Automatic Transcription System for Meetings of the Japanese National Congress

Yuya Akita *Masato Mimura* *Tatsuya Kawahara*

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

Abstract

This paper presents an automatic speech recognition (ASR) system for assisting meeting record creation of the National Congress of Japan. The system is designed to cope with spontaneous characteristics of meeting speech, as well as a variety of topics and speakers. For acoustic model, minimum phone error (MPE) training is applied with several normalization techniques. For language model, we have proposed statistical style transformation to generate spoken-style N-grams and their statistics. We also introduce statistical modeling of pronunciation variation in spontaneous speech. The ASR system was evaluated on real congressional meetings, and achieved word accuracy of 84%. It is also suggested that the ASR-based transcripts with this accuracy level is usable for editing meeting records.

Index Terms: Spontaneous speech recognition, congressional speech, minimum phone error training, statistical style transformation

1. Introduction

Recently, research targets of automatic speech recognition (ASR) have been extended to spontaneous speech such as meetings. Major research projects on meetings include the NIST Rich Transcription (RT) project [1] and the AMI/AMIDA project [2], which have dealt with ASR of multi-party meetings. The TC-STAR project investigated speech translation of congressional meetings in European Parliament, and ASR systems were developed by several institutes [3, 4, 5], which reported WER of around 10% for European Parliament Plenary Speeches (EPPS).

We have also been developing an ASR system for the National Congress (Diet) of Japan. In the National Congress, every meeting is transcribed for a record. At present, utterances are taken down in shorthand, and then edited afterwards by professional stenographers, who are specially trained in the national institute. Especially in Japanese, it is difficult to transcribe in real time by typing, since a large number of homonyms appear in Japanese sentences and selection of correct words (in *kanji* notation) takes much time. The ASR technology is expected to be useful for the speech-to-text transcription process. In fact, the House of Representatives has been seriously considering introduction of ASR for its next-generation transcription system.

Majority of meetings of the National Congress of Japan are held in a number of specialized committees. Unlike plenary meetings such as EPPS, the committee meetings are more interactive and spontaneous, because most of the sessions are done in the question-answer format, and many utterances are not reading manuscripts and sometimes excited. The spontaneous characteristics cause much difficulty in ASR, especially in acoustic and language modeling.

In this paper, we present our current ASR system for the National Congress. First, we describe our speech corpus of the National Congress which is used to build and test the system, then we explain the system in detail. Specifically, minimum phone error (MPE) training and several normalization methods are incorporated to acoustic modeling. For language and pronunciation modeling, we have proposed a scheme of statistical style transformation [6, 7] to efficiently cover topics and spontaneous characteristics. Finally, experimental evaluations of the ASR system on real congressional meetings are reported. We also made evaluations of ASR results by professional stenographers in terms of usability as drafts of meeting minutes.

2. Speech corpus of the National Congress

We have been preparing a speech corpus of meetings in the National Congress (the House of Representatives). Participants of the meetings are mainly members of the Cabinet, members of the National Congress and government officials. The corpus includes a variety of specialized committee meetings such as foreign affairs, national security, judicial affairs and agriculture, as well as the committee of budget, in which a variety of domestic and international issues are discussed. Thus, the corpus covers a variety of topics and speakers. To date, we have collected audio data of 61 meetings that were held during 2003 to 2007. The total duration of the audio data is 236 hours.

Audio data were recorded via close-talking microphones. The National Congress has several meeting rooms, which have different acoustic conditions. Therefore, meetings were chosen in order to cover not only topics but also all meeting rooms. In each turn, a speaker starts to talk after designation by the chairperson. Therefore, overlapping of speech by multiple speakers is rarely observed, but some noises such as handclaps and heck-

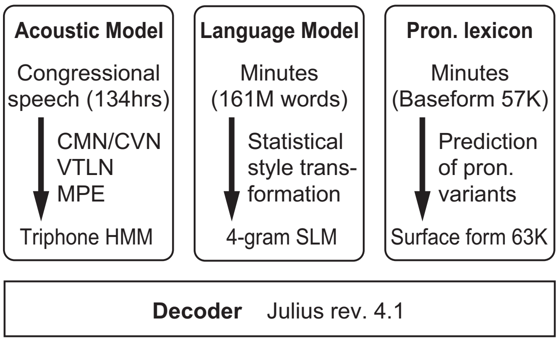

Figure 1: Overview of the ASR system

ling are occasionally observed. All utterances were transcribed manually and faithfully. Audio data is also manually segmented into speaker turns by detection of speaker changes.

Minutes are edited by the National Congress for every meeting, therefore, we aligned these minutes with faithful transcripts as a parallel corpus. Specifically, we have compared minutes and transcripts, and then annotated different portion in the transcripts.

We also use all minutes from 1999 to 2007 for training of language model.

3. ASR system for the National Congress

Figure 1 shows an overview of our ASR system for the National Congress meetings. As a decoder, our Julius [8] rev.4.1 is used. The acoustic model is based on MPE training with 134-hour speech data. The language model is a 4-gram model, which is transformed from 161M-word document-style texts in the minutes to a spoken-style model. The baseform lexicon is extended by adding surface forms, which are predicted by statistical transformation of pronunciation variations. Note that a part of the corpus of the National Congress is reserved as an evaluation test set, and the rest is used for training of models.

3.1. Acoustic modeling

As acoustic features, we adopt MFCC coefficients, their Δ and $\Delta\Delta$ coefficients together with ΔEnergy and $\Delta\Delta$Energy. The total number of features is 38. Cepstrum mean normalization (CMN), cepstrum variance normalization (CVN) and vocal tract length normalization (VTLN) are applied to these features. Normalization is individually performed for every speaker turn.

The acoustic model is three-state, left-to-right, diagonal-covariance triphone HMM. In this system, we introduce MPE training [9] to estimate HMM parameters. The objective function F of MPE training is defined as:

$$F(\lambda) = \sum_r \frac{\sum_s p_\lambda(O_r|s)^\kappa P(s) RawAcc(s)}{\sum_s p_\lambda(O_r|s)^\kappa P(s)}, \quad (1)$$

where O_r is a sequence of observation data, $P(s)$ is a linguistic score for a sentence hypothesis s, and κ is a scaling factor. $RawAcc(s)$ is an estimate of phone accuracy for s, which is calculated with some approximation; we generate a large number of competing sentence hypotheses by doing ASR roughly, i.e., applying loose linguistic constraint to ASR by using a bigram language model. Then, sufficient statistics are estimated by the forward-backward algorithm over a hypothesis lattice and a reference phone transcript.

3.2. Language modeling

When modeling the Japanese language for spontaneous speech recognition, we have to take into account that spoken Japanese is much different from written Japanese. One of typical spoken-style expressions are end-of-sentence (EOS) expressions (such as *desu ne*) that are used not only as true EOS but like fillers. There are various colloquial words which should be corrected in the written-style language, as well as many fillers and some disfluencies. A large amount of faithful transcripts are necessary to model these kinds of spoken-style expressions, however, collection of such transcripts is limited in text size.

To build a language model for spoken Japanese, we have proposed a novel scheme of statistical transformation of language model [6]. The transformation is based on the framework of statistical machine translation, where sentence Y of the target language is generated from sentence X of the source language, which maximizes posterior probability $P(Y|X)$ based on Bayes' rule.

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \quad (2)$$

In this work, we consider document-style and spoken language as different ones, denoted by X and Y, respectively, and estimate spoken language model $P(Y)$, which is formulated as Equation (3) by rewriting Equation (2).

$$P(Y) = P(X)\frac{P(Y|X)}{P(X|Y)} \quad (3)$$

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

$$N_{LM}(y) = N_{LM}(x)\frac{P(y|x)}{P(x|y)} \quad (4)$$

Here, x and y are individual patterns that are transformed and $N_{LM}(x)$ and $N_{LM}(y)$ are N-gram entries including them. Transformation patterns x and y contain preceding and following words as contexts. To alleviate the data sparseness problem, part-of-speech (POS) contexts are also introduced.

The word-based and POS-based transformation models are applied to N-gram entries of a document-style language model using Equation (4) in a back-off scheme. First, each word-based pattern is applied to input document-style N-gram entries in turn, and a spoken-style N-gram entry is generated with its estimated occurrence count when the pattern is matched. If not matched, then POS-based pattern is applied.

3.3. Pronunciation modeling

Spoken Japanese also has variations of pronunciation, for example, longer and shorter vowels such as “/m o ch i i r u/” to “/m o ch i: r u/” (= use) and “/h o N t o: n i/” to “/h o N t o n i/” (= truly), omission of phones like “/s o r e/” to “/s o e/” (= that) and other changes such as “/k e r e d o m o/” to “/k e: d o m o/” (= though) and “/m a i n i ch i/” to “/m a i N ch i/” (= everyday). To generate a better pronunciation lexicon for spontaneous speech, we proposed a prediction method of pronunciation variations [7]. In this method, pronunciation variation is modeled by statistical mapping of phone sequences between baseforms and surface forms. Since the corpus of the National Congress does not have surface forms and is not good for training of the statistical model, we make use of the Corpus of Spontaneous Japanese (CSJ) [10] to derive this transformation model. The model is flexibly applicable to any new lexicon, and their surface forms can be generated with appropriate probabilities.

In training of the transformation model, patterns of pronunciation variations are detected, and necessary statistics of variation patterns and their phone contexts are estimated in the training corpus (CSJ). Then, a set of rewrite rules, i.e., a transformation model, is derived with appropriate contexts and probabilities from the variation statistics. We use at most two phones as preceding and following contexts, respectively. A threshold is introduced for occurrence counts to determine an adequate length of the context so that the model has reliable statistics. Rules are defined in a descending order, from the longest context set to a context-independent rule. The contextual patterns eliminated by the threshold are backed-off to shorter-context rules. As a result, we obtained 265 kinds of variation patterns and 1,381 rules.

These rules are applied to baseforms to generate new pronunciation entries (surface forms). Rules with longer context are applied with higher priority, and then backed-off to shorter contexts if necessary. Probabilities of the resulting new pronunciation entry and the original entry are updated by variation probabilities of the applied rules. To suppress false matching by infrequent entries, pronunciation entries whose probabilities are smaller than a threshold are eliminated.

4. Experimental evaluation

4.1. Experimental setup

We evaluated the performance of our ASR system by using meeting speech in the corpus of the National Congress. Seven committee meetings in 2007, which were held after all meetings used for training, were selected for a test set. The total number of words in the test set is 306,988. To meet requirements from the House of Representatives, we set real time factor (RTF) of decoding as around three in this experiment. Decoding parameters were tuned to realize this criteria on a Core2 Extreme Q6850 processor (3.0GHz).

Table 1: Improvements of word accuracy by incorporated techniques

Techniques	Word accuracy	Abs. gain
(Baseline)	78.4%	—
AM:+VTLN	79.7%	1.3%
AM:+MPE	81.6%	1.9%
LM:+Transform	83.0%	1.4%
LM:+4-gram	83.6%	0.6%
PM:+Surface form	84.0%	0.4%

The acoustic model was trained with the 134-hour speech data in the corpus. For comparison, we prepared another model by maximum likelihood (ML) training.

For language model training, we used minutes of the congress from 1999 to 2007 (immediately before the test set) for statistical style transformation. The transformation model was trained using transcripts from 2003 to 2005 in the corpus. The total sizes of training data for the language model and the transformation model were 161M and 2.8M words, respectively. For comparison, we prepared other two models. One is a simple mixture-based model made by interpolating these two data. The interpolation weight was chosen afterwards to provide best perplexity on the test set. The other is a transformed 3-gram model, which was previously used in our system.

The original pronunciation lexicon was based on baseforms given by a morphological dictionary, and it was extended by predicting and adding surface forms. The size of the vocabulary, the baseform lexicon and the surface form lexicon were 53,791, 57,050 and 62,928, respectively. The vocabulary of the mixture-based and 3-gram models is exactly same as that of the transformation-based 4-gram model. The out-of-vocabulary (OOV) rate on the test set is 0.13%.

4.2. ASR results

First, we preliminarily compared the performance of ASR with RTF ≈ 3 and RTF > 6 where we considered the search space was wide enough and the performance was saturated. The degradation of performance by the former against the latter was 0.37% absolute on average. We consider it permissible to realize faster decoding required by the House.

The baseline system was composed of the ML-trained acoustic model, the mixture-based language model and the baseform lexicon. The word accuracy by this system was 78.4%. Table 1 shows improvements of word accuracy by the techniques described above. VTLN and MPE training gained 1.3% and 1.9%, respectively, and by both techniques absolute improvement of 3.2% was obtained. As for language model, Table 2 shows perplexity by var-

Table 2: Reduction of test-set perplexity

Models	Perplexity	%Reduction
Mixed, 3-gram	58.1	—
Mixed, 4-gram	52.6	9.6%
Transformed, 3-gram	45.4	21.9%
Transformed, 4-gram	43.3	25.5%

ious models. The reduction of perplexity by the transformed 4-gram model over mixture-based 3-gram model was 25.5%. As shown in Table 1, word accuracy was also improved by 1.4% absolute with the model transformation. By extending the length of N-grams from three to four, we also obtained 0.6% absolute improvement on word accuracy. The language model transformation well covered spoken-style expressions, thus the model realized higher prediction performance. Furthermore, the 4-gram model suppressed inappropriate hypotheses in decoding. The extended lexicon improved word accuracy by 0.4%. The baseforms in this experiment were generated with the latest morphological dictionary which contained more spoken-style pronunciations than our former lexicon, and yet the prediction method could still generate effective surface forms, and improved word accuracy. All these improvements on word accuracy are statistically significant at $p < 0.01$. By this RTF ≈ 3 system, we finally achieved word accuracy of 84.0% and character accuracy of 86.4%.

4.3. Usability test by professional stenographers

Based on these ASR results, an experiment was conducted to investigate the usability of the ASR-based transcription for real application. We prepared three sets of ASR-based transcripts whose word accuracy were 85%, 80% and 75%, and eighteen professional stenographers in the House of Representatives edited these automatic transcripts. As a result, the lower word accuracy became, the more time was needed to edit. Especially for 75% case, some stenographers commented that it was better to manually transcribe from scratch, rather than to edit ASR results. In contrast, no such comments were made in the case of 85%. The fact demonstrates that the word accuracy we achieved in the previous experiments is sufficient for this kind of real application.

Next, we compared the time required to make meeting minutes based on ASR-based transcripts and manual stenography currently used in the House. For sixteen 5-minute speech segments, the former spent around 70 minutes for each, while the latter needed about 100 minutes. Note that it took more than one hour to edit minutes for 5-minute speech, because the editing work included confirmation of content in the speech as well as correction of errors. The automated system successfully reduced editing time by 30%, which demonstrates the effectiveness of the system.

5. Conclusions

We have described our ASR system for the National Congress of Japan. The system consists of an acoustic model with state-of-the-art techniques, and a language model and pronunciation lexicon transformed from orthographic style to spoken style by our statistical transformation methods. The system achieved word accuracy of 84% on real congressional meetings. It was also suggested that this ASR-based transcripts were useful for efficient editing of minutes by professional stenographers. We hope the ASR system is further applicable to other applications in the future, such as automatic captioning and audio/video search.

Acknowledgment: This work was partly supported by Strategic Information and Communications R&D Promotion Programme (SCOPE), Ministry of Internal Affairs and Communications, Japan.

6. References

[1] J.S. Garofolo, C.D. Laprun, and J.G. Fiscus, "The Rich Transcription 2004 Spring Meeting Recognition Evaluation," in *Proc. ICASSP Meeting Recognition Workshop*, 2004.

[2] S. Renals, T. Hain, and H. Bourlard, "Recognition and Understanding of Meetings: the AMI and AMIDA Projects," in *Proc. ASRU*, 2007, pp. 238–247.

[3] 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.

[4] J. Loof, M. Bisani, Ch. Gollan, G. Heigold, B. Hoffmeister, Ch. Plahl, R. Schluter, and H. Ney, "The 2006 RWTH Parliamentary Speeches Transcription System," in *Proc. ICSLP*, 2006, pp. 105–108.

[5] 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. ICSLP*, 2006, pp. 1225–1228.

[6] Y. Akita and T. Kawahara, "Topic-independent Speaking-style Transformation of Language Model for Spontaneous Speech Recognition," in *Proc. ICASSP*, 2007, vol. 4, pp. 33–36.

[7] Y. Akita and T. Kawahara, "Generalized Statistical Modeling of Pronunciation Variations using Variable-length Phone Context," in *Proc. ICASSP*, 2005, vol. 1, pp. 689–692.

[8] T. Kawahara, A. Lee, K. Takeda, K. Itou, and K. Shikano, "Recent Progress of Open-Source LVCSR Engine Julius and Japanese Model Repository," in *Proc. ICSLP*, 2004, pp. 3069–3072.

[9] 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.

[10] 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.