



Semi-automated Update of Automatic Transcription System for the Japanese National Congress

Yuya Akita^{†‡} Masato Mimura[†] Graham Neubig[‡] Tatsuya Kawahara^{†‡}

[†] Academic Center for Computing and Media Studies, Kyoto University,

[‡] School of Informatics, Kyoto University,
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

Update of acoustic and language models is vital to maintain performance of automatic speech recognition (ASR) systems. To alleviate efforts for updating models, we propose a “semi-automated” framework for the ASR system of the Japanese National Congress. The framework consists of our speaking-style transformation (SST) and lightly-supervised training (LSV) approaches, which can automatically generate spoken-style training texts and labels from documents like meeting minutes. An experimental evaluation demonstrated that this update framework improved the ASR performance for the latest meeting data. We also address an estimation method of the ASR accuracy based on SST, which uses minutes as reference texts and does not require verbatim transcripts.

Index Terms: Spontaneous speech recognition, congressional speech, lightly-supervised training, speaking-style transformation

1. Introduction

Spontaneous speech recognition research has been extending its targets to a variety of public speaking such as academic presentations [1, 2], classroom lectures [3, 4] and congressional meetings [5]. We have been developing an ASR system for the National Congress (Diet) of Japan [6], which is deployed by the Congress for the new-generation editing system of meeting minutes.

These kinds of ASR systems need to catch up new topics and new people, by reflecting these new characteristics to acoustic and language models. However, updating such models in an ASR system is not easy, because of difficulties in preparing faithful transcripts for additional speech data collected during trials and/or operations. Update of running ASR systems is often conducted in spoken dialogue systems (SDS), but it is oriented for rapid development and adaptation of ASR at the earliest stage of operation. These can be intensively done, and the performance will usually saturate very quickly. On the other hand, topics and speakers in the congressional speech are changing over years. Moreover, room acoustic characteristics are often changed. The ASR system needs continuous maintenance with model update, but it is practically impossible to prepare faithful transcripts for that purpose.

The National Congress makes minutes for every meeting, however, they have been edited to enhance readability of the texts; fillers are completely removed and colloquial expressions are often replaced with formal expressions. Apparently, we cannot use minutes directly as training texts for language model nor training labels for acoustic model.

We have proposed a framework of the speaking-style transformation (SST) [7] which generates spoken-style N-gram statistics from document texts. This framework was successfully used for making the language model for the transcription system. Moreover, we applied SST to generate labels for acoustic model [8], realizing lightly-supervised training (LSV). In this paper, we demonstrate these schemes can be extended to continuous model updating. Since the SST does not require any parameters to be tuned by hand, the update process is “semi-automated,” namely, the update process can be performed automatically, once hand-edited documents are provided with the corresponding speech. We also address estimation of ASR accuracy without faithful transcripts for continuous monitoring of the system performance.

2. Semi-automated System Update

2.1. Overall Framework

Figure 1 shows an overview of the proposed updating framework applied to ASR of the National Congress meetings. For every meeting, speech is transcribed by the ASR system. Then, the ASR output is edited by professional stenographers in the Congress (the House of Representatives), so that resulting minutes meet the strict orthographic standard of the House. This edit includes disfluency removal and replacement of expressions, as well as correction of ASR errors, but does not include any summarization of the content. Once the minutes and corresponding speech are archived, then update is performed for acoustic and language models by applying SST and LSV, which are described below.

The Japanese National Congress has typically two or three sessions in a year. A session lasts two months to six months, and during a session, several meetings take place almost every day. Therefore, we cannot suspend the ASR system nor even replace models during a session.

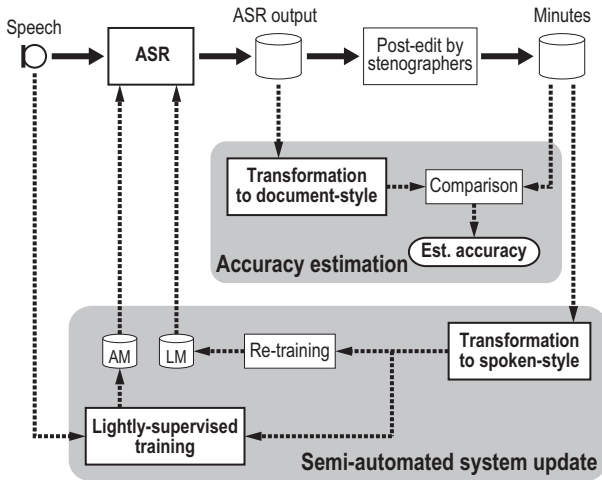


Figure 1: Proposed update framework

We suppose that update is performed at every interval of sessions using archived data of the previous session.

2.2. Speaking-Style Transformation (SST) for Language Model

The speaking-style transformation method [7] is based on the framework of statistical machine translation [9], where sentence V of the target language is generated from sentence W of the source language, which maximizes posterior probability $P(V|W)$ based on Bayes' rule.

$$P(V|W) = \frac{P(W|V)P(V)}{P(W)} \quad (1)$$

In this work, we consider document-style and spoken languages as different ones, denoted by W and V , respectively, and estimate spoken language model $P(V)$, which is formulated as Equation (2) by rewriting Equation (1).

$$P(V) = P(W) \frac{P(V|W)}{P(W|V)} \quad (2)$$

The conditional probabilities $P(V|W)$ and $P(W|V)$, i.e., transformation model, can be estimated using a parallel aligned corpus of faithful transcripts and their corresponding meeting minutes. For N-gram language model, transformation is actually performed on N-gram occurrence counts (N_{LM}).

$$N_{LM}(v) = N_{LM}(w) \frac{P(v|w)}{P(w|v)} \quad (3)$$

Here, w and v are individual patterns that are transformed and $N_{LM}(w)$ and $N_{LM}(v)$ are N-gram entries including them. Transformation patterns w and v contain preceding and following words as contexts. To alleviate the data sparseness problem, part-of-speech (POS) contexts are also introduced [7]. Using the estimated N-gram entries and occurrence counts, the spoken-style language model is trained in a standard manner.

2.3. Lightly-Supervised Training (LSV) of Acoustic Model

SST generates spoken-style N-gram statistics from minutes, however, it cannot recover the original verbatim transcript which is necessary for acoustic model training. Therefore, in our LSV method [8], we generate a dedicated language model by SST for decoding every speech (=speaker turn). As a result of ASR with this model, we can obtain a verbatim transcript for the speech with high accuracy, which can be used to train the acoustic model.

The procedure of LSV is as follows. First, for each speaker turn of the speech, we compute N-gram counts from the corresponding text segment in the minutes. The N-gram entries and counts are then converted to the spoken style using SST. The model is very constrained and still expected to predict spontaneous phenomena such as filler insertion. Then, ASR is conducted using the dedicated model to produce a verbatim transcript. Finally, the best phone hypothesis produced by ASR is used as phone labels for standard HMM training. For discriminative training such as the minimum phone error (MPE) criterion, we also generate competing hypotheses using a baseline language model.

2.4. System Update

As a baseline system, the system we reported in [6] was slightly revised for meetings of the 172nd and 173rd sessions¹ in 2009. The language model is word trigram model. We used texts of 168M words from all minutes of sessions 145–171 (years 1999–2009) for training. These texts were transformed into spoken-style language model by SST. The size of vocabulary is 64K. The acoustic model is triphone HMM trained with the MPE criterion [10]. For training of this model, we used speech data of 225 hours collected from meetings held in years 2003–2007.

In the summer of 2009, the general election of the House of Representatives was called after the 171st session, and resulted in replacement of more than a hundred members. The government was also changed at the 172nd session, i.e., the prime minister and cabinet members were replaced entirely. Consequently, the baseline system does not cover these new members well.

The baseline system was updated for the 174th session (starting January 2010), by using meeting data of the 172nd and the 173rd sessions (September–December, 2009). Specifically, speech of 95 hours was collected from meetings in the 173rd session for acoustic model training. Corresponding minutes of the speech were transformed by SST to spoken-style language model, and then phone labels were generated by decoding the speech with this model. The acoustic model was re-trained using these phone labels, together with the 225-hour fully-transcribed speech used in the baseline model. For lan-

¹The 172nd session is a special session only to elect a new prime minister. The duration of this session is four days, thus, these two sessions are jointly handled.

Table 1: Improvements of word and character accuracy by the model update

Systems	Word		Character	
	Corr.	Acc.	Corr.	Acc.
Baseline (for 172nd/173rd)	82.5%	79.4%	88.0%	85.5%
AM adaptation	82.0%	78.7%	87.6%	85.0%
AM re-training	83.5%	80.5%	88.9%	86.5%
LM re-training	82.6%	79.5%	88.1%	85.6%
AM&LM re-training (for 174th)	83.6%	80.6%	89.0%	86.6%

guage model training, texts of 1.3M words from all minutes of the 172nd and the 173rd sessions were simply merged into the 168M-word training texts of the baseline model. Resulting 170M-word texts were transformed by SST to new statistics, which were then used to re-train the language model.

2.5. Experimental Evaluation

We evaluated the effect of the update by using latest meeting speech of the Congress. Three committee meetings of the 174th session in 2010 were selected for a test set. The total number of words and characters in the test set is 123,405 and 230,979, respectively. The out-of-vocabulary rate was 0.48% by the baseline model, and was unchanged by the updated model, although 310 words were newly added to the latter through the update process. As a decoder, our Julius [11] rev.4.1 is used.

Table 1 shows word-based and character-based correctness and accuracy by the baseline system and the updated systems. For comparison, we also built an acoustic model which was adapted from the baseline model by MLLR using the 95-hour data with the same labels. The adapted acoustic model did not improve ASR performance; instead it degraded slightly. MLLR adaptation might have distorted the model trained by MPE. The re-training scheme made the model consistent, and fully exploited the additional data. It achieves relative reduction of word and character errors by 5.1% and 6.5%, respectively. Relative error reduction by updating language model was 0.6% for both words and characters. By combining re-trained acoustic and language models, we finally obtained 5.9% and 7.3% of relative word and character error reduction. This result demonstrates that the proposed framework successfully worked to improve ASR performance for the latest session.

3. Accuracy Estimation using Minutes

When running an ASR system, monitoring its performance is essential for maintenance. Performance measurement is also useful for selection of training data. In fact, discussions are occasionally thrown into confusion,

and such portions are sometimes excluded even from the official record. These segments of disorders will be harmful for the update of the system and should be removed from training data, since ASR tends to be inaccurate there.

However, exact accuracy level cannot be calculated, because we do not have access to verbatim transcripts. Here, we propose an alternative method to estimate accuracy of ASR transcripts by making use of minutes. As shown in Figure 1, we apply SST [12] to convert each ASR transcript into the document style, then compare it with corresponding minutes.

3.1. Style Transformation for Minutes

The document-style transformation method is also based on SMT. Here, languages W and V in Equation (1) are swapped:

$$P(W|V) = \frac{P(V|W)P(W)}{P(V)}, \quad (4)$$

hence, the formulation can be denoted as:

$$\hat{W} = \arg \max_W P(V|W)P(W) \quad (5)$$

where $P(V|W)$ is a transformation model which defines possible transformation from the spoken style to the document style, and $P(W)$ is a document-style language model which ensure the appropriateness of transformed word sequence, and can be simply trained with the minutes.

As a transformation model, we introduce word-based conditional probabilities which take into account both document- and spoken-style word histories $\{w\}$ and $\{v\}$:

$$P(V|W) \approx \prod_{i=1}^K P(v_i|v_1, \dots, v_{i-1}, w_1, \dots, w_K). \quad (6)$$

Actually, the context length in Equation (6) is limited to three. These probabilities can be estimated by using a parallel aligned corpus. Moreover, we extend the simple noisy-channel model to a log-linear model which can take into account joint probabilities and additional features [12].

For input word sequence $\{w_i\}$, possible hypotheses are generated and the best one is searched with a decoder based on weighted finite-state transducer (WFST).

3.2. Experimental Evaluation

For training of SST models, we used meeting minutes of 158M words and the parallel corpus of 2.8M words. The test set and the ASR system were different from those in Section 2. We prepared three committee meetings as the testing data. There are 332 speaker turns in total. The total numbers of words and characters in verbatim transcripts of the test set are 77,007 and 126,335, respectively. Minutes contain 71,115 words and 115,693 characters in total, which are 8% smaller than those in the verbatim transcripts because of the edit by stenographers.

Table 2: Estimated and actual statistics

	Actual w/ transcripts	Estimation w/ minutes	Differ- ence	Corre- lation
Word corr.	84.6%	84.9%	+0.3%	0.92
Word acc.	82.5%	80.7%	-1.8%	0.88
Char. corr.	87.2%	86.6%	-0.6%	0.88
Char. acc.	85.3%	82.6%	-2.7%	0.88

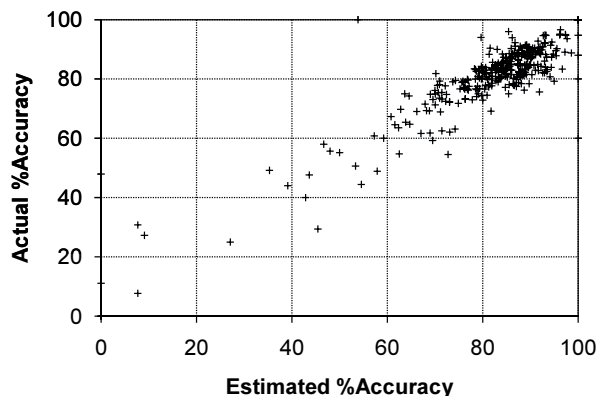


Figure 2: Correlation between estimated and actual character accuracy

Average accuracy and correctness of words and characters are listed in Table 2. Actual statistics were calculated for raw ASR results by comparing with faithful transcripts, while estimated statistics were derived for transformed ASR results by comparing with minutes. As for word and character correctness, the SST-based estimation predicted close values to the actual correctness. The proposed method underestimated word and character accuracy, because some inserted portions were beyond word-level and not predictable by SST. The difference between actual and estimated character accuracy is 2.7% on the average.

We also calculated accuracy and correctness for each speaker turn, then correlation coefficients between estimated and actual values, as shown in Table 2. Figure 2 shows correlation between estimated and actual character accuracy. For 332 speaker turns in the test set, high correlation was achieved for all measures. These results demonstrate that the estimated accuracy can be useful for data selection based on speaker turns.

4. Conclusions

We have presented a semi-automated updating framework for the ASR system of the National Congress. The key techniques in acoustic and language modeling are SST and LSV which automatically produce training data for the update. Experimental results demonstrated that the update process successfully improved ASR performance. We also addressed accuracy estima-

tion of ASR results using SST, and it was confirmed by experiments that the transformation-based comparison could estimate correctness of ASR. Though this paper described a framework for ASR of congressional speech, this will be applicable to different domains such as lectures.

5. Acknowledgements

This work was supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

6. References

- [1] L. Lamel, G. Adda, E. Bilinski, and J.L. Gauvain, "Transcribing Lectures and Seminars," in *Proc. Eurospeech*, 2005, pp. 1657–1660.
- [2] H. Nanjo and T. Kawahara, "Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition," *IEEE Trans. Speech & Audio Proc.*, vol. 12, no. 4, pp. 391–400, 2004.
- [3] 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.
- [4] 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.
- [5] L. Lamel, J.-L. Gauvain, G. Adda, C. Barras, E. Bilinski, O. Galibert, A. Pujol, H. Schwenk, and X. Zhu, "The LIMSI 2006 TC-STAR EPPS Transcription Systems," in *Proc. ICASSP*, 2007, vol. 4, pp. 997–1000.
- [6] Y. Akita, M. Mimura, and T. Kawahara, "Automatic Transcription System for Meetings of the Japanese National Congress," in *Proc. Interspeech*, 2009, pp. 84–87.
- [7] Y. Akita and T. Kawahara, "Statistical Transformation of Language and Pronunciation Models for Spontaneous Speech Recognition," *IEEE Trans. Audio, Speech, & Language Proc.*, (accepted for publication).
- [8] T. Kawahara, M. Mimura, and Y. Akita, "Language Model Transformation Applied to Lightly Supervised Training of Acoustic Model for Congress Meetings," in *Proc. ICASSP*, 2009, pp. 3853–3856.
- [9] P. Brown, S. Pietra, V. Pietra, and R. Mercer, "The Mathematics of Statistical Machine Translation: Parameter Estimation," *Computational Linguistics*, vol. 19, no. 2, pp. 263–311, 1993.
- [10] D. Povey and P.C. Woodland, "Minimum Phone Error and I-smoothing for Improved Discriminative Training," in *Proc. ICASSP*, 2002, vol. 1, pp. 105–108.
- [11] A. Lee and T. Kawahara, "Recent Development of Open-Source Speech Recognition Engine Julius," in *Proc. AP-SIPA*, 2009, pp. 131–137.
- [12] G. Neubig, Y. Akita, S. Mori, and T. Kawahara, "Improved Statistical Models for SMT-based Speaking Style Transformation," in *Proc. ICASSP*, 2010, pp. 5206–5209.