Phoneme Set Design for Second Language Speech Recognition

Xiaoyun Wang

ID No. 4G141101

January 2017

Graduate School of Science and Engineering
Department of Information and Computer Science
Doshisha University
Kyoto, JAPAN

Thesis Committee:
Seiichi Yamamoto, Chair
Shigeru Katagiri
Graham Wilcock
Kazuhiko Takahashi
Masashi Okubo

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

To my beloved parents and families
for their support and endless love

Acknowledgments

> *Gratitude bestows reverence, allowing us to encounter everyday epiphanies, those transcendent moments of awe that change forever how we experience life and the world.*
>
> John Milton, *English poet and polemicist*

First and foremost, I would like to express my deepest gratitude to my excellent advisor Professor Seiichi Yamamoto, for his encouragement, selfless support and continual guidance throughout my Ph.D. journey. I started my research journey with him and found what I am interested in. His insights, meticulous reading, editing of this dissertation and every other publication resulting from this research, have definitely improved the quality of my work. Seiichi always provides me freedom to investigate and explore this field, guides my future directions. Without my dear advisor, I might not be inspired with this study, not enjoy it and be excited for this long journey.

I must thank Prof. Jinsong Zhang for his encouragement and help on the earlier stage of this research. His technical expertise and detailed understanding of speech recognition systems, which have made it possible for the development of the first Japanese English speech recognition system during my early study. I also want to thank the other members of my supervisory committee: Prof. Masafumi Nishida, Prof. Tsuneo Kato, Prof. Masuzo Yanagida, Prof. Shigeru Katagiri, Prof. Graham

Wilcock, Prof. Kristiina Jokinen, Dr. Ichiro Umada, Dr. Xugang Lu, Dr. Peng Shen. Thanks to their multi-disciplinary expertise and insightful advice.

My thanks also go to the comprehensive doctoral program – Global Resource Management (GRM) of Doshisha University. It led me to have had additional opportunities to learn about social science based on practical knowledge and experience, widened my sight of the world, elongated my tentacles to different disciplines, and heightened my foothold in career promotion, besides pursuing a pure academic life. The financial support of visiting University of Helsinki, UCLA and Indiana University broadened my horizon in many ways as well as providing me with the most useful feedback and extension for my research.

I also want to thank my undergraduate advisor Prof. Terumasa Ehara, who brought me into the speech and language processing area and motivated me to continue my study.

Then, I would like to thank all of my friends and research colleagues who helped me collect the speech corpus to complete this research work. Thank you for everything you have done to make my nine years of life in Japan so memorable.

Last but not least, I would like to express my deep gratitude to my dearest Mom and Dad, to my husband for their unconditional love, enormous support and encouragement in all aspects of my life. Without the constant love and support from my dear families, I will not have been free to chase my dream until now.

Xiaoyun Wang, Kyoto

Abstract

In today's environment of rapid globalization, people have increasing opportunities for speaking in foreign languages, and the ability to communicate in foreign language is more important than ever. Various applications (dialogue-based computer assisted language learning (CALL) systems, car navigation system, hotel reservation systems, mobile platforms, etc.) are incorporating spoken language interfaces for non-native speakers to provide more convenient life. The key role for these kinds of interfaces is non-native automatic speech recognition (ASR) or second language (L2) speech recognition.

Non-native speakers usually have a limited vocabulary and a less than complete knowledge of the grammatical structures of the target language. This limited vocabulary forces speakers to express themselves in basic words, making their speech sound unnatural to native speakers. In addition, non-native speech includes less fluent pronunciation and mispronunciation even in cases in which it is well composed without grammatical errors. Therefore, non-native speakers represent a significant challenge for state-of-the-art ASR.

Two major problems need to be addressed for non-native ASR: (1) For given speech with limited vocabulary and less knowledge of grammatical structures, how can a speech recognizer take advantage of these characteristics of the speakers? (2) For given speech with different pronunciation variations, how can a system recognize speaker's utterance correctly? In order to tackle the above problems, this dissertation proposes to derive a customized phoneme set that is different from the canonical one but suited to non-native speech, particularly when the mother tongue of users is known. In addition, we build a proficiency-dependent phoneme set to capture their

different pronunciation variations, based on the analysis of efficiency of the derived phoneme set for non-native speakers with different proficiency levels.

The dissertation focuses on several important aspects for the above two problems: *the phonological knowledge of differences between mother tongue (L1) and target language (TL), acoustic and linguistic features of non-native speech, proficiency of non-native speakers.* The first part of the dissertation proposes the statistical method using integrated acoustic and linguistic features on the phonetic decision tree (PDT) to derive the phoneme set for L2 speech recognition. As the results of the first part of the dissertation show, the effect of the derived phoneme set is different depending on the speakers' proficiency in L2. To further improve the second language ASR, the second part of the dissertation investigates the relation between proficiency of speakers and a derived phoneme set customized for them. The investigated results are then used as the basis of a novel speech recognition method using a lexicon in which the pronunciation of each lexical item is represented by multiple phoneme sets for each L2 speaker with various proficiency levels.

The dissertation verifies the efficacy of the proposed methods using second language speech collected with a translation game type dialogue-based English CALL system. In conclusion, the dissertation shows that a speech recognizer with the proposed methods that is able to alleviate the problem caused by confused mispronunciation by L2 speakers. As a result, the ASR system can achieve a higher recognition accuracy with the derived phoneme set than that with the canonical phoneme set which is used in the traditional English speech recognition system.

Keywords

Automatic Speech Recognition (ASR)

Acoustic Likelihood

Computer Assistant Language Learning (CALL) Systems

Integrated Acoustic and Linguistic Features

Proficiency Dependent Reduced Phoneme Set

Phonetic Decision Tree (PDT)

Reduced Phoneme Set (RPS)

Second Language (L2)

Second Language Speech Recognition

Target Language (TL)

Linguistic Discrimination Ability

Language Proficiency

Multiple Reduced Phoneme Sets

Mother Tongue (L1)

Non-native Speech Recognition

Unified Acoustic and Linguistic Objective Function

List of Abbreviations

AM	Acoustic mode
ASR	Automatic speech recognition
ATR	Advanced Telecommunications Research Institute International
CALL	Computer assisted language learning systems
CD	Context-dependent
CI	Context-independent
ERJ	English read by Japanese
GMM	Gaussian mixture model
HTK	Hidden Markov model toolkit
HMM	Hidden Markov model
IPA	International phonetic alphabet
JPE	Japanese pronunciation of English
L1	Native language or mother tongue of speakers
L2	Second language of speakers
LL	Log likelihood
LM	Language model
LPC	Linear Prediction Coding
MFCC	Mel-Frequency Cepstral Coefficients
PDF	Probability density function
PDT	Phonetic decision tree
PLP	Perceptual linear prediction
ROVER	Recognizer output voting error reduction
RPS	Reduced phoneme set
SLA	Second language acquisition

TIMIT	Acoustic phonetic continuous speech database created by Texas Instruments and MIT
TL	Target language
WAR	Word accuracy rate

Contents

List of Figures

List of Tables

1 Introduction

> *Novel intelligent recognition strategies and the incorporation of knowledge about human speech communication that so far has been unknown or ignored.*

Alex Weibel, *Director of interACT & Professor at Carnegie Mellon University*

The rapid progress in transportation systems and information technologies has increased the opportunities for worldwide communication. People have more opportunities than ever before for speaking in foreign languages in addition to their mother tongue (L1) [1], [2], [65]. Meanwhile, automatic speech recognition (ASR) systems, machines that can automatically recognize words spoken by human being (Figure 1.1), are being applied increasingly rapidly in modern communication [24]. For example, car navigation systems usually offer drivers the possibility to search for a destination globally in English. Call centers are used for shopping websites, banks or other kinds of companies, where ASR based systems provide information guidance or help customers queries in 24 hours a day to process all calls. Smart home devices, such as Amazon Alexa[1]and Google Home[2], are voice-activated assistants that provide more convenient life, which can add an event to the calendar, set temperature of thermostat, stream customers' favourite song, going so far as to build in smart home capabilities. Moreover, in many of these applications of ASR, there are cases when the speaker is not expressing themselves in their L1.

[1]https://developer.amazon.com/alexa
[2]https://madeby.google.com/home/

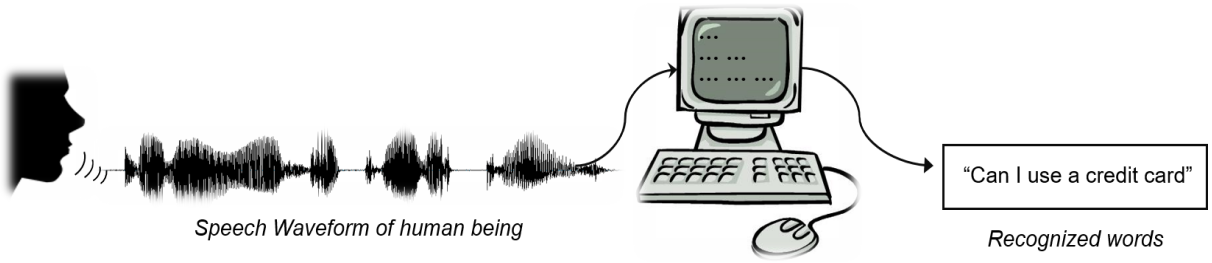


Figure 1.1: An ASR system that recognizes the speech waveform of a human utterance as "Can I use a credit card".

Despite a number of successful systems in industry (e.g. BMW's car navigation systems[3], Hotel reservation systems used in "Hotel.com" [4], and Advanced Telecommunications Research Institute International (ATR) computer assisted language learning (CALL) system – BRIX system[5]), speech recognition system are still very brittle when the speech is produced by second language (L2) speakers [15]. Therefore, when developing an ASR system for L2 speakers, recognition of their speech is still a challenging task even for state-of-the-art ASR systems. The biggest challenges of L2 speech recognition therefore result from speakers' various mispronunciation, limited vocabulary, incorrect grammar and low proficiency in speaking skill.

In this introductory chapter, we first introduce how an ASR system is developed for L2 and articulate the research problem existing in the current development procedure. We then introduce the general goal and contributions of this thesis to address the challenges. Finally, we give a chapter by chapter review of the rest of this dissertation.

1.1 Introduction

People from different countries usually speak a second language with various accents, partly depending on their L1 and L2. In comparison to native speakers, non-native speakers have different pronunciation affected by their L1 [3], [4], [5], [6], [75], [78], confused pronunciation [9],

[3]http://www.bmw.com/com/en/insights/technology/technology_guide/articles/navigation_system.html
[4]http://www.irishexaminer.com/examviral/technology-and-gaming
[5]http://www.atr-lt.jp/products/brix

[10], less knowledge of grammatical structures [10], [66] and a smaller vocabulary size [11], [12]. These issues result in non-native speakers delivering mispronunciation or less fluent pronunciation, confusing listeners with far-fetched sentences, and expressing themselves in limited vocabulary [41]. In addition, the limited vocabulary usually forces speakers to express themselves in basic words, making their speech sound unnatural to native speakers. Celce-Murcia et al. [80] showed that it is difficult to communicate effectively without correct pronunciation because different phonetics and prosody render speech sounds unnatural to native speakers and impede comprehension of the utterance.

A speech recognition system traditionally recognizes words as the phoneme sequences defined in the pronunciation dictionary. However, it is difficult to match a non-native speaker's utterances correctly which deviate from the standard pronunciations, because of insertions, deletions, and substitutions of phonemes or incorrect grammar.

A kind of dialogue-based CALL system [13], [14], [30], [36], [39], acting as automated interlocutors that prompt learners to produce speech in the target language (TL), was used to collect a transcribed conversation of a non-native speaker. The following example of transcribed conversations was produced by a native speaker of Japanese who was an undergraduate student in a Japanese University and learned English as L2. This conversation was picked from a dialogue-based CALL system that we developed [30], [36], [39] which will be described in Chapter 4 in detail.

Shopping scenario abroad:

(Here, "System" acts as a shop assistant and "Customer" is the non-native speaker)

System: May I help you?

Customer: Wow, this is a cute, would you show me this red item?

System: Yes, of course, would you like to try it on?

Customer: Yeah, is there match my size?

System: What size do you wear?

Customer: Medium Japanese.

System: OK, I'll bring one right over.

System: Let's see how this one looks on you.

Customer: ei..to.. (Japanese)... Okay, where is a clothes room?

System: Right this way.

System: How is it?

Customer: I like this it very much. Is this jacket made from truth kawa ("kawa" is a Japanese word and means "leather" in English)?

System: ... ... (System can not understand and recognize what is "kawa" actually.)

This CALL system allows learners to utter spontaneous utterances, through simulating conversation in real-life situations, such as a shopping scenario. But the quality of L2 speech usually differs significantly from L1 speech in terms of phonemes, prosody, lexicon, disfluencies, [15], [16], [17], [27], and also covers an enormous range of proficiencies [64] and speech types. In the above conversation, "ei..to.." is frequently used in Japanese when speakers think about and prepare their description; "kawa" is a Japanese word which has the meaning of "leather" in English. These phenomena are usually caused by the interference from the L1. This interference may lead to incorrect recognition of L2 by L1 speaker, because they will interpret the sound system of L2 based on their L1 phonology. We will discuss this in more detail in Section 2.2. The following conversation is the same as the above one, but with response by native speakers.

Shopping scenario abroad:

(Here, "System" acts as a shop assistant and "Customer" is the native speaker)

System: May I help you?

Customer: Wow, it's so cute, can I see that red one?

System: Yes, of course, would you like to try it on?

Customer: Yes, do you have something in my size?

System: What size do you wear?

Customer: I wear a Japanese medium.

System: OK, I'll bring one right over.

System: Let's see how this one looks on you.

Customer: Thank you! Where is the dressing room?

System: Right this way.

System: How is it?

Customer: I really like this. Is this jacket made from genuine leather?

System: ... ... (Continue the conversation)

Comparing the spontaneous speech by non-native and native speakers, we could find that non-native speaker's speech includes many grammatical errors and the combination of recognition errors. These must render weaker the feedback function of CALL systems aiming to correct grammatical errors and infelicitous phrases. The errors caused by different levels when comparing non-native and native speech cover an enormous range of talkers' styles and various proficiency levels. These problems would cause deterioration in the effectiveness of general speech-driven applications, such as question-answering systems and spoken dialogue systems.

1.2 The Problem

Human beings can understand phonologically confused L2 speakers' speech by guessing an intended spoken word and sometimes correcting it from the context-dependent (CD) conversation after the listener gets used to the style of the talker, i.e., the various insertions, deletions, and substitutions of phonemes or incorrect grammar [24]. However, this function is beyond the ability of even state-of-the-art ASR technologies that only exploit a short-range language model to predict the words that follow, and as a result, the performance of the ASR deteriorates for L2 speech. The dissertation mainly focuses on the following problematic issues:

The first problematic issue is that it is difficult to collect enough L2 speech to train the acoustic model (AM) directly in L2 speech recognition. It continues to be the case that transcriptions of L2 speech are less available than those of native speech, and it is extremely difficult to collect enough data on each conversation topic by a considerable number of L2 speakers with various language proficiencies.

The second problematic issue is when L2 speakers' confused pronunciations become an issue for spoken dialogue systems that target tourists, such as travel assistance systems, hotel reservation systems, and systems in which consumers purchase goods through a network. Meanwhile, the vocabulary and grammar of L2 speakers is often limited and simple, but a speech recognizer takes no or only a little advantage of this and is confused by the different phonetics.

The third problematic issue is that the speech quality of L2 speakers overall depends on their proficiency levels in TL [20], [21], and there are different patterns in accent among inexperienced, moderately experienced, and experienced speakers [23], [25]. As a result, influence of the L1 on their pronunciation is unavoidable when building a second language ASR system, because this dramatically degrades the speech recognition performance.

Hence, the recognition of second language speech presents several challenges for ASR. The target of this thesis is to provide a method that can adjust an ASR system to be more tolerant to the acoustic and linguistic variations produced by L2 speakers, so that it could recover some of the errors caused by their pronunciations.

1.3 Thesis Statement

As revealed by many second language acquisition studies [3], [4], [5], [6], [75], [78], pronunciation by L2 speakers is usually significantly influenced by the mother tongue of the speakers. Therefore, speech recognition performance for L2 speakers, especially when the mother tongue of speakers is known, can be significantly improved by the method considering knowledge of phonological differences between L1 and L2, the different phonetic features, and the relations

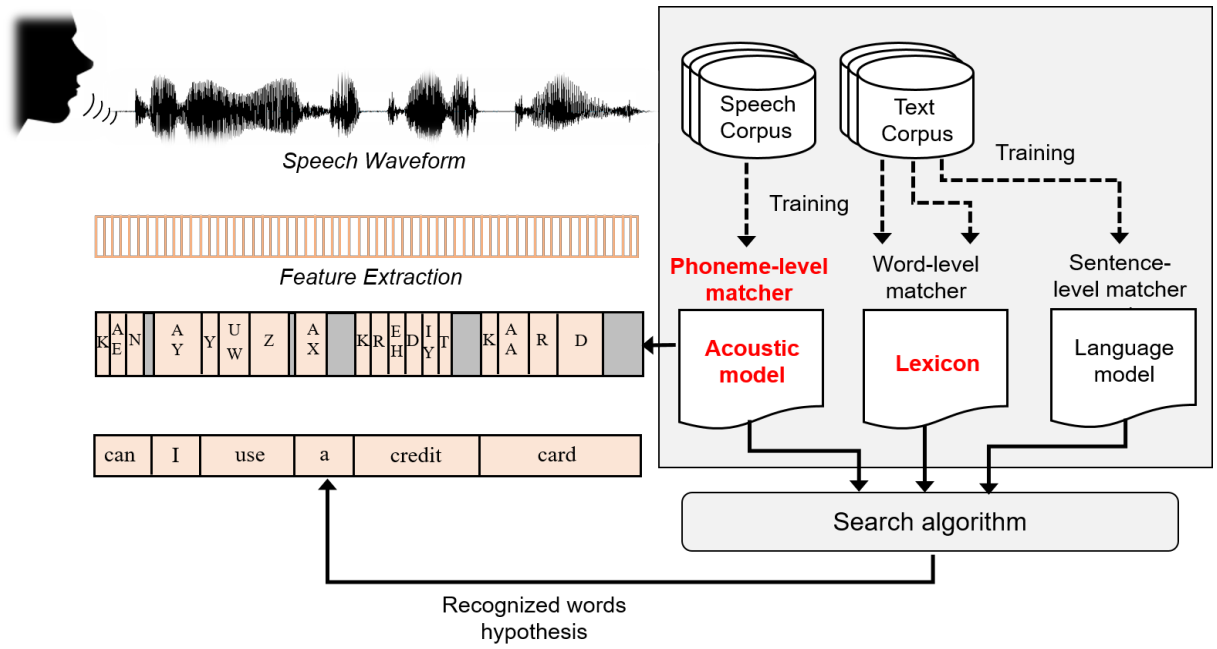


Figure 1.2: A generic ASR system, composed of five components: feature extraction module, acoustic model, pronunciation lexicon, language model and search algorithm.

between acoustic features and linguistic features of L2.

The main purpose of this work is to develop the acoustic model with customized phoneme set for second language ASR systems that is useful for phoneme-level matching despite speech with confused pronunciation or mispronunciation by L2 speakers. An example of the corresponding phoneme set used in a generic ASR system, as shown in Figure 1.2, is composed of five components – feature extraction, acoustic model, pronunciation lexicon, language model and search algorithm. The parts marked in red colour are the main target of this document. Based on the acoustic models with customized phoneme set, we then re-construct the pronunciation dictionary with constraints for recognizing speech by L2 speakers.

For the customized phoneme set, there are two close relations considered at the first step of this work – integrated **acoustic features** and **linguistic features**. The proficiency of L2 speakers varies widely, as does the influence of L1 on their pronunciation. As a result, the effect of the customized phoneme set is different depending on the speakers' proficiency in L2. In order to further improve the customized phoneme set, there are two important stages to investigate

– **the relation between proficiency and the phoneme set** and **multiple-pass decoding** with proficiency-dependent phoneme sets.

In conclusion, the thesis demonstrates the validity of building a second language ASR system with the acoustic model based on the customized phoneme set for L2 speakers. The main contribution of the dissertation is presenting the potential for reducing the mismatch on the phoneme-level for confused pronunciation or mispronunciation and showing the feasibility of improving the scalability and efficiency of ASR system development for L2 speakers when the mother tongue of the speakers is known.

1.4 Thesis Structure

The dissertation is organized as follows:

■ **Chapter 2 – Background and Related Work**

This chapter gives a brief introduction to the architecture and components of L2 speech recognition. Subsequently, we review background knowledge and survey the history of the studies related to general second language acquisitions (SLA) as well as research in speech recognition for L2 speakers. Related works in the domain of L2 speech recognition and the focus of this work are presented at the end of this chapter.

■ **Chapter 3 – Phonology of Japanese English**

This chapter focuses on the introduction of specific domains in Japanese English which are related to our focus on knowledge resources. The remainder of this chapter is divided into two sections – accent of Japanese English and pronunciation variation of Japanese English.

■ **Chapter 4 – Japanese English Speech Database**

This chapter describes speech databases of Japanese English that are used for model training and testing for the baseline system and the systems using our proposed methods. Part of this research related to our data collection has been presented in the following publications [39], [36] :

– **Xiaoyun Wang**, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, "A Dialogue-Based English CALL System for Japanese." *in Proceedings of the 12th National Conference on Man-Machine Speech Communication (NCMMSC 2013)*, Guiyang, China, 2013.

– **Xiaoyun Wang**, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, "Phoneme Set Design Using English Speech Database by Japanese for Dialogue-Based English CALL Systems." *in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

■ Chapter 5 – Phoneme Set Design with Integrated Acoustic and Linguistic Features

This chapter describes a method for L2 speech recognition incorporating integrated acoustic and linguistic features into the recognizer, which corresponds to the goal of recovering errors caused by the pronunciation of L2 speakers. Some related contributions have been presented in the following publications [67], [68]:

– **Xiaoyun Wang**, Tsuneo Kato, and Seiichi Yamamoto, "Phoneme Set Design Considering Integrated Acoustic and Linguistic Features of Second Language Speech." *in Proceedings of The 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, USA, 2016.

– **Xiaoyun Wang**, Tsuneo Kato, and Seiichi Yamamoto, "Phoneme Set Design Based on Integrated Acoustic and Linguistic Features for Second Language Speech Recognition." 2016. Submitted to *IEICE TRANS. INF. & SYST.*

■ Chapter 6 – Analysis of the Effect of Acoustic and Linguistic Features

This chapter focuses on analysis of the effect of acoustic and linguistic features in detail. The analysis of the efficiency of splitting methods and phoneme occupation probabilities was used to verify the effectiveness of acoustic features; the recognition performance of phoneme sets by different methods and detailed phonemes in final clusters using different methods were investigated to verify the effectiveness of linguistic feature. Some related publications [87], [36], [33] have been published as follows :

– **Xiaoyun Wang**, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, "Efficient Phoneme Set Design Using Phonetic Decision Tree in Dialogue-Based English CALL Systems for Japanese Students." *IEICE Technical Report*, Tsukuba, Japan, Vol. 113, No. 366, pp.47-51, Dec 2013.

– **Xiaoyun Wang**, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, "Phoneme Set Design Using English Speech Database by Japanese for Dialogue-Based English CALL Systems." *in Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)*, Reykjavik, Iceland, 2014.

– **Xiaoyun Wang**, Jinsong Zhang, Masafumi Nishida, and Seiichi Yamamoto, "Phoneme Set Design for Speech Recognition of English by Japanese." *IEICE TRANS. INF. & SYST.*, 98(1):148–156, 2015.

■ Chapter 7 – Analysis of the Relation Between Language Proficiency Level and the Customized Phoneme Set

This chapter investigates the relation between the language proficiency and customized phoneme sets by the proposed method. Part of this work has been presented in the following publications [89], [72]:

– **Xiaoyun Wang**, Jinsong Zhang and Seiichi Yamamoto, "Multiple reduced phoneme sets for second language speech recognition." *in Proceedings of Acoustical Society of Japan (ASJ Autumn 2014)*, Hokkaido, Japan, 2014.

– **Xiaoyun Wang**, and Seiichi Yamamoto, "Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets." *in Proceedings of The 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, 2015.

■ Chapter 8 – Multiple-Pass Decoding with Lexicon Represented by Multiple Phoneme Sets

This chapter focuses on the acoustic modeling with consideration of different proficiency levels of L2 speakers. Some of these contributions have been presented in the following publications

[72], [73]:

- **Xiaoyun Wang**, and Seiichi Yamamoto, "Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets." *in Proceedings of The 16th Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, 2015.
- **Xiaoyun Wang**, and Seiichi Yamamoto, “Speech Recognition of English by Japanese using Lexicon Represented by Multiple Reduced Phoneme Sets." *IEICE TRANS. INF. & SYST.*, 98(12):2271–2279, 2015.

■ Chapter 9 – Conclusion and Future Work

This chapter summarizes the main contribution of this dissertation and discusses interesting directions for future work.

2 Background and Related Work

> *When we study human language, we are approaching what some might call the human essence, the distinctive qualities of mind that are, so far as we know, unique to humans.*

Avram Noam Chomsky, *American linguistic, philosopher, cognitive scientist, historian, social critic, and political activist.*

Automatic speech recognition (ASR) systems which convert speech from a recorded audio signal to text are being incorporated in several types of device (e.g. car navigation systems, call centres, smart phones, smart home devices, language learning systems and travel assistance systems), especially using second language speech. This chapter starts with a brief introduction to the architecture and components of typical speech recognition and L2 modeling in ASR in Section 2.1. Section 2.2 gives a discussion of second language acquisition (SLA), which has influenced the approach to statistical modeling of L2 speech. Section 2.3 presents related work in the domain of L2 speech recognition. The focus of this work is briefly introduced at the end of this chapter.

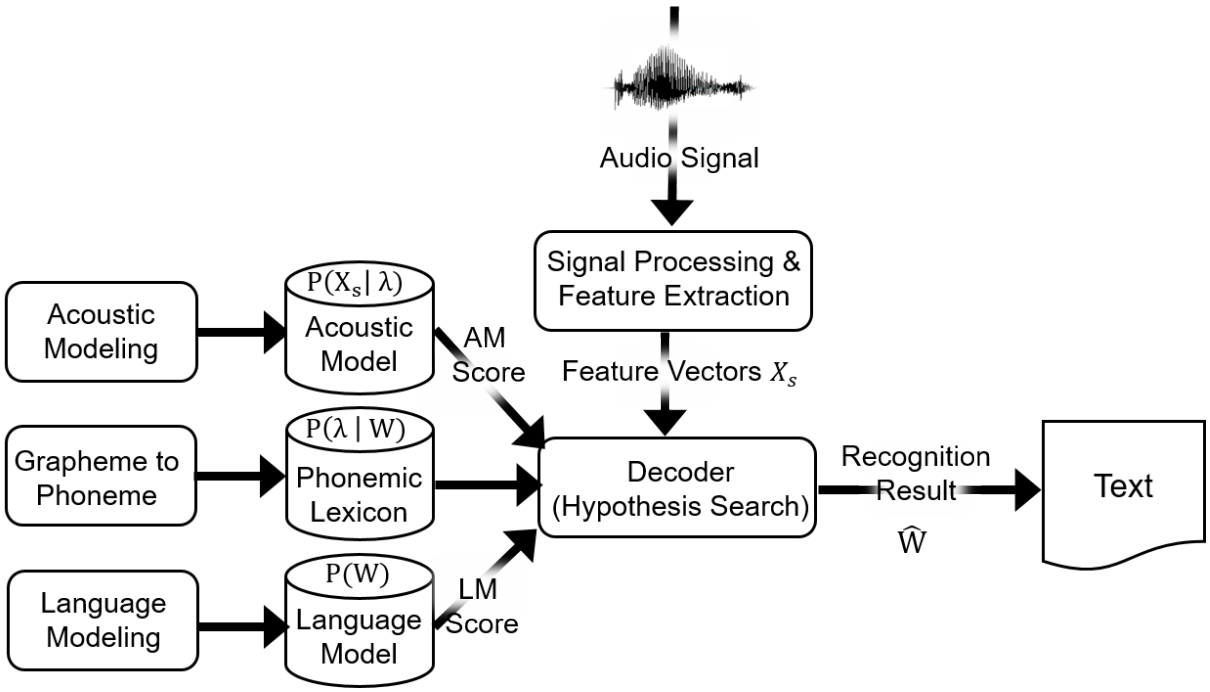


Figure 2.1: A typical architecture of speech recognition systems.

2.1 ASR for Second Language Speakers

2.1.1 Architecture of a Speech Recognition System

An ASR system aims to infer the input words given the observed speech signal and deliver an output text transcription. The speech signal could correspond to any single word or word sequences in the vocabulary with a certain probability. This process can be regarded as mapping a speech signal into a set of meaningful strings of words, and can be performed using multiple level pattern recognition, since the acoustic speech signals can be structured into a hierarchy of speech units such as sub-units (phonemes), words, and word sequences (sentences) [92]. Therefore, a score can be calculated for this matching process when assuming a single word w or sequence of words W was spoken, and the calculation is based on the acoustic properties of sub-units (acoustic model (AM) score) and linguistic knowledge by learning from the correlation of words in the context (language model (LM) score).

The basic architecture of an ASR system is illustrated in Figure 2.1. The process of speech recognition can be divided into the following components: feature extraction module, acoustic model (AM), lexicon, language model (LM), and decoder (hypothesis search). In the following subsections, we will briefly look at each components in turn.

2.1.2 Feature Extraction

Pre-processing aims to transform the input signal speech into a constrained network of hypothesized segments, in order to limit and reduce the size of the search space. The process usually includes pre-emphasis, filtering, sampling, and quantization. 16 kHz as a sufficient sampling frequency is used to represent human speech intelligibly.

Feature extraction aims to process the speech signal into a set of observation feature vectors X_s by removing redundant information such as the fundamental frequencies or noise. These observation features are extracted over time frames of uniform length. The length of frames is typically around 25 msec for the calculated acoustic samples which are in the window with multi-dimensional feature vectors. The time frames are typically overlapping and shifted by 10ms.

There have been many variants of feature extraction techniques, e.g., Linear Prediction Coding (LPC) [52], cepstrum analysis [56], and perceptrual linear prediction (PLP) [93]. The techniques based on Mel-Frequency Cepstral Coefficient (MFCCs) [82] are the most widely used feature extraction techniques in speech recognition.

2.1.3 Acoustic Modeling

Acoustic modeling aims to estimate the acoustic probability $P(X_s|\lambda)$ that the observation feature vectors X_s have been generated by the sub-unit (phoneme) model λ and generates an AM score for the variable length feature sequences. It also integrates knowledge between phonetics and

acoustics, and takes as input the features extracted from the last processing step mentioned in Section 2.1.2.

Building a robust acoustic model is one of the main challenges for second language speech recognition. Two issues when dealing with acoustic features for L2 speech by the AM components are variable length feature vectors and variations in the speech signals. The variations in the speech signals by L2 speakers usually manifest at less fluent pronunciation, mispronunciation, and variability of pronunciation by L2 speakers, when considering that read speech without grammatical errors produced by them has only different acoustic features in comparison to that by native speakers. In addition, Huang et al. [81] showed that context variability could appear at the sentence, word and phonetic level. Therefore, mispronunciation can exist in sub-units (phonemes) when it is realized in different contexts at the phoneme level. The variability of pronunciation is most noticeable given the various language proficiency on a speaker-by-speaker basis.

Hidden Markov Models (HMMs) are typically used as representations of acoustic models, where the short-term speech spectral characteristics are modeled with HMM state probability, while the temporal speech characteristics are governed by HMM state transitions [63].

2.1.4 Pronunciation Lexicon

Pronunciation lexicon aims to estimate the word probabilities $P(\lambda|W)$ when gives the sub-unit (phoneme) sequences generated by its model λ. Acoustic models define speech units using phonetic features which are related to articulators (mouth, tongue or vocal tract) and others from speech. The lexicon describes all word items in the vocabulary defined by phoneme sequences for the pronunciation [63], [24], which is generated by the implementation of a lexical tree-based search [19]. Based on the definition of lexical tree-based search, an example of a lexicon tree using sub-unit (phoneme) is illustrated in Figure 2.2. Here, the lexical tree includes the words along with the pronunciation dictionary including the multiple pronunciations of words.

If there is no definition of the manner of pronunciation, the rules for converting graphemes to

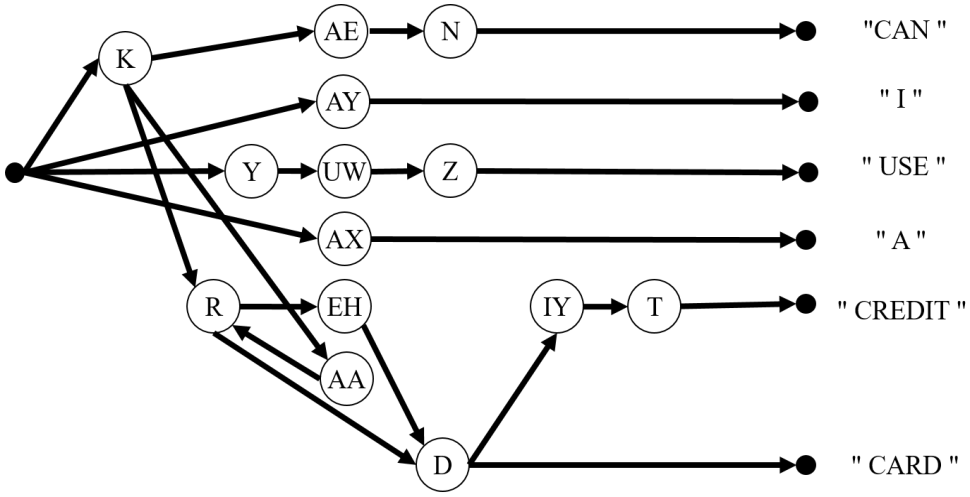


Figure 2.2: An example of a phoneme tree-based pronunciation lexicon for word sequences "Can I use a credit card".

phonemes have to be created based on the knowledge of the language characteristics [95]. The words used in Figure 2.2 can be typically modeled in the English pronunciation lexicon as follows:

CAN	K AE N
I	AY
USE	Y UW Z
A	AX
CREDIT	K R EH D IY T
CARD	K AA R D

Variations of pronunciation for L2 speaker and multi-pronunciations can also be added into the pronunciation lexicon as follows:

CAN	K AE N
CAN	K AX N
USE	Y UW S
USE	Y UW Z

A	AX
A	EY
CREDIT	K R EH AX IY T
CREDIT	K R EH AX AX T
CREDIT	K R EH AX IH T
CREDIT	K R EH D IY T
CREDIT	K R EH D AX T
CREDIT	K R EH D IH T
CREDIT	K R EH DH IY T
CREDIT	K R EH DH AX T
CREDIT	K R EH DH IH T
CREDIT	K R EH T IY T
CREDIT	K R EH T AX T
CREDIT	K R EH T IH T

2.1.5 Language Modeling

Language modeling aims to define rules that arrange the competent use of a language such as syntax and morphology to present the language grammar. The ways to represent the language grammar can be divided into formal language model and stochastic language model, which are the approach based on linguistic knowledge and the statistical data-driven approach, respectively. The most popular approach is the latter one, such as N-gram model, which typically estimates the prior probability $P(W)$ of hypothesized word sequences $W = w_1, w_2, ..., w_n$ or LM score, by learning the correlation between words from a text corpora.

The LM score can often be estimated more accurately if prior knowledge about the limited do-

main or tasks is known. Using the Bayes' rule of probability theory, the prior probability $P(W)$ can be formally decomposed as

$$P(\boldsymbol{W}) = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \tag{2.1}$$

where $P(w_i|w_1, w_2, ..., w_{i-1})$ is the probability that the word w_i will be spoken given that the word sequences $w_1, w_2, ..., w_{i-1}$ were said previously. Here, the previously spoken $w_1, w_2, ..., w_{i-1}$ is frequently referred to as a history of the word sequences.

2.1.6 Decoding

Hypothesized search based decoding aims to determine the most probable string of words $\hat{W}$ among all possible word strings given the observation feature vector X_s. This component combines AM score (acoustic probability score $P(X_s|W)$) and LM score (prior probabilities of each utterance $P(W)$) given the feature vector sequences and the hypothesized word sequences, respectively. Finally, the word sequences with the highest score were outputted as the recognition result.

During the hypothesized search process, the words $\hat{W}$ that fit best to the observation vector X_s were given in the following fundamental equation:

$$\hat{W} = \arg\max_{W}(P(X_s|W)P(W)) \tag{2.2}$$

with $P(W)$ is from the LM and $P(X_s|W)$ is calculated by the sub-unit (phoneme) sequences generated by its model λ and used to define word W in the lexicon:

$$P(X_s|W) = \prod_{i} P(x|\lambda_i)P(\lambda_i) \tag{2.3}$$

In theory, it is necessary to consider $P(x)$, but as this term is the same for all competing hypotheses, and it can be omitted for the computing.

2.2 Second Language Acquisition

Second language acquisition (SLA) is the process by which people attain a second or additional language in addition to their mother tongue (L1). The simplest answer for "how do people learn a second, a third or more languages?" is "with great difficulty". The book edited by Gass [83] discusses the questions from a variety of perspectives, such as why most L2 learners cannot easily achieve the same degree of proficiency in a new language as they do in their L1; why some individuals usually appear to achieve native-like proficiency in more than one language; how learners create a new second language learning system with only limited exposure.

However, there are many reasons why it might be important to understand how second languages are learned. One is that SLA has a closer relationship with other relevant fields, such as development of language learning systems, ASR or spoken dialogue systems. Before we look at L2 speech recognition, it is important to first understand how learners master a new language (Q1), and how the differences between L1 and L2 affect the ASR system (Q2), so that approaches can be proposed to improve the recognition performance for L2 speakers.

2.2.1 Q1: How do learners master a new language?

Language acquisition basically involves the study of phonology, pronunciation, grammars, and vocabulary. Therefore, most SLA studies discuss strategies that L2 learners need in order to overcome the difficulty of language acquisition under four major subject headings – *phonology, morphology, syntax, and lexicon* – of a language. These subjects also interact with each other, for example a book edited by Paul [86] discusses the interaction of different subjects, e.g. the relation of syntactic phrases to phonological phrases, the relation of morphological to phono-

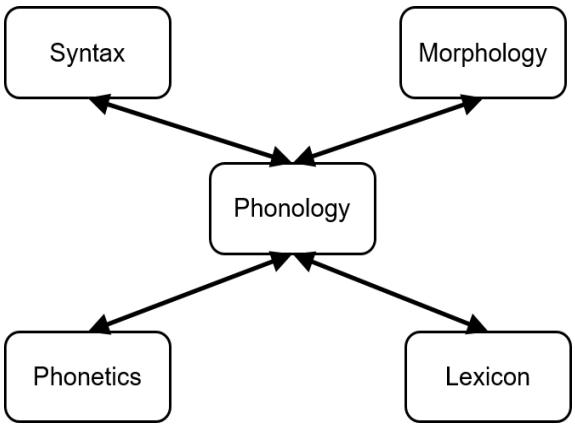


Figure 2.3: Interaction with different subjects of language acquisition.

logical structure, and the relation of phonological structures to phonetic ones, or the relation of phonological structures to lexical items (the relations can be seen in Figure 2.3).

(1) Phonology

Phonology can be regarded as a complex knowledge of people's sound system. It is used to distinguish different words, and shows what sounds are possible and impossible in a language, so that L1 speakers know that a word contains a sound that is not in their mother tongue. This knowledge is reflected in recognition as well as in production [83]. Flege [58] also showed the interference from the L1 may lead to incorrect perception of L2, because they will interpret sound system of L2 based on their L1 phonology. This interference usually results in incorrect speech production, and sometimes cannot be repaired. Therefore, it can be one of the reasons which affects the performance of recognition system.

On the other hand, phonological knowledge involves knowing what happens to word items in the fast speech as opposed to more carefully articulated speech [86]. For the native speech of American English, the speaker wanted to express the following idea (assuming a typical utterance called "UTT"):

UTT_1: I am going to write a letter.

that person would possibly say the sentences like the following:

UTT_2: I'm gonna wriDa leDer.

In the fast speech, speakers usually know when to combine or not to combine the words, but in clearer, more articulated speech we do not. For another example, consider the sound of /ŋ/ (phoneme "NG" used in speech recognition) which can appear at the end of the words but cannot appear in the beginning of the words in English, although it can be in other languages, such as Japanese.

(2) Morphology

Morphology is the study of word formation. In many cases, words are made up of more than one part [86]. For example the word "unforeseen" is made up of three parts: *un*, which has a negative function; *fore*, which means earlier in time; and *seen*, which means to visualize [83]. Each part is referred to as a morpheme, which can be defined as the minimal unit of meaning. In comparison to native speakers, morphology in L2 is a hard task because basic word order is typically non-problematic past the initial stage of acquisition [66], but even the most basic morphology is often lacking from the speech of classroom learners who can not monitor themselves effectively [76], [79].

(3) Syntax

Syntax of the language is frequently known as grammar referring to the knowledge that we have about the order of word items in a sentence. It is generally divided into *prescriptive grammar* and *descriptive grammar* [83]. The rules of prescriptive grammars are generally leaned at school and do not consider the way that a language is actually used by native speakers. On the other hand, linguists are concerned with descriptive grammars, in which they attempt to describe language as it is actually used. The rules of descriptive grammars should always be true, but native speakers may violate the prescriptive rules. L2 speakers usually have a less than complete knowledge of the grammatical structures of the target language [10]. This limited vocabulary forces L2 speakers to express themselves in basic words, making their speech sound unnatural to native speakers and deviating from native speakers.

(4) Lexicon

Lexicon usually refers to the vocabulary in language acquisition, and aims to describes the pronunciation of all words in the vocabulary. L2 speakers usually have a limited vocabulary size of

the target language in comparison to native speakers. This limited vocabulary forces speakers to express themselves in basic words, and making their speech sound unnatural to native [24]. It is difficult to communicate effectively, when L2 speakers give incorrect usage or pronunciation of vocabulary which affects actual meaning. Such as an example mentioned in the "Introduction", Japanese customer said the following sentence:

System: How is it?

Customer: I like this it very much. Is this jacket made from truth *kawa* ("kawa" is an Japanese word which is the meaning of "leather" in English)?

For native speakers, they can eventually understand speech by non-native even though with grammatical errors. But for incorrect vocabulary, especially when cause by interference from L1, such as "kawa", which do not appear in TL, native speakers who don't understand Japanese will not get the meaning of their speech.

2.2.2 Q2: How do the differences between L1 and L2 affect the ASR system?

Focusing on the relationship between language acquisition and pronunciation produced by L2 speakers, we start the discussion from the question – *Does speaker's pronunciation of L1 penetrate in their L2 pronunciation?* – that affects the weaker speech recognition performance of L2 due to substitution of L1 phonetic units to L2 phonetic ones.

At the early stage of research about the relationship between in L1 and L2, Patkowski [70] presented data suggesting that the strength of foreign accents increases greatly if L2 learners begin to master a language after the age of 15 years. Poulissee et al. [4] investigated the study undertaken to provide data relevant to modeling development of bilingual speech production. They found L1 use in L2 production, and showed that L2 speakers usually rely on their L1 syllables or rules when they produce an unfamiliar pronunciation. The development errors that happen with L2 learners are also observed in language acquisition by children of L1 [71]. Flege

et al. [3] discussed how the variation in amount of L1 use influences L2 and found that the degree of activation of the L1 or the strength of its representations may influence the accuracy of L2 production. In Fung and Liu's study [91], they expounded that the articulation habits of the speakers in their L1 produce most of the accent of their L2 pronunciation.

For complex and difficult pronunciations, both children of L1 and L2 learners may resort to simplify to familiar pronunciation by deleting, inserting, or substituting certain sounds with another one. However, the relation between age and L2 pronunciation accuracy remains unresolved until now, and the same is true for the fact that a speakers' pronunciation of L1 based on their articulation of phones is normal, but is different with their pronunciation of L2, especially in the phonetic level. It must be one of the reasons influencing the weaker performance of ASR systems by both L2 speakers and children of L1.

2.3 Related Works in the Domain of L2 Speech Recognition

The technologies to process the native speech have matured in state-of-the-art ASR. L2 speech as we seen in previous sections is different from native speech in terms of phonology, morphology, syntax, and lexicon. These differences give rise to more problematic issues when ASR systems process L2 speech. Hence, the recognition of L2 speech remains a challenging task even for state-of-the-art speech recognition technology. In order to make ASR systems more tolerant to variational speech produced by L2 speakers in comparison to that produced by native speakers, various methodologies have been proposed. They mainly aim to improve the system performance from three important components: acoustic model, language model, and pronunciation model. Mismatch caused by the negative transfer of the phonology of L1 speakers to L2 speech affects the performance of the target acoustic model and pronunciation model when recognizing speech by L2 speakers. Incorrect usage of vocabulary and grammars may also influence the performance of ASR systems.

Considering the relation between phonology and lexicon, Schaden [28] designed a set of post-

lexical rules to transform canonical phonetic dictionaries of L2 into adapted dictionaries for L1 speakers based on the consideration of each L1 and L2 pair. Finally, this study presented an extended lexicon adaptation method using a set of rewriting rules based on the study of phonological properties of the native language and the target language. Bouselmi et al. [40] developed a phonetic confusion scheme consisting in associating to each spoken phone several sequences of confused phones by non-native speakers. Then they used different combinations of acoustic models representing the canonical and the foreign pronunciation.

Considering the relation between phonology and syntax, Kenstowicz [26] drew a distinction between distinctive and redundant phonological properties.

Considering phonological structures, Livescu's work [27] used an acoustic model interpolating with native and non-native models in order to cover various pronunciations and accents and achieved 8.1% relative word error reduction on a non-native test set to the baseline system using models trained on pooled native and non-native data. Wang and Waibel [29] investigated how is bilingual model, speaker adaptation, acoustic mode interpolation useful for non-native ASR, and proposed an interpolation acoustic model based on the polyphone decision tree specialization method [96].

Considering the relation between phonology and phonetics, Oh et al. [32] proposed an acoustic model adaptation method for L2 speech with a variant phonetic unit obtained by analysing the variability of L2 speech pronunciation.

2.4 Focus of This Work

L2 speakers do not articulate their speech like native speakers, since their speech is often influenced by their native phonology. Therefore, speech recognition performance for L2 speakers, specially when the mother tongue of speakers is known, can be significantly improved by the method considering phonological knowledge among L1, L2 and their phonetic features, acoustic features and linguistic features of L2.

The focus of this work is to develop the acoustic model with customized phoneme set for second language ASR systems that is effective for phoneme-level matching despite speech with confused pronunciation or mispronunciation by L2 speakers. Based on the acoustic models with customized phoneme set, we then reconstruct the pronunciation dictionary constraints for recognition of L2. In this dissertation, we verified the efficacy of the proposed method for L2 speech recognition in Japanese English.

3 Phonology of Japanese English

> *A scientific research is a search after truth, and it is only after discovery that the question of applicability can be usefully considered.*

Henri Moissan, *Nobel Prize in Chemistry winner*

English as a global language is taught as an international communication tool in many countries. Japan is far behind in terms of introducing and delivering bilingual education, lets alone effective immersion programs in comparison to other countries in Asia [94].

English education in Japan has traditionally been focused on an articulatory phonetics approach. In non-native pronunciation, English spoken by Japanese is a representative case with familiar issues. Looking at Japanese pronunciation of English (JPE) can help us to understand the use of English all over the world. This is not only because some similar expressions of Japanese English are used in Korean English and other languages' English, but also because words and expressions classified in a certain way form a fascinating prism that we can use to see how English change when it is uttered by non-native speakers.

This chapter gives an introduction of a specific domain in JPE that relates to Japanese English ASR. It is divided into two parts: accent of Japanese English and pronunciation variation of Japanese English.

Table 3.1: English vowels versus Japanese vowels in international phonetic alphabet notation.

English vowels	Japanese vowels
/ɑ/, /æ/, /ʌ/	/a/
/ɪ/, /i/, /ɝ/	/i/
/ʊ/, /ʊ/	/u/
/e/, /ei/, /aɪ/	/e/
/o/, /ɔ/, /ɔɪ/, /aʊ/	/o/
/ə/, /ɚ/	

3.1 Accent of Japanese English

The issue of foreign accent is one of the most noticeable marks of SLA. According to Flege's study [69], the recognition of foreign accent is related to acoustic differences between native and non-native speakers' segmental articulations and supra-segmental levels, which are the main components of accent.

English is a stress-accent language that is expressed by a combination of pitch, duration, intensity, and vowel quality [97], [98], [99]. People from different countries have different accents because lexical accent differs among different languages. The type of accent variations mainly depends on the mother tongue (L1) and the target language (TL) [100], [101]. The accent can sometimes have adverse consequences for the L2 speakers [25], [102]. First, listeners might have difficulty understanding speech that differs from the patterns of their accustomed production. Second, interlocutors might respond negatively to accented speech because of impatience related to their inexperience with L2 speakers. In this section, we will look at Japanese accented English.

There are various word pairs that differ only in one sound, and for each accent specific word pairs can be found where the discriminating sound is especially difficult to produce for people with specific accents [24]. There are at least six new vowels when considering English spoken by Japanese in comparison to the five vowels in Japanese. The English vowels versus Japanese vowels in international phonetic alphabet (IPA) notation are listed in Table 3.1. For example, /æ/ is usually produced by Japanese with low intelligibility due to some spectro temporal properties

that differ from the sound produced by a native speaker [6], [61]. Aikawa et al. [61] investigated spectro temporal properties for the reduced intelligibility of vowels of Japanese English. Their experimental results showed that the pairs /i-ɪ/, /æ-ʌ/, and /ɪ-e/ were spectrally similar for Japanese English, in comparison with the pairs /e-ɜ˞/, /i-ɜ˞/, /ɪ-æ/, and /i-æ/, which were dissimilar. They mentioned that the most confusing pairs consist of /æ-a/, /æ-ʌ/, and /ʌ-a/ for Japanese accented English vowels. Vance [62] reviewed five Japanese vowels and showed that short /i/ and /u/ are often devoiced or deleted between voiceless sounds in Japanese. By contrast, English vowels vary widely in comparison to Japanese vowels, and have allophonic long forms before voiced consonants [31].

Another specific pronunciation *Schwa* has been the focus of many SLA studies, specifically, its occurrence in unstressed syllables that JPE may substitute for almost any other English vowel [85].

Considering English consonants spoken by Japanese, the most frequent mispronunciation is the discrimination between /l/ and /r/. The English consonant /b/ may sometimes be pronounced almost like /v/, and /v/ may also be pronounced as /b/ by Japanese [84], [85]. In order to identify more of the confusing and accented sounds of Japanese, it can be helpful to look at the pronunciation variation of Japanese English, as discussed in the following section.

3.2 Pronunciation Variation of Japanese English

The differences between how Japanese pronounce words that are derived from English and how native speakers pronounce those words can be investigated from the following viewpoints.

First, different pronunciation by Japanese and English can be considered in the syllable structure. According to Riney's study on JPE [6], Japanese and English syllables are respectively called "open" (i.e., ending in a vowel, which is characteristic of syllables in Japanese) or "close" (i.e., ending in a consonant, which is characteristic of syllables in English). Japanese is a syllabic language, which means all consonants apart from a final /N/ must be followed by a vowel. This

is the most noticeable difference when an English word finishes with a consonant that is not /N/, for example, "boat", which would be pronounced as "bouto", and "home", which would be pronounced as "houmu" [74]. Syllable structures are known as "phonotactics", and phonemes can be used to create order in a language [6]. Japanese and English present multiple contrasts in this consideration.

Second, different pronunciation with Japanese and English can be considered in terms of consonant clusters. Japanese is often described as a language without consonant clusters or obstruents like "stops" and "fricatives" at the end of a syllable. English usually presents a dramatic contrast: for example, there are 47 consonant clusters in the initial position and 169 consonant clusters in the final position for its syllable structure [31]. The most investigated consonant pairs in Japanese English are /r/ and /l/. In an early study of English acquisition by Japanese, Brown [77] and Basson [90] found that the English /r/ was one of the most difficult English consonants to speak for Japanese. Some consonants in Japanese have allophones, e.g., a nasal conventionally represented as /N/ to /m/ before /b/, /p/, /m/, and /n/ before /t/, /d/, /n/. Sheldon et al. [60] found that the pronunciation variation of Japanese English of /r/ more easily appears in the initial and intervocalic positions.

As for the other investigations, Esling and Wong [18] found that voice quality setting can be used to describe English accented by L2 speakers and to improve English pronunciation by L2 speakers. Vance's study [62] reviewed Japanese articulatory settings and concluded that "lip rounding" is weaker in Japanese than in English; "jaw position" is more open in Japanese than in English; and the tongue blade articulator is used in Japanese while the tongue tip articulator is used in English. These articulatory settings and voice quality strongly affect the pronunciation by L2 speakers.

Last, considering the "katakana" scripts in Japanese (also known as Katakana-English), there are many words borrowed from European languages that are written in the "katakana" scripts, which are a style of Japanese characters. These sometimes have an influence on the fluency of JPE. For example, Japanese often borrow "enquete" (the meaning of questionnaire or survey) from French, which is pronounced "anketto" in "katakana", and use it in English conversation.

Table 3.2: Examples of "katakana" in Japanese (loanwords) and corresponding words in English.

"katakana" word	English word
anchi	anti
bredo	bread
karaa	collar
kizzu	kids
mazaa	mother
meitaa	meter
monkii	monkey
shieta	theatre
saado	third
baiorin	violin

Table 3.2 shows some examples of loanwords written as "katakana" in Japanese along with the corresponding words in English.

4 Japanese English Speech Database

> *The goal is to turn data into information, and information into insight.*
>
> Carly Fiorina, *former executive, president, and chair of Hewlett-Packard Co.*

4.1 Introduction

A different challenge about non-native speech processing lies in prosodic effects and other non-phonetic differences. People from different countries speak English with different accents, depending mainly on speaker's L1 and L2 [24], especially between different language systems. Therefore, there are more linguistic and phonetic differences between Japanese and English [7], [8] in comparison with other Asian language. It makes Japanese more difficult to speak English normally in comparison with other Asian countries' people. In order to analyse the speech and language technology based on the general statistical methods, large scale English database by Japanese speakers are necessary. This chapter describes speech database of Japanese English that will be used at model training and testing for the ASR system in this document. Another part of this chapter introduces recorded and annotated Japanese learner corpus collected by our developed dialogue-based CALL system [30], [36], [39]. We will summarize the English speech database by Japanese students and describe the importance of collecting our own database.

Table 4.1: English word and sentence sets prepared in terms of the segmental aspect of English pronunciation.

Set	Size
Phonetically balanced words	300
Minimal pair words	600
Phonetically balanced sentences	460
Sentences including phoneme sequence difficult for Japanese to pronounce correctly	32
Sentences designed for test set	100

Table 4.2: Word and sentence sets prepared in terms of the prosodic aspect of English pronunciation.

Set	Size
Words with various lexical accent patters	109
Sentences with various intonation pronunciation	94
Sentences with various rhythm patterns	121

4.2 ERJ Database

In order to improve the proficiency of Japanese pronunciation of English (JPE), a Japanese national project of "Advanced Utilization of Multimedia for Education" has started in 2000 [50]. "English Learners' Speech Database" is called as "English Read by Japanese" database (ERJ) [48] collected for this project which made in view of CALL system development consisting of words and sentences read by 200 Japanese students. All important details about speech recoding setup and properties of ERJ database collected are described in this section which refers to Minematsu's introduction [48].

4.2.1 Information of Participants

In order to cover a wide range of English pronunciation ability, ERJ database was collected carefully by English of both good and poor speakers. Database collection was cooperated with twenty organizations such as universities and colleges. It is considered gender-balanced, and is

Table 4.3: Phonemic symbols (phoneme in alphabet notation) assigned to reading material.

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

recorded by 100 male and 102 female Japanese students who learn English as their L2. All of the sentences are divided into eight groups and all of the words are divided into five groups (Table 4.1 and Table 4.2). There are at least 120 sentences in a sentence group and at least 220 words in a word group required to record by each participant. For each sentence and each word, it is read by 12 speakers and 20 speakers of both genders, respectively.

4.2.2 Data Contents and Specification

Collection of ERJ database used phonemic symbols of TIMIT database [47] and those of CMU pronunciation dictionary as reference sets. The phonemic symbols of vowels and consonants assigned to reading material were shown in Table 4.3. Selection of reading material is based on the syllabuses of English pronunciation training which is divided into segmental and prosodic. Details of two aspects are shown in Table 4.1 and Table 4.2.

Participant have enough time to practice their reading before actually recoding. They were asked to read the materials what they think is correct pronunciation for them.

4.3 Learner Corpus Collected by Dialogue-based CALL System

We described a work on developing a human-machine dialogue-based English CALL system [30], [36], [39]. The system aims at eliciting more speech production from Japanese learners in order to improve their speaking skills, by involving them in man-machine dialogues.

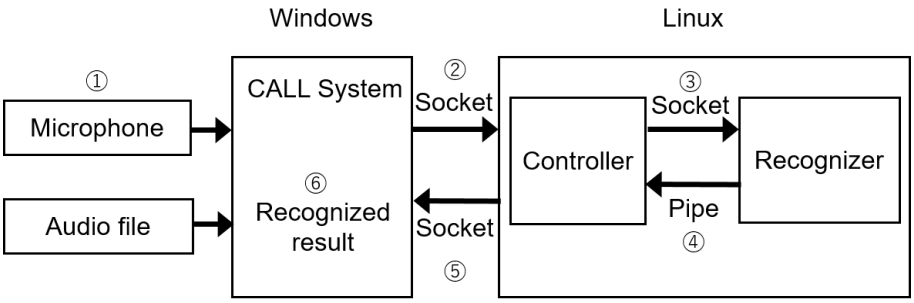


Figure 4.1: Block diagram of the Dialogue-based CALL system.

Spoken responses by L2 speakers varied widely. It is difficult to befittingly collect the effective speech for them. Therefore, the methodologies for constraining L2 speakers' responses is an important task for data collection of spontaneous speech by L2 speakers. Various methodologies have been proposed for CALL systems to constrain L2 speakers' response and make their speech in the controllable range, such as giving users hint stimuli in the form of a keyword or incomplete sentences, having users do a pre-exercise of typical conversational examples before using CALL systems, and so on [41], [42], [43], [44], [45].

Another CALL application is called as a "translation game" [46] presents sentences in the learners' native language, asks them to provide a spoken translation in the target language, and then gives feedback on grammatical and vocabulary errors. This methodology can relatively constrain diverse spoken responses by L2 speakers in comparison to conventional spoken dialogue systems for non-native speakers. In our development of English CALL system for Japanese, we referred the translation game type method mentioned above and provided the interface which prompted text in Japanese, in order to help them construct utterances which are easy to be recognized and collected.

4.3.1 System Structure

Our system is based on the Java agent development framework. Figure 4.1 shows the block diagram of the system, in which the recognizer uses Tokyo Tech Transducer-based Decoder T^3

Figure 4.2: A screenshot of the CALL interface: (1) dialogue scenario selection (shopping, restaurant and hotel); (2) prompt by system; (3) hint stimulus; (4) recognition result; (5) corrected feedback

[49]. The work-flow is as follows:

1. Speech input through microphone.
2. Transmission of speech data to the dialogue controller.
3. Doing speech recognition.
4. Sending recognition results to the controller.
5. Sending recognition results to the CALL interface.
6. Displaying the results on the screen.

Table 4.4: Examples of translations and evaluated scores in five grades.

Grade	Example 1	Example 2	Example 3
5	Could you keep my luggage till 3 o'clock?	Can you guarantee the quality?	I didn't notice it when I bought it.
4	Can you keep this baggage until 3 o'clock?		I didn't realize this when I bought this.
3	Would you keep these luggage until 3 o'clock?	Do you insure the quality?	I didn't know when I bought.
2	Can I deposit this luggage until 3	Can you promise the quality?	I didn't know when it buy.
1	Can you lent this bag by 3 o'clock?	Can you after care it?	I didn't know that time.

4.3.2 Man-Machine Interface

Figure 4.2 illustrates a screen shot of the man-machine interface of the developed CALL system. At the beginning, the user selects a topic from the menu, which currently supports dialogue scenarios of shopping, menu ordering at a restaurant, and room booking at a hotel, as a window 1 shows. The system usually initiates a dialogue by asking the user a question, which is shown in the window 2. The window 3 usually provides hint information for the answer in Japanese. The window 4 gives the speech recognition results and the window 5 shows a corrected feedback. Colours are used to pinpoint word errors: the blue for insertions, and the red for deletions and substitutions. All utterances including the system question, user's response and the corrected feedback can be reproduced as many times as the user requests.

4.3.3 Learner Corpus

We used the developed dialogue-based CALL system to collect English speech data totally uttered by 65 Japanese students on topics related to shopping, ordering at a restaurant, and hotel booking. Each participant uttered orally translated English speech corresponding to Japanese sentences displayed on a screen. We define "speech orally translated English corresponding to Japanese sentences" to "semi-spontaneous speech" in this document.

Table 4.5: Distribution of scores of translation quality (grade 5 to grade 1) in a part of collected learner corpus including of 10 females and 10 males.

Grade	800–900	700–800	600–700	500–600	400-500	Average
5	17.5	16.6	19.1	11.6	10.0	15.4
4	14.6	11.9	9.8	7.7	8.6	9.8
3	45.4	50.6	51.6	49.3	40.8	46.0
2	9.3	8.0	9.5	15.5	18.7	13.0
1	13.2	12.8	10.0	16.1	21.9	15.8

The utterances were transcribed and their translation quality was evaluated and scored one of five grades by one or two native English speakers with a subjective evaluation method used at the International Workshop on Spoken Language Translation [51]. Table 4.4 shows examples of translations and evaluated scores in five grades. Grade 5 is the meaning of perfect transcriptions like native speaker's expressions, grade 4 means good sentences, grade 3 shows non-native expressions, grade 2 is the meaning of disfluency transcriptions, and grade 1 is incomprehensible expressions. Expressions regarded as ungrammatical and unacceptable in the learner corpus were also given comments for generating effective feedback in our system.

In our data collection, the communication levels of English by Japanese were measured using the Test of English for International Communication (TOEIC) [53]. Their scores ranged from 300 to 910 (990 being the highest score that can be attained). Table 4.5 shows the distribution of scores of translation quality in a part of collected learner corpus including of 10 females and 10 males, but except one person who ranged over 900 and one person who ranges less than 300. Table 4.5 shows that there is a tendency that utterances produced by participants scoring in the low TOEIC level range were generally given by the translation quality with low grade (grade 1 and grade 2). According to the distribution of scores, the translation quality with grade 3 accounts for half of all counting results, and that with other grades account from approximately 10% to 16%. It is almost equally distributed for the construction of learner corpus. This database was used as evaluation data set in the experiments.

5 Phoneme Set Design with Integrated Acoustic and Linguistic Features

> *In automatic speech recognition, the acoustic signal is the only tangible connection between the talker and the machine. While the signal conveys linguistic information, this information is often encoded in such a complex manner that signal exhibits a great deal of variability.*

Victor Zue, *Professor at Massachusetts Institute of Technology*

When building an ASR system for non-native speech, most of the ASR technologies have been developed to handle the subject of pronunciation variations in terms of acoustic modeling [27], [32], [33], lexical modeling [34] and extended lexicon [28], and grammatical relations in terms of language modeling [35] mentioned in Chapter 2. This chapter focuses on the pronunciation variations in terms of acoustic modeling with the customized phoneme set design which considers that L2 speech usually includes less fluent pronunciation and more frequent pronunciation mistakes.

In the remaining part of this chapter, we describe the introduction of the work in Section 5.1. Section 5.2 describes the reduced phoneme set (RPS) for second language speech. In Section 5.3, we illustrate the criterion of the phoneme set design. Section 5.4 presents theory of phonetic

decision tree (PDT)-based clustering splitting. Then in Section 5.5, we introduce the designed discrimination rules used in PDT method. Section 5.6 presents framework of the phoneme set design in detail. In Section 5.7 and Section 5.8, the experiments and discussions are presented in detail, respectively. Finally, we conclude and summarize this proposed method.

5.1 Introduction

Read speech produced by non-native speakers has only different acoustic features in comparison to that by native speakers. On the other hand, utterances produced by non-native speakers on their own have different features from the native speech not only in acoustic features but also lexical or grammatical features. These acoustic and linguistic features of non-native speech share a close relation when they are combined and relate to the performance of ASR systems, and both features should be taken into consideration when designing non-native speech ASR systems. We propose a novel method to derive a reduced phoneme set (RPS) to improve the recognition performance for second language speech.

There have been several previous studies on using a RPS for speech recognition. For example, Vazhenina et al. proposed a method to generate an initial confusion matrix of phonemes which combined phonological and statistical information of Russian and then manually merge some easily confused phones by referencing phonological knowledge [37]. Although this approach has a good performance, it did not consider acoustic variation of each phone but only depended on the phonological knowledge.

Zhang et al. [88] proposed an efficient phoneme set of tone-dependent units to build a Chinese ASR system, by iteratively merging a pair of tone-dependent units according to the principle of minimal loss of the mutual information between the word items and their phoneme transcriptions based on the text corpus. This approach had a capability to keep discriminative tonal and phoneme contrasts that are most helpful for disambiguating homophone words due to the lack of tones, and merge those tonal and phoneme contrasts that are not importance for word disam-

biguation for the recognition task. Recognition results showed that the proposal was effective for Chinese ASR systems, but the proposed method using the statistical information were only measured based on the text corpus, and acoustic features and spectral properties fell into neglect in Zhang's study.

There was also a study on measuring the distance between acoustic models to merge language-dependent phones using a hierarchical phone clustering algorithm [38]. However, this approach does not consider the acoustic characteristics of the phonemes in real utterances, because of each phone in the use of different levels in spoken language. These previous studies performed well with native speech, neither consider the characteristics of the second language speech: specifically, that the mapping applicable to the alignment between phonetic symbols and the native speaker's speech does not in some cases apply to the second language speech, which contains inherently overlapping distributions of phonetics and phonemes that do not exist in the canonical phoneme set.

In this study, we propose a novel method considering integrated acoustic and linguistic features to derive a reduced phoneme set for L2 speech recognition. The phoneme set is created with a phonetic decision tree (PDT)-based top-down sequential splitting method that utilizes the phonological knowledge between L1 and L2, and delivers a better recognition performance for second language speech.

Approach proposed in this chapter considers both acoustic and linguistic features in a unified way and optimizes the weighted total of both factors. We evaluated the recognition performance of the proposed method in the L2 speech corpus collected by the English CALL system mention in Chapter 4.

5.2 Reduced Phoneme Set for Second Language Speech

The RPS usually alleviates acoustic discrimination ability – the mapping between the sequence of acoustic feature vectors and phonemes for the L2 speakers. There are two reasons the RPS is

effective for ASR for L2 speech when the L1 of users is known. **One**, the reduced phoneme set can create suitable phonological decoding for the second language speech because the reduced set can be designed to characterize the acoustic features of the second language speech more correctly. **Two**, we can obtain more reliable estimate values as parameters of acoustic models, because there is more speech data for training the acoustic model of each phoneme in the reduced set than in the canonical one.

However, the reduced size of phonemes has in principle a weaker linguistic discrimination ability – the mapping between phoneme sequences and word sequences – in comparison to the canonical phoneme set which is used in general English ASR system. The effect of the reduced phoneme sets improved acoustic discrimination ability outweighs the drop in its linguistic one compared with the canonical phoneme set.

5.3 Criterion of the Phoneme Set Design

In this work, we adopt maximization of the weighted sum of a phoneme set's acoustic likelihood and its linguistic discrimination ability to derive the optimal phoneme set S, as

$$\Psi_S = \arg\max[\lambda \cdot \triangle L_S + (1 - \lambda) \cdot \mathcal{F}(S)], \tag{5.1}$$

where $\triangle L_S$ is the increased acoustic likelihood of the reduced phoneme set S compared with the canonical one, $\mathcal{F}(S)$ represents its linguistic discrimination ability, and Ψ_S is an unified objective function corresponding the optimal reduced phoneme set S over all reduced ones. Details will be described in the following sections.

5.3.1 Acoustic Likelihood

We use as the acoustic objective function the accumulated log likelihood of probabilities generating the second language speech observation data $\boldsymbol{O}_t = [O_1, O_2, ..., O_T]$ by the probabilistic density functions ($PDFs$) defined by the parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\sigma}}$. It is defined by

$$L(P_S) \approx \sum_{t=1}^{T} \log[P(\boldsymbol{O}_t; \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\sigma}}_s)] \cdot \gamma_s(\boldsymbol{O}_t), \tag{5.2}$$

where S represents a phoneme set and P_S is the node pdf of a phoneme set S. $\hat{\boldsymbol{\mu}}_s$ and $\hat{\boldsymbol{\sigma}}_s$ represent the mean vector and the covariance matrix of phonemes s assigned to the phoneme set S, respectively, and describe the shape of the Gaussian distributions. $\gamma_s(\boldsymbol{O}_t)$ is a posteriori probability of the observation data $\boldsymbol{O}_t$ being generated by phoneme s. In here, it is calculated by the canonical phonemes s typically used in Japanese English speech utterances.
Consequently, increased acoustic likelihood $\triangle L_S$ with the reduced phoneme set is defined as

$$\triangle L_S = L(P_s) - L(P_c), \tag{5.3}$$

where P_s and P_c represent the log likelihood defined in Eq. (5.2) for the reduced phoneme set and the canonical phoneme set, respectively.

5.3.2 Linguistic Discrimination Ability

Various words w_1, w_2, ..., w_n of originally discriminated phoneme sequences ordered by the canonical phoneme set are re-figured as one word w^R of the same phoneme sequence by the reduced phoneme sets. Hence, the words represented by the reduced phoneme set include more homophones, which are words with the same pronunciation but different meaning and spelling, than those by the canonical one. The phoneme sequences by the reduced phoneme set worsen the word discrimination ability in the lexicon.

These homophones decrease linguistic discrimination ability, but they are usually disambiguated with contextual information in human-to-human communications and are partly done with a language model in ASR. We should therefore consider the effect of language model that partly disambiguates homophones to measure linguistic discrimination ability of the reduced phoneme set by collecting a huge transcription of non-native speech data, as word probabilities in utterances by non-native speakers differ from those by native speakers. Unfortunately, transcriptions of non-native speech are less available than those of native speech, so we use as an approximate approach, word discrimination ability – $\mathcal{F}_{Lex}(S)$ – the ratio of perplexity $PP(W_{M_{diff}(S)})$ of words with discriminated phoneme sequences in the reduced phoneme set S to perplexity $PP(W_N)$ of words with discriminated phoneme sequences in the canonical one, to define the linguistic discrimination ability in this study. The word discrimination ability of the reduced phoneme set is generally written as

$$\begin{aligned} \mathcal{F}_{Lex}(S) &= \frac{PP(W_{M_{diff}(S)})}{PP(W_N)} \\ &= \frac{2^{H(W_{M_{diff}(S)})}}{2^{H(W_N)}}, \end{aligned} \tag{5.4}$$

where $W_{M_{diff}(S)}$ is the words with discriminated phoneme sequences in the lexicon represented by the reduced phoneme set S and W_N is the words with discriminated phoneme sequences in the original lexicon represented by the canonical phoneme set. $H(W_{M_{diff}(S)})$ and $H(W_N)$ are the entropy of words $W_{M_{diff}(S)}$ and that of words W_N, respectively. Assuming each word has a single pronunciation, the entropy of words $W_{M_{diff}}$ and W_N can be calculated with

$$\begin{aligned} H(W_M) &= -\sum_{m=1}^{M_{diff}} P(w_m) log P(w_m) \\ H(W_N) &= -\sum_{n=1}^{N} P(w_n) log P(w_n), \end{aligned} \tag{5.5}$$

where w_m is the homomorphic word with different pronunciation included in $W_{M_{diff}}$. w_n is the homomorphic word with different pronunciation included in W_N.

Unfortunately, transcriptions of non-native speech are less available than those of native speech, and it is extremely difficult to collect enough data on each conversation topic by a considerable number of non-native speakers with various language proficiencies. Considering the difficulty of satisfying this requirement for non-native speech, the probability of each word $P(w)$ is simplified to be equal, and satisfies the following condition in the phoneme set design, as

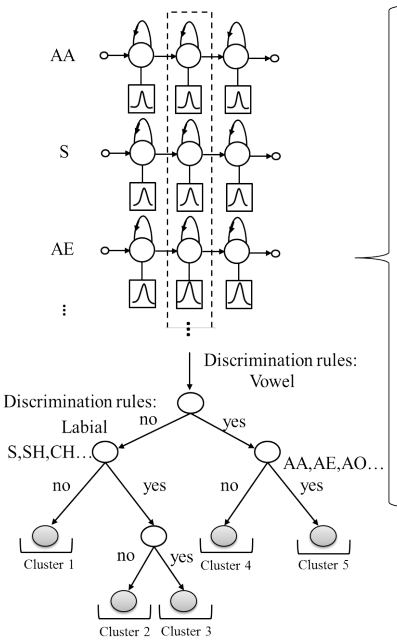
$$\begin{aligned} P(w_m) = \frac{1}{M_{diff}(S)}, \quad P(w_n) = \frac{1}{N} \\ (1 \leqslant m \leqslant M_{diff}(S), 1 \leqslant n \leqslant N), \end{aligned} \tag{5.6}$$

where $M_{diff}(S)$ is the total number of discriminated phoneme sequences in the lexicon represented by the reduced phoneme set S. N is the total number of discriminated phoneme sequences in the original lexicon represented by the canonical phoneme set. The simple assumption of the equal probabilities lead to a simplified $\mathcal{F}_{Lex}(S)$ as,

$$\begin{aligned} \mathcal{F}_{Lex}(S) &= \frac{PP(W_{M_{diff}(S)})}{PP(W_N)} \\ &\doteqdot \frac{M_{diff}(S)}{N}, \end{aligned} \tag{5.7}$$

Regarding the words with discriminated phoneme sequences in Eq. (5.5), w_m has polyphonic ways of pronunciation. The entropies mentioned in Eq. (5.5) can be extended to the following equations,

$$\begin{aligned} H(W_{M_{diff}(S)}) &= -\sum_{m=1}^{M_{diff}} \sum_{k=1}^{C(w_m)} \frac{P(w_m)}{C(w_m)} log \frac{P(w_m)}{C(w_m)} \\ H(W_N) &= -\sum_{n=1}^{N} \sum_{k=1}^{C(w_n)} \frac{P(w_n)}{C(w_n)} log \frac{P(w_n)}{C(w_n)}, \end{aligned} \tag{5.8}$$



Discrimination rules	Contents
Affricates	R, V, JH, Z, TH, ZH, DH
Affricate1	DH, ZH
Affricate2	TH, Z
…	…
Alveolars	L, N, D, T
Alveolar1	N, L
Alveolar2	T, D, N
…	…
Apicals	DH, TH, ZH, Z, S
…	…
Confusing-Consonant	T, DH ,D
Confusing-Vowel	AX, IY, IH
Labials	M, P, B, V, F
…	…
Retroflex1	R, SH
Retroflex2	R, ZH
Retroflex3	R, DH
…	…
Stops	B, P, D, T, G, K
…	…

A part of "Discrimination Rules"

Figure 5.1: PDT-based top-down cluster splitting and a part of the discrimination rules.

In here, M_{diff} is the shorthand notation of $M_{diff}(S)$. $C(w_m)$ is the number of pronunciations of a homomorphic word w_m with different pronunciation included in M_{diff}. $C(w_n)$ is the number of pronunciations of a homomorphic word w_n with different pronunciation included in N.

5.4 Theory for PDT-Based Cluster Splitting

The PDT is a top-down binary sequential splitting process that uses the phonetic acoustic features of speech by L2 speakers and the occurrence distributions of each phoneme as the splitting criterion and uses the relation between the phonological structure of the mother tongue and target languages of the L1 speakers as a set of discrimination rules. Figure 5.1 shows an example of a PDT that partitions the initial phoneme cluster into five terminal clusters with our designed discrimination rules which will be introduced in the following section in detail.

5.5 Discrimination Rules Design

As revealed by many second language acquisition studies, pronunciation by L2 speakers is usually significantly influenced by the mother tongue of the speakers, particularly when the number of phonemes of the mother tongue is less than that of the target language [4]. There are five vowels in common use in Japanese, each of which has a long form functioning as a separate phoneme. In contrast, there are 17 different vowels usually used in English, including several diphthongs such as [ɔɪ], [aʊ], and [aɪ]. There are also some consonants in Japanese that do not appear in English, such as the voiceless palatal fricative [ç] and voiceless bilabial fricative [Φ], e.g., "hito" (human) and "Fujisann" (Mt. Fuji) [6]. Using such phonemes for English speaking undoubtedly creates a high number of mispronunciations.

We designed 166 discrimination rules based on this knowledge of the phonetic relations between the Japanese and English languages and the actual pronunciation inclination of English utterances by Japanese. The full designed discrimination rules were shown in Appendix C. More specifically, we referred to the literature of linguistic knowledge [4], [6] and phoneme confusion matrix for Japanese speakers of English [24] to design the discrimination rules. Discrimination rules that categorize each phoneme on the basis of phonetic features such as the manner and position of articulation were utilized to carry on the preliminary splitting before actually splitting. The preliminary splitting will be described in Section 5.6.1 in detail. A part of the discrimination rules is shown in Figure 5.1, where the first rule in the list, "Affricates" denotes that phonemes R, V, JH, Z, TH, ZH, and DH have an affricate feature, making them suitable to discriminate the native speech. Other sets of phonemes are listed as phonemes with "affricate" in the "Affricate1", "Affricate2", etc. rules, considering the inclination of mispronunciation by the second language speakers. All phonemes listed in each discrimination rule based on other phonetic features depict the similar phonological characteristics and have the possibility to be merged into a cluster.

Table 5.1 shows the canonical phoneme set of English in the Alphabet notation which will be used as our baseline system. A list of the phonemic symbols of English corresponding to the IPA notation and word examples can be found in appendix A. The assigned phonemic symbols

Table 5.1: Canonical phoneme set of English in Alphabet notation.

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

of English are adopted in our experiment as our initial phoneme set.

5.6 Framework of the Phoneme Set Design

We followed an incremental procedure in our design of the phoneme set with a PDT-based top-down clustering method to obtain the optimal reduced phoneme set. Figure 5.2 shows the overall procedural diagram of the phoneme cluster splitting with the unified acoustic and linguistic objective function mentioned in Section 5.3.

5.6.1 Initialization Conditions

■ Initial phoneme cluster

To set a cluster including all phonemes of the canonical set listed in Table 5.1 as a root cluster and use the mid-state of the context-independent English HMMs of each phoneme as their acoustic model.

■ Lexicon

To prepare the words with discriminated phoneme sequences in the original lexicon represented by the canonical phoneme set.

■ Discrimination rules

To use the designed discrimination rules to carry out the preliminary splitting process. The cluster is split heuristically by the discrimination rules, which were defined by the phonetic features and phonological properties of Japanese English on the linguistic level. The preliminary splitting

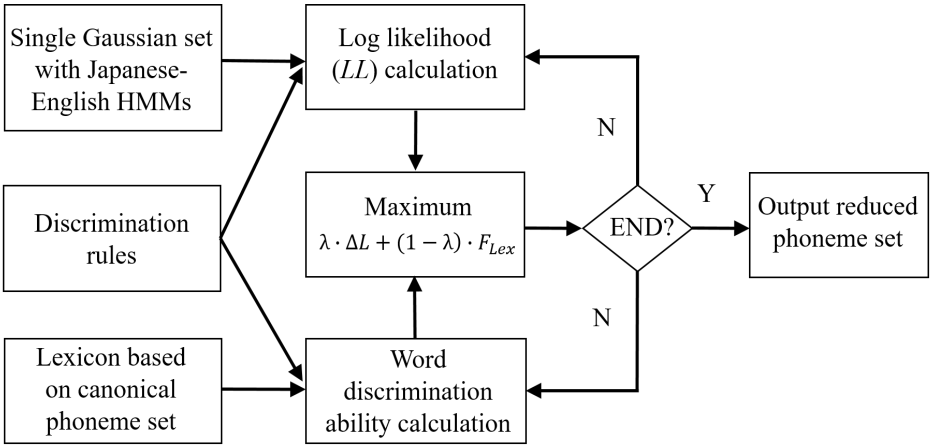


Figure 5.2: Phoneme cluster splitting with a PDT-based top-down method using both log likelihood (acoustic part) and word discriminating ability (linguistic part) as criteria.

process based on designed discrimination rules is used to further calculate log likelihood of each phoneme cluster and word discrimination ability in each renewed lexicon, as described in the following section.

5.6.2 Phoneme Cluster Splitting Procedure

Step 1 Calculate Log Likelihood

Assuming that the phoneme cluster s is partitioned into $s_y(r)$ and $s_n(r)$ by one of the discrimination rules r, the increase of log likelihood $\triangle L_{s,r}$ is calculated as

$$\triangle L_{s,r} = L(s_y(r)) + L(s_n(r)) - L(s) \tag{5.9}$$

$\triangle L_{s,r}$ is the increased log likelihood of the phoneme cluster, which is calculated for all discrimination rules r applicable to each cluster. The detailed calculation is explained in the subsection 5.6.3.

Step 2 Renew Lexicon

The lexicon will be renewed by the current phoneme set based on all discrimination rules r. Here, phonemes existing in the same clusters/rules will be temporarily merged into one phoneme for renewing the lexicon.

Step 3 Calculate Word Discrimination Ability

The probability of words with discriminated phoneme sequences in each renewed lexicon by one of the discrimination rules r is based on Eq. (5.6) and calculated as

$$\mathcal{F}_{Lex}(s, r) = \frac{M_{diff}(s, r)}{N} \tag{5.10}$$

where N is the total number of discriminated phoneme sequences in the original lexicon represented by the canonical phoneme set and $M_{diff}(s, r))$ is the number of discriminated phoneme sequences in the renewed lexicon represented by the current phoneme set based on the discrimination rule r.

Step 4 Select the Optimal Splitting Rule and Phoneme Cluster to Split

The rule r^* and the phoneme culster s^* are chosen when it brings about the maximum of the following formula:

$$\Psi^*_{s^*,r^*} = \arg\max_{all\ s,r}[\lambda \cdot \triangle L_{s^*,r^*} + (1 - \lambda) \cdot \mathcal{F}_{Lex}(s^*, r^*)] \qquad (0 \leqslant \lambda \leqslant 1) \tag{5.11}$$

where λ is across on the interval [0, 1] with evenly 0.1 (0, 0.1, 0.2, ..., 0.9, 1.0).

Step 5 Split Phoneme Clusters

The phoneme cluster s^* is split into two clusters, $s^*_y(r^*)$ and $s^*_n(r^*)$, in accordance with rule r^* selected in Step 4.

Step 6 Check Convergence

Check whether the stop criterion is satisfied. If yes, the splitting process is terminated. If not, steps 1 to 5 are repeated.

5.6.3 Calculation of Log Likelihood

We used as the splitting criterion the log likelihood (LL) defined by the logarithm of the probabilistic density functions ($PDFs$) of an acoustic model generating the speech observation data. It can be achieved by using Eq. (5.1) when setting $\lambda = 1$ and also defined by

$$L(P_m) \approx \sum_{t=1}^{T} \log[P(\boldsymbol{O}_t; \hat{\boldsymbol{\mu}}_m, \hat{\boldsymbol{\sigma}}_m)] \cdot \gamma_m(\boldsymbol{O}_t) \tag{5.12}$$

where P_m represents the m^{th} phoneme or phoneme cluster and P is the joint node PDF of the phoneme cluster. The mean vector $\hat{\boldsymbol{\mu}}_m$ and covariance matrix $\hat{\sigma}_m$ are calculated with equations (5.13) and (5.14), respectively:

$$\hat{\boldsymbol{\mu}}_m = \sum_{i \in P_m} \frac{\gamma_i \boldsymbol{\mu}_m}{\gamma_i} \tag{5.13}$$

$$\hat{\sigma}_m = \sum_{i \in P_m} \frac{\gamma_i (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_m)^2 + \gamma_i \sigma_i}{\gamma_m} \tag{5.14}$$

where $\boldsymbol{\mu}_i$ and σ_i represent the mean vector and the covariance matrix of phoneme i, respectively, which is an element of class P_m, and γ_i represents the phonetic occupation counts of phoneme i. $\gamma_m(\boldsymbol{O}_t)$ is a posteriori probability of the model generating the observation data $\boldsymbol{O}_t = [O_1, O_2, \ldots, O_T]$. It is calculated by the canonical phoneme set used in typical Japanese English speech utterances.

We can compute the log likelihood of each phoneme cluster by substituting equations (5.13) and (5.14) into (5.12).

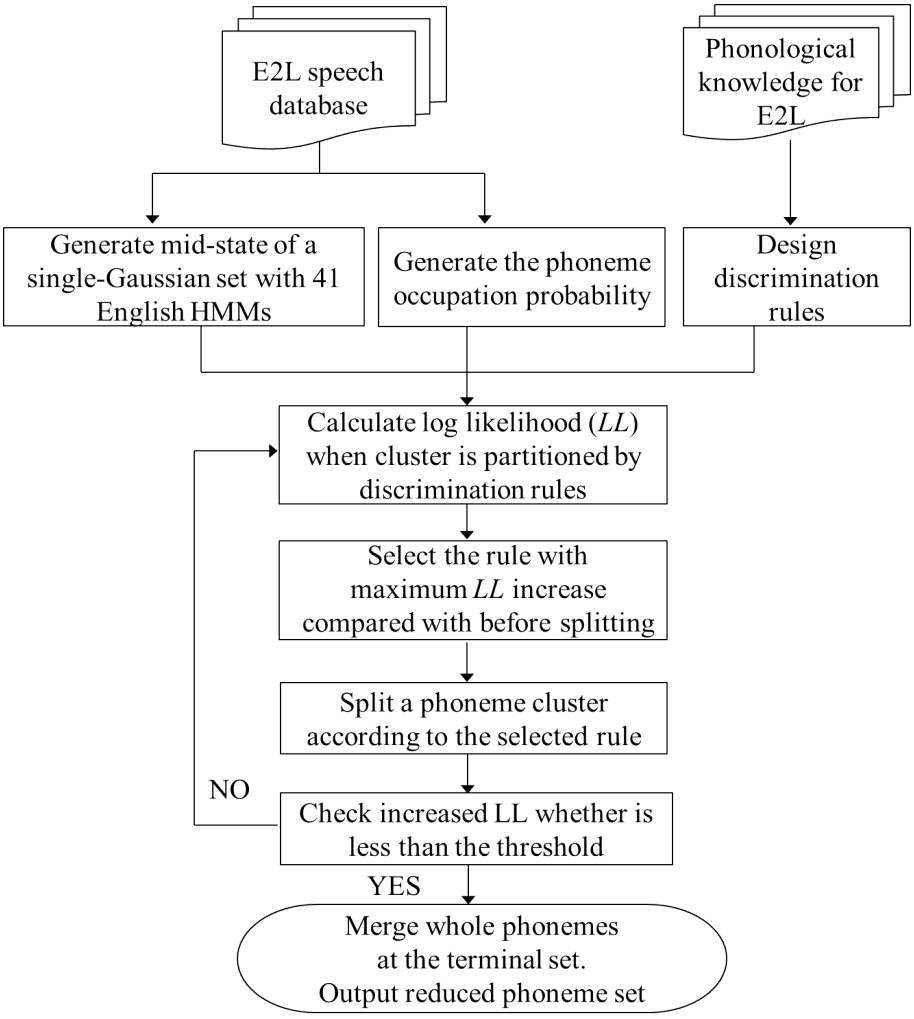


Figure 5.3: Overall procedural diagram of the phoneme cluster splitting with a phonetic decision tree (PDT)-based top-down method using a maximum log likelihood criterion.

Figure 5.3 shows the overall procedural diagram of the phoneme cluster splitting with the PDT-based top-down method using a maximum log likelihood criterion. The detailed procedure can be summarized:

1. Set all merging phonemes as a root cluster at the initial state. Calculate increased log likelihood $\triangle L_R$ according to equations (5.12) and (5.15), assuming that cluster s is partitioned into $s_y(r)$ and $s_n(r)$ by discrimination rule r.

$$\triangle L_{s,r} = L(s_y(r)) + L(s_n(r)) - L(s) \tag{5.15}$$

where $\triangle L_{s,r}$ is the increased log likelihood of the phoneme cluster, which is calculated for all discrimination rules applicable to every cluster.

2. Select phoneme cluster s and the rule r with maximum log likelihood.

$$L_{s^*,r^*} = \arg\max_{all \quad s,r} \triangle L_{s^*,r^*} \tag{5.16}$$

The rule r^* and the phoneme cluster s^* are chosen when it causes the maximum log likelihood increasing compared with before splitting.

3. Split a phoneme cluster according to the selected discrimination rule r^*.

4. Check whether increased log likelihood is less than the threshold. If yes, output the final reduced phoneme set. If no, repeat the splitting process.

5.7 Experiments

5.7.1 Experimental Setup

Experiments were carried out to compare the recognition performance using the canonical phoneme set and the reduce phoneme sets by the proposed method. The phonemic symbols of the TIMIT database were used as a reference set [47]. There are 41 phonemes in the canonical phoneme set, including 17 vowels and 24 consonants. Table 5.1 lists the phonemes of English in Arpabet notation and IPA notation. The baseline is ASR using the canonical phoneme set in the experiment.

Table 5.2: Condition of acoustic analysis and HMM specifications.

AA	Sampling rate	16kHz
	Feature vector	0-12 mel-cepstral
		energy+Δ + $\Delta\Delta$ (CMN) 39 dimension
	Frame length	20ms
	Frame shift	10ms
	Window type	Hamming window
HMM	Number of states	5 states 3 loops
	Learning method	Concatenated training
	Type	Left to right continuous HMM

For the initial phoneme cluster, an English speech database read by Japanese students (ERJ) [48] mentioned in Chapter 4 was used to train context-independent (CI) 3-state monophone gaussian mixture model (GMM)-HMMs of a left-to-right state topology. Table 5.2 shows the experimental conditions for acoustic analysis and the HMM specifications.

5.7.2 Acoustic Model, Language Model, and Lexicon

For the ASR system, the ERJ speech database was used to train context-dependent (CD) state-tying triphone GMM-HMM acoustic models of various numbers of phoneme sets. We developed a bigram language model using about 5,000 transcribed utterances taken from the learner corpus mentioned in Chapter 4. We used a pronunciation lexicon related to conversation about travel abroad. It consisted of about 45,660 phoneme sequences for 28,000 word types with different meanings. There are approximately 43,100 discriminated phoneme sequences in the original pronunciation lexicon represented by the canonical phoneme set.

5.7.3 Evaluation Data

We collected speech from 20 participants uttering orally translated English speech corresponding to visual prompts also from the CALL system as evaluation data. The participants were Japanese students who had acquired Japanese as their L1 and learned English as their L2. Their speak-

ing styles ranged widely from ones similar to conversation to ones closer to read speech. The communication levels of participants in English were measured using the Test of English for International Communication (TOEIC) [53]. Their scores ranged from 380 to 910. Table 4.5 shows the distribution of scores of translation quality for the evaluation data except for one person who ranged over 900 and one person who ranges less than 300. Briefly summarized that there were a total of 1,420 utterances recorded by each participant in response to 71 visual prompts.

5.7.4 Experimental Results

In order to verify the performance of the derived RPSs by the proposed method, we heuristically chose 25-, 28-, and 32-phoneme sets that are reliable proficiency-dependent phoneme sets will be introduced in Chapter 7 [1] and used them for recognition experiments. We used the hidden markov model toolkit (HTK) toolkit [54] to compare the performance on ASR implementing the proposed method with that of the canonical phoneme set and the reduced phoneme sets generated by the PDT when setting $\lambda = 1$, which means only used as the splitting criterion the log likelihood given by an acoustic model. The recognition results can be achieved by using Eq. (5.1) when setting $\lambda = 1$.

Figure 5.4 shows the word accuracy of the canonical phoneme set, the reduced phoneme sets by PDT when setting $\lambda = 1$, and the reduced phoneme sets by PDT based on the weighted λ considering unified acoustic and linguistic objective function. We observed the following:

- The reduced phoneme sets with weight λ delivered a better performance than the canonical phoneme set and other reduced phoneme sets by PDT when setting $\lambda = 1$.
- The recognition performance using the proposed method which considering the weighted λ was improved more for fewer numbers of phonemes than for greater numbers of phonemes in comparison to that only based on the method setting $\lambda = 1$.

[1]The optimal RPS corresponding to the English proficiency of speakers was determined to be 25-RPS for speakers with a TOEIC score of less than 500, 28-RPS for those with a 500–700 score, and a 32-RPS for those with scores higher than 700.

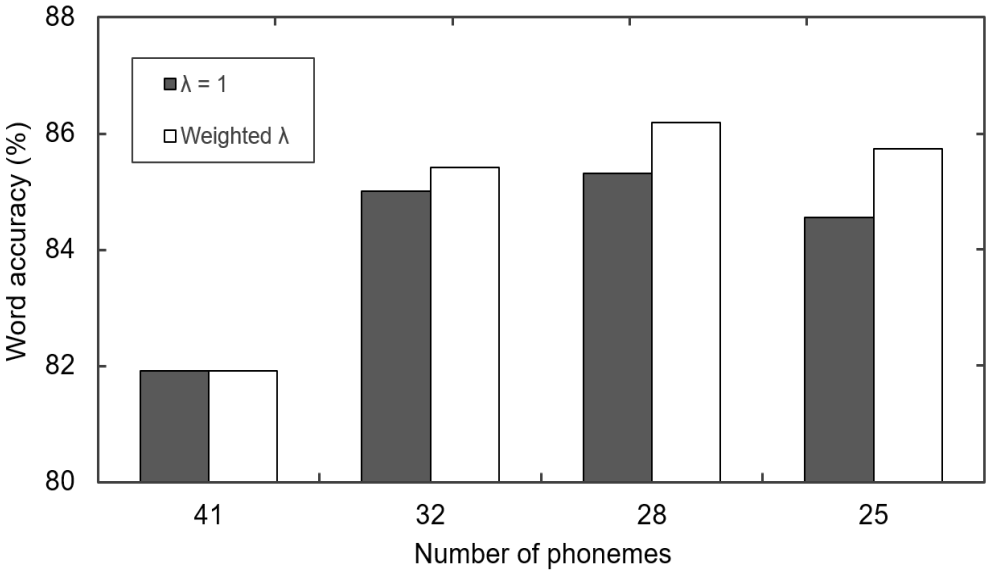


Figure 5.4: Word accuracy of the canonical phoneme set and various reduced phoneme sets by PDT only based on the acoustic likelihood ($\lambda = 1$) and PDT based on the proposed method (weighted λ).

5.8 Discussions

In this section, we investigate from two aspects—one, why is efficiency of the reduced phoneme set based on the unified acoustic and linguistic objective function; two, word discrimination ability considering equal/estimated probability of occurrence of each word—to evaluate the efficiency of adopting linguistic discrimination ability for improving speech recognition accuracy for the second language speech.

5.8.1 Efficiency of the Reduced Phoneme Set based on the Unified Acoustic and Linguistic Objective Function

In order to evaluate the efficiency of the proposed method based on unified acoustic and linguistic objective function, we investigated the relation between the recognition performance of various numbers of phonemes and different weighting factors.

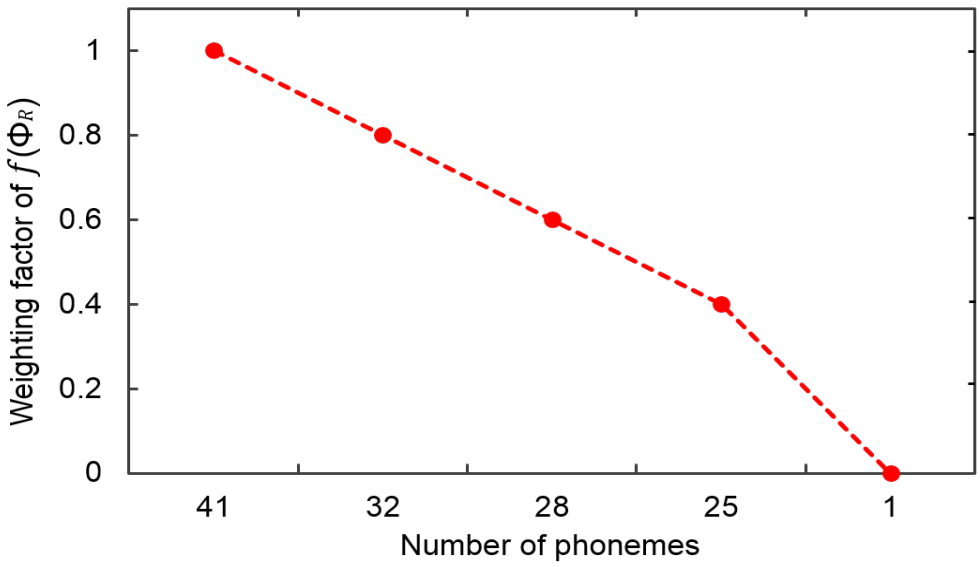


Figure 5.5: The best recognition performance of various numbers of phonemes corresponding to weighting factor of word discrimination ability ($\mathcal{F}_{Lex}(s^*, r^*)$).

Figure 5.5 shows the best recognition performance corresponding to the weighting factor ($1 - \lambda$) of word discrimination ability ($\mathcal{F}_{Lex}(s^*, r^*)$ in Eq. (5.11)) for various numbers of phonemes generated by our proposed method. It is clear that

- The most efficient weighting factor of word discrimination ability is different depending on the number of phonemes in the set.
- There is a trend of reducing the weighting factor of word discrimination ability with numbers ranging from 41 to 1 in decreasing order for the best recognition performance.

5.8.2 Word Discrimination Ability Considering the Equal/Estimated Occurrence Probability of Each Word

The occurrence probability of each word is thought to be largely different. Designing the reduced phoneme set in consideration of the occurrence probability of each word would affect the linguistic discrimination ability, although it is still difficult to collect transcriptions of non-native

Table 5.3: Word discrimination ability (%) for discriminated phoneme sequences of **all lexicon items** represented by the canonical phoneme set and various numbers of phoneme sets considering **the equal occurrence probability** of each word. The reduction rate in comparison to the canonical phoneme set is given in parentheses.

λ	Number of phonemes in the set	Original (Canonical set)	Only based on acoustic likelihood	Proposal
0.2	32	94.4	92.1 (2.3)	93.2 (1.2)
0.4	28		88.9 (5.5)	89.6 (4.8)
0.6	25		87.9 (6.5)	89.1 (5.3)

Table 5.4: Word discrimination ability (%) for discriminated phoneme sequences corresponding to words used in **evaluation data** represented by the canonical phoneme set and various numbers of ones considering **occurrence probability estimated with the learner corpus**. The reduction rate in comparison to the canonical phoneme set is given in parentheses.

λ	Number of phonemes in the set	Original (Canonical set)	Only based on acoustic likelihood	Proposal
0.2	32	92.6	89.4 (3.2)	89.9 (2.7)
0.4	28		88.3 (4.3)	89.7 (2.9)
0.6	25		85.7 (6.9)	87.3 (5.3)

speech. We temporarily check the effect of the occurrence probability of each word using a small corpus.

In the case of the equal occurrence probability of each word, we utilized the same computational method for the reduced phoneme set design (refer to Eq. (5.6)). In the case of different occurrence probability of each word, we estimated the probability using the text corpus of evaluation data mentioned in Section 5.7.3 and the learner corpus mentioned in Section 4.3.3. The occurrence probability of each word for each corpus satisfies

$$\sum_{m=1}^{M_{diff}} P(w_m) = \sum_{n=1}^{N} P(w_n) = 1. \tag{5.17}$$

Table 5.3 shows the word discrimination ability for the discriminated phoneme sequences of all

lexicon items by the canonical phoneme set and various numbers of the reduced phoneme sets, considering the equal probability of occurrence of each word. Even if the number of the phoneme set is reduced to 25 (39% reduction of phoneme numbers), only 5.3% of lexical items are merged into a confusable word class.

Table 5.4 shows the word discrimination ability for discriminated phoneme sequences of all lexicon items used in the evaluation data by the canonical phoneme set and various numbers of the reduced phoneme sets, considering word occurrence probability estimated in the learner corpus. In this case, 5.3% of lexical items are also merged into a confusable word class. This is smaller than expected in light of the number of reduced phonemes, which indicates that the phoneme occurrence distribution is largely distributed.

The vocabulary size of the lexicon used in the experiment is 28,000, which we feel is sufficiently large for the productive vocabulary of second language speakers. The literature on English as a foreign language for Japanese learners [55], [57] reported the mean vocabulary size of the aural and written case to be approximately 5,000, consisting of 3,000 words with different categories. Therefore, we use other vocabulary sizes for the lexicon, 14,000 and 7,000, to verify the efficacy of the proposed method.

Figure 5.6 shows the word accuracy of the canonical phoneme set and various reduced phoneme sets by the proposed method with different vocabulary sizes of the lexicon. Experimental results show that the lexicon with smaller vocabulary size achieved better recognition performance than the larger ones.

5.9 Summary

This chapter proposes a method of designing the reduced phoneme sets for L2 speech maximizing a unified acoustic and linguistic objective function of second language speakers and implements the method as a decision tree to derive a reduced phoneme set. We apply the reduced phoneme set developed with the proposed method to English utterances spoken by Japanese

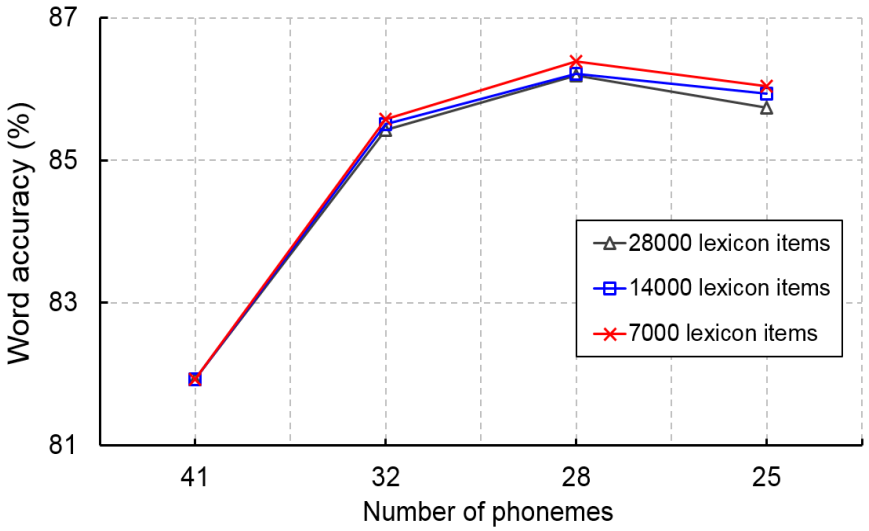


Figure 5.6: Word accuracy of canonical phoneme set and various reduced phoneme sets by proposed method with different vocabulary size of the lexicon.

collected with a translation game type dialogue-based CALL system. The experimental results show that it achieves a greater improvement in speech recognition performance than the canonical phoneme set and the reduced ones by PDT only based on the acoustic likelihood. We verified that the proposed method is effective for ASR that recognizes L2 speech when the mother tongue of speakers is known.

6 Analysis of Effect of Acoustic and Linguistic Features

> *Discovery consists of looking at the same thing as everyone else and thinking something different.*

Albert Szent-Györgyi, *Nobel Prize in Physiology or Medicine winner*

We adopt maximization of the weight total of a phoneme set's acoustic likelihood and its linguistic discrimination ability to derive the optimal phoneme set in Chapter 5. In order to verify the efficiency of proposed method, the analysis of effect of acoustic feature and linguistic feature is introduced and investigated in this chapter in detail.

6.1 Analysis of Effect of Acoustic Feature

In order to examine the effectiveness of acoustic feature, we compared the performance of ASR implementing the proposed method with that of the canonical phoneme set and analyzed the experimental results from the following two aspects: the effect of splitting methods and the effect of phoneme occupation probabilities.

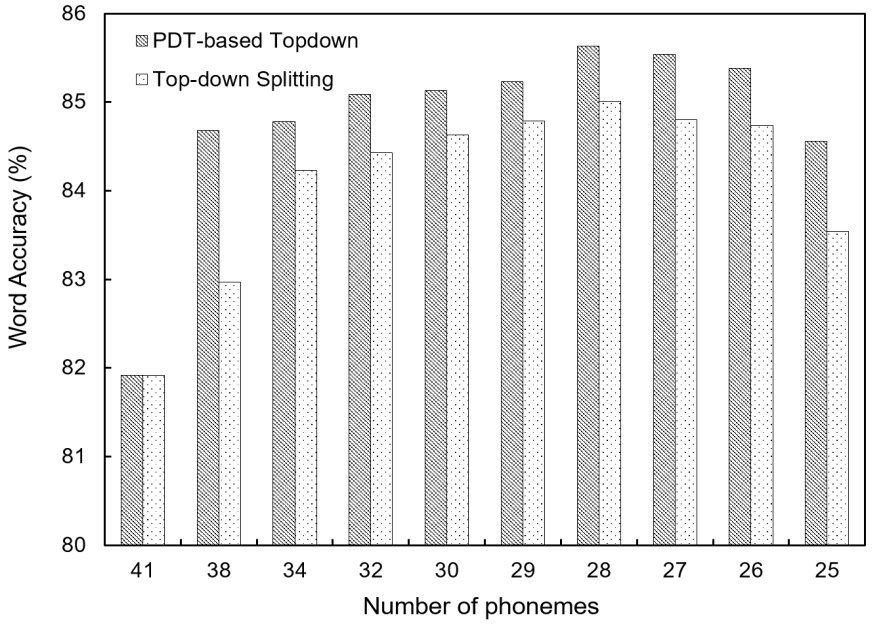


Figure 6.1: Word accuracy of different numbers of phoneme sets using the PDT-based top-down method and the top-down splitting method.

6.1.1 Experimental Setup

We took English speech data uttered by 55 Japanese students based on the conversations on shopping from learner corpus which was described in Chapter 4. Each participant uttered orally translated English speech corresponding to Japanese sentences displayed on a screen. 20 participants uttered orally translated English speech corresponding to about 200 Japanese sentences and the other 35 participants uttered speech corresponding to about 100 utterances. Then, we developed a 2-gram language model from 5,000 transcribed utterances spoken by these 55 Japanese university students and transcribed utterances spoken by English native speakers. The pronunciation lexicon consisted of about 35,000 word types related to conversations about travel abroad.

6.1.2 Efficiency of Splitting Methods

In order to evaluate the efficiency of proposal which using a PDT-based top down method for improving speech recognition accuracy for the second language speech, we compared the performance of the reduced phoneme set with the PDT-based top-down method and that of the reduced phoneme set splitting in the manner of the top-down method using only the phonetic distance between each phoneme. We used the Linde-Buzo-Gray algorithm [103] as the top-down splitting method to obtain the reduced phoneme set. Euclidean distance was used to calculate every two mean vectors of phoneme. The splitting process was repeated until obtaining the final cluster number. We used the same evaluation data with that described in Section 5.7.3.

Figure 6.1 shows the word accuracies by different numbers of phoneme sets that were determined with the PDT-based top-down method and a top-down splitting method using only phonetic distance. We can observe the following:

- The reduced phoneme sets that were determined with the PDT-based top-down method achieved better performance than that of the top-down splitting method using only phonetic distance.
- There were significant differences between word accuracies obtained with our method and the method using only phonetic distances for the phoneme set of 38, 28, 25 (paired t-test, $t_{(19)} = 4.11$, $p < 0.001$ for 38, 28, 25 phonemes; $p < 0.05$ for 32, 27, 26 phonemes).

6.1.3 Effect of Phoneme Occupation Probabilities

We conducted an experiment to examine the effect of phoneme occupation probability γ on recognition accuracy and compared it with the performance in a case in which phoneme occupation probability γ was not used as a reference. We used two corpora for calculating γ: the domain-independent one based on ERJ database mentioned in Chapter 4 and the domain-dependent one consisting of 3,464 transcribed utterances by 34 university students. This experiment used the same evaluation data as that described in Section 5.7.3.

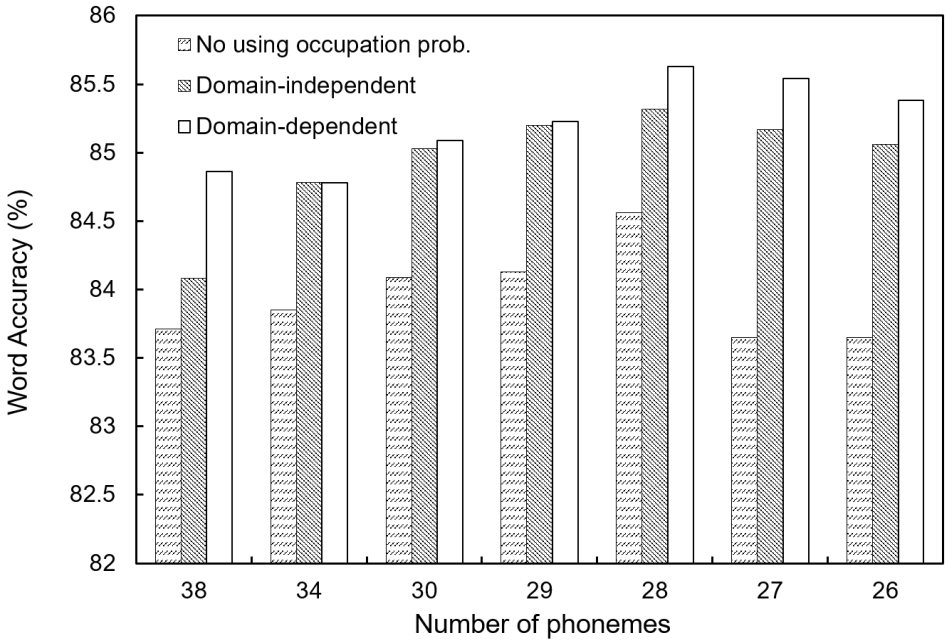


Figure 6.2: Word accuracy with different phoneme occupation probabilities in the same number of phoneme sets.

Figure 6.2 shows the word accuracies with various numbers of phoneme sets trained with two different phoneme occupation probabilities and with not using occupation probability. The experiments showed that:

- The 28-phoneme set rendered the highest word accuracy in the reduced phoneme set for both phoneme occupation probabilities and also for the one without using it.
- The phoneme occupation probability trained with domain-independent data achieved higher word accuracies than that without using it.
- There were no significant differences between word accuracies trained with domain-dependent data and domain-independent data for all reduced phoneme sets.

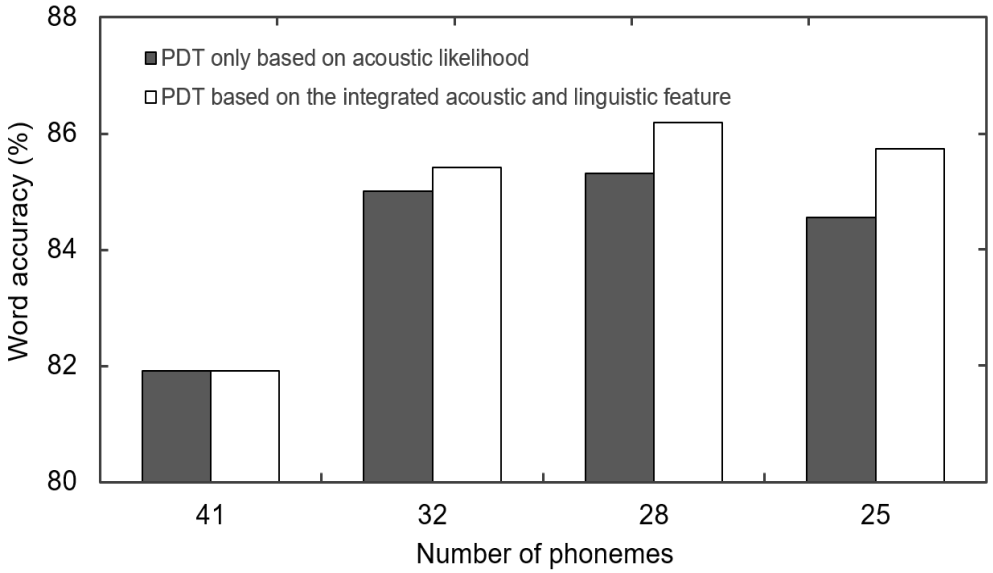


Figure 6.3: Word accuracy of the canonical phoneme set and various reduced phoneme sets by PDT only based on the acoustic likelihood and PDT based on the integrated acoustic and linguistic features.

6.2 Analysis of Effect of Linguistic Feature

In order to examine the effectiveness of linguistic feature, we compared the performance of ASR implementing the proposed method and analyzed the experimental results from the following two aspect: RPSs by different methods, which is phonemes generated by the proposal only using the acoustic likelihood compared with those generated by the integrated method mentioned in Chapter 5 and analysis of detailed phonemes in final clusters with different methods.

6.2.1 Reduced Phoneme Set by Different Methods

We used the hidden markov model toolkit (HTK) toolkit [54] to compare the performance on ASR implementing the proposed method based on the integrated acoustic and linguistic feature with that of the canonical phoneme set and the reduced phoneme sets generated by the method only based on the acoustic likelihood, which was introduced in Section 5.6.3.

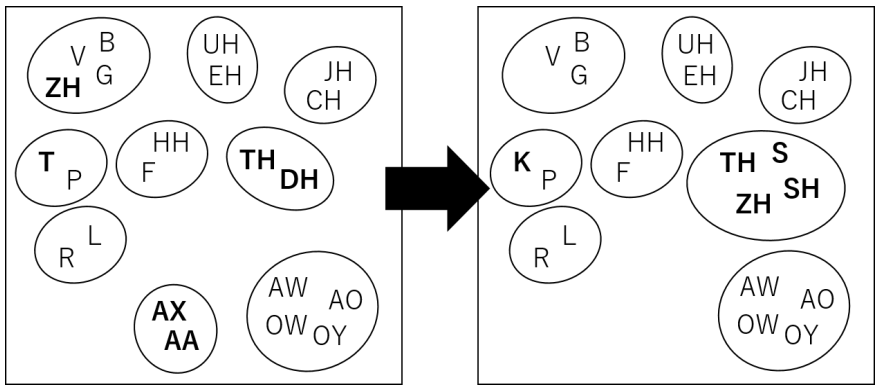


Figure 6.4: Final clusters of 28-phoneme sets generated by both PDT only based on the acoustic likelihood (left) and PDT based on unified acoustic and linguistic objective function (right). The different phonemes are shown in bold. (Non-merged phonemes are not included in the figure.)

Figure 6.3 shows the word accuracies of the canonical phoneme set and various reduced phoneme sets by PDT only based on the acoustic likelihood and PDT based on the integrated acoustic and linguistic features. The experimental results showed that the reduced phoneme sets with integrated acoustic and linguistic features delivered a better performance than the canonical phoneme set and other reduced phoneme sets by PDT only using acoustic features.

6.2.2 Analysis of Detailed Phonemes in Different Methods

The previous experimental results in Section 5.7.4 showed that the reduced phone sets by the method based on the unified acoustic and linguistic objective function delivered better performance than those with PDT only based on the acoustic likelihood. In order to further clarify the efficiency of the method based on the unified acoustic and linguistic objective function, we compare phoneme sets in the final clusters created with the proposed method and PDT only based on acoustic likelihood.

Figure 6.4 shows an example of phoneme sets in final clusters, which are merged phonemes in the leaves of a decision tree when generating 28-phoneme sets. The left figure depicts the clusters obtained by PDT only based on the acoustic likelihood, and the right one depicts the results obtained by the method based on the unified acoustic and linguistic objective function.

Some phonemes are differently merged into phoneme sets in the left and right figures depending on the difference of criterion of both methods.

One of the specific features of the method based on the unified acoustic and linguistic objective function compared to PDT only based on the acoustic likelihood is clear from the result that phonemes /P/, /T/, and /K/ are differently merged in the right and left figures in Figure 6.4. Phonemes /P/, /T/, and /K/ have the same manner of articulation (plosive), which forces them to be merged into a cluster based on the acoustic feature, but they have a different place of articulation (labial, dental, palatal). In the left figure, which shows the results with PDT only based on the acoustic likelihood, phonemes /T/ and /P/ are merged because the place of articulation between /T/ and /P/ is nearer than that between /K/ and /P/. On the other hand, phonemes /K/ and /P/ are merged in the right figure on the basis of proposed method. This difference can be explained by the fact that the probability of homophones that have the same phoneme strings when /K/ and /P/ are merged achieved 0.9% absolute reduction in comparison to that of homophones that have the same phoneme strings when /T/ and /P/ are merged. Experimental results show that the linguistic discriminating ability decreases more when /T/ and /P/ are merged.

6.3 Discussion

The experiment results described in Chapter 6 can be summed up as follows.

- The reduced phoneme sets provided better word accuracies than the canonical one, and the 28-phoneme set obtained the best performance. Compared to the canonical phoneme set (41 phonemes), the PDT-based top-down method reduced word errors from 18.1% to 14.7%, a relative error reduction rate of 18.5%. There was a significant difference between the word accuracy of the canonical phoneme set and that of the 28-phoneme set (paired t-test, $t_{(19)} = 4.04$, $p < 0.001$ for 28 phonemes).
- The word accuracy of the proposed PDT-based top-down method is better for each reduced phoneme set than that of the top-down splitting method using only the phonetic distance

between each phoneme, as discussed in 6.2.2. The recognition results demonstrate that the discrimination rules function effectively in designing the reduced phoneme set.

- The recognition results discussed in Section 6.1.3 show that calculating the phoneme occupation probability improved recognition accuracy when using it as the weight of the splitting criterion. The phoneme occupation probability trained with the domain-dependent corpus gave a slightly better performance than that of the domain-independent corpus for all numbers of phoneme clusters. However, the phoneme occupation probabilities did not significantly change regardless of task domain, and it is not necessary to re-train HMMs of different phoneme clusters depending on each target task.

Appendix Figure B.1 and B.2 show examples of detailed cluster splitting with the PDT-based top-down method and that with a top-down splitting method using only phonetic distances to obtain a phoneme set with 28 phonemes, respectively.

We evaluated the effect of word accuracy rate (WAR) increase on the performance of the dialogue-based CALL system from the viewpoint of providing effective feedback, which is one of the features contributing to the system performance [104], [105]. We compared the ratio of returning effective feedback corresponding to each ungrammatical/unacceptable expression between the two conditions. This method showed that the ratio of returning correct feedback was 68% even with a simple exact pattern matching between the recognition results and ungrammatical/unacceptable expressions stored in the learner corpus. The relative error reduction is 9.8% compared with the conventional method using the canonical phoneme set.

6.4 Summary

We adopt maximization of the weight total of a phoneme set's acoustic likelihood and its linguistic discrimination ability to derive the optimal phoneme set in Chapter 5. In order to verify the efficiency of proposed method, the analysis of effect of acoustic feature and linguistic feature is introduced and investigated in this chapter in detail.

In this Chapter, we analyse the effect of acoustic and linguistic features based on the reduced phone sets by the method with PDT only based on the acoustic likelihood and that by the method with PDT-based on the unified acoustic and linguistic objective function. The speech recognition results obtained for second language speech collected with a translation game type dialogue-based CALL system showed the effectiveness of acoustic and linguistic features which were both used in our proposal method.

7 Analysis of the Relation Between Proficiency Level and the Phoneme Set

> *Whenever you look at a piece of work and you think the fellow was crazy, then you want to pay some attention to that. One of you is likely to be, and you had better find out which one it is. It makes an awful lot of difference.*

Charles Franklin Kettering, *U.S. engineer and inventor*

The speech quality of second language speakers overall depends on their proficiency level in the second language [20], [21], and there are different patterns in accent among inexperienced, moderately experienced, and experienced speakers [23], [3], [25], [58]. As a result, influence of the mother tongue on pronunciation varies widely.

In this chapter, we investigate the relation between our reduced phoneme set and the English proficiency level of speakers. On the basis of the results of this investigation, we propose a novel method to improve the second language speech recognition performance when the mother tongue of speakers is known. We evaluated the proposed method by using speech data collected by a previously developed dialogue-based English CALL system in the form of a translation exercise for Japanese students.

The rest of this chapter is organized as follow. Section 7.1 describes the relation between the

reduced phoneme set and the English proficiency level of L2 speakers is discussed. In Section 7.2, we propose using multiple reduced phoneme sets for second language speech recognition. Then in Section 7.3, we discuss the experimental results. Finally, we summarize this chapter.

7.1 Relation Between Optimal Phoneme Set and L2 Speakers with Different Proficiencies

Both results of Chapter 5 and 6 demonstrated that phoneme mismatches resulting in mis-recognition of L2 speech can be improved by reducing the number of phonemes. The tendency of mispronunciation depends on the average proficiency level of L2 speakers. It is generally expected that phoneme mismatch is more frequent in speech by those with low level proficiency and that the optimal number of reduced phonemes may vary depending on speaker proficiencies. In order to examine this hypothesis, we conducted an experiment to determine the relation between the reduced phoneme set and the proficiencies of speakers classified by both top-down and bottom-up methods, as described below. The experimental conditions and results are presented in the next subsections.

7.1.1 Participant Information

We used 3,150 orally translated speech items from learner corpus mention in Chapter 4 as evaluation data. Their speaking styles range widely from ones similar to conversation to ones closer to read-speech. There were a total of 45 participants aged 18 to 24 who had acquired Japanese as their mother tongue and learned English as their second language. The Test of English for International Communication (TOEIC) [53] score was used for measuring the English proficiency of the speakers.

The participants' English proficiencies ranged from 300 to 910 (990 being the highest score that can be attained) in our experiments. We divided their scores into 5 levels based on TOEIC score

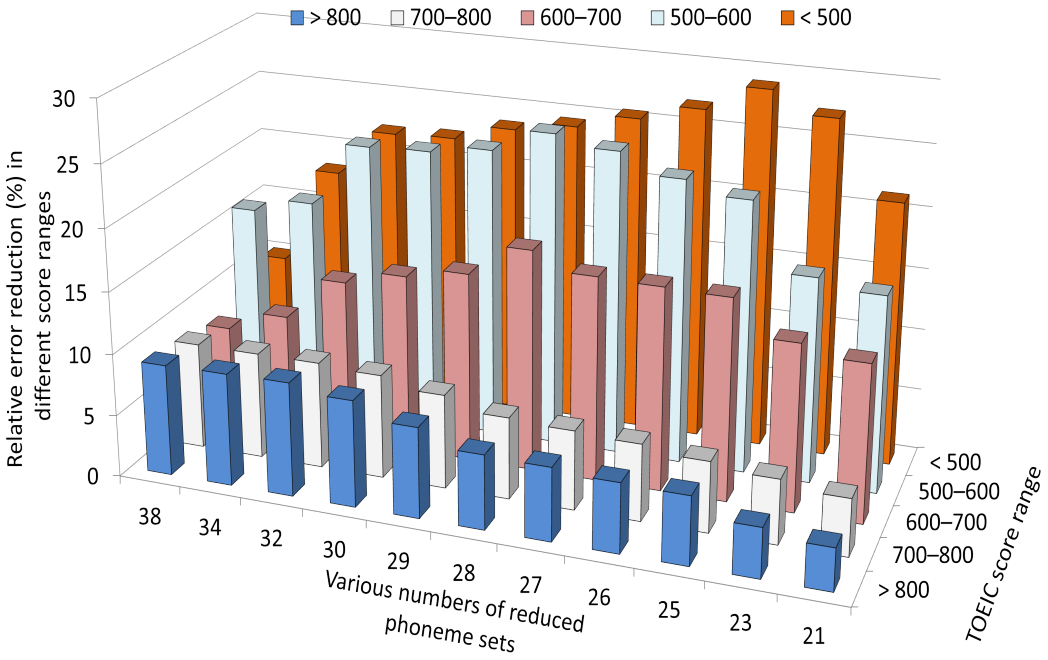


Figure 7.1: Relative error reduction for speech by participants in different TOEIC score ranges.

range: lower than 500, 500−600, 600−700, 700−800, and higher than 800, with 10, 10, 10, 8, and 7 participants in each respective score range. Based on the research on proficiency evaluation of foreigner [23], [25], [64], we define that the score rang which is lower than 500 pertains as "lower-level proficiency", a 500–700 score is the part of "middler-level proficiency", and scores higher than 700 belong to "higher-level proficiency" in our study.

7.1.2 Recognition Performance with Proficiencies-based Clustering

We used the HTK toolkit [54] to compare the ASR performance using canonical and reduced phoneme sets for speech by Japanese participants at each level of the 5 TOEIC score ranges. The reduced phoneme sets investigated in this study were derived by the proposal method described in Chapter 6. Figure 7.1 shows the relative error reduction of various reduced phoneme sets compared with the canonical one for speech by participants of each of the 5 TOEIC score ranges. It shows that

- All reduced phoneme sets achieved error reduction compared with the canonical phoneme set for all TOEIC score ranges.
- The optimal phoneme number of reduced phoneme sets varies depending on the English proficiency level of the speakers.
- The recognition performance of the reduced phoneme set was improved more for speech by participants with lower-level proficiency than for those with higher-level proficiency.
- The optimal number of phonemes for the speakers with lower-level proficiency was smaller than that for those with higher-level proficiency.

The results of Figure 7.1 suggest that the optimal reduced phoneme set corresponding to the English proficiency of speakers is a 25-phoneme set for speakers with a TOEIC score of less than 500, a 28-phoneme set for those with a 500−700 score, and a 32-phoneme set for those with scores higher than 700.

This investigation result showed that the effect of the reduced phoneme set is different depending on the proficiency of second language speakers. In addition, the proficiency of second language speakers various widely, it is inadequate to use single phoneme set to recognize input speech by all second language speakers.

7.1.3 Recognition Performance with Speaker-by-Speaker Basis

We assume that overall language proficiency would correlate roughly with goodness of pronunciation to obtain the results in Section 7.1.2. That is to say, we assume that the speech quality collected from a group of L2 speakers of higher proficiency would be better, on average, than that of lower proficiency. Various researches have already clarified the factors affecting goodness of pronunciation by L2 speakers [3], [25], [59]. However, there is still controversy when it comes to using a standardized test such as TOEIC to classify participants into groups from the perspective of correlation with goodness of speech quality and proficiency [64].

In this work, we used a method in which the language proficiency of participants was not utilized

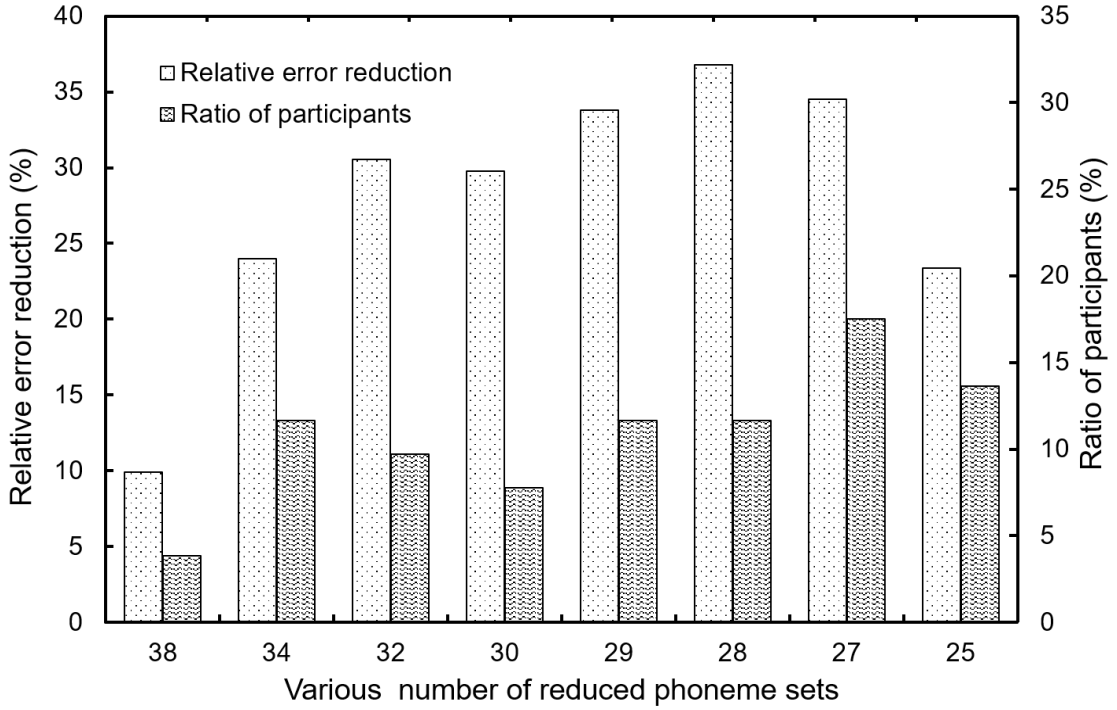


Figure 7.2: Ratio of participants in their optimal reduced phoneme set and relative error reduction for speech by speakers achieved the best recognition accuracy.

for classifying groups, as opposed to the top-down method. In this method, which we call the *"bottom-up method"*, we count the number of participants achieving the best recognition accuracy for each reduced phoneme set ranging from 38 to 21. The black bars and white bars in Figure 7.2 depict the ratios of participants (the number of participants achieving the best recognition accuracy divided by the total number of participants) and relative error reduction compared with the canonical one for speech by participants who achieved the best recognition accuracy for the corresponding reduced phoneme set, respectively. We found that

- The distribution of the set number of the optimal reduced phoneme set seems to have multiple peaks.
- It is not sufficiently clear because of a shortage of data, but 34-RPS, 27-RPS, and 25-RPS achieved higher ratios of participants than their surrounding phoneme sets. The numbers of reduced phoneme set achieving more relative error reduction are a little shifted from the numbers achieving higher ratios, and 32-RPS, 28-RPS, and 25-RPS achieved more relative

error reduction than their surrounding phoneme sets.

- The bottom-up method showed almost the same result for selecting multiple reduced phoneme sets as the top-down method for our collected speech data.

7.2 Summary

In this Chapter, we clarified the relation between the second language speakers and an optimal reduced phoneme set. On the basis of this analysis, we then investigated the recognition performance with speaker-by-speaker basis. The analysis results showed that multiple reduced phoneme set maybe further improve the recognition performance of L2 speech.

8 Multiple-Pass Decoding with Lexicon Represented by Multiple Sets

> *Knowledge has three degrees–opinion, science, and illumination. The means or instrument of the first is sense; of the second, dialectic; of the third, intuition. This last is absolute knowledge founded on the identity of the mind knowing with the object known.*

Plotinus, *founder of Neoplatonism*

In this chapter, we propose an ASR system using multiple reduced phoneme sets to further improve the recognition performance of second language speech considering speakers' proficiency level. Based on the experimental results in Chapter 7, we selected 25-, 28- and 32-phoneme sets as the components of multiple reduced phoneme sets to capture the various proficiency levels of second language speakers.

We assumed that one of three multiple phoneme sets, not a mixture, is used to recognize input speech by a single second language speaker in order to resolve the problem caused by Viterbi decoding (detailed in Section 8.2). To fulfill this constraint, we developed a lexicon in which the pronunciation of each lexical item is represented by the multiple reduced phoneme sets and a language model implementing the constraint, as described in the following.

8.1 Lexicon

Some phonemes in the canonical phoneme set are differently distributed among the 25-phoneme, 28-phoneme, and 32-phoneme sets. Specifically, some lexical items are represented by a single phoneme set sequence consisting only of phonemes without merging or merged into the same clusters by the phonetic decision tree (PDT) in three reduced phoneme sets. Other lexical items in the lexicon are represented with three different phoneme set sequences in the multiple reduced phoneme sets. Figure B.3 in Appendix B shows the result of cluster splitting with PDT in which 25, 28, and 32 phonemes were obtained as the final phoneme set and depicts phonemes of single and different phoneme set sequences.

In the lexicon, 62.9% of the lexical items have a single phoneme set sequence and 37.1% have multiple phoneme set sequences used for the experiment. We added a symbol that differentiates words of the single phoneme set sequence from those of the multiple phoneme set sequences.

8.2 Language Model

A simple method for training a stochastic language model is to train a language model independently of the structure of the lexicon. This can be done easily by counting word occurrences in the training corpus, assuming that words represented with multiple phoneme set sequences have multiple pronunciations.

Since probabilities leaving the start arc of each word must add up to 1.0, each of these pronunciation paths through this multiple-pronunciation HMM word model will have a smaller probability than the path through a word with only a single pronunciation path. A Viterbi decoder can only follow one of these pronunciation paths and may ignore a word with many pronunciations in favour of an incorrect word with only one pronunciation path. It is well known that the Viterbi approximation penalizes words with many pronunciations [22].

In order to resolve degradation of speech recognition performance stemming from multiple pro-

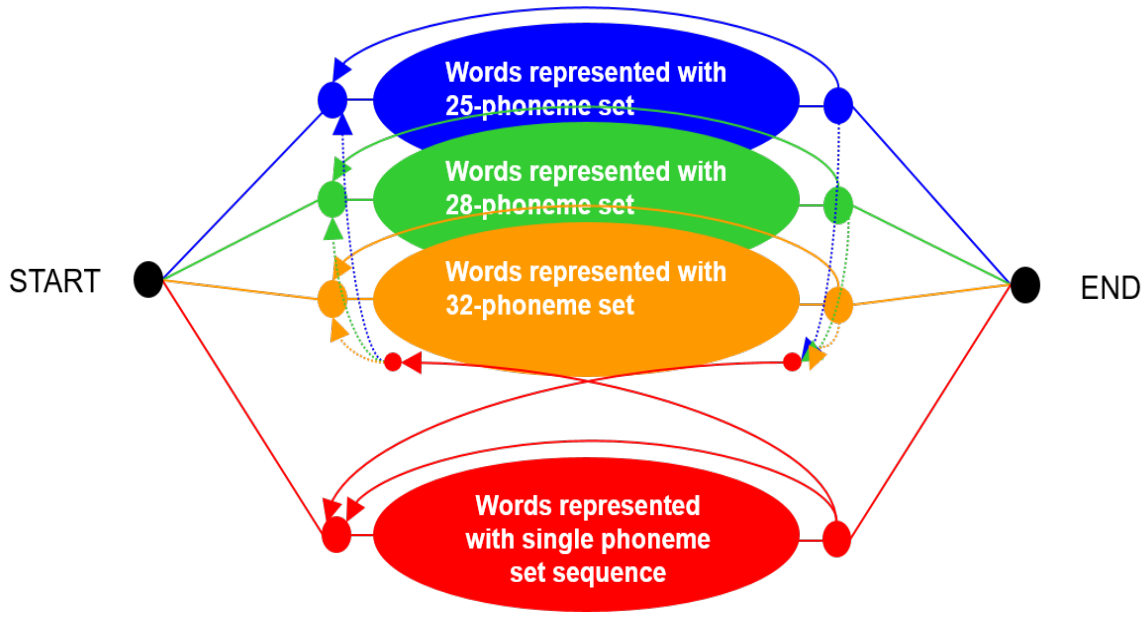


Figure 8.1: Schematic diagram of language model for words represented by 25-, 28-, and 32-phoneme sets and for words with single phoneme set sequence. Arcs depict transition between words represented by the same reduced phoneme set and words of single pronunciation. The transition among words represented by different reduced phoneme sets is inhibited.

nunciations, our language model only permits transition between words represented by the same reduced phoneme set and words of the single pronunciation while inhibiting transition among words represented by different reduced phoneme sets, as shown in Figure 8.1.

8.3 Multiple-Pass Decoding

In order to take complete advantage of language model implementing the lexical items with increased multiple pronunciations, we use a multiple-pass decoding algorithm that modifies the Viterbi algorithm to return the N-best word sequences for a given speech input. The bigram grammar mentioned in Section 8.2 is used in the first pass with an N-best Viterbi algorithm to return the 50 most highly probable sentences. The 50-best list is used to create a more sophisticated LM that allows transition among the words with multiple reduced phoneme set sequences. This LM allows transition among the words with different reduced phoneme set calculates like-

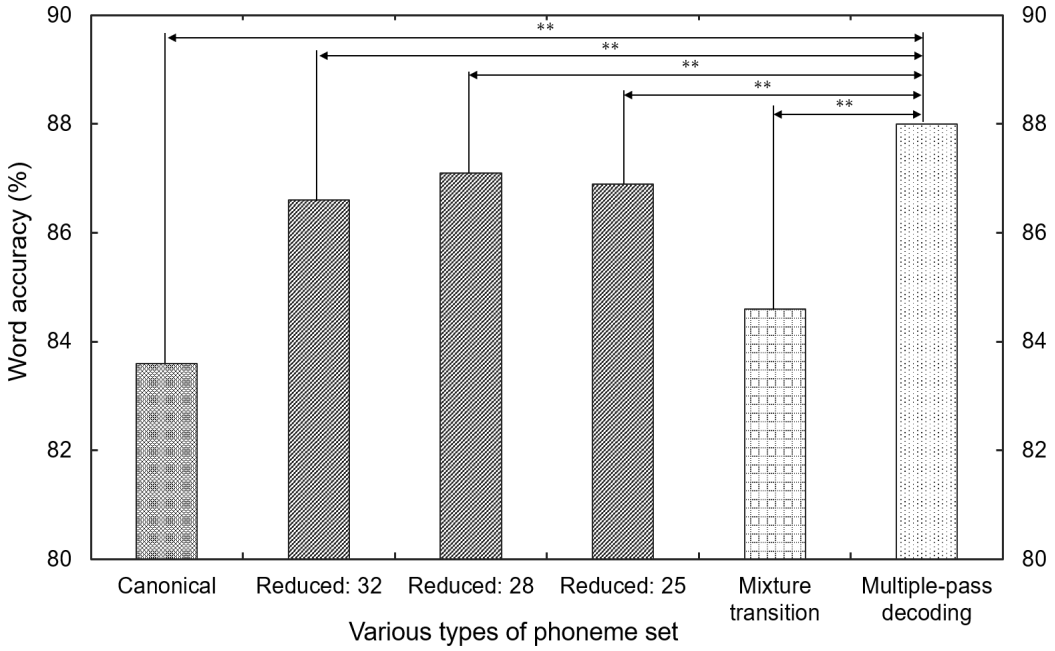


Figure 8.2: Word accuracy of canonical phoneme set, three single reduced phoneme sets, mixture transition and multiple-pass decoding with multiple reduced phoneme sets. ** indicates a significant difference between the word accuracy of multiple-pass decoding with the multiple reduced phoneme sets and other ones ($p < 0.01$).

lihood of concatenating words with different reduced phoneme set in the second pass. Sentences are thus re-scored and re-ranked according to this more sophisticated probability [22].

8.4 Experimental Results

For evaluating the proposed method, we compared the performance on ASR implementing the proposed method with that of the canonical phoneme set and three single reduced phoneme sets. We also compared the multiple-pass decoding with a mixture transition that allows transitions among all words represented with multiple reduced phoneme set sequences in the first-pass.

Figure 8.2 shows the word accuracy of the canonical phoneme set, various single reduced phoneme sets, mixture transition, and multiple-pass decoding with the multiple reduced phoneme sets. We

observed the following:

- Multiple-pass decoding with multiple reduced phoneme sets had a better performance than the canonical phoneme set, all of the single reduced phoneme sets, and the mixture transition.
- There were significant differences between the word accuracy of multiple-pass decoding with multiple reduced phoneme sets and the canonical phoneme set, the 32-phoneme set, the 28-phoneme set, the 25-phoneme set, and the mixture transition (paired t-test, $t_{(44)} = 2.02$, $p < 0.01$).

8.5 Discussion

8.5.1 Efficacy of the Multiple Reduced Phoneme Sets

In order to explore the efficiency of the multiple reduced phoneme sets further, we compared the relative error reduction of the proficiency-dependent reduced phoneme set and multiple reduced phoneme sets. A comparison of the relative error reduction of the proficiency-dependent reduced phoneme set and multiple reduced phoneme sets for speech by speakers in each TOEIC score range is shown in Figure 8.3. We find that:

- The multiple reduced phoneme sets achieved better performance than the proficiency-dependent reduced phoneme set for speech by speakers at all language proficiency levels.
- There were significant differences between the relative error reduction of the multiple reduced phoneme sets and the proficiency-dependent reduced phoneme set (paired t-test, $t_{(9)} = 2.26$, $p < 0.05$) for scores lower than 500 and (paired t-test, $t_{(6)} = 2.57$, $p < 0.05$) for scores higher than 800.

The results of Figure 8.3 showed that the most highly improved performances appeared at the lowest and highest proficiency levels. The experimental results of our previous studies showed

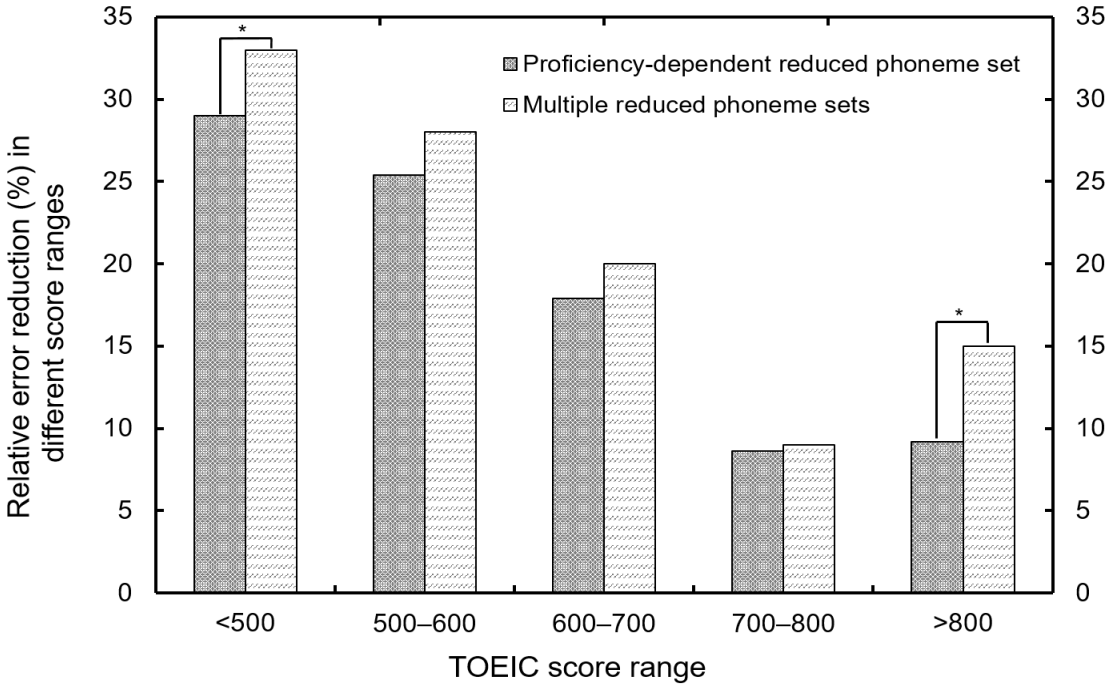


Figure 8.3: Relative error reduction of proficiency-dependent phoneme set and multiple reduced phoneme sets in different TOEIC score ranges. * indicates a significant difference between the relative error reduction of multiple reduced phoneme sets and the proficiency-dependent one ($p < 0.05$).

that a single 28-reduced phoneme set provided better performance in comparison with other reduced ones for speech by L2 speakers on average [33]. 25-phoneme and 32-phoneme sets, which have slightly different characters from the 28-phoneme set, were assigned as the proficiency-dependent reduced phoneme set to L2 speakers of lowest and highest proficiency levels, respectively. However, goodness of pronunciation varies among L2 speakers of the same proficiency level. As a result, because the optimal reduced phoneme set is selected for speech by each L2 speaker, the multiple reduced phoneme set can accurately recognize more utterance of speakers of lowest and highest proficiency levels in comparison with the proficiency-dependent reduced phoneme set.

Table 8.1: Word error rates by speech recognizers using the proposed method, parallel processing of distinct speech recognizers, and language model allowing mixture of reduced phoneme sets.

The proposed method	Parallel processing of distinct speech recognizers	Language model allowing mixture of reduced phoneme sets
12.4%	12.3%	15.4%

8.5.2 Efficacy of Language Modeling

As discussed in subsection 4.2, the best way to avoid problems stemming from multi-pronunciations is to recognize speech distinctively with multiple speech decoders with distinct language models represented by 25-, 28-, and 32-phoneme sets. In this experiment, the parallel structure at the phoneme level enables the decoder to match the acoustic models of the 25-, 28-, and 32-phoneme sets during the decoding process. There have been several previous studies on using multiple recognizers for recognition improvement. One of the more popular approaches is ROVER (recognizer output voting error reduction) developed by NIST [106], which is mainly used to reduce the word error rates of ASR by majority vote from multiple speech recognizers. Other conventional methods are to select the most likely recognition result from different ones through multiple recognizers of various acoustic and/or language models by comparing the model likelihood of different recognition results [107], [108].

A language model that allows transitions among all words whose pronunciation is represented with multiple reduced phoneme sets was also trained for a speech recognition system assuming that speech by a single speaker is represented with a mixture of multiple reduced phoneme sets. These methods were compared with the proposed method of multiple reduced phoneme sets.

A comparison of the word error rates by the three methods (Table 8.1) showed that

- The language model allowing a mixture of multiple reduced phoneme sets achieved a lower recognition performance than other methods.
- The proposed method of multiple reduced phoneme sets achieved almost the same recog-

nition performance as parallel processing by distinct acoustic models of multiple reduced phoneme sets. There was no significant difference between the word accuracy of the proposed method and parallel processing of distinct speech recognizers.

In the proposed method, the transfer probability from the start state to each word of multiple pronunciation is reduced by one-third. However, the results in Table 8.1 suggest that this effect is negligible.

A disadvantage of the parallel processing of the distinct speech recognizer is that it increased the amount of recognition processing required.

8.6 Summary

In this chapter, on the basis of the analysis of the relation between the proficiencies of speakers and an optimal reduced phoneme set, we proposed a novel speech recognition technique using multiple-pass decoding with multiple reduced phoneme sets. Multiple reduced phoneme sets with three different reduced phoneme sets were constructed to capture the various proficiency levels of second language speakers. The proposed method was able to further improve the recognition performance for second language speech for each proficiency level compared with the canonical phoneme set and various single reduced phoneme sets.

9 Conclusion and Future Work

> *As a technology, the book is like a hammer. That is to say, it is perfect: a tool ideally suited to its task. Hammers can be tweaked and varied but will never go obsolete. Even when builders pound nails by the thousand with pneumatic nail guns, every household needs a hammer.*

James Gleick, *american author, historian of science, and Pulitzer Prize winner*

9.1 Conclusions

This dissertation consists in second language speech recognition from phonological knowledge among mother tongue, second language and their phonetic features, acoustic features and linguistic features of second language, specially when the mother tongue of speakers is know. Owing to the speech by second language speakers different from that by native, recognition of second language speech is a challenging task even for state-of-the-art ASR systems. The thesis address three challenges of second language ASR: the lack of advantage of L2 speech with limited vocabulary and less knowledge of grammatical structures, mismatching issues on the phoneme-level for confused pronunciation or mispronunciation by L2 speakers, and various proficiency levels of L2 speakers.

To tack above challenges, we proposed a method of designing a customized phoneme set for second language speech maximizing a unified acoustic and linguistic objective function of second language speakers and implemented the method as a decision tree to derive a reduced phoneme set.

To further investigate the performance of second language ASR system with customized phoneme sets, this dissertation exploits the relation between proficiency level and the customized phoneme set to build a proficiency-dependent phoneme set to capture different pronunciation variations by second language speakers.

We applied the reduced phoneme set developed with the proposed method to English utterances spoken by Japanese collected with a translation game type dialogue-based CALL system. The experimental results presents a greater improvement in speech recognition performance than the canonical phoneme set and other single customized phoneme set and verified that the proposed method is effective for ASR that recognizes second language speech when the mother tongue of users is known.

9.2 Future Works

This dissertation presents the feasibility of building a speech recognizer with the proposed methods is able to alleviate the problem caused by confused mispronunciation by second language speakers. It points to the potential research direction, such as to carry on examining linguistic discrimination ability based on more accurate word occurrence probability in other corpora. Since collecting a huge amount of speech data of non-native speakers of various proficiencies is still quite difficult, a direction to use the occurrence probability of each word in a native speech corpus or its interpolation with the probability can be considered to obtain in a small corpus of non-native speakers as an approximate approach.

Bibliography

[1] C. Kramsch and S. Thorne, "Foreign Language Learning As Global Communicative Practice," *Globalization and language teaching*, pp. 83–100, 2002. 1

[2] R. Kubota, "The Impact of Globalization on Language Teaching in Japan," *Globalization and language teaching*, pp. 13–28, 2002. 1

[3] J.E. Flege, E. M. Frieda, and T. Nozawa, "Amount of Native-language (L1) Use Affects the Pronunciation of An L2," *The Journal of Phonetics*, vol. 25, no. 2, pp. 169–186, 1997. 2, 6, 24, 73, 76

[4] N. Poulisse and T. Bongaerts, "First Language Use in Second Language Production," *in Handbook of Applied Linguistics, Oxford University Press*, 15.1, pp. 36–57, 1994. 2, 6, 23, 49

[5] N. Minematsu, "Perceptual and Structural Analysis of Pronunciation Diversity of World Englishes," *In Proceedings of International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (Oriental-COCOSDA)*, Keynote 2, 2014. 2, 6

[6] T. Riney and J. Anderson-Hsieh, "Japanese Pronunciation of English," *The Journal of Japan Association for Language Teaching (JALT)*, vol.15, no. 1, pp. 21–36, 1993. 2, 6, 29, 30, 49

[7] A.C. Gimson, "An Introduction to the Pronunciation of English," *The Journal of the International Phonetic Association*, vol. 10, Issue 1-2, pp. 80–81, 1980. 33

[8] F. Ramus, "Acoustic correlates of linguistic rhythm: Perspectives," *In Proceedings of International Conference Speech Prosody*, pp. 115–120, 2002. 33

[9] K.E. Schairer, "Native Speaker Reaction to Non-native Speech," *The Journal of the Mod-*

ern Language, vol.76, no. 3, pp. 309–319, 1992. 2

[10] E.M. Varonis and S. Gass, "The Comprehensibility of Non-native Speech," *the Journal of Studies in Second Language Acquisition (SSLA)*, vol.4, no.2, pp. 114–136, 1982. 3, 22

[11] A. Mizumoto and T. Shimamoto, "A Comparison of Aural and Written Vocabulary Size of Japanese EFL University Learners," *Language Education & Technology*, vol. 45, pp. 35–51, 2008. 3

[12] S. Ishikawa, T. Uemura, M. Kaneda, S. Shimizu, N. Sugimori, Y. Tono, and M. Murata, "JACET 8000: JACET List of 8000 basic words," *Tokyo: JACET*, 2003. 3

[13] F.Ehsani and E. Knodt, "Speech Technology in Computer-aided Language Learning: Strengths and Limitations of A New CALL Paradigm," *Language Learning & Technology*, vol. 2, no. 1, pp. 45–60, 1998. 3

[14] H. Wang, T. Kawahara, and Y. Wang, "Improving Non-native Speech Recognition Performance by Discriminative Training for Language Model in a CALL System," *In Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2011. 3

[15] W. Byrne, E. Knodt, S. Khudanpur, & J. Bernstein, "Is Automatic Speech Recognition Ready for Non-native Speech? A Data Collection Effort and Initial Experiments in Modeling Conversational Hispanic English," *In Proceedings of Speech Technology in Language Learning (STiLL)*, vol. 1, no. 99, 1998. 2, 4

[16] P. Langlais, A.M. Öster, and B. Granström, "Automatic Detection of Mispronunciation in Non-native Swedish Speech," *In Proceedings of The European Speech Communication Association (ESCA) workshop focusing on Speech Technology in Language Learning (STiLL)*, Marholmen, Sweden, 1998. 4

[17] D.A. Van Leeuwen and R. Orr, "Speech Recognition of Non-native Speech Using Native and Non-native Acoustic Models," *The Netherlands: TNO Human Factors Research Institute*, 2000. 4

[18] J. Esling and R. Wong, "Voice Quality Settings and the Teaching of Pronunciation," *TESOL Quarterly*, vol. 17, pp. 89–95, 1983. 30

[19] P. Fetter, "Detection and Transcription of OOV Words," *Ph.D. Dissertation*, TU Berlin, Germany, 1998. 16

[20] K. Osaki, N. Minematsu, and K. Hirose, "Speech Recognition of Japanese English Using Japanese Specific Pronunciation Habits," *IEICE Technical report*, SP2002-180, 2003. 6, 73

[21] J. Van Doremalen, C. Cucchiarini, and H. Strik, "Optimizing Automatic Speech Recognition for Low-proficient Non-native Speakers," *EURASIP Journal on Audio, Speech, and Music Processing*, 2010. 6, 73

[22] D. Jurafsky and J.H. Martin, "Speech & Language Processing," *Pearson Education India*, 2000. 80, 82

[23] P. Trofimovich and W. Baker, "Learning Second Language Suprasegmentals: Effect of L2 Experience on Prosody and Fluency Characteristics of L2 Speech," *The Journal of Studies in Second Language Acquisition (SSLA)*, vol. 28, no. 01, pp. 1–30, 2006. 6, 73, 75

[24] R.E. Gruhn, W. Minker, and S. Nakamura, "Statistical Pronunciation Modeling for Non-native Speech Processing," *Springer Berlin Heidelberg*, 2011. 1, 5, 16, 23, 28, 33, 49

[25] J.E. Flege, "Factors Affecting Degree of Perceived Foreign Accent in English Sentences," *The Journal of the Acoustical Society of America*, vol. 84, no. 1, pp. 70–79, 1988. 6, 28, 73, 75, 76

[26] M. Kenstowicz, "Phonology in Generative Grammar," *Cambridge/MA, Oxford: Blackwell*, 1994. 25

[27] K. Livescu, "Analysis and Modeling of Non-native Speech for Automatic Speech Recognition," *Diss. Massachusetts Institute of Technology*, 1999. 4, 25, 41

[28] S. Schaden, "Generating Non-native Pronunciation Lexicons by Phonological Rule," *In Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2004. 24, 41

[29] Z. Wang, T. Schultz, and A. Waibel, "Comparison of Acoustic Model Adaptation Techniques on Non-native Speech," *In Proceedings of International Conference on Acoustics,*

Speech, and Signal Processing (ICASSP), 2003 25

[30] Y. Nagai, T. Senzai, S, Yamamoto, and M. Nishida, "Sentence Classification with Grammatical Errors and Those Out of Scope of Grammar Assumption for Dialogue-Based CALL Systems." *in Proceeding of International Conference on Text, Speech and Dialogue (TSD)*, Springer Berlin Heidelberg, 2012. 3, 33, 35

[31] C.Prator and B. W. Robinett "Manual of American English Pronunciation (4th Edition)," *Tokyo: Harcourt Brace Jovanovich Japan, Inc.*, 1986. 29, 30

[32] Y.R. Oh, J.S. Yoon, and H.K. Kim, "Acoustic Model Adaptation Based on Pronunciation Variability Analysis for Non-native Speech Recognition," *The Journal of Speech Communication*, vol. 49, no. 1, pp. 59–70, 2007. 25, 41

[33] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme Set Design for Speech Recognition of English by Japanese," *The Journal of IEICE Transactions on Information and Systems*, vol. E98-D, no. 1, pp. 148–156, 2015. 9, 41, 84

[34] K. Livescu and J. Glass, "Lexical Modeling of Non-native Speech for Automatic Speech Recognition," *In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, pp. 1683–1686, 2000. 41

[35] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED corpus lectures," *In Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. I-232–I-235, 2003. 41

[36] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme Set Design Using English Speech Database by Japanese for Dialogue-based English CALL Systems," *In Proceedings of International Conference on Language Resources and Evaluation (LREC)*, Reykjavik, Iceland, pp. 3948–3951, 2014. 3, 8, 9, 33, 35

[37] D. Vazhenina and K. Markov, "Phoneme Set Selection for Russian Speech Recognition." *In Proceedings of International Conference on Natural Language Processing and Knowledge Engineering (NLP-KE)*, Tokushima, Japan, pp. 475–478, Nov. 2011. 42

[38] B. Mak and E. Barnard, "Phone Clustering Using the Bhattacharyya Distance." *In Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Philadel-

phia, PA, vol. 4, pp. 2005-2008, Oct. 1996. 43

[39] X. Wang, J. Zhang, M. Nishida & S. Yamamoto, "A Dialogue-Based English CALL System for Japanese." *In Proceedings of National Conference on Man-Machine Speech Communication (NCMMSC)*, Guiyang, China, Aug. 2013. 3, 8, 33, 35

[40] G. Bouselmi, D. Fohr, and I. Illina, "Combined Acoustic and Pronunciation Modelling for Non-native Speech Recognition," *Computation and Language*, 2007. 25

[41] M. Eskenazi, "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and A Prototype," *Language Learning & Technology*, vol. 2, no. 2, pp. 62–76, 1999. 3, 36

[42] O. Kweon, A. Ito, M. Suzuki & S. Makino, "A grammatical error detection method for dialogue-based CALL system," *The Journal of Natural Language Processing*, vol. 12, no. 4, pp. 137–156, Dec. 2005. 36

[43] A. Ito, R. Tsutsui, S. Makino & M. Suzuki, "Recognition of English Utterances with Grammatical and Lexical Mistakes for Dialogue-Based CALL System," *In Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Australia, pp. 2819–2822, Sept. 2008. 36

[44] M. Eskenazi, "An Overview of Spoken Language Technology for Education," *The Journal of Speech Communication*, vol. 51, issue. 10, pp. 832–844, Oct. 2009. 36

[45] T. Kawahara, N. Minematsu, "Computer-Assisted Language Learning (CALL) Based on Speech Technologies." *The Journal of IEICE Transactions on Information and Systems*, vol. J96-D, no. 7, pp. 1549–1565, 2013. 36

[46] C. Wang, S. Seneff, "Automatic Assessment of Student Translations for Foreign Language Tutoring." *in Proceedings of Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT)*, Rochester, NY, pp. 468–475, April. 2007. 36

[47] Copyright 1993 Trustees of the University of Pennsylvania, "TIMIT Acoustic–Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S1, accessed April 8, 2016. xii, 35, 55, 103

[48] 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," *in Proceedings of International Congress on Acoustics (ICA)*, vol. 1, pp. 557–560, 2004. 34, 56

[49] P. R. Dixon, D. A. Caseiro, T. Oonish, and S. Furui, "The TITECH Large Vocabulary WFST Speech Recognition System," *in Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2007. 37

[50] N. Minematsu, G. Kurata, and K. Hirose, "Corpus-Based Analysis of English Spoken by Japanese Students in View of the Entire Phonemic System of English." *in Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2002. 34

[51] E. Sumita, Y. Sasaki, and S. Yamamoto, "Frontier of Evaluation Method for MT Systems," *IPSJ Magazine*, vol. 46, no. 5, 2005. 39

[52] J. D. Markel and A. H. Gray, "Linear Prediction of Speech," *Springer Verlag*, New York, 1976. 15

[53] TOEIC, "Mapping the TOEIC and TOEIC Bridge Tests on the Common European Framework of Reference for Languages," https://www.ets.org/toeic/research/mapping_toeic, accessed April 8, 2016. 39, 57, 74

[54] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, and D. Povey, *HTK Speech Recognition Toolkit Version 3.4, Cambridge University Engineering Department*, 2006. 57, 67, 75

[55] A. Mizumoto and T. Shimamoto, "A Comparison of Aural and Written Vocabulary Size of Japanese EFL University Learners," *Language Education & Technology*, vol. 45, pp. 35–51, 2008. 61

[56] A. Oppenheim, et al., "Digital Signal Processing," *Research Laboratory of Electronics (RLE) at the Massachusetts Institute of Technology (MIT)*, 1987. 15

[57] S. Ishikawa, T. Uemura, M. Kaneda, S. Shimizu, N. Sugimori, Y. Tono, and M. Murata, "JACET 8000: JACET List of 8000 basic words," *Tokyo: JACET*, 2003. 61

[58] J.E. Flege, M.J. Munro, and I.R.A. MacKay, "Factors Affecting Strength of Perceived Foreign Accent in A Second Language," *The Journal of the Acoustical Society of America*, vol. 97. no. 5, pp. 3125–3134, 1995. 21, 73

[59] G.J. Ockey, D. Koyama, E. Setoguchi, and A. Sun, "The Extent to Which TOEFL iBT Speaking Scores Are Associated with Performance on Oral Language Tasks and Oral Ability Components for Japanese University Students," *Language Testing*, vol. 32, issue. 1, pp. 39–62, 2015. 76

[60] A. Sheldon and W. Strange, "The Acquisition of /r/ and /l/ by Japanese Learners of English: Evidence that Speech Production Can Precede Speech Perception," *The Journal of Applied Psycholinguistics*, vol. 3, no.03, pp. 243–261, 1982. 30

[61] D. Kewley-Porr, R. Akahane-Yamada, and K. Aikawa, "Intelligibility and Acoustic Correlates of Japanese Accented English Vowels," *In Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 1996. 29

[62] T. Vance, "An Introduction to Japanese Phonology," *Albany: State University of New York Press*, 1987. 29, 30

[63] S. Sakti, K. Markov, S. Nakamura, and W. Minker, "Incorporating Knowledge Sources into Statistical Speech Recognition," *Spring*, vol. 42, 2009. 16

[64] D.E. Powers, "Assessing English-Language Proficiency in All Four Language Domains: Is It Really Necessary?" *Compendium Study, ETS, TOEIC*, 2013. 4, 75, 76

[65] D. Nunan, "The Impact of English as A Global Language on Educational Policies and Practices in the Asia Pacific Region," *TESOL quarterly*, vol. 37, no. 4, pp. 589–613, 2003. 1

[66] R.M. DeKeyser, "What Makes Learning Second Language Grammar Difficult? A Review of Issues," *The Journal of Language Learning*, vol. 55, S1, pp. 1–25, 2005. 3, 22

[67] X. Wang, T. Kato, and S. Yamamoto, "Phoneme Set Design Considering Integrated Acoustic and Linguistic Features of Second Language Speech," *in Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, San Francisco, USA, 2016. 9

[68] X. Wang, T. Kato, and S. Yamamoto, "Phoneme Set Design Based on Integrated Acoustic and Linguistic Features for Second Language Speech Recognition," *The Journal of IEICE Transactions on Information and Systems*, vol. E98-D, no. 1, pp. 148–156, 2016. 9

[69] J. E. Flege, "The Detection of French Accent by American Listeners," *the Journal of the Acoustical Society of America*, vol. 76, pp. 692–707, 1984. 28

[70] M. Patkowski, "Age and Accent in A Second Language: A Reply to James Emil Flege," *The Journal of Applied Linguistics*, vol. 11, no. 1, pp.73–89, 1990. 23

[71] W. O'Grady and J. Archibald, "Contemporary Linguistic Analysis: An Introduction," *Pearson Canada*, 2015. 23

[72] X. Wang and S. Yamamoto, "Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets," *in Proceedings of Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany, 2015. 10, 11

[73] X. Wang and S. Yamamoto, "Speech Recognition of English by Japanese using Lexicon Represented by Multiple Reduced Phoneme Sets," *The Journal of IEICE Transactions on Information and Systems*, vol. E98-D, no. 12, pp. 2271–2279, 2015. 11

[74] "Pronunciation Changes in Japanese English," http://www.tefl.net/elt/articles/home-abroad/Japanese English-pronunciation-changes/, accessed Oct 8, 2016. 30

[75] C. T. Best, G. W. McRoberts, and E. Goodell, "Discrimination of Non-native Consonant Contrasts Varying in Perceptual Assimilation to the Listener's Native Phonological System," *The Journal of the Acoustical Society of America*, vol. 109, no. 2, pp. 775–794, 2001. 2, 6

[76] S. D. Krashen and P. Pon, "An Error Analysis of An Advanced ESL Learner," *Working Papers in Bilingualism*, vol. 7, pp. 125–129, 1975. 22

[77] V. Brown, "Improving Your Pronunciation," *Tokyo: Meirindo*, 1960. 30

[78] M. D. Tyler, C. T. Best, A. Faber, and A. G. Levitt, "Perceptual Assimilation and Discrimination of Non-native Vowel Contrasts," *The Journal of Phonetica*, vol. 71, no. 1, pp.

4–21, 2014. 2, 6

[79] E. E. Tarone, "Variability in Interlanguage UseA Study of Style-Shifting in Morphology and Syntax," *The Journal of Language Learning*, vol. 35, pp. 373–404, 1985. 22

[80] M. Celce-Murcia and L. McIntosh, "Teaching English as A Second or Foreign Language," *Boston, MA: Heinle & Heinle*, 1991. 3

[81] X. Huang, A. Acero, and H. W. Hon, "Spoken Language ProcessingA Guide to Theory, Algorithm, and System Development," *New Jersey, Prentice Hall PTR*, 2001. 16

[82] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *In Proceedings of IEEE transactions on acoustics, speech, and signal processing (IEEE/ACM)*, vol. 28, no. 4, pp. 357–366, 1980. 15

[83] S. M. Gass, "Second Language Acquisition: An Introductory Course," *Fourth Edition, Routledge*, 2013. 20, 21, 22

[84] J. Kenworthy, "Teaching English Pronunciation," *Longman*, 1992. 29

[85] I. Thompson, "Japanese Speakers," *Learner English: A teacher's guide to interference and other problems*, pp. 296–309, 2001. 29

[86] P. De Lacy, ed, "The Cambridge Handbook of Phonology," *Cambridge University Press*, 2007. 20, 21, 22

[87] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Efficient Phoneme Set Design Using Phonetic Decision Tree in Dialogue-Based English CALL Systems for Japanese Students." *IEICE Technical Report*, Tsukuba, Japan, Vol. 113, No. 366, pp.47-51, Dec 2013. 9

[88] J. Zhang, X. Hu, and S. Nakamura, "Using Mutual Information Criterion to Design An Efficient Phoneme Set for Chinese Speech Recognition." *The Journal of IEICE transactions on information and systems*, vol. 91, no. 3, pp. 508–513, 2008. 42

[89] X. Wang, J. Zhang, and S. Yamamoto, "Multiple Reduced Phoneme Sets for Second Language Speech Recognition." *in Proceedings of Acoustical Society of Japan, Autumn (ASJ)*,

Hokkaido, Japan, 2014. 10

[90] "Patterns of Pronunciation Errors in English by Native Japanese and Hebrew Speakers: Interference and Simplification Processes," *City University of New York*, 1986 30

[91] L. W. Kat, and P. Fung, "Fast Accent Identification and Accented Speech Recognition." *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp.221–224, 1999. 24

[92] L. Rabiner and B. H. Juang, "Fundamentals of Speech Recognition," *Prentice Hall*, First edition, 1993. 14

[93] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–1752, 1990. 15

[94] M. Sakamoto, "Moving Towards Effective English Language Teaching in Japan: Issues and Challenges," *The Journal of Multilingual and Multicultural Development*, vol. 33, no. 4, pp. 409–420, 2012. 27

[95] T.T. P, "Automatic Speech Recognition for Non-Native Speakers," *UNIVERSITÉ JOSEPH FOURIER, Ph.D. Dissertation*, 2008. 17

[96] T. Schultz, A. Waibel, "Polyphone Decision Tree Specialization for Language Adaptation," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000. 25

[97] D. B. Fry, "Duration and Intensity As Physical Correlates of Linguistic Stress," *The Journal of Acoustic Society of America*, vol. 27, pp. 765–768, 1955. 28

[98] T. Gay, "Physiological and Acoustic Correlates of Perceived Stress," *The Journal of Language and Speech*, vol. 21, pp. 347–353, 1978. 28

[99] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness Predicts Prominence: Fundamental Frequency Lends Little," *the Journal of Acoustical Society of America*, vol. 118, no. 2, pp. 1038–1054, 2005. 28

[100] C. Fries, "Teaching and Learning English As A Foreign Language," *Ann Arbor: The University of Michigan Press*, 1945. 28

[101] J. Jenkins, "The Phonology of English As An International Language," *Oxford: Oxford University Press*, 2000. 28

[102] M. J. Munro, T. M. Derwing, and S. L. Morton, "The Mutual Intelligibility of L2 Speech," *The Journal of Studies in Second Language Acquisition (SSLA)*, vol. 28, no. 1, pp. 111–131, 2006. 28

[103] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *The Journal of IEEE Transactions on Communications (TCOM)*, vol. 28, no. 1, pp. 84–95, 1980. 65

[104] H.B. Basiron, "Corrective Feedback in Dialogue-based Computer Assisted Language Learning." *in Proceedings ofNew Zealand Computer Science Research Student Conference (NZCSRSC)*, pp. 192–195, 2008. 70

[105] W.L. Johnson, S. Marsella and H. Vilhjalmsson, "The DARWARS Tactical Language Training System." *in Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, 2004. 70

[106] J.G. Fiscus, "A Post-processing System to Yield Reduced Word Error Rates: Recognizer Output Voting Error Reduction (ROVER)," *in Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 347–354, Dec. 1997. 85

[107] M.A. Zissman, "Language identification using phoneme recognition and phonotactic language modeling," *in Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 5, pp. 3503–3506, May. 1995. 85

[108] T. Isobe, K. Itou, and K. Takeda, "A Likelihood Normalization Method for the Domain Selection in the Multi-Decoder Speech Recognition System," *The Journal of IEICE transactions on information and systems (in Japanese Edition)*, vol. J90-D, no. 7, pp. 1773–1780, 2007. 85

Appendices

Phonemic Symbols of English

Table A.1: List of phonemic symbols of English (41 phonemes) corresponding to IPA notation and word examples [47]

Phone label	IPA	Word example	Phone label	IPA	Word example
AA	[ɑ]	b***o***b	HH	[h]	***h***ay
AX	[ə]	***a***bout	K	[k]	***k***ey
AW	[aʊ]	b***ou***t	L	[l]	***l***ay
AO	[ɔ]	b***ou***ght	M	[m]	***m***o***m***
OW	[o]	b***oa***t	N	[n]	***n***oo***n***
OY	[ɔɪ]	b***oy***	P	[p]	***p***ea
AH	[ʌ]	b***u***t	R	[r]	***r***ay
AXR	[ɚ]	butt***er***	S	[s]	***s***ea
AE	[æ]	b***a***t	TH	[θ]	***th***in
AY	[aɪ]	b***i***te	V	[v]	***v***an
EH	[e]	b***e***t	W	[w]	***w***ay
UH	[ʊ]	b***oo***k	Y	[j]	***y***acht
IH	[ɪ]	b***i***t	Z	[z]	***z***one
ER	[ɝ]	b***ir***d	SH	[ʃ]	***sh***e
UW	[ʊ]	b***oo***t	ZH	[ʒ]	a***z***ure
EY	[ei]	b***ai***t	B	[b]	***b***ee
IY	[i]	b***ee***t	D	[d]	***d***ay
CH	[ʧ]	***ch***oke	F	[f]	***f***in
DH	[ð]	***th***en	G	[g]	***g***ay
NG	[ŋ]	si***ng***	JH	[ʤ]	***j***oke
T	[t]	***t***ea			

B Result of Cluster Splitting

In Figure B.1, B.2 and B.3, "C" refers to terminal nodes that indicate a cluster. Vowels are contained inside the area shown by a dotted curve in Appendix Figure B.1. The two specific cluster splitting processes clearly demonstrate that:

- The PDT-based top-down method provides clusters holding major distinctive features, e.g., vocalic and consonantal, and the top-down splitting method using only phonetic distances could not perform well when it came to distinguishing phonemes

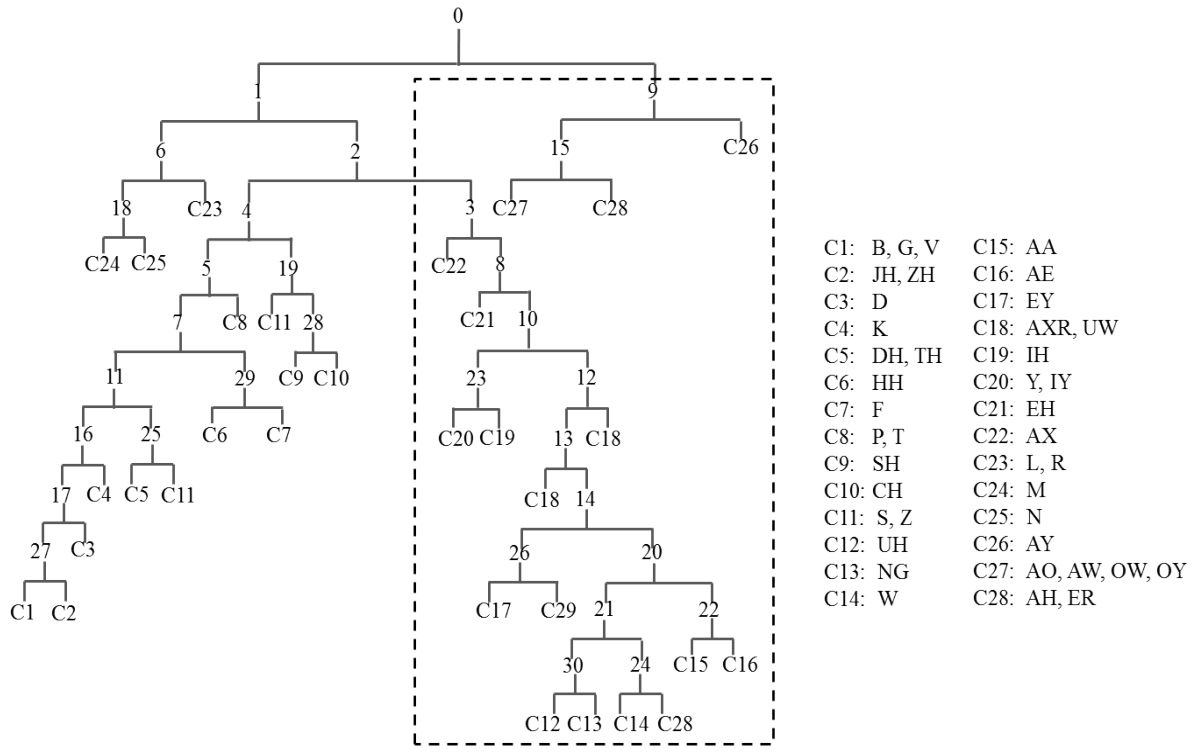


Figure B.1: Result of cluster splitting with PDT-based top-down method in which 28 phonemes were obtained as the final phoneme set. Terminal nodes use "C" to indicate a cluster.

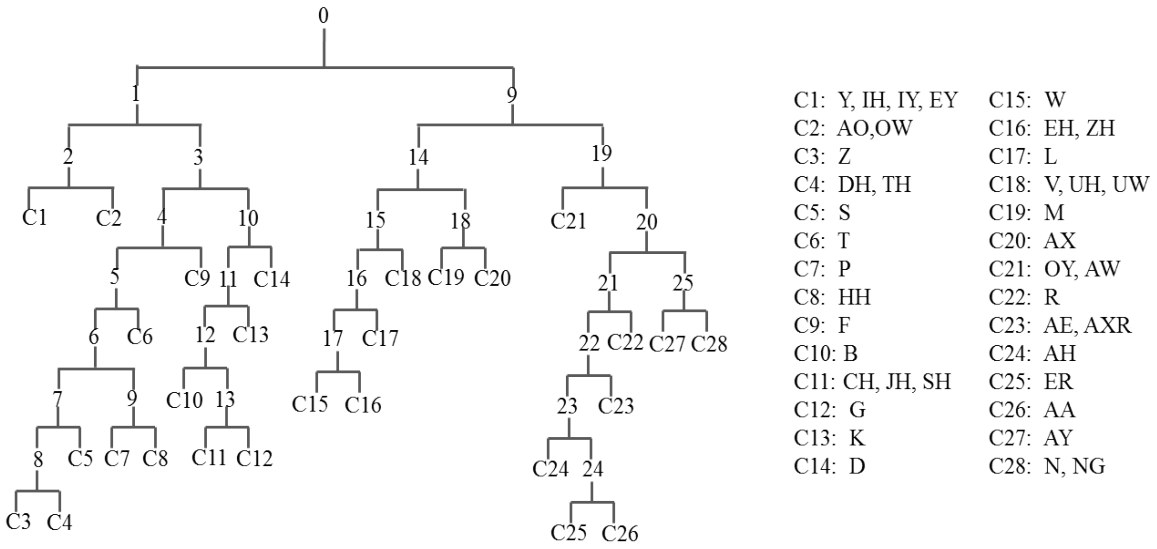
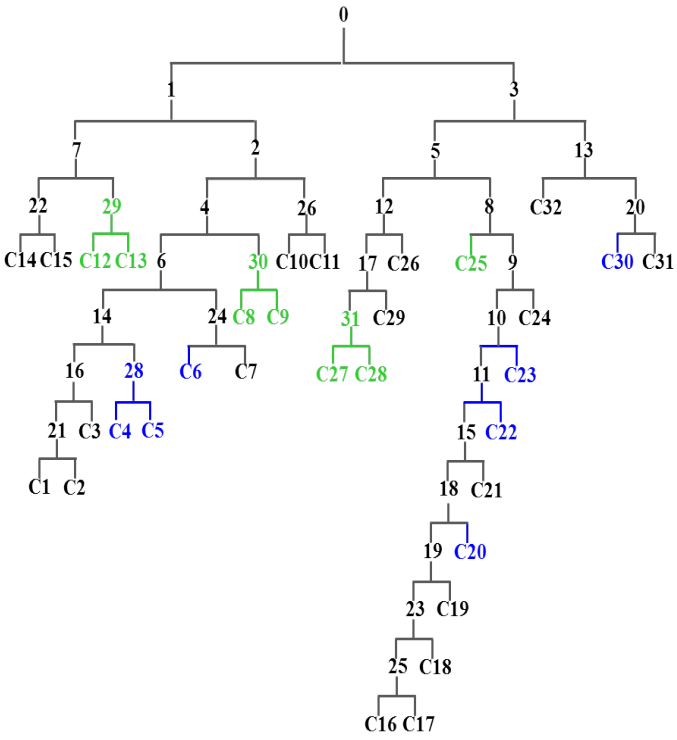


Figure B.2: Result of cluster splitting with the top-down splitting method in which 28 phonemes were obtained as the final phoneme set.

based on major distinctive features.

- Based on the Appendix Figure B.1, cluster 11 (C11) and cluster 28 (C28) appeared twice and were then merged as a terminal cluster, respectively, because combined $\triangle L$s in these clusters were less than the threshold.
- /B/ [b] and /V/ [v] are, according to previous research and basic phonological knowledge, phonemes that are often confused. Phoneme /B/ [b] is also confused with /G/ [g], as with the words "bought" and "got".

Figure B.3 shows an example of the detailed cluster splitting process to obtain a phoneme set with 32 phonemes as the final phoneme set. The coloured fonts show the phonemes of different phoneme set sequences that have been differently merged among 25-, 28-, and 32-phoneme sets, with green fonts depicting the cluster merging for obtaining 28 phonemes based on the cluster splitting step of the 32-phoneme set and blue fonts depicting the cluster merging for obtaining 25 phonemes based on the cluster splitting step of the 28-phoneme set. Black font indicates phonemes of the single phoneme set sequence.



C1: G, V, B, ZH	C9: T	C17: NG	C25: OY
C2: K	C10: S	C18: AE	C26: AY
C3: F, HH	C11: Z	C19: W	C27: AO
C4: D	C12: L	C20: Y	C28: AW, OW
C5: TH, DH	C13: R	C21: EY	C29: AH
C6: CH, JH	C14: M	C22: AXR	C30: IY
C7: SH	C15: N	C23: ER	C31: IH
C8: P	C16: EH, UH	C24: UW	C32: AA, AX

Figure B.3: The result of cluster splitting with PDT in which 25, 28, and 32 phonemes were obtained as the final phoneme set. The phonemes of single and different phoneme set sequences are depicted.)

C Discrimination Rules Design

There are 166 discrimination rules designed based on the knowledge of the phonetic relations between the Japanese and English languages and the actual pronunciation inclination of English utterances by Japanese. The full rules are shown in the following Tables.

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

Contents of Designed Discrimination Rules
Apicals
DH,ZH
S,DH
S,TH
S,Z,DH
S,Z,TH,DH,ZH
S,Z,TH
S,Z,ZH
S,ZH
S,Z
TH,DH,ZH
TH,DH
TH,ZH
Z,DH,ZH
Z,DH
Z,TH,DH
Z,TH,ZH
Z,TH
Z,ZH
Alveolars
D,L
D,N
N,L
T,D,L
T,D,N,L
T,D
T,L
T,N,D
T,N,L
T,N
Velars
G,K,HH
G,K
G,HH
K,HH
Palatals
SH,JH,CH
SH,JH
JH,CH
SH,CH

Contents of Designed Discrimination Rules
Retroflex
R,SH
R,ZH
R,JH
R,DH
R,CH
Nasals
M,N,NG
M,N
M,NG
N,NG
Articulation manners
Stops
B,D,G,sil
B,D,G
B,D,sil
B,D
B,G,sil
B,G
B,P,D,T,G,K,sil
B,P,D,T,G,K
D,G,sil
D,G
P,K,sil
P,K
P,T,K,sil
P,T,K
P,T,sil
P,T
T,K,sil
T,K

Contents of Designed Discrimination Rules
Affricates
DH,ZH,JH,sil
DH,ZH,JH
DH,ZH,TH,Z,JH,V,R,sil
DH,ZH,TH,Z,JH,V,R
DH,ZH,TH,Z,JH,sil
DH,ZH,TH,Z,JH
DH,ZH,TH,Z,sil
DH,ZH,TH,Z
DH,ZH,sil
DH,ZH
TH,Z,JH,sil
TH,Z,JH
TH,Z,sil
TH,Z
ZH,JH,sil
ZH,JH
ZH,TH,Z,JH,sil
ZH,TH,Z,JH
ZH,TH,Z,sil
ZH,TH,Z
ZH,V,sil
ZH,V
Fricatives
F,HH,SH,CH,sil
F,HH,SH,CH
F,HH,SH,sil
F,HH,SH
F,HH,sil
F,HH
S,CH,sil
S,CH
S,F,HH,SH,CH,sil
S,F,HH,SH,CH
S,F,HH,SH,sil
S,F,HH,SH
S,F,SH,sil
S,F,SH
S,F,sil

Contents of Designed Discrimination Rules
Fricatives
S,F
S,SH,CH,sil
S,SH,CH
S,SH,sil
S,SH
SH,CH,sil
SH,CH
Sonorants
L,M,N,R,NG
L,M,N,R
L,M,N
L,N,R
L,N
L,R
M,N,NG
M,N,R,NG
N,NG
Context for Finals
AA,AE,AY,AH,AW,AO
AA,AE,AY,AH,AW
AA,AE,AY,AH
AE,AY,AH,AW,AO
AE,AY,AH,AW
AH,AW,AO
AH,AW
AW,AO
AX,AXR,AA,AE,AY,AH,AW,AO
AX,AXR,AA,AE,AY,AH,AW
AXR,AA,AE,AY,AH,AW,AO
AXR,AA,AE,AY,AH,AW
AY,AH,AW,AO
AY,AH,AW
AH,AW

Contents of Designed Discrimination Rules
Context for Finals
EH,ER,EY EH,ER EH,EY ER,EY
IY,IH,Y IY,IH IH,Y IY,Y
OW,OY,W OW,OY OY,W OW,W
UW,UH
IH,IY,AX D,DH,T