LANGUAGE MODEL ADAPTATION FOR ACADEMIC LECTURES USING CHARACTER RECOGNITION RESULT OF PRESENTATION SLIDES

Yuya Akita, Yizheng Tong and Tatsuya Kawahara

School of Informatics, Kyoto University, Kyoto 606-8501, Japan

ABSTRACT

For automatic speech recognition (ASR) of lectures, texts of presentation slides are expected to be useful for adapting a language model, while slide texts are not always available in a machine-readable form. In this paper, we propose a language model adaptation framework that uses character recognition results of slide images in a lecture video. Since character recognition results contain many errors, we introduce a filtering method based on morphological and topic information. Then we perform linear interpolation of the baseline language model with the filtered results and also relevant texts which are selected automatically from a text database using the filtered results. We further conduct a cache-based adaptation method on the resulting language model, in which keywords in the filtered results are cached and used to boost the word probability. In an experimental evaluation over real lectures, we obtained a significant improvement of ASR performance by this adaptation framework.

Index Terms— Language model, adaptation, lectures, character recognition, presentation slides

1. INTRODUCTION

Lecture videos are nowadays archived and opened to the public by several projects such as OpenCourseWare (OCW)[1] and massive online open courses (MOOC)[2]. As the amount of these archives is constantly increasing, efficient browsing should be implemented, for example, by indexing these videos using audio transcripts. Transcripts are also helpful as captions to understand technical terms, which are often observed in academic lectures, as well as to support hearing-impaired [1] and non-native viewers. However, it takes much cost to transcribe audio and give captions by hand. Manual compilation is virtually impossible for a large archive.

For efficient transcription of lectures, automatic speech recognition (ASR) can be used [2, 3, 4, 5]. A distinctive characteristic of lecture speech is a variety of technical terms depending on the topic of each lecture. For example, names of new diseases such as "severe fever with thrombocytopenia syndrome" or "SFTS" often appear in the field of medical science. Two kinds of approaches have been adopted to deal with the variety in a vocabulary, i.e., a huge-vocabulary (e.g. millions of words) universal language model and language model adaptation to the target lecture. However, newest words such as "SFTS" are not expected to be included in the vocabulary of the universal model, unless constantly maintained. Furthermore, it is more difficult to cover contextual information of technical terms, rather than the terms themselves, in the universal model. Thus, for better ASR of lectures, language model should be adapted to each lecture using relevant text materials.

For language model adaptation in lecture ASR, newspaper articles [1, 6], technical books and documents [1, 7, 8], presentation slides [9, 10, 11, 12, 13] and web texts [14, 15] are often used. In this study, we particularly focus on presentation slides, because slides are used in many lectures and the content of the slides is directly reflected in the wording of utterances. Several studies have been reported on slide-based adaptation, for example, Yamazaki et al. [9] proposed a linear interpolation framework which combined a baseline model with a model trained with whole slide texts and another model trained with a single slide using the timestamp of slides. We also proposed a combination of probabilistic latent semantic analysis (PLSA), a collection of web texts, and a cache model using slide texts [10]. Miranda et al. [11] proposed a rescoring framework using word lattices made from ASR results and slide texts.

These studies assume that slide texts are obtained in a machine-readable form such as an electronic file. However, this is not often the case with online archives (e.g. YouTube videos) in which only a lecture video can be used for ASR. For this kind of lecture videos, a possible solution is to use optical character recognition (OCR) to obtain textual information from slide images in a lecture video. Language model adaptation using OCR results has been recently investigated by Martínez-Villaronga et al. [12] and Wiesler et al. [13]. In the video archive[3] used by these studies, slide images to which OCR was conducted were separately provided with good readability for education purpose. In contrast, we suppose a common scenario (e.g. YouTube videos), in which only a lecture video is available, i.e., slide images are recorded in the video together with other scenes such as the lecturer and the audience. The slide images are noisy and the quality of them is low, hence OCR produces much more recognition errors, compared to the previous studies. Therefore, it is needed to perform removal of errors from OCR texts and adaptation using fragmented OCR results. For the former, we propose a filtering method dedicated to OCR texts. For the latter, we extend the adaptation framework proposed in the previous work [10], and investigate its effectiveness when using erroneous OCR texts.

In this paper, we first examine how much extent OCR can work on lecture slides, and then investigate its usability in terms of lexical coverage. Next, we describe the proposed framework of language model adaptation using OCR results, followed by an experimental evaluation.

2. OCR OF PRESENTATION SLIDES

2.1. OCR system for lecture videos

A usual OCR system first detects an image region where some characters are found, then perform pattern matching with character templates. In this study, we need to detect segments of slide images in a video before the OCR process, as there are several patterns of video segments that include slide images and they occasionally switch to others. Moreover, by determining a timestamp of each slide, we can conduct more precise adaptation using the texts of the slide corresponding to each of the lecturer's utterances.

From this point of view, we adopt "TalkMiner[4]" [16] developed by Fuji Xerox. It is an online system opened to the public, and gives indices to online videos by OCR. First, it detects video segments where no visual motions are observed, and extract them as slide images. Then, it performs OCR on slide images to generate texts. It

[1] www.oeconsortium.org.

[2] E.g. edX (www.edx.org) founded by MIT and Harvard University.

[3] videolectures.net.

[4] www.talkminer.com.

Table 1. Specifications of the set of lectures

Lectures	6
Total duration	180.3 minutes
Total number of words	38,638
Accuracy by baseline ASR system	78.4%

Table 2. Error and recall rates in the OCR results

Substitution errors	27.8%	Recall rate	64.6%
Deletion errors	7.7%	False alarm rate	35.4%

finally associates each slide image and corresponding texts as a slide index and saves them to a database. This OCR output contains timestamps of slides, i.e., beginning time and duration of each slide, which enable time alignment of slides with an input audio stream. Note that the TalkMiner system does not provide confidence measure scores.

2.2. Usability of OCR results

When performing OCR on slide images from lecture videos, recognition accuracy significantly degrades because of the quality of extracted images; the resolution of online videos is often low. Another reason is that OCR sometimes treats objects in pictures as texts since some characters often match with some image portion, which result in false alarms. Here, we conducted a preliminary investigation on the performance and the usability of OCR results using real lecture videos. We used six lectures, whose specifications are listed in Table 1, from academic symposia held by the Center for iPS cell Research and Application (CiRA) of Kyoto University, in the years of 2010 and 2011. The topics of these lectures were state-of-the-art stem cell research and related backgrounds in biology and medical sciences. All lectures were made in Japanese language, and are available at the Kyoto University OCW website[5]. The lecture videos were processed by TalkMiner. The average recall rate was 64.6%, but the false alarm rate was 35.4%, as shown in Table 2. The false alarms were caused mainly at figures and pictures in the slide images. The total error rate of this result (70.9%) is significantly higher than that of the previous work [12] (43%). We also performed speech recognition using the baseline system, which will be described in Section 5.2, and obtained the ASR accuracy of 78.4%.

If OCR texts cover out-of-vocabulary (OOV) words of the language model in the ASR system and incorrectly recognized keywords by the ASR system, OCR texts may contribute to improvement of ASR performance. Thus, we calculated the coverage of OOV words and that of keywords which were not recognized by the baseline ASR system.

Table 3 shows these coverages. To calculate the coverage of OOV words, we used manual transcription of the lecture videos. The language model in the ASR system was trained with a collection of lectures, but does not cover state-of-the-art stem cell research. Therefore, several technical terms such as "iPS cell[6]" were not covered by the model, and the OOV rate was 3.1%. Among these OOV words, 53.1% of the words were covered by correct slide texts, which were made by hand from slide images. Even using OCR results, 45.7% of the OOV words were covered. This result suggests that the OOV rate can be reduced by using OCR results of slide images.

Keywords are defined as characteristic words in a particular lecture. Specifically, we use tf-idf scores to pick them up. For calculation of document frequency, we used a set of articles from the Mainichi newspaper in the year 2011 (total 81,768 articles) as a document set. We calculated tf-idf scores for all words in each lecture using manual transcripts, sorted the words by these scores, and then extracted the top one-tenth words as the keywords. Here, some function words such as particles, disfluency phenomena, and stop words were excluded from the keyword set. Table 3 shows the coverage on the keywords. By ASR, 44.9% of the keywords were incorrectly recognized. Correct slide texts and OCR results covered 93.1% and 91.2% of these misrecognized keywords, respectively. Thus, it is expected to improve ASR performance by adaptation of language model even using OCR results.

Table 3. Coverages of OOV words and misrecognized keywords by slide texts

OOV rate	3.1%
OOV words covered by OCR slide texts	45.7%
OOV words covered by correct slide texts	53.1%

Misrecognized keywords among all keywords	44.9%
Misrecognized keywords covered by OCR slide texts	91.2%
Misrecognized keywords covered by correct slide texts	93.1%

3. FILTERING OF OCR RESULTS

The process flow of the proposed language model adaptation is shown in Fig. 1. In this section, we explain the first step, i.e., filtering based on morphological and topic information to remove OCR errors.

3.1. Filtering based on morphological information

As mentioned in Section 2.2, there are a large number of false alarms which are mainly caused by objects in figures and pictures in input slide images. Since many of them do not constitute a lexical unit, we first apply a morphological analyzer to detect them. The advantage of the use of morphological analyzer is its coverage. A morphological dictionary usually has a wider coverage than those used in ASR. Even for the newest words and technical terms which are difficult to be covered in the word level, it can often be covered by decomposing into morphemes. Units rejected by a morphological analyzer are likely to be OCR errors, hence we can remove them.

In this work, we use KyTea[7] [17] as a Japanese morphological analyzer. It gives a dedicated tag "NA" to non-alphabet signs, marks and symbols, as the common character set of Japanese has thousands of characters including a variety of non-alphabet symbols. Another tag "UNK" is also given to unknown units which are not recognized by the analyzer. In our morphological filtering, we simply remove all words that have NA or UNK tag.

3.2. Filtering based on topic information

Even after the morphological filtering, there may be OCR errors in the filtered texts which are lexically correct but irrelevant to the slides. To remove this kind of errors, we further apply topic-based filtering. The basic idea of this filtering is to restrict the vocabulary of the OCR results to the relevant topic words. To emphasize topic words, several probabilistic models such as PLSA [18] and latent Dirichlet allocation (LDA) [19] were proposed. However, the probabilistic approaches are not suitable for our purpose, which is not to precisely predict some specific word but just to decide to keep or discard each input word. Thus, we adopt simple definition of vocabulary by collecting relevant documents to the OCR results.

In this step, we first select relevant documents from a text database which covers a wide range of topics. These documents are selected based on the similarity to the OCR results, thus the words in the set are likely to be observed in slide texts. Then, we extract a set of words which appear in these documents. Using this set, the OCR

[5] ocw.kyoto-u.ac.jp.

[6] iPS cell stands for "induced pluripotent stem cell."

[7] www.phontron.com/kytea/.

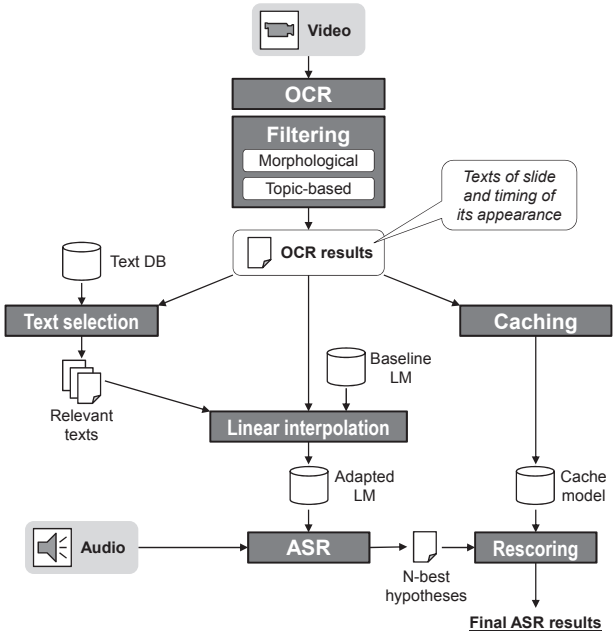

Fig. 1. Proposed adaptation framework

results are again filtered so that only the words in the set are kept and the other words are discarded. The similarity of the document and the OCR results is defined as a cosine distance of vectors which are composed of tf-idf scores of keywords defined in Section 2.2. We again use the articles of the Mainichi newspaper in 2011 as a document database, and the total number of articles is 81,768. We empirically define the number of articles to be selected as 5,000 to provide a good coverage.

4. LANGUAGE MODEL ADAPTATION USING OCR RESULTS AND RELEVANT DOCUMENTS

Using the filtered texts, adaptation is conducted on the baseline model. The adaptation method consists of linear interpolation for "global" adaptation to the topics throughout the lecture, and a cache model for "local" adaptation to the specific segment of speech.

4.1. Linear interpolation of baseline language model

The baseline language model is adapted by linearly interpolating with the filtered OCR texts, as shown Equation (1):

$$P_{\mathrm{mix\_slide}}(w) = (1-\alpha)P_{\mathrm{base}}(w) + \alpha P_{\mathrm{slide}}(w). \quad (1)$$

Here, $P_{\mathrm{mix\_slide}}(w)$ is the adapted model, $P_{\mathrm{base}}(w)$ is the baseline model and $P_{\mathrm{slide}}(w)$ is a model trained with the OCR texts. According to Table 3, OCR texts can cover a significant number of OOV words and keywords, thus the interpolation is expected to improve the performance.

As the amount of the filtered OCR texts is usually small, we also incorporate relevant documents for adaptation. In this work, we use news articles selected in the same manner in Section 3.2, but the number of the selected articles is significantly reduced to cover only relevant topics. The selected articles are interpolated with the baseline model together with the filtered OCR texts based on Equation (2):

$$P_{\mathrm{mix\_docs}}(w) = (1-\alpha-\beta)P_{\mathrm{base}}(w) + \alpha P_{\mathrm{slide}}(w) + \beta P_{\mathrm{docs}}(w), \quad (2)$$

where $P_{\mathrm{mix\_docs}}(w)$ is the adapted model and $P_{\mathrm{docs}}(w)$ is a model trained with the relevant documents. For Equations (1) and (2), the interpolation weights α and β are optimized using the development set which will be described later.

4.2. Adaptation based on cache model

Linear interpolation using the OCR results is considered as "global" adaptation for the entire lecture because it uses texts of all slides of the lecture. In contrast, by using only the slide which was presented when the target utterance was made, "local" adaptation for a specific speech segment can be performed, since words appeared in these slides are more likely to be used in actual utterances. To assign a higher linguistic score to these kinds of words, we introduce a cache model framework [20], and rescore N-best hypotheses provided by the adapted language model described in the previous section.

In the original cache model framework [20], words in previous utterances are pooled and assigned a higher linguistic score. In this work, we use a context of slide texts instead of a context of uttered words, and we "cache" words in the slide which corresponds to the target utterance, and its previous and following slides. Here, words to be cached are limited to content words. The reason to use multiple slides is that a lecturer sometimes speaks about the previous or the following slide. For example, even after switching slides forward, a lecturer often refers to the content of the previous slide. When showing and explaining a figure or a picture in a slide, its description is sometimes found in the adjacent slide. Consequently, a word probability by this cache model is calculated based on Equation (3):

$$P(w|S_p) = \frac{1}{\sum_{i=p-1}^{p+1}|S_i|}\sum_{i=p-1}^{p+1}\sum_{w_s\in S_i}\delta(w,w_s), \quad (3)$$

where S_i is the i-th slide and $|S_i|$ is the total number of words in the slide S_i. $P(w|S_p)$ is a cache model probability of word w in the p-th slide S_p. $\delta(w,w_s)$ is the Kronecker delta, i.e., the function returns 1 when $w = w_s$ and 0 otherwise. Using this cache model probability, rescoring is performed according to Equation (4):

$$P(w_i|w_{i-1},w_{i-2},S_p) = \gamma P(w_i|S_p) + (1-\gamma)P(w_i|w_{i-1},w_{i-2}). \quad (4)$$

Here, we assume the language model is word trigram model, and $P(w_i|w_{i-1},w_{i-2})$ denotes a trigram probability by this model. This rescoring is controlled by a weight γ.

5. EXPERIMENTAL EVALUATION

We made an experimental evaluation of the proposed filtering and adaptation methods. The test set was six lectures in the CiRA symposia at Kyoto University in 2010 and 2011, which was described in Section 2.2.

5.1. Performance of filtering of OCR results

As evaluation measures of filtering of OCR results, we calculated the recall and precision rates of all words in the test-set lectures, which are listed in Table 4. We also calculated recall rates of OOV words and keywords before and after applying our filtering method, as shown in Table 5.

In Table 4, the precision rate was improved by 18.5 points, while the recall rate was dropped by 2.4 points. In Table 5, the degradation of recall rates were only 5.5 and 0.6 points for OOV words and keywords, respectively. These results demonstrate that the proposed filtering method effectively eliminated OCR errors without discarding correct OCR results so much.

Table 4. Recall and precision rates of all words before and after filtering

	Recall	Precision	F-measure
Before filtering	59.5%	31.7%	41.4
After filtering	57.1%	50.3%	53.5

Table 5. Recall rates of OOV words and keywords before and after filtering

	OOV words	Keywords
Before filtering	86.1%	97.9%
After filtering	80.5%	97.3%

5.2. Global adaptation of language model

For ASR experiments, we used our Julius 4.1.5 decoder. The baseline language model was a word trigram model, and trained using all lecture transcripts in the Corpus of Spontaneous Japanese (CSJ) [21]. As mentioned in Section 2.2, the CSJ is a collection of lectures, but it does not cover the topics of the test set, i.e., biology and medical science. The amount of the training texts was 7.7M words, and the vocabulary size was 37K words. Throughout the experiments in this work, the interpolation weights were optimized based on perplexity over the transcripts of the development set, which consisted of three lectures in the same symposium as the test set but held in a different year (2009). For reference, we tested the proposed adaptation using the correct slide texts as well as the OCR results. We also conducted adaptation without relevant documents.

Table 6 lists ASR accuracy of all words (character-based) and keywords by the baseline model, the interpolated model only with the OCR results, and the interpolated model with both the OCR results and the relevant documents. ASR results by the models adapted using the correct slide texts are also listed. Note that keywords are defined in Section 2.2. When interpolating with the OCR results, the interpolation weight α in Equation (1) was determined to be 0.15. When interpolating with the OCR results and relevant articles, the weights α and β in Equation (2) were determined to be 0.05 and 0.20, respectively.

As shown in Table 6, the adapted models significantly improved accuracy. The major portion of the improvement was realized only with the OCR results. Specifically, for all words, we obtained an absolute gain of 4.8% (a relative error reduction of 22%) by the OCR results. For keywords, the gain was 30.5% absolute (67% relative). Using the OCR results, OOV words and technical terms such as "iPS cell" were covered, and thus these words were successfully recognized by the adapted model. Incorporation of relevant documents resulted in further improvement, and absolute gains of 5.8% for all words and 31.7% for keywords were obtained. Compared to the results with the correct slide texts, the improvements with the OCR results on all words were degraded to only 88.3%–95.1%. The same tendency was observed on the recall rate of keywords. The proposed adaptation framework effectively worked, since we used the OCR texts whose F-measure score was 53.5.

The proposed adaptation might take effect on language model by optimizing trigram probabilities, and on word lexicon by adding word entries. To investigate which had larger effect, we conducted adaptation only on language model and only on word lexicon, using the OCR results. When adapting the language model (i.e., trigram probabilities) and not adding words into the lexicon, ASR accuracy for all words was improved by 3.3 points to 81.7%. When adding words into the lexicon and not adapting language model probabilities, the gain was 1.1 points. When the proposed method was fully applied, the respective improvements were jointly observed, and the ASR accuracy reached 83.2% as shown in Table 6. This result shows that adaptation on trigram probabilities and lexicon entries were synergistically worked.

Table 6. Effect of language model adaptation in ASR

		All words	Keywords
Baseline		78.4%	54.8%
w/ slides	OCR results	83.2%	85.3%
(Equation (1))	(cf.) correct texts	83.8%	87.1%
w/ slide & docs	OCR results	84.2%	86.5%
(Equation (2))	(cf.) correct texts	84.5%	87.3%

Figures above are ASR accuracy for all words, and word recall for keywords.

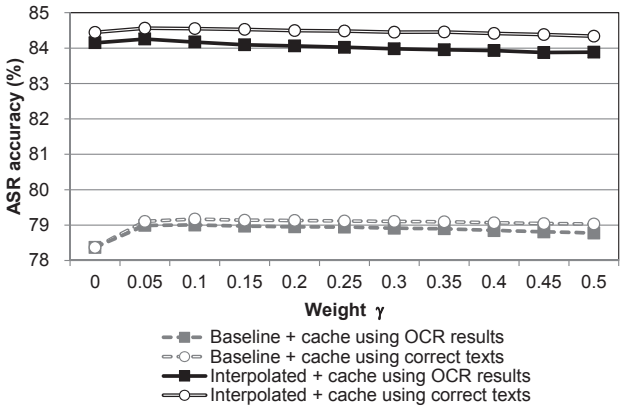

Fig. 2. ASR accuracy on all words by cache model

5.3. Local adaptation based on the cache model

In the evaluation of cache model, we applied it to the baseline model and the adapted model using the OCR results and relevant documents. We generated 100-best ASR hypotheses, and performed rescoring with the cache model on them. Cache model using the correct slide texts was also tested for the reference purpose.

Fig. 2 shows ASR accuracy on all words by the rescoring. We calculated the accuracy by changing the weight γ in Equation (4) from 0 to 0.5 by 0.05. When γ was 0.05, the proposed cache model realized the highest improvements, which were 0.74 points with the correct slide texts and 0.62 points with the OCR results, for the baseline model. On the other hand, when applying the cache model to the adapted language model, the improvement was around 0.1 points in all cases. Since OOV words and keywords have already been added and boosted by the interpolation, there was little room of improvement by additional adaptation with the cache model. We observed a similar tendency for keywords.

6. CONCLUSIONS

We have proposed an adaptation framework of language model for academic lectures, by using OCR results of presentation slides. In this paper, we first demonstrated that many of OOV words and keywords could be covered even with erroneous OCR results. Then, we proposed a filtering method to eliminate OCR errors using a morphological analyzer and a newspaper database. With the filtered OCR results and relevant newspaper articles, language model adaptation realized significant improvements of ASR performance.

7. ACKNOWLEDGMENT

This research work was partly supported by JST CREST and ERATO programs, and JSPS Grant-in-aid for scientific research.

8. REFERENCES

[1] P. Cerva, J. Silovsky, J. Zdansky, J. Nouza, and J. Malek, "Real-Time Lecture Transcription using ASR for Czech Hearing Impaired or Deaf Students," in *Proc. Interspeech*, 2012, pp. 3343–3344.

[2] J. Glass, T.J. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent Progress in the MIT Spoken Lecture Processing Project," in *Proc. Interspeech*, 2007, pp. 2553–2556.

[3] S. Togashi and S. Nakagawa, "A Browsing System for Classroom Lecture Speech," in *Proc. Interspeech*, 2008, pp. 2803–2806.

[4] R. Ranchal, T. Taber-Doughty, Y. Guo, K. Bain, H. Martin, J. Robinson, and B. Duerstock, "Using Speech Recognition for Real-Time Captioning and Lecture Transcription in the Classroom," *IEEE Trans. Learning Technologies*, vol. 6, no. 4, pp. 299–311, 2013.

[5] H. Liao, E. McDermott, and A. Senior, "Large Scale Deep Neural Network Acoustic Modeling with Semi-supervised Training Data for YouTube Video Transcription," in *Proc. ASRU*, 2013.

[6] P. Bell, H. Yamamoto, P. Swietojanski, Y. Wu, F. McInnes, C. Hori, and S. Renals, "A Lecture Transcription System Combining Neural Network Acoustic and Language Models," in *Proc. Interspeech*, 2013, pp. 3087–3091.

[7] A. Park, T. Hazen, and J. Glass, "Automatic Processing of Audio Lectures for Information Retrieval: Vocabulary Selection and Language Modeling," in *Proc. ICASSP*, 2005, vol. 1, pp. 497–500.

[8] Y. Akita, M. Watanabe, and T. Kawahara, "Automatic Transcription of Lecture Speech using Language Model Based on Speaking-Style Transformation of Proceeding Texts," in *Proc. Interspeech*, 2012, pp. 3343–3344.

[9] H. Yamazaki, K. Iwano, K. Shinoda, S. Furui, and H. Yokota, "Dynamic Language Model Adaptation Using Presentation Slides for Lecture Speech Recognition," in *Proc. Interspeech*, 2007, pp. 2349–2352.

[10] T. Kawahara, Y. Nemoto, and Y. Akita, "Automatic Lecture Transcription by Exploiting Presentation Slide Information for Language Model Adaptation," in *Proc. ICASSP*, 2008, pp. 4929–4932.

[11] J. Miranda, J. Neto, and A.W. Black, "Improving ASR by Integrating Lecture Audio and Slides," in *Proc. ICASSP*, 2013, pp. 8131–8134.

[12] A. Martínez-Villaronga, M.A. del Agua, J. Andrés-Ferrer, and A. Juan, "Language Model Adaptation for Video Lectures Transcription," in *Proc. ICASSP*, 2013, pp. 8450–8453.

[13] S. Wiesler, K. Irie, Z. Tüske, R. Schlüter, and H. Ney, "The RWTH English Lecture Recognition System," in *Proc. ICASSP*, 2014, pp. 3310–3314.

[14] R. Masumura, S. Hahm, and A. Ito, "Training a Language Model Using Webdata for Large Vocabulary Japanese Spontaneous Speech Recognition," in *Proc. Interspeech*, 2011, pp. 1465–1468.

[15] E. Cho, C. Fügen, T. Hermann, K. Kilgour, M. Mediani, C. Mohr, J. Niehues, K. Rottmann, C. Saam, S. Stuker, and A. Waibel, "A Real-World System for Simultaneous Translation of German Lectures," in *Proc. Interspeech*, 2013, pp. 3473–3477.

[16] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L.A. Rowe, "TalkMiner: A Search Engine for Online Lecture Video," in *Proc. ACM Multimedia*, 2010, pp. 1507–1508.

[17] G. Neubig, Y. Nakata, and S. Mori, "Pointwise Prediction for Robust, Adaptable Japanese Morphological Analysis," in *Proc. ACL-HLT*, 2011, pp. 529–533.

[18] T. Hofmann, "Probabilistic Latent Semantic Indexing," in *Proc. SIGIR*, 1999, pp. 50–57.

[19] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[20] R. Kuhn and R. De Mori, "A Cache-based Natural Language Model for Speech Recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 12, no. 6, pp. 570–583, 1990.

[21] S. Furui, K. Maekawa, and H. Isahara, "Toward the Realization of Spontaneous Speech Recognition —Introduction of a Japanese Priority Program and Preliminary Results—," in *Proc. ICSLP*, 2000, pp. 518–521.