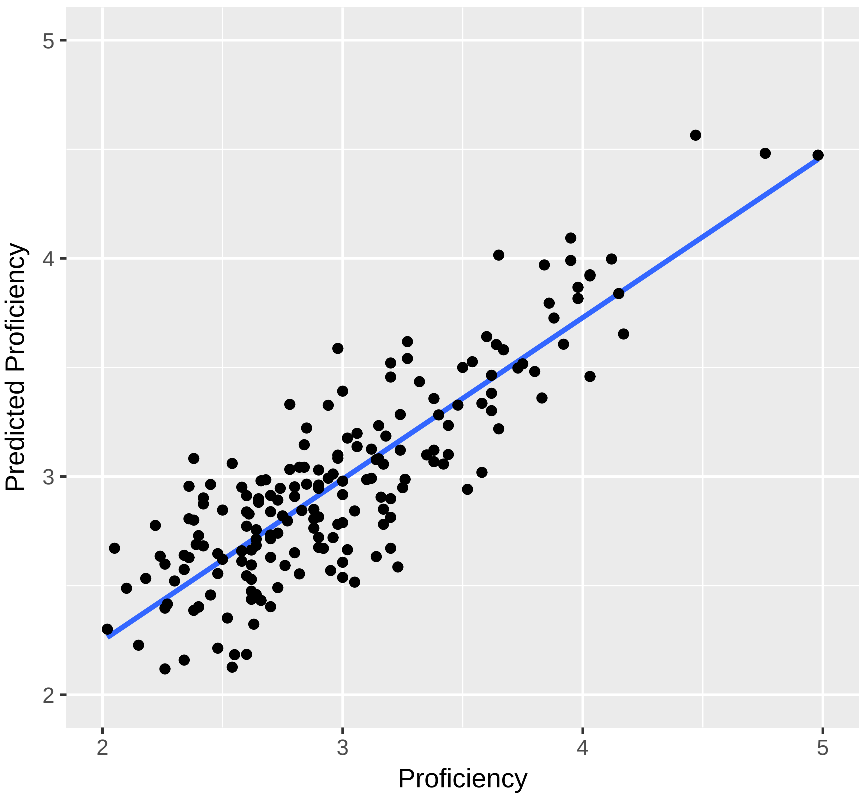

Fig. 3.9 Scatter plot of the proficiency and predicted proficiency on ERJ database. N = 190.

the average segmental score among the three human raters (H1, H2 and H3) in TMNED was calculated as proficiency score. The difference on evaluation from ERJ database is that the test unit is each of the dialogue here, while ERJ database is each of the students. The design of the test unit is based on the reason that TMNED is rather small in terms of only 13 students, and unlike reading tasks, the performance of the restricted conversation speech is always deviated by the different scenes. Figure 3.10 shows the histogram of proficiency in TMNED. The average value of proficiency is 2.82, and the standard deviation is 0.68. Table 3.9 summarize the inter-rater correlation on TMNED. The correlation among the three raters is around 0.5, which means low agreement among them. Proficiency is the average segmental score among three human raters. We also calculated other combination scores for analysis. For example, H1H2 means the average score from human rater 1 and human rater 2, the meaning is the same with H2H3 and H1H3. The correlation coefficients (r(H1-H2H3), r(H2-H1H3)

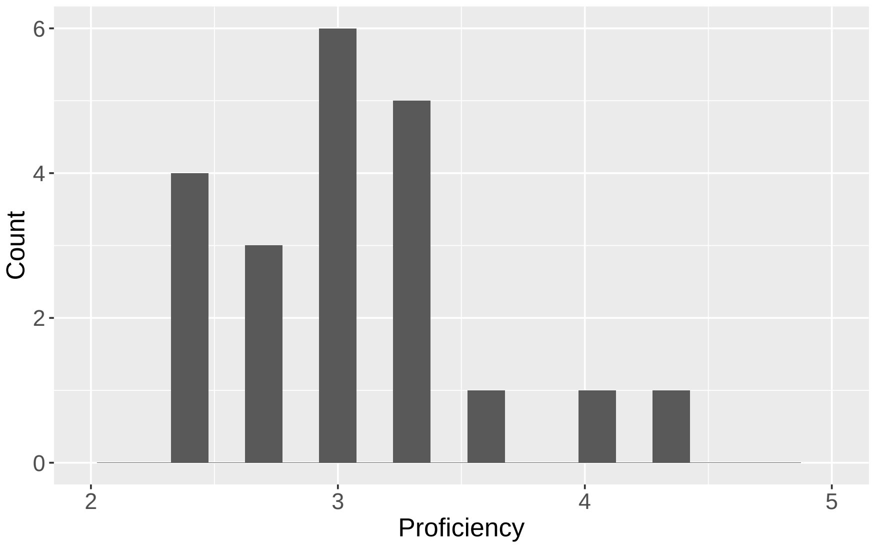

Fig. 3.10 Histogram of proficiency scores in TMNED corpus. The average value of proficiency is 2.82, and the standard deviation is 0.68. $N = 26$.

Table 3.9 Inter-rater correlation on TMNED corpus. Proficiency is the average segmental score among three human raters. H1H2 means the average score from human rater 1 and human rater 2. H2H3 means the average score from human rater 2 and human rater 3. H1H3 means the average score from human rater 1 and human rater 3. $N = 26$.

	H1	H2	H3	H1H2	H2H3	H1H3
H1	*	0.50, $p = 0.009$	0.47, $p = 0.016$	0.81, $p < 0.001$	0.56, $p = 0.003$	0.87, $p < 0.001$
H2	0.50, $p = 0.009$	*	0.49, $p = 0.010$	0.91, $p < 0.001$	0.92, $p < 0.001$	0.58, $p = 0.002$
H3	0.47, $p = 0.016$	0.49, $p = 0.010$	*	0.55, $p = 0.003$	0.80, $p < 0.001$	0.84, $p < 0.001$

and *r*(H3-H1H2)) range from 0.55 to 0.58, which is higher than that of H-H. The correlation between the two-rater average score to the individual rater score (calculated in the two-rater average score) has no meaning. Each human rater has a high agreement with proficiency(0.78, $p < 0.001$; 0.87, $p < 0.001$; 0.77, $p < 0.001$).

The speech from the student in each of the 26 dialogues was recognized by our non-native and native ASR systems. Since each speech in TMNED is essentially produced by memory, the human

Table 3.10 Average value of WER and RER from the ASR systems by both GMM-HMM and DNN-HMM on TMNED.

	MONO	TRI	NN
WER1[%](Non-native ASR)	25.03	16.18	16.72
WER2[%](Non-native ASR)	26.60	17.45	17.72
WER1[%](native ASR)	53.70	52.20	50.96
WER2[%](native ASR)	54.82	53.34	52.88
RER[%]	54.94	55.56	55.39

transcribed text would have more or less different places from the original reference text. Therefore, we calculated WER1 based on the comparison between native ASR output and the human transcribed reference text, and WER2 based on the comparison between native ASR output and the prompt reference text. Table 3.10 shows the average value of WER and RER from the output of these two ASR systems. Compared with the results of WER2, there is almost 1% improvement in WER1. Figure 3.11 shows the absolute correlation between the proficiency and ASR-based machine scores for the evaluation on TMNED. Unlike the correlation results in ERJ database, this time the WER and RER scores have more agreement with proficiency than log-likelihood scores. It is positive to see that RER_TRI has the highest correlation ($r = 0.692, p < 0.001$) among all the machine scores. E_LL_NN has the highest value ($r = 0.363, p = 0.067$) among log-likelihood scores. Finally, we used nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN) for combination as the prediction score, excluding the WER-based scores. Figure 3.12 shows a scatter plot of the proficiency score and the prediction score for the evaluation on TMNED. The correlation between the prediction score and the proficiency score is $0.803, p < 0.001, N = 26$. We checked the test for the difference between human-human correlation and machine-human correlation. Three different tests are set as presented in Table 3.11. H1 is human rater 1. H2 is human rater 12. H3 is human rater 3. H1H2 means the average score from H1 and H2. H2H3 means the average score from H2 and H3. H1H3 means the average score from H1 and H3. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO,

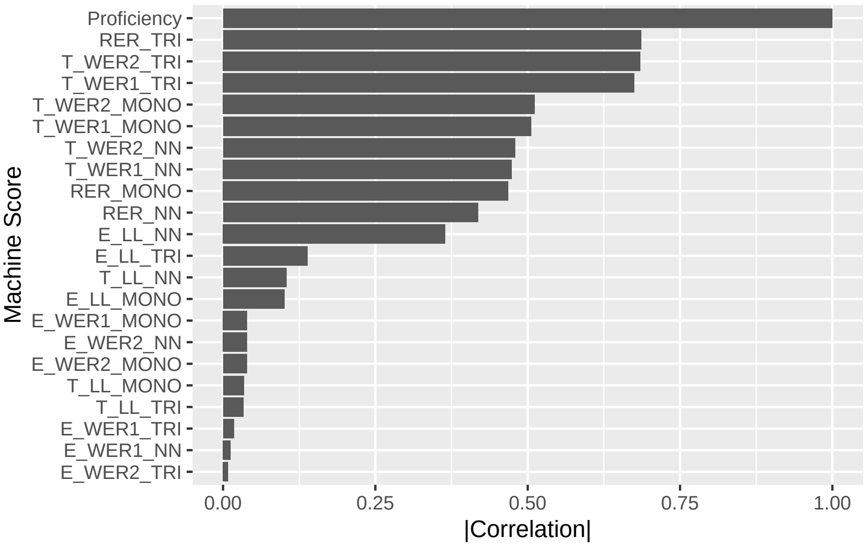

Fig. 3.11 Absolute correlation between the proficiency score and machine scores on TMNED corpus.

Table 3.11 Test for the difference between human-human correlation vs. machine-human correlation on TMNED corpus. H1 is human rater 1. H2 is human rater 12. H3 is human rater 3. H1H2 means the average score from H1 and H2. H2H3 means the average score from H2 and H3. H1H3 means the average score from H1 and H3. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). r(PS-H2H3) = 0.786, $p < 0.001$. r(PS-H1H3) = 0.798, $p < 0.001$. r(PS-H1H2) = 0.735, $p < 0.001$. $N = 26$.

Test for difference between r(human-human) and r(machine-human)
r(H1-H2H3) *vs.* r(PS-H2H3): $z = -1.89, p = 0.063$
r(H2-H1H3) *vs.* r(PS-H1H3): $z = -1.85, p = 0.064$
r(H3-H1H2) *vs.* r(PS-H1H2): $z = -1.67, p = 0.094$

T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). The correlation between the prediction score and average score from two raters ranges from 0.735 to 0.798, which is significantly larger than the inter-rater correlation.

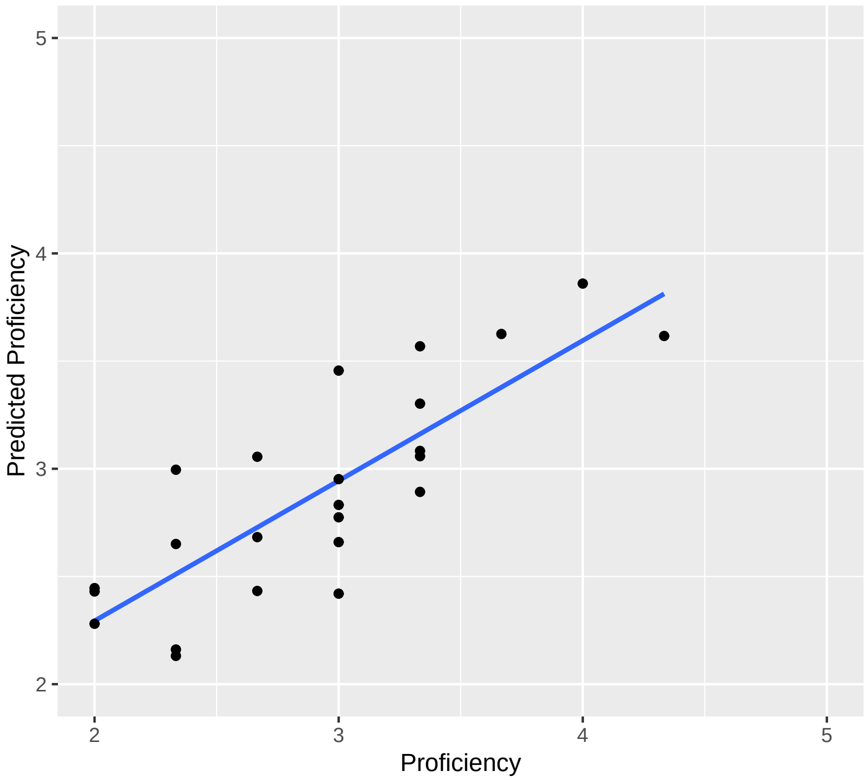

Fig. 3.12 Scatter plot of the proficiency and predicted proficiency on TMNED corpus. N = 26.

e) Automatic assessment on TEC corpus

In this section, TEC corpus was used for the spontaneous non-native English speech evaluation. The experimental conditions and evaluation steps are kept the same as those in the previous Section of evaluation on TMNED corpus.

The proficiency score is defined as the average segmental score among the three American English native raters (A1, A2 and A3) in TEC corpus. The test unit is the speech from one student in each dialogue. Figure 3.13 shows the histogram of proficiency in TEC. The average value of proficiency is 2.84, and the standard deviation is 0.48. The inter-rater correlation on TEC is presented in Table 3.12. Proficiency is the average segmental score among three human raters. A1A2 means the average score from human rater 1 and human rater 2. A2A3 means the average score from human rater 2 and human rater 3. A1A3 means the average score from human rater 1 and human rater 3. The correlation

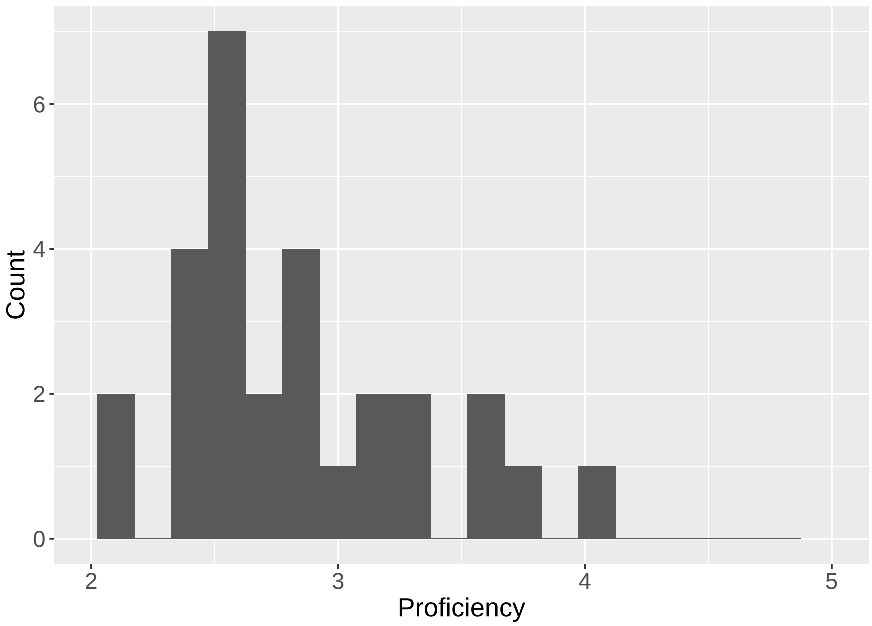

Fig. 3.13 Histogram of proficiency scores in TEC corpus. The average value of proficiency is 2.84, and the standard deviation is 0.48. N = 28.

Table 3.12 Inter-rater correlation on TEC corpus. Proficiency is the average segmental score among three human raters. A1A2 means the average score from human rater 1 and human rater 2. A2A3 means the average score from human rater 2 and human rater 3. A1A3 means the average score from human rater 1 and human rater 3. N = 28.

	A1	A2	A3	A1A2	A2A3	A1A3
A1	*	0.62, $p < 0.001$	0.60, $p < 0.001$	0.90, $p < 0.001$	0.68, $p < 0.001$	0.89, $p < 0.001$
A2	0.62, $p < 0.001$	*	0.62, $p < 0.001$	0.90, $p < 0.001$	0.89, $p < 0.001$	0.69, $p < 0.001$
A3	0.60, $p < 0.001$	0.62, $p < 0.001$	*	0.68, $p < 0.001$	0.91, $p < 0.001$	0.90, $p < 0.001$

between every two of the three raters is around 0.6, which is a little higher than that in TMNED. The correlation coefficients (r(A1-A2A3), r(A2-A1A3) and r(A3-A1A2)) range from 0.68 to 0.69, which shows the same tendency in TMNED.

There is no golden standard reference sentence in TEC corpus, this time, we directly calculated the

Table 3.13 Average value of RER from the ASR systems by both GMM-HMM and DNN-HMM on TEC.

	MONO	TRI	NN
RER[%]	49.01	55.92	40.24

RER score of each test unit from the outputs of non-native and native ASR systems. Table 3.13 shows the average value of RER in TEC. Compared with the results in Table 3.10, the average value of RER in TEC with the DNN-based acoustic model is rather lower than that in TMNED. Figure 3.14 shows the absolute correlation between the proficiency and ASR-based machine scores for the evaluation of TEC. RER_TRI has the highest correlation ($r = 0.708, p < 0.001$) among all the machine scores, and T_LL_NN has the highest value ($r = 0.436, p = 0.020$) among log-likelihood scores. Nine machine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI, RER_NN) are combined as the predicted proficiency score. Figure 3.15 shows a scatter plot of the proficiency score and the prediction score for the evaluation on TEC. The correlation between the prediction score and the proficiency score is $0.799, p < 0.001, N = 28$. We did the same test for the difference between human-human correlation and machine-human correlation as Section d). Three different tests are set as presented in Table 3.14. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). The correlation between the prediction score and the average score from two raters ranges from 0.756 to 0.775. Compared with the results in Table 3.11, the difference is not significant. From these results, it is indicated that the proposed RER is useful and acceptable for automatic evaluation on the spontaneous non-native English conversation speech.

(5) Analysis and discussion

Table 3.15 shows the comparison among different automatic pronunciation evaluation system. Our training data include non-native speech as English read by Japanese and native American English speech; therefore, they involve much phonemic category in both L1 and L2 speakers. We applied the

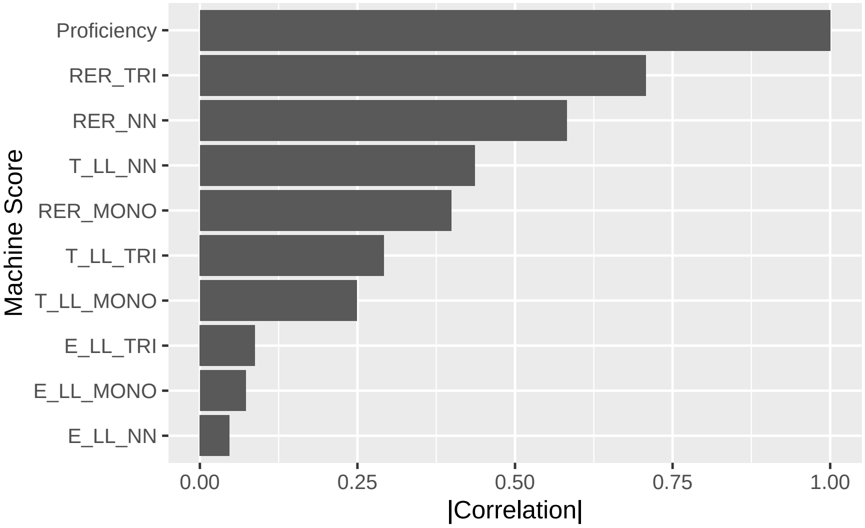

Fig. 3.14 Absolute correlation between the proficiency score and machine scores on TEC corpus.

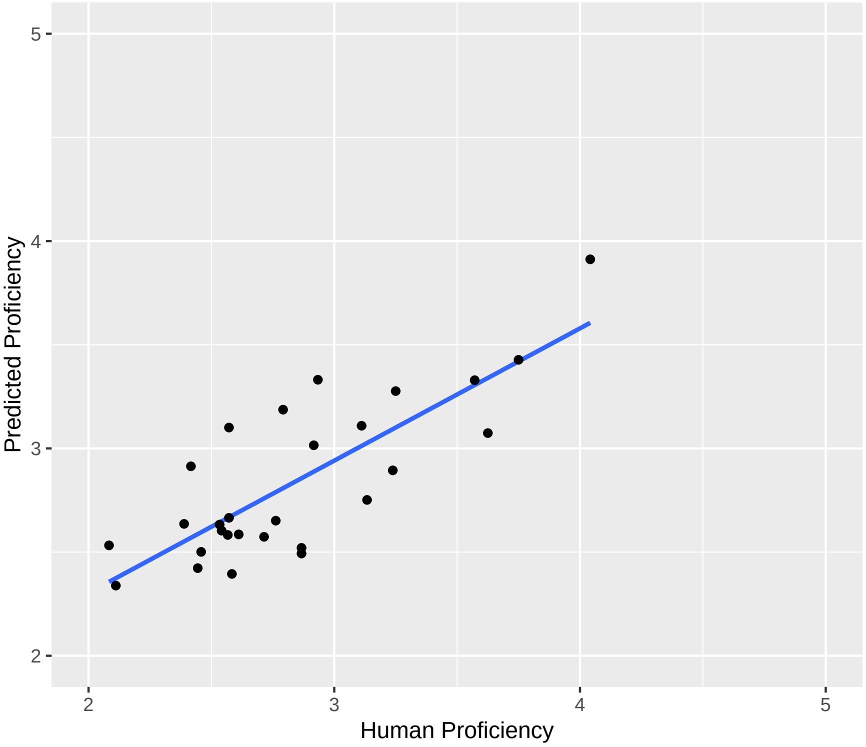

Fig. 3.15 Scatter plot of the proficiency and predicted proficiency on TEC corpus. N = 28.

Table 3.14 Test for the difference between human-human correlation vs. machine-human correlation on TEC corpus. PS (predicted score) is the best machine score from the combination of nine scores (T_LL_MONO, T_LL_TRI, T_LL_NN, E_LL_MONO, E_LL_TRI, E_LL_NN, RER_MONO, RER_TRI and RER_NN). r(PS-A2A3) = 0.768, $p < 0.001$. r(PS-A1A3) = 0.756, $p < 0.001$. r(PS-A1A2) = 0.775, $p < 0.001$. $N = 28$.

Test for difference between r(human-human) and r(machine-human)
r(A1-A2A3) *vs.* r(PS-A2A3): $z = -0.95, p = 0.340$
r(A2-A1A3) *vs.* r(PS-A1A3): $z = -0.69, p = 0.491$
r(A3-A1A2) *vs.* r(PS-A1A2): $z = -0.98, p = 0.326$

Table 3.15 Automatic pronunciation evaluation system.

		Feature	Reference text
GOP	Native AM, Forced Alignment Native AM, Phone Recognition	Likelihood	YES
[de Wet+, 2009]$^{10)}$	Non-native AM, Sentence Recognition	WER	YES
[Zechner+, 2009]$^{12)}$	Non-native AM, Sentence Recognition Native AM, Forced Alignment	Likelihood Duration	NO
[Moustroufas+, 2007]$^{14)}$	Non-native AM, Sentence Recognition Native AM, Sentence Recognition	Likelihood	NO
Tohoku University	Non-native AM, Sentence Recognition Native AM, Sentence Recognition	Likelihood RER	NO

recent state-of-the-art DNN-based training methods for the native and non-native acoustic models. As for the test sets in this study, we prepared three different kinds of speech for evaluation: reading aloud speech, constrained interactive dialogue speech and spontaneous English conversation speech. The correlation results show the positive usage in the CALL system.

It is common to use both non-native and native acoustic models in an automatic assessment language proficiency system, some related works are listed in Section 1. To obtain the RER, it costs double efforts to establish two ASR systems, compared with only one ASR. It also needs multi-models to decode the speech several times in terms of the RER's calculation. For the automatic assessment on non-native speech without knowing reference sentences, RER shows its superiority, especially on the experiments of TMNED and TEC, where the log-likelihood scores are rather poorly related with proficiency scores.

(6) Summary

In this chapter, we performed an automatic pronunciation evaluation system for English utterances from Japanese native speakers by utilizing both native and non-native acoustic models. Our system can evaluate the utterances from Japanese learners without knowing the reference text in advance. To this end, we proposed the RER as a new machine score for the regression-based prediction of pronunciation proficiency. As a result, the correlation between the proposed machine score and the human score is very strong.

4 . Language modeling in speech recognition for grammatical error detection

To develop a speech-based CALL system for grammatical exercise, first and foremost, the speech recognizer in the system should have the ability to recognize L2 learners' utterance including grammatical errors as accurately as possible for providing effective and corrective feedback. In this chapter, unlike the Chapter 3 only concentrating on the acoustic modeling in ASR for English pronunciation evaluation, for the grammatical error detection (GED) task, not only the acoustic model with extremely high performance is needed, but also language modeling is very important as the language model contains a lot of verbal information (correct or wrong grammar and expression). A good language model helps to improve the speech recognition rate, which further makes the GED system better.

(1) Previous method

The quality of ASR is always dependent on its two main components: acoustic models (AMs) and language models (LMs). The AMs describe how the sounds are pronounced in a language and the LMs cover the probabilities of words or sequences of words. Training data, as one of the important factors, significantly impacts the performance in each of these two different models. Normally, the training data is always obtained from the available corpus basically containing the correct contents, which might imply that it would be awful for some specific domain recognition tasks. So far, with the widespread application of deep learning, a large number of researches about deep neural networks (DNN) based acoustic models have increasingly improved speech recognition rate for non-native

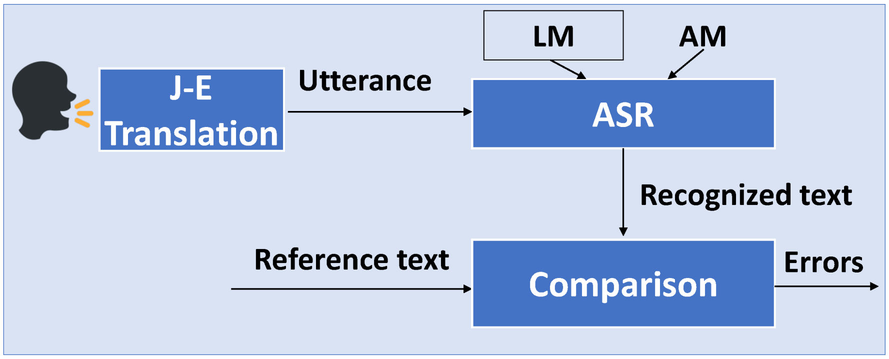

Fig. 4.1 The process of grammatical error detection in speech recognition.

speech[33),36)]. However, in these ordinary DNN-based systems, less attention has been paid to how to establish language models in ASR for improving recognition of L2 speech with grammatical errors.

First, we explain the overview of language modeling of the ASR system for speech-based grammatical error detection.

To be able to perform the grammatical error detection for speech, it is necessary to obtain information about such grammatical errors for language modeling in ASR. Since an ordinary English corpus does not contain many grammatical mistakes, a language model trained with such kind of corpus can not model grammatical errors and assign the probabilities of these errors, yielding the failure of speech errors recognition. Thus, previous researches concentrate on the methods for sentence generation with errors. Most of them established handcrafted grammatical error rules and implemented these rules on the prompted right source sentences for generating new sentences with errors[39),15),77)]. For instance, Anzai *et al.*[39)] proposed an approach to artificially generate sentences with grammatical errors that Japanese learners tend to make. They developed error rules according to the corpus [65)] that contains mistakes in spoken English from Japanese, and applied those errors to the correct sentences.

A problem with their method is that it is unclear whether the proposed rules are "optimum" in some senses. Because their rules were heuristically established, the generated sentences do not necessarily reflect the comprehensive situation of English spoken by Japanese learners.

Therefore, we examine to use the neural machine translation (NMT) to convert correct English sentences into those with grammatical errors. The proposed 3-gram language model is trained from these NMT generated sentences. To this end, on developing our speech-based CALL system, we design that Japanese Learners conduct a pre-exercise on the basic grammatical study in a specific domain and then do a Japanese-to-English oral translation task. Since we also assume that the system has the translated reference English correct sentences, the system can point out errors by comparing the correct sentences and the recognition result as long as the recognition result matches what the learner really have uttered. The process of grammatical error detection in our CALL system is showed in Figure 4.1.

(2) Evaluation metrics on language modeling for GED

For a traditional written GED system, it is common to use the default metrics as showed in Chapter 2. However, this time, to the spoken GED in our dialogue-based CALL system, as the performance of ASR is the precondition before the actual GED module, if the recognized output of the ASR is near to the human transcribed text, then we can take it as the written text in the traditional GED system, namely a full range of automated evaluation system. On the other hand, normally, it is difficult for the system to prepare the human-annotated correct text from the English speech spoken by the learners in advance. There is no such correct text as a standard reference text for comparison. Therefore, the evaluation metrics in this research need to be adjusted.

Regarding our evaluation on language modeling in ASR for the spoken GED system, we constructed a restricted Japanese to English oral translation corpus as explained in Chapter 2. Considering the evaluation research as an initial attempt with the characteristics of the cornerstone to the spoken GED, the TJTEOT corpus makes it suitable for the evaluation and experiment. Table 4.1 shows an example of evaluation in our system.

To evaluate the effect of the generated sentences from NMT for language modeling, here are four modified evaluation metrics. Consider a learner utters a sentence that contains several grammatical errors, and the CALL system points out the errors. Let N_e denote the number of actual grammatical errors included in an uttered sentence. Then the system points out the wrong words. Let N_{TP} and N_{FP} be the number of correctly detected error words and wrongly detected error words, respectively.

Table 4.1 An example for evaluation in our speech-based GED system.

Learner	he	likes	speak	English	on	*	classroom
Reference	he	likes	speaking	English	in	the	classroom
ASR output	he	likely	speak	English	on	the	classrooms
Evaluation	TN	FP	TP	TN	TP	FN	TN
Detection	✓	X	X	✓	X	✓	✓

1. Word error rate (WER).

$$WER = \frac{S + D + I}{N} \tag{4.1}$$

WER shows the performance with the language model in ASR by comparing the recognized result from ASR to the human transcription. The definition of WER has the same meaning to that in Chapter 3 with a more intuitive statistics. In which, S is the number of substitutions, D is the number of deletions, I is the number of insertions, N is the number of words in the human transcription.

2. Precision (P).

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \tag{4.2}$$

P is the proportion of actual recognized error words among all the recognized error words. It can be used for measuring the quality of the ASR.

3. Recall (R).

$$R = \frac{N_{TP}}{N_e} \tag{4.3}$$

R is the proportion of actual recognized error words among all the error words in real utterance. It can be used for measuring the coverage of the ASR.

4. F-measure.

$$F = \frac{2PR}{P + R} \tag{4.4}$$

F-measure shows the harmonic mean of P and R.

Table 4.2 Settings in DNN training for speech-based GED task.

	NN1	NN2	NN3
Input Feature	40MFCCs	40MFCCs	40MFCCs
Hidden Type	tanh	*p*-norm	*p*-norm E.L.
Input Nodes	360	360	360
Hidden Layer	4	4	4
Hidden Dim	1024	1000/200	1000/200
Output Nodes	1551	1551	1551

(3) Data pre-processing

Before digging into the evaluation of language modeling from the sentences generated by NMT, a standard ASR should be first established.

ERJ database[1)] was used to train the acoustic models in this part. TJTEOT corpus was used as the evaluation set. All the sentences in the ERJ database were selected out for the training set. The utterances from TJTEOT were determined as test set. HMM-based monophone and triphone acoustic models were primarily trained as the basis of DNN-based acoustic models. The settings of the conventional GMM-HMM acoustic models were kept same as those in Chapter 3. Here only selected NNET2 methods for DNN training. The details of settings in three different DNN training are listed in Table 4.2.

By constructing the default language model, we have made a statistical analysis on the TJTEOT corpus. Figure 4.2 shows the counted number of translated English sentences in TJTEOT corpus. In this figure, the horizontal axis from S1 to S42 means the 42 translation units. Totally, there are 144 different sentences in the corpus.

The CMU pronouncing dictionary is modified as the baseline pronunciation lexicon by changing all the stressed phonemes to non-stressed phonemes.

We trained a trigram language model by using all the sentences in the ERJ database and the 144 English sentences including the target correct sentences and human transcription from TJTEOT corpus. Thus, this trigram language model is a closed gold condition language model.

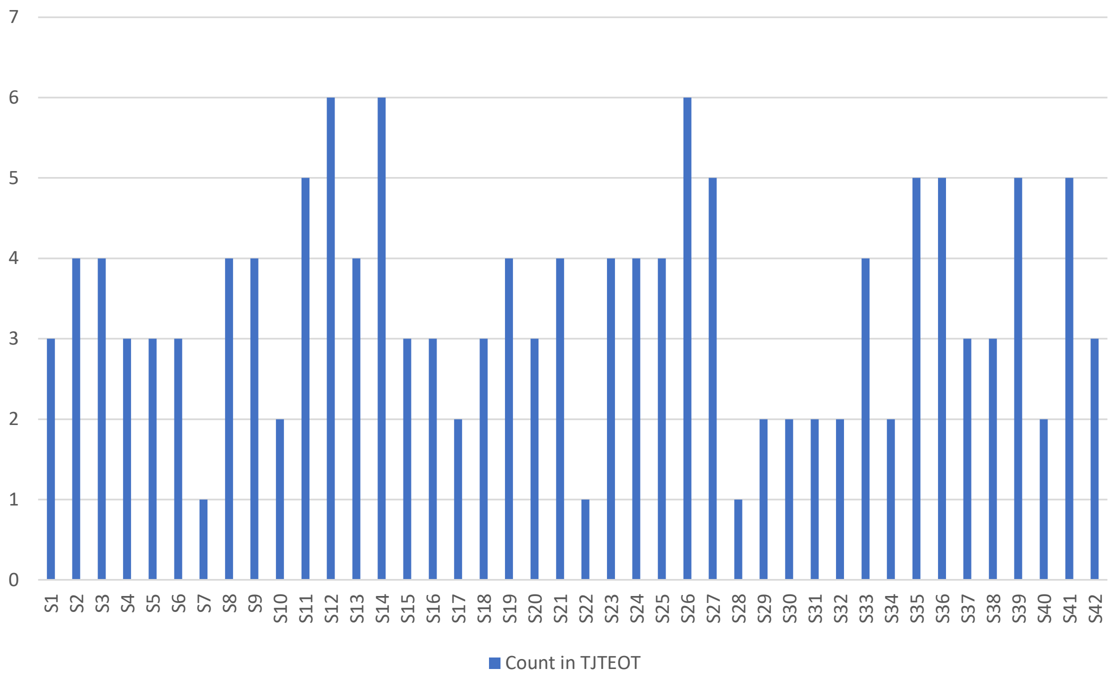

Fig. 4.2 Statistical information about counted number of translated English sentences in TJTEOT corpus. $N = 42$.

The recognition results are presented in Figure 4.3. From these results, we found that the NN1 method performed well among the acoustic models. Therefore, we only choose the NN1-based acoustic model for the next experiments.

(4) Conditions of NMT

From the point of view of parameter learning, the problem of recurrent neural networks (RNNs) is that there is a vanishing or exploding gradient[78)], which makes it difficult for RNNs to deal with long distance dependency. Therefore, the NMT model did not perform well with the traditional RNNs.

a) LSTM & GRU

Long short-term memory (LSTM) is a special RNNs, mainly to solve the exploding and vanishing problem in long sequence training[79),80)]. With the help of the gate mechanism (allowing the deletion

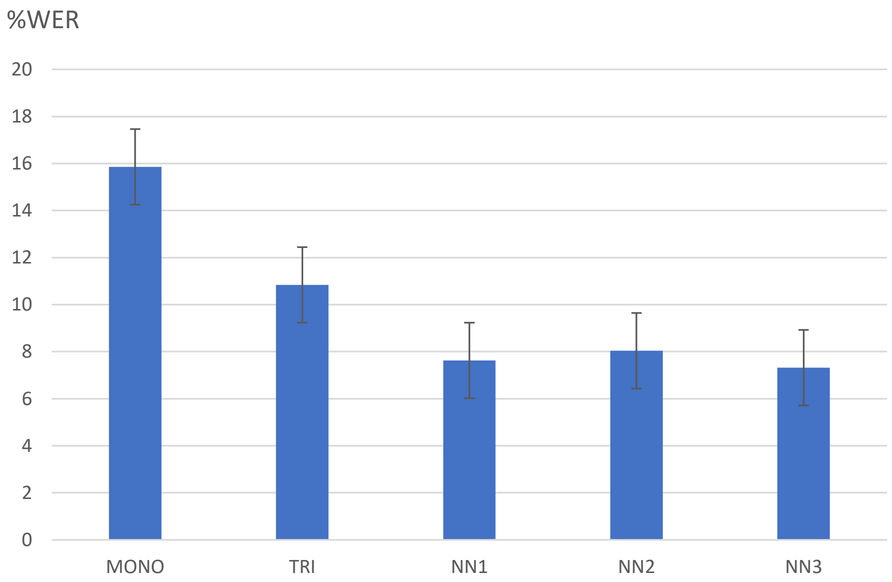

Fig. 4.3 The avarage value of WER with different acoustic models in TJTEOT corpus. Here, language model is under closed condition.

and updating of explicit memory in the LSTM), that problem is controlled, allowing the model to get far better access to the "long distance" in the sentence. Simply put, LSTM can perform better than vanilla RNNs in longer sequences.

Gate recurrent unit (GRU)[81)] is one kind of RNNs. Like LSTM, it is also proposed to solve the problem such as gradient in long-term memory and back propagation. Compared to LSTM, the use of GRU can achieve comparable results, and it is easier to train with less parameters, which can greatly improve training efficiency, therefore in most of time, it is more inclined to use GRU.

Figure 4.4 shows the structure of LSTM and GRU with the input and output. The LSTM has two transfer states, a cell state c_t, and a hidden state h_t. The input and output structure of GRU is the same as that of normal RNNs. There is a current input x_t, and only a hidden state h_{t-1} passed by the previous node. This hidden state contains information about the previous node. Combined with x_t and h_{t-1}, the GRU gets the output y_t of the current hidden node and the hidden state h_t passed to the next node.

The introduction of LSTM and GRU solves the problem of "long-distance dependency" and turns

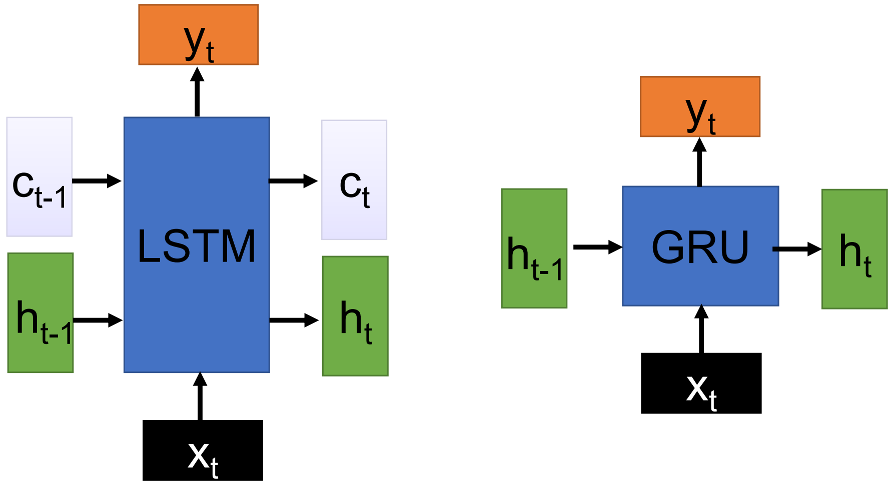

Fig. 4.4 The structure of LSTM and GRU with the input and output.

the main problem of NMT into a "fixed-length vector" problem: regardless of the length of the source sentence, this neural network needs to compress it into a fixed-length vector, which brings more complexity and uncertainty in the decoding process, especially when the source sentence is very long.

b) Attention mechanism

The hidden state of the last neuron of the encoder RNNs is used as the initial state of the decoder RNNs. That is to say, the last hidden state vector of the encoder needs to carry all the information of the source sentence, which becomes the information bottleneck of the entire model. Since the attention mechanism for NMT in 2014 was introduced[82)], the fixed-length vector problem has also begun to be resolved. As Figure 4.5 shows, a content-based attention mechanism can be applied to dynamically generate a (weighted) context vector from the source sentence. The neural network then predicts the word based on the context vector rather than a fixed-length vector.

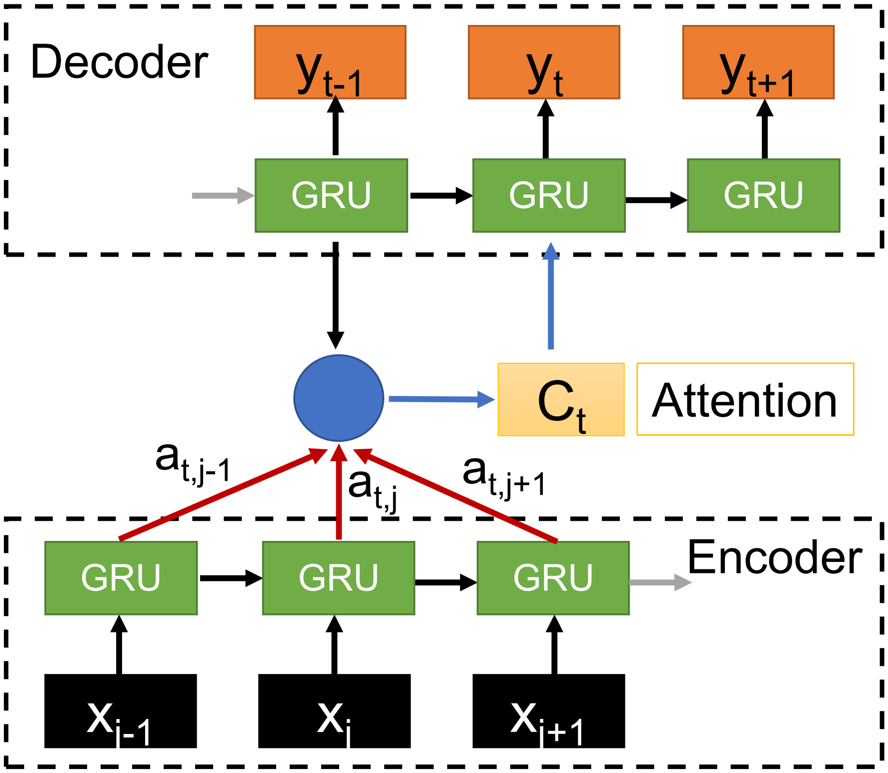

Fig. 4.5 The attention mechanism in neural machine translation with encoder-decoder structure.

(5) Experiments and results

a) Generating sentences

NMT generating We generated sentences with errors based on an encoder-decoder model implemented with Keras[1]. In this experiment, we propose sentence generation with errors based on the recent NMT method showed in Figure 4.6. This is neural network sequence modeling with attention mechanism[82)] that takes a correct source sentence into a vector and generates a sentence with grammatical errors using the decoder. As a first step for training an initial NMT to generate identical sentences, we used 201935 identical sentence pairs to train the NMT model. Table 4.3 shows the experimental conditions. We set the time steps with the maximum number of words in an English sentence to 22 and the GRUs (Gated Recurrent Units) to 256 units. The number of epochs in the

[1] thttps://github.com/keras-team/keras

Table 4.3 Settings of our proposed NMT for language modeling.

RNN architecture	GRUs
No. of units in hidden layer	256
Time steps of input layer	22
Optimizer	Adam
Activation functions	softmax
Trainable parameters	16434516

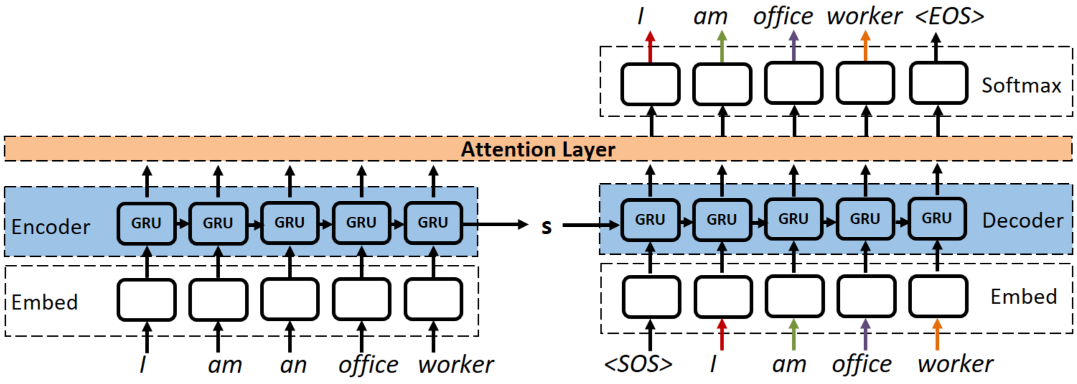

Fig. 4.6 An encoder-decoder model with attention mechanism for generation sentences with errors.

initial model is 50.

When the initial NMT model was established, it was fine-tuned to the prepared 8476 right-to-wrong sentence pairs applied by running additional training epochs. The vocabulary size was 7746. 10 different epochs (10, 20, 30, 40, 50, 60, 70, 80, 90, 100) were set in this process for updating the parameters of the initial NMT model, which yielded 10 new NMT models for generating sentences with errors.

Table 4.4 shows some samples of the sentences generated from our NMT system. From the first two rows of the table, it is obvious that the NMT models generated the sentences with article errors. In the third row is a kind of word missing errors and in the forth is the preposition errors. Figure 4.7 presents the NMT attention alignment visualization in a heat map. In this figure, the attention weight between the word “office” in the source and the word “a office” in the output is strong, where the

Table 4.4 Samples of sentences generated from NMT.

Source sentence	NMT-based sentence
I’m an office worker	I’m a office worker
I’m an office worker	I’m office worker
how much is it	how much is
how much were they	how much about they
. . .	. . .

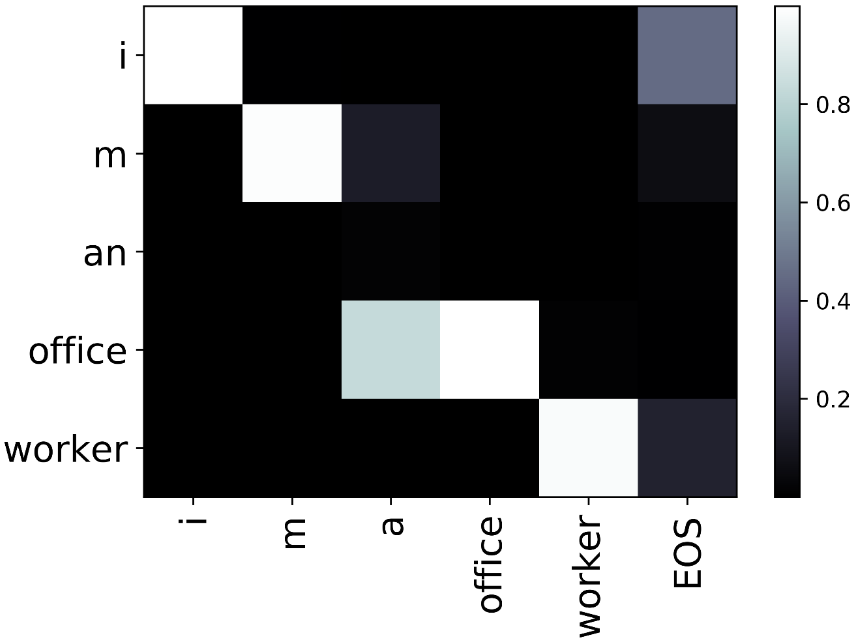

Fig. 4.7 Heat map of NMT attention alignment visualization. Here, the source sentence is “I’m an office worker” and the output sentence is “I’m a office worker”.

word “an” in the source nearly does not correlate with the target output.

Rule-based generating To develop the rule-based error generator, we aligned the right and wrong sentence pairs in NICT-JLE corpus using the IBM Model 1 alignment[52). We used word.alignment package for R. Then we extracted 5841 kinds of error phrase pairs. The top 20 right-wrong pairs with the highest frequency are shown in Table 4.5. Then we randomly replaced the “right” part included in an input sentence with the “wrong” phrase. A probability to apply any error rules in a sentence was set to 0.9. The insertion errors and other errors were treated differently. First, we determined whether insertion error is applied or not with a probability proportion to the frequency of the insertion error. If an insertion error was applied, the position of insertion was determined randomly, and the word to insert was chosen according to the relative frequency of insertion errors. If an insertion error was not applied, other errors (substitution and deletion errors) were chosen according to the relative frequency, and applied to the input sentence.

We obtained 420 English sentences from those 10 fine-tuned NMT models and another 420 ones from the rule-based error generator, respectively. Table 4.6 shows some generated sentences from a target one. We can find that most errors generated by the rules are not related to the actual errors that appeared in the speech.

To evaluate the proposed method for language modeling in ASR, we fixed the acoustic model referred to our previous work[36]. This DNN-based AM was trained from ERJ corpus with the conditions showed in Table 4.7.

b) Experimental results with different language models

For the experiment, we prepared five LMs for comparison, as shown in Table 4.8. The baseline model was trained from the sentences in ERJ and the target sentences of TJTEOT, considering the situation where we know the sentences the learners were supposed to say. The LM_rule and LM_NMT models were trained by adding the generated sentences by the rule-based error generation and the NMT, respectively. We merged the two sets of generated sentences together to train the LM_rule & NMT model. We doubled the amount of generated sentences in LM_rule and LM_NMT respectively. Thus, the number of sentences in the middle three LMs that appeared in Table 4.8 is the same. Finally, the LM_gold model used the uttered sentences as a part of the training data to investigate the performance of the closed model.

1. Language Model baseline (LM_baseline). 980 ERJ (English read by Japanese) database sentences + 42 target correct sentences in the translation task.

Table 4.5 Top 20 errors in the parallel right-to-wrong sentence pairs from NICT JLE corpus. “-” stands for no word should be there.

Ranking	Correct	Wrong	Frequency
1	-	the	620
2	-	to	181
3	-	a	177
4	-	is	151
5	are	is	98
6	-	in	51
7	were	are	48
8	an	a	46
9	had	have	46
10	movies	movie	40
11	-	so	33
12	-	and	29
13	-	I	22
14	the	of	21
15	friends	friend	19
16	kinds	kind	18
17	students	student	18
18	places	place	17
19	cats	cat	16
20	trains	train	16

2. Rule-based Language Model (LM_rule). 980 ERJ database sentences + 42 target correct sentences in the translation task + 420 rule-based generated sentences.

3. NMT generated Language Model (LM_NMT). 980 ERJ database sentences + 42 target correct sentences in the translation task + 420 NMT model generated sentences.

Table 4.6 Examples of generated sentences.

Target	They visited me on Sunday.
rule-based	They visited me on the Sunday.
rule-based	They the visited me on Sunday.
rule-based	They a visited me on Sunday.
NMT model	They visited at Sunday.
NMT model	They visited Sunday.
NMT model	They visited me on Sunday.

Table 4.7 DNN parameters of acoustic modeling in ASR.

Input feature	40 MFCCs
Type of hidden layer	hyperbolic tangent
No. of input nodes	360
No. of hidden layer	4
Hidden dimension	1024
No. of output nodes	1551

Table 4.8 The three different language models in speech recognition experiment.

Training data	LM_ baseline	LM_ rule	LM_ NMT	LM_ rule & NMT	LM_ gold
ERJ sentences	✓	✓	✓	✓	✓
Target sentences	✓	✓	✓	✓	✓
Rule sentences		✓		✓	
NMT sentences			✓	✓	
Human transcription					✓

4. Rule-based and NMT generated Language Model (LM_rule & NMT). 980 ERJ database sentences + 42 target correct sentences in the translation task + 420 rule-based generated sentences + 420 NMT model generated sentences.

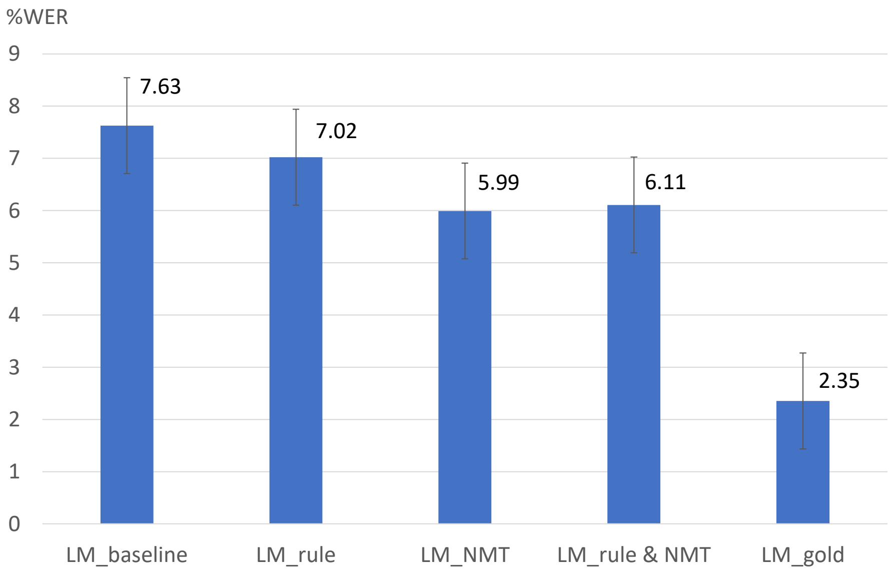

Fig. 4.8 Average speech recognition results in TJTEOT corpus by using the five different language models.

5. NMT generated Language Model (LM_NMT). 980 ERJ database sentences + 42 target correct sentences in the translation task + 144 human transcription sentences in the translation task (correct sentences and sentences with grammatical errors transcribed from the non-native English speech).

Figure 4.8 shows the speech recognition results by applying the five different language models. By comparing with the result of baseline LM, the other four LMs contribute more or less recognition accuracy in ASR. Especially in the closed-condition LM as gold LM, the value of WER decreased to 2.35%. By investigating the results of the middle three LMs, the previous rule-based LM promoted the recognition a little and our proposed LM gave more accuracy than the previous method. By combining the rule-based and NMT-based sentences, the combined LM even did a bit worse than the NMT-based one, which indicated that the rule-based method produced some errors not related to the real ones. The recognition results show that if the LM includes the real grammatical errors from the learners, the ASR will perform better.

Table 4.9 Evaluation results by different language modeling for grammatical error detection.

	P	*R*	*F*
LM_baseline	0.47	0.54	0.50
LM_rule	0.48	0.44	0.46
LM_NMT	0.54	0.52	0.53
LM_rule & NMT	0.52	0.52	0.52
LM_gold	0.73	0.83	0.78

(6) Analysis and discussion

Table 4.9 shows the evaluation results of precision, recall and F-measure by using different language models in ASR. The closed-conditional gold LM is able to conduct GED task in the speech to an acceptable degree, however, in a real CALL system, it is always not easy to obtain the human transcriptions in advance. Compared with the results of baseline LM, both rule-based generated sentences and NMT-based generated sentences have promoted the performance of detecting grammatical errors. The value of *R* in rule-based language modeling is lower than that in the baseline, which proves many not related rule-based errors reduced recognizer performance. From these results, it is indicated that the sentences generated by NMT play a beneficial role in speech-based GED task.

(7) Summary

In this chapter, by using the five different language models, the contents containing more related grammatical errors will help to improve the speech recognition accuracy. We introduced the process of generation of text with grammatical errors by using neural machine translation (NMT).

Normally, the NMT only translates one sentence to one different sentence. In order to cope with this limitation in our GED system, we set different epochs in NMT for generating more sentences. The results show that the proposed method did work for generated related errors to lower the WER. In the evaluation of the five language models to GED task, the generated errors by NMT are more realistic.

5 . Conclusions

In this thesis, a new dialogue-based computer-assisted language learning (CALL) system has been built for English language study to Japanese learners. Language proficiency evaluation is the key component in our dialogue-based CALL system as well as the target research in this thesis. For the language proficiency evaluation, pronunciation evaluation and grammatical error detection were described in the first two chapters. The method and idea proposed for the former objective in the CALL system were clarified in chapter 3. And the approaches in the other one were explained in chapter4.

In chapter 3, regarding the aspect on phonemic-level pronunciation evaluation, traditional evaluation system often assume that there are reference texts prepared, which is unavailable for a free-conversation language learning system. To this end, we proposed the RER as a new machine score for the regression-based prediction of proficiency, based on the two assumed conditions:

1. non-native ASR has sufficiently high recognition accuracy

2. WER from native ASR is effective to evaluate language proficiency

To evaluate the correlation between the proficiency scores and the proposed RER scores, we firstly analyzed the inter-rater correlation on ERJ database, TMNED corpus and TEC corpus. Results show the raters have very high correlation among themselves in ERJ database ($r \approx 0.8$), while not a good correlation in TMNED ($r \approx 0.5$) or TEC ($r \approx 0.6$). The speech data size of TMNED or TEC is rather smaller than that of ERJ database. Especially in TMNED corpus, the raters gave scores (three different kinds: speech-related, emotion-related and gesture-related) by directly watching the recording video. Therefore, the segmental score analyzed in our work partly influenced by other aspects from the learner in the video. The results of the evaluation on ERJ database show the strong effectiveness of RER for language proficiency assessment. RER derived from DNN-based acoustic models already has a high correlation coefficient with the proficiency score as 0.721. On the evaluation of TEC corpus, one RER-based machine score (RER_TRI) reaches a higher correlation to proficiency than WER-based machine scores. We conducted the linear regression prediction of proficiency with combined machine scores. In all experiments, when all the machine scores (excluding WER-based scores) were combined, the correlation coefficient reached to extremely high with the proficiency score (ERJ database: $r = 0.826, p < 0.001, N = 190$; TMNED corpus: $r = 0.803, p < 0.001, N = 26$; TEC corpus: $r = 0.799, p < 0.001, N = 28$) , which means this prediction method could improve human-to-machine correlation significantly.

In chapter 4, for the grammatical error detection in our CALL system, firstly, we assumed that the Japanese learners could conduct English grammar practice which can be regarded as a sub-module in our system. To this end, we developed a Japanese-to-English oral translation task as the TJTEOT corpus for test and evaluation. Therefore, our CALL system has the reference target English sentences in advance, which can be compared with the utterances from the learners to detect the grammatical errors. To recognize the grammatical errors in speech with high accuracy, we introduced the sentence generation with errors by utilizing the state-of-the-art NMT models for language modeling. These NMT models were developed based on the sequence-to-sequence model with the current outstanding attention mechanism. Normally, one NMT model only translates one sentence to one different sentence. In order to tackle with this limitation in our CALL system, we set different epochs in NMT training, which updates the parameters in NMT for generating a new NMT model, yielding more sentences with errors for language modeling. Compared with the traditional rule-based method, our proposed method can be established more quickly without spending a lot of time modeling the error rules. Since the NMT can make use of the context information due to its specific inner structure, the generated sentences from NMT have more comprehensive information about grammatical errors. According to the evaluation results, we can found that the language model trained by NMT generated sentences contributes lower WER than the rule-based language model, which indicated that the NMT models generated sentences with more realistic errors happened in the real non-native speech.

In this thesis, regarding a CALL system, speech recognition and automatic pronunciation assessment for L2 learners are vital tasks. The former needs L2 acoustic models and the latter needs L1 acoustic models. Hence, our proposed RER obviously bridges the gap between the two different things. The RER demonstrated a farsighted meaning to the evaluation of spontaneous non-native speech in the dialogue-based CALL system.

By comparing the results of different language models in the grammatical error detection task, our proposed NMT-based method has a significant advantage over the rule-based one, yielding the more related errors happened in the real non-native English speech. Moreover, this deep learning-based method fills a gap in language modeling for speech-based grammatical error detection in the CALL system.

REFERENCES

1) N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, “Development of English speech database read by Japanese to support CALL research,” Proc. Int. Cong. Acoust., pp.557–560, 2004.

2) H. Wu, Y. Chiba, T. Nose, and A. Ito, “Analyzing effect of physical expression on English proficiency for multimodal computer-assisted language learning,” Proc. Interspeech, pp.1746–1750, 2018.

3) A.C. Baugh and T. Cable, A history of the English language, Routledge, 1993.

4) D. Crystal, “Two thousand million?,” English today, vol.24, no.1, pp.3–6, 2008.

5) D. Crystal, English as a global language, Cambridge university press, 2012.

6) K. Beatty, Teaching & researching: Computer-assisted language learning, Routledge, 2013.

7) M. Warschauer and D. Healey, “Computers and language learning: An overview,” Language teaching, vol.31, no.2, pp.57–71, 1998.

8) D. Larsen-Freeman, Techniques and principles in language teaching, Oxford University, 2000.

9) H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, “Combination of machine scores for automatic grading of pronunciation quality,” Speech Communication, vol.30, no.2-3, pp.121–130, 2000.

10) F. de Wet, C. Van der Walt, and T. Niesler, “Automatic assessment of oral language proficiency and listening comprehension,” Speech Communication, vol.51, no.10, pp.864–874, 2009.

11) S. Seneff, C. Wang, and J. Zhang, "Spoken conversational interaction for language learning," Proc. InSTIL/ICALL Symposium 2004, 2004.

12) K. Zechner, D. Higgins, X. Xi, and D.M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," Speech Communication, vol.51, no.10, pp.883–895, 2009.

13) S. Lee, H. Noh, J. Lee, K. Lee, and G. Lee, "Postech approaches for dialog-based english conversation tutoring," Proc. APSIPA ASC, pp.794–803, 2010.

14) N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," Computer Speech and Language, vol.21, no.1, pp.219–230, 2007.

15) A. Ito, R. Tsutsui, S. Makino, and M. Suzuki, "Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system," Proc. Interspeech, pp.2819–2822, 2008.

16) M. Tu, A. Grabek, J. Liss, and V. Berisha, "Investigating the role of L1 in automatic pronunciation evaluation of L2 speech," Proc. Interspeech, pp.1636–1640, 2018.

17) M. Suzuki, T. Konno, A. Ito, and S. Makino, "Automatic evaluation system of english prosody based on word importance factor," Journal of Systemics, Cybernetics and Informatics, vol.6, no.4, pp.83–90, 2008.

18) F. Hönig, A. Batliner, and E. Nöth, "Automatic assessment of non-native prosody annotation, modelling and evaluation," Proc. IS ADEPT, pp.21–30, 2012.

19) Q.T. Truong, T. Kato, and S. Yamamoto, "Automatic assessment of l2 english word prosody using weighted distances of f0 and intensity contours.," Proc. Interspeech, pp.2186–2190, 2018.

20) L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality," Speech Communication, vol.30, no.2-3, pp.83–93, 2000.

21) C. Molina, N.B. Yoma, J. Wuth, and H. Vivanco, "Asr based pronunciation evaluation with automatically generated competing vocabulary and classifier fusion," Speech Communication, vol.51, no.6, pp.485–498, 2009.

22) Y. Kim, H. Franco, and L. Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction," Proc. Eurospeech, pp.649–652, 1997.

23) S.M. Witt and S.J. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," Speech Communication, vol.30, no.2-3, pp.95–108, 2000.

24) S. Witt and S. Young, "Computer-assisted pronunciation teaching based on automatic speech recognition," Language Teaching and Language Technology Groningen,, vol.The Netherlands, pp.25–35, 1997.

25) H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic detection of phone-level mispronunciation for language learning," Proc. Eurospeech, pp.851–854, 1999.

26) A. Ito, Y.L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree," Acoustical science and technology, vol.28, no.2, pp.131–133, 2007.

27) W. Hu, Y. Qian, F.K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," Speech Communication, vol.67, pp.154–166, 2015.

28) W. Li, S.M. Siniscalchi, N.F. Chen, and C.H. Lee, "Improving non-native mispronunciation detection and enriching diagnostic feedback with DNN-based speech attribute modeling," Proc. ICASSP, pp.6135–6139, 2016.

29) S. Mao, X. Li, K. Li, Z. Wu, X. Liu, and H. Meng, "Unsupervised discovery of an extended phoneme set in l2 english speech for mispronunciation detection and diagnosis," Proc. ICASSP, pp.6244–6248, 2018.

30) K. Li, S. Mao, X. Li, Z. Wu, and H. Meng, “Automatic lexical stress and pitch accent detection for L2 english speech using multi-distribution deep neural networks,” Speech Communication, vol.96, pp.28–36, 2018.

31) S. Witt and S.J. Young, “Language learning based on non-native speech recognition,” Proc. Eurospeech, pp.633–636, 1997.

32) J. Yue, F. Shiozawa, S. Toyama, Y. Yamauchi, K. Ito, D. Saito, and N. Minematsu, “Automatic scoring of shadowing speech based on DNN posteriors and their DTW,” Proc. Interspeech, pp.1422–1426, 2017.

33) W. Hu, Y. Qian, and F.K. Soong, “A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL),” Proc. Interspeech, pp.1886–1890, 2013.

34) N. Minematsu, G. Kurata, and K. Hirose, “Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition,” Proc. ICSLP, pp.529–531, 2002.

35) H. Franco, L. Neumeyer, Y. Kim, and O. Ronen, “Automatic pronunciation scoring for language instruction,” Proc. ICASSP, pp.1471–1474, 1997.

36) J. Fu, Y. Chiba, T. Nose, and A. Ito, “Evaluation of English speech recognition for Japanese learners using DNN-based acoustic models,” Pan, J.-S., Ito, A., Tsai, P.-W. and Jain, L. C. (Eds.) Recent Advances in Intelligent Information Hiding and Multimedia Signal Processing, pp.93–100, Springer, 2018.

37) G. Kawai and K. Hirose, “A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training,” Proc. ICSLP, pp.1823–1826, 1998.

38) Y. Ohkawa, M. Suzuki, H. Ogasawara, A. Ito, and S. Makino, “A speaker adaptation method for non-native speech using learners ’ native utterances for computer-assisted language learning systems,” Speech Communication, vol.51, no.10, pp.875–882, 2009.

39) T. Anzai and A. Ito, “Recognition of utterances with grammatical mistakes based on optimization

of language model towards interactive call systems," Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pp.1–4, IEEE, 2012.

40) T. Anzai, S. Hahm, A. Ito, M. Ito, and S. Makino, "Grammatical error detection from english utterances spoken by japanese," Proc. APSIPA ASC, pp.482–485, 2010.

41) L.R. Bahl, P.F. Brown, P.V. De Souza, and R.L. Mercer, "Maximum mutual information estimation of hidden markov model parameters for speech recognition," Proc. ICASSP, pp.49–52, 1986.

42) K.F. Lee, "On large-vocabulary speaker-independent continuous speech recognition," Speech communication, vol.7, no.4, pp.375–379, 1988.

43) M. Gales, S. Young, *et al.*, "The application of hidden markov models in speech recognition," Foundations and Trends® in Signal Processing, vol.1, no.3, pp.195–304, 2008.

44) L. Muda, M. Begam, and I. Elamvazuthi, "Voice recognition algorithms using mel frequency cepstral coefficient (mfcc) and dynamic time warping (dtw) techniques," arXiv preprint arXiv:1003.4083, 2010.

45) G. Hinton, L. Deng, D. Yu, G.E. Dahl, A.r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T.N. Sainath, *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," IEEE Signal processing magazine, vol.29, no.6, pp.82–97, 2012.

46) J. Norrish, Language learners and their errors, Macmillan Hong Kong, 1983.

47) N. Macdonald, L. Frase, P. Gingrich, and S. Keenan, "The writer's workbench: Computer aids for text analysis," IEEE Transactions on Communications, vol.30, no.1, pp.105–110, 1982.

48) H. Oyama and Y. Matsumoto, "Automatic error detection method for japanese particles," Ritsumeikan Asia Pacific University Polyglossia, vol.18, pp.55–63, 2010.

49) C. Brockett, W.B. Dolan, and M. Gamon, "Correcting esl errors using phrasal smt techniques," Proc. ACL, pp.249–256, Association for Computational Linguistics, 2006.

50) Z. Yuan and M. Felice, “Constrained grammatical error correction using statistical machine translation,” Proc. CoNLL, pp.52–61, 2013.

51) M. Chodorow, M. Dickinson, R. Israel, and J. Tetreault, “Problems in evaluating grammatical error detection systems,” Proc. COLING 2012, pp.611–628, 2012.

52) P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer, “The mathematics of statistical machine translation: Parameter estimation,” Computational linguistics, vol.19, no.2, pp.263–311, 1993.

53) N. Kalchbrenner and P. Blunsom, “Recurrent continuous translation models,” Proc. EMNLP, pp.1700–1709, 2013.

54) T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, “Recurrent neural network based language model,” Proc. Interspeech, 2010.

55) D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, *et al.*, “The Kaldi speech recognition toolkit,” Proc. ASRU, IEEE Signal Processing Society, 2011.

56) T. Makino and R. Aoki, “English read by Japanese phonetic corpus: an interim report,” Research in Language, vol.10, no.1, pp.79–95, 2012.

57) D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, “Regularized-MLLR speaker adaptation for computer-assisted language learning system,” Proc. Interspeech, pp.594–597, 2010.

58) N. Minematsu, K. Okabe, K. Ogaki, and K. Hirose, “Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database,” Proc. Interspeech, pp.1481–1484, 2011.

59) X. Wang, T. Kato, and S. Yamamoto, “Phoneme set design based on integrated acoustic and linguistic features for second language speech recognition,” IEICE TRANSACTIONS on Information and Systems, vol.100, no.4, pp.857–864, 2017.

60) Y. Oshima, S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, “Non-native text-to-speech preserving speaker individuality based on partial correction of prosodic and phonetic characteristics,” IEICE TRANSACTIONS on Information and Systems, vol.99, no.12, pp.3132–3139, 2016.

61) J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, and D.S. Pallett, “DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1,” NASA STI/Recon technical report, vol.93, 1993.

62) A. Graves and J. Schmidhuber, “Framewise phoneme classification with bidirectional LSTM and other neural network architectures,” Neural Networks, vol.18, no.5-6, pp.602–610, 2005.

63) K. Greff, R.K. Srivastava, J. Koutník, B.R. Steunebrink, and J. Schmidhuber, “LSTM: A search space odyssey,” IEEE transactions on neural networks and learning systems, vol.28, no.10, pp.2222–2232, 2017.

64) C. Danescu-Niculescu-Mizil and L. Lee, “Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs,” Pro. ACL, pp.76–87, Association for Computational Linguistics, 2011.

65) E. Izumi, K. Uchimoto, and H. Isahara, “The nict jle corpus: Exploiting the language learners ’ speech database for research and education,” International Journal of the Computer, the Internet and Management, vol.12, no.2, pp.119–125, 2004.

66) J. Bernstein, J. De Jong, D. Pisoni, and B. Townshend, “Two experiments on automatic scoring of spoken language proficiency,” Proc. InSTIL, pp.57–61, 2000.

67) C. Cucchiarini, H. Strik, and L. Boves, “Quantitative assessment of second language learners ’ fluency: Comparisons between read and spontaneous speech,” The Journal of the Acoustical Society of America, vol.111, no.6, pp.2862–2873, 2002.

68) C. Cucchiarini, H. Strik, and L. Boves, “Different aspects of expert pronunciation quality ratings

and their relation to scores produced by speech recognition algorithms," Speech Communication, vol.30, no.2-3, pp.109–119, 2000.

69) J. Tao, K. Evanini, and X. Wang, "The influence of automatic speech recognition accuracy on the performance of an automated speech assessment system," Proc. Spoken Language Technology Workshop (SLT), pp.294–299, IEEE, 2014.

70) V.I. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," Soviet physics doklady, pp.707–710, 1966.

71) R.A. Spears, B.J. Birner, S.R. Kleinedler, and L. Nisset, Conversational American English, McGraw-Hill Education, 2010.

72) X. Zhang, J. Trmal, D. Povey, and S. Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," Proc. ICASSP, pp.215–219, 2014.

73) J. Pan, C. Liu, Z. Wang, Y. Hu, and H. Jiang, "Investigation of deep neural networks (DNN) for large vocabulary continuous speech recognition: Why DNN surpasses GMMs in acoustic modeling," Proc. ISCSLP, pp.301–305, 2012.

74) A.r. Mohamed, G. Hinton, and G. Penn, "Understanding how deep belief networks perform acoustic modelling," Proc. ICASSP, pp.4273–4276, 2012.

75) T. Yoshioka, X. Chen, and M.J. Gales, "Impact of single-microphone dereverberation on DNN-based meeting transcription systems," Proc. ICASSP, pp.5527–5531, 2014.

76) X.L. Meng, R. Rosenthal, and D.B. Rubin, "Comparing correlated correlation coefficients," Psychological bulletin, vol.111, no.1, p.172, 1992.

77) H. Strik, J.v. Doremalen, J.v.d. Loo, and C. Cucchiarini, "Improving asr processing of ungrammatical utterances through grammatical error modeling," Proc. SLaTE, pp.109–112, 2011.

78) R. Pascanu, T. Mikolov, and Y. Bengio, "On the difficulty of training recurrent neural networks," Proc. ICML, pp.1310–1318, 2013.

79) I. Sutskever, O. Vinyals, and Q.V. Le, “Sequence to sequence learning with neural networks,” Proc. NeurIPS, pp.3104–3112, 2014.

80) K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, “Learning phrase representations using rnn encoder-decoder for statistical machine translation,” arXiv preprint arXiv:1406.1078, 2014.

81) J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” arXiv preprint arXiv:1412.3555, 2014.

82) D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” arXiv preprint arXiv:1409.0473, 2014.

List of publications

Journal Papers

- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Automatic assessment of English proficiency for Japanese learners without reference sentences based on deep neural network acoustic models," Speech Communication, Vol. 116, pp. 86-97, 2020.
- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Language modeling in speech recognition for grammatical error detection based on neural machine translation," Acoustical Science and Technology (Conditional acceptance).

International Conferences (refereed)

- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Evaluation of English speech recognition for Japanese learners using DNN-based acoustic models," in International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 93-100, 2018.

International Workshops

- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "DNN-based speech recognition of English spoken by Japanese learners," in the Proceedings of Tohoku-Germany Workshop, January 2018.
- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Evaluation of English speech recogni-

tion based on DNN acoustic models for Japanese learners," in the Proceedings of JCK Workshop, 4 pages, November 2018.

Domestic Publications

- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Evaluation of DNN-based Speech Recognition for English Spoken by Japanese Learners," in the Proceedings of Spoken Language Processing (SLP), Vol. 26, pp.1-5, 2018.
- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Automatic English Proficiency Assessment for Japanese Learners without Reference Transcriptions," in the Proceedings of Acoustical Society of Japan (ASJ Spring 2019), pp.1276-1278, 2019.
- Jiang Fu, Yuya Chiba, Takashi Nose, and Akinori Ito, "Automatic Generation of Text with Errors using Neural Machine Translation for Grammatical Error Detection," in the Proceedings of Acoustical Society of Japan (ASJ Autumn 2019), pp.837-840, 2019.