論文 / 著書情報
Article / Book Information

Title	Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition
Authors	Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, SADAOKI FURUI, Haruo Yokota
Citation	Proc. INTERSPEECH 2007, Vol. , No. , pp. 2349-2352,
Pub. date	2007, 8
Copyright	(c) 2007 International Speech Communication Association, ISCA
DOI	http://dx.doi.org/

Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition

Hiroki Yamazaki, Koji Iwano, Koichi Shinoda, Sadaoki Furui and Haruo Yokota

Department of Computer Science, Tokyo Institute of Technology, Japan
yamazaki@ks.cs.titech.ac.jp, {iwano, shinoda, furui, yokota}@cs.titech.ac.jp

Abstract

We propose a dynamic language model adaptation method that uses the temporal information from lecture slides for lecture speech recognition. The proposed method consists of two steps. First, the language model is adapted with the text information extracted from all the slides of a given lecture. Next, the text information of a given slide is extracted based on temporal information and used for local adaptation. Hence, the language model, used to recognize speech associated with the given slide changes dynamically from one slide to the next. We evaluated the proposed method with the speech data from four Japanese lecture courses. Our experiments show the effectiveness of our proposed method, especially for keyword detection. The F-measure error rate for lecture keywords was reduced by 2.4%.
Index Terms: language model adaptation, speech recognition, classroom lecture speech.

1. Introduction

Recent advancements in computer and storage technology enable archiving large multimedia databases. The databases of classroom lectures in universities and colleges are particularly useful knowledge resources, and they are expected to be used in education systems.

Recently much effort has been made to construct educational systems that use the multimedia content of classroom lectures to support distant-learning [1, 2, 3, 4, 5]. Among the various kinds of content related to lectures, the transcription of speech data is expected to be the most important for indexing and searching lecture contents [2, 6]. Therefore, high-level speech recognition engine for lectures is required. Lecture speech recognition has been studied extensively. Many research projects for lecture transcriptions, such as the European project CHIL (Computers in the Human Interaction Loop) [8], and the American iCampus Spoken Lecture Processing project [9], have been conducted. Trancoso *et. al.* [7] investigated the automatic transcription of classroom lectures in Portuguese.

Large databases of conference presentations, such as the Corpus of Spontaneous Japanese (CSJ) [10, 11], and the TED corpus [12] have been collected to improve speech recognition accuracy. With the use of these databases, a state-of-the-art speech recognition systems for conference presentations achieves accuracy of 70-80%. Hence, the recognition results provided by these systems are good enough to be used for speech summarization and speech indexing [13]. The speaking style of classroom lectures is, however, much different from that of lectures in meetings or conferences. Classroom lectures are not always practiced in advance, and the same phrases are repeated many times for emphasis. The lecture speaking style is closer to that in dialogue because lecturers are always ready to be interrupted by questions from students. The spontaneity of this kind of speech is much higher than other kinds of presentations; the lectures are characterized by strong coarticulation effects, non-grammatical constructions, hesitations, repetitions, and filled pauses. For these reasons, speech recognition for classroom lecture speech is generally more difficult than that of speeches in conferences or meetings; its recognition accuracy is around 40-60%. Furthermore, no large database of classroom lecture speech is available for training acoustic and language models.

In classrooms, lecturers often use various materials, e.g., textbooks or slides, to help their students understand. Since those materials include many keywords that also appear in lecture speech, they are expected to be useful for language modeling in speech recognition. Several adaptation methods for language models using such content have already been proposed for lecture speech recognition. For example, Togashi *et. al.* [14] proposed a method of using the text information in presentation slides.

If lecture speech is accompanied by slides, a strong correlation can be observed between slides and speech. In particular, the speech corresponding to a given slide contains most of the text information presented in the slide. We expect this relation between speech and text information of the slide can improve the model adaptation for lecture speech recognition. We propose a *dynamic* adaptation method for language modeling that applies text information from slides. In this method, a slide-dependent language model is constructed for each slide, and this model is used afterwards to recognize the speech associated with the given slide. The language model is changed dynamically as the lecture progresses.

This paper is organized as follows. In Section 2, the base system applied in our studies is introduced. In Section 3, the proposed language model adaptation method is explained, and in Section 4, the effectiveness of the proposed method is discussed.

2. UPRISE: Unified Presentation Contents Retrieval by Impression Search Engine

UPRISE (Unified Presentation Contents Retrieval by Impression Search Engine) [1, 15] is a lecture presentation system to support distant-learning. It stores many types of multimedia materials, such as texts, pictures, graphs, images, sounds, voices, and videos, and provides a unified presentation view (Figure 1) as a lecture video retrieval system. The retrieval system returns appropriate lecture video scenes to match given keywords. Since the speech information in lectures is used to narrow down the search candidates [6], a high level of speech recognition accuracy is strongly required.

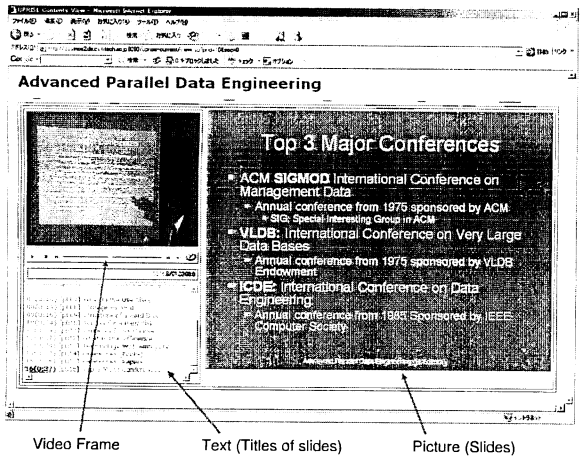

Figure 1: *Unified presentation view in UPRISE.*

The content in lectures stored in UPRISE are synchronized as the class progresses; that is, each event in a lecture is marked with temporal information. This synchronized content motivated us to investigate a dynamic adaptation method along the time-axis by using the content strongly related to speech information. We propose a dynamic language model adaptation method using one of the kinds of content, slide information.

3. Dynamic Language Model Adaptation

In UPRISE, the temporal information of when each slide is shown in the class is detected automatically and recorded [16]. This function enable us to investigate adaptation methods by using temporal information from a class. Here, we propose a dynamic adaptation method for language modeling that uses slides with this temporal information. By using slide information, the proposed method changes the language model parameters so that they are fitted to the technical terms that characterize the corresponding part of the lecture. To update the language model, n-gram counts of the adaptation data extracted from the slides are added to those of the baseline training data with a weighting coefficient. The vocabulary of the adapted language model consists of words from the original training data and the adaptation data.

The detailed algorithm of the proposed adaptation method is as follows. First, *Global Adaptation* (GA) is conducted, in which the text data of all slides used in a course are used as adaptation data. The frequency $F_G(N_i)$ of each n-gram N_i is calculated as follows.

$$F_G(N_i) = F(N_i) + w_1 G(N_i), \quad (1)$$

where $F(N_i)$ is the frequency of n-gram N_i that appear in the baseline training data, $G(N_i)$ is the frequency of n-gram N_i that appear in the adaptation data, and w_1 is a weight coefficient that should be optimized experimentally. Next, in *Local Adaptation* (LA), the language model from GA is further adapted locally to each slide. The frequency $F_L(N_i)$ of each n-gram N_i, corresponding to each slide, is calculated as follows.

$$F_L(N_i) = F_G(N_i) + w_2 L(N_i), \quad (2)$$

where $L(N_i)$ is the number of appearances of n-gram N_i in each slide, and w_2 is a weighting factor. A new language model for each slide is constructed by using the frequency $F_L(N_i)$ for all the n-grams.

Table 1: Details of the collected lecture database.

	LEC1	LEC2	LEC3	LEC4
Lecturer	A	A	B	C
Size of keyword list	136	56	53	135
Num. of classes	11	10	7	11
Lecture length (min.)	766	831	545	952
Lecture length using slide (min.)	754 (98.4%)	803 (96.7%)	352 (64.7%)	948 (99.6%)
Num. of slides	265	282	136	270
Num. of Words per slide	52.1	60.2	52.5	13.0
Num. of Keywords per slide	9.3	5.5	8.3	3.4

4. Recognition Experiments

4.1. Experimental Conditions

We collected audio and video of four Japanese lecture courses (LEC1, LEC2, LEC3, and LEC4) for evaluation. Each lecture course consisted of 12 classes, where each class duration was around 80 minutes. The audio data was recorded using a close-talking microphone. The lecture courses LEC1 and LEC2 were given by the same lecturer. The data from nine classes were excluded because their recording quality was poor. We collected PowerPoint slides used in those lectures along with the temporal information. The speech data were segmented using the temporal information such that each speech segment had the same boundary as its corresponding slide. When an utterance occurred at the exact boundary of two slides, the speech data was cut just after the utterance. We manually transcribed all the speech data for evaluation except the utterances of speakers other than the lecturer. Keywords, which were mostly technical terms characterizing the lectures, were selected subjectively by several researchers. The details of the collected lecture database are listed in Table 1.

The best way to improve classroom speech recognition accuracy is to use a fairly large amount of speech data with transcriptions as training sets for the acoustic and language models. However, as already discussed in Introduction, no such large corpus for classroom lectures has been available. Therefore, to construct our models, we used the CSJ database, which consists of presentation speech data from academic conferences. The CSJ data are expected to share similar properties with classroom speech to some extent as they are speech data of monologues with specified themes.

As the initial language model, we constructed a trigram model by using 967 academic presentation transcriptions (3M morphemes) from CSJ. We call this model the *baseline* model. We used ChaSen [17], a Japanese morphological analyzer, for preprocessing the text data. The recognition vocabulary consisted of 25,000 words, which appeared most frequently in the training set. We used the Witten-Bell method for back-off smoothing [18].

The initial acoustic model was also constructed from the CSJ data set, 953 academic presentations, and 1,543 extemporaneous presentations that include both male and female speak-

Table 2: Results of speech recognition and keyword detection achieved using the baseline language model (%).

	Word acc.	Recall	Precision	F-measure
LEC1	39.2	37.2	59.1	45.7
LEC2	37.3	36.1	66.2	46.7
LEC3	57.7	56.4	82.4	67.0
LEC4	60.1	49.6	71.9	58.7
Avg.	47.4	45.0	69.1	54.5

Table 3: Results of speech recognition and keyword detection by Global Adaptation (GA) (%).

	Word acc.	Recall	Precision	F-measure
LEC1	41.1	50.6	65.9	57.2
LEC2	38.7	49.8	71.6	58.7
LEC3	59.8	65.8	83.1	73.4
LEC4	61.4	57.1	74.4	64.6
Avg.	49.0	55.5	72.8	63.0

ers. We used 25 dimensional acoustic features: 12 dimension MFCC, 12 ΔMFCC, and Δ power. Cepstral Mean Subtraction (CMS) was used to filter each utterance. We used left-to-right 3-state triphone HMMs, which have 3000 states and 16 mixtures per state, and used HTK [19] as the training tool. We also conducted unsupervised adaptation by using the MLLR method [20]. As adaptation data, we used the opening 10 minutes speech of collected classes. The number of regression classes was 64. MLLR adaptation is expected to change the acoustic model parameters to fit the speaker or noises in the test set. We used Julius, an open-source real-time large vocabulary speech recognition engine, as the decoder [21]. We evaluated the recognition results in terms of word accuracy. We evaluated the keyword detection in terms of recall and precision rates, which are important for evaluating retrieval performance. In the evaluation of keyword detection, we compared the beginning time t_1 of a keyword in the recognition result with that in the reference, t_2. When the difference between t_1 and t_2 was less than 500 ms, we assumed that the keyword was correctly recognized.

4.2. Results

First, we investigated the initial language model and the initial acoustic model performance. The average recognition accuracy was 41.9%, the keyword recall was 36.8%, and the keyword precision was 62.6%. The recognition results achieved by using the acoustic model adapted by unsupervised MLLR are given in Table 2. The average recognition accuracy was 47.4%, the keyword recall was 45.0%, and precision was 69.1%. In all the following experiments, we used the acoustic models adapted by MLLR.

Next, we investigated the efficiency of Global Adaptation (GA). The relationship between the weight w_1 and the recognition rate is shown in Figure 2. We found that the recognition rate was not much influenced by the change of w_1 value. According to these results, we set the value of w_1 to 20. The results obtained using GA are presented in Table 3. The average word error rate was reduced by 3.0%. The average error rates for recall and precision of keyword detection were also reduced by 19.1% and 12.0%, respectively. These results indicate that GA

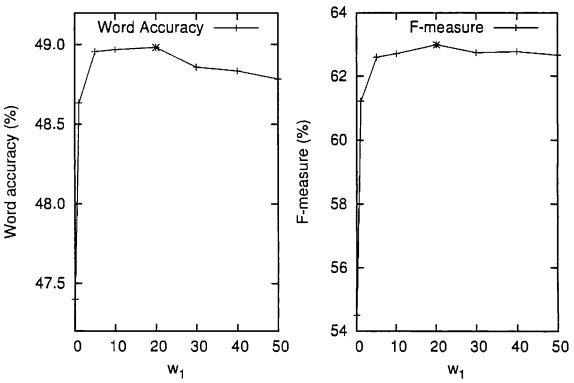

Figure 2: Relationship between the weight w_1 and recognition performance.

Table 4: The effect of the expansion of the vocabulary by Global Adaptation (GA).

LM	Dictionary	Acc	FM	PP	OOV
Baseline	Baseline	47.4%	54.5%	167.5	4.3%
Baseline	Extended	47.8%	55.2%	167.5	3.4%
GA	Extended	49.0%	63.0%	161.7	3.4%

was effective for speech recognition and keyword detection.

In GA, the words included in the slides were added to the recognition dictionary. In order to investigate the effect of vocabulary expansion, we performed an experiment in which we used the baseline language model and the extended dictionary. The recognition results, the perplexities, and the OOV rates are shown in Table 4. The obtained results indicate that the effectiveness of GA was the result of both the expansion of the vocabulary and the change of the statistical parameters of the language models.

Finally, we evaluated the effectiveness of Local Adaptation (LA). The relationship between the weight w_2 and recognition performance is presented in Figure 3. The recognition accuracy was not much influenced by w_2, so we set it to 9,000. The recognition results are listed in Table 5. While the recognition accuracy was not much changed from the results obtained by GA, the keyword detection rate was significantly improved. The keyword F-measure error rate was reduced by 2.4% on average. Thus we confirmed the effectiveness of the proposed dynamic language model adaptation method for keyword detection.

The results shown in Table 5 indicate that the effectiveness of LA was not the same for all lectures. While in LEC1 and LEC2, the error rate of keyword F-measure was reduced by 3.8%, that was only reduced by 1.2% in LEC3 and LEC4. In the case of LEC3, this difference might be due to the fact that the times of slide presentations was particularly short compared to those in the other lectures (Table 1). When slides were not presented, the GA model was used to recognize speech. This means that only a small amount of speech in LEC3 was recognized with the LA model, and thus LA did not have much impact as it did in the case of LEC1 and LEC2. In LEC4, the slides contain a relatively small amount of words or keywords (Table 1). This might be the reason that the effectiveness of LA for LEC4 is less than for the other lectures.

Table 5: Results of speech recognition and keyword detection by Local Adaptation (LA) (%).

	Word acc.	Recall	Precision	F-measure
LEC1	41.1	52.5	66.3	58.6
LEC2	38.6	51.9	73.1	60.7
LEC3	59.4	66.3	83.3	73.8
LEC4	61.2	57.7	74.3	65.0
Avg.	48.8	56.7	73.1	63.9

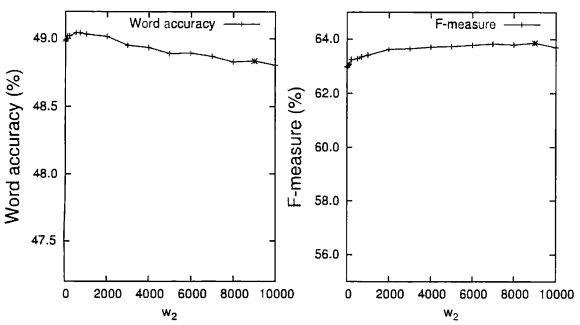

Figure 3: Relationship between the weight w_2 and recognition performance ($w_1 = 20$).

5. Conclusions and Future Work

We have proposed a dynamic adaptation method of language modeling that exploits slide information from lecture speech with temporal information. We evaluated the proposed method with the speech data of four lecture courses in Japanese. The results showed the effectiveness of our method, especially for keyword detection.

In future, we need to collect more lecture data because the size of our present database is still small for fair evaluation. In addition, we must investigate more efficient use of slides for adaptation. We also plan to extend our framework to other content such as lecture speech in meetings or TV broadcasts.

6. Acknowledgements

This study was supported by the 21st COE program, Frame work for Systemization and Application of Large-scale Knowledge Resources.

7. References

[1] H. Yokota, T. Kobayashi, T. Muraki and S. Naoi, "UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine," IEICE Transactions on Information and Systems, vol. E87-D, no. 2, pp. 307-406, 2004.

[2] A. Fujii, K. Itou and T. Ishikawa, "LODEM: A system for on-demand video lectures," Speech Communication 48, pp. 516-531, 2006.

[3] R. Müller and T. Ottmann, "The Authoring on the Fly system for automated recording and replay of (tele)presentations," Multimedia Systems, vol. 8 no. 3, pp. 158-176, 2000.

[4] Informedia ii digital video library, Carnegie Mellon University The Informedia Project, http://www.informedia.cs.cmu.edu/.

[5] G. D. Abowd, "Classroom 2000: an experiment with the instrumentation of a living educational environment," IBM Systems Journal, vol. 38, no. 4, pp. 508-530, 1999.

[6] H. Okamoto, W. Nakano, T. Kobayashi, S. Naoi, H. Yokota, K. Iwano and S. Furui, "Unified presentation contents retrieval using voice information," Proc. DEWS2006, 6c-o1, 2006 (in Japanese).

[7] I. Trancoso, R. Nunes and L. Neves, "Recognition of classroom lectures in European Portuguese," Proc. INTERSPEECH 2006 - ICSLP, pp. 281-284, 2006.

[8] L. Lamel, G. Adda, E. Bilinski and J. L. Gauvain, "Transcribing lectures and seminars," Proc. INTERSPEECH 2005, pp. 1675-1660, 2005.

[9] J. Glass, T. Hazen, I. Hetherington and C. Wang, "Analysis and processing of lecture audio data: Preliminary investigations," Proc. Human Language Technology NAACL, Speech Indexing Workshop, Boston, 2004.

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

[11] The Corpus of Spontaneous Japanese, National Institute for Japanese Language, http://www2.kokken.go.jp/csj/public/.

[12] L. Lamel, F. Schiel, A. Fourcin, J. Mariani and H. Tillmann, "The Translanguage English Database TED," Proc. ICSLP, vol. 4, pp. 1795-1798, 2004.

[13] S. Furui, "Recent progress in corpus-based spontaneous speech recognition," IEICE Transactions on Information and Systems, vol. E88-D, no. 3, pp. 366-375, 2005.

[14] S. Togashi, N. Kitaoka and S. Nakagawa, "Speech recognition of lecture documents with LM adapted by lecture slides," Proc. Acoustical Society of Japan Spring Meeting, 1-P-24, pp. 191-192, 2006 (in Japanese).

[15] H. Yokota, T. Kobayashi, H. Okamoto and W. Nakano, "Unified contents retrieval from an academic repository," Proc. International Symposium on Large-scale Knowledge Resources (LKR2006), Tokyo, Japan, pp. 41-46, 2006.

[16] N. Ozawa, H. Takebe, Y. Katsuyama, S. Naoi and H. Yokota, "Slide identification for lecture movies by matching characters and images," Proc. SPIE, vol. 5296-10, Document Recognition and Retrieval XI, pp.74-81, 2004.

[17] ChaSen (ver. 2.2.3) and ipadic (ver. 2.4.4), http://chasen.naist.jp/hiki/ChaSen/.

[18] I. H. Witten and T. C. Bell, "The zero-frequency problem: Estimating the probabilities of novel events in adaptive text compression," IEEE Transactions on Information Theory, vol. 37, no. 4, pp. 1085-1094, 1991.

[19] HMM Tool Kit (HTK) (ver. 3.2), http://htk.eng.cam.ac.uk/.

[20] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," Computer Speech and Language, vol. 9, no. 2, pp. 171-185, 1995.

[21] Julius (ver. 3.5), http://julius.sourceforge.jp/en/julius.html.