Detection of Precisely Transcribed Parts from Inexact Transcribed Corpus

Kengo Ohta [#1], Masatoshi Tsuchiya [*2], Seiichi Nakagawa [#3]

[#] *Department of Information and Computer Sciences / [*] Information and Media Center,*
Toyohashi University of Technology,
1–1 Hibarigaoka, Tenpakucho, Toyohashi-shi, Aichi, 441–8580 Japan
[1] kohta@slp.cs.tut.ac.jp
[2] tsuchiya@imc.tut.ac.jp
[3] nakagawa@slp.cs.tut.ac.jp

***Abstract*—Although large-scale spontaneous speech corpora are crucial resource for various domains of spoken language processing, they are usually limited due to their construction cost especially in transcribing precisely. On the other hand, inexact transcribed corpora like shorthand notes, meeting records and closed captions are widely available. Unfortunately, it is difficult to use them directly as speech corpora for learning acoustic models, because they contain two kinds of text, precisely transcribed parts and *edited* parts. In order to resolve this problem, this paper proposes an automatic detection method of precisely transcribed parts from inexact transcribed corpora. Our method consists of two steps: the first step is an automatic alignment between the inexact transcription and its corresponding utterance, and the second step is a support vector machine based detector of precisely transcribed parts using several features obtained by the first step. Experiments using the Japanese National Diet Record shows that automatic detection of precise parts is effective for lightly supervised speaker adaptation, and shows that it achieves reasonable performance to reduce the converting cost from inexact transcribed corpora into precisely transcribed ones.**

I. Introduction

Large-scale spontaneous speech corpora are crucial resource for various domains of spoken language processing. For example, the simplest approach to construct a language model, which covers spoken-style expressions as well as the specified domain topics, is training it from a large-scale spontaneous corpus including many precise transcriptions of spontaneous speech in the specified domain. However, the available corpora are usually limited because their construction cost are quite expensive especially in transcribing speech precisely.

On the other hand, inexact transcribed corpora like shorthand notes, meeting records and closed captions are more widely available than precisely transcribed ones, because their inexactness reduces their construction cost. Fig.1 shows an example snippet of the Japanese National Diet Record, which is an inexact transcribed corpus of debates in the Japanese National Diet. It is continuously maintained by the Japanese National Diet Library[1], and covers the debates over the past 60 years. The corpus consists of *edited* transcriptions shown in the lower part of Fig.1, and there are three kinds of editing operations between the *edited* transcription and the precise transcription in the upper part of Fig.1. The first is removing redundant expressions (e.g., "ですね/*desune*/", "と/*to*/"), disfluencies such as filled pauses (e.g., "えー/*ee*/", "いー/*ii*/"), and hesitations (e.g., "け/*ke*/"). The second is that colloquial expressions (e.g., "てる/*teru*/") are replaced by literary expressions (e.g., "ている/*teiru*/"), and that omission of particles (e.g., "を/*wo*/") are recovered. The third is that certain commas are added or removed according to the shorthand writers' intuition. Because this corpus contains both precisely transcribed parts and *edited* parts, it is difficult to use it directly as a speech corpus for learning acoustic models.

ところ が ですね， えー，この 資料，見 て み ます と 神奈川 県 の 場合 は， け，結果 と し て 財政 的 に いー 豊か に なっ てる と．
(tokoro ga desu ne, ee, kono shiryou, mi te mi masu to kanagawa ken no baai wa, ke, kekka to shi te zaisei teki ni ii yutaka ni naq teru to.)

(i) Precise transcription

ところ が， この 資料 を 見 て み ます と，神奈川 県 の 場合 は， 結果 と し て 財政 的 に 豊か に なっ ている．
(tokoro ga, kono shiryou wo mi te mi masu to, kanagawa ken no baai wa, kekka to shi te zaisei teki ni yutaka ni naq te iru.)

(ii) Edited transcription

Fig. 1. Example of precise and edited transcriptions (in Japanese)

In order to resolve this problem, this paper proposes an automatic detection method of precisely transcribed parts from *edited* transcriptions. Our method consists of two steps: the first step is an automatic alignment between the *edited* transcription and its corresponding utterance, and the second step is a support vector machine based detector of precise parts using several features obtained by the first step.

There are two major directions of related works. The first

[1] http://kokkai.ndl.go.jp/

TABLE I
WORD OVERLAP

	Sub	Del	Ins	Err
Audiobook[3]	0.4%	1.4%	3.6%	5.4%
Japanese National Diet Records	1.0%	6.3%	0.7%	8.0%

direction is adaptation of acoustic models. Watanabe et al.[1] proposed that the common fragments between the outputs of several LVCSR systems were used as training labels of adaptation of acoustic models. Lamel et al. [2] applied an automatic alignment for lightly supervised acoustic model training. In their work, automatic alignment is used to filter out unreliable training data in acoustic model training. Norbert et al. [3] also employed lightly supervised recognition for automatic alignment between text and speech of Audiobooks. Table I shows word overlap between precise transcriptions and *edited* transcriptions. As shown in Table I, Japanese National Diet Records contains more deletion errors than Audiobooks, and these errors make alignment more difficult. Our experimental result shows that our proposed method achieves high accuracy of detecting precise parts, and shows that precise detected parts are effective for training labels of lightly supervised speaker adaptation.

The second direction is reducing the construction cost of precisly transcribed corpora. Roy et al. [4] utilized an acoustic score obtained from automatic alignment to estimate the accuracy and difficulty in transcribing speech recordings. Maruyama et al. [5] suggested using automatic alignment for timing detection of closed captioning in documentary programs. Our experimental result shows that it is quite difficult to detect *edited* parts accurately, but still shows that our proposed method achieves enough performance to reduce the conversion time from *edited* transcriptions into precise ones.

The remainder of this paper is organized as follows. In Section II, we introduce the automatic alignment method between the edited transcription and its corresponding utterance. Section III describes the edited part detector based on a support vector machine. The evaluation experiment on the Japanese National Diet Record is presented in Section IV. Finally, Section V concludes the paper.

II. AUTOMATIC ALIGNMENT BETWEEN EDITED TRANSCRIPTION AND ITS CORRESPONDING UTTERANCE

As the first step in our proposed method, we perform an automatic alignment between the edited transcription and its corresponding utterance. In this alignment, words in the edited transcription and the corresponding utterances are aligned by allowing the insertion of a short pause or filled pause between words. Such alignment is implemented by automatic speech recognition together with a constraint based on a bigram language model as shown in Fig.2. Here, w_i represents the *i*-th word in the transcription, and sp_i and $Filler_i$ denote a short pause and a filled pause, respectively, occurring immediately after word w_i. As a result of this constraint, the output of the automatic speech recognition is restricted to the same word sequence as in the transcription.

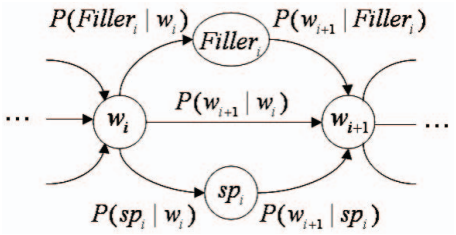

Fig. 2. Bigram constraint for automatic alignment

An example of the automatic alignment between an precise/edited transcription and its corresponding utterance is shown in Fig.3. As shown in this figure, when the precise transcription is aligned with the corresponding utterance, the syllable durations resemble their inherent values. On the other hand, when the edited transcription is aligned with the corresponding utterance, automatic alignment makes the best effort possible to align the input utterance with the transcription. As a result, frames of spurious syllables are absorbed by a silence segment or another syllable segment. This causes the syllable duration to be distorted and the acoustic score in the alignment degrades because of a mismatch between the syllable and the aligned model. In the example in Fig.3, the model */ga/* is forced to align with the frames of syllable */de/*. Besides, the short pause models are also forced to align with the frames of syllables */ga/* and */su ne/*. Additionally, a filled pause */oh/* is inserted. Hence, if syllable durations are overly long or short compared with their inherent values, or if acoustic scores are worse than a standard value, it may suggest that there are mismatches between the text and the utterances due to the text having been edited.

III. AUTOMATIC DETECTION OF PRECISE PARTS

As the second step in our proposed method, a support vector machine based detector is employed to detect the precise parts using certain features obtained by the automatic alignment described in Section II.

In this study, we formalize the detection of precise parts as a binary classification problem for each word in the edited transcriptions. Each word is classified either as a non-edited word or a edited word based on the features obtained by the automatic alignment. We used TinySVM (ver 0.09) [6] as the support vector machine implementation with a polynomial kernel.

A. Features

In our method, nine categories of features are adopted. We used the features in [7] as a reference for feature selection. Details of each feature are described in the following sections. All these features for the focused word, the preceding two words, and the succeeding two words are combined into a feature vector for each word.

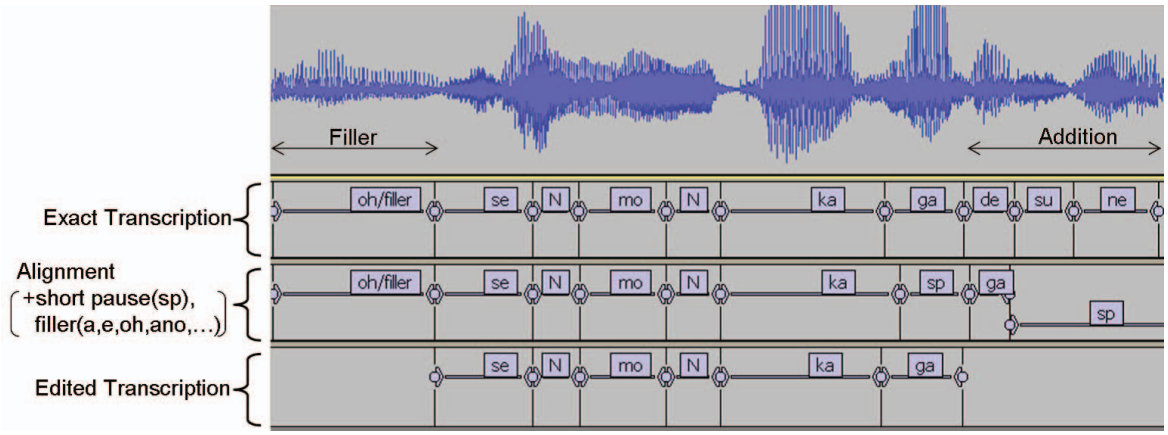

Fig. 3. Example of alignment between edited transcription and its corresponding utterance

1) HMM based log likelihood: The acoustic score obtained by automatic alignment is used for the first feature. This score is normalized by the duration of the word, and then subtracted from an acoustic score obtained from HMM based continuous syllable recognition. This corresponds to the posterior log-probability as the feature.

2) Posterior probability: The posterior probability of each word obtained from the confusion network is used as the second feature. A confusion network is constructed in the large vocabulary continuous speech recognition process, which is executed in parallel with the automatic alignment. The feature value "0" is assigned to a word that does not occur in the confusion network.

3) Number of competing words: The number of competing words in the confusion network is used as the third feature. For a word that does not occur in the confusion network, the maximum value in the network is assigned.

4) Mean of syllable duration: The mean of the duration of syllables in the word is used as the fourth feature. The duration of each syllable is normalized by two factors: the local and global means.

The mean of syllable duration normalized by the local mean is defined as follows:

$$Local\_d = \frac{1}{N}\sum_{i=1}^{N}\frac{dur(s_i)}{\frac{1}{6}\sum_{j=i-3(j\neq i)}^{i+3} dur(s_j)}, \quad (1)$$

where N is the number of syllables in the word, and $dur(s)$ is the duration of syllable s. The duration of each syllable is normalized based on the preceding three syllables and the succeeding three syllables.

The mean of syllable duration normalized by the global mean is defined as follows:

$$Global\_d = \frac{\frac{1}{N}\sum_{i=1}^{N} dur(s_i)}{\frac{1}{|U|}\sum_{s\in U} dur(s)}, \quad (2)$$

where U is the set of syllables in the utterance. The duration of each syllable is normalized based on all the syllables in the utterance.

5) Variance in syllable duration: The variance in the duration of syllables in the word is used as the fifth feature. The duration of each syllable is normalized based on all the syllables in the utterance (*Global_d*). The definition is as follows:

$$Var\_d = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{dur(s_i)}{\frac{1}{|U|}\sum_{s\in U} dur(s)} - Global\_d\right)^2 \quad (3)$$

6) Score of statistical syllable duration model: In [8], a statistical phone duration model is utilized to filter out corrupt user utterances for a computer-assisted pronunciation training system. Motivated by their work, we build a statistical syllable duration model and use the score of this model as the sixth feature.

The definition of this feature is given below.

$$Score\_d = \frac{1}{|W|}\log\left(\prod_{s\in W}\frac{P(dur(s)|s)}{P_{anti-model}(dur(s))}\right), \quad (4)$$

where W is the set of syllables in the word. Here, the statistical syllable duration model $P(dur(s)|s)$ is the Gamma distribution trained from syllable durations obtained by automatic alignment between precise transcriptions and their corresponding utterances. On the other hand, the anti-model $P_{anti-model}(dur(s))$ is also a Gamma distribution, which models the distortion of syllable duration in automatic alignment between edited transcriptions and utterances. This anti-model is trained from syllable durations obtained by automatic forced alignments between *shuffled* transcriptions and utterances (as a result, each transcription is aligned to an uncorresponding utterance), which simulate the mismatches between transcriptions and utterances.

7) Word identity: The word identity is used as the seventh feature. There are several words that have a greater tendency to be edited, such as auxiliary verbs "ます/*masu*/" or particles "が/*ga*/". This feature captures such a tendency.

8) Word length: The total number of syllables in the word is used as the eighth feature. We consider that this feature acts

as a negative feature based on the hypothesis that a longer word has a greater tendency to be involved in an edited part. For example, the word "けれども/*keredomo*/" is often replaced by "けど/*kedo*/", and the word "やっぱり/*yaqpari*/" is often replaced by "やはり/*yahari*/".

9) Word duration: The word duration obtained by automatic alignment is used as the ninth feature. This feature is also based on the previous hypothesis that a longer word has a greater tendency to be involved in an edited part.

B. Target performance for reducing conversion time

In this study, we consider the target performance of detecting edited parts as 33% for precision and 50% for recall. Because this target performance looks quite bad, we have to show that this target performance is still enough to reduce the conversion time from edited transcriptions into precision ones. Here, as an example, consider a short transcription consisting of 100 words, of which 10 words are edited. According to this target, we can detect 15 words, 5 of which are truly edited. Thus, human transcribers can find and correct 5 edited words by checking only 15 words. In other words, if a user wants to find 10 edited words, he/she needs to check only 30 candidate words. This is three times more efficient than the situation without our detection method.

IV. Experiment

In this section, we discuss our evaluation experiments using the Japanese National Diet Record.

A. Experimental Setup

The Japanese National Diet Record contains numerous edited transcriptions, in which spoken-style expressions and disfluencies such as filled pauses and hesitations have been edited.

We conducted the evaluation experiments under two conditions, the semi-closed speaker condition and the open speaker condition. In the semi-closed speaker experiment, two speakers out of four are included in both the training set and test set. On the other hand, in the open speaker experiment, speakers in each set are completely separated. The data statistics for each condition are shown in Table II.

As the decoder for automatic alignment and continuous syllable recognition, we used the in-house large vocabulary continuous speech recognition system, SPOJUS++ (SPOken Japanese Understanding System) [9], and its acoustic analysis condition is shown in Table III. The 116 Japanese context-independent syllable-based acoustic models [10] (a left-to-right topology, 4 emitting states, and a single Gaussian mixture with full covariance matrix) were trained from academic presentation speech data and simulated public speech data in the CSJ (Corpus of Spontaneous Japanese) [11].

Based on a preliminary experiment, we set $P(Filler_i|w_i) = 0.05$, $P(w_{i+1}|w_i) = 0.475$ and $P(sp_i|w_i) = 0.475$ for the alignment constraint (Fig.2).

TABLE II
Data statistics

	Semi-closed speaker		Open speaker	
	Train	Test	Train	Test
Speech Length (min)	22	20	42	60
# of Speakers	5	4	7	11
# of Words	3.6k	3.6k	7.2k	10.8k
# of Edited Words	347	257	604	426
Editted Ratio (%)	9.6	7.1	8.4	3.9

TABLE III
Conditions of acoustic analysis for input speech

Sampling Rate	16kHz
Preemphasis	0.98
Analysis Window	Hamming Window
Analysis Frame Length	25ms
Analysis Frame Shift	10ms
Feature Parameter	MFCC + △MFCC + △△MFCC + △Pow + △△Pow (38 dimensions)

B. Detection Results of Edited Parts

1) Semi-closed speaker experiment: The recall-precision curves for edited part detection for all 4 speakers, 2 closed speakers and 2 open speakers, are shown in Fig.4. As shown in this figure, there is a large difference in performance between closed speakers and open speakers. Specifically, reasonable performance (recall=50%, precision=33%) was obtained for the closed speakers, while much worse results were obtained for the open speakers. This result suggests that either each speaker has his/her own specific characteristic for edits or the amount of training data was insufficient. More training data are not desired for our purpose, which is to detect edited parts from a large-scale corpus with only a small amount of supervision.

2) Open speaker experiment: The recall-precision curve for the open speaker experiment is also shown in Fig.4. As shown in this figure, the performance is significantly lower than that under the semi-closed speaker condition. This

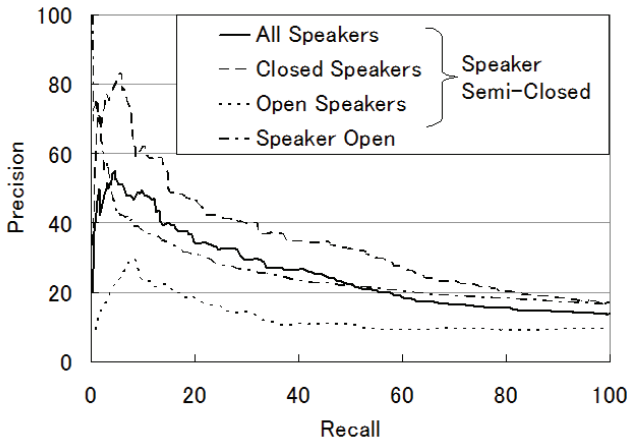

Fig. 4. Recall-precision curve for detection of edited parts

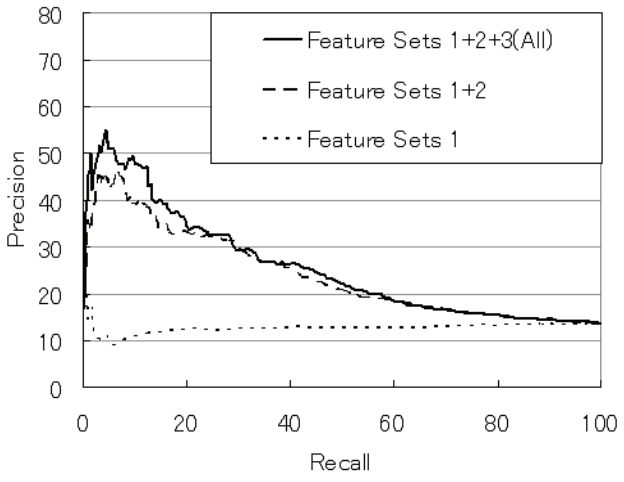

Fig. 5. Recall-Precision Curve with Different Feature Sets in Speaker Semi-closed Condition

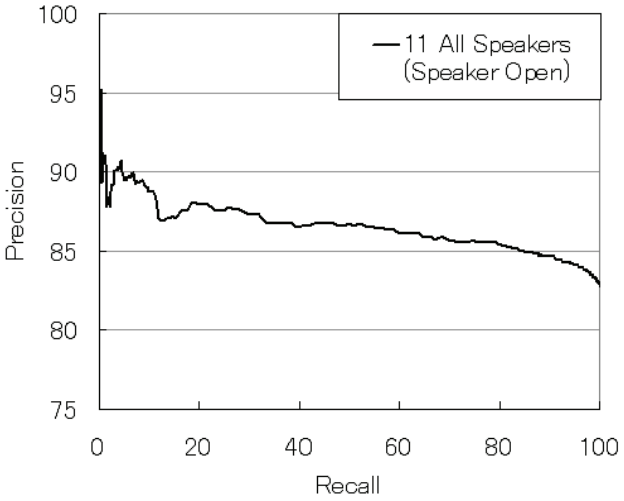

Fig. 6. Recall-precision curve for detection of precise parts (syllable units)

result reinforces the previous suggestion that each speaker has his/her own specific characteristics for edits. We consider that the difference between the results under the open speaker condition and those for open-speakers under the semi-closed speaker condition is caused by two factors: the difference in edited ratio, and the difference in the recording environment.

3) Feature Analysis: We analysed the effect of each feature set both in speaker semi-closed condition and speaker open condition. Here, we divided the features into three sets.

- Feature set 1: acoustic score feature.
- Feature set 2: mean of syllable duration, variance of syllable duration, score of statistical syllable duration model.
- Feature set 3: word identity, word length, word duration.

The effects of adding each feature set are shown in Fig. 5. As shown in Fig. 5, the syllable length features (mean and variance of syllable duration, score of statistical syllable duration model) were found to be the most effective of all the features. On the other hand, the acoustic features (HMM based log likelihood, posterior probability, number of competing words) and the word level features (word identity, word length, word duration) provided a supplementary effect.

C. Detection Results of Precise Parts

We conducted an experiment to detect precise parts, with the results shown in Fig.6. As shown in this figure, for whole speakers under the open speaker condition, the exact ratio without our detection method (recall=100%) is 83.7%. This improves to 87.1% by filtering 60% of the whole transcription (recall=60%). This corresponds to an improvement from 82.7% to 86.2% at the syllable level.

D. Experiment of Lightly Supervised Speaker Adaptation

This section describes the result of the lightly supervised speaker adaptation experiment using automatically extracted precise parts as training labels.

Maximum A Posteriori (MAP) estimation method is employed for speaker adaptation of HMMs[12]. All training and adaptation of HMMs were performed using the HTK HMM toolkit ver 3.4.1[13]. Because transcriptions of Japanese National Diet Record contains many ideographic characters, CRF-based Japanese morphological analyzer MeCab ver 0.96[2] (with UniDic ver1.3.12[3]) is employed to convert them into syllable sequence.

As the decoder of automatic speech recognition, we used SPOJUS++ [9], which has many novel features including a dynamic expansion of linear dictionary, a use of likelihood index for efficient handling of the inter-word dependency and one pass decoding. Its input speech analysis condition is shown in Table III. Its acoustic model is the 928 Japanese context-dependent syllable-based acoustic models with 8 left contexts (5 vowels, silence, /N/, and short pause including /q/), which were trained from academic presentation speech data in the CSJ [11]. Each continuous density HMM had 5 states, and 4 of them had pdfs of output probability. Each pdf consisted of 64 Gaussians with diagonal covariance matrices.

As the language model, a word-based trigram model with Witten-Bell backoff is trained on the Japanese National Diet Record contains 38,668K words in 1,083 meetings. Considering that the Japanese National Diet Record does not contains transcriptions of filled pause and silent pause, we applied our previously proposed filler prediction model[14] and pause insertion model[15] for estimating the probability of filled pause and silent pause. The filler prediction model consists of two sub-models: the filler insertion model and the filler selection model. The filler insertion model predicts places where filled pauses should be inserted, and is represented by a CRF. The filler selection model selects appropriate filled

[2]http://mecab.sourceforge.net/
[3]http://www.tokuteicorpus.jp/dist/

TABLE IV
RESULTS OF SPEAKER ADAPTATION

Adaptation method	Training Label Statistics			ASR Performance	
	# of labels	Prec. (%)	Coverage (%)	Cor. (%)	Acc. (%)
No adaptation	0	—	0	71.5	67.2
Using edited transcriptions	5,119 (90.7%)	82.7	76.7	74.0	69.5
Using extracted precise parts	4,221 (74.8%)	84.6	71.1	74.6	70.9
Using precise transcriptions	5,642 (100.0%)	100.0	92.6	76.2	71.6

pauses for given places and is represented by a simple conditional distribution. The pause insertion model predicts places where silent pauses should be inserted, and is represented by a CRF. These models are trained from academic presentation speech data and simulated public speech data in the CSJ. The CRF that represents the filler insertion model and the pause insertion model is trained with the toolkit CRF++[4], which uses LBFGS, a quasi-Newton algorithm for large scale numerical optimization problems, to estimate parameters and a Gaussian prior to avoid over-fitting. We used a combination of the preceding two words, current words, succeeding two words, their POSs, and the preceding two moras as features of the CRF.

We prepared 5,642 syllables of 3 speakers for training and 4,411 syllables for testing. The result is shown in Table IV. The column "*Prec.*" shows the ratio of the correct labels in the training data, and the column "*Coverage*" shows the ratio of learned syllables by the training data in the testing syllables. The first line "*No adaptation*" shows the baseline result without speaker adaptation. The second line "*Using edited transcriptions*" shows the another baseline result, in which the whole edited transcriptions are used for training labels. The fourth line "*Using precise transcriptions*" shows the upper bound of our proposed method, in which the precise transcriptions are manually prepared and used for training labels[5]. The third line "*Using extracted precise parts*" shows the result of our proposed method using the automatically detected precise parts under the condition $recall = 90\%$. As shown in Table IV, our proposed method achieves higher performance than baselines, although our proposed method uses less training data. It means that automatic detection of precise parts is effective to refine edited transcriptions for lightly supervised speaker adaptation of acoustic models.

V. CONCLUSION

In this paper, we proposed an automatic detection method of precise parts from inexact transcribed corpora. The evaluation experiments using the Japanese National Diet Record showed that our proposed method achieves 97.5% precision under the condition $recall = 80\%$. And more, the experiment showed that precise parts extracted automatically by our proposed method was effective as the training data of lightly supervised speaker adaptation of acoustic models.

[4]http://chasen.org/~taku/software/CRF++/

[5]Because editing operations of Japanese National Diet Record decrease words as shown in Table I, there are more training labels in the manually prepared precise transcriptions than the edited transcriptions.

Acknowledgements: We would like to thank the Global COE program "Frontiers of Intelligent Sensing" for supporting our research.

REFERENCES

[1] T. Watanabe, H. Nishizaki, T. Utsuro, and S. Nakagawa, "Unsupervised speaker adaptation using high confidence portion recognition results by multiple recognition systems," in *Proceedings of International Conference Spoken Language Processing*, 2004, pp. 1989–1992.

[2] L. Lamel, J. L. Gauvain, and G. Adda, "Investigating lightly supervised acoustic model training," in *Processing of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 477–480.

[3] N. Braunschweiler, M. J. F. Gales, and S. Buchholz, "Lightly supervised recognition for automatic alignment of large coherent speech recordings," in *Proceeding of INTERSPEECH-2010*, 2010, pp. 2222–2225.

[4] B. C. Roy, S. Vosoughi, and D. Roy, "Automatic estimation of transcription accuracy and difficulty," in *Proceeding of Interspeech*, 2010, pp. 1902–1905.

[5] I. Maruyama, Y. Abe, T. Ehara, and K. Shirai, "A study on detecting time of superimposing captions in documentary programs," in *Proceeding of the Autumn Meeting of Acoustical Society of Japan (ASJ)*, 1999, pp. 177–178, (in Japanese).

[6] T. Kudoh, "TinySVM," http://chasen.org/~taku/software/TinySVM/.

[7] X. Huang, A. Acero, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. Prentice Hall PTR, 2001.

[8] W. Lo, A. M. Harrison, and H. Meng, "Statistical phone duration modeling to filter for intact utterances in a computer-assisted pronunciation training system," in *Proceeding of International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 5238–5241.

[9] Y. Fujii, K. Yamamoto, and S. Nakagawa, "Large vocabulary speech recognition system: Spojus++," in *Proceeding of 11th WSEAS International Conference MUSP-11*, 2011.

[10] S. Nakagawa, K. Hanai, K. Yamamoto, and N. Minematsu, "Comparison of syllable-based hmms and triphone-based hmms in japanese speech recognition," in *Proceeding of International Workshop on Automatic Speech Recognition and Understanding*, 1999, pp. 393–396.

[11] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *Proceeding of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*, 2003, pp. 7–12.

[12] Y. Tsurumi and S. Nakagawa, "An unsupervised speaker adaptation method for cotinuous parameter hmm by maximum a posteriori probability estimation," in *Proceedings of ICSLP'94*, 1994, pp. 431–434.

[13] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department, 2006.

[14] K. Ohta, M. Tsuchiya, and S. Nakagawa, "Evaluating spoken language model based on filler prediction model in speech recognition," in *Proceedings of Interspeech2008*, Brisbane, Australia, September 2008, pp. 1558–1561.

[15] K. Ohta, T. Masatoshi, and S. Nakagawa, "Effective use of pause information in language modelling for speech recognition," in *Proceedings of Interspeech2009*, Brighton, UK, September 2009, pp. 2691–2694.