Automatic Transcription of Parliamentary Meetings and Classroom Lectures
– A Sustainable Approach and Real System Evaluations –

Tatsuya Kawahara
School of Informatics, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
Email: kawahara@i.kyoto-u.ac.jp

***Abstract*—Applications of automatic speech recognition (ASR) have been extended to a variety of tasks and domains, including spontaneous human-human speech. We have developed an ASR system for the Japanese Pariliament (Diet), which is deployed this year. By exploiting official records made by human stenographers, we have realized an efficient training scheme of acoustic and language models, which does not require faithful transcripts and thus is scalable to enormous data. Evaluation results of the semi-automated model update are presented. We are also working on an ASR system for classroom lectures, which is intended for assisting hearing impaired students. As the classroom lectures in universities are very technical, efficient adaptation methods of acoustic and language models are investigated. A trial of real-time captioning for a hearing impaired student in our university is reported.**

I. Introduction

Applications of automatic speech recognition (ASR) have been extended to a variety of tasks and domains, including spontaneous human-human speech [1]. A number of projects have been conducted for the tasks of oral presentations [2][3], classroom lectures [4][5], telephone conversations [6][7], in-house meetings [8][9], and parliamentary meetings [10][11]. In many of these projects, successful ASR performance is obtained once a large corpus of typically hundreds of hours is constructed.

On the other hand, it is not a practical assumption that we can prepare such a large-scale training database for every application. Even if we can prepare the initial training database, the acoustic and language models should be updated to reflect the change of speakers and topics. Therefore, one of the most fundamental problems in this kind of spontaneous speech recognition is preparation of a sufficient amount of matched training data to cover wide variation of the acoustic and linguistic characteristics in the target task domain [1].

To be accurate, there is a huge scale of speech data in real world; meetings, lectures and conversations are made every day, anywhere in the world. The real problem is that we need faithful transcripts for the speech, but it is too costly to prepare them because it involves manual transcription of utterances with many disfluencies, compared to the reading of prepared materials. Therefore, it is vitally important to explore a new ASR scheme which does not rely on supervised training as in the conventional manner.

We have been working on an automatic transcription (ASR) system for the Japanese Parliament (Diet) [11]. The goal of the system is to replace the conventional manual short-hand scheme for efficient generation of official meeting records. Note that the stenographers are not totally replaced by the system, but they will be engaged in correction and post-editing of the ASR outputs. The system is now deployed in the House of Representatives (Lower House) of the Diet, and will be operated soon. We have constructed a corpus of the meeting speech and its faithful transcripts of about 300 hours to train the acoustic and language models. On the other hand, meetings are held almost every day, which total up to approximately 2500 hours in a year. We should make use of them, not only to make better baseline acoustic and language models, but also to update the models periodically to reflect the change of speakers and topics. Although faithful transcripts are not made for them, there are official meeting records made by stenographers. By exploiting them, we can realize a "sustainable" system, which can evolve, i.e. update/re-train the models, only with speech and text generated during the system operation. To this end, we have proposed a scheme of language model transformation [12][13][14], which generates a language model and also (training labels for) an acoustic model, without extra human interventions. In this paper, following the analysis of the differences between faithful transcripts and official meeting records in Section II, the proposed scheme of language model transformation is reviewed in Section III. Then, the evaluation in the real system setting is reported in Section IV.

We are also working on an automatic transcription (ASR) system for classroom lectures. The goal of the system is to assist hearing impaired students by partially replacing the current hand-writing or PC-typing scheme which is operated by two or more volunteer persons. [1] While lectures given at universities are very technical, a course of lectures is usually given by same professors for a long time period. Thus, we should also design a "sustainable" scheme in this scenario. Preliminary efforts and system evaluations are reported in Section V.

[1]Real-time typing is not easy for Japanese language, because we need to convert kana (phonetic) symbols to kanji (Chinese) characters.

ISBN 978-1-4244-6245-2/10/$26.00 ©2010 IEEE

TABLE I
ANALYSIS OF EDITS MADE BY STENOGRAPHERS FOR OFFICIAL RECORDS

edit type	category	ratio
Deletion	Fillers (ex.) *maa, ano:*	50.1%
	Discourse markers (ex.) *desu-ne, ano*	23.5%
	*Repair	3.0%
	*Word fragments	2.8%
	Syntax correction (ex.) *wo(suru)*	1.8%
	Extraneous expressions	1.7%
Insertion	Function words (ex.) *wo, wa, (te)i(ru)*	7.8%
Substitution	Colloquial expressions (ex.) *dewa/ja, keredomo/kedomo*	6.4%
*Reordering		1.3%

* means not-simple edits, which are hard to model with WFSTs.

II. DIFFERENCES BETWEEN "ACTUALLY" SPOKEN LANGUAGE AND "NORMALLY" TRANSCRIBED TEXT

In any languages universally, there is a difference between spoken style and written style. [2] Moreover, spontaneous speech essentially includes disfluency phenomena such as fillers and repairs as well as redundancy such as repeats and redundant words. Therefore, faithful transcripts of spontaneous speech are not good in terms of readability and documentation. As a result, there is a significant mis-match between faithful transcripts and official meeting records of the Diet because of the editting process by stenographers. Similar phenomena are observed in TV programs (between spoken utterances and their captions) and public speaking (between actual speeches and scripts).

Unlike European Parliament Plenary Sessions (EPPS), which were targeted by the TC-Star project [10][15], the great majority of sessions in the Japanese Diet are in committee meetings. They are more interactive and thus spontaneous, compared to plenary sessions. This results in many differences between the actual utterances and the official meeting records. It is observed that the difference in the words (edit distance or word error rate) between the two transcripts ranges 10-20% (15.5% on average). Table I lists the breakdown of the edits made by stenographers. The analysis is made for transcripts of eight meetings, which consist of 380K words in total. Among them, approximately a half are by deletion of fillers. The rests are not so trivial, but their majority are by deletion of redundant words such as discourse markers and extraneous words. When we measure the word error rate, deletion of non-fillers accounts more than fillers, because they often involve more than one word. It should also be noted that most of the edits (around 90%) can be modeled by simple substitution/deletion/insertion of a word (or two words), which can be dealted with WFSTs (Weighted Finite State Transducers).

III. SCHEME OF LANGUAGE MODEL TRANSFORMATION

We have proposed a scheme of language model transformation to cope with the differences between spontaneous utterances (verbatim text: V) and human-made transcripts (written-style text: W) [12][13][14]. In this scheme, the two are regarded as different languages and statistical machine translation (SMT) is applied (Fig. 1). It can be applied in both directions: to clean a faithful transcript of the spoken utterances to a document-style text, and to recover the faithful transcript from the human-made text.

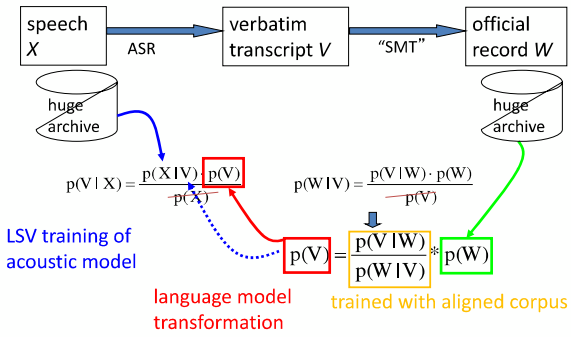

Fig. 1. Overview of the proposed scheme

The decoding process is formulated in the same manner as SMT, which is based on the following Bayes' rule.

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

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

Here the denominator is usually ignored in the decoding.

We have extended the simple noisy channel model to log-linear model which can incorporate joint probability $p(W, V)$ and contextual information [16], for the task of cleaning transcripts (eq. 1).

A. Estimation of Verbatim-style Language Model

On the other hand, the process to uniquely determine V (eq. 2) is much more difficult than the cleaning process (eq. 1) because there are more arbitrary choices in this direction; for example, fillers can be randomly inserted in (eq. 2) while all fillers are removed in (eq. 1). Therefore, we are more interested in estimating the statistical language model of V, rather than recovering the text of V. Thus, we derive the following estimation formula.

$$p(V) = p(W)\cdot\frac{p(V|W)}{p(W|V)} \quad (3)$$

The key point of this scheme is that the available text size of the document-style texts W is much larger than that of the verbatim texts V needed for training ASR systems. For the Diet meetings, we have huge archives of official meeting records. Therefore, we fully exploit their statistics $p(W)$ to estimate the verbatim-style language model $p(V)$ for ASR (Fig. 1 right-hand side).

[2] In Japanese language, the difference is very large, for example, moods of verbs are totally different.

The transformation is actually performed on occurrence counts of N-grams as below.

$$N_{gram}(v_1^n) = N_{gram}(w_1^n) \cdot \frac{p(v|w)}{p(w|v)} \quad (4)$$

Here v and w are individual transformation patterns. We model substitution $w \rightarrow v$, deletion of w, and insertion of v, by considering their contextual words. [3] $N_{gram}(w_1^n)$ is an N-gram entry (count) including them, thus to be revised to $N_{gram}(v_1^n)$. Estimation of the conditional probabilities $p(v|w)$ and $p(w|v)$ requires an aligned corpus of verbatim transcripts and their corresponding document-style texts. We have constructed the "parallel" corpus by using a part of the official records of the Japanese Diet meetings. The conditional probabilities are estimated by counting the corresponding patterns observed in the corpus. Their neighboring words are taken into account in defining the transformation patterns for precise modeling. For example, an insertion of a filler "ah" is modeled by $\{w = (w_{-1}, w_{+1}) \rightarrow v = (w_{-1}, ah, w_{+1})\}$, and the N-gram entries affected by this insertion are revised. A smoothing technique based on POS (Part-Of-Speech) information is introduced to mitigate the data sparseness problem. Please refer to [13] for implementation details and evaluation.

B. Lightly Supervised Training of Acoustic Model

The language model transformation scheme is also applied to lightly supervised training (LSV) of an acoustic model [14]. For the Diet meetings, we have large archives of speech which are not faithfully transcribed but have edited texts in the official records. Since it is not possible to uniquely recover the original verbatim transcripts from texts of the official records, as mentioned in the previous sub-section, we generate a dedicated language model for decoding the speech using the corresponding text segment of the official record. As a result of ASR, we expect to obtain a verbatim transcript with high accuracy. (Fig. 1 left-hand side).

For each turn (typically ten seconds to three minutes, and on the average one minute) of the meetings, we compute N-gram counts from the corresponding text segment of the official record. Here, we adopt a turn as a processing unit, because the whole session (typically two to five hours) is too long, containing a variety of topics and speakers. The N-gram entries and counts are then converted to the verbatim style using the transformation model (eq. 4). Here we count up to trigrams. Insertion of fillers and omission of particles in the N-gram chain are also modeled considering their context in this process. Then, ASR is conducted using the dedicated model to produce a verbatim transcript. The model is very constrained and still expected to accurately predict spontaneous phenomena such as filler insertion. It is also compact compared with the former methods of lightly supervised training [17][18][19][20], which interpolate the cleaned text with the baseline language model, resulting in a very large model.

[3]Unlike ordinary SMT, permutation of words is not considered in this transformation.

With the proposed method applied to turn-based segments, we can get character accuracy of 94.3% (baseline ASR 83.1%) for the additional training data set, which is used for re-training of the acoustic model. The best phone hypothesis is used as the label for the standard HMM training based on ML (Maximum Likelihood) criterion. For discriminative training such as MPE (Minimum Phone Error) criterion, we also generate competing hypotheses using the baseline language model.

IV. Evaluation of Semi-automated Update of ASR System for Parliamentary Meetings

The baseline ASR system for the Diet meetings [11] is as follows: The language model is a word trigram model. We used texts of 168M words from all official meeting records of sessions 145–171 (years 1999–2009) for training. These texts were transformed into verbatim-style language model. The size of vocabulary is 64K. The acoustic model is triphone HMMs trained with the MPE criterion. For training of this model, we used speech data of 225 hours collected from the 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 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 official meeting records of the speech were transformed to verbatim-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 language model training, texts of 1.3M words from all official meeting records 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 converted to new statistics, which were then used to re-train the language model.

We evaluated the effect of the update by using new test data. 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 through the update process.

Table II 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 (Maximum Likelihood Linear Regression) using the 95-hour data with the same labels. The adapted acoustic model did not improve ASR

TABLE II
IMPROVEMENT OF ASR ACCURACY FOR THE DIET MEETINGS BY SEMI-AUTOMATED MODEL UPDATE

Systems	Word		Chacacter	
	Corr.	Acc.	Corr.	Acc.
Baseline	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	83.6%	80.6%	89.0%	86.6%

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 scheme successfully worked to improve ASR performance.

Note again that the scheme can be applied forever, to be scalable to thousands of hours speech and to keep up with the change of speakers and topics, without requiring extra manual transcription.

V. ASR SYSTEM FOR ASSISTING HEARING IMPAIRED STUDENTS IN CLASSROOMS

Recently, more and more hearing impaired students are admitted to colleges and universities. It is imperative that schools provide necessary means to these students so that they can study just as alike as non-handicapped students. Conventionally, possible solutions are among sign language, PC captioning, and hand-writing. Currently in colleges and universities in Japan, hand-writing is most widely-used for note-taking for hearing impaired students, because it is the easiest to learn and deploy, while some schools adopt PC keyboard-typing. Note-taking is conducted by student volunteers, not professional stenographers because of the budget problem. And real-time transcription of lectures is so difficult and stressing that it is widely known that only 20-30% of utterances can be transcribed even with two volunteers engaged in turn in a lecture.

Moreover, many lectures at universities are so technical that "out-of-field" volunteers cannot catch the content or technical words, for example, engineering students cannot help medical students. Actually in our university, note-takers for lectures in higher grades even in the undergraduate can be collected only from the senior students of the same department, but it is very difficult to get a sufficient number of volunteers for all time slots. Therefore, the ASR technology is expected to provide an alternative solution to assist note-taking for hearing impaired students.

An overview of the proposed system is depicted in Fig. 2. The lecturer's speech is input to a microphone. Usually in

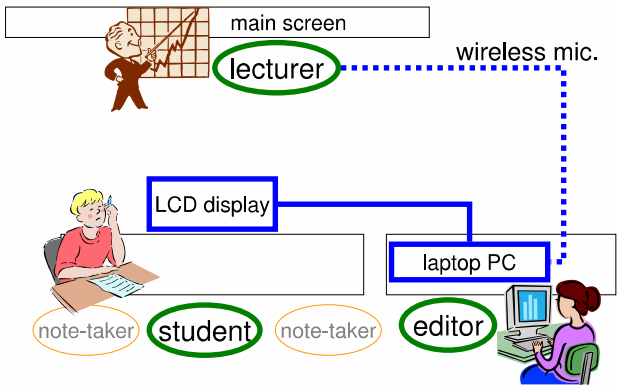

Fig. 2. System overview (human note-takers are present during the trial)

classrooms, lecturers want to move around with their hands free, so we adopt a wireless pin microphone. Speech is transmitted to a receiver in the same room, and then fed into a PC the system resides. The input speech is processed by the ASR system which is adapted to the lecture. Specifically, we adapt the acoustic model to the speaker, and the language model to the content of the lecture as well as to the speaker. The system is based on our open-source ASR engine Julius [21] [4], which anyone can easily install for free. The ASR output is generated by the utterance unit, segmented by a long pause, and given to the post-editing system, with which selection and correction is conducted. Final presentation is done through another free software program IPtalk [5], which is widely used in Japan as a PC captioning program for hearing impaired people.

A. Adaptation of ASR Models

For real-time transcription, or efficient decoding with high accuracy, we should prepare compact acoustic and language models matched to lectures. Thus, we adopt a scheme to adapt acoustic and language models to every course and lecturer offline. Normally, one lecture course lasts a dozen of weeks, and it is often the case that the course is taught by the same lecturer for many years.

Thus, we investigate effective methods of ASR model adaptation. If we can access to digital text media of the textbooks or slides used in the lecture, we can exploit it for adaptation of the language model. We proposed several adaptation methods based on PLSA and relevant Web texts [5]. They are effective, especially in improving keyword detection accuracy, but are limited by nature because of the small size of matched texts.

A more effective but costly method is to use speech data given by the same lecturers, for example, in previous lectures. It is possible to record up to dozens of hours of lectures, and supervised adaptation using them would drastically improve the ASR accuracy as shown in [4]. However, it is not a

[4] http://julius.sourceforge.jp/
[5] http://iptalk.hp.infoseek.co.jp/

practical assumption, at least in terms of budget, to prepare manual transcripts of such a large amount of data for every course. And we need to operate the system even before a large amount of data is collected.

In the proposed system, the ASR results are selected and corrected by a human editor. As the corrected text is usually cleaned and shortened, it is not a faithful transcript of the utterance, thus not suitable for model adaptation. Moreover, many ASR results are discarded if they contain too many errors or do not contain meaningful content. Still, we can make use of the information of human selection and correction. Here we assume that the human editor selects ASR results based on the word accuracy, and thus we use only selected texts for the model adaptation.

The naive unsupervised adaptation which uses all ASR results as they are realizes a modest improvement in accuracy. We can conduct another way of unsupervised adaptation by filtering the ASR results based on the confidence measure (CM) of the recognizer. But the CM-based selection is not effective at all, while the oracle selection brought improvement in keyword detection. However, the improvement is not so significant as the case using manual transcripts or supervised adaptation. The cost of manual transcription can be reduced by using the information of human editor's selection. If we use 50% of the ASR results selected by oracle (human editor), and manually transcribe the other half, we can achieve almost comparable performance to the supervised adaptation with the entire manual transcripts. Note again that the selection is manual, but naturally included in the system operation.

B. Field Trials of the System

We conducted trials of the system in real lectures of a course on material science for civil engineering in our university. A hearing impaired student attendeding this course was assisted by two human note-takers, who sat down next to the student and wrote down the content of the lecture by hand. In the trials, we set up our system nearby and placed an LCD screen in fron of the student. Thus, he could see either the screen or the paper notes. The experimental scene is just as described in Fig. 2.

The amount of texts transcribed and shown to the student is compared in Table III. We compute the ratio of the number of output words against that of all uttered words. The set of keywords are defined as content words that appeared in the slide text used in the lecture. [6] It is shown that the amount of the texts made by our system and one human editor is significantly larger than the texts transcribed by two persons in cooperation. The hand-written texts constituted less than 20% of the original utterances, as often pointed out in Japanese. The ASR-based system restored 30-40% of the content, which might be comparable to well-trained type-writers. But most of the people cannot type-in for so long without break, so usually two or four persons are necessary for a lecture. On the other hand, our system could be operated by a single person for 90 minutes. This is a clear advantage of the system. The result also suggests that if the system was operated by two persons in cooperation, most of the content could be presented.

[6]Many of the uttered words are redundant or non-sense, so even the perfect note-taking would make much smaller than 100%.

TABLE III
COMPARISON OF AMOUNT OF PRESENTED TEXTS IN A CLASSROOM LECTURE

	words	keywords
hand-writing note-takers	16.4%	16.4%
ASR-based system	29.3%	42.9%

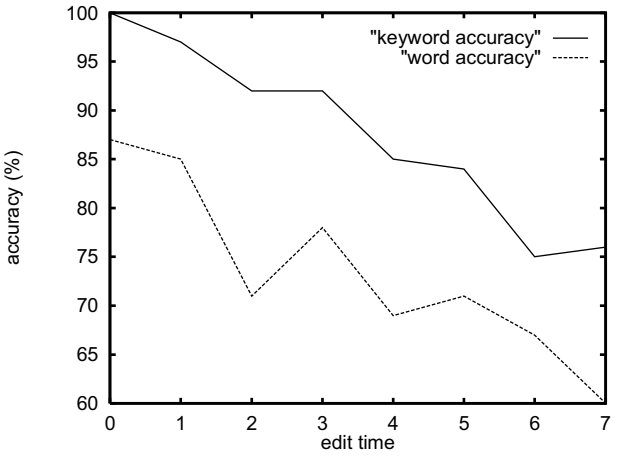

Fig. 3. Distribution of correction time in terms of ASR accuracy

We also investigate the latency time caused by the post-editing, as the ASR itself was performed almost in real time. The system records the exact time (1) when it receives the ASR output, (2) when the human editor selects utterances for correction, and (3) when the editor finishes the correction and outputs the text for presentation. The average time for selecting texts ((2)-(1)) was 4.07 seconds and it is not so correlated with the ASR accuracy. The average time for correcting texts ((3)-(2)) was 4.84 seconds, and distributions of the correction time and the ASR accuracy is shown in Fig. 3. We can see the general tendency that more time is needed when the ASR accuracy is lower. We can expect prompt output when the word accuracy exceeds 80%.

In order to get a feedback by the actual user, we asked the university staff interview the student after the lectures. His overall impression was the system generated significantly more content than the current note-takers, but he would like much faster output though the delay by our system is not so bad as the current manual scheme. Therefore, it is foremost important to improve the ASR accuracy.

VI. CONCLUSIONS

This paper addresses a "sustainable" approach for ASR systems, which can evolve without requiring manual transcription. The proposed scheme is not totally unsupervised, but assumes the final output which is not a faithful transcript. This assumption holds in many speech transcription applications. In many cases, however, the relationship between the actually

spoken utterances and the available resource may not be strong and thus its modeling may not be straightforward.

Still, the semi-supervised training scheme is one of the most important topics in ASR so that it can be applied to a variety of real-world applications, since an enormous amount of speech data are there.

Acknowledgment

A part of the works presented in this article is indebted to Dr. Yuya Akita, Mr. Masato Mimura and Mr. Graham Neubig. This work is supported by JST CREST and JSPS Grant-in-Aid for Scientific Research.

References

[1] Sadaoki Furui and Tatsuya Kawahara. Transcription and distillation of spontaneous speech. In J.Benesty, M.M.Sondhi, and Y.Huang, editors, *Springer Handbook on Speech Processing and Speech Communication*, pages 627–651. Springer, 2008.

[2] 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*, volume 3, pages 518–521, 2000.

[3] H.Nanjo and T.Kawahara. Language model and speaking rate adaptation for spontaneous presentation speech recognition. *IEEE Trans. Speech & Audio Process.*, 12(4):391–400, 2004.

[4] 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*, pages 2553–2556, 2007.

[5] T.Kawahara, Y.Nemoto, and Y.Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *Proc. IEEE-ICASSP*, pages 4929–4932, 2008.

[6] S.Matsoukas, J.-L.Gauvain, G.Adda, T.Colthurst, Chia-Lin Kao, O.Kimball, L.Lamel, F.Lefevre, J.Z.Ma, J.Makhoul, L.Nguyen, R.Prasad, R.Schwartz, H.Schwenk, and B.Xiang. Advances in transcription of broadcast news and conversational telephone speech within the combined EARS BBN/LIMSI system. *IEEE Trans. Audio, Speech & Language Process.*, 14(5):1541–1556, 2006.

[7] S.F.Chen, B.Kingsbury, L.Mangu, D.Povey, G.Saon, H.Soltau, and G.Zweig. Advances in speech transcription at IBM under the DARPA EARS program. *IEEE Trans. Audio, Speech & Language Process.*, 14(5):1596–1608, 2006.

[8] J.Fiscus, J.Ajot, and J.S.Garofol. The rich transcription 2006 evaluation overview and speech-to-text results. In *NIST Meeting Recognition Workshop*, 2006.

[9] S.Renals, T.Hain, and H.Bourlard. Recognition and understanding of meetings: The AMI and AMIDA projects. In *Proc. IEEE Workshop Automatic Speech Recognition & Understanding*, 2007.

[10] C.Gollan, M.Bisani, S.Kanthak, R.Schluter, and H.Ney. Cross domain automatic transcription on the TC-STAR EPPS corpus. In *Proc. IEEE-ICASSP*, volume 1, pages 825–828, 2005.

[11] Y.Akita, M.Mimura, and T.Kawahara. Automatic transcription system for meetings of the Japanese national congress. In *Proc. INTERSPEECH*, pages 84–87, 2009.

[12] Y.Akita and T.Kawahara. Efficient estimation of language model statistics of spontaneous speech via statistical transformation model. In *Proc. IEEE-ICASSP*, volume 1, pages 1049–1052, 2006.

[13] Y.Akita and T.Kawahara. Statistical transformation of language and pronunciation models for spontaneous speech recognition. *IEEE Trans. Audio, Speech & Language Process.*, 18(6):1539–1549, 2010.

[14] T.Kawahara, M.Mimura, and Y.Akita. Language model transformation applied to lightly supervised training of acoustic model for congress meetings. In *Proc. IEEE-ICASSP*, pages 3853–3856, 2009.

[15] B.Ramabhadran, O.Siohan, L.Mangu, G.Zweig, M.Westphal, H.Schulz, and A.Soneiro. The IBM 2006 speech transcription system for European parliamentary speeches. In *Proc. INTERSPEECH*, pages 1225–1228, 2006.

[16] G.Neubig, Y.Akita, S.Mori, and T.Kawahara. Improved statistical models for SMT-based speaking style transformation. In *Proc. IEEE-ICASSP*, pages 5206–5209, 2010.

[17] L.Lamel, J.Gauvain, and G.Adda. Investigating lightly supervised acoustic model training. In *Proc. IEEE-ICASSP*, volume 1, pages 477–480, 2001.

[18] H.Y.Chan and P.Woodland. Improving broadcast news transcription by lightly supervised discriminative training. In *Proc. IEEE-ICASSP*, volume 1, pages 737–740, 2004.

[19] L.Nguyen and B.Xiang. Light supervision in acoustic model training. In *Proc. IEEE-ICASSP*, volume 1, pages 185–188, 2004.

[20] M.Paulik and A.Waibel. Lightly supervised acoustic model training EPPS recordings. In *Proc. INTERSPEECH*, pages 224–227, 2008.

[21] A.Lee and T.Kawahara. Recent development of open-source speech recognition engine Julius. In *Proc. APSIPA ASC*, pages 131–137, 2009.