UNSUPERVISED LANGUAGE MODEL ADAPTATION FOR LECTURE SPEECH RECOGNITION

Hiroaki Nanjo and Tatsuya Kawahara

School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
{nanjo,kawahara}@kuis.kyoto-u.ac.jp

ABSTRACT

This paper addresses speaker adaptation of language model in large vocabulary spontaneous speech recognition. In spontaneous speech, the expression and pronunciation of words vary a lot depending on the speaker and topic. Therefore, we present unsupervised methods of language model adaptation to a specific speaker by (1) making direct use of the initial recognition result for generating an enhanced model, and (2) selecting similar texts, utterance by utterance, based on the model. We also investigate the pronunciation variation modeling and its adaptation in the same framework. It is confirmed that all proposed adaptation methods and their combinations reduced the perplexity and word error rate in transcription of real lectures.

1. INTRODUCTION

Under the Science and Technology Agency Priority Program in Japan (1999-2004) [1], a large scale spontaneous speech corpus has been collected and we have started extensive studies on large vocabulary spontaneous speech recognition. Our main goal is the automatic transcription of live lectures such as oral presentations in conferences.

For spontaneous speech recognition, the speaker dependent models obtain better performance than the speaker independent models. Thus, a variety of model adaptation methods have been studied. As for acoustic model, the speaker adaptation methods have been studied extensively and successful in improving accuracy. We also have investigated unsupervised acoustic model adaptation methods and confirmed their effects [2].

On the other hand, conventional studies on language model adaptation mainly aimed at how to adapt the model to the specific domains or topics [3][4]. As for lecture speech recognition, the adaptation to each speaker is required because the preference of expressions and their pronunciation are quite different among the speakers. Fortunately, lectures have relatively longer speech and their transcriptions, which will make the speaker adaptation possible. In this paper, we present several methods of unsupervised language model adaptation to speaker's characteristics in expression. We also address the pronunciation variation modeling and its adaptation in the framework.

Table 1. Test-set lectures and baseline result

lecture ID	#words (#pauses)	duration (min.)	WER (%)	perplexity
A1	7355 (688)	28	38.5	72.40
A2	6109 (482)	27	31.3	83.72
A3	5269 (426)	23	39.2	58.90
A4	7747 (739)	42	33.4	67.78
A5	3561 (227)	15	29.7	68.94
B1	5413 (798)	30	22.5	55.33
B2	2843 (253)	12	24.4	61.21
B3	11781 (1334)	57	35.4	78.13
B4	3179 (350)	15	29.7	52.83
B5	3227 (238)	14	36.7	63.01
total	56484 (5535)	263	33.1	68.18

A1(A01M0035), A2(A05M0031), A3(A06M0134), A4(KK99DEC005), A5(YG99MAY005), B1(A01M0007), B2(A01M0074), B3(A02M0117), B4(A03M0100), B5(YG99JUN001)

2. TASK AND BASELINE SYSTEM

2.1. Corpus and Test-set

The Corpus of Spontaneous Japanese (CSJ) developed by the project consists of a variety of oral presentations at technical conferences and informal monologue talks on given topics. The test-set for evaluation consists of ten lecture presentations specified in Table 1, which have been commonly used [5]. Many of them are invited lectures at technical meetings, thus relatively longer than simple paper presentations. They were given by experienced lecturers who did not prepare drafts.

2.2. Language Model

For language model training, all transcribed data available (in Feb. 2002) are used. There are 1099 presentations and

the text size in total is 3.15M words (=Japanese morphemes) including pause punctuations.

We trained backoff word trigram model as a baseline language model using CMU-Cambridge SLM toolkit ver.2. The vocabulary is defined by 16029 words which are found more than 3 times in the training data. Test-set OOV rate is 2.10% with this vocabulary. In the baseline language model training process, the pronunciation information is ignored. Thus, the words that have same spelling are regarded as one entry. The variation of pronunciation is simply handled by adding multiple baseforms in the lexicon.

2.3. Acoustic Model

As for acoustic model training, only male speakers at technical conferences are used in this work. We use 394 presentations that amount to 60 hour speech. We trained context-dependent triphone models. Decision-tree clustering was performed to set up 3000 shared states. We also adopt PTM (phonetic tied-mixture) modeling [6], where triphone states of the same phone share Gaussians but have different weights. Here, 129 codebooks of 192 mixture components are used. The specification of acoustic model (speech analysis, phone set, etc.) is same as the baseline model of [2].

2.4. Specification of Baseline System

We use the large vocabulary speech recognition decoder Julius rev.3.2 that was developed at our laboratory [7].

The average word error rate (WER) with the baseline system is 33.1% and the average of test-set perplexity is 68.18. WER and perplexity for each speaker are listed in Table 1. For the test-set perplexity computation, OOV words are excluded but pause punctuations are included. For the calculation of WER, OOV words are included but pauses are excluded. The total number of words except pauses is 50949.

3. LANGUAGE MODEL ADAPTATION USING INITIAL RECOGNITION RESULT

At first, we introduce a simple adaptation method using the initial recognition result, because lectures have relatively longer speech and their transcriptions, which contains speaker's topics and characteristics in expression. The adaptation process is shown in part of Fig. 1.

Backoff word trigram model ($LM1$) is trained with the initial recognition result which contains not a few errors. Bigram and trigram entries which are found only once are discarded.

Then, linear interpolation with the baseline model ($LM0$) is performed for adaptation according to the equation (1),

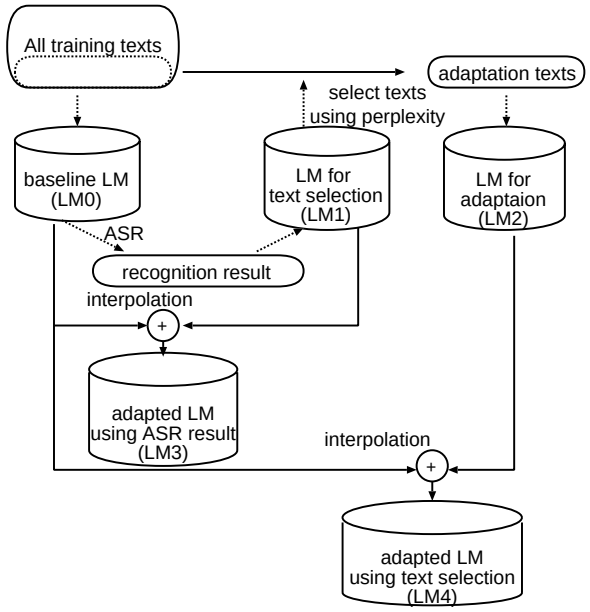

Fig. 1. Flowchart of language model adaptation

$$P_{adapt}(w) = \lambda \cdot P_{rec}(w) + (1-\lambda) \cdot P_{base}(w) \quad (1)$$

where $P_{base}(w)$ is a probability for word sequence w of the baseline language model and $P_{rec}(w)$ is a probability of the language model generated by the initial recognition result. $P_{adapt}(w)$ is a probability of the adapted language model. The interpolation coefficient λ is estimated using EM algorithm denoted in equation (2),

$$\hat{\lambda} = \sum_{i=1}^{N} \frac{\lambda \cdot P_{rec}(w_i)}{\lambda \cdot P_{rec}(w_i) + (1-\lambda) \cdot P_{base}(w_i)} \quad (2)$$

where w_i is the i-th word of the correct transcription of the corresponding test-set lecture, which is actually unavailable. If we substitute the initial recognition result for the correct transcription, the value of λ becomes large and we cannot reduce perplexity of the correct transcription.

So, we introduced a development-set for estimation of λ. We select a half of the test-set lectures as the development-set which is used so as to estimate interpolation coefficient λ for the other half of the test-set lectures (evaluation-set). In this paper, we divided the test-set lectures into two groups; A1 to A5 and B1 to B5 as listed in Table 1. The λ estimation is performed according to the equation (2) for each lecture of the development-set until convergence, and then their average is set to the final λ value.

The development-set and evaluation-set are swapped and the same process is performed. For this procedure (interpolation of the baseline model $LM0$ and model $LM1$

Table 2. Result of adaptation using initial recognition result

	WER(%)	perplexity
baseline ($LM0$)	33.1	68.18
adapted using ASR result ($LM3$)	31.0	52.37

trained with the ASR result), we use complementary back-off algorithm [8], which works well in the case there is a quite difference of the N-gram entries between the models.

The result of the adaptation method is shown in Table 2. The simple unsupervised method reduced WER by 2.1% absolute.

4. LANGUAGE MODEL ADAPTATION USING TEXT SELECTION

Next, we present another method to enhance the language model by weighting similar texts to a test-set lecture based on the initial recognition.

There is a method to select similar texts based on a-priori knowledge such as use of preprints of the corresponding lecture and transcriptions of presentations by the same speaker. We once tried incorporation of preprint texts for adaptation and got accuracy improvement by 0.5% (for A1) and 3.0% (for B1) absolute. However, we cannot assume that preprints are always available.

In this paper, we explore a method without a-priori knowledge. Conventional studies made use of texts selected based on topic-dependent word counts such as $tf \cdot idf$ [4][9] or based on the perplexity or coverage. We define the perplexity as a similarity measure in this work. Text selection is performed for every utterance unit which is defined by pause segments. We also tried selecting lecture by lecture [9], but could not achieve perplexity reduction.

The adaptation process is also shown in Fig. 1. At first, language model $LM1$ using the initial recognition result, which is the same as the one described in previous section except no cutoff of bigram and trigram entries is done, is set up for text selection. Then, perplexity of each training text is computed using the model $LM1$. In this computation, OOV words are included. Texts that have lower perplexity than a threshold th are selected and language model $LM2$ is generated. Then, linear interpolation with the baseline model ($LM0$) is performed to generate an adapted model $LM4$. Again, we make use of the development-set for determining the threshold th and estimating the interpolation coefficient. Optimal values of the threshold th and coefficient λ are estimated for each lecture, and averaged for the whole development-set.

Using A1 to A5 as a development-set and B1 to B5 as an evaluation-set, the perplexity threshold th and the interpolation coefficient λ were estimated as 110 and 0.472, respectively. When we swapped the development-set and the evaluation-set, the estimated values were 92 and 0.479. With the adapted model ($LM4$), the overall test-set perplexity was reduced from 68.18 to 65.60. We also reduced WER by 0.5% absolute as shown in Table 3.

Table 3. Result of adaptation using text selection

	WER(%)	perplexity
baseline ($LM0$)	33.1	68.18
adapted using text selection ($LM4$)	32.6	65.60

Table 4. Results of adaptation including pronunciation

	WER(%)	perplexity
baseline ($LM0$)	33.1	68.18
pronunciation ($LMP0$)	30.8	70.76
pronun + ASR result ($LMP3$)	28.8	53.69
pronun + text selection ($LMP4$)	30.4	67.77
adapted using all methods	28.7	53.20

We investigated texts selected for adaptation and found that they contained not a few carrier and filler phrases which often appear at the beginning and end of sentences in Japanese, such as *"desu ne"*, *"de ano:"* and *"e: ma:"*. They are considered as representing speaker's characteristics in expression. This suggests that our proposed adaptation method properly extracts such features.

5. ADAPTATION INCLUDING PRONUNCIATION

5.1. Pronunciation Modeling

Then, we also incorporate the factor of the pronunciation variation in the modeling and adaptation. In spontaneous Japanese, one of the main causes of pronunciation variation is a word context, especially adjacent words. The baseline system coped with the pronunciation variation problem by simply adding variant baseform entries to the lexicon. The method often encounters harmful side effects, specifically, false matching increases especially with short functional words which tend to have more pronunciation variations. Actually, for a certain word, adding its all pronunciations extracted from the training text corpus (including some extraction errors) to the lexicon resulted in the increase of WER by 6.6% absolute.

Thus, we introduce an approach in which the pronunciation variation modeling is combined with the language model. Usually, this approach is implemented by unigram modeling of pronunciation variation (probability of baseform entries) for each lexical entry [10][11]. Here we adopt trigram modeling which considers the contextual effect. We distinguish the entries whose pronunciation forms are dif-

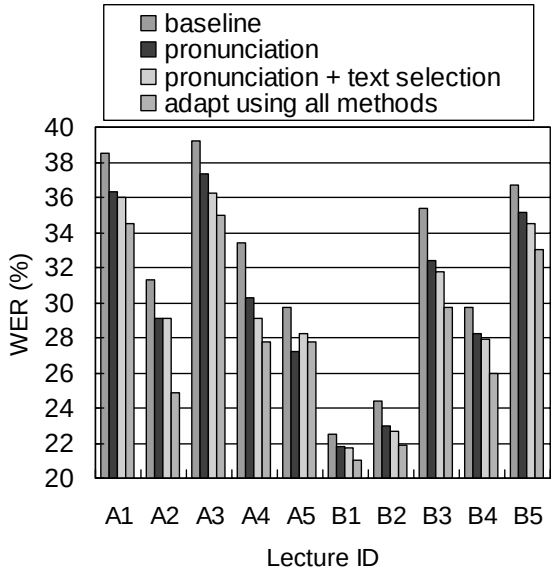

Fig. 2. WER of each test-set lecture with proposed methods

ferent even if they have a same spelling and part of speech. Using these entries, trigram language model is trained with the same corpus. The result is shown in Table 4 (row of "pronunciation"). WER was reduced significantly (2.3% absolute). This demonstrates the pronunciation variation much depends on the word context.

5.2. Combination of All Adaptation Methods

We also investigate the combination of the adaptation methods in order to adapt each speaker's tendency of talks and pronunciation variation. The result is shown in the lower part of Table 4. In these combinations, we obtained much the same error reduction as in the previous experiments which do not consider pronunciation variation. The fact shows that the language model is adapted in both variation of pronunciation and expressions.

Lastly, we combine all proposed methods and evaluate the effect. The procedure is almost same as shown in Fig. 1, but the baseline model ($LM0$) is substituted with the model adapted using the initial ASR result ($LM3$) to generate the adapted model ($LM4$). The synergetic effect was confirmed and WER was reduced to 28.7% from 33.1% (total error reduction rate is 13.3%). Fig. 2 shows WER for each test-set lecture. Except for A5, we achieved WER reduction of 1% to 6% absolute, which confirms the significant effects of the proposed unsupervised adaptation methods.

6. CONCLUSIONS

We have presented methods that adapt a language model to speaker's characteristics in expression and pronunciation variation. Especially, adapting pronunciation variation together with the language model is vital in spontaneous speech recognition. All proposed adaptation methods and their combinations effectively reduced the perplexity and WER.

Acknowledgment: The authors are grateful to Prof. S. Furui, Dr. A.Yamada and Mr. K. Uchimoto of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology". The authors are thankful to Prof. H. G. Okuno of Kyoto University for useful comments. This paper is supported in part by Informatics Research Center for Development of Knowledge Society Infrastructure (COE program of the Ministry of Education, Culture, Sports, Science and Technology, Japan).

7. REFERENCES

[1] S.Furui, K.Maekawa, and H.Isahara, "Toward the realization of spontaneous speech recognition – introducing of a japanese priority program and preliminary results –," in *Proc. ICSLP*, 2000, vol. 3.

[2] H.Nanjo and T.Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," in *Proc. IEEE-ICASSP*, 2002, vol. 1, pp. 725–728.

[3] S.F.Chen, K.Seymore, and R.Rosenfeld, "Topic Adaptation for Language Modeling Using Unnormalized Exponential Models," in *Proc. IEEE-ICASSP*, 1998, vol. 2, pp. 681–684.

[4] M.Mahajan, D.Beeferman, and X.D.Huang, "Improved Topic-Dependent Language Modeling using Information Retrieval Techniques," in *Proc. IEEE-ICASSP*, 1999, vol. 1, pp. 541–544.

[5] T.Shinozaki and S.Furui, "Towards automatic transcription of spontaneous presentations," in *Proc. EUROSPEECH*, 2001, pp. 491–494.

[6] A.Lee, T.Kawahara, K.Takeda, and K.Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proc. IEEE-ICASSP*, 2000, vol. 3, pp. 1269–1272.

[7] A.Lee, T.Kawahara, and K.Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. EUROSPEECH*, 2001, pp. 1691–1694.

[8] A.Lee, T.Kawahara, K.Takeda, M.Mimura, A.Yamada, A.Ito, K.Itou, and K.Shikano, "Continuous Speech Recognition Consortium — an open repository for CSR tools and models —," in *In Proc. IEEE Int'l Conf. on Language Resources and Evaluation*, 2002.

[9] T.Niesler and D.Willett, "Unsupervised Language Model Adaptation for Lecture Speech Transcription," in *Proc. ICSLP*, 2002.

[10] B.Peskin, M.Newman, D.McAllaster, V.Nagesha, H.Richards, S.Wegmann, M.Hunt, and L.Gillick, "Improvements in Recognition of Conversational Telephone Speech," in *Proc. IEEE-ICASSP*, 1999, vol. 1, pp. 53–56.

[11] H.Schramm and X.Aubert, "Efficient integration of multiple pronunciations in a large vocabulary decoder," in *Proc. IEEE-ICASSP*, 2000, vol. 3, pp. 1659–1662.