PAPER

Phoneme Set Design Based on Integrated Acoustic and Linguistic Features for Second Language Speech Recognition

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

SUMMARY Recognition of second language (L2) speech is 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. Considering that the expressions of non-native speakers are usually simpler than those of native ones, and that second language speech usually includes mispronunciation and less fluent pronunciation, we propose a novel method that maximizes unified acoustic and linguistic objective function to derive a phoneme set for second language speech recognition. The authors verify the efficacy of the proposed method using second language speech collected with a translation game type dialogue-based computer assisted language learning (CALL) system. In this paper, the authors examine the performance based on acoustic likelihood, linguistic discrimination ability and integrated objective function for second language speech. Experiments demonstrate the validity of the phoneme set derived by the proposed method.

key words: *second language (L2) speech recognition, unified acoustic and linguistic objective function, reduced phoneme set (RPS), linguistic discrimination ability*

1. Introduction

With the current wave of rapid globalization, people have more opportunities than ever before for speaking in foreign languages in addition to their mother tongue (L1) [1], [2]. However, in comparison to native speakers, non-native speakers have different pronunciation due to their L1 [3], [4], less knowledge of grammatical structures, and a smaller vocabulary size [5]. These issues result in non-native speakers delivering mispronunciation or less fluent pronunciation, confusing listeners with far-fetched sentences, and expressing themselves in basic words. Celce-Murcia et al. showed that it is difficult to communicate effectively without correct pronunciation because different phonetics and prosody render speech sounds unnatural to native speakers and impede comprehension of the utterance [6].

Human beings can eventually understand non-native speech easily because after a while the listener gets used to the style of the talker, e.g., the various insertions, deletions, and substitutions of phonemes or incorrect grammar [7]. More problematic is when non-native pronunciations become an issue for spoken dialogue systems that target tourists, such as travel assistance systems, hotel reservation systems, and systems in which consumers purchase goods through a network. The vocabulary and grammar of non-native speakers is often limited and simple, but a speech recognizer takes no or only little advantage of this and is confused by the different phonetics. Hence, the recognition of second language speech remains 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. Livescu used an acoustic model interpolating with native and non-native acoustic models to cover various pronunciations and accents [8]. 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 [9]. 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 [10]. We also proposed a reduced phoneme set (RPS) created with a phonetic decision tree (PDT) method [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 separately in acoustic modeling [8], [10], [11], lexical modeling [12] and extended lexicon [9], and grammatical relations in terms of language modeling [13] for non-native speech ASR. Read speech produced by non-native speakers has only different acoustic features in comparison to that by native speakers. On the other hand, utterances produced by non-native speakers on their own have different features from the native speech not only in acoustic features but also lexical or grammatical features. These acoustic and linguistic features of non-native speech share a close relation when it comes to the performance of ASR systems, and both features should be taken into consideration when designing non-native speech ASR systems. In this paper, we propose a novel method that maximizes an unified acoustic and linguistic objective function to derive the phoneme set for second language speech recognition.

Our proposal is based on research results obtained with our previously proposed reduced phoneme set and is a natural extension for handling the acoustic and linguistic features of non-native speech in a unified way.

Manuscript received May 19, 2016.
Manuscript revised September 18, 2016.
Manuscript publicized December 29, 2016.

[†]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.2016EDP7207

The previously proposed reduced phoneme set [11] was created with a decision tree based top-down sequential splitting method that utilizes the phonological knowledge among L1, target languages, and their phonetic features, delivering a better recognition performance for non-native speech. The reduced phoneme set alleviates acoustic discrimination ability–the mapping between the sequence of acoustic feature vectors and phonemes for the second language (L2) speakers–but unfortunately it has in principle a weaker linguistic discrimination ability–the mapping between phoneme sequences and word sequences–in comparison to the canonical one. The effect of its improved acoustic discrimination ability outweighs the drop in its linguistic one compared with the canonical phoneme set. Our new approach considers both acoustic and linguistic features in a unified way and optimizes the weighted total of both factors. We evaluate the proposed method by using speech data collected by our previously developed dialogue-based English CALL system [14] in the form of a translation exercise for Japanese students.

In Sect. 2 of this paper, we give an overview of the phoneme set design. The procedure of the phoneme set design is introduced in Sect. 3. Section 4 reports the experimental results. Section 5 is a discussion of the experimental results. We close with a conclusion and a brief mention of our future work in Sect. 6.

2. Overview of the Phoneme Set Design

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

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

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

2.1 Acoustic Likelihood

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

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

where S represents a phoneme set and P_S is the node pdf of a phoneme set S. $\hat{\boldsymbol{\mu}}_s$ and $\hat{\boldsymbol{\sigma}}_s$ represent the mean vector and the covariance matrix of phonemes s assigned to the phoneme set S, respectively. $\gamma_s(\boldsymbol{O}_t)$ is a posteriori probability of the observation data $\boldsymbol{O}_t$ being generated by phoneme s. In here, it is calculated by the canonical phonemes s typically used in Japanese-English speech utterances.

Consequently, increased acoustic likelihood $\triangle L_S$ with the reduced phoneme set is defined as

$$\triangle L_S = L(P_s) - L(P_c), \tag{3}$$

where P_s 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 Discrimination Ability

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

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

$$\begin{aligned}\mathcal{F}_{Lex}(S) &= \frac{PP(W_{M_{diff}(S)})}{PP(W_N)}\\ &= \frac{2^{H(W_{M_{diff}(S)})}}{2^{H(W_N)}},\end{aligned} \tag{4}$$

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

$$H(W_M) = -\sum_{m=1}^{M_{diff}} P(w_m) logP(w_m)$$

$$H(W_N) = -\sum_{n=1}^{N} P(w_n) log P(w_n), \tag{5}$$

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

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

$$\begin{gathered} P(w_m) = \frac{1}{M_{diff}(S)}, \quad P(w_n) = \frac{1}{N} \\ (1 \leqslant m \leqslant M_{diff}(S), 1 \leqslant n \leqslant N), \end{gathered} \tag{6}$$

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

$$\begin{aligned} \mathcal{F}_{Lex}(S) &= \frac{PP(W_{M_{diff}(S)})}{PP(W_N)} \\ &\div \frac{M_{diff}(S)}{N}, \end{aligned} \tag{7}$$

Regarding the words with discriminated phoneme sequences in Eq. (5), w_m has polyphonic ways of pronunciation. The entropies mentioned in Eq. (5) can be extended to the following equations,

$$\begin{aligned} H(W_{M_{diff}(S)}) &= -\sum_{m=1}^{M_{diff}} \sum_{k=1}^{C(w_m)} \frac{P(w_m)}{C(w_m)} log \frac{P(w_m)}{C(w_m)} \\ H(W_N) &= -\sum_{n=1}^{N} \sum_{k=1}^{C(w_n)} \frac{P(w_n)}{C(w_n)} log \frac{P(w_n)}{C(w_n)}, \end{aligned} \tag{8}$$

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

2.3 Discrimination Rules

As discrimination rules for producing PDT, 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 [11] was used to carry out the preliminary splitting process for both the acoustic discrimination ability 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, phonological properties between the target language and the mother tongue. In the splitting method, all phonemes listed in each discrimination rule based on the phonetic features depict similar phonological characteristics and have the possibility to be merged into a cluster.

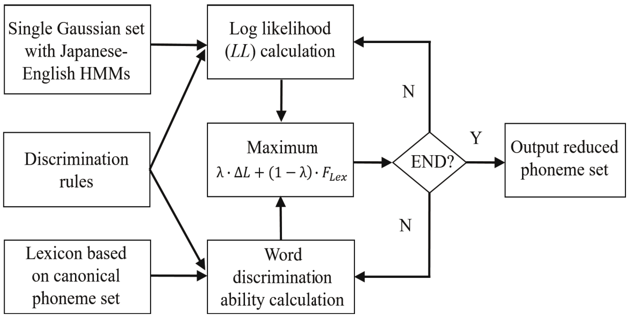
Single Gaussian set with Japanese-English HMMs
Discrimination rules
Lexicon based on canonical phoneme set
Log likelihood (LL) calculation
Maximum λ · ΔL + (1 − λ) · F_Lex
Word discrimination ability calculation
END?
Y
N
N
Output reduced phoneme set

Fig. 1 Phoneme cluster splitting with a PDT-based top-down method using both log likelihood (acoustic part) and word discriminating ability (linguistic part) as criteria.

3. Procedure of the Phoneme Set Design

We followed an incremental procedure in our design of the phoneme set with a PDT-based top-down clustering method to obtain the optimal reduced phoneme set. Figure 1 shows the overall procedural diagram of the phoneme cluster splitting with the unified acoustic and linguistic objective function mentioned in Sect. 2.

3.1 Initialization Conditions

■ Initial phoneme cluster
To set a cluster including all phonemes of the canonical set listed in Table 1 as a root cluster and use the mid-state of the context-independent English HMMs of each phoneme as their acoustic model.

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

■ Discrimination rules
To use the designed discrimination rules (detailed in Sect. 2.3) to carry out the preliminary splitting process. The cluster is split heuristically by the discrimination rules, which were defined by the phonetic features and phonological properties of Japanese-English on the linguistic level. The preliminary splitting process based on designed discrimination rules is used to further calculate log likelihood of each phoneme cluster and word discrimination ability in each renewed lexicon, as described in the following section.

3.2 Phoneme Cluster Splitting Procedure

Step 1 Calculate log likelihood

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 /ʊ/, 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/

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

$$\triangle L_{s,r} = L(s_y(r)) + L(s_n(r)) - L(s) \qquad (9)$$

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

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

Step 3 Calculate word discrimination ability
The probability of words with discriminated phoneme sequences in each renewed lexicon by one of the discrimination rules r is based on Eq. (6) and calculated as

$$\mathcal{F}_{Lex}(s,r) = \frac{M_{diff}(s,r)}{N} \qquad (10)$$

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

Step 4 Select the optimal splitting rule and phoneme cluster to split
The rule r^* and the phoneme culster s^* are chosen when it brings about the maximum of the following formula:

$$\Psi^*_{s^*,r^*} = \underset{all\ s,r}{\arg\max}[\lambda \cdot \triangle L_{s^*,r^*} + (1-\lambda) \cdot \mathcal{F}_{Lex}(s^*,r^*)] \quad (0 \leqslant \lambda \leqslant 1) \qquad (11)$$

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

The phonemic symbols of the TIMIT database were used as a reference set [15]. There are 41 phonemes in the canonical phoneme set, including 17 vowels and 24 consonants. Table 1 lists the phonemes of English in Arpabet notation and IPA notation. 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 (ERJ) [16] 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 2 lists the specific features of the ERJ speech database.

Table 2 English word and sentence sets spoken by 200 Japanese students [19].

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

4.2 Learner Corpus

We used our previously developed dialogue-based CALL system [14] to collect English speech data uttered by 65 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 utterances were transcribed and their translation quality was evaluated and scored one of five grades by native English speakers with a subjective evaluation method used at the International Workshop on Spoken Language Translation [17]. Expressions regarded as ungrammatical and unacceptable in the learner corpus were given comments for generating effective feedback.

4.3 Acoustic Model, Language Model, and Lexicon

The ERJ speech database mentioned in Sect. 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 using about 5,000 transcribed utterances taken from the learner corpus. We used a pronunciation lexicon related to conversation about travel abroad. It consisted of about 45,660 phoneme sequences for 28,000 word types with different meanings. There are

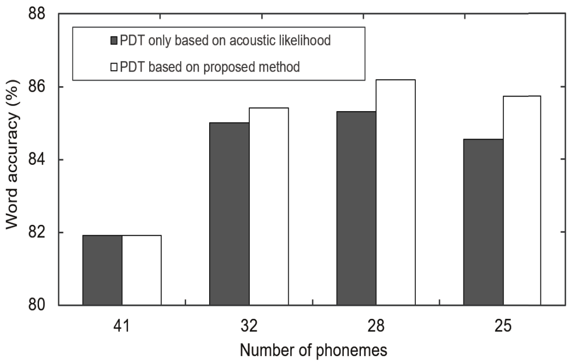

Fig. 2 Word accuracy of the canonical phoneme set and various reduced phoneme sets by PDT only based on the acoustic likelihood and PDT based on the proposed method.

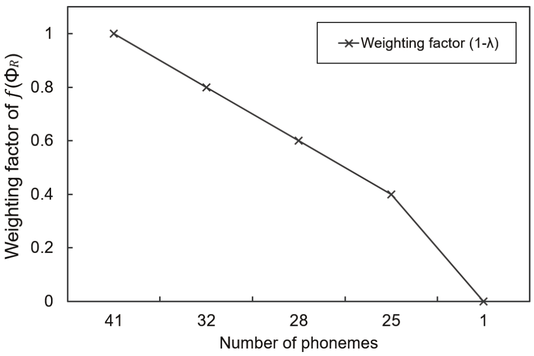

Fig. 3 The best recognition performance of various numbers of phonemes corresponding to weighting factor of word discrimination ability ($\mathcal{F}_{Lex}(s^*, r^*)$).

approximately 43,100 discriminated phoneme sequences in the original pronunciation lexicon represented by the canonical phoneme set.

4.4 Evaluation Data

We collected speech from 20 participants uttering orally translated English speech corresponding to visual prompts from the CALL system as evaluation data. The participants 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) [18]. 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 Experimental Results

4.5.1 Speech Recognition Results

In order to verify the performance of the derived phoneme set by the proposed method, we heuristically chose 25-, 28-, and 32-phoneme sets that had been verified as reliable proficiency-dependent phoneme sets in our previous study [19]† and used them for recognition experiments. We used the HTK toolkit [20] 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 likelihood. This was proposed in our previous study [11], which only used as the splitting criterion the log likelihood given by an acoustic model. The results of the reduced phoneme sets created with the PDT only based on the acoustic likelihood 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 only based on the acoustic likelihood, and the reduced phoneme sets by PDT based on the unified acoustic and linguistic objective function. 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 PDT only based on the acoustic likelihood.
- 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 only based on the acoustic likelihood.

4.5.2 Efficiency of the Reduced Phoneme Set Based on the Unified Acoustic and Linguistic Objective Function

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

Figure 3 shows the best recognition performance corresponding to the weighting factor $(1-\lambda)$ of word discrimination ability ($\mathcal{F}_{Lex}(s^*, r^*)$ in Eq. (11)) for various numbers of phonemes generated by our proposed method. It is clear that

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

5. Discussion

In this section, we investigate from two aspects—one,

[†] 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.

phonemes generated by the proposal in comparison to those generated only by using the acoustic likelihood, and two, word discrimination ability considering equal/different probability of occurrence of each word—to evaluate the efficiency of adopting linguistic discrimination ability for improving speech recognition accuracy for the second language speech.

5.1 Reduced Phoneme Set by Different Methods

The experimental results in Sect. 4.5.1 showed that the reduced phone sets by the proposed method delivered better performance than those with PDT only based on the acoustic likelihood ($\lambda = 1$). To further clarify the efficiency of the proposed method based on the unified acoustic and linguistic objective function, we compare phoneme sets in the final clusters created with the proposed method and PDT only based on acoustic likelihood.

Figure 4 shows an example of phoneme sets in final clusters, which are merged phonemes in the leaves of a decision tree when generating 28-phoneme sets. The left figure depicts the clusters obtained by PDT only based on the acoustic likelihood, and the right one depicts the results obtained by our proposed method. Some phonemes are differently merged into phoneme sets in the left and right figures depending on the difference of criterion of both methods.

One of the specific features of the proposed method compared to PDT only based on the acoustic likelihood is clear from the result that phonemes /P/, /T/, and /K/ are differently merged in the right and left figures in Fig. 4. Phonemes /P/, /T/, and /K/ have the same manner of articulation (plosive), which forces them to be merged into a cluster based on the acoustic feature, but they have a different place of articulation (labial, dental, palatal). In the left figure, which shows the results with PDT only based on the acoustic likelihood, phonemes /T/ and /P/ are merged because the place of articulation between /T/ and /P/ is nearer than that between /K/ and /P/. On the other hand, phonemes /K/ and /P/ are merged in the right figure on the basis of proposed method. This difference can be explained by the fact that the probability of homophones that have the same phoneme strings when /K/ and /P/ are merged achieved 0.9% absolute reduction in comparison to that of homophones that have the same phoneme strings when /T/ and /P/ are merged. Experimental results show that the linguistic discriminating ability decreases more when /T/ and /P/ are merged.

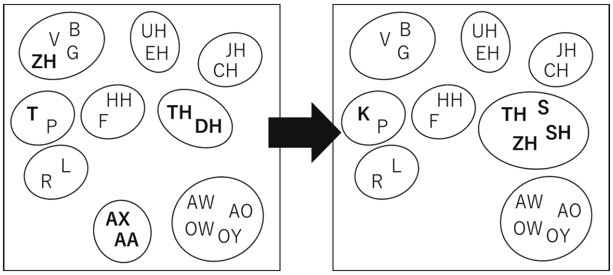

Fig. 4 Final clusters of 28-phoneme sets generated by both PDT only based on the acoustic likelihood (left) and PDT based on unified acoustic and linguistic objective function (right). The different phonemes are shown in bold. (Non-merged phonemes are not included in the figure.)

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

The occurrence probability of each word is thought to be largely different. Designing the reduced phoneme set in consideration of the occurrence probability of each word would affect the linguistic discrimination ability, although it is still difficult to collect transcriptions of non-native speech. We temporarily check the effect of the occurrence probability of each word using a small corpus.

In the case of the equal occurrence probability of each word, we utilized the same computational method for the reduced phoneme set design (refer to Eq. (6)). In the case of different occurrence probability of each word, we estimated the probability using the text corpus of evaluation data mentioned in Sect. 4.4 and the learner corpus mentioned in Sect. 4.2. The occurrence probability of each word for each corpus satisfies

$$\sum_{m=1}^{M_{diff}} P(w_m) = \sum_{n=1}^{N} P(w_n) = 1. \tag{12}$$

Table 3 shows the word discrimination ability for the discriminated phoneme sequences of all lexicon items by the canonical phoneme set and various numbers of the reduced phoneme sets, considering the equal probability of occurrence of each word. Even if the number of the phoneme set is reduced to 25 (39% reduction of phoneme numbers), only 5.3% of lexical items are merged into a confusable word class. Table 4 shows the word discrimination ability for discriminated phoneme sequences of all lexicon items used in the evaluation data by the canonical phoneme set and various numbers of the reduced phoneme sets, considering word occurrence probability estimated in the learner corpus. In this case, 5.3% of lexical items are also merged into a confusable word class. This is smaller than expected in light of the number of reduced phonemes, which indicates that the phoneme occurrence distribution is largely distributed.

The vocabulary size of the lexicon used in the experiment is 28,000, which we feel is sufficiently large for the productive vocabulary of second language speakers. The literature on English as a foreign language for Japanese learn-

Table 3 Word discrimination ability (%) for discriminated phoneme sequences of **all lexicon items** represented by the canonical phoneme set and various numbers of phoneme sets considering **the equal occurrence probability** of each word. The reduction rate in comparison to the canonical phoneme set is given in parentheses.

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

Table 4 Word discrimination ability (%) for discriminated phoneme sequences corresponding to words used in **evaluation data** represented by the canonical phoneme set and various numbers of ones considering **occurrence probability estimated with the learner corpus**. The reduction rate in comparison to the canonical phoneme set is given in parentheses.

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

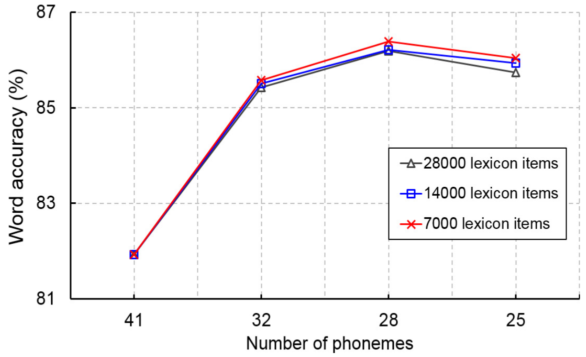

Fig. 5 Word accuracy of canonical phoneme set and various reduced phoneme sets by proposed method with different vocabulary size of the lexicon.

ers [21], [22] reported the mean vocabulary size of the aural and written case to be approximately 5,000, consisting of 3,000 words with different categories. Therefore, we use other vocabulary sizes for the lexicon, 14,000 and 7,000, to verify the efficacy of the proposed method.

Figure 5 shows the word accuracy of the canonical phoneme set and various reduced phoneme sets by the proposed method with different vocabulary sizes of the lexicon. Experimental results show that the lexicon with smaller vocabulary size achieved better recognition performance than the larger ones.

6. Conclusion and Future Work

In this study, we proposed a method of designing a phoneme set for second language speech maximizing a unified acoustic and linguistic objective function of second language speakers and implemented the method as a decision tree to derive a reduced phoneme set. We applied the reduced phoneme set developed with the proposed method to English utterances spoken by Japanese collected with a translation game type dialogue-based CALL system. The experimental results showed that it achieved a greater improvement in speech recognition performance than the canonical phoneme set and the reduced ones by PDT only based on the acoustic likelihood. We have verified that the proposed method is effective for ASR that recognizes second language speech when the mother tongue of users is known.

In future, we will carry on examining linguistic discrimination ability based on more accurate word occurrence probability in other corpora. Collecting a huge amount of speech data of non-native speakers of various proficiencies is still quite difficult, so we plan to use the occurrence 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.

Acknowledgments

This research was supported by a contract with MEXT number 15K02738.

References

[1] C. Kramsch and S. Thorne, "Foreign language learning as global communicative practice," Globalization and Language Teaching, pp.83–100, 2002.

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

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

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

[5] S. Krashen, Principles and practice in second language acquisition, 2nd ed., Pergamon, Oxford, 1982.

[6] M. Celce-Murcia and L. McIntosh, Teaching English as a second or foreign language, Heinle & Heinle, Boston, MA, 1991.

[7] R.E. Gruhn, W. Minker, and S. Nakamura, "Statistical pronunciation modeling for non-native speech processing," Springer Berlin Heidelberg, 2011.

[8] K. Livescu, "Analysis and modeling of non-native speech for automatic speech recognition," Diss. Massachusetts Institute of Technology, pp.25–30, 1999.

[9] S. Schaden, "Generating non-native pronunciation lexicons by phonological rule," Proc. ICSLP, pp.2545–2548, 2004.

[10] 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.

[11] 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.

[12] K. Livescu and J. Glass, "Lexical modeling of non-native speech for automatic speech recognition," Proc. ICASSP, vol.3, pp.1683–1686, 2000.

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

[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 April 8, 2016.

[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] E. Sumita, Y. Sasaki, and S. Yamamoto, "Frontier of evaluation method for MT systems," IPSJ Magazine, vol.46, no.5, pp.552–557, 2005.

[18] TOEIC, "Mapping the TOEIC and TOEIC Bridge Tests on the Common European Framework of Reference for Languages,"

https://www.ets.org/toeic/research/mapping_toeic, accessed April 8, 2016.

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

[20] 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.

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

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

Xiaoyun Wang 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 Doshisha University. Her research interests include computer assisted language learning systems, speech recognition, language acquisition, and spoken language processing. She is a member of the ASJ, the IEICE Japan, and the ISCA.

Tsuneo Kato received his B.E., M.E., and Ph.D. degrees from the University of Tokyo in 1994, 1996, and 2011. He joined Doshisha University in 2015, where he is currently an associate professor in the Department of Intelligent Information Engineering. He had previously worked at KDDI R&D Laboratories Inc. since 1996. He has been engaged in the research and development of automatic speech recognition and intelligent user interfaces. He received an IPSJ Kiyasu Special Industrial Achievement Award in 2011. He is a member of IPSJ, ASJ, IEICE, and IEEE.

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).