INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

Phoneme Set Design Considering Integrated Acoustic and Linguistic Features of Second Language Speech

Xiaoyun Wang[1], *Tsuneo Kato*[1], *Seiichi Yamamoto*[1]

[1]Graduate School of Science and Engineering, Doshisha University, Kyoto, Japan

euo1101@mail4.doshisha.ac.jp, {tsukato, seyamamo}@mail.doshisha.ac.jp

Abstract

Recognition of second language speech is still a challenging task even for state-of-the-art automatic speech recognition (ASR) systems. Considering that second language speech usually includes less fluent pronunciation and mispronunciation even when it is grammatically correct, we propose a novel phonetic decision tree (PDT) method considering integrated acoustic and linguistic features to derive the phoneme set for second language speech recognition. We verify the efficacy of the proposed method using second language speech collected with a translation game type dialogue-based English CALL system. Experimental results demonstrated that the derived phoneme set achieved higher accuracy recognition performance than the canonical one.

Index Terms: second language speech recognition, phonetic decision tree (PDT), reduced phoneme set (RPS), acoustic variation, lexical discrimination

1. Introduction

Due to the present wave of globalization, there are more opportunities to use foreign languages than ever before. However, in comparison to native speakers, non-native speakers have slightly different pronunciation affected by their mother tongue [1, 2], less knowledge of grammatical structures, and a smaller vocabulary size [3]. These issues result in non-native speakers delivering less fluent pronunciation or mispronunciation, distracting listeners with far-fetched sentences, and expressing themselves in basic words. Celce-Murcia and Goodwin showed that it is difficult to communicate effectively without correct pronunciation because different phonetics and prosody render their speech sounds unnatural to native speakers and impede comprehension of their utterance [4].

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 incorrect grammar. More problematic is when non-native pronunciations become an issue for speech dialogue systems. The vocabulary and grammar of non-native speakers is often limited and simple, but a speech recognizer takes no or only a little advantage of this and is confused by the different phonetics. Hence, recognition of second language speech is still a challenging task even for state-of-the-art automatic speech recognition (ASR) systems.

In order to make ASR systems more tolerant to the acoustic and linguistic variations produced by second language speakers, various methodologies have been proposed to improve the speech recognition accuracy in these aspects. Livescu used an acoustic model interpolating with both native and non-native acoustic models to cover the various pronunciations and accents [5]. Schaden presented an extended lexicon adaptation method using a set of rewriting rules based on the study of phonological properties of the native language and the target language [6]. Oh et al. proposed an acoustic model adaption method for second language speech with a variant phonetic unit obtained by analyzing the variability of second language speech pronunciation [7]. We also proposed using a reduced phoneme set (RPS) created with a phonetic decision tree (PDT) method [8], to improve the recognition accuracy for non-native speech in cases where the mother tongue of the speaker is known, particularly for dialogue-based computer assisted language learning (CALL) systems or mobile platforms [9]-[11]. This method was applied to the recognition of English utterances spoken by Japanese speakers and the experimental results demonstrated that the reduced phoneme set was more effective than the canonical one.

As mentioned previously, most of the ASR technologies have been developed to handle the subject of pronunciation variations in terms of acoustic modeling [5][7] or extended lexicon [6] and grammatical relations in terms of language modeling [12] for non-native speech ASR. However, there are almost no methods that handle the difference between acoustic and linguistic features of non-native and native speech in a uniform way, even if both features share a close relation and should be simultaneously taken into consideration.

In this paper, we propose a novel phoneme set design method, based on the research results obtained with our previously proposed reduced phoneme set from the perspective of handling the acoustic and linguistic features of non-native speech in a uniform way. Our previously proposed reduced phoneme set was created with a phonetic decision tree (PDT) based top-down sequential splitting method [8] that utilizes the phonological knowledge between mother and target languages and their phonetic features, delivering a better recognition performance for non-native speech. Although the reduced phoneme set has in principle a weaker linguistic discriminating performance than the canonical one, the effect of improving its acoustic discriminating performance outweighs the anti-effect of degradation with its linguistic discriminating performance compared with the canonical one. Our new approach considers acoustic and linguistic discriminating performance in a uniform way and optimizes the weighted total of both discriminating performances. We evaluate the proposed method by using speech data collected by our previously developed dialogue-based English CALL system [13] in the form of a translation exercise for Japanese students.

The rest of this paper is structured as follows. In Section 2, we describe the criterion of the phoneme set design. The procedure of designing the reduced phoneme set is introduced in Section 3. Section 4 presents the experiments. Section 5 is a

http://dx.doi.org/10.21437/Interspeech.2016-663

discussion of the experimental results. We conclude with a brief summary in Section 6.

2. Criterion of phoneme set design

The criterion of selecting the reduced phoneme set S is to maximize the weighted total of its acoustic discriminating performance and its linguistic discriminating performance, as

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

where $\triangle L_S$ is the increased acoustic discriminating performance of the reduced phoneme set S compared with the canonical one, $f(S)$ represents its linguistic discriminating performance, and Ψ_S is the set of optimal reduced phoneme set over all reduced phoneme sets calculated with weighted total discriminating performances. Acoustic and linguistic discriminating performances are evaluated in the following procedure.

2.1. Acoustic discriminating performance

We use as the criterion of acoustic discrimination the log likelihood defined by the logarithm of the probability distribution function (*pdf*) of an acoustic model generating the second language speech observation data $\boldsymbol{O}_t = [O_1, O_2, ..., O_T]$. It is defined by

$$L(P_S) \approx \sum_{t=1}^{T} \log[P(\boldsymbol{O}_t, \hat{\boldsymbol{\mu}}_s, \hat{\boldsymbol{\sigma}}_s)] \cdot \gamma_{s,t}, \tag{2}$$

$$\gamma_{s,t} = \sum_{i \in P_s} P(\gamma_{i,t}), \tag{3}$$

where S represents a phoneme set and P is the joint node *pdf* of a phoneme set. $\hat{\boldsymbol{\mu}}_s$ and $\hat{\boldsymbol{\sigma}}_s$ represent the mean vector and the covariance matrix of phonemes assigned to the phoneme set, respectively. $\gamma_{s,t}$ is defined with equation (3), which means a posteriori probability of the model generating the observation data $\boldsymbol{O}_t$. It is used to predict the occupancy frequency of the canonical phonemes that are used in typical Japanese-English speech utterances.
Acoustic discriminating performance $\triangle L_S$ with the reduced phoneme set is defined as

$$\triangle L_S = L(P_r) - L(P_c), \tag{4}$$

where P_r and P_c represent the log likelihood defined in Eq. (2) for the reduced phoneme set and the canonical phoneme set, respectively.

2.2. Linguistic discriminating performance

Various words $w_1, w_2, \ldots, w_n$ of originally different phoneme sequences ordered by the canonical phoneme set are re-figured as one word w^R of the same phoneme sequences by the reduced phoneme sets. Hence, the lexicon labeled by the reduced phoneme set includes more homonyms than that by the canonical one, which worsens linguistic discriminating performance.

Each word has a different probability of occurrence $P(w)$ in utterances by non-native speakers. These probabilities should be considered to estimate the linguistic discrimination of using the reduced phoneme set by collecting a huge transcription of non-native speech data. However, the transcription of non-native speech is more limited than that of the native one and it is extremely difficult to collect enough data of each conversation topic by a considerable number of non-native speakers

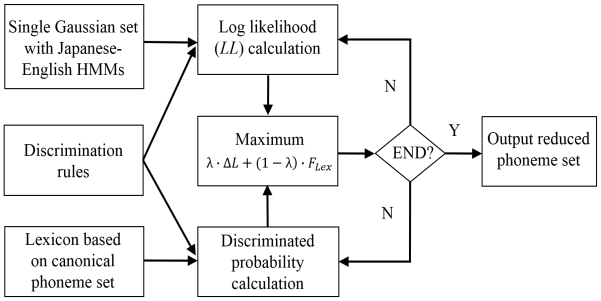

Figure 1: *Phoneme cluster splitting with a phonetic decision tree based top-down method using both log likelihood and discriminating probability as criterion.*

with various language proficiencies. It is difficult to satisfy this requirement for non-native speech, so we use $\mathcal{F}_{Lex}(S)$, the ratio of the number of lexical items of discriminated phoneme sequences in the reduced phoneme set to that of the canonical one, as an approximate approach. Therefore, the linguistic discriminating performance of the reduced phoneme set is written as

$$\mathcal{F}_{Lex}(S) = \frac{\mathrm{C}(w_{diff})}{\mathrm{C}(w)}, \tag{5}$$

where $\mathrm{C}(w_{diff})$ is the count of discriminated lexical items in the lexicon represented by the reduced phoneme set. $\mathrm{C}(w)$ is the total amount of discriminated lexical items in original lexicon represented by the canonical phoneme set.

3. Procedure of designing reduced phoneme set

We used an incremental procedure to design a reduced phoneme set using a phonetic decision tree based top-down method to obtain the optimal reduced phoneme set. Figure 1 shows the overall procedural diagram of the phoneme cluster splitting with the two criteria mentioned above.

■ Initialization condition
1. *Initial phoneme cluster*
To set a cluster of all phonemes of the canonical set as a root cluster and select the mid-state of the context-independent English HMMs of all phonemes as the acoustic model of each phoneme.

2. *Lexicon*
To prepare the discriminated lexical items represented by the canonical phoneme set.

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

4. *Discrimination rules*
To use the discrimination rules designed in our previous study [8], a detailed description is provided below.

As discrimination rules, we used the knowledge of phonetic relations between the Japanese and English languages and the actual pronunciation inclination of English utterances by Japanese. A total of 166 discrimination rules designed in our previous study [8] was used to carry on the preliminary splitting

process for both the acoustic discriminating performance and the linguistic one. The set of rules was designed to categorize each phoneme on the basis of phonetic features such as the manner, position of articulation, and phonological properties between the target language and the mother tongue. In the 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.

■ Phoneme cluster splitting procedure
Step 1 *Calculate log likelihood*
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 log likelihood $\triangle L_R$ is calculated as

$$\triangle L_R = L(S_y(R)) + L(S_n(R)) - L(S) \tag{6}$$

$\triangle L_R$ is the increased log likelihood of the phoneme cluster, which is calculated for all discrimination rules R applicable to each cluster.

Step 2 *Renew lexicon*
The lexicon will be renewed by the current phoneme set based on all discrimination rules R. Here, phonemes existing in the same clusters/rules will be temporarily merged into one phoneme for renewing the lexicon.

Step 3 *Calculate discrimination probability*
The probability of discriminated words in each renewed lexicon by one of the discrimination rules R is calculated as

$$f_{Lex}(R) = \frac{\mathrm{C}(w_{diff}(R))}{\mathrm{C}(w)} \tag{7}$$

Step 4 *Select optimization discrimination rule*
The rule R^* is chosen as the splitting rule when it brings about the maximum of the following formula:

$$\Psi_{R^*} = \arg\max_{all\ \ R}[\lambda \cdot \triangle L_{R^*} + (1-\lambda) \cdot f_{Lex}(R^*)] \qquad (0 \leqslant \lambda \leqslant 1) \tag{8}$$

Step 5 *Split phoneme clusters*
The phoneme cluster S is split into two clusters, $S_y(R^*)$ and $S_n(R^*)$, in accordance with rule R^* selected in Step 4.

Step 6 *Check convergence*
Check whether the stop criterion is satisfied. If yes, the splitting process is terminated. If not, steps 1 to 5 are repeated.

4. Experiments

4.1. Phoneme set

In this study, the phonemic symbols of the TIMIT database were used as a reference set [14]. There are 41 phonemes in the canonical phoneme set, including 17 vowels and 24 consonants (detailed in Table 1). The baseline is ASR using the canonical phoneme set in the experiment.

For the initial phoneme cluster, an English speech database read by Japanese students (E2L) [15] was used to train context-independent 3-state monophone HMMs of a left-to-right state topology. This database includes phonetic symbols as well as prosodic ones assigned to various words and sentences. It contains 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.

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

4.2. Learner corpus

We used our previously developed dialogue-based CALL system [13] to collect English speech data uttered by 55 Japanese students on topics related to shopping, ordering at a restaurant, hotel booking, and others. Each participant uttered orally translated English speech corresponding to Japanese sentences displayed on a screen. The selected utterances were transcribed and their translation quality was evaluated as one of five grades by native English speakers with a subjective evaluation method used at the International Workshop of Spoken Language Translation [16]. Expressions regarded as ungrammatical and unacceptable in the learner corpus were given comments for generating effective feedback.

4.3. Acoustic model & Language model

The E2L speech database mentioned in Section 4.1 was used to train context-dependent state-tying triphone HMM acoustic models of various numbers of phoneme sets. We developed a bigram language model from 5,000 transcribed utterances taken from the learner corpus. The pronunciation lexicon included about 35,000 vocabulary words related to conversation about travel abroad.

4.4. Evaluation data

We collected speech from 20 participants uttering orally translated English speech corresponding to the visual prompt from the CALL system as evaluation data. There were Japanese students who had acquired Japanese as their mother tongue and learned English as their second language. Their speaking styles ranged widely from ones similar to conversation to ones closer to read-speech. The communication levels of participants in English were measured using the Test of English for International Communication (TOEIC) [17]. Their scores ranged from 380 to 910 (990 being the highest score that can be attained). In this study, there were a total of 1,420 utterances recorded by each participant in response to 71 visual prompts.

4.5. Recognition results

In order to verify the performance of the phoneme set with the proposed method, we heuristically chose 25-, 28-, and 32-phoneme sets which are the reliable proficiency-dependent phoneme sets [18][1], for the recognition experiment. We used

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

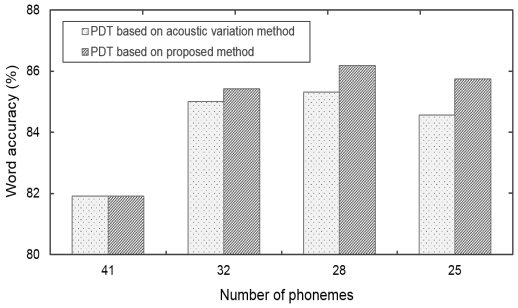

Figure 2: *Word accuracy of canonical phoneme set and various reduced phoneme sets by PDT only based on the acoustic variation method and PDT based on the proposed method.*

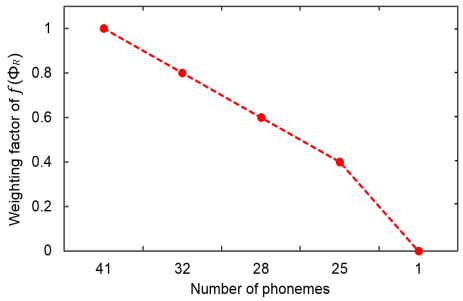

Figure 3: *The best recognition performance of various numbers of phonemes corresponding to weighting factor of probability of discriminated words ($f_{Lex}(R)$).*

the HTK toolkit [19] to compare the performance on ASR implementing the proposed method with that of the canonical phoneme set and the reduced phoneme sets generated by the PDT only based on the acoustic variation method. The results of the reduced phoneme sets created with the PDT based on the acoustic variation method can be achieved by using Eq. (1) when setting $\lambda = 1$.

Figure 2 shows the word accuracy of the canonical phoneme set, the reduced phoneme sets by PDT based on the acoustic variation method and the reduced phoneme sets by PDT based on both acoustic variation and lexicon discrimination methods. We observed the following:

- The reduced phoneme sets with the proposed method delivered a better performance than the canonical phoneme set and other reduced phoneme sets by using PDT based on the acoustic variation method.
- The recognition performance using the proposed method was improved more for fewer numbers of phonemes than for greater numbers of phonemes in comparison to that using the acoustic variation method.

5. Discussion

In order to evaluate the efficiency of the proposed method in consideration of both acoustic and linguistic features, we investigated the relation between the recognition performance of various numbers of phonemes and different weighting factors.

Figure 3 shows the best recognition performance corresponding to the weighting factor of probability of discriminated words ($f_{Lex}(R)$) for various numbers of phonemes generated by our proposed method. It is clear that

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

The probability of occurrence is thought to be largely different depending on the word. We designed the reduced phoneme set assuming the probability is equal among each word, and we obtained experimental results indicating that the proposed method achieved a better improvement of the speech recognition than the canonical phoneme set and the reduced ones by PDT only based on the acoustic variation method. Designing the reduced phoneme set in consideration of the probability of occurrence of each word would decrease the anti-effect of linguistic discrimination, although it is still difficult to collect transcriptions of non-native speech.

6. Conclusion and Future work

In this study, we presented a novel phonetic decision tree (PDT)-based algorithm to derive a phoneme set for second language ASR systems considering both acoustic variations and lexical discrimination. The speech recognition results obtained for English spoken by Japanese collected with a translation game type dialogue-based CALL system showed that the phoneme set created by the proposed method achieved better improvement of speech recognition than the canonical phoneme set and the reduced ones by PDT only based on the acoustic variation method. We have verified that the proposed method is effective for ASR that recognizes second language speech when the mother tongue of users is known.

In the future, we plan to take into consideration the linguistic discriminating performance based on the probability of occurrence of each word. Collecting a huge amount of speech data of non-native speakers of various proficiencies is still difficult, so we will use the probability of each word in a native speech corpus or its interpolation with the probability obtained in a small corpus of non-native speakers as an approximate approach.

7. References

[1] N. Poulisse and T. Bongaerts, *First language use in second language production*, Applied Linguistics: Oxford University Press, vol. 15, no. 1, pp. 36–57, 1994.

[2] N. Minematsu, "Perceptual and Structural Analysis of Pronunciation Diversity of World Englishes," *Proceedings of O-COCOSDA*, Keynote 2, 2014.

[3] S. Krashen, *Principles and practice in second language acquisition (2nd ed.)*, Pergamon: Oxford, 1982.

[4] M. Celce-Murcia and J. Goodwin, *Teaching English as a second or foreign language (2nd ed.)*, New York: Newbury House, 1991.

[5] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," Diss. Massachusetts Institute of Technology, 1999.

[6] S. Schaden, "Generating non-Native pronunciation lexicons by phonological rule," *Proceedings of ICSLP*, 2004.

[7] Y. R. Oh, J. S. Yoon, and H. K. Kim, "Acoustic model adaptation based on pronunciation variability analysis for non-native speech

recognition," *Speech Communication*, vol. 49, no. 1, pp. 59–70, 2007.

[8] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme set design for speech recognition of English by Japanese," *IEICE Transactions on Information and Systems*, vol. E98-D, no. 1, pp. 148–156, 2015.

[9] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," *Proceedings of the HLT–NAACL*, pp. 468–475, Apr. 2007.

[10] F. Xu, S. Schmeier, R. Ai, and H. Uszkoreit, "Yochina: Mobile Multimedia and Multimodal Crosslingual Dialogue System," in Natural Interaction with Robots, Knowbots and Smartphones, New York: Springer, pp. 51–57, 2014.

[11] T. Kawahara and N. Minematsu, "Computer-assisted language learning (CALL) based on speech technologies," *IEICE Transactions on Information and Systems*, vol. J96-D, no. 7, pp. 1549–1565, 2013.

[12] E. Leeuwis, M. Federico, and M. Cettolo, "Language modeling and transcription of the TED corpus lectures," *Proceedings of ICASSP*, pp. I-232–I-235, 2003.

[13] X. Wang, J. Zhang, M. Nishida, and S. Yamamoto, "Phoneme Set Design Using English Speech Database by Japanese for Dialogue-based English CALL Systems," *Proceedings of LREC*, pp. 3948–3951, 2014.

[14] Copyright 1993 Trustees of the University of Pennsylvania, "TIMIT Acoustic–Phonetic Continuous Speech Corpus," https://catalog.ldc.upenn.edu/LDC93S1, accessed Feb. 21 2016.

[15] 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," *Proceedings of ICA*, vol. 1, pp. 557–560, 2004.

[16] E. Sumita, Y. Sasaki, and S. Yamamoto, "Frontier of evaluation method for MT systems," *IPSJ Magazine*, vol. 46, no. 5, 2005.

[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 Feb. 21 2016.

[18] X. Wang and S. Yamamoto, "Second Language Speech Recognition Using Multiple-Pass Decoding with Lexicon Represented by Multiple Reduced Phoneme Sets," *Proceedings of INTERSPEECH*, 2015.

[19] 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.