IEICE TRANS. INF. & SYST., VOL.E98–D, NO.1 JANUARY 2015

PAPER

Phoneme Set Design for Speech Recognition of English by Japanese

Xiaoyun WANG[†a)], ***Member***, **Jinsong ZHANG**[††], ***Nonmember***, **Masafumi NISHIDA**[†], ***Member***, ***and*** **Seiichi YAMAMOTO**[†], ***Fellow***

SUMMARY This paper describes a novel method to improve the performance of second language speech recognition when the mother tongue of users is known. Considering that second language speech usually includes less fluent pronunciation and more frequent pronunciation mistakes, the authors propose using a reduced phoneme set generated by a phonetic decision tree (PDT)-based top-down sequential splitting method instead of the canonical one of the second language. The authors verify the efficacy of the proposed method using second language speech collected with a translation game type dialogue-based English CALL system. Experiments show that a speech recognizer achieved higher recognition accuracy with the reduced phoneme set than with the canonical phoneme set.

key words: *phonetic decision tree (PDT), Phoneme set, Second language speech recognition*

1. Introduction

The rapid progress in transportation systems and information technologies has increased the opportunities for worldwide communication. Many people have more opportunities for speaking in a foreign language, 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 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 usually includes less fluent pronunciation and mispronunciation even in cases in which it is well composed. Actual human beings can eventually understand non-native speech quite easily because after a while the listener gets 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 a speech recognizer takes no or only a little advantage of this and is confused by the different phonetics [1].

In order to improve the speech recognition accuracy for non-native speech, various methodologies have been proposed for adapting to the acoustic features of non-native speech, including a speaker adaptation method for second language speech recognition [2], a method using the state-tying of acoustic modeling (AM) for second language speech with a variant phonetic unit obtained by analyzing the variability of second language speech pronunciation [3], AM interpolating with both native and non-native acoustic models [4], and others. Automatic speech recognition (ASR) technology for non-native speech adopted in various speech dialogue systems was developed assuming that the mother tongues of users were both unknown and various. However, recent studies have shown that the mother tongue of users can be known for ASR adopted for certain applications, such as dialogue-based computer assisted language learning (CALL) systems or mobile terminals for personal users, and that the phonetic relation between the target language and the mother tongue of users can be used to improve the recognition accuracy of ASR adopted for such applications.

In this paper, we propose a novel speech recognition method that uses a reduced phoneme set to improve the recognition accuracy for English utterances by Japanese. There have been several previous studies on using a reduced phoneme set for speech recognition. For example, Vazhenina et al. proposed a method that generates an initially confusing table of phonemes based on logical and statistical information and then manually merges some easily confused phones by referencing phonological knowledge [5]. Although this approach has a good performance, it did not consider the spectral properties of the phone models. There was also a study on measuring the distance between acoustic models to merge language-dependent phones using a hierarchical phone clustering algorithm [6]. However, this approach does not consider the acoustic characteristics of the phonemes in real utterances. Although both these methods 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.

Our recognition method using a reduced phoneme set

Manuscript received May 20, 2014.
Manuscript revised August 22, 2014.
Manuscript publicized October 1, 2014.

[†]The authors are with the Graduate School of Science and Engineering, Doshisha University, Kyotanabe-shi, 610–0321 Japan.
[††]The author is with Beijing Language and Culture University, No.15 Xueyuanlu, Haidian, Beijing, 100083, China.
a) E-mail: ougyouun@gmail.com

DOI: 10.1587/transinf.2014EDP7168

created with a phonetic decision tree (PDT)-based top-down sequential splitting utilizes not only the phonological knowledge between mother and target languages and their phonetic features but also the occurrence distribution of the phonemes of the target language produced by second language speakers. We evaluated the recognition performance of the proposed method in the second language speech corpus collected by a previously developed dialogue-based English CALL system in the form of a translation exercise for Japanese students [7].

The rest of this paper is structured as follows. Section 2 describes the reduced phoneme set for recognizing second language speech. In Sect. 3, we illustrate our proposed phoneme set construction in detail. Section 4 presents the speech recognition experiments conducted on the proposed method. Section 5 is a discussion of the experimental results, and in Sect. 6 we close with a conclusion and brief mention of our future work.

2. Reduced Phoneme Set for Second Language Speech

There are two reasons the reduced phoneme set is effective for ASR for second language speech when the mother tongue 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, because there is more speech data for training the acoustic model of each phoneme in the reduced set than in the canonical one, we can obtain more reliable estimate values as parameters of acoustic models.

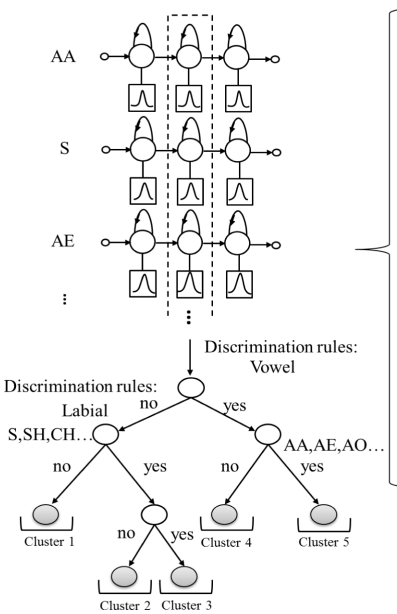
The reduced phoneme set is expected to be even more effective in cases where speakers' utterances are fairly restricted and predictable, such as dialogue-based CALL systems. A human being can understand phonologically confused non-native speech by guessing at the intended spoken word and sometimes correcting it from the context after the listener gets used to the style of the talker. 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 non-native speech. However, if a user's utterances can be designed to be highly constrained, the confused words are limited to a small number.

Various methodologies for constraining spoken responses by students have been proposed for dialogue-based CALL systems, 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 [8]–[12]. A CALL application called a "translation game" [13] 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 improve the accuracy of the ASR by reducing the variety of spoken responses compared with conventional spoken dialogue systems.

3. Effective Phoneme Set Construction

3.1 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 second language speakers and the occurrence distributions of each phoneme as the splitting criterion and uses the relation between the phonological structure of the mother and target languages of the second language speakers as a set of discrimination rules. Figure 1 shows an example of a PDT that partitions the initial phoneme cluster into five terminal



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"

Fig. 1 PDT-based top-down cluster splitting and a part of the discrimination rules.

clusters with our designed discrimination rules.

3.1.1 Splitting Criterion

We used as the splitting criterion the log likelihood (LL) defined by the logarithm of the probability distribution function (pdf) of an acoustic model generating the speech observation data. It is defined by

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

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 Eqs. (2) and (3), respectively:

$$\hat{\boldsymbol{\mu}}_m = \sum_{i \in P_m} \frac{\gamma_i \boldsymbol{\mu}_i}{\gamma_i} \tag{2}$$

$$\hat{\sigma}_m = \sum_{i \in P_m} \frac{\gamma_i (\boldsymbol{\mu}_i - \hat{\boldsymbol{\mu}}_m)^2 + \gamma_i \sigma_i}{\gamma_i} \tag{3}$$

$$\gamma_m = \sum_{i \in P_m} P(\gamma_i) \tag{4}$$

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. γ_m is defined with Eq. (4), which means a posteriori probability of the model generating the observation data $\boldsymbol{O}_t = [O_1, O_2, \ldots, O_T]$, which is a good prediction of the occupancy frequency of the canonical phonemes that are used in typical Japanese-English speech utterances.

We can compute the log likelihood of each phoneme cluster by substituting Eqs. (2) and (3) into (1).

3.1.2 Discrimination Rules Design

As revealed by many second language acquisition studies, pronunciation by second language 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 [14]. 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) [15]. 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. More specifically, we referred to the literature of linguistic knowledge [14], [15] and phoneme confusion matrix for Japanese speakers of English [1] 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 effectively. A part of the discrimination rules is shown in Fig. 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 1 Canonical phoneme set of English in Alphabet notation.

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

Table 1 shows a canonical phoneme set of English in the Alphabet notation. 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 of English are adopted in our experiment as our initial phoneme set.

3.2 Procedure of Designing Reduced Phoneme Set

We used a 4-step procedure to design the reduced phoneme set using a phonetic decision tree-based top-down method. Figure 2 shows the overall procedural diagram of the phoneme cluster splitting with the PDT-based top-down method using a maximum log likelihood criterion.

■ Initialization condition

1. *Initial phoneme cluster*

To set a cluster of the 41 merging phonemes listed in Table 1 as a root cluster and select the mid-state of the context-independent English HMMs of the 41 phonemes as the acoustic model of each phoneme.

2. *Phonetic occupation counts*

To select the counts of each phoneme that appeared in the training data as the phoneme occupation probabilities.

3. *Discrimination rules*

To use the discrimination rules described previously (Sect. 3.1.2).

■ Phoneme cluster splitting procedure

Step 1 *Calculate LL*

Assuming that the cluster S is partitioned into $S_y(R)$ and $S_n(R)$ by one of the discrimination rules R, the increase of

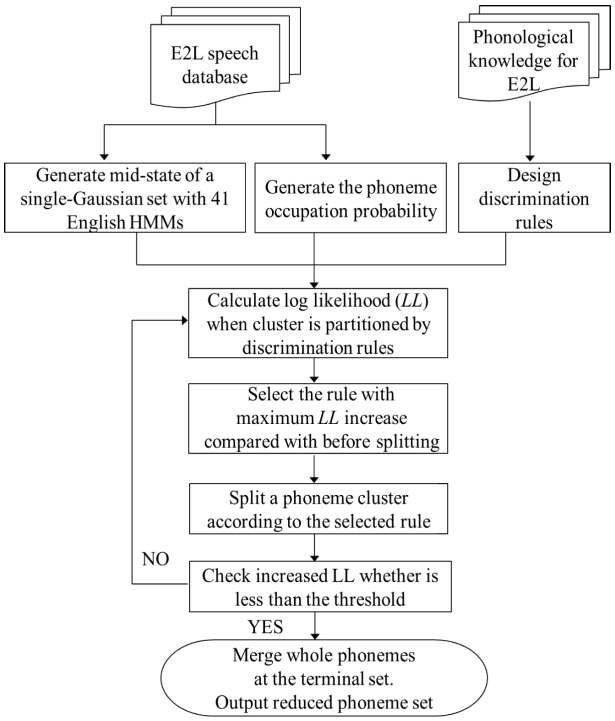


Fig. 2 Overall procedural diagram of the phoneme cluster splitting with a phonetic decision tree (PDT)-based top-down method using a maximum log likelihood criterion.

log likelihood $\triangle L_R$ is calculated as

$$\triangle L_R = L(S_y(R)) + L(S_n(R)) - L(S) \quad (5)$$

$\triangle L_R$ is calculated for all discrimination rules applicable to every cluster.

Step 2 *Select optimization discrimination rule*
The rule R^* is chosen as the splitting rule when it brings about the maximum increase:

$$L_{R^*} = \arg\max_{all\ R} \triangle L_R \quad (6)$$

Step 3 *Split phoneme clusters*
The phoneme cluster S is split into two clusters, $S_y(R^*)$ and $S_n(R^*)$, according to rule R^*.

Step 4 *Convergence check*
If the stop criterion is satisfied, the splitting process is terminated. If not, steps 1 to 3 are repeated.

4. Experiment

4.1 Experimental Setup

4.1.1 Phoneme Set

In this study, we used the phonemic symbols of the TIMIT database as a reference set [16]. The size of the initial phoneme set was 41 and consisted of 17 vowels and 24 consonants (Table 1). After iterative splitting, we chose various phoneme sets with numbers ranging from 38 to 25 in decreasing order for the recognition experiment. An English speech database read by Japanese students (E2L) was used to train context-independent 3-state monophone HMMs of a left-to-right state topology. This database included phonetic symbols as well as prosodic ones assigned to various words and sentences. It had a total of 80,409 utterances consisting of both individual words and sentences spoken by 200 Japanese students (100 males and 100 females). All sentences and words were respectively divided into 8 sets (about 120 sentences/part) and 5 sets (about 220 words/part). Each sentence and each word was read by about 12 and 20 speakers, respectively, for each set. Table 2 lists the features of the database [17].

Table 2 English word and sentence sets spoken by 200 Japanese students [17].

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

4.1.2 Learner Corpus

We collected English speech data uttered by 55 Japanese students based on shopping, ordering at a restaurant, hotel booking, and other scenarios. Each participant uttered orally translated English speech corresponding to Japanese sentences displayed on a screen. Twenty participants uttered orally translated English speech corresponding to about 200 Japanese sentences and the other 35 participants uttered speech corresponding to about 100 utterances. These utterances were transcribed and their translation quality was evaluated in five grades by English native speakers based on subjective evaluation method used at the International Workshop of Spoken Language Translation [18]. Expressions regarded to be ungrammatical and unacceptable in the learner corpus were given comments for generating effective feedback.

4.1.3 Automatic Speech Recognition

In this study, we used the HTK toolkit [19] to compare the ASR performance using a canonical phoneme set as the baseline with generated reduced phoneme sets using our proposed method considering the real time factor (RTF). We set the RTF to less than 1 second for each recognition result as the experimental condition.

We built context-dependent state-tying triphone HMM acoustic models of various numbers of reduced phoneme sets using the same speech data as the phoneme set design. We developed a 2-gram language model from 5,000 transcribed utterances spoken by 55 Japanese university stu-

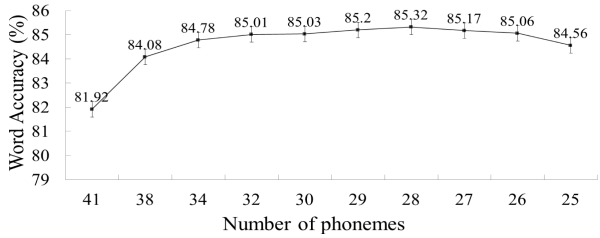


Fig. 3 Word accuracies for various numbers of phoneme sets.

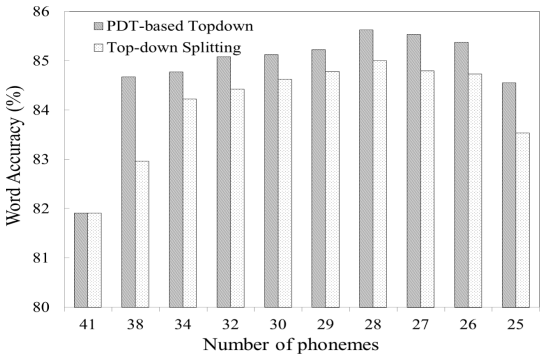


Fig. 4 Word accuracy of different numbers of phoneme sets using the PDT-based top-down method and the top-down splitting method.

dents [7] and transcribed utterances spoken by English native speakers. The pronunciation lexicon consisted of about 35,000 word types related to conversations about travel abroad.

For evaluating our proposed method, we recruited 20 participants between the ages of 18 and 24 and had them to utter orally translated English speech corresponding to the visual prompt from the CALL system. They were Japanese students who had acquired Japanese as their L1 and learned English as their L2. Their communication levels in English were measured using the Test of English for International Communication (TOEIC). Their scores ranged from 380 to 910 (990 being the highest score that can be attained). We recorded the first utterance by each participant to 71 visual prompts in the shopping scene. After collecting all orally translated speech, we had them read out 71 grammatically correct English sentences corresponding to each visual prompt.

4.2 Experimental Results

In order to examine the effectiveness of the proposed method, 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 three aspects: the effect of splitting method, the effect of phoneme occupation probability, and the influence of language model on the increase of homophone words.

4.2.1 Speech Recognition Results

Figure 3 shows the word accuracies for various phoneme sets whose number ranges from 38 to 25 in decreasing order. As shown, all reduced phoneme sets provided better word accuracies than the canonical one, and the reduced phoneme set of 28-phoneme clusters obtained the best performance.

4.2.2 Efficiency of Splitting Method

In order to evaluate the efficiency of our proposed 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 [20] 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.

Figure 4 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).

4.2.3 Effect of Phoneme Occupation Probabilities

We also conducted an experiment to examine the effect of phoneme occupation probability γ_i on recognition accuracy and compared it with the performance in a case in which phoneme occupation probability γ_i was not used as a reference. We used two corpora for calculating γ_i: the domain-independent one [17] used in the previous experiment and the domain-dependent one consisting of 3,464 transcribed utterances by 34 university students. This experiment used the same evaluation data as the one described in Sect. 4.1.3.

Figure 5 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 accu-

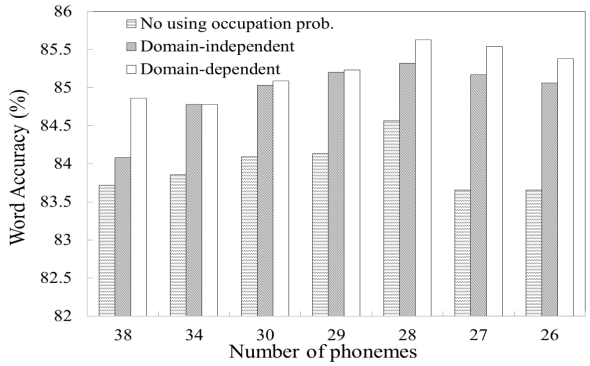


Fig. 5 Word accuracy with different phoneme occupation probability in the same number of phoneme sets.

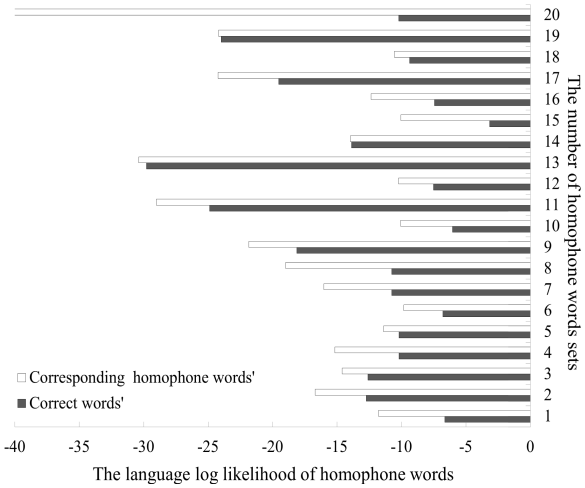


Fig. 6 Language log likelihood differences of homophone words.

racies 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.

4.2.4 Effect of LM for Distinguishing Homophone Words

Reducing the size of the phoneme sets improved the performance of the acoustic model but increased the number of homophones, which are words pronounced the same as another word but differing in meaning and having the same phoneme labeled in the lexicon. This generally causes confusion with language decoding. However, due to the previously mentioned features of the dialogue-based CALL system, homophones can be well distinguished by a good language model. Figure 6 illustrates the difference of the language log likelihoods of some example homophone words. Homophone words and examples of corresponding speech recognized sentences are shown in appendix B.

5. Discussion

The experiment results described in Sect. 4.2 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 4.2.2. The recognition results demonstrate that the discrimination rules function effectively in designing the reduced phoneme set.
- The recognition results discussed in Sect. 4.2.3 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.

Figures 7 and 8 show examples of 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. "C" refers to terminal nodes that indicate a cluster. Vowels are contained inside the area shown by a dotted curve in Fig. 7. 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 based on major distinctive features.
- In Fig. 7, 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".

Our method is generally effective for ASR of second language speech when the mother tongue of users is known. One of the promising applications of our method is dialogue-based CALL systems. We evaluated the ef-

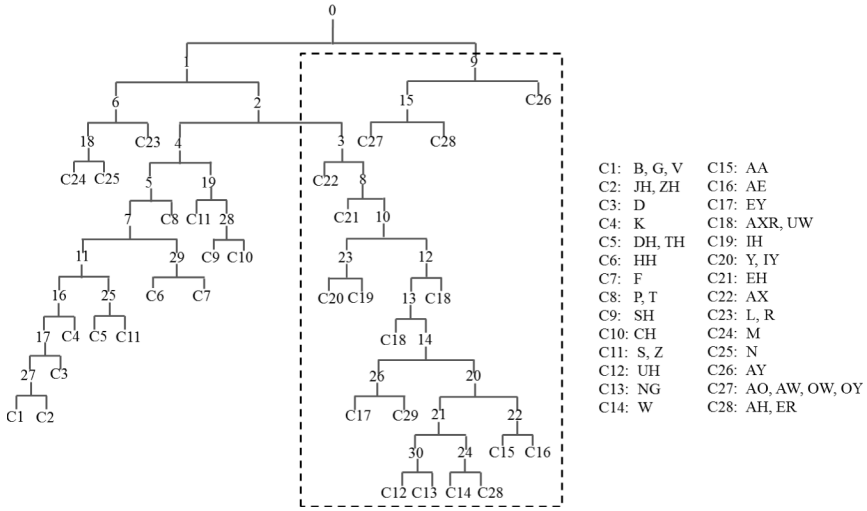


Fig. 7 Result of cluster splitting with PDT-based top-down method in which 28 phonemes were obtained as the final phoneme set. Terminal nodes use "C" to indicate a cluster.

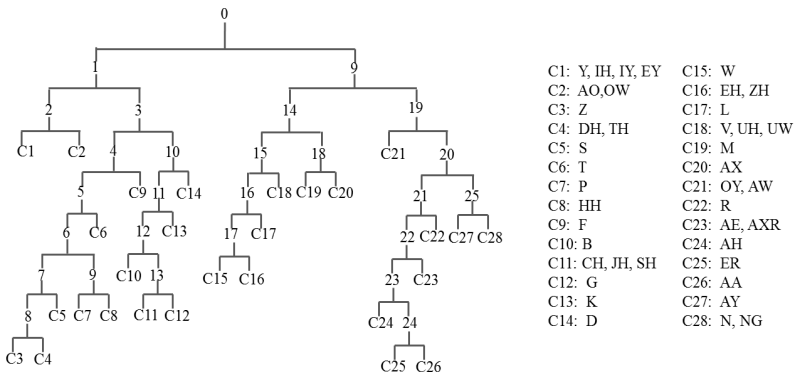


Fig. 8 Result of cluster splitting with the top-down splitting method in which 28 phonemes were obtained as the final phoneme set.

fect of 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 [21], [22]. We compared the ratio of returning effective feedback corresponding to each ungrammatical/unacceptable expression between the two conditions. Our 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. Conclusion and Future Works

In this paper, we presented a novel method of designing a reduced phoneme set for recognizing English spoken by Japanese. The speech recognition results obtained for second language speech collected with a translation game type dialogue-based CALL system showed that the reduced phoneme set with the proposed method achieved a better improvement of speech recognition than the canonical one. The proposed method is effective for ASR that recognizes second language speech when the mother tongue of users is known.

The improvement of the recognition accuracy is dependent on user proficiency. We plan to investigate the relation between the number of phoneme sets and the proficiency of users by collecting more speech data by individuals of various proficiencies. We also plan to evaluate the performance of the proposed method for speech collected with spoken dialogue system in which users' utterances are less restricted and less predictable than translation game type CALL systems.

Acknowledgments

I would like to express my deepest gratitude to Dr. Ichiro Umata of NICT and Emeritus Professor Masuzo Yanagida of Doshisha University for their invaluable comments and helpful discussions.

References

[1] R. Gruhn, W. Minker, and S. Nakamura, "Statistical pronunciation modeling for non-native speech processing," Diss. University of Ulm, 2008.
[2] D. Luo, Y. Qiao, N. Minematsu, Y. Yamauchi, and K. Hirose, "Regularized-MLLR speaker adaptation for computer-assisted language learning system," Proc. ISCA, pp.594–597, Chiba, Japan, Sept. 2010.
[3] Y. Oh, J. Yoon, and H. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech recognition," Speech Communication, vol.49, no.1, pp.59–70, Jan. 2007.
[4] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," Ph.D Thesis, Massachusetts Institute of Technology, 1999.
[5] D. Vazhenina and K. Markov, "Phoneme set selection for Russian speech recognition," Proc. IEEE NLP-KE, pp.475–478, Tokushima, Japan, Nov. 2011.
[6] B. Mak and E. Barnard, "Phone clustering using the Bhattacharyya distance," Proc. ICSLP 96. Proceedings. 4th International Conf. on, vol.4, pp.2005–2008, Philadelphia, PA, Oct. 1996.
[7] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "A dialogue-based English CALL system for Japanese," Proc. NCMMSC, Guiyang, China, Aug. 2013.
[8] 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.
[9] O. Kweon, A. Ito, M. Suzuki, and S. Makino, "A grammatical error detection method for dialogue-based CALL system," Journal of Natural Language Processing, vol.12, no.4, pp.137–156, Dec. 2005.
[10] A. Ito, R. Tsutsui, S. Makino, and M. Suzuki, "Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system," INTERSPEECH, pp.2819–2822, Australia, Sept. 2008.
[11] M. Eskenazi, "An overview of spoken language technology for education," Speech Communication, vol.51, no.10, pp.832–844, Oct. 2009.
[12] 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.
[13] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," Proc. NAACL/HLT, pp.468–475, Rochester, NY, April 2007.
[14] N. Poulisse and T. Bongaerts, "First language use in second language production," Applied linguistics, vol.15, no.1, pp.36–57, Oxford University Press, 1994.
[15] T. Riney and J. Anderson-Hsieh, "Descriptions of Japanese pronunciation of English," JALT Journal, vol.15, no.1, pp.21–36, May 1993.
[16] TIMIT, Acoustic Phonetic Continuous Speech Corpus. http://www.ldc.upenn.edu/Catalog/LDC93S1.html
[17] 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.
[18] E. Sumita, Y. Sasaki, and S. Yamamoto, "Frontier of evaluation method for MT systems," IPSJ Magazine, vol.46, no.5, 2005.
[19] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, and G. Moore, HTK Speech Recognition Toolkit ver. 3.2, Cambridge Univ.
[20] Y. Linde, A. Buzo, and R.M. Gray, "An algorithm for vector quantizer design," IEEE Trans. Commun., vol.COM-28, no.1, pp.84–95, 1980.
[21] H.B. Basiron, "Corrective Feedback in Dialogue-based Computer Assisted Language Learning," Proc. NZCSRSC, pp.192–195, 2008.
[22] W.L. Johnson, S. Marsella, and H. Vilhjalmsson, "The DARWARS Tactical Language Training System," Proc. I/ITSEC, 2004.

Appendix

Table A·1 List of phonemic symbols of English (41 phonemes) corresponding to IPA notation and word examples [16].

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

Table A·2 List of increased homophone word sets and corresponding speech recognized sentences.

Set number	Correct words	Corresponding homophone words	Correct recognized sentence example
1	a	ah	Do you have a more bright color?
2	ah	a	Ah, that is fine. Where is the fitting room?
3	in	on	Can I use an insurance for in japan?
4	in	on	Can you check in a maker?
5	on	in	This is suit on me.
6	are	or	Yes, here you are.
7	big	give	This is a little big, so please give me smaller one.
8	give	big	This is a little big, so please give me smaller one.
9	bought	got	There is a point I did not notice when I bought it.
10	buy	by	Where can I buy it?
11	buy	by	I like this but I will buy it when it is on special sale.
12	by	buy	What is it made by?
13	fold	hold	I want a brief case which is big enough not to fold normal envelope.
14	hold	fold	Please hold this for a present.
15	know	no	I do not know my size.
16	no	know	There is no middle size.
17	tight	type	It is tight for me, so do you have the bigger one?
18	type	tight	Is there another type?
19	wait	weight	I like it but I will wait this gets reasonable price.
20	well	wool	Well, yes, where is fitting room?

Xiaoyun Wang was born in China on September 2, 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.

Jinsong Zhang was born in China on October 4, 1968. He received a B.E. in Electronic Engineering from Hefei University of Technology, China in 1989, an M.E. from the University of Science and Technology of China (USTC) in 1992, and a Ph.D. from the University of Tokyo, Japan in 2000. From 1992 to 1996 he worked as a teaching assistant and lecturer in the Department of Electronic Engineering at USTC. Since 2000, he has been with ATR Spoken Language Translation Research Laboratories as an invited researcher. His main research interests include speech recognition, prosody information processing, and speech synthesis.

Masafumi Nishida received a B.E. in 1997, an M.E. in 1999, and a Ph.D. in 2002, all in Electronics and Informatics, and all from Ryukoku University, Shiga, Japan. From 2002 to 2003, he was a post-doctoral researcher at PRESTO, Japan Science and Technology Corporation. In 2003, he became a Research Associate at the Department of Information Science, Chiba University. From 2007 to 2008, he was an Assistant Professor at the Graduate School of Advanced Integration Science, Chiba University. Currently, he is an Associate Professor at the Department of Information Systems Design, Doshisha University. His research interests include speech recognition, speaker recognition, spoken dialogue systems, and well-being information technology. He received the 2011 Yamashita SIG Research Award from IPSJ. He is now an Associate Editor for IEICE Transactions on Information and Systems.

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 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 the IEICE Japan (Fellow).