Phoneme Set Design Using English Speech Database by Japanese for Dialogue-Based English CALL Systems

Xiaoyun WANG[1], Jinsong ZHANG[2], Masafumi NISHIDA[1], Seiichi YAMAMOTO[1]

[1]Graduate School of Science and Engineering, Doshisha University, 1-3 Miyakodani, Tatara, Kyoto 610-0321, Japan,
[2]Beijing Language and Culture University, No. 15 Xueyuanlu, Haidian, Beijing, 100083, China2
ougyouun@gmail.com, jinsong.zhang@blcu.edu.cn, mnishida@mail.doshisha.ac.jp, seyamamo@mail.doshisha.ac.jp

Abstract

This paper describes a method of generating a reduced phoneme set for dialogue-based computer assisted language learning (CALL) systems. We designed a reduced phoneme set consisting of classified phonemes more aligned with the learners' speech characteristics than the canonical set of a target language. This reduced phoneme set provides an inherently more appropriate model for dealing with mispronunciation by second language speakers. In this study, we used a phonetic decision tree (PDT)-based top-down sequential splitting method to generate the reduced phoneme set and then applied this method to a translation-game type English CALL system for Japanese to determine its effectiveness. Experimental results showed that the proposed method improves the performance of recognizing non-native speech.

Keywords: Phoneme set, Japanese English, Dialogue-based CALL system

1. Introduction

With the integration of automatic speech recognition (ASR) technology, computer assisted language learning (CALL) systems have become highly popular in recent years. Dialogue-based CALL systems act as automated interlocutors that prompt learners to elicit speech in the target language and provide informative feedback that is of enormous educational value in terms of improving the learners' language communication skills (Kawai, 1997; Ito, 2008).

The characteristics of second language speech—the phonemes, prosody, lexicon, grammar, disfluencies, and so on—usually differ significantly from first language speech, and ASR of second language speech is still somewhat of a challenge. There are essentially three approaches that have been considered for dealing with this challenge to improve the recognition accuracy for second language speech: one specific approach and two general approaches.

The specific approach is to design a CALL system that constrains user utterances and takes full advantage of the characteristics of visual systems by giving learners hint stimuli in the form of a keyword, graphical style, or incomplete sentence. One type of dialogue-based CALL system called a "translation game" simulates real-life conversation exercises and constrains user utterances by presenting sentences in their first language as their responses (Wang, 2007; Rayner, 2010).

The first general approach is to adapt models that suit the character of a particular user. There has been considerable research pertaining to appropriate acoustic modeling (AM) for second language speech recognition. Several methods have been proposed, including an acoustic model adaption based on MLLR, which is a popular transformation technique for reducing mismatch in ASR (Leggetter, 1995; Luo, 2010), and acoustic modeling interpolating native and non-native acoustic models (Livescu, 1999).

The second general approach is to use a reduced phoneme set instead of the canonical one. There have been several studies on using reduced sets and analyzing their effectiveness for speech recognition. One approach used a method to generate an initially confusing table of phonemes by logical and statistical deduction and then manually merge some easily confused phones by referencing phonological knowledge (Vazhenina, 2011). Although this approach had 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 (Huang, 2007). However, this type of approach does not consider the acoustic characteristics of the phonemes in real utterances.

These general methods do not utilize the characteristics of dialogue-based CALL systems, such as phonetic knowledge relations between first and second languages and highly constrained user utterances. In an earlier study, considering that erroneous pronunciations have been produced by second language speakers for dialogue-based CALL systems, and that these utterances are highly predictable by a language model, we proposed a reduced phoneme set created with a phonetic decision tree (PDT) using a Japanese-English speech database (Wang, 2013). This method utilizes not only the phonetic knowledge and acoustic characteristics of the phonemes but also the occurrences of English phonemes produced by Japanese. In this paper, we describe our analyses of CALL system performance using our proposed method.

This paper is structured as follows. Section 2 describes our proposed method for phoneme set construction and Section 3 presents the speech recognition experiments in which the performance of the reduced phoneme set is compared with that of the canonical one. Section 4 is a discussion of the experimental results. We conclude in Section 5 with a brief summary.

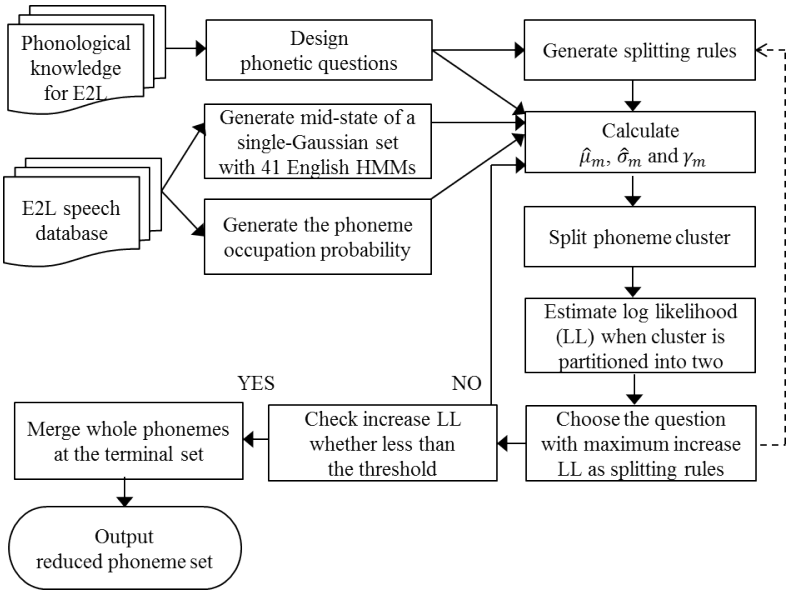

Figure 1: Overview of phoneme cluster splitting procedure.

2. Phoneme set design

Phonetically labeled training data are usually based on the canonical phoneme set for the target language. Mapping that is appropriate between phonetic symbols and native speakers' speech is not always best with second language learners' speech, which contains inherently overlapping distributions of phonetic acoustics and indistinguishable pronunciation in the canonical phoneme set. In actual human-to-human communication, people are constantly exploiting context information, but this function is beyond the ability of even state-of-art ASR technologies that can only exploit a language model at a short range to predict the following words. As a result, the performance of ASR deteriorates for non-native speech. In dialogue-based CALL systems, in which user utterances are highly constrained, the ASR systems of a reduced phoneme set might still function as well as those of canonical phoneme sets because of the narrow search space of the hypotheses. Based on these considerations, we propose creating a reduced phoneme set with a phonetic decision tree-based top-down splitting using a maximum log likelihood criterion in order to maintain the most effective phoneme set for recognizing second language speech.

2.1. Phonetic decision tree

The phonetic decision tree (PDT) is a binary tree using a top-down sequential optimization splitting process predefined by a set of phonetic questions. The log likelihood (LL) is used as the splitting criterion defined by the following equation:

$$L(P_m) \approx \sum_{t=1}^{T} \log[P(\boldsymbol{O_t}, \widehat{\boldsymbol{\mu}}_{\boldsymbol{m}}, \widehat{\sigma}_m)] \cdot \gamma_m$$

In the above equation, $L(P_m)$ represents the log likelihood of the m^{th} phoneme or the phoneme cluster's probability density function. $\widehat{\boldsymbol{\mu}}_{\boldsymbol{m}} and \widehat{\sigma}_m$ refer to the mean vector and the covariance matrix, respectively. γ_m is the 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 canonical phonemes.

2.2. Procedure design

We adopted a 4-step procedure to design the reduced phoneme set using a phonetic decision tree-based top-down method. An overview of the phoneme cluster splitting with a PDT-based top-down method using a maximum log likelihood criterion is shown in Fig. 1.

■ Initialization condition

1. Initial phoneme cluster

We selected the mid-state of 41 context-independent English HMMs as the initial phoneme cluster.

2. Phonetic occupation counts

We selected the counts of each phoneme that had been calculated with the Japanese-English speech database (the training data in our experiments) as the phoneme occupation probabilities.

3. Phonetic questions

Each questions defined by phonological knowledge describes the phonetic effect by neighboring phonemes (Riney, 1993; Poulisse, 1994).

■ Phoneme cluster splitting procedure

1. Estimate LL

Assuming that a parent cluster S is partitioned into two sub-clusters $S_y(Q)$ and $S_n(Q)$ by a question, the increase of log likelihood $\triangle L_Q$ is calculated as

$$\triangle L_Q = L(S_y(Q)) + L(S_n(Q)) - L(S)$$

2. Select optimization phonetic question

Question Q^* is chosen as the splitting question rule when it brings about the maximum increase:

$$L_Q^* = \underset{allQ}{\operatorname{argmax}} \triangle L_Q$$

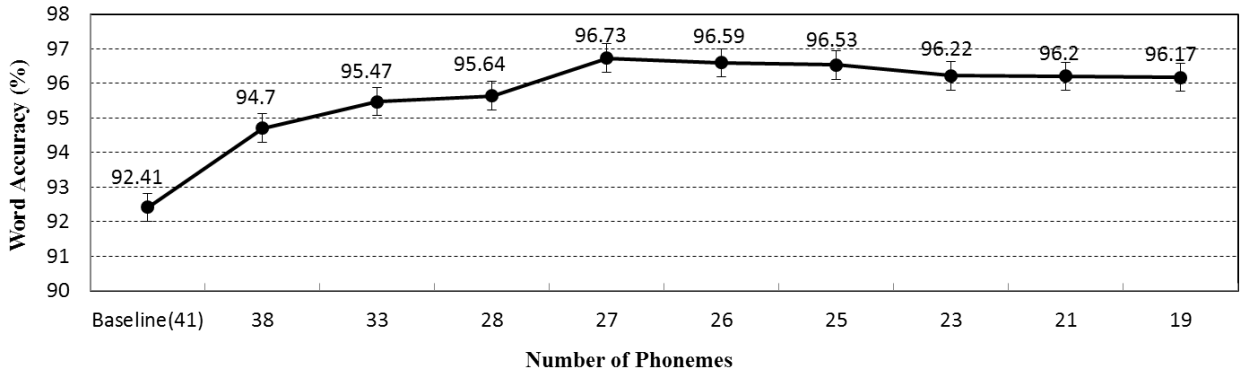

Figure 2: Word accuracy of different phoneme sets.

3. Phonetic questions
Phoneme cluster S is split into two clusters, $S_y(Q^*)$ and $S_n(Q^*)$, according to question Q^*.
4.Convergence check
If the stop criterion is satisfied, the merging process is terminated. If not, step 1-3 are repeated.

3. Experiments

3.1. Experimental conditions

3.1.1. Speech database

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. It had a total of 80,409 utterances consisting of both individual words and longer sentences. The words and sentences were read by 200 Japanese students (100 males and 100 females). All sentences were respectively divided into 8 sets (about 120 sentences/part) and all words were divided into 5 sets (about 220 words/part). Each sentence and each word was read by about 12 and 20 speakers, respectively. Table 1 lists the features of the database(Minematsu, 2004).

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 1: Word and sentence sets prepared from the viewpoint of segmental aspect of English pronunciation.

3.1.2. Phoneme set & AMs & LM

We set the size of the initial canonical phoneme set to 41 by referring to the TIMIT database. We used the HTK Toolkit (Young, HTKToolkit ver. 3.2) and the same speech database as the phoneme reduction experiment to develop state-tying triphone HMM acoustic models of sets with various numbers of phonemes. We then compared the ASR performance using a canonical phoneme set as the baseline with reduced phoneme sets generated using our proposed method to evaluate performance considering the real time factor (RTF). Here, the RTF is the ratio of the overall CPU time required to process all utterances on an Intel Xeon 2.8 GHz PC with 94.4 GB RAM. We set the RTF to less than 1 second for each recognition result as the experimental condition. The pronunciation lexicon consisted of about 35,000 vocabulary words related to conversations about travel abroad.
The language model was a 2-gram model trained from 1,092 grammatically correct utterances and 2,372 spontaneous utterances that were translated orally from Japanese sentences shown on a screen by 34 university students. Evaluation data included 994 utterances read by 14 speakers (7 males and 7 females). Each participant produced 71 utterances. The test set perplexity was 7.9. In the experiments, both the canonical phoneme set and the reduced phoneme set used the same lexicon and language model.

3.2. Experimental results

The word accuracies with the canonical phoneme set and different numbers of the reduced phoneme set determined with the proposed method are shown in Fig. 2.
The results clearly showed that:
• The reduced phoneme set provided better word accuracies than the canonical one.
• The reduced phoneme set consisting of 27 phonemes clusters obtained the best performance.

4. Discussion

Experimental results showed that the reduced phoneme sets provided better word accuracies than the canonical phoneme set and that the reduced phoneme set consisting of 27 phoneme clusters obtained the best performance. Compared to the canonical phoneme set (41 phonemes), the reduced phoneme set decreased the word recognition error rate from 7.59% to 3.27%, a relative error reduction of 56.9%. This clearly demonstrates the suitability of the proposal method for dialogue-based CALL systems. There was a significant difference between the word accuracy of the canonical phoneme set and that of the 27-phoneme set (paired t-test, t(13) = 6.02, $p < 0.001$ for 27 phonemes).
In the overall speech recognition process, the performance of the acoustic model was significantly improved by reducing the number of phonemes with the proposed method.

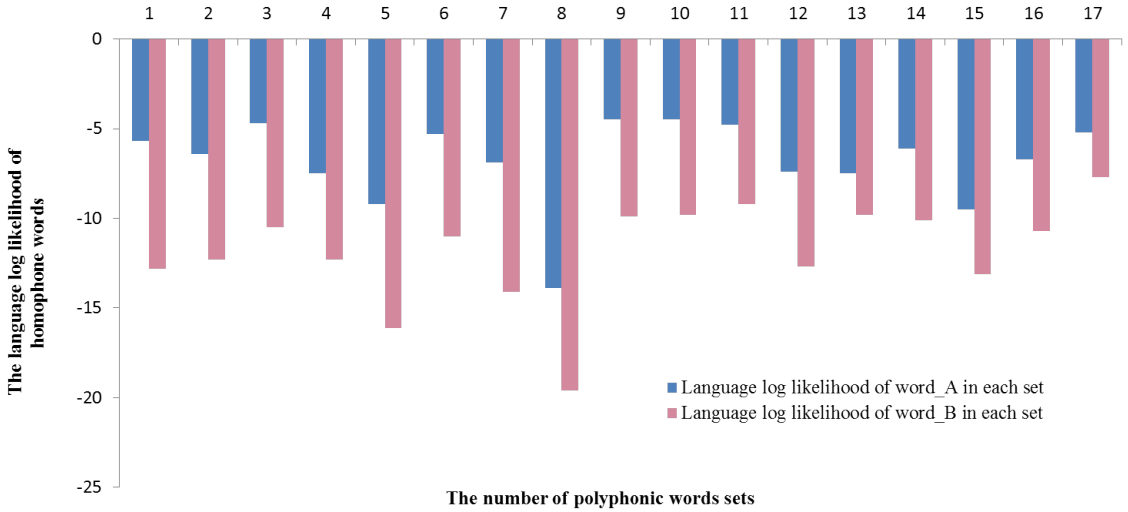

Figure 3: Language log likelihood differences of homophone words.

We should point out, however, that the number of homophone wordsthat is, words with different meanings labeled with the same phonemes has increased in the lexicon. This generally causes confusion with language decoding. However, due to the fact that dialogue-based CALL systems feature user utterances that are highly constrained, homophone words can be distinguished by grammatically related context information. Figure 3 shows the language log likelihood difference of homophone words, where the blue and pink bars denote the language log likelihood of different words labeled with the same phoneme sequence. These results show that our language model can adequately distinguish the increased homophone words resulting from the reduced number of phonemes.

5. Conclusion and future works

In this study, we presented a method of designing a reduced phoneme set for the use of dialogue-based English CALL systems by Japanese. The results of speech recognition experiments showed that the derived phoneme set is sufficient for this type of system using second language speech recognition. We intend to apply this method as the tasks of other second language learning systems to check whether it is still valid and if it can again achieve a better performance than the canonical set.

6. References

G. Kawai. 1997. A CALL system using speech recognition to train the pronunciation of Japanese long vowels, the mora nasal and mora obstruents. *EUROSPEECH*, Rhodes, Greece.

A. Ito. 2008. Recognition of English utterances with grammatical and lexical mistakes for dialogue-based CALL system. *INTERSPEECH*, pages 2819-2822, Brisbane, Australia.

C. Wang. 2007. Automatic Assessment of Student Translations for Foreign Language Tutoring. *Proceedings of NAACL HLT*, pages 468-475, Rochester, NY.

E. Rayner. 2010. A multilingual CALL game based on speech translation. *Proceedings of LREC*, Valetta, Malta.

C. Leggetterr. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech Language.*, pages 171-185.

D. Luo. 2010. Regularized-MLLR speaker adaptation for computer-assisted language learning system. *INTERSPEECH*, pages 594-597, Chiba, Japan.

K. Livescu. 1999. Analysis and modeling of non-native speech for automatic speech recognition. Ph.D. theses, Massachusetts Institute of Technology.

D. Vazhenina. 2011. Phoneme set selection for russian speech recognition. *Natural Language Processing and-Knowledge Engineering (NLP-KE), 2011 7th International Conference on. IEEE*, pages 475-478, Tokushima, Japan.

C. Huang. 2007. Phone Set Generation Based on Acoustic and Contextual Analysis for Multilingual Speech Recognition. *Proceedings of ICASSP*, pages 1017-1020, Hawaii, USA.

X. Wang. 2013. A Dialogue-Based English CALL systemfor Japanese. *Proceedings of NCMMSC*, Guiyang, China.

T. Riney. 1993. Descriptions of Japanese pronunciation of English *JALT Journal*, 15(1):21–36.

N. Poulisse. 1994. *First language use in second language production*. in Handbook of Applied linguistics, Oxford University Press.

N. Minematsu. 2004. Development of English speech database read by Japanese to support CALL research. *Proc. ICA*, pages 557-560.

S. Young. HTK Speech Recognition Toolkit ver. 3.2. Cambridge Univ.