Unsupervised Cross-Adaptation Approach for Speech Recognition by Combined Language Model and Acoustic Model Adaptation

Tetsuo Kosaka*, Taro Miyamoto and* Masaharu Kato*
* Graduate School of Science and Engineering, Yamagata University, Yonezawa, Japan
E-mail: tkosaka@yz.yamagata-u.ac.jp Tel/Fax: +81-238-263369

***Abstract*—The aim of this study is to improve speech recognition with a combination of language model (LM) and the acoustic model (AM) adaptation. The proposed adaptation techniques are based on cross-system adaptation or cross-validation (CV) adaptation. The principle is to use complementary information derived from several systems or data sets. Because language information and acoustic information differ completely, the combined approach is expected to be effective. We evaluate the performance of the proposed methods by conducting speech recognition experiments using the Corpus of Spontaneous Japanese (CSJ). Both cross-system adaptation and CV adaptation give better performance than the conventional adaptation method; the cross-system adaptation method was found to exhibit the best recognition performance.**

I. Introduction

It is well known that automatic speech recognition can be improved by a process known as adaptation. Various adaptation techniques have been investigated for many years. These techniques are referred to as language model (LM) adaptation or acoustic model (AM) adaptation. In unsupervised adaptation, hypotheses determined in a preliminary decoding pass are used as the reference for the adaptation process. Because there is no guarantee that these hypotheses are correct, the adaptation procedure does not always lead to an improvement in overall recognition performance. In general, the adaptation process is performed iteratively. Because the hypotheses which have similar error patterns are taken as the reference in each iteration step, the performance quickly reaches saturation.

In this study, we aim to improve the recognition performance further by using a combination of LM and AM adaptation in an unsupervised manner. To address the issue of saturation, we focus on cross-system adaptation [1][2] and cross-validation adaptation [3]. The principle behind these approaches is to use complementary information derived from several systems or data sets. Although the two approaches are similar in a lot of ways, they have the following differences. In the former approach, the same data are used by different systems, whereas in the latter, different sets of data are used by the same system. We propose an unsupervised adaptation technique that combines the LM and the AM adaptation on the basis of cross-system adaptation or CV adaptation.

Cross-system adaptation makes use of two different systems. One system $S1$ is adapted by using the output of a different system $S2$. Since these systems have different error patterns, they can provide complementary information. For example, Stuker et al. proposed a cross-system adaptation method where different phoneme sets and acoustic front-ends are used to obtain complementary information [2]. In the present work, we combine results from both the LM adaptation and the AM adaptation. The LM considers linguistic information such as the occurrence rates of word sequences, whereas the AM considers acoustic information generated during speech. Because the origins of these data are completely different, the data complement one another. A similar approach, involving both LM and AM adaptations was investigated by Liu et al. [4]. They performed LM and AM adaptations sequentially within a cross-site adaptation framework; however, the recognition outputs of the LM and AM were not crossed, as explained above.

The use of the unsupervised CV adaptation algorithm proposed by Shinozaki et al. [3] reduces the propagation of errors while the model parameters are updated by avoiding data overlap between the decoding step and the model updating step. In the study, the maximum-likelihood linear regression (MLLR) technique [5] was used to adapt AMs on the basis of CV adaptation. In our work, CV adaptation is used not only for AM adaptation but also for LM adaptation. CV-based LM adaptation (LMCV) can be performed in a similar manner to CV-based AM adaptation (AMCV). The LMCV adaptation can be used to obtain complementary information among text data. In addition, complementary information between LM and AM can also be obtained by performing AMCV and LMCV sequentially.

In order to evaluate the performance of the proposed cross-system and CV adaptations, we conduct recognition experiments on a spontaneous speech database, the Corpus of Spontaneous Japanese (CSJ). This corpus is the largest speech corpus in Japan and consists of approximately 7M words with a total speech length of 650 h [6].

II. Unsupervised Model Adaptation

In this section, we explain the LM and AM adaptation techniques used in cross-system and CV adaptations.

A. Language model adaptation

The drawback of LM adaptation is the difficulty in preparing enough data for updating the LM parameters. Class-based

LM adaptation methods have been proposed to cope with the sparseness of the LM data [7][8]. The unsupervised LM adaptation method we employ is based on the statistics of occurrence rates of part-of-speech (POS) classes [9]. This is a kind of class-based LM adaptation. For unsupervised LM adaptation with limited adaptation data, over-training may result from the use of word occurrence statistics. The number of classes can be reduced by using POS instead of words as the units of adaptation. Fig. 1 shows the block diagram of POS-based LM adaptation. The words in a lexicon are categorized according to their POS. The LM adaptation then proceeds as follows:

1) A baseline N-gram is trained by using a large amount of training data. The frequency of occurrence of the POS sequences is also counted from the N-gram count.
2) Entire utterances are recognized by using the baseline N-gram, and these recognition results are then used for adaptation. We are aware of the fact that recognition errors are included in the results.
3) The occurrences of POS sequences are counted from the recognition results.
4) A POS N-gram LM is calculated by using the number of occurrences of POS sequences from both steps 1 and 3. The N-gram probability of the POS sequences is calculated as

$$P(c_i|c_{i-N+1}\cdots c_{i-1}) = \frac{N_0(c_{i-N+1}\cdots c_i) + W\cdot N(c_{i-N+1}\cdots c_i)}{N_0(c_{i-N+1}\cdots c_{i-1}) + W\cdot N(c_{i-N+1}\cdots c_{i-1})}, \quad (1)$$

where N_0 and N are the number of POS sequences counted in the training texts and in the recognition results, respectively. W is a linear interpolation coefficient determined by a preliminary experiment.

5) An adapted LM is created by linearly interpolating the baseline LM obtained in step 1 and the POS LM obtained in step 4, as follows:

$$P'(w_i|w_{i-N+1}\cdots w_{i-1}) = \lambda P(w_i|w_{i-N+1}\cdots w_{i-1}) + (1-\lambda)P(w_i|c_i)P(c_i|c_{i-N+1}\cdots c_{i-1}). \quad (2)$$

The first term on the right-hand side is the probability of the word N-gram and the second term that of the POS N-gram. $P(w_i|c_i)$ is the occurrence probability of a word for each POS class and λ is a linear interpolation coefficient. On the basis of preliminary results, λ was set to 0.7 and the number of POS classes was set to 379, considering both the type and the form of inflection.

B. Acoustic model adaptation

For AM adaptation, we employ the standard MLLR technique [5]. We determine the number of regression classes automatically based on the amount of adaptation data. Our experiments guarantee more than 16 seconds of adaptation data per class. The maximum number of classes is 33, i.e., the number of phoneme classes. In the adaptation process, mixture weights and mean values of HMMs are adapted.

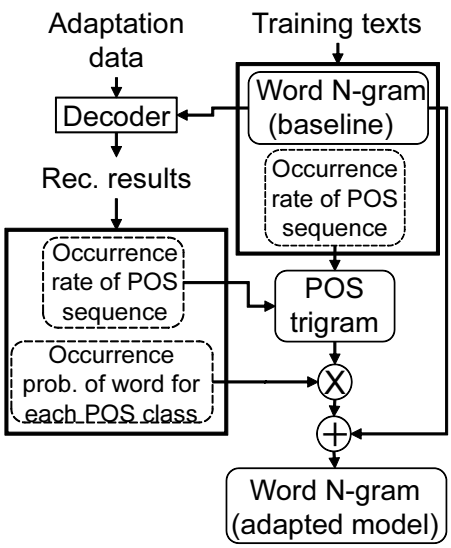

Fig. 1. A block diagram of the LM adaptation procedure.

III. Cross-system adaptation

The proposed cross-system adaptation method uses information from two different sources, namely, the results from LM and AM adaptations. Because of their completely different features, they constitute complementary pieces of information. In the system combination method, we investigated the significance of differences in error patterns between different systems expressed in term of the phoneme mismatch rate (PMR) between the two systems [10]. To calculate the PMR, the two phoneme recognition results are aligned, and the differences in the phoneme level are counted. If the PMR equals zero, the two results are considered to be the same. We found that the PMR correlated strongly with the performance improvement rate. From the results of the preliminary experiment, the PMR between LM and AM adaptations was 6.32 % (The details of this experiment are given in Section VI). Referring to the earlier results, this value signifies that the information obtained is sufficiently complementary. Fig. 2 shows an overview of the proposed cross-system adaptation method. First, recognition results are generated by using the LM and AM (LM_0, AM_0). The LM and AM adaptation procedures are then applied to these initial recognition results, respectively, to yield the adapted models (LM_1, AM_1). Next two new sets of recognition results A and B are obtained by using LM_0 and AM_1, and LM_1 and AM_0, respectively. The former (A) is considered to be the results of the AM adaptation; the latter (B), those of the LM adaptation. The next step is an essential part of cross-system adaptation. Result A (AM adaptation) is used for LM adaptation to yield the adapted LM (LM_2). Similarly, results B (LM adaptation) is used for AM adaptation to give AM_2. This procedure is then repeated.

IV. Cross-validation adaptation

CV adaptation was originally proposed for adapting acoustic models [3]. We applied this method to LM adaptation. Fig. 3 outlines the procedure of the K-fold CV adaptation method. The procedure consists of a decoding step and an updating

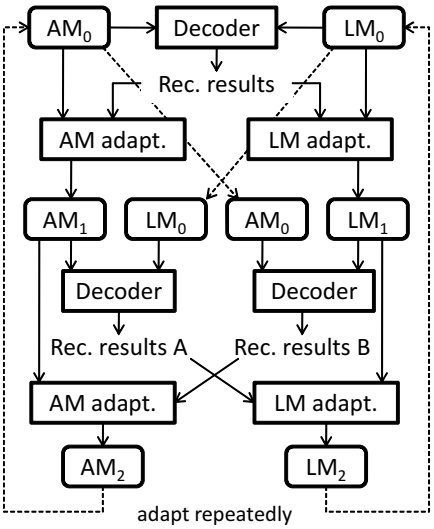

Fig. 2. Cross-system adaptation between LM and AM.

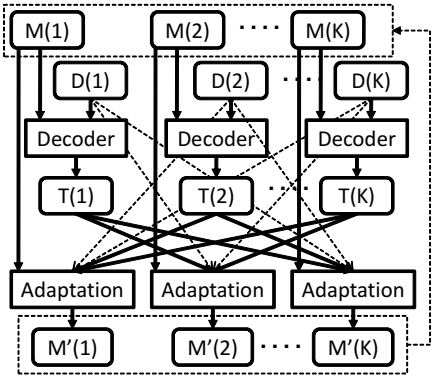

Fig. 3. Unsupervised cross-validation adaptation.

step. First, the input speech data is partitioned into K subsets $(D(1), \cdots, D(K))$. The k-th transcript $(T(k))$ is obtained by decoding the k-th subset by using the k-th AM $(M(k))$. The k-th CV model is updated by excluding the k-th transcript using an adaptation method. The updated model $M'(k)$ are used for recognizing $D(k)$. In the LMCV procedure, the POS-class-based LM adaptation (described in Section II-A) is used. The AMCV procedure uses standard MLLR adaptation. To obtain complementary information between LM and AM, AMCV and LMCV are conducted sequentially.

V. Experimental conditions

In this section, we describe the LVSCR system used for recognition experiments. The speech-analysis module digitizes a speech signal at a sampling frequency of 16 kHz with a quantization size of 16 bits. The length of the analysis frame is 25 ms and the frame period is set to 8 ms. A 13-dimensional feature (12-dimensional MFCC and log power) is derived from the digitized samples for each frame. Further, delta and delta-delta features are calculated from the MFCC feature and the log power, respectively. Then, the total number of dimensions becomes 39. The 39-dimensional parameters are normalized by using the cepstral mean normalization (CMN) method. A two-pass search decoder using a bigram and a trigram is used for recognition. Decoding is performed using a one-pass algorithm in which a frame-synchronous beam search and a tree-structured lexicon are applied in the first pass. The bigram and trigram models are trained from text data containing 6.68M words in the CSJ [6]. Trained language models comprise 47,099 word-pronunciation entries.

The AM is trained using lecture speech data, and the total speech length is 275 h. A set of shared-state triphones is used as the AM. Each triphone is modeled by block-diagonal HMM in which the correlations between static, delta or delta-delta coefficients are assumed to be zero. The total number of states is 3000, and the number of mixture components is 32. In CV adaptation procedures, the number of CV folds is set to 30 for both LMCV and AMCV adaptations. All adaptation procedures are conducted iteratively until there is no further improvement, that is, until there is saturation.

We use the "testset1" evaluation set which consists of academic presentations given by 10 male speakers. This is one of the standard test sets available in the CSJ corpus. Experimental results for each research group can be compared by using this test set. The total speech length is 1.7 h. Each academic presentation is given by one speaker. Since an adaptation procedure is conducted on each presentation, the adaptation can be considered to be done in speaker-adaptation mode. The average length of each presentation is 10.2 min.

VI. Results

For cross-system adaptation, it is important for the error patterns to differ between two systems. Therefore, we investigate the difference in the recognition results of LM and AM adaptations. The baseline recognition performance, in terms of the word error rate (WER) for the speaker-independent model, was 19.62%. The WER after one pass of LM or AM adaptation was 18.52% and 17.62%, respectively. AM adaptation is slightly more effective than LM adaptation. In order to investigate the difference in error patterns between the adaptations, we calculated the PMR, as described in Section III. The PMR between the LM and AM adaptations was 6.32%, which indicates significantly different error patterns between the two methods and the possibility of obtaining complementary information by using cross-system adaptation.

In order to demonstrate the effectiveness of CV adaptation, the following seven methods are compared:

SI: A baseline result is obtained by using the speaker-independent (SI) model. The SI model is also used as an initial model for adaptation.

AM: Only the acoustic model is adapted.

AMCV: Only the acoustic model is adapted by using the CV technique.

LM: Only the language model is adapted.

LMCV: Only the language model is adapted by using the CV technique.

AM-LM: AM and LM adaptations are conducted sequentially.

AMCV-LMCV: AMCV and LMCV adaptations are conducted sequentially.

TABLE II
TREND OF WERS FOR EACH ITERATION. AM_x OR LM_x REFER TO THE X-TH ITERATION OF AM OR LM ADAPTATION, RESPECTIVELY.

System	SI	AM_1	LM_1	AM_2	LM_2	AM_3
AM-LM	19.62	17.62	16.75	16.68	16.63	**16.60**
AMCV-LMCV	19.62	16.93	16.55	**16.54**	**16.54**	16.56

TABLE I
WERS FOR VARIOUS ADAPTATION TECHNIQUES APPLIED TO TESTSET 1.

Method	WER (%)	Total number of iterations
SI	19.62	-
AM	17.50	5
AMCV	16.93	1
LM	18.47	3
LMCV	18.43	1
AM-LM	16.60	AM:3, LM:2
AMCV-LMCV	16.54	AM:2, LM:1

Adaptation procedures were conducted iteratively until the improvement in performance reached saturation. Table I shows the results of various adaptation methods. The CV approach is effective for both AM and LM adaptations. The WER is 17.50% for AM, 16.93% for AMCV, 18.47% for LM, and 18.43% for LMCV. The total numbers of iterations suggest that the performance reaches its maximum level faster with CV adaptation than with the conventional adaptation. The sequential adaptation techniques (AM-LM and AMCV-LMCV) give a better performance than the simple AM or LM approach. Complementary information between AM and LM is thought to be obtained in the sequential approach. AMCV-LMCV gives better performance (16.54%) than AM-LM (16.60%). Table II shows the evolution of the WER with the number of iterations. In the AM-LM method, three iterations for AM adaptation and two iterations for LM adaptation are needed to reach performance saturation. In contrast, AMCV-LMCV needs fewer iterations than AM-LM. We conclude that the CV approach is effective for both AM and LM adaptations.

The cross-system adaptation between LM and AM adaptations was performed to provide a comparison with the results of CV-based adaptation. The performance saturated after three iterations with a WER of 16.50%. Since LM and AM adaptations were performed in parallel in the cross-system adaptation, the total number of adaptation procedures was six. Compared with AMCV-LMCV, cross-system adaptation gives a better performance.

The experimental results are summarized in Fig. 4. The combination of LM and AM adaptations is more effective than the use of only LM or AM adaptation. Cross-system adaptation yields the best results. However, compared with the AM-LM method, its improvement is limited. In our experiments, the language model weight and insertion penalty are fixed. In the cross adaptation approach, optimization of those parameters would be necessary to improve the performance further.

VII. CONCLUSIONS

We proposed an unsupervised adaptation technique that combines LM and AM adaptations on the basis of cross-system adaptation or CV adaptation. In essence, these approaches use complementary information derived from several systems or data sets. The proposed methods were applied to the CSJ. The results indicate that cross-system adaptation yields the best recognition performance.

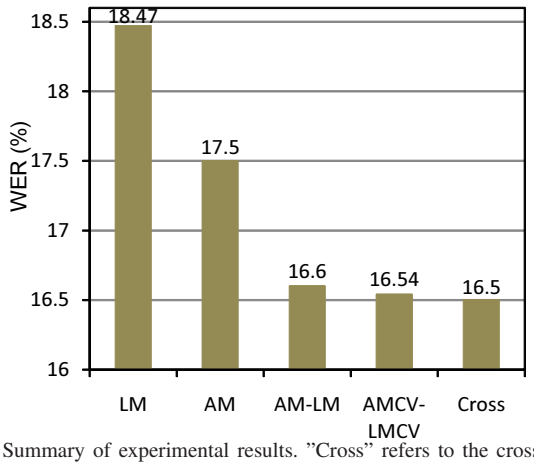

Fig. 4. Summary of experimental results. "Cross" refers to the cross-system adaptation method.

In order to improve the recognition performance even further, we plan to combine the cross-system adaptation approach and the CV adaptation approach in a future study.

ACKNOWLEDGMENT

This work was supported by KAKENHI (22500144).

REFERENCES

[1] H.Soltau, B.Kingsbury, L.Mangu, D.Povey, G.Saon, and G.Zweig, "The IBM conversational telephony system for rich transcription," in *Proc. of ICASSP2005*, 2005, pp. 205–208.

[2] S.Stuker, C.Fugen, S.Burger, and M.Wolfel, "Cross-system adaptation andcombination for continuous speech recognition: The influence of phoneme set and acoustic front-end," in *Proc. of INTERSPEECH2006*, 2006, pp. 521–524.

[3] T.Shinozaki, Y.Kubota, and S.Furui, "Unsupervised cross-validation adaptation algorithms for improved adaptation performance," in *Proc. of ICASSP2009*, 2009, pp. 4377–4380.

[4] X.Liu, M.J.F.Gales, and P.C.Woodland, "Lanugage model cross adaptation for LVCSR system combination," in *Proc. of INTERSPEECH2010*, 2010, pp. 342–345.

[5] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249–264, 1996.

[6] S.Furui, M.Nakamura, T.Ichiba, and K.Iwano, "Analysis and recognition of spontaneous speech using corpus of spontaneous Japanese," *Speech Communication*, vol. 47, pp. 208 – 219, 2005.

[7] G.Moore and S.Young, "Class-based language model adaptation using mixtures of word-class weights," in *Proc. of ICSLP2000*, 2000, pp. 512–515.

[8] H.Yamamoto and Y.Sagisaka, "A language model adaptation using multiple varied corpora," in *Proc. of ASRU2001*, 2001, pp. 389–392.

[9] R.Tsutsumi, M.Katoh, T.Kosaka, and M.Kohda, "Lecture speech recognition using pronunciation variant modeling," *IEICE Trans. on Information and Systems*, vol. J89-D, no. 2, pp. 305–313, 2006.

[10] T.Kosaka, K.Goto, T.Ito, and M.Kato, "Lecture speech recognition by combining word graphs of various acoustic models," in *Proc. of INTERSPEECH2010*, 2010, pp. 2978–2981.