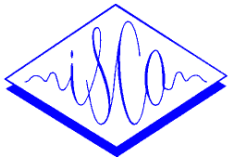

ISCA & IEEE Workshop on Spontaneous
Speech Processing and Recognition
Tokyo Institute of Technology, Tokyo, Japan
April 13-16, 2003

UNSUPERVISED LANGUAGE MODEL ADAPTATION USING WORD CLASSES FOR SPONTANEOUS SPEECH RECOGNITION

T. Yokoyama, T. Shinozaki, K. Iwano and S. Furui

Tokyo Institute of Technology
Department of Computer Science
2-12-1 Ookayama, Meguroku, Tokyo, 152-8552 Japan
{tadasuke, staka, iwano, furui}@furui.cs.titech.ac.jp

ABSTRACT

This paper proposes an unsupervised, batch-type, class-based language model adaptation method for spontaneous speech recognition. The word classes are automatically determined by maximizing the average mutual information between the classes using a training set. A class-based language model is built based on recognition hypotheses obtained using a general word-based language model, and linearly interpolated with that general language model. All the input utterances are re-recognized using the adapted language model. It was confirmed that the proposed method is effective in improving the recognition accuracy in spontaneous presentation recognition. The proposed method was combined with acoustic model adaptation, and it was found that the effects of language model adaptation and acoustic model adaptation are additive. The optimum number of classes is 100 irrespective of whether the acoustic model adaptation is combined or not, and in this condition the language model adaptation yields approximately 2% absolute value improvement in the word accuracy.

1. INTRODUCTION

Although speech of reading written texts can be recognized with a high recognition accuracy using state-of-the art speech recognition technology, the recognition accuracy of freely spoken spontaneous speech is still poor. For example, currently, mean recognition accuracy of spontaneous presentation recognition using "Corpus of Spontaneous Japanese (CSJ)[1]" can only reach roughly 70%[2]. The principal cause of the problem is a mismatch between trained acoustic/language models and input speech due to a limited amount of training data in comparison with a vast variation of spontaneous speech. Spontaneous presentation utterances are both acoustically and linguistically variable according to speakers and topics. To cope with this problem, automatic adaptation is essential for both acoustic and language models.

Adaptation techniques can be classified into supervised and unsupervised methods. Since unsupervised methods can use recognition data itself for adaptation, they are more flexible than supervised methods. However, unsupervised methods are usually more difficult to develop than supervised methods, especially for spontaneous speech having a relatively high recognition error rate. Assuming that the presentation recognition is performed off-line, we have recently investigated a batch-type unsupervised acoustic model adaptation for this task, in which first all the presentation utterances are recognized using a speaker-independent model, then that model is adapted by using the recognition results, and used again to re-recognize the utterances. This process is repeated until recognition results converge. We have achieved roughly 5% improvement of word accuracy by this method[2]. Although various useful unsupervised acoustic model adaptation methods have been proposed, unsupervised language model adaptation has not proved to be highly successful in improving recognition accuracy[3]. This is because the language model space is usually very sparse and therefore it is very difficult to obtain reliable information from recognition results with a relatively high recognition error rate.

In order to cope with the sparseness of the language model space, class-based language model adaptation methods have been proposed[4, 5]. However, they have never been used for unsupervised adaptation in spontaneous speech recognition.

This paper proposes a class-based batch-style unsupervised language model adaptation for spontaneous speech recognition and presents its effectiveness in spontaneous presentation recognition using the CSJ. This paper also investigates the effectiveness of combining this language model adaptation with unsupervised acoustic model adaptation.

The paper is organized as follows. Section 2 describes the unsupervised language model adaptation method. Section 3 explains a method of combining the proposed method with unsupervised acoustic model adaptation. In Section 4 the experimental conditions are described. In Section 5 we

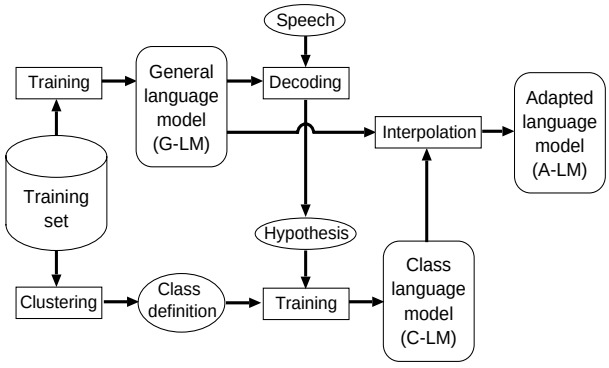

Fig. 1. An overview of the unsupervised class-based language model adaptation method.

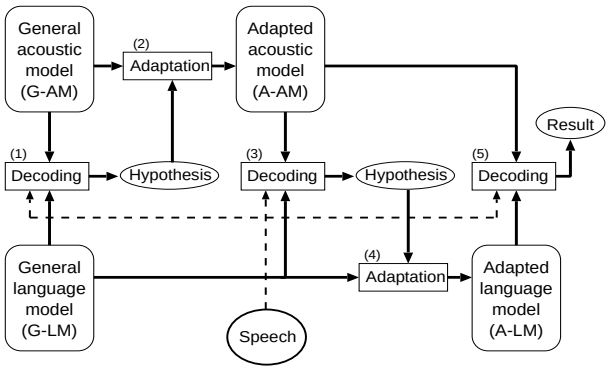

Fig. 2. An overview of the recognition flow using both the proposed language model adaptation and acoustic model adaptation.

describe and discuss recognition experiments performed using our adaptation method. Finally, in Section 6, main conclusions are presented.

2. UNSUPERVISED LANGUAGE MODEL ADAPTATION

Figure 1 shows the overview of the proposed class-based unsupervised language model adaptation method.

Using many transcriptions in the training data set, a general language model (G-LM) consisting of word-based n-grams is built. Word classes approximately maximizing the average mutual information between classes are also made by applying a clustering algorithm, the "incremental greedy merging algorithm[6]", to the training data set. Our proposed adaptation method consists of the following three steps.

(1) Whole utterances of a presentation are recognized using the G-LM.

(2) A class-based language model (C-LM) is trained using the recognition results and the word-class information.

(3) An adapted language model (A-LM) is obtained by linearly interpolating the G-LM and the C-LM. An adapted language model for a word w with word history h, $P_a(w|h)$, is calculated as follows:

$$P_a(w|h) = (1-\lambda)P_g(w|h) + \lambda P_c(w|h) \qquad (1)$$

where $P_g(w|h)$ and $P_c(w|h)$ represent language models, G-LM and C-LM, respectively, and λ indicates a linear interpolation coefficient.

3. COMBINATION WITH ACOUSTIC MODEL ADAPTATION

The proposed language model adaptation method was combined with an unsupervised acoustic model adaptation method. Figure 2 shows the overview of the recognition flow including both the language and acoustic model adaptation processes. The overall steps are as follows:

(1) Recognizing all utterances using the general language model (G-LM) and a general speaker independent acoustic model (G-AM),

(2) Building a speaker adapted acoustic model (A-AM) by adapting the G-AM by the MLLR technique using the recognition results obtained in (1),

(3) Obtaining improved recognition hypothesis by re-recognizing the utterances using G-LM and A-AM,

(4) Building an adapted language model (A-LM) by the language model adaptation method described in Section 2 using the recognition hypotheses obtained in (3),

(5) Re-recognizing the utterances using A-LM and A-AM.

4. EXPERIMENTAL CONDITIONS

4.1. Training and test sets

The training data set consists of 1,289 presentations in the CSJ with approximately 3M words. The test set consists of 10 presentations in the CSJ, having no overlap with the training set. All the 10 presentations are given by male speakers. Each presentation's ID in the CSJ, conference

Table 1. List of the test set data.

ID	Conference name	Number of words	Word accuracy (%)
A01M0007	Acoust. Soc. Jap.	4,610	73.19
A01M0035	Acoust. Soc. Jap.	6,151	59.03
A01M0074	Acoust. Soc. Jap.	2,479	75.67
A02M0076	Soc. Jap. Linguistic	5,045	70.11
A02M0098	Soc. Jap. Linguistic	3,817	64.46
A02M0117	Soc. Jap. Linguistic	9,887	67.03
A03M0100	Assoc. Natural Lang. Proc.	2,735	66.27
A03M0111	Assoc. Natural Lang. Proc.	3,376	57.20
A05M0031	Phonetics Soc. Jap.	5,288	66.40
A06M0134	Assoc. Socioling. Science	4,585	58.18

name, number of words, and baseline word accuracy when using G-LM are shown in Table 1. The total number of words in the test set is approximately 48k and the average word accuracy is 65.6%.

4.2. Language model

The general language model (G-LM) consists of word-based bi-grams and reverse tri-grams. Bi-grams and reverse tri-grams are used for the first path and the second path of decoding, respectively. Unseen n-grams are estimated using the Katz's back-off smoothing technique[7]. The approximately 35k words that appear twice or more in the training data set are selected as vocabulary words.

The class-based language model (C-LM) consists of class-based bi-grams and reverse bi-grams. Probabilities of class transition and word occurrence in each class are estimated using the recognition results. Therefore, the vocabulary covers only the words appearing in the recognition hypotheses.

The adapted language model (A-LM) consists of word-based bi-grams and reverse tri-grams. The reverse tri-gram is obtained by interpolating between the reverse tri-gram of G-LM and the reverse bi-gram of C-LM.

All language models are made using the SRI Language Modeling Toolkit[8].

4.3. Acoustic model

The acoustic features are 25 dimensional vectors consisting of 12MFCC, their delta and delta log energy. The CMS (cepstral mean subtraction) is applied to each utterance. A general speaker independent acoustic model (G-AM) is made using 455 presentations, having a length of approximately 94 hours, taken from the training data set. The model is a tied-state tri-phone HMM having 3k states and 16 Gaussian mixtures in each state. HTK v2.2 is used for the acoustic modeling.

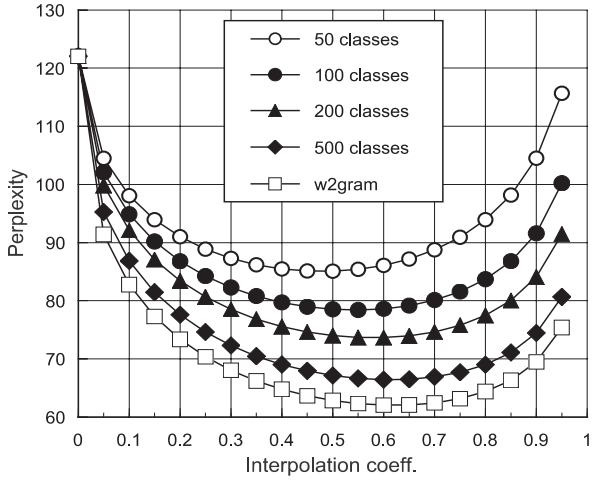

Fig. 3. Test-set word perplexity as a function of the interpolation coefficient λ.

5. EXPERIMENTAL RESULTS

The Julius v3.2 decoder[9] was used for speech recognition. Insertion penalty and language model weight were optimized for the recognition condition using the G-LM.

Figure 3 shows the test-set word perplexity after language model adaptation as a function of the interpolation coefficient λ at various conditions of the number of word classes; 50, 100, 200, and 500. Each curve indicates the result averaged over all the presentations. In the "w2gram" condition, the C-LM is equivalent to the word-based bi-gram and reverse bi-gram modeling with no word classes. The perplexity decreases with adaptation in all the word-class conditions. When a C-LM with 500 classes is used, the perplexity becomes almost a half of that before adaptation at the best condition of the interpolation coefficient λ.

Figure 4 shows word accuracy averaged over all the pre-

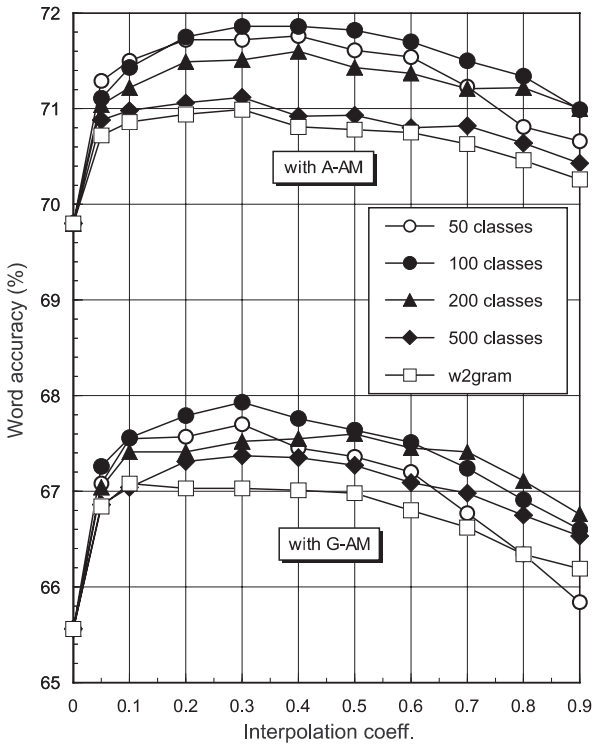

Fig. 4. Word accuracy as a function of the interpolation coefficient λ.

sentations as a function of the interpolation coefficient λ, in various class conditions. "with A-AM" and "with G-AM" indicates the results when the adapted acoustic model (A-AM) and the general, speaker-independent acoustic model (G-AM) are used, respectively. The results without language model adaptation are shown in the leftmost position, where the interpolation coefficient λ is set at 0. It is shown that the acoustic model adaptation yields 4.3% improvement of the mean word accuracy.

In all conditions, the recognition accuracy is improved by the language model adaptation. Although the word-based adaptation method "w2gram" yields better test-set perplexity than the class-based adaptation methods, its recognition performance is much worse than the class-based method.

In both "with A-AM" and "with G-AM" conditions, the best improvement of recognition accuracy, approximately 2% in the absolute value, is obtained by the proposed language model adaptation method. The best results are obtained by setting λ at 0.3 and the number of classes at 100, irrespective of whether acoustic model adaptation is combined or not. This means that the effects of acoustic and language model adaptation are additive. By combining both adaptation processes, the word accuracy is improved from 65.6% to 71.8%.

6. CONCLUSION

This paper has proposed a batch-type unsupervised language model adaptation method using a class-based language model built based on recognition hypotheses obtained using a general word-based language model. The word classes are automatically determined by maximizing the average mutual information between the classes using a training set. The class-based model is linearly interpolated with the general language model and used for re-recognizing the speech. This method is effective in improving the word accuracy of spontaneous presentation speech recognition. It has also been confirmed that the effects of acoustic and language model adaptation are additive.

Future research includes automatic optimization of the number of word classes and the interpolation coefficient λ.

7. REFERENCES

[1] K. Maekawa, H. Koiso, S. Furui and H. Isahara "Spontaneous speech corpus of Japanese," *Proc. LREC2000*, Athens, Greece, vol.2, pp.947–952, 2000.

[2] T. Shinozaki, C. Hori and S. Furui, "Towards automatic transcription of spontaneous presentation," *Proc. Eurospeech2001*, Aalborg, Denmark, vol.1, pp.491–494, 2001.

[3] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," *Proc. ICSLP2002*, Denver, pp.1413–1416, 2002.

[4] G. Moore and S. Young, "Class-based language model adaptation using mixtures of word-class weights," *Proc ICSLP2000*, Beijing, China, vol.4, pp.512–515, 2000.

[5] H. Yamamoto and Y. Sagisaka, "A language model adaptation using multiple varied corpora," *Proc. ASRU2001*, Madonna di Campiglio, Trento, Italy, 2001.

[6] P. F. Brown, V. J. Della Pietra, P. V. deSouza, J. C. Lai, R. L. Mercer, "Class-based n-gram models of natural language," *Computational Linguistics*, vol.18, no.4, pp.467–479, 1992.

[7] S. M. Katz, "Estimation of probabilities from sparse data for language model component of a speech recognizer," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol.35, no.3, pp.400–401, 1987.

[8] A. Stolcke, "SRILM - an extensible language modeling toolkit," *Proc. ICSLP2002*, Denver, pp.901–904, 2002. http://www.speech.sri.com/projects/srilm/

[9] A. Lee, T. Kawahara and K. Shikano. " Julius – an open source real-time large vocabulary recognition engine," *Proc. Eurospeech2001*, Aalborg, Denmark, vol.3, pp.1691–1694, 2001.