# UNSUPERVISED CLASS-BASED LANGUAGE MODEL ADAPTATION 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 the general language model. All the input utterances are re-recognized using the adapted language model. The proposed method was applied to the recognition of spontaneous presentations and was found to be effective in improving the recognition accuracy for all the presentations. The best condition was found to be using 100 word classes, and in this condition 2.3% of the absolute value improvement in the word accuracy averaged over all the speakers was achieved.

## 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 low. 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 the limited amount of training data in comparison with the 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. Supposing 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, and then a model adapted using the recognition results is used 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 is organized as follows. Section 2 describes the unsupervised language model adaptation method. In Section 3 experimental conditions are described. In Section 4 we describe and discuss recognition experiments performed using our adaptation method. Finally, in Section 5 the main conclusions are presented.

## 2. UNSUPERVISED LANGUAGE MODEL ADAPTATION

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

0-7803-7663-3/03/$17.00 ©2003 IEEE

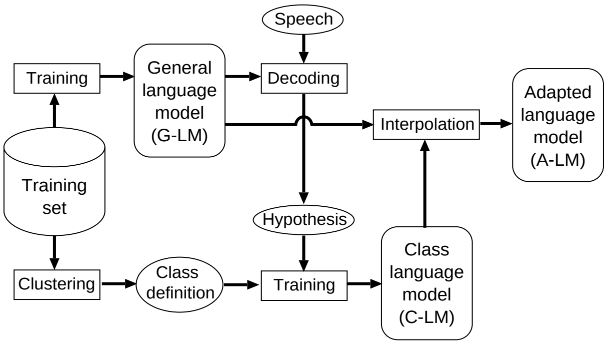

**Fig. 1**. An overview of the unsupervised class-based language model adaptation method.

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 $\lambda$ indicates a linear interpolation coefficient.

## 3. EXPERIMENTAL CONDITIONS

### 3.1. 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 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 66%.

### 3.2. 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].

### 3.3. 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 speaker independent acoustic model is made using 455 presentations, having a length of approximately 94 hours, taken from the training data set. The model is a tied-state triphone HMM having 3k states and 16 Gaussian mixtures in each state. HTK v2.2 is used for the acoustic modeling.

## 4. 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 2 shows the test-set word perplexity as a function of the interpolation coefficient $\lambda$ at various conditions of the number of word classes; 50, 100, 500, and 1000. 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 or more is used, the perplexity becomes almost a half of that before adaptation at the best condition of the interpolation coefficient $\lambda$.

Figure 3 shows the word accuracy averaged over all presentations as a function of the interpolation coefficient $\lambda$, in various class conditions. Results without adaptation are obtained under the condition that the interpolation coefficient $\lambda$ is set at 0. In all the conditions, the recognition accuracy is improved by the adaptation. The best improvement of the mean accuracy, 2.3% in the absolute value, is achieved

**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 |

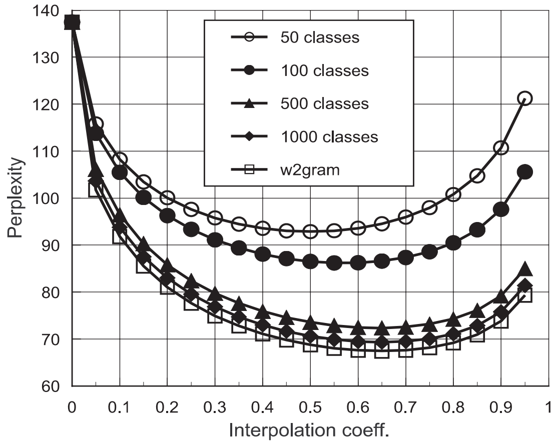

**Fig. 2**. Test-set word perplexity as a function of the interpolation coefficient $\lambda$.

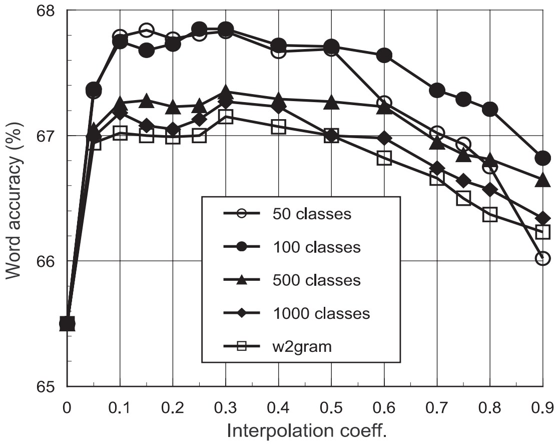

**Fig. 3**. Word accuracy as a function of the interpolation coefficient $\lambda$.

when $\lambda$ is 0.3 and the number of classes is 100. Although the word-based adaptation method “w2gram” yields better test-set perplexity than the class-based adaptation methods, its recognition performance is the worst.

Figure 4 shows absolute percentage improvement of the word accuracy for each presentation when the C-LM with 100 classes is used. Under the condition that the interpolation coefficient $\lambda$ is less than 0.5, accuracy is improved by the class-based adaptation for all the presentations. The best improvement, approximately 6% in the absolute value, is achieved for the presentation A05M0031 when $\lambda$ is 0.3.

### 4.1. Discussion

Figures 2 and 3 show that there is a discrepancy between the improvement of the test-set perplexity and that of the word accuracy. That is, the adaptation conditions with a large number of classes which yield smaller test-set perplexity do not necessarily achieve better recognition performance. This is probably because the adaptation methods having a large number of classes are over-tuned to the recognition results including errors, and therefore, the contribution to reduce the perplexity of correct hypotheses becomes less significant.

Figure 4 shows the wide variety of improvement of recognition accuracy among presentations. Figure 5 shows the relationship between the improvement of recognition accuracy by the adaptation and the difference of perplexities between recognition hypotheses and correct transcriptions using the general language model, G-LM. Twenty presentations including the test set were used, and the number of classes and the interpolation coefficient $\lambda$ were set at 100 and 0.3, respectively. There exists a clear correlation between the difference of the perplexities and the improvement of word accuracy; its correlation coefficient is 0.76, which is significant at a 0.1% significance level. The strong correlation means that the proposed adaptation method is more effective for presentations having a larger mismatch

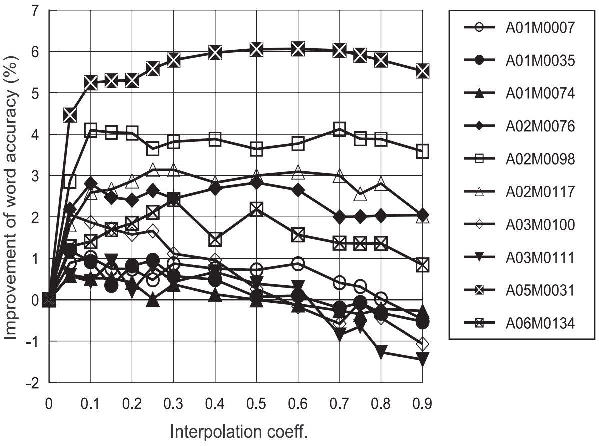

**Fig. 4**. Improvement of the word accuracy as a function of the interpolation coefficient $\lambda$ for each presentation.

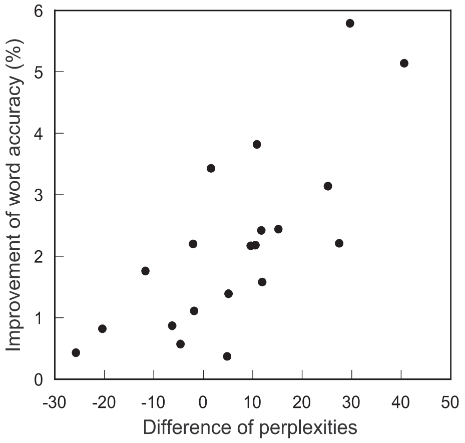

**Fig. 5**. Relationship between the difference of the perplexities between recognition hypothesis and correct transcription calculated using G-LM and the improvement of word accuracy.

with the G-LM.

## 5. 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. Using a class-based language model having 100 classes, 2.3% improvement of word accuracy averaged over speakers in the absolute value has been achieved.

Future research includes automatic optimization of the number of word classes and the interpolation coefficient $\lambda$, and combination with acoustic model adaptation.

## 6. 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.