Transcription System Using Automatic Speech Recognition for the Japanese Parliament (Diet)

Tatsuya Kawahara
Kyoto University, Academic Center for Computing and Media Studies
Sakyo-ku, Kyoto 606-8501, Japan

Abstract

This article describes a new automatic transcription system in the Japanese Parliament which deploys our automatic speech recognition (ASR) technology. To achieve high recognition performance in spontaneous meeting speech, we have investigated an efficient training scheme with minimal supervision which can exploit a huge amount of real data. Specifically, we have proposed a lightly-supervised training scheme based on statistical language model transformation, which fills the gap between faithful transcripts of spoken utterances and final texts for documentation. Once this mapping is trained, we no longer need faithful transcripts for training both acoustic and language models. Instead, we can fully exploit the speech and text data available in Parliament as they are. This scheme also realizes a sustainable ASR system which evolves, i.e. update/re-train the models, only with speech and text generated during the system operation. The ASR system has been deployed in the Japanese Parliament since 2010, and consistently achieved character accuracy of nearly 90%, which is useful for streamlining the transcription process.

Introduction

Since the Japanese Parliament (Diet) was founded in 1890, verbatim records had been made by manual shorthand over a hundred years. However, early in this century, the government terminated recruiting stenographers, and investigated alternative methods [1]. The House of Representatives has chosen our ASR technology for the new system. The system was deployed and tested in 2010, and it has been in official operation from April 2011. This is the first automatic transcription system deployed in national Parliaments.

Requirements for the ASR system are as follows. The first is high accuracy; over 90% is preferred. This can be easily achieved in plenary sessions, but is difficult in committee meetings, which are interactive, spontaneous, and often heated. The second requirement is fast turn-around. In the House, reporters are assigned speech for transcription in 5-minute segments. ASR should be performed almost in real-time, so Parliamentary reporters can start working promptly even during the session. The third issue is compliance to the standard transcript guidelines of the House. This can be guaranteed by using only the past Parliamentary meeting records for building the lexicon and language model.

In order to achieve high recognition performance, acoustic and language models must be customized to Parliamentary speech; that is, they need to be trained with a large amount of the matched data. Fortunately, there is a large amount of data of Parliamentary meetings. There is a huge archive of official meeting records in text, amounting to 15M words per year, which is comparable to newspapers. There is also a huge archive of meeting speech, which amounts to 1200 hours per year. However, official meeting records are different from actual utterances due to the editing process by Parliamentary reporters. There are several reasons for this: differences between spoken-style and written-style, disfluency phenomena such as fillers and repairs, redundancy such as discourse markers, and grammatical corrections.

From these reasons, we need to build a corpus of Parliamentary meetings, which consists of faithful transcripts of utterances including fillers. We prepared this kind of corpus in the size of 200 hours in speech or 2.4M words in text. The corpus is vital for satisfactory performance, but very costly. The methodology has been adopted in the conventional ASR research and development (Furui and Kawahara 2008),(J.Glass et al. 2007),(S.Matsoukas et al. 2006),(S.F.Chen et al. 2006),(S.Renals, T.Hain, and H.Bourlard 2007),(C.Gollan et al. 2005). In fact, the amount of our corpus is large, even compared with corpora prepared in these projects, however, uses only a fraction of the available data as mentioned above. Moreover, the corpus needs to be updated; otherwise, the performance would degrade in time.

In order to exploit the huge archives of Parliamentary meetings in a more efficient manner, we have investigated a novel training scheme by focusing on the differences between the official meeting record and the faithful transcript. We found that majority of the differences can be computationally modeled by a scheme of statistical machine translation (SMT). With the statistical model of the difference, we can predict what is uttered from the official records. By applying the SMT model to a huge scale of the past Parliamentary meeting records, a precise language model is generated. Moreover, by matching the audio data with the

[1] Similar changes have taken place in many countries over last decades.

model predicted for every speaker turn, we can reconstruct what was actually uttered. This results in an effective lightly-supervised training of the acoustic model, by exploiting a huge speech archive that is not faithfully transcribed. As a result, we could build precise models of spontaneous speech in Parliament, and these models will evolve in time, reflecting the change of Members of Parliament (MPs) and topics discussed.

In this article, following the analysis of the differences between faithful transcripts and official meeting records, the scheme of language model transformation is reviewed. Then, the specification of the transcription system deployed in the Japanese Parliament and its evaluations through trials and operations in the real setting are reported.

Differences between "Actually" Spoken Language and "Normally" Transcribed Text

In any languages universally, there is a difference between spoken style and written style. 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 Parliament because of the editing process by Parliamentary reporters. 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 (C.Gollan et al. 2005)(B.Ramabhadran et al. 2006), the great majority of sessions in the Japanese Parliament are in committee meetings. They are more interactive and thus spontaneous, compared to plenary sessions. The characteristics pose a big challenge to ASR; no existing systems could meet the requirements which were demanded by Parliament and stated in Introduction. The characteristics of the meetings also result 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% (13% on average). Table 1 lists the breakdown of the edits made by Parliamentary reporters. 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, and rests are not so trivial. However, most of the edits (around 93%) can be modeled by simple substitution/deletion/insertion of a word (or two words).

When we compare the analysis with the European Parliament Plenary Sessions, we observe Japanese has more disfluency and redundancy, but less grammatical corrections, because the Japanese language has a relatively free grammatical structure. Still, the above conclusion generally holds.

Table 1: Analysis of edits made for official records

edit type	category	ratio
Deletion	Fillers	50.1%
	Discourse markers	23.5%
	*Repair	3.0%
	*Word fragments	2.8%
	Syntax correction	1.8%
	Extraneous expressions	1.7%
Insertion	Function words	7.8%
Substitution	Colloquial expressions	6.4%
*Reordering		1.3%

* means not-simple edits, which are not dealt in the proposed scheme.

Scheme of Language Model Transformation

We have proposed a statistical scheme to cope with the differences between spontaneous utterances (verbatim text: V) and human-made transcripts (written-style text: W) (Y.Akita and T.Kawahara 2006)(Y.Akita and T.Kawahara 2010). In this scheme, the two are regarded as different languages and statistical machine translation (SMT) is applied (Figure 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 a human-made text.

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

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

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

Here the denominator is usually ignored in the decoding.

We have extended the simple noisy channel model to a log-linear model which can incorporate joint probability $p(W, V)$ and contextual information (G.Neubig et al. 2010), for the task of cleaning transcripts (eq. 1).

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.

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

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 Parliamentary meetings, we have a huge archive of official meeting records. Therefore, we fully exploit their statistics

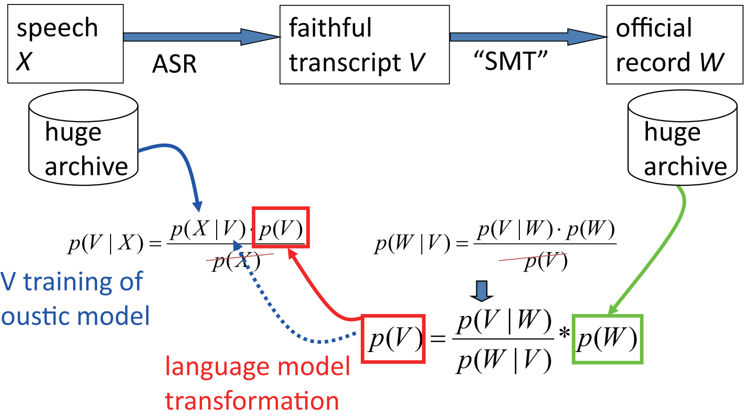
speech X
ASR
faithful transcript V
"SMT"
official record W
huge archive
huge archive
$p(V|X) = \frac{p(X|V)\cdot \boxed{p(V)}}{\cancel{p(X)}}$
$p(W|V) = \frac{p(V|W)\cdot p(W)}{\cancel{p(V)}}$
$\boxed{p(V)} = \frac{p(V|W)}{p(W|V)} * \boxed{p(W)}$
LSV training of acoustic model
language model transformation

Figure 1: Overview of the proposed scheme of ASR model training

$p(W)$ to estimate the verbatim-style language model $p(V)$ for ASR (Figure 1 right-hand side).

The transformation is actually performed on occurrence counts of N-grams as below. Here we count up to trigrams.

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

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. [2] $N_{gram}(w_1^n)$ 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 Parliamentary meetings, which was mentioned in Introduction. 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 (Y.Akita and T.Kawahara 2010) for implementation details and evaluation.

Lightly Supervised Training of Acoustic Model

The language model transformation scheme is also applied to lightly supervised training (LSV) of an acoustic model (T.Kawahara, M.Mimura, and Y.Akita 2009). For the Parliamentary meetings, we have a huge archive of speech which are not faithfully transcribed but have edited texts in the official records. Since it is not possible to uniquely recover the faithful 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 faithful transcript with high accuracy. (Figure 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). 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 faithful 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 (L.Lamel, J.Gauvain, and G.Adda 2001)(H.Y.Chan and P.Woodland 2004)(L.Nguyen and B.Xiang 2004)(M.Paulik and A.Waibel 2008), which interpolate the cleaned text with the baseline language model, resulting in a very large model.

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.

Transcription System for the Japanese Parliament – Deployment and Evaluations –

The ASR system was deployed as a core part of the new transcription system in the House of Representatives. By applying the scheme of language model transformation to a huge scale of the past Parliamentary meeting records (200M words in text over 12 years), a faithfully spoken-style lan-

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

guage model is generated. Moreover, the lightly-supervised training of an acoustic model was applied to 1000 hours of speech that are not faithfully transcribed. As a result, we could build precise models of spontaneous speech in Parliament.

These acoustic and language models, developed by Kyoto University, have been integrated into the recognition engine or decoder of NTT Corporation, which is based on the fast on-the-fly composition of WFST (Weighted Finite State Transducers) (T.Hori and A.Nakamura 2005). The overview of the entire system is depicted in Figure 2. The module integration was enabled effectively and smoothly since we have developed an open-source ASR platform Julius[3], which decouples acoustic and language models from the recognition engine in an open format. [4]

Speech is captured by stand microphones in meeting rooms for plenary sessions and committee meetings. Separate channels are used for interpellators and ministers. Channel selection and speaker segmentation modules, which were also developed in NTT, were also incorporated. The development of the entire system took almost one year, after the research presented in this paper was conducted over a couple of years.

Trials and evaluations of the system have been conducted since the system was deployed in March 2010. The accuracy defined by the character correctness compared against the official record [5] is 89.3% for 60 meetings done in 2010. When limited to plenary sessions, it is over 95%. No meetings got accuracy of less than 85%. The processing speed is 0.5 in real-time factor, which means it takes about 2.5 minutes for a 5-minute segment assigned to each Parliamentary reporter. The system can also automatically annotate and remove fillers, but automation of other edits is still under ongoing research (G.Neubig et al. 2010). Furthermore, the semi-automated update of the acoustic and language models using the data throughout the trial operations has brought an additional gain in accuracy by 0.7% absolute.

After the trials in FY 2010, the system has been in official operation from April 2011. The new system now handles all plenary sessions and committee meetings. The speaker-independent ASR system generates an initial draft, which is corrected by Parliamentary reporters. The averaged character correctness measured for 118 meetings held throughout 2011 is 89.8%, and the additional insertion errors, excluding fillers, account for 8.2%. It translates that, roughly speaking, the system's recognition error rate is around 10%, and disfluencies and colloquial expressions to be deleted or corrected also account for 10%. Thus, Parliamentary reporters still play an important role although the entire transcription process is streamlined by the new system compared with the conventional short-hand scheme.

The post-editor used by Parliamentary reporters is vital for efficient correction of ASR errors and cleaning transcripts. Designed by reporters, it is a screen editor, which is similar to the word-processor interface. The editor provides easy reference to original speech and video, by time, by utterance, and by character. It can speed up and down the replay of speech. A side effect of the ASR-based system is all of text, speech, and video are aligned and hyperlinked by speakers and by utterance. It will allow for efficient search and retrieval of the multi-media archive.

For system maintenance, we continuously monitor the ASR accuracy, and update ASR models. Specifically, the lexicon and language model are updated once a year to incorporate new words and topics. Note that new words can be added by Parliamentary reporters at any time. The acoustic model will be updated after the change of the Cabinet or MPs, which usually takes place after the general election. Note that these updates can be semi-automated without manual transcription in our lightly-supervised training scheme. The updates are done by software engineers contracted with Parliament, not by the researchers and developers of the system. We expect the system will improve or evolve with more data accumulated.

Concluding Remarks and Future Perspectives

The article addresses the new transcription system based on our novel "sustainable" approach to ASR systems, which can evolve without requiring manual transcription. It is not totally unsupervised, but assumes the final output which is generated during the system operation. With this new scheme, we believe we have realized one of the highest-standard ASR system dedicated to Parliamentary meetings. Specifically, the accuracy reaches nearly 90%, which the conventional systems based on manually prepared corpora would never achieve.

There was a drastic change from the manual short-hand to this fully ICT-based system. Although it needed some time, Parliamentary reporters have been almost accustomed to the new system and provide more positive feedbacks in time. We also expect that training of new reporters will be much easier compared with the traditional short-hand stenographers, who needed one or two years before the actual assignment of transcription.

The system was ported to a software package by NTT Corporation, and has been adopted in a number of local governments in Japan.

Acknowledgment

The work presented in this article is indebted to Dr. Yuya Akita and Mr. Masato Mimura of Kyoto University as well as a number of researchers and engineers of NTT Corporation. This work has also been supported by Japanese Governmental funds including MIC SCOPE, JST CREST and JSPS Grant-in-Aid for Scientific Research.

[3] http://julius.sourceforge.jp/

[4] Julius itself was not adopted in the running system because it is no-warranty, and the NTT decoder performs much faster.

[5] This is slightly different from accuracy normally used in ASR evaluations, which is computed against faithful transcripts. Preparing faithful transcripts costs very much and cannot be done for a long period. On the other hand, computing character correctness against official meeting records is consistent with the goal of the system.

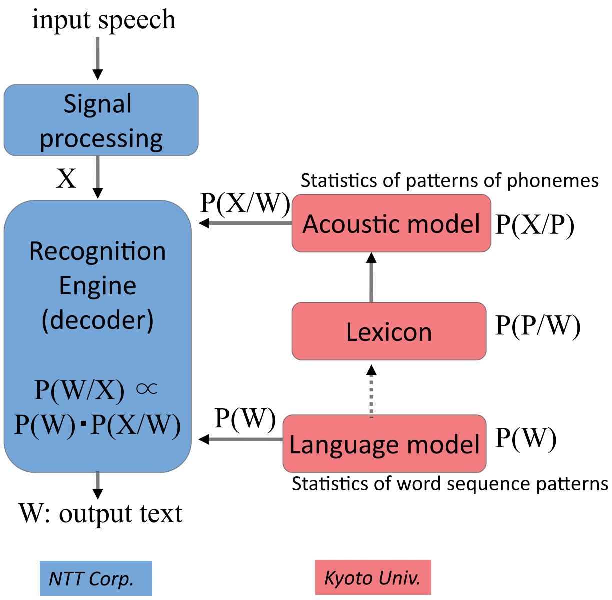

Figure 2: Overview of the ASR system

References

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

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

Furui, S., and Kawahara, T. 2008. Transcription and distillation of spontaneous speech. In J.Benesty; M.M.Sondhi; and Y.Huang., eds., *Springer Handbook on Speech Processing and Speech Communication*. Springer. 627–651.

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

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

J.Glass; Hazen, T.; S.Cyphers; I.Malioutov; D.Huynh; and R.Barzilay. 2007. Recent progress in the MIT spoken lecture processing project. In *Proc. INTERSPEECH*, 2553–2556.

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

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

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

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

S.Matsoukas; J.-L.Gauvain; G.Adda; T.Colthurst; Kao, C.-L.; O.Kimball; L.Lamel; F.Lefevre; J.Z.Ma; J.Makhoul; L.Nguyen; R.Prasad; R.Schwartz; H.Schwenk; and B.Xiang. 2006. 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.

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

T.Hori, and A.Nakamura. 2005. Generalized fast on-the-fly composition algorithm for wfst-based speech recognition. In *Proc. Interspeech*, 557–560.

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

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

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