Improvement of Non-native Speech Recognition by Effectively Modeling Frequently Observed Pronunciation Habits

Nobuaki MINEMATSU[*†], *Koichi OSAKI*[*], *and Keikichi HIROSE*[**]

[*]Graduate School of Information Science and Technology, University of Tokyo
[†]Department of Speech, Music and Hearing, Royal Institute of Technology (KTH), Stockholm
[**]Graduate School of Frontier Sciences, University of Tokyo
{mine,koichi,hirose}@gavo.t.u-tokyo.ac.jp

Abstract

In this paper, two techniques are proposed to enhance the non-native (Japanese English) speech recognition performance. The first technique effectively integrates orthographic representation of a phoneme as an additional context in state clustering in training tied-state triphones. Non-native speakers often learned the target language not through their *ears* but through their *eyes* and it is easily assumed that their pronunciation of a phoneme may depend upon its grapheme. Here, correspondence between a vowel and its grapheme is automatically extracted and used as an additional context in the state clustering. The second technique elaborately couples a Japanese English acoustic model and a Japanese Japanese model to make a parallel model. When using triphones, mapping between the two models should be carefully trained because phoneme sets of both the models are different. Here, several phoneme recognition experiments are done to induce the mapping, and based upon the mapping, a tentative method of the coupling is examined. Results of LVCSR experiments show high validity of both the proposed methods.

1. Introduction

Large vocabulary continuous speech recognition (LVCSR) in an ideal condition, such as read, clean, monotonous, and native speech, has realized its performance of approximately 98% and research targets have been shifted to real-world speech recognition. Recognition of spontaneous, noisy, expressive, and/or non-native speech is considered as challenging task and a lot of efforts are being made to improve the performance. Among these tasks, we are focusing on non-native speech recognition. Why does non-native speech have to be recognized automatically ? One of LVCSR applications is automatic captioning of lecture speech and it is especially required to realize a barrier free society. The lectures are not always done in native languages and, in international meetings, more than half lecturers have to talk in a non-native language, which is English most of the cases. Also, public and international statements are often done in English irrespective of the speaker's native language. Public talk and private talk, which is required to be transcribed automatically and correctly ? The authors consider that more practical attention should be paid to non-native *English* speech recognition. Non-native English speech can be treated as speech of a single language ? As non-native speech is regarded as interfusion between the speaker's native language and the target language[1], we consider that it should be defined as a pair of the two languages, such as English spoken by people whose native language is Japanese, and researches should be made by focusing on phenomena dependent on either of the two languages.

To recognize non-native speech, larger distortions must be handled adequately in acoustic modeling[2], language modeling, lexical modeling[3], and/or decoding strategy[4]. In the current paper, acoustic modeling is mainly investigated. Some researches have been already reported where non-native speech samples were used as adaptation data to adapt native acoustic models. In this case, however, phonetic structure embedded in the model set in state-clustering of triphone training or in mixture-clustering of MLLR adaptation often mismatches with the structure found in non-native speech. One solution is capturing the structure more coarsely and some performance improvements were reported[2]. The state-level or mixture-level phonetic structure was introduced into the acoustic models to realize their efficient training or adaptation, and the introduction is based on an assumption that the structure shall be commonly found in input speakers. But this is invalid without question to non-native speakers. Another and maybe better solution is to build the baseline speaker-independent models only with non-native speech samples and adapt them with a particular non-native speaker's utterances. Our previous study[5] examined this method with a Japanese English read speech database[6]. However, we found that Japanese speakers had extremely various levels of pronunciation proficiency. The baseline models were their *averaged* models and the structure in the model set was not necessarily adequate for input Japanese speakers. To solve this problem, a novel technique was proposed where interpolated models between Japanese English (JE) and American English (AE) models with the weighting factor producing the maximum likelihood score were estimated and the structure suited for the particular speaker was estimated from the interpolated models. This technique improved recognition performance compared to that obtained in the models adapted to the input speaker from the baseline models with MLLR techniques.

Our previous study, however, had two clear drawbacks. 1) The previous study focused upon only the mixture-level clustering performed in the MLLR adaptation. The state-level clustering in the baseline triphone training remained intact. 2) The interpolated models were generated between JE and AE models and it means that the above technique only helps Japanese speakers with better pronunciation proficiency. For the first problem, in this paper, more adequate preparation of a question set for the state clustering is examined by carefully considering how Japanese learn English in their school days. The second problem is discussed by making multi-path parallel triphone models from JE triphones and Japanese Japanese (JJ), which is Japanese spoken by Japanese, triphones. This is because non-native pronunciation is considered as interfusion between the target language and the native language[1].

2. Development of an English database read by Japanese students

Most of the current speech and language techniques are based upon statistical methods and they naturally require databases. No databases produce no outcomes. The first author of this paper and his co-workers designed and developed a database of English words and sentences *read* by Japanese students[6]. This database was designed mainly to be used in CALL (Computer Aided Language Learning) system development and in English pronunciation education. The database is divided into two sets. The first set is related to segmental aspect of pronunciation and the other is related to its prosodic aspect. Table 1 shows a list of word/sentence sets in the database. A unique recording strategy was adopted where a speaker was asked to repeat reading a given word or sentence until he/she judged that the correct pronunciation was done. Even under this strategy, in a subsequent analysis, a great number of pronunciation errors were found according to pronunciation proficiency of the individual speakers. This somewhat artificial recording certainly came to add some distortions on the recorded speech samples. No examples of illegal or ungrammatical wording, very few occurrences of word fragments, restarts, interjections and so on. However, these phenomena are often treated in other modules of speech recognition such as language models and decoders. In this meaning, the above database can be said to be rather optimal for training acoustic models of English spoken by Japanese.

The database contains speech samples of 100 male and 102 female students. The total number of sentence utterances is approximately 12,000 and that of word utterances is 22,000 for each gender. In other words, the amount of speech samples per speaker is about 120 sentences and 220 words. In this paper, a part of the sentence data were used for training HMMs and the remaining part were used for evaluating the proposed methods.

3. Extended vowel acoustic models based upon their graphemes

3.1. Phonetic structure embedded in the model set

In training of standard triphone models, top-down state-level tying is commonly used to reduce the number of free parameters and increase robustness of the resulting models. The state tying is usually realized as a decision tree clustering to avoid the unseen data problem and a question set are prepared only with regard to left and right phoneme contexts. This is based upon an assumption that acoustic realization of a phoneme (allophone) is fully characterized by its preceding and succeeding phonemes. If this assumption is not valid enough, an additional context should be introduced. What kind of information source affects the acoustic realization of phonemes in JE?

Table 1: *Word and sentence sets contained in the database*

set	size
Phonemically-balanced words	300
Minimal pair words	600
TIMIT-based phonemically-balanced sentences	460
Sentences including phoneme sequences difficult for Japanese to pronounce fluently	32
Sentences designed as a test set	100
Words with various accent patterns	109
Sentences with various intonation patterns	94
Sentences with various rhythm patterns	121

Most of Japanese start to learn English as their second language when they are 12 years old. And most of the cases, they learn English not from their *ears* but from their *eyes*. Further, the Japanese language has English-alphabetic representation of its sounds, so-called *rōmaji* representation. Children learn this representation before they get 12. In other words, when they start to learn English, each of English alphabet is already and strongly mapped to a particular Japanese sound in their brains. PRONLEX pronunciation lexicon shows that above, common, and useful should have the same vowel /ax/ (schwa). But it is difficult for Japanese to assign a single phone to these segments.

These facts easily let us expect spelling of words strongly influences their pronunciation by Japanese. If correspondence between a phoneme and its grapheme is easily extracted from spelling of a word and its phonemic sequence, it can be used as an additional context in the clustering. But English is infamous for its difficult correspondence between the two. In this paper, only vowels are focused for the following two reasons. One is differences between the Japanese phonemic system and the English one. Difference in the vowel system is much larger than that in the consonant system between the two. This implies that acoustic distortion found in vowel pronunciation is much larger than that in consonant pronunciation in JE and the larger distortion should be handled primarily. The other reason is easiness to capture phoneme-to-grapheme correspondence in vowels compared to that in consonants, which is described shortly.

3.2. Correspondence between vowels and their graphemes

For a given word, the correspondence between vowels and their graphemes are automatically extracted by the following simple rules using the word's spelling and its phonemic sequence found in PRONLEX pronunciation lexicon.

1. Number of vowels N_v of the word is obtained from the phonemic sequence.
2. The spelling is separated into vowel segments and consonant ones. The former are defined as segments comprised only by any of aiueo and the latter are segments which are not vowel segments. Number of the vowel segments comprised by only a single e is obtained (n_e).
3. If a consonant segment has letter y and its preceding and succeeding letters, if exist, are not any one of aiueo or y, the y is treated as a new vowel segment.
4. If N_v is equal to the number of vowel segments, the vowel-to-grapheme correspondence is extracted as it is.
5. If N_v+n_e is equal to the number of vowel segments, the vowel-to-grapheme correspondence is extracted by ignoring the e segments.
6. If a vowel in the word is a diphthong and its corresponding vowel segment precedes y or w, then it is included in the vowel segment.
7. If the above procedure cannot find the correspondence, vowels in the word are labeled as *missing* vowels.

These rules are incomplete. But they could show the vowel-to-grapheme correspondence of about 90% of the words in the database and their accuracy was about 95%. The procedure converted vowels into extended ones considering their graphemes. Table 2 shows examples of the extended vowels. Suffixes mean the graphemes and the order of the extended vowels reflects the frequency. Some extended vowels had few examples and these extended vowels should be merged into missing ones for stable

Table 2: *Examples of the extended vowels*

original	extended vowels (_? means that it is *missing*)
ax	ax_a, ax_e, ax_o, ax_u, ax_i, ax_ou, ax_?
eh	eh_e, eh_ea, eh_a, eh_ai, eh_?
ih	ih_i, ih_e, ih_io, ih_a, ih_y, ih_o, ih_ui, ih_ee, ih_ia, ih_?
ay	ay_i, ay_y, ay_ie, ay_ui, y_uy, y_y, y_?

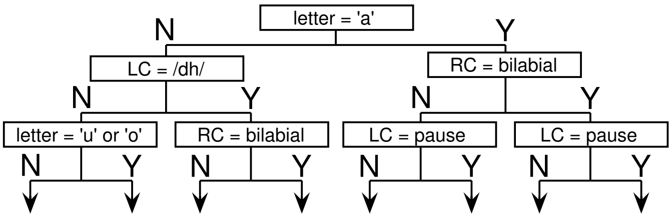

Figure 1: *Decision tree of the central state of /ax/*

Table 3: *Experimental conditions*

sampling	16bit / 16kHz
window	25 ms length and 10 ms shift
parameters	12MFCC + 12ΔMFCC + ΔPower
training data	JE male samples (68 speakers)
testing data-A	JE male samples (32 speakers, PP<200)
testing data-B	JE male samples (10 speakers, PP<500)
HMMs	tied-state triphones (#state=1000, #mix=16)
LM	WSJ-based bigram and trigram
decoder	Julius v3.3p2
vocabulary	20 K

Table 4: *Correct recognition rates (task A)*

BL1	BL2	full	θ=1.0	θ=0.8	θ=0.5
69.5	68.2	70.5	71.5	70.4	70.7

estimation of parameters. Here, for each original vowel, accumulated occurrence of its more frequent extended vowels was counted and 98% of them were finally defined as extended and the remainings were labeled as missing. As a result, the number of vowels was increased from 16 to 95.

3.3. Extended tree-based clustering

In addition to questions on left and right phoneme contexts, those on graphemes of the central phonemes were introduced as for vowels. If an original vowel is converted into n extended and missing vowels, (2^n-2)/2 different binary divisions of the vowels are possible based upon their graphemes. The number of questions was increased from 118 in the baseline clustering to 1,414 in the extended clustering. Figure 1 shows an example of the obtained trees. The tree shows that the primary factor to acoustically characterize an schwa sound is its grapheme. About 60 % of the original vowels adopted the extended questions for their trees at depth less than 4. These indicate validity of using the graphemes as context in the state clustering.

Top-down clustering firstly models seen triphones individually and they are merged subsequently based upon the state-tying. The proposed method increases the number of vowels and the amount of training data for each of the triphones is reduced. To avoid this problem, we examined adequate selection of the extended vowels by referring to the extended trees. The extended questions were re-ordered based on their frequency of being used in the clustering. Only the extended vowels so as to cover θ percentage of all the occurrences of the extended questions were adopted and the others were re-labeled as missing.

3.4. Evaluation of the proposed method

Two kinds of baseline triphone models were prepared. One was common non-extended triphones. The other was also non-extended triphones which were generated from extended triphones by merging their extended vowels of an original vowel into one in the state-tying. Difference of the performance between the two baseline models shows influence of the separate modeling of each of the seen triphones with their graphemes. Table 3 shows the experimental conditions and two kinds of testing data (tasks A and B) were prepared. As for threshold θ for covering the extended questions, we examined three cases of θ=1.0, 0.8, and 0.5. Table 4 shows correct recognition rates for task A, where 'BL1' and 'BL2' are the two baseline models and 'full' is the extended models without any post-selection, which is different from 'θ=1.0' models. This is because some extended triphones in 'full' models are completely irrelevant to any extended questions actually used. As expected, the separate modeling of the extended triphones with a reduced amount of data degrades the performance. But considering questions on graphemes, the performance is improved to be even better than that of BL1. The table also indicates that all the extended triphones relevant to the extended questions should be introduced. As for task B, performance of the extended triphones overcame that of the normal triphones. Although the detail results are not listed here due to limit of space, the best performance was obtained in the extended triphones in seven speakers out of ten.

4. Parallel models of JE and JJ triphones

As discussed in Section 1, non-native speech can be viewed as interfusion between the native language of the speaker and the target language[1]. In our previous study, the 'interfused' models between JE and AE models were used to estimate phonetic structure suited for the speaker[5]. The resulting models, however, were only applicable to speakers with better pronunciation proficiency. This section investigates another interfused model set, which are implemented as parallel models of JE and JJ.

4.1. Phoneme replacement analysis of JE with JJ

JE includes various phoneme errors (replacements, insertions, and deletions). Part of speech samples of all the training speakers were recognized with phoneme network grammars of JE and JJ, which represented possible phoneme errors expected by considering the intended sentence and characteristics of JE. Each phoneme was allowed to be replaced with an articulatory-similar or graphemic-similar JJ phoneme. JJ models were provided by CSRC[7]. All the analyses were done with JJ and JE monophones. Results showed that about 30 % of the instances of an English phoneme were replaced by JJ phonemes on average and, for several phonemes, more than 50 % were replaced although all of the input speakers were training speakers of the JE models[8]. This finding strongly indicates that for some phonemes of perhaps very poor speakers, JJ phoneme models are much more adequate to recognize their English speech.

4.2. Mapping between JE and JJ triphones

Mapping between JE and JJ models has to be carefully designed because they have different phoneme sets and it was done by referring to results of the above phoneme replacement analy-

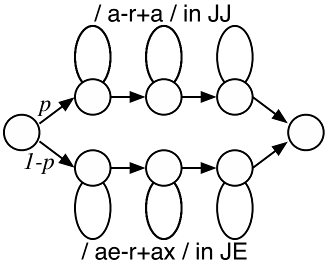

Figure 2: *An example of the parallel models*

Table 5: *Examples of the mapping between JE and JJ phonemes*

JE	JJ	JE	JJ	JE	JJ	JE	JJ
ae	a	ah	a	ch	ch	dh	z
eh	e	nx	n	ih	e	jh	j
oy	o i	er	a	sh	sh	th	s
uh	u	aw	a u	ay	a	zh	sh

Table 6: *Performance improvement by the parallel models*

speaker	GOP	JE[%]	JE+JJ[%]
A	-1086.25	62.26	62.26
B	-904.88	67.52	68.15
C	-800.35	74.50	74.50
D	-749.12	72.03	77.12
E	-689.38	64.78	66.04
F	-600.19	65.13	67.76
G	-550.26	66.14	66.14
H	-505.33	66.45	64.47
I	-447.81	63.95	61.22
J	-195.04	54.78	52.23

ses. For each of JE phonemes, the JJ phoneme which was used most frequently for its replacement was adopted as its mapping phoneme. Table 5 shows some examples of the mapping, with which, however, every JE triphone was not mapped to a JJ triphone because Japanese has its own phonological constraint that every consonant is followed by a vowel with some exceptions. If a corresponding JJ triphone was not found, a biphone, if not exist, a monophone was considered for the mapping.

4.3. Branching probabilities of the parallel models

After the mapping was completed, branching probability p was estimated for each central phoneme of the parallel models in the following way. Here, branching factor p=0.0 means that the parallel model is the same as the JE model and p=1.0 means that it is the same as the JJ model (See Figure 2). Using adaptation data, p was estimated separately for each of the central phonemes so as to maximize the averaged likelihood score of speech segments of the phoneme. Only if the difference between the maximum score and the score obtained by the normal JE triphones was larger than threshold α, the parallel triphone models were adopted (p>0.0) to characterize the central phoneme. α was determined experimentally to be 0.3. The procedures gave us a unique value of p for all the triphones with the same central phoneme but, with minor modifications, it is also possible to assign different values to different triphones.

4.4. Evaluation of the proposed method

Evaluation was done only in task B, where about 100 utterances were used to adaptively estimate p for each speaker in advance. Since the proposed method was expected to improve the performance in the cases of poorer speakers, we calculated Goodness Of Pronunciation (GOP) scores[9] for each speaker using the adaptation data. Table 6 shows correct recognition rates for all the speakers in ascending order of GOP scores. Speaker A is the poorest speaker and J is the best speaker, who is bilingual. In the case of J, the baseline (JE) performance is extremely low because he is too good in speaking English.

For the best three speakers in GOP, as expected, the performance was degraded by the proposed method. The performance improvements were only found in the speakers in a middle or a low range of GOP. The authors consider that the proposed method was a rather rough method to model the interfusion because it used only one-to-one mapping between JE and JJ phonemes and branching probability was assigned equally to all the different triphones with the same central phoneme. However, the recognition experiments showed validity of the proposed method for intermediate or poor speakers. The above refinements are expected to bring about further improvement.

5. Conclusions

In this paper, two techniques were proposed to improve the Japanese English speech recognition performance. The first technique effectively introduced graphemes of vowels as an additional context into the state-level clustering. Here, considering Japanese learning process of English, only the phoneme-to-grapheme correspondence of vowels was used. It was found that many decision trees exploited questions on the graphemes and evaluation experiments showed the performance improvement. The second technique coupled a JE model and a JJ model to constitute a parallel model. The mapping between the two and their branching probabilities were estimated so as to maximize the likelihood scores. Recognition experiments showed that the improvement was observed especially for speakers with lower pronunciation proficiency. Although this paper focused upon non-native speech recognition, the first technique of using additional contexts of the central phoneme can be widely applied to other tasks of large vocabulary continuous speech recognition.

6. References

[1] D. Compernolle, "Recognition speech of goats, wolves, sheep, and.... non-natives," Speech Communication, 35, pp.71–79 (2001)

[2] X. He, *et al.,* "Fast model adaptation and complexity selection for non-native English speakers," Proc. ICASSP'2002, pp.577–580 (2002)

[3] C. Huang, *et al.,* "Accent modeling based on pronunciation dictionary adaptation for large vocabulary Mandarin speech recognition," Proc. ICSLP'2000, pp.838–841 (2000)

[4] N. Binder, *et al.,* "Recognition of non-native speech using dynamic phoneme lattice processing," Proc. spring meeting of ASJ, 3-P-19, pp.203–204 (2002)

[5] N. Minematsu, *et al.,* "Integration of MLLR adaptation with pronunciation proficiency adaptation for non-native speech recognition," Proc. ICSLP'2002, pp.529–532 (2002)

[6] N. Minematsu *et al*., "English speech database read by Japanese learners for CALL system development," Proc. LREC'2002, pp.896–903 (2002)

[7] http://www.lang.astem.or.jp/CSRC

[8] K. Osaki *et al*., "Speech recognition of Japanese English using Japanese specific pronunciation habits," Technical report of IEICE, SP2002-180, pp.7–12 (2003, in Japanese)

[9] S. Witt *et al*., "Language learning based on non-native speech recognition," Proc. EUROSPEECH'1997, pp.633–636 (1997)