INTEGRATION OF MLLR ADAPTATION WITH PRONUNCIATION PROFICIENCY ADAPTATION FOR NON-NATIVE SPEECH RECOGNITION

Nobuaki MINEMATSU† *Gakuto KURATA*† *Keikichi HIROSE*‡

† Graduate School of Information Science and Technology, University of Tokyo
‡ Graduate School of Frontier Sciences, University of Tokyo
7-3-1, Hongo, Bunkyo-ku, Tokyo, 113-0033, JAPAN
{mine, gakuto, hirose}@gavo.t.u-tokyo.ac.jp

ABSTRACT

To recognize non-native speech, larger acoustic/linguistic distortions must be handled adequately in acoustic modeling, language modeling, lexical modeling, and/or decoding strategy. In this paper, a novel method to enhance MLLR adaptation of acoustic models for non-native speech recognition is proposed. In the case of native speech recognition, MLLR speaker adaptation was successfully introduced because it enables efficient adaptation with a small number of adaptation data by using a regression tree of Gaussian mixtures of HMMs. However, as for non-native speech, most of the cases, the regression tree built from the baseline HMMs does not match with pronunciation proficiency of a speaker. This paper provides a solution for this problem, where the speaker's proficiency is automatically estimated and the tree suited for the proficiency is built, which can be viewed as proficiency adaptation. Recognition experiments show that MLLR with the new tree raises the averaged error reduction rate up to about 30 % from the baseline MLLR performance of approximately 20 %.

1. INTRODUCTION

Rapid globalization requires Japanese people to speak English in many situations and rapid advances of computation and speech technologies enable spoken dialogue systems to appear in the real world. This indicates that the systems will have to recognize non-native speech adequately in the near future. As is well-known, non-native speech includes much larger acoustic/linguistic distortions compared to native speech. To achieve high recognition performance with non-native speech, the distortions have to be canceled or normalized in acoustic modeling, language modeling, lexical modeling, and/or decoding strategy in the current paradigm of speech recognition. Many studies were already done to handle the distortions in one or more of the above four modules in a speech recognizer[1, 2, 3]. In this paper, research focus is put on the acoustic modeling, where adaptation techniques were often applied to the models to treat the distortions in previous works.

If we have HMMs trained with native speakers' speech and some speech samples of a particular non-native speaker, the HMMs can be adapted to the speaker by using speaker adaptation techniques such as MAP, MLLR, and so on. If we have a relatively large database of non-native speech and some speech samples of a particular non-native speaker, we can built speaker-independent non-native speakers' HMMs and adapt them to the particular speaker with the adaptation techniques. However, these two strategies have clear drawbacks even if all the non-native speakers have the same mother tongue. The first strategy tries to do the speaker adaptation and the mother tongue adaptation at the same time. Since it is expected that the distortions caused by non-nativeness are as large as or maybe larger than those caused by speaker differences, the first strategy will give us only a little improvement of the performance. Furthermore, in the widely-used procedures of training tied-state triphones, which are the most common form of triphones, we can find a hypothesis that speakers of training data and those of testing data should have the same mother tongue. State-tying is usually done by top-down clustering with a prepared set of questions on right/left phoneme context or the central phoneme. It is easily understood that the resulting clustering made by training speakers of a mother tongue are not adequate for testing speakers of another mother tongue. This is because non-native speakers often do phoneme replacement/deletion/insertion. Without them, phonetic characteristics of the phonemes are easily changed in non-native utterances. For these reasons, the widely-used tied-state triphones trained by native speakers are thought to have a definite limit to be used for improving non-native speech recognition performance.

The second strategy uses the adaptation techniques only to do the speaker adaptation. In this sense, this strategy is better than the first one. However, this strategy also has a weakness when only a small number of adaptation data are available. In native speech recognition, MLLR adaptation is successfully used with a small number of adaptation data because it utilizes a phoneme regression tree to cluster HMM mixtures and can modify the HMM parameters even when phonemes of the HMMs are not seen in the adaptation data. But structure of the regression tree built with the baseline non-native HMMs does not always match with pronunciation proficiency of a particular non-native speaker. In this paper, broader distributions of the non-native HMM parameters caused by speaker differences compared to the native HMMs is also shown. This is why immediate use of MLLR will give only a limited improvement even in the second strategy. To solve this problem, in this work, a regression tree suited for the speaker's proficiency is estimated automatically and used for MLLR adaptation, which is considered as pronunciation proficiency adaptation.

In the rest of this paper, firstly, development of an English speech database read by 200 Japanese students is summarized and then, corpus-based analysis of the Japanese English is described very briefly due to the limit of space. This analysis gives us very interesting findings, which are reported in detail in another paper of this conference[4]. After that, the proposed method is described to enhance MLLR adaptation with experimental results of recognizing English speech spoken by Japanese students.

2. DEVELOPMENT OF AN ENGLISH DATABASE READ BY JAPANESE STUDENTS

Multimedia technologies have brought about many applications in education. In Japan, a big national project of "Advanced Utilization of Multimedia for Education" has started in 2000, under which several research groups are presently aiming at developing CALL (Computer Aided Language Learning) systems. To develop the systems, non-native speech databases are required and the first author of this paper and his co-workers designed and developed a database of English words and sentences read by Japanese students[5]. This database was designed mainly to be used in English pronunciation education and it can be divided into two subsets. The first one is related to segmental aspect of pronunciation and the other is 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 thought that the correct pronunciation was done. Even under this strategy, a great number of pronunciation errors were expected to be included in the database according to pronunciation proficiency of the individual speakers.

The database contains speech samples of 100 male students and 100 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 study, part of the sentence data in the database were used for training HMMs and evaluating the proposed method. This is because a series of experiments were done before the completion of the database development. However, the amount of speech samples used in the experiments was large enough and therefore, the reliability of the obtained results is considered to be quite high.

3. CORPUS-BASED ANALYSIS OF ENGLISH SPOKEN BY JAPANESE STUDENTS

3.1. Analysis of American English and Japanese English through their HMMs

Before describing the proposed adaptation method, it is beneficial to show some results of corpus-based analysis done with the database because the obtained findings are closely related to our proposal. In previous studies of phonetics, many studies can be easily found where rather example-based comparisons were done between native English speech samples and Japanese English ones. In this study, a comparison between the two types of English using the large database was done with speech recognition techniques, namely, HMM-based acoustic modeling. Firstly we built two sets of HMMs; American English (AE) and Japanese English (JE). For the former, 25,652 sentences spoken by 245 male speakers in WSJ database were used and 8,282 sentences of 68 male speakers in the above new database were used for the latter. To enable easy-to-understand visualization of results of the analysis, monophones with a single mixture were adopted here. As is explained later, more complex form of HMMs such as tied-state triphones is not adequate for the proposed method. It should be noted that the obtained findings with the simply-structured HMMs can effectively improve the performance of the widely-used complexly-structured HMMs of tied-state triphones with Gaussian mixtures.

For each set of HMMs of AE and JE, distance between any two states of the HMM set was calculated where Bhattacharrya distance measure was used. With the obtained distance matrix, a regression tree (tree diagram) was built for each HMM set with Ward's method, which is one of hierarchical clustering methods. Leaf nodes of the tree corresponded to states of the HMMs (state-level regression tree). Comparison between these two trees showed us many interesting characteristics of JE. **Figure. 1** shows an example of a subtree in JE. Clearly indicated in the figure, states of phonemes of /r/ and /l/ are found very close to each other and we can definitely say that spectral shapes of /r/ and /l/ in JE are almost the same although they show clear spectral differences in AE. The same findings can be observed in cases of /s/ and /th/, /f/ and /h/, /z/ and /dh/, and so on. The analysis also gave us quantitative distance between the corresponding states of the two phonemes in each case, which is shown in **Table. 2** for AE and JE. Further, frequent insertion of a vowel after a consonant, which is one of the well-known characteristics of JE, can also be seen in the entire tree. Additional findings as well as those mentioned above are described in more detail in another paper of this conference[4].

Figure. 2 shows ratios of averaged variances of cepstrum coefficients in JE to those in AE. Here, the averaged variances were calculated for each state over cepstrum dimensions. The figure shows that the variances in JE are larger than those in AE although

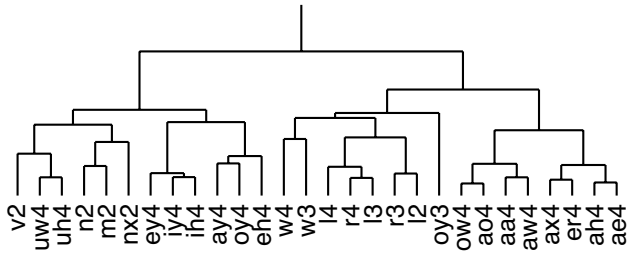

Fig. 1. An example of a subtree drawn with JE HMMs

Table 1. Word and sentence sets contained in the database

set	size
Phonetically-balanced words	300
Minimal pair words	600
MOCHA-TIMIT phonetically-balanced sentences	460
Sentences including phoneme sequences difficult for Japanese to pronounce correctly	32
Sentences designed as test set	100
Words with various accent patters	109
Sentences with various intonation patterns	94
Sentences with various rhythm patterns	121

Table 2. Bhattacharrya distance between corresponding two states of two phoneme paris of AE and JE

	/s2/ & /th2/	/s3/ & /th3/	/s4/ & /th4/
AE	1.00	1.35	0.95
JE	0.30	0.24	0.30
	/f2/ & /h2/	/f3/ & /h3/	/f4/ & /h4/
AE	1.08	1.44	1.14
JE	0.61	0.75	0.87

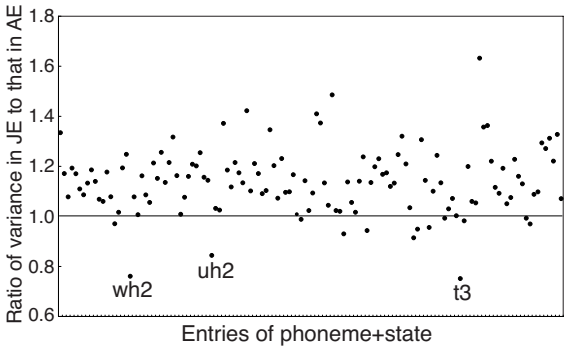

Fig. 2. Ratio of averaged variance in JE to that in AE

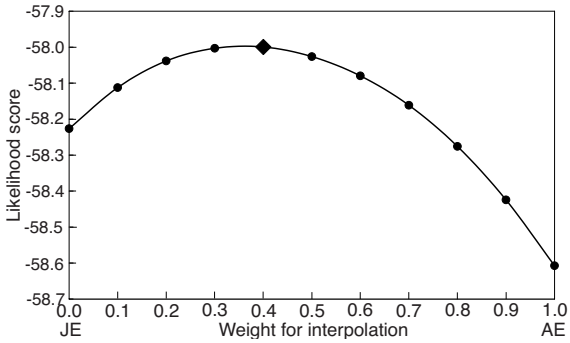

Fig. 3. Likelihood scores calculated by the forced alignment using the interpolated HMMs with various weightings

the JE training data size is much smaller. Considering that the JE database contains carefully read speech only, the above fact implies that the larger broadness of parameter distributions in JE is due to inter-speaker variations of pronunciation proficiency. This finding led us to necessity of introducing pronunciation proficiency adaptation into the conventional adaptation techniques.

3.2. Problems in applying MLLR to non-native speech

MLLR adaptation looks up a mixture-level regression tree to cluster and merge Gaussian mixtures and the same transformation matrix is applied to the merged mixtures. The mixture-level tree includes the state-level regression tree as shown in **Figure. 1** and therefore, structure of the mixture-level tree drawn with AE HMMs is completely different from that of the tree with JE HMMs. It clearly indicates that adaptation of AE HMMs to a Japanese speaker with the AE tree will have only a small improvement or may degrade the recognition performance in the worst case. This is the case even when JE HMMs are adapted to a particular Japanese speaker by using the JE tree. This is because the tree structure of English spoken by individual Japanese speakers are different from each other according to their pronunciation proficiency. In other words, MLLR techniques assume that the tree drawn from the baseline HMMs matches with adaptation speech uttered by a speaker. This assumption is valid when the speaker is native and the HMMs are trained with native speech. In the case of non-native speech, however, it can be easily understood that the assumption is not valid due to large differences in proficiency among speakers.

4. ESTIMATION OF THE REGRESSION TREE SUITED FOR A PARTICULAR SPEAKER

To solve the problem, it is necessary to estimate the pronunciation proficiency of the speaker and build the regression tree suited for the proficiency. In this study, interpolated HMMs (monophones with a single mixture) between AE and JE HMMs with optimal weightings were examined, which is a similar method to MAP adaptation. Since diagonal matrices were used in HMMs, variances were interpolated as well as average vectors in the HMM set. **Figure. 3** shows likelihood scores of adaptation data of a Japanese speaker, which were calculated by the forced alignment using the interpolated HMMs with various weightings. The case of w=0 means that JE HMMs were used and the case of w=1 represents that AE HMMs were used for the alignment. In the figure, 0.4 is the optimal weighting and the regression tree drawn by the HMM

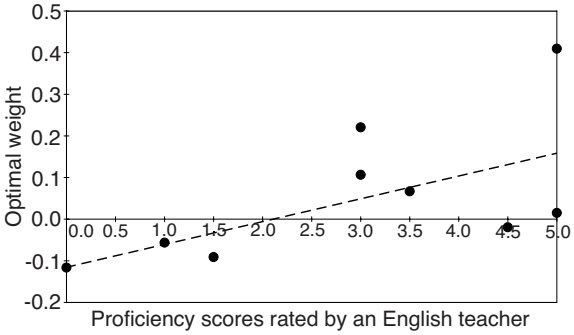

Fig. 4. Relationship between optimal weights of Japanese speakers and their pronunciation proficiency scores rated by a human English teacher

set interpolated at this weighting will be adequate for the pronunciation proficiency of that particular speaker.

Figure. 4 shows relationship between the optimal weightings of Japanese speakers and their pronunciation proficiency scores rated by a human English teacher. In the case that no peak was found between w=0 and w=1 in the forced-alignment likelihood graph, a set of dots at w ($0 \leq w \leq 1$) was approximated by a quadratic curve and the peak was obtained on the curve in the area at w ($0 > w$, $1 < w$). Relatively good correlation can be seen in this figure and it implies that the interpolated HMMs with the optimal weighting characterizes the proficiency of the speaker rather well.

5. MLLR ADAPTATION WITH THE NEW TREE AND ITS EXPERIMENTAL EVALUATION

5.1. Experimental conditions

Recognition experiments were carried out to verify the validity of the proposed method to enhance MLLR adaptation for non-native speech recognition. Speech samples were digitized at 16bit/16kHz sampling and MFCC-based parameters were extracted from the signals with 25 ms frame length and 10 ms frame shift. 12 MFCCs, 12 ΔMFCCs, and Δpower were used to train HMMs of AE and JE. Monophones with a single mixture were adopted to estimate the regression tree suited for the speaker's own pronunciation proficiency and tied-state triphones (#state=2000) with 16 mixtures were adopted to recognize input speech. AE HMMs were trained

with WSJ database and JE HMMs were trained with the database described in section **2**. 30 sentences were used per test speaker to estimate the regression tree and to adapt the baseline HMMs to the speaker. To do the MLLR adaptation, a mixture-level regression tree is required. Here, a state-level tree is used under an assumption that all the mixtures in a state should be merged into a mixture cluster and therefore, the same transformation matrix is applied to mixtures belonging to a state. The corpus-based analysis described in section **3** is possible with tied-state triphones with multiple mixtures. However, the following two reasons led us not to use the complexly-structured HMMs. Triphones are usually built with top-down clustering where one physical state is shared among some logical states. The correspondence between a physical state and logical ones is naturally quite different between AE and JE. This difference in the state correspondence may deduce wrong interpolation between AE and JE HMMs. This is true of HMMs with multiple mixtures. Even if monophones with multiple mixtures are used, the interpolation will not work so well because the mixture index is not an ordinal scale. Then, correct correspondence between a mixture in an AE HMM state and a mixture in a JE HMM state is theoretically impossible. For these two reasons, we adopted the HMMs with the simplest structure, namely, monophones with a single mixture. However, use of diagonal covariance matrices was not required for the analysis. A previous work reported that HMMs with a single mixture of full covariance matrices worked as well as HMMs with 3 to 5 mixtures of diagonal matrices. Use of full covariance matrices is expected to give us more adequate structure of the regression tree.

Number of the test speakers is 10 and that of the test sentences per speaker is 30. Bigram perplexity of the test sentences, which were different from the adaptation sentences, ranged from 300 to 1000 with the vocabulary size of 20K. Unknown word rate was about 8 %. Julius v3.2 was used with forward word bigrams and backward word trigrams. **Table. 3** shows 5 experimental conditions with respect to the baseline HMMs and their adaptation strategies. Depth of the regression tree used in CASEs 3 to 5 was the same and it was tuned in CASE 3, not in CASE 5. Therefore, if it is allowed to tune the depth in CASE 5, the performance of the proposed method may be further improved.

5.2. Results and discussions

Since JE HMMs were trained with speakers of a wide range of pronunciation proficiency, about half of the test speakers were expected to show their peaks between w=0 and w=1. Experiments showed that likelihood peaks were found for five speakers (spk-1 to spk-5) out of ten and the proposed method was applied to these five speakers. **Table. 4** shows word accuracy for each condition and each speaker. It should be noted that, for spk-1, the proposed method was applied to AE HMMs not to JE HMMs because the performance of the JE HMMs is quite low with this speaker. Averaged error reduction rate of the conventional MLLR over the five speakers is calculated to be 21.8 % and that of the proposed method is improved to 30.0 %, which clearly indicates the validity of the proposed method. As for the other 5 speakers (spk-6 to spk-10), parameter extrapolation was preliminary examined only for average vectors of HMMs. In this cases, the optimal weights were less than 0.0, which is shown in **Figure. 4**. Recognition experiments, however, showed no improvement. Pronunciation proficiency adaptation for these speakers was left as a future work.

Table 3. Experimental conditions with respect to the baseline HMMs and their adaptation strategies

condition	baseline HMM	adaptation
CASE-1	JE HMM	no adaptation
CASE-2	AE HMM	no adaptation
CASE-3	JE HMM	conventional MLLR
CASE-4	AE HMM	conventional MLLR
CASE-5	JE HMM	proposed MLLR

Table 4. Word accuracy for each condition and each speaker [%]

speaker	C-1	C-2	C-3	C-4	C-5
spk-1	56.2	72.2	66.7	84.9	85.4
spk-2	92.7	44.8	93.2	62.1	93.2
spk-3	91.3	62.1	92.1	72.6	94.1
spk-4	90.9	53.9	94.5	69.4	94.1
spk-5	88.6	47.5	89.5	53.4	91.8
spk-6	89.0	38.8	90.0	48.4	—
spk-7	68.5	40.2	81.3	51.6	—
spk-8	95.4	39.7	95.4	52.5	—
spk-9	89.0	39.3	88.1	52.5	—
spk-10	70.8	42.5	83.6	60.3	—

6. CONCLUSIONS

This paper proposed a novel method to enhance MLLR-based adaptation techniques for non-native speech recognition by adapting a regression tree to the speakers' own pronunciation proficiency. This study introduced a new axis along which acoustic features tend to be distributed according to one of the speakers' properties, pronunciation proficiency, and the adaptation technique along the axis was proposed. The adaptation along this axis can be applied to other speech recognition modules than acoustic models such as language models, pronunciation lexicon, and decoding strategy. Although high validity of the proposed method was shown, several issues were left unsolved such as speakers who don't have their peaks between w=0 and w=1, and the regression tree generation and the optimal weight estimation with full covariance matrices.

7. REFERENCES

[1] L. M. Tomokiyo, "Lexical and acoustic modeling of non-native speech in LVCSR," Proc. ICSLP'2000, vol.4, pp.346–349 (2000)

[2] C. Huang, et al., "Accent modeling based on pronunciation dictionary adaptation for large vocabulary mandarin speech recognition," Proc. ICSLP'2000, vol.3, pp.818–821 (2000)

[3] I. Amdal, et al., "Joint pronunciation modelling of non-native speakers using data-driven methods," Proc. ICSLP'2000, vol.3, pp.622–625 (2000)

[4] N. Minematsu et al., "Corpus-based analysis of English spoken by Japanese students in view of the entire phonemic system of English," Proc. ICSLP'2002 (2002, accepted)

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