PAPER

Speech Recognition of English by Japanese Using Lexicon Represented by Multiple Reduced Phoneme Sets

Xiaoyun WANG[†a)], ***Student Member and*** **Seiichi YAMAMOTO**[†], ***Fellow***

SUMMARY Recognition of second language (L2) speech is still a challenging task even for state-of-the-art automatic speech recognition (ASR) systems, partly because pronunciation by L2 speakers is usually significantly influenced by the mother tongue of the speakers. The authors previously proposed using a reduced phoneme set (RPS) instead of the canonical one of L2 when the mother tongue of speakers is known, and demonstrated that this reduced phoneme set improved the recognition performance through experiments using English utterances spoken by Japanese. However, the proficiency of L2 speakers varies widely, as does the influence of the mother tongue on their pronunciation. As a result, the effect of the reduced phoneme set is different depending on the speakers' proficiency in L2. In this paper, the authors examine the relation between proficiency of speakers and a reduced phoneme set customized for them. The experimental 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 reduced phoneme sets, and the implementation of a language model most suitable for that lexicon is described. Experimental results demonstrate the high validity of the proposed method.

key words: *second language (L2) speech recognition, proficiency of L2 speakers, reduced phoneme set (RPS), multiple reduced phoneme sets, proficiency-dependent reduced phoneme set*

1. Introduction

In today's environment of rapid globalization, people have increasing opportunities for speaking in foreign languages, and the ability to communicate in foreign languages is now more important than ever. Non-native speakers have a limited vocabulary and a less than complete knowledge of the grammatical structures of a foreign 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 usually includes less fluent pronunciation and mispronunciation even in cases in which it is otherwise delivered well. Human beings can eventually understand non-native speech quite easily because after a while the listeners get used to the style of the talker, i.e., the various insertions, deletions, and substitutions of phonemes or the wrong grammar. More problematic is when non-native pronunciations become an issue for speech dialogue systems that target tourists, such as travel assistance systems, hotel reservation systems, and systems in which consumers purchase goods through a network. The vocabulary and grammar of non-native speakers is often limited and therefore basic, but speech recognizers take no or only a little advantage of this and are confused by the different phonetics [1]. Non-native speech poses several challenges for automatic speech recognition.

In order to improve the speech recognition accuracy for non-native speech, various methodologies have been proposed, including acoustic model adaptation for second language speech with a variant phonetic unit obtained by analyzing the variability of second language speech pronunciation [2], an acoustic model interpolating from both native and non-native acoustic models [3], data driven generation of pronunciation variants for lexical modeling [4], and others. These automatic speech recognition technologies for non-native speech have been developed assuming the mother tongues of users to be unknown and various. However, recent studies have shown that the mother tongue of speakers can be predetermined for ASR to improve recognition results in certain applications, such as dialogue-based computer assisted language learning (CALL) systems or mobile platforms [5]–[7].

To improve the recognition accuracy for non-native speech in cases where the mother tongue of the speaker is known, we previously proposed using a reduced phoneme set (RPS) created with a phonetic decision tree based top-down sequential splitting method [8]. We applied this method to the recognition of English utterances spoken by Japanese speakers of various English proficiencies and demonstrated that the reduced phoneme set was more effective than the canonical phoneme set conventionally used in English ASR (refer to Sect. 3.1 for details).

As revealed by many second language acquisition studies, the pronunciation of target language by second language speakers generally deviates from native speech and is significantly influenced by the mother tongue of the speakers [9]. At the same time, the speech quality of second language speakers overall depends on their proficiency level in the second language [10], [11], and there are different patterns in accent among inexperienced, moderately experienced, and experienced speakers [12], [13]. As a result, influence of the mother tongue on pronunciation varies widely.

In the current work, 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

Manuscript received February 16, 2015.
Manuscript revised June 25, 2015.
Manuscript publicized September 10, 2015.

[†]The authors are with Graduate School of Science and Engineering, Doshisha University, Kyotanabe-shi, 610–0321 Japan.
a) E-mail: ougyouun@gmail.com

DOI: 10.1587/transinf.2015EDP7061

dialogue-based English CALL system [14] in the form of a translation exercise for Japanese students.

The rest of this paper is organized as follows. The reduced phoneme set for Japanese accented English is briefly introduced in Sect. 2. In Sect. 3, the relation between the reduced phoneme set and the English proficiency level of L2 speakers is discussed. In Sect. 4, we propose using multiple reduced phoneme sets for second language speech recognition and describe the implementation of a language model suitable for it. Section 5 reports the experimental results and we discuss these results in Sect. 6. We close with a conclusion and a brief mention of future work in Sect. 7.

2. Reduced Phoneme Set

The reduced phoneme set proposed in our previous work was created with a phonetic decision tree based top-down sequential splitting method. The phonetic decision tree is a binary tree in which discrimination rules are attached to each splitting step. As discrimination rules, we use the knowledge of phonetic relations between the Japanese and English languages and the actual pronunciation inclination of English utterances by Japanese. The splitting process uses the phonetic acoustic features of speech by L2 speakers and the occurrence distributions of each phoneme as the splitting criterion.

2.1 Criterion of Determining Phoneme Set

We used as the splitting criterion the log likelihood (1) defined by the logarithm of the probability density function of an acoustic model generating the L2 speech observation data.

$$L(P_m) \approx \sum_{t=1}^{T} \log P(\boldsymbol{O}_t) \cdot \gamma_m \tag{1}$$

where P_m represents the m^{th} phoneme or phoneme cluster. P is the joint node pdf of the phoneme cluster. $\boldsymbol{O}_t$ means the observation data of $[O_1, O_2, \ldots, O_T]$. γ_m is the phonetic occupation counts of the model generating $\boldsymbol{O}_t$, which is a good prediction of the occupancy frequency of the canonical phonemes that are used in typical Japanese-English speech utterances.

To carry out the preliminary splitting, we utilized 166 discrimination rules [8] that were designed to categorize each phoneme on the basis of phonetic features such as the manner and position of articulation. A selection of these discrimination rules is shown in Table A· 1 (appendix). For example, one of the discrimination rules of "Apicals" denotes that phonemes DH /ð/, TH /θ/, ZH /ʒ/, Z /z/, and S /s/ have an apical feature, making them suitable to discriminate non-native speech. All rules are based on knowledge of the phonetic relation between the Japanese and English languages and the actual pronunciation inclination of English utterances by Japanese. In this splitting method, all phonemes listed in each discrimination rule based on other phonetic features depict similar phonological characteristics and have the possibility to be merged into a cluster.

2.2 Procedure of Determining Phoneme Set

We primarily used a 4-step procedure to design a reduced phoneme set:

1. Set all merging phonemes as a root cluster at the initial state. Calculate increased log likelihood $\triangle L_R$ according to equations (1) and (2), assuming that cluster S is partitioned into $S_y(R)$ and $S_n(R)$ by discrimination rule R.

$$\triangle L_R = L(S_y(R)) + L(S_n(R)) - L(S) \tag{2}$$

   where $\triangle L_R$ is the increased log likelihood of the phoneme cluster, which is calculated for all discrimination rules applicable to every cluster.
2. Select the rule with maximum log likelihood increase compared with before splitting.

$$L_{R^*} = \arg\max_{all\ \ R} \triangle L_R \tag{3}$$

   The rule R^* causing the maximum increase is chosen as the splitting rule.
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.

There are two reasons the reduced phoneme set can create suitable phonological decoding when the mother tongue of speakers is known. The first is that it can be designed to characterize the acoustic features of Japanese accented English more correctly. The second is that there is more speech data for training the acoustic model of each phoneme in the reduced phoneme set than in the canonical one, so we can obtain more reliable estimate values as parameters of acoustic models.

3. Relation between Optimal Phoneme Set and L2 Speakers with Different Proficiencies

Our previous research demonstrated that phoneme mismatches resulting in mis-recognition of second language speech can be improved by reducing the number of phonemes [8]. 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 are presented in the next subsections.

3.1 Phoneme Set

The phonemic symbols of the TIMIT database were used as the canonical phoneme set [15]. Table 1 lists the phonemes of English in Arpabet notation and IPA notation as the canonical phoneme set. The baseline is ASR using the canonical phoneme set consisting of 41 phonemes. In order to evaluate the relation between the optimal phoneme set number and second language speakers, we also used various reduced phoneme sets with numbers ranging from 21 to 38 obtained from our previous study [8].

3.2 Acoustic and Language Models

An English as a Second Language (ESL) speech database [16] was used to train the acoustic models. The database contained utterances read by 200 Japanese students (100 male and 100 female) covering different degrees of English proficiency and included phonemic symbols as well as prosodic ones assigned to various words and sentences. The total number of sentence utterances was approximately 12,000 and the total number of word utterances was approximately 22,000 for each gender. Table 2 lists the specific features of the ESL speech database. In this study, all sentences and word utterances were used to train context-dependent state-tying triphone HMM acoustic models of various numbers of phoneme sets. Table 3 shows the experimental conditions for acoustic analysis (AA) and the HMM specifications.

We developed a 2-gram language model from 5,350 transcribed utterances by 62 Japanese university students. The pronunciation lexicon included about 35,000 vocabulary words related to conversation about travel abroad.

Table 1 Canonical phoneme set of English in Arpabet notation and IPA notation.

Vowels	Consonants
AE /æ/, AH /ʌ/, EH /e/, IH /ɪ/, OY /ɔɪ/, ER /ɝ/, UH /ʊ/, AW /aʊ/, AY /aɪ/, AA /ɑ/, AO /ɔ/, EY /ei/, IY /i/, OW /o/, UW /u/, AX /ə/, AXR /ɚ/	CH /ʧ/, DH /ð/, NG /ŋ/, JH /ʤ/, SH /ʃ/, TH /θ/, ZH /ʒ/, B /b/, D /d/, F /f/, G /g/, HH /h/, K /k/, L /l/, M /m/, N /n/, P /p/, R /r/, S /s/, T /t/, V /v/, W /w/, Y /j/, Z /z/

Table 2 English word and sentence sets spoken by 200 Japanese students [16].

Set	Size
Phonetically balanced words	300
Minimal pair words	600
TIMIT-based phonetically balanced sentences	460
Sentences including phoneme sequence difficult for Japanese to pronounce correctly	32
Sentences designed for test set	100
Words with various accent patterns	109
Sentences with various intonation patterns	94
Sentences with various rhythm patterns	121

3.3 Evaluation Data

We collected orally translated speech using a dialogue-based CALL system [14] as evaluation data. There were a total of 45 participants between the ages of 18 and 24 who had acquired Japanese as their mother tongue and learned English as their second language. In this study, the Test of English for International Communication (TOEIC) score was used for measuring the overall English proficiency of the speakers. Specifically, TOEIC was used as a measure to verify the variety of L2 speech of overall language proficiency and to classify participants in accordance with their level of pronunciation. According to the standard of the Educational Testing Service (ETS) [17], a TOEIC score of 500 is the minimum score for new university graduates recruited by general Japanese companies, while 700 is the minimum preferred score for studying and working abroad. A score of more than 900 is expected for skillful business and daily conversation in English. The participants' English proficiencies ranged from 300 to 910 (990 being the highest score that can be attained) in our experiments.

3.4 Recognition Performance with Proficiency-Based Clustering

We divided participants into 5 groups based on their TOEIC score range: lower than 500, 500−600, 600−700, 700−800, and higher than 800, with 10, 10, 10, 8, and 7 participants in each score range. We call this method the *"top-down method"* in the following.

We used the HTK toolkit [18] to compare the ASR performance using a single canonical phoneme set and reduced phoneme sets for speech by a group of participants classified by the top-down method at each level of the five TOEIC score ranges. Figure 1 shows the relative error reduction of various reduced phoneme sets compared with the canonical one for speech by participants of each of the five TOEIC score ranges. We found 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 is different depending on the English proficiency level of the speakers.
- The recognition performance of the reduced phoneme

Table 3 Condition of acoustic analysis (AA) and HMM specifications.

AA	Sampling rate	16kHz
	Feature vector	0-12 mel-cepstral energy+Δ + ΔΔ (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

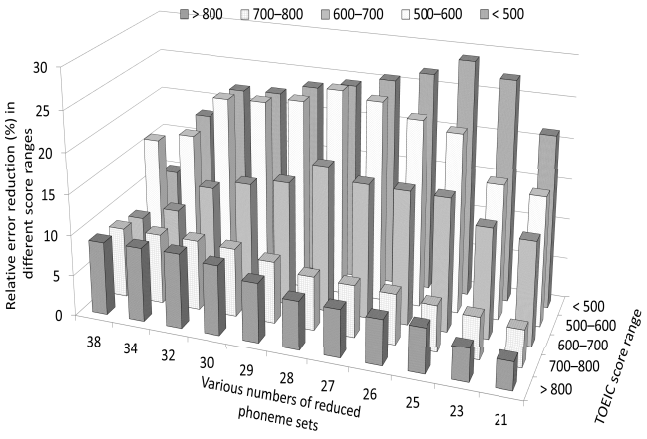


Fig. 1 Relative error reduction for speech by participants in different TOEIC score ranges.

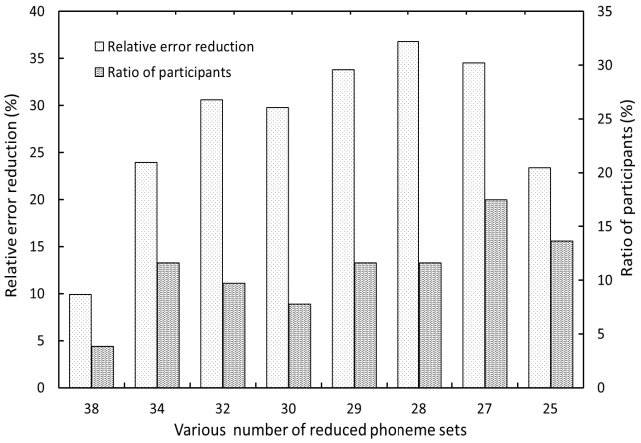


Fig. 2 Ratio of participants in their optimal reduced phoneme set and relative error reduction for speech by speakers achieved the best recognition accuracy.

set was improved more for the 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.

According to the results of Fig. 1, 25-RPS is the optimal reduced phoneme set for the English proficiency of speakers who have a TOEIC score lower than 500, a 28-RPS is the optimal reduced phoneme set for those with a score of 500–700, and a 32-RPS is the optimal one for those with a score of higher than 700.

The optimal reduced phoneme set (25-, 28-, or 32-phoneme) corresponding to English proficiency is selected using the top-down method. We refer to this optimal phoneme set as the *proficiency-dependent reduced phoneme set* henceforth.

3.5 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 Sect. 3.4. 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 [13], [19], [20]. 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 [21].

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

4. ASR System with Multiple Reduced Phoneme Sets

Both of the results discussed in Sect. 3 showed that the optimal reduced phoneme set is different depending on the second language speakers. It seems to be inadequate to use a single phoneme set to recognize input speech by all second language speakers. In this section, we propose an ASR system using multiple reduced phoneme sets to further improve the recognition performance of second language speech considering the various proficiency levels of second language speakers.

On the basis of the experimental results in Sects. 3.4 and 3.5, 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 set the constrain that one of three multiple phoneme sets, not a mixture, be used to recognize input speech by a single second language speaker. 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 by a language model implementing the constraint, as described in the following.

4.1 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. Fig. A· 2 (appendix) 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.

4.2 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 simply 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 favor 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 pronunciations, 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 Fig. 3.

[†]The experimental results of the bottom-up method showed that 28- and 32-phoneme sets can also be considered as the components of multiple reduced phoneme sets. An additional experimental result is shown in Fig. A· 1 (appendix) to compare the performance of multiple reduced phoneme sets constructed by 28- and 32-phoneme sets with multiple reduced phoneme sets constructed by 25-, 28-, and 32-reduced phoneme sets

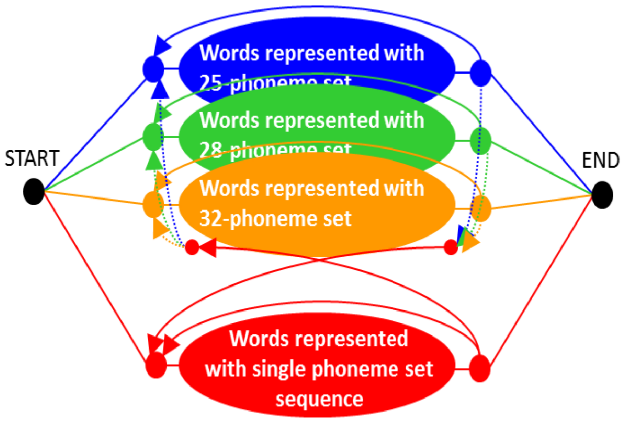


Fig. 3 Schematic diagram of language model for words represented by 25-, 28-, and 32-phoneme sets and for words with single phoneme set sequence. Arcs depict transition between words represented by the same reduced phoneme set and words of single pronunciation. The transition among words represented by different reduced phoneme sets is inhibited.

5. Experimental Results

To evaluate the proposed method, we investigated the efficacy of the derived lexicon represented by multiple reduced phoneme sets and the language model designed for the represented lexicon. We compared the performance on ASR implementing the proposed method with that of the canonical phoneme set and three single reduced phoneme sets considering the real time factor (RTF). We set the RTF to less than 1 for each recognition result as the experimental condition.

Figure 4 shows the word error rates of various numbers of phoneme sets including the canonical, the single, and the multiple phoneme sets. Multiple sets in Fig. 4 are constructed by 25-, 28-, and 32-phoneme sets. We observed the following:

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

6. Discussion

6.1 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

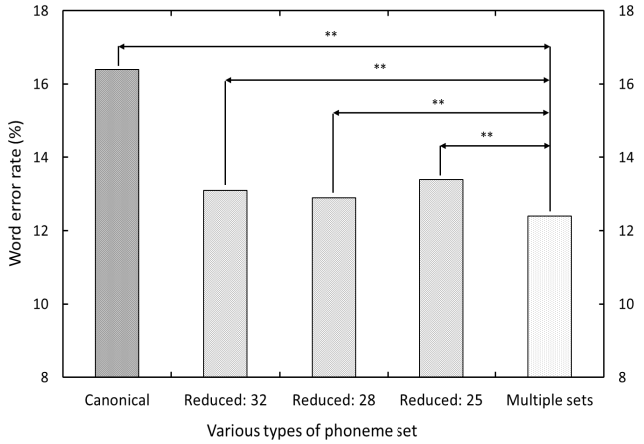


Fig. 4 Word error rates of canonical phoneme set, three single reduced phoneme sets, and the multiple reduced phoneme sets. ** indicates a significant difference between the word accuracy of multiple reduced phoneme sets and the other ones ($p < 0.01$).

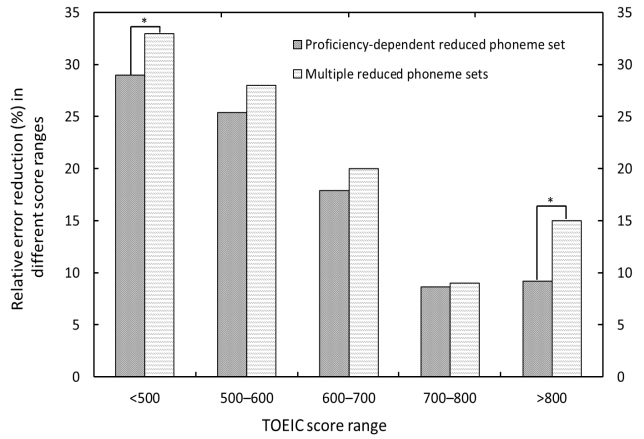


Fig. 5 Relative error reduction of proficiency-dependent phoneme set and multiple reduced phoneme sets in different TOEIC score ranges. * indicates a significant difference between the relative error reduction of multiple reduced phoneme sets and the proficiency-dependent one ($p < 0.05$).

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 Fig. 5. 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 Fig. 5 showed that the most highly improved performances appeared at the lowest and highest proficiency levels. The experimental results of our previous studies showed that a single 28-reduced phoneme set provided better performance in comparison with other reduced ones for speech by L2 speakers on average [8]. 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 4 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%

6.2 Efficacy of Language Modeling

As discussed in Sect. 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 [23], 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 [24]–[26].

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 4) 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 recognition 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 4 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.

7. Conclusion and Future Work

We clarified the relation between the second language speakers and an optimal reduced phoneme set. On the basis of this analysis, we then proposed a novel speech recognition technique using 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 collected with a translation game type dialogue-based CALL system for each proficiency level compared with the canonical phoneme set and single reduced phoneme sets.

The proficiency-dependent phoneme set was designed to be optimal for recognizing speech by speakers in the corresponding score ranges. The experimental results demonstrated that the multiple reduced phoneme sets achieved better recognition performance than the proficiency-dependent phoneme set for speech by speakers at all language proficiency levels. We plan to investigate the relation between the features of speech by each speaker and the optimal reduced phoneme set in future.

Acknowledgments

We express our deepest gratitude to Dr. Ichiro Umata of NICT, Professor Tsuneo Kato of Doshisha University and Professor Jinsong Zhang of the Beijing Language and Culture University for providing us with invaluable comments and helpful discussions.

References

[1] R.E. Gruhn, W. Minker, and S. Nakamura, Statistical pronunciation modeling for non-native speech processing, Springer Berlin Heidelberg, 2011.
[2] Y.R. Oh, J.S. Yoon, and H.K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," Speech Commun., vol.49, no.1, pp.59–70, Jan. 2007.
[3] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," Diss. Massachusetts Institute of Technology, 1999.
[4] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," Proc. ICASSP, vol.3, pp.1683–1686, 2000.
[5] T. Kawahara and N. Minematsu, "Computer-assisted language learning (CALL) based on speech technologies," IEICE Tans. Inf. & Syst. (Japanese Edition), vol.J96-D, no.7, pp.1549–1565, July 2013.
[6] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," Proc. HLT-AACL, pp.468–475, Rochester, NY, April 2007.
[7] F. Xu, S. Schmeier, R. Ai, and H. Uszkoreit, "Yochina: Mobile multimedia and multimodal crosslingual dialogue system," Natural Interaction with Robots, Knowbots and Smartphones, Springer, New York, pp.51–57, 2014.
[8] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme set design for speech recognition of English by Japanese," IEICE Tans. Inf. & Syst., vol.E98-D, no.1, pp.148–156, Jan. 2015.
[9] N. Poulisse and T. Bongaerts, "First language use in second language production," Handbook of Applied Linguistics, vol.15, no.1, pp.36–57, 1994.
[10] K. Osaki, N. Minematsu, and K. Hirose, "Speech recognition of Japanese English using Japanese specific pronunciation habits," IEICE Technical Report, SP2002-180, 2003 (in Japanese).
[11] 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, vol.2010, pp.1–13, 2010.
[12] P. Trofimovich and W. Baker, "Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech," Studies in Second Language Acquisition, vol.28, no.1, pp.1–30, 2006.
[13] J.E. Flege, "Factors affecting degree of perceived foreign accent in English sentences," J. Acoust. Soc. Am., vol.84, no.1, pp.70–79, 1988.
[14] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme set design using English speech database by Japanese for dialogue-based English CALL Systems," Proc. LREC, pp.3948–3951, Reykjavik, Iceland, 2014.
[15] Copyright 1993 Trustees of the University of Pennsylvania, "TIMIT acoustic-phonetic continuous speech corpus," https://catalog.ldc.upenn.edu/LDC93S1, accessed Jan. 27, 2015.
[16] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," Proc. ICA, vol.1, pp.557–560, 2004.
[17] 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 Jan. 27, 2015
[18] 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.
[19] J.E. Flege, M.J. Munro, and I.R.A. MacKay, "Factors affecting strength of perceived foreign accent in a second language," J. Acoust. Soc. Am., vol.97. no.5, pp.3125–3134, 1995.
[20] 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, no.1, pp.39–62, 2015.
[21] D.E. Powers, "Assessing English-language proficiency in all four language domains: Is it really necessary?," Compendium Study, ETS, TOEIC, 2013.
[22] D. Jurafsky and J.H. Martin, Speech & language processing, Pearson Education India, 2000
[23] J.G. Fiscus, "A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER)," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding, pp.347–354, Dec. 1997.
[24] M.A. Zissman, "Language identification using phoneme recogni-

tion and phonotactic language modeling," Proc. ICASSP, vol.5, pp.3503–3506, May 1995.

[25] D.P.P. López and C. García-Mateo, "Application of confidence measures for dialogue systems through the use of parallel speech recognizers," Proc. INTERSPEECH, pp.2785–2788, Sept. 2005.

[26] T. Isobe, K. Itou, and K. Takeda, "A likelihood normalization method for the domain selection in the multi-decoder speech recognition system," IEICE Tans. Inf. & Syst. (Japanese Edition), vol.J90-D, no.7, pp.1773–1780, July 2007.

Appendix

In Fig. A· 1, Multiple reduced sets (A) are constructed by 28- and 32-phoneme sets and Multiple reduced sets (B) are constructed by 25, 28- and 32-phoneme. Figure A· 1 shows that the Multiple reduced sets (B) achieved better performance than the Multiple reduced sets (A).

Figure A· 2 shows an example of the detailed cluster splitting process to obtain a phoneme set with 32 phonemes as the final phoneme set. "C" refers to terminal nodes that indicate a cluster. The colored 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.

Table A· 1 A selection of discrimination rules used to design the reduced phoneme set for Japanese-English.

Discrimination Rules	Contents
Affricates	R, V, JH, Z, TH, ZH, DH
Affricate1	DH, ZH
Affricate2	TH, Z
...	...
Alveolar	L, N, D, T
Alveolar1	L, N
Alveolar2	N, D, T
...	...
Apicals	DH, TH, ZH, Z, S
Apical1	TH, Z, S
Apical2	TH, ZH, Z
Confusing-Consonant	T, DH ,D
Confusing-Vowel	AX, IY, IH
Fricatives	S, F, HH, SH, CH
...	...
Labials	M, P, B, V, F
...	...
Retroflex1	R, SH
Retroflex2	R, ZH
Retroflex3	R, DH
...	...
Stops	B, P, D, T, G, K

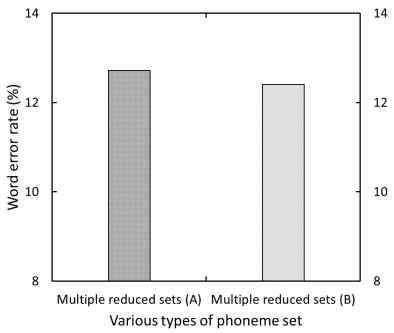
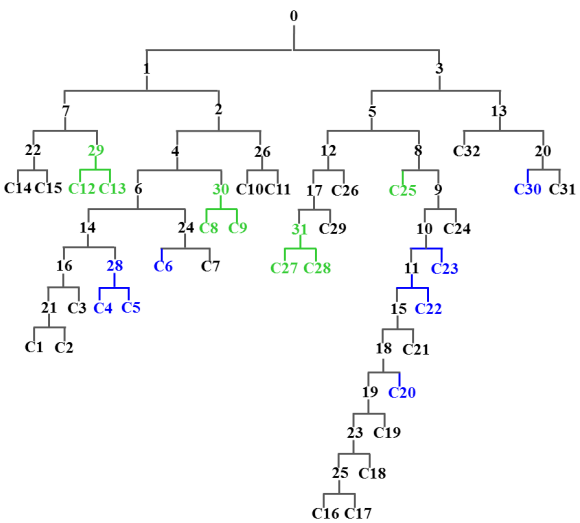


Fig. A· 1 Word error rates of the multiple reduced sets of 28- and 32-phoneme sets and the multiple sets of 25-, 28- and 32-phoneme sets.



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

Fig. A· 2 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.

Xiaoyun Wang was born in China in 1989. She received a B.S. in Information Science from Yamanashieiwa University, Yamanashi, Japan in 2012, and an M.S. from the graduate school of Science and Engineering, Doshisha University, Kyoto, Japan. She is now a Ph.D. student at the graduate school of Science and Engineering, Doshisha University, Kyoto, Japan.

Seiichi Yamamoto received B.S., M.S., and Ph.D. degrees from Osaka University in 1972, 1974, and 1983. He joined Kokusai Denshin Denwa Co. Ltd. in April 1974 and ATR Interpreting Telecommunications Research Laboratories in May 1997. He was appointed president of ATR-ITL in 1997. He is currently a Professor in the Department of Information Systems Design, Faculty of Science and Engineering, Doshisha University, Kyoto, Japan. His research interests include digital signal processing, speech recognition, speech synthesis, natural language processing, spoken language processing, spoken language translation, and multi-modal dialogue processing. He received Technology Development Awards from the Acoustical Society of Japan in 1995 and 1997, a best paper award from the Information and Systems Society of IEICE in 2006, and a telecom-system technology award from the Telecommunications Advancement Foundation in 2007. Dr. Yamamoto is a member of the ASJ, the IPSJ, the IEEE (Fellow), and IEICE Japan (Fellow).