PAPER *Special Section on Corpus-Based Speech Technologies*

An Unsupervised Speaker Adaptation Method for Lecture-Style Spontaneous Speech Recognition Using Multiple Recognition Systems

Seiichi NAKAGAWA[†a)], ***Member***, **Tomohiro WATANABE**[†b)], ***Nonmember***, **Hiromitsu NISHIZAKI**[††c)], ***Member***, ***and*** **Takehito UTSURO**[†††d)], ***Nonmember***

SUMMARY This paper describes an accurate unsupervised speaker adaptation method for lecture style spontaneous speech recognition using multiple LVCSR systems. In an unsupervised speaker adaptation framework, the improvement of recognition performance by adapting acoustic models remarkably depends on the accuracy of labels such as phonemes and syllables. Therefore, extraction of the adaptation data guided by confidence measure is effective for unsupervised adaptation. In this paper, we looked for the high confidence portions based on the agreement between two LVCSR systems, adapted acoustic models using the portions attached with high accurate labels, and then improved the recognition accuracy. We applied our method to the Corpus of Spontaneous Japanese (CSJ) and the method improved the recognition rate by about 2.1% in comparison with a traditional method.

key words: *spontaneous speech recognition, unsupervised speaker adaptation, confidence measure, multiple LVCSR models*

1. Introduction

Since current speech recognizers' outputs are far from perfect and always include a certain amount of recognition errors, it is quite desirable to have an estimate of confidence for each hypothesized word. This is especially true for many practical applications of speech recognition systems such as keyword based speech understanding, and recognition error rejection confirmation in spoken dialogue systems. Most of previous works on confidence measure [1] are based on features available in a single LVCSR system. We experimentally evaluated the agreement among the outputs of multiple Japanese LVCSR models*, with respect to whether it is effective as an estimate of confidence for each hypothesized word [2]. Our previous study reported that the agreement between the outputs with two different acoustic models can achieve quite reliable confidence, and also showed that the proposed measure of confidence outperforms previously studied features for confidence measure such as the *acoustic stability* and the *hypothesis density* [1].

On the other hand, a speaker adaptation for acoustic models used in an LVCSR model, which is one of the methods to improve recognition performance, has been researched, and many adaptation techniques have been proposed [3]–[5]. When utterances of speakers that are used to adapt acoustic models can not be prepared in advance, an unsupervised adaptation technique may be effective on clean speech. We proposed an unsupervised speaker adaptation method using speech recognition results for a context-free grammar driven continuous speech recognition system and showed the effectiveness [6]. On the other hand, we also showed that an unsupervised speaker adaptation is not effective when syllable recognition rate is about 60–70%, in other words, when a recognizer does not use any language constraints. Therefore, unsupervised adaptation is very difficult with spontaneous speech such as lecture speech because of the very lower recognition accuracy.

Zhang et al. [3] have proposed a clustering technique of speakers and an unsupervised learning method of acoustic models using the speaker clustering results for recognizing news stories with many speaker turns in on-line processing of recognition. In [3], the recognition performances of the news stories ranged from 80% to 90% in word-based accuracy. Zhang et al. showed that the performance of unsupervised adaptation of acoustic models using highly accurate labels 90% or more in syllable-based accuracy, which was obtained from recognition results using a trigram language model, was almost equivalent to the recognition performance of supervised adaptation. However, the performance of spontaneous speech as the lecture-style ranges from 60% to 70% in word-based recognition accuracy, which correspond to about 70–80% in syllable-based recognition accuracy. Therefore, a traditional unsupervised speaker adaptation method is not suitable under such a condition. An adaptation approach using words or word sequences transcribed

Manuscript received July 1, 2004.
Manuscript revised September 28, 2004.

[†] The authors are with the Department of Information and Computer Sciences, Toyohashi University of Technology, Toyohashi-shi, 440–8580 Japan.

[††] The author is with the Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Kofu-shi, 400–8511 Japan.

[†††] The author is with the Graduate School of Informatics, Kyoto University, Kyoto-shi, 606–8501 Japan.

a) E-mail: nakagawa@slp.ics.tut.ac.jp
b) E-mail: watanabe@cl.ics.tut.ac.jp
c) E-mail: hnishi@yamanashi.ac.jp
d) E-mail: utsuro@pine.kuee.kyoto-u.ac.jp

DOI: 10.1093/ietisy/e88–d.3.463

*In this paper, we define "an LVCSR model" as a set of a decoder, a language model, and an acoustic model which is trained under the various conditions such as triphone, syllable, kinds of features. So we regard an LVCSR model as one of various LVCSR systems if decoders used in among the LVCSR systems are same, but the type of acoustic model is different from others.

by an LVCSR model which consist of only high confidence portions can be considered on the task of recognizing spontaneous speech.

Kemp et al. [4] proposed an unsupervised speaker adaptation method similar to bagging based on a high confidence score. Ogata et al. [5] proposed an unsupervised speaker adaptation method by extracting reliable portions based on *a posterior* probability. Yokoyama et al. [7] proposed an unsupervised speaker adaptation method based on unsupervisd batch-type topic adaptation for language models and unsupervised adaptation of acoustic models. However, there is still a gap on these methods between unsupervised-driven results and supervised-driven results. In [5], there was no difference between the recognition rates for unsupervised adaptation methods using all transcribed labels of training speech and using a part of transcribed labels which had high confidence based on *a posterior* probability. We guess that the accuracy of the labels based on *a posterior* probability described in [5] was low. Therefore, it needs to use labels having the higher accuracy for unsupervised adaptation to achieve more improvement of the recognition performance. We have proposed the high confidence measure described in [3], and it may be useful for unsupervised adaptation methods.

This paper describes an accurate unsupervised speaker adaptation method for lecture-style spontaneous speech recognition. We earlier proposed the method of how to look for the high confidence portions of transcriptions based on the agreement between two LVCSR models. As the result of a recognition experiment using acoustic models adapted by only using the high confidence portions of lecture speech, our unsupervised adaptation achieved the improvement of 2.1% in word-based accuracy in comparison with a traditional unsupervised adaptation method.

2. Specification of Japanese LVCSR Models

2.1 Decoders

We use the decoder named *Julius* Ver.3.3 among the Japanese LVCSR models, which is provided by the IPA Japanese dictation free software project [8], as well as the one named *SPOJUS*, which has been developed in our laboratory [9], [10]. Both decoders are composed of two decoding passes, where the first pass uses the word bigram, and the second pass uses the word trigram.

2.2 Acoustic Models

The acoustic models of Japanese LVCSR models are based on Gaussian mixture HMM. We evaluate phoneme-based HMMs as well as syllable-based HMMs. Speaker-independent acoustic models were trained by using read speech (about 20000 sentences uttered 180 male speakers; JNAS) and read speech & lecture-style spontaneous speech (115 lectures uttered by 115 male speakers; lecture speech).

2.2.1 Acoustic Models with the Decoder `Julius`

As the acoustic models used with the decoder Julius, we evaluate phoneme-based HMMs as well as syllable-based HMMs. The following two types of HMMs are evaluated: i) triphone model, and ii) syllable model [11]. Every HMMs are gender-dependent (male). The feature parameters consist of 12 dimensional mel frequency cepstrum coefficients (MFCC), delta 12 dimensionals, and delta powers (henceforth "MFCC-frm"). The sampling frequency is 16 kHz and the frame is shifted by 10 ms at every frame.

A typical triphone HMM consists of 5 states with 3 self-loops and 3 output distributions. Each distribution is composed of 16 Gaussian mixtures having diagonal covariance matrices. The total number of distributions is 2000. On the other hand, a typical syllable-based HMM consists of 7 states with 5 self-loops and 5 output distribution. Each distribution is composed of 16 Gaussian mixtures having diagonal covariance matrices. The total number of distributions is 600.

2.2.2 Acoustic Models with the Decoder `SPOJUS`

The acoustic models used with the decoder SPOJUS are based on syllable HMMs, which have been developed in our laboratory [12]. The acoustic models are gender-dependent (male) syllable unit HMMs. We evaluated five types of HMMs which differ in feature parameters: In 16 kHz sampling, 24 dimensional mel frequency cepstrum coefficients (MFCC) segmented from 4 successive frames of 12 dimensions (12 dim. × 4 frm. = 48 dim.) (lower dimensions reduction by K-L expansion) delta 12 dimensions calculated over 9 successive frames, delta delta 12 dimensions and delta, delta delta powers (henceforth "MFCC-seg"); 12 dimensional mel frequency cepstrum coefficients (MFCC), delta, delta delta 12 dimensions, and delta, delta delta powers (henceforth "MFCC-frm").

In 12 k sampling, 20 dimensional mel frequency cepstrum coefficients segmented from 4 successive frames of 12 dimensions (12 dim. × 4 frm. = 40 dim.) (lower dimensions reduction by K-L expansion), delta 10 dimensions calculated over 9 successive frames, delta delta 10 dimensions and delta, delta delta powers; The sampling frequency is 12 kHz or 16 kHz and the frame is shifted by 8 ms or 10 ms at every frame.

Each syllable-based HMM consists of 5 states with 4 self-loops and 4 output distributions. Each distribution is composed of 4 Gaussian mixtures having full covariance matrices.

2.3 Language Models

The language model is used from the Corpus of Spontaneous Japanese (CSJ) Project [13], which was trained using a text made by correctly transcribing lecture speech. From the training data which contain 612 lectures (1480834 words), 20000 high-frequent words are independently used to produce a word-based bigram/trigram model. The OOV rate for the evaluation set is 9.7%.

2.4 Combinations of Decoder and Acoustic Model/Language Model

Combinations of the following models are evaluated on the experiment to extract high confidence portions from the speech:

1. Decoder Julius (sampling frequency is 16 kHz, frame shift length is 10 msec) and triphone-based acoustic models,
2. Decoder Julius (16 kHz, 10 msec) and syllable-based acoustic models [11],
3. Decoder SPOJUS (feature vector is MFCC-seg, 16 kHz, 10 msec) and syllable-based acoustic models,
4. Decoder SPOJUS (MFCC-frm, 16 kHz, 10 msec) and syllable-based acoustic models,
5. Decoder SPOJUS (MFCC-seg, 16 kHz, 8 msec) and syllable-based acoustic models,
6. Decoder SPOJUS (MFCC-frm, 16 kHz, 8 msec) and syllable-based acoustic models, and
7. Decoder SPOJUS (MFCC-seg, 12 kHz, 8 msec) and syllable-based acoustic models.

2.5 Extraction of High Confidence Portions [2]

This section gives the extraction method of high confidence portions and the definition of our metric for evaluating confidence. In principle, the task of estimating confidence for each hypothesized word is to have an estimate of which words of the outputs of LVCSR models are likely to be correct and which are not reliable. In this paper, however, we focus on estimating correctly recognized words and evaluate confidence according to recall/precision rates of estimating correctly recognized words. The following gives a procedure for evaluating the agreement among the outputs of multiple LVCSR models as an estimate of correctly recognized words. First, let us suppose that we have two outputs Hyp_1 and Hyp_2 of two LVCSR models, each of which is represented as a sequence of hypothesized words. Next, two sequences Hyp_1 and Hyp_2 of hypothesized words are aligned by Dynamic Time Warping. Then, words that are aligned together and have an identical lexical form are collected into a list named agreed word list. Suppose that we have two sequences Hyp_1 and Hyp_2 of hypothesized words as below:

$$Hyp_1 = w_{11}, \cdots, w_{1i}, \cdots, w_{1k}$$
$$Hyp_2 = w_{21}, \cdots, w_{2j}, \cdots, w_{2l}$$

Then, the agreed word list is constructed by collecting those words w_{1i} $(= w_{2j})$ that satisfy the constraint: w_{1i} and w_{2j} are aligned together by DP matching, and w_{1i} and w_{2j} are lexically identical. Figure 1 illustrates an example of above procedure. The figure shows the case of outputs of the two LVCSR models, and the term "bon" commonly outputted from the two models is incorrect (substitution error). Finally, the following recall/precision rates are calculated by comparing the agreed word list with the reference sentence considering both the lexical form and the position of each word.

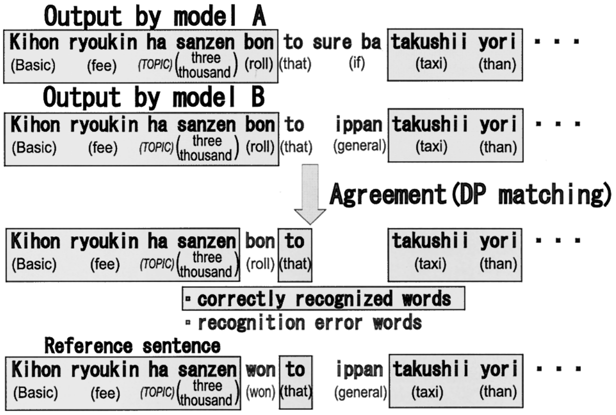

Output by model A
Kihon ryoukin ha sanzen bon to sure ba takushii yori ・・・
(Basic) (fee) (TOPIC) (three thousand) (roll) (that) (if) (taxi) (than)
Output by model B
Kihon ryoukin ha sanzen bon to ippan takushii yori ・・・
(Basic) (fee) (TOPIC) (three thousand) (roll) (that) (general) (taxi) (than)
Agreement(DP matching)
Kihon ryoukin ha sanzen bon to takushii yori ・・・
(Basic) (fee) (TOPIC) (three thousand) (roll) (that) (taxi) (than)
▪ correctly recognized words
▫ recognition error words
Reference sentence
Kihon ryoukin ha sanzen won to ippan takushii yori ・・・
(Basic) (fee) (TOPIC) (three thousand) (won) (that) (general) (taxi) (than)

Fig. 1 An example of agreement between two outputs.

$$Recall = \frac{\textit{\# of correct words in the agreed word list}}{\textit{\# of words in the reference sentences}}$$

$$Precision = \frac{\textit{\# of correct words in the agreed word list}}{\textit{\# of words in the agreed word list}}$$

In the case of Fig. 1, the recall and precision are 0.778 (= 7/9) and 0.875 (= 7/8) respectively.

3. Experimental Results for Speaker-Independent LVCSR

3.1 Training/Evaluation Data Sets

A training data set for acoustic modeling constitutes 115 lectures from the CSJ uttered by 115 male speakers at the meeting of the Acoustic Society of Japan (ASJ), because the speech evaluated in this paper was only recorded at the ASJ meetings, although the CSJ contains numerous lectures recorded at various academic meetings.

Four lecture speech utterances at the meeting of the Acoustic Society of Japan (ASJ) described as follows are used as the evaluation set: *a01m0035* (male speaker, 2610 words), *a01m0007* (male speaker, 2341 words), *a01m0074* (male speaker, 780 words), and *a05m0031* (male speaker, 1604 words). Sentence boundary detection for spontaneous speech such a lecture-style is a very difficult problem. In this experiment, sentence boundaries are automatically detected by the duration of pause. We use two kinds of pauses, one is long pauses more than 400 ms and the other is short pauses more than 200 ms. Table 1 summarizes the statistics of segmented utterances. A short utterance segmented by a threshold of 200 ms for pauses is composed of about 10 words on the average, on the other hand, a long utterances about 31 words.

3.2 Performance of Each LVCSR Model

Table 2 shows word-based recognition performances for the whole lecture speech (i.e., sentences in "total" in Table 1 were used). We evaluate two types of acoustic models, one trained using only lecture-style spontaneous speech described in Sect. 3.1 (denoted by "lecture speech" in Table 2), the other trained using read speech (180 persons, about 20000 sentences) from the Japanese Newspaper Article Sentences (JNAS) [14] (denoted by "JNAS" in Table 2).

By comparing Table 2 (a) and (b), we find that recognition of a long utterance is more difficult than that of a short utterance. Segawa et al. [15] proposed a continuous speech recognition method which does not need the explicit speech end-point detection while avoiding recognition of long utterances and showed the effectiveness in spoken dialog transcription experiments. So we use utterances segmented by the threshold of 200 ms hereafter. Using the acoustic models trained from lecture speech achieves improvement of recognition rates in comparison with the acoustic models by the JNAS, but it is not enough. Therefore, it is necessary to improve the recognition performances utilizing an unsupervised adaptation technique of the acoustic model. However, we can readily assume that the accuracy of the labels, which are necessary for an adaptation, is much poorer because of the low word accuracy as shown in Table 2. Thus, using only the labels extracted from the high-confidence portions of transcriptions of lecture speech may be very useful for adapting acoustic models, because these labels may be refined.

Table 1 Speech materials for evaluation.

(a) Number of long utterances segmented by long pause.

lecture ID	adaptation	test	total
a01m0035	100	31	131
a01m0007	100	121	221
a01m0074	50	11	61
a05m0031	100	60	160

(b) Number of short utterances segmented by short pause.

lecture ID	adaptation	test	total
a01m0035	300	273	573
a01m0007	300	352	652
a01m0074	150	69	219
a05m0031	300	117	417

Table 2 Comparison of the word-based recognition rate of an individual LVCSR model (the average of all evaluation data) [%].

(a) Long utterances segmented by long pause.

training data	JNAS		lecture speech	
model (acoustic model)	Cor.	Acc.	Cor.	Acc.
Julius (16 k, 10 ms, triphone)	52.7	43.2	64.3	55.9
Julius (16 k, 10 ms, syllable)	55.9	48.1	64.3	60.1
SPOJUS (16 k, 10 ms, seg)	55.7	50.1	60.7	53.5
SPOJUS (16 k, 10 ms, frm)	53.8	47.8	57.7	50.1
SPOJUS (16 k, 8 ms, seg)	45.6	38.8	62.4	54.1
SPOJUS (16 k, 8 ms, frm)	56.7	50.4	61.8	54.5
SPOJUS (12 k, 8 ms, seg)	57.7	51.4	62.3	55.3

(b) Short utterances segmented by short pause.

training data	JNAS		lecture speech	
model (acoustic model)	Cor.	Acc.	Cor.	Acc.
Julius (16 k, 10 ms, triphone)	64.3	55.9	68.9	61.1
Julius (16 k, 10 ms, syllable)	64.3	60.1	64.7	60.6
SPOJUS (16 k, 10 ms, frm)	52.9	48.1	60.8	56.1
SPOJUS (16 k, 10 ms, seg)	38.6	34.5	62.4	57.4
SPOJUS (16 k, 8 ms, frm)	56.8	50.2	64.7	58.5
SPOJUS (16 k, 8 ms, seg)	60.5	53.9	66.3	59.8
SPOJUS (12 k, 8 ms, seg)	58.2	52.0	63.9	57.4

4. Evaluation of Unsupervised Speaker Adaptation Using High Confidence

In an unsupervised speaker adaptation framework, the improvement of recognition performance by the adaptation remarkably depends on the accuracy of labels in syllable-formed or phoneme-formed portions.

4.1 Data Set

In the experiments of the following Sect. 3.2, each lecture speech described in Sect. 3.1 is divided into two sets, one for adaptation for adapting acoustic models, the other for evaluation of recognition performance. The training set contains 300 sentences in the first half of the lecture speech[†], while the remainder comprises the evaluation set (see Table 1 (b)).

(1) Adaptation data: the number of sentences used for adaptation is also shown in Table 1 (b), i.e. *a01m0035* (300 sentences, 3434 words), *a01m0007* (300 sentences, 1846 words), *a01m0074* (150 sentences, 1668 words), and *a05m0031* (300 sentences, 3566 words).

(2) Test data: the number of sentences used for evaluation is summarized in Table 1 (b), i.e. *a01m0035* (273 sentences, 2610 words), *a01m0007* (352 sentences, 2341 words), *a01m0074* (69 sentences, 780 words), and *a05m0031* (117 sentences, 1605 words).

4.2 Agreement between Multiple LVCSR Models

Agreement between outputs of multiple LVCSR models is defined as the agreement portions obtained by Dynamic Time Warping between the different outputs of two LVCSR models. Table 3 shows the performances of the agreement between outputs of two LVCSR models. The baseline in Table 3 indicates the result of the single LVCSR model (decoder Julius, sampling frequency is 16 kHz, frame shift of 10 msec, and acoustic model using triphone model) with the highest recognition performance for transcribing test speech among the single systems shown in Table 2.

Table 3 also summarizes following two results of the model combinations with the highest precision rate among 10 combinations in different decoders or among 11 combinations in the same decoder.

[†] However, the first 150 sentences are used for training only in the lecture speech "a01m0074", because of only 219 sentences in total.

Table 4 Word-based recognition rates of unsupervised adaptation (the average of four speakers). The number of training (adaptation) and test sentences is shown in Table 1 [%].

adaptation method	baseline		unsupervised		supervised	
LVCSR models	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
Julius-triphone, 16 kHz-10 msec	67.4	60.5	68.6	62.5	69.2	62.6
Julius-syllable, 16 kHz-10 msec	61.7	58.1	63.9	60.6	64.3	61.0
SPOJUS, MFCC-frm-16 kHz-10 msec	57.2	52.7	58.7	54.5	65.2	61.2
SPOJUS, MFCC-seg-16 kHz-10 msec	59.8	54.8	62.9	57.6	67.4	62.1
SPOJUS, MFCC-frm-16 kHz-8 msec	61.7	55.8	65.2	59.3	69.3	63.8
SPOJUS, MFCC-seg-16 kHz-8 msec	64.0	58.0	67.5	60.6	70.5	64.4
SPOJUS, MFCC-seg-12 kHz-8 msec	60.9	54.7	64.2	57.3	67.8	61.9

Table 3 Performance of the recognition rates of the portions based on the agreement between multiple LVCSR models (all evaluation data) [%].

$$Cor. = \frac{\textit{\# of correct words}}{\textit{\# of high confidence words}},$$

$$Acc. = \frac{\textit{\# of correct words–\# of insertion words}}{\textit{\# of high confidence words}},$$

where *Correct is equivlent to Precision*

(a) Recognition rates. (word)

LVCSR models	Cor.	Acc.
baseline (Julius-triphone)	68.9	61.1
agreement (same decoder)	86.5	85.2
agreement (different decoder)	87.8	87.6

(b) Recognition rates. (syllable)

LVCSR models	Cor.	Acc.
baseline (Julius-triphone)	82.2	75.5
agreement (same decoder)	94.3	90.0
agreement (different decoder)	94.5	92.2

1. Combination of different decoders
   - Decoder Julius (16 kHz, 10 msec) and triphone-based acoustic models.
   - Decoder SPOJUS (MFCC-seg, 16 kHz, 8 msec) and syllable-based acoustic models.
2. Combinations of same decoder
   - Decoder Julius (16 kHz, 10 msec) and triphone-based acoustic models.
   - Decoder Julius (16 kHz, 10 msec) and syllable-based acoustic models.

As shown in Table 3, the agreement among outputs of two LVCSR models has high confidence. The average recognition rates are about 90% in syllables. The rates are enough for unsupervised speaker adaptation as described in Sect. 1.

To show the effectiveness of our confidence measure [2], we compare our confidence measure with the other one. We extract words[†] with high confidence using Julius Ver.3.4 which can calculate confidence measure for each word based on *a posterior* probability [16]. The syllable-based accuracy based on Julius is 87.4% in these portions. This is worse than the accuracy (89.0%) based on our confidence masure as shown in Table 5. Therefore, we can claim that our confidence measure for unsupervised adaptation is more effective than the typical measure based on *a posterior* probability. In this work, these high confidence portions are used for speaker adaptation.

4.3 Results of Unsupervised Adaptation Experiments

To investigate our proposed approach, first, we compare our unsupervised adaptation approach which uses only the high confidence portions of the recognized labels from lecture speech with another approach using whole recognized labels from lecture speech. Those two approaches to speaker adaptation use a MAP adaptation technique [17] (mean vector & full covariance matrix) for the SPOJUS and an MLLR [3], [5] (only mean vector) for the Julius[††]. Table 4 shows the recognition performances when the speaker adaptation approach uses whole labels.

The performances for "baseline" indicate the results for the speaker-independent acoustic models, just as in Table 3. The rows of "unsupervised" and "supervised", on the other hand, denote the results of using adapted acoustic models. In Table 4, the unsupervised adaptation approach using whole labels slightly improves the recognition accuracy against the baseline. However, the unsupervised adaptation is no match for the supervised adaptation, especially, for SPOJUS.

Next, in the experiments on speaker adaptation only using the portions based on the agreement between the outputs of two LVCSR models, we select two combinations of the LVCSR models; the one is a combination between the same decoders, while the other is between different decoders. We investigated syllable correct/accuracy rates of the transcribed lecture speech as shown in Table 5. From the table, we can find that syllable sequences of only high confidence portions are refined, and its syllable-based correct and accuracy achieve improvements of 11.4% and 10.7%, respectively, in combinations between the same decoders. Furthermore, the accuracy performance of syllable sequences improves 2.2% or more in combinations between different decoders compared with combinations of the same decoder. In comparison with *a posterior* probability based confidence measure by Julius Ver3.4 (denoted as "Julilus-syllable (*pos-*

[†] 65% of all words were extracted. This is the same as the case of our confidence measure.

[††] We performed the MAP adaptation method for the SPOJUS and the global MLLR for the Julius. Those adaptation methods have been used for each LVCSR system. Especially, our MAP adaptation method can adapt the mean vectors and full covariance matrices for every our syllable-based HMM.

Table 6 Experimental results of unsupervised adaptation in word-based recognition rates by the single LVCSR model.
“high confidence adpt. (same)” = the agreement portion of Julius (Julius-triphone and Julius-syllable), “high confidence adpt. (diff)” = the agreement portion of Julius (16 kHz-10 msec-tri) and SPOJUS (MFCC-seg-16 kHz-8 msec) from the adaptation data described in Table 1 [%].

(a) Julius (triphone)

	baseline		unsupervised		high confidence adpt (same)		high confidence adpt (diff)		supervised	
speaker ID	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
a01m0035	62.1	53.1	63.1	55.0	62.8	54.2	63.0	54.0	63.5	55.0
a01m0007	71.8	64.3	72.5	64.5	72.4	64.7	72.8	65.4	72.6	64.8
a01m0074	78.5	73.3	77.6	73.7	77.8	74.4	78.0	73.9	78.5	74.5
a05m0031	62.1	53.1	61.4	57.1	61.2	56.6	61.4	57.3	62.1	56.9
average	67.4	60.5	68.6	62.5	68.6	62.5	68.8	62.7	69.2	62.8

(b) SPOJUS (MFCC-seg, 10 kHz, 8 ms).

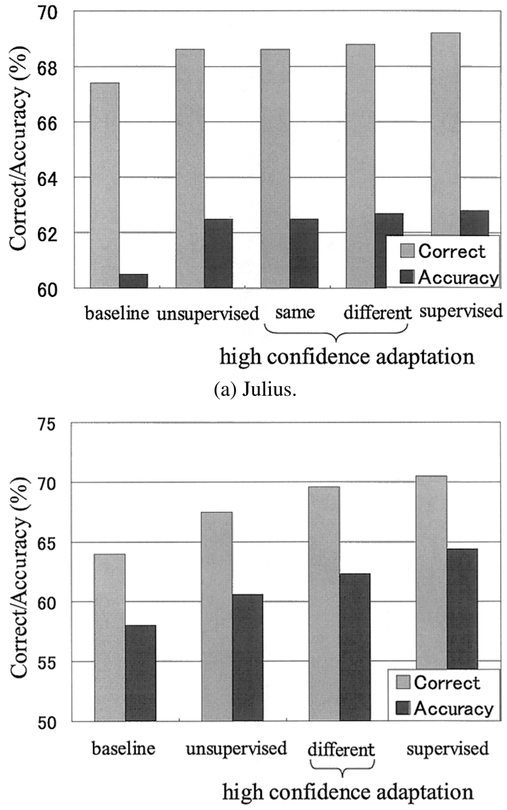
	baseline		unsupervised		high confidence adpt (diff)		supervised	
speaker ID	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.	Cor.	Acc.
a01m0035	57.8	50.4	59.8	52.0	63.1	55.2	63.4	55.9
a01m0007	69.5	62.0	71.4	63.7	72.2	65.0	72.5	65.4
a01m0074	73.9	69.1	78.2	71.0	79.2	73.0	80.1	74.4
a05m0031	54.9	50.3	60.4	55.7	62.8	58.4	66.1	62.0
average	64.0	58.0	67.5	60.6	69.2	62.7	70.5	64.4

Table 5 Recognition rates of adaptation data for speaker adaptation (syllables) [%].

LVCSR models	Cor.	Acc.	Prec.	Rec.
Julius-triphone (all portions)	82.9	76.1	—	—
Julius-syllable (all portions)	78.3	74.2	—	—
SPOJUS (all portions)[†]	80.4	74.5	—	—
agreement (same)	94.3	86.8	94.3	73.0
agreement (different)	94.5	89.0	94.5	73.2
Julius-syllable (*posterior* prob.)	89.4	87.4	89.4	60.0

teriori prob.)”), in addition, our confidence measure remarkably outperforms it in the syllable-based recognition rates.

Table 6 shows the recognition rates for the unsupervised adaptation approach using only high confidence portions. Figure 2 illustrates the average recognition rates in the table. In Table 6 and Fig. 2, the performance of “baseline”, “unsupervised” and “supervised” are equivalent to Table 4. It is surprising to find no significant difference between our proposed unsupervised adaptation and the supervised adaptation in recognition rate (especially, except for the speaker a05m0031), and the acoustic models adapted by our technique increase recognition performance to 62.7% from 58.0% against the baseline models in word-based accuracy for SPOJUS. We guess that this fact is caused by high precision as shown in Table 5. The results clearly indicate that our unsupervised adaptation technique, which uses only the labels forming the high confidence portions, is very effective for dictating spontaneous speech that is difficult to correctly transcribe using an LVCSR model. In addition, in spite of using two LVCSR systems in which the different adaptation techiniques are used, the learning effectiveness of the unsupervised adaptaion is nearly equal (except for a05m0031) to the one of the supervised adaptation. This proves the validity of our proposed technique.



(a) Julius.

(b) SPOJUS.

Fig. 2 Comparison of the adaptation methods.

5. Iteration of Unsupervised Adaptation

In the speaker adaptation experiment described in Sect. 4.3, we supposed that acoustic models are adapted in on-line

[†] The condition of feature parameters used in SPOJUS is “MFCC-seg-16 kHz-8 msec.”

processing, so the lecture speech was divided into training (adaptation) and test sets. This experiment aims at sequential adaptation of acoustic models. However, it is not necessary to prepare a training data set for adapting acoustic models and we used a best combination between two LVCSR models in off-line recognition processing.

Considering iteration in adapting acoustic models in off-line processing, all lecture speech sentences can be recognized, where the lecture speech does not need to be split into two data sets such as for training and test. So, in this case, the test data is equivalent to training data for adapting acoustic models. We look for the high confidence portions of whole sentences, and can adapt acoustic models using the reliable portions of the speech. Then, the sentences are recognized again by adapted acoustic models. The recognition performances in accordance with iterating speaker adaptation are shown in Table 7 and Fig. 3, in which the rates described in Table 7 (e) are the same as the graphs. The row of "baseline" corresponds to the results in Table 2 (b). The row of "unsupervised" denotes the results by unsupervised adaptation using whole of recognized labels. The number of iterations is up to 4 times[†]. The results are from the two LVCSR models: "Julius-triphone" and "SPOJUS-seg-16 kHz-8 msec". Adaptation methods are the MLLR for the Julius, and the MAP for the SPOJUS. The results termed "baseline" denote the performance for speaker-independent acoustic models, while the results named "supervised" are supervised adaptation in which the acoustic models are adapted by the test data. The recognition accuracies improve remarkably in the first iteration of adapting acoustic models. Whenever the number of iteration increases after the first adaptation, recognition accuracy rises slightly. As to the precision and recall rates of the agreement portions, the more the iteration number increases, the more the recall rate improves even if the precision rate remains virtually the same. We can consider this to be the improvement factor in the 2nd iteration. Those results indicate that the unsupervised speaker adaptation using the high confidence portions based on the agreement between two LVCSR models is also very effective in the iteration of speaker adaptation.

We should notice that there are still large differences between results by our proposed speaker adaptation and supervised adaptation, in spite of small differences in the case of Table 6. This is caused by the reason of the usage of the same data for adaptation and testing. In other words, acoustic models adapted by supervised learning is tuned for the adapted data (same as test data in this case).

Surprisingly, the recognition rates by unsupervised adaptation in Table 7 overperform the rates by supervised adaptation in Table 4.

Table 8 summarizes the statisics of coverage of syllables in adaptation data. For example, the first row "0" in "number of occurrences" denotes the number of syllables which did not appear in adaptation data. About half of syllables appear in adaptation data more than 10 times. We guess that more than 10 syllable samples are enough for roughly speaker adaptation for the syllable. There is no big differ-

Table 7 Experimental results for iteratively adaptating AMs in word-based recognition rates for unsupervised adaptation using only high cofidence portion.

training data = test data (all data): combination of different decoders [%], (J): Julius, (S): SPOJUS, (syl) = syllable.

(a) a01m0035.

adapt. methods	Julius-tri		SPOJUS-seg		agreement (syl)	
# of iteration	Cor.	Acc.	Cor.	Acc.	Prec.	Rec.
baseline	63.4	55.0	58.8	51.1	—	—
unsupervised	63.5	55.1	59.8	52.0	79.9	76.7 (J)
					77.9	72.4 (S)
1st iteration	63.6	55.3	64.8	57.2	93.4	65.5
2nd iteration	63.6	55.4	65.1	57.5	92.6	69.4
3rd iteration	63.7	55.4	65.5	57.7	92.9	70.1
4th iteration	63.7	55.5	65.6	57.9	92.9	70.0
supervised	63.9	56.4	73.9	68.8	100	100

(b) a01m0007.

adapt. methods	Julius-tri		SPOJUS-seg		agreement (syl)	
# of iteration	Cor.	Acc.	Cor.	Acc.	Prec.	Rec.
baseline	72.3	65.4	70.9	63.6	—	—
unsupervised	73.0	65.9	71.1	63.5	85.9	86.3 (J)
					85.2	84.4 (S)
1st iteration	73.5	66.7	74.6	68.4	95.1	79.2
2nd iteration	73.5	66.9	74.6	68.4	94.6	82.8
3rd iteration	73.6	66.9	74.6	68.4	94.4	82.7
4th iteration	73.6	66.9	74.6	68.5	94.4	82.7
supervised	73.8	67.2	81.3	77.2	100	100

(c) a01m0074.

adapt. methods	Julius-tri		SPOJUS-seg		agreement (syl)	
# of iteration	Cor.	Acc.	Cor.	Acc.	Prec.	Rec.
baseline	75.8	67.6	73.7	67.2	—	—
unsupervised	76.1	67.8	74.4	67.9	85.1	87.5 (J)
					84.7	84.8 (S)
1st iteration	77.5	69.7	79.0	72.1	94.9	79.8
2nd iteration	77.5	69.7	79.8	73.1	94.0	84.5
3rd iteration	77.5	69.8	80.4	73.6	94.0	85.1
4th iteration	77.5	69.8	80.6	73.8	94.0	84.9
supervised	77.7	70.0	85.2	80.2	100	100

(d) a05m0031.

adapt. methods	Julius-tri		SPOJUS-seg		agreement (syl)	
# of iteration	Cor.	Acc.	Cor.	Acc.	Prec.	Rec.
baseline	64.0	56.4	61.6	57.4	—	—
unsupervised	66.3	58.2	62.8	58.3	80.5	78.3 (J)
					80.1	75.0 (S)
1st iteration	67.1	60.6	68.7	65.1	93.7	67.3
2nd iteration	67.1	60.8	69.8	66.1	93.1	74.8
3rd iteration	67.2	60.8	70.6	66.8	93.3	75.4
4th iteration	67.2	60.9	70.7	67.0	93.3	75.4
supervised	67.4	61.2	78.8	76.1	100	100

(e) average.

adapt. methods	Julius-tri		SPOJUS-seg		agreement (syl)	
(# of iteration)	Cor.	Acc.	Cor.	Acc.	Prec.	Rec.
baseline	68.9	61.1	66.3	59.8	—	—
unsupervised	69.7	61.8	67.0	60.4	82.9	82.2 (J)
					82.0	79.2 (S)
1st iteration	70.4	63.1	71.8	65.7	94.3	73.0
2nd iteration	70.4	63.2	72.3	66.3	93.6	77.9
3rd iteration	70.5	63.2	72.8	66.6	93.7	78.3
4th iteration	70.5	63.3	72.9	66.8	93.7	78.3
supervised	70.7	63.8	79.8	75.6	100	100

[†] The performance converged in the iteration number 4, although we tried to perform more numbers of iteration.

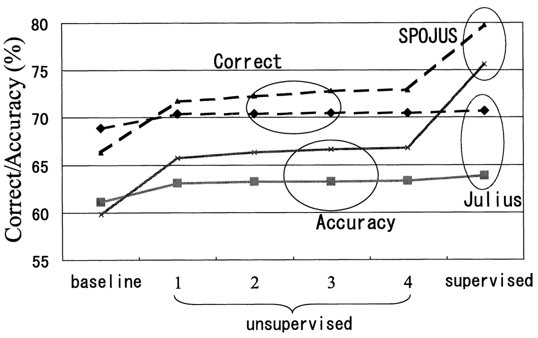


Fig. 3 Performances by increasing the number of iteration of adaptation.

Table 8 Coverage of syllables for speaker adaptation (SPOJUS, MFCC-seg-16 kHz-8 msec).

(a) All portions.

# of occurrences	a01m0035	a01m0007	a01m0074	a05m0031
0	30	32	33	28
1	4	6	5	9
2	3	2	1	4
3	4	3	5	0
4~5	3	1	5	1
6~10	11	5	13	9
11~	61	67	54	66

(b) High confidence portions.

# of occurrences	a01m0035	a01m0007	a01m0074	a05m0031
0	30	34	34	34
1	5	5	6	6
2	5	2	1	1
3	5	2	4	1
4~5	7	3	4	2
6~10	3	3	14	8
11~	61	67	53	64

ence on occurrence distributions between all portions and high confidence postions. This is the reason why unsupervised adaptation works well like supervised adaptation.

6. Conclusions

In this paper, we proposed the unsupervised speaker adaptation method for acoustic models. The speaker adapted acoustic models using labels from only the high confidence portions of a speech transcription remarkably improves the recognition performance, which is almost the same performance as the one when supervised adaptation is used.

In future work, we intend not only adapt acoustic models, but also language models using the high confidence portions.

References

[1] T. Kemp and T. Schaaf, "Estimating confidence using word lattices," Proc. EUROSPEECH'97, pp.827–830, 1997.

[2] T. Utsuro, H. Nishizaki, Y. Kodama, and S. Nakagawa, "Estimating high-confidence portions based on agreement among output of multiple LVCSR models," Syst. Comput. in Japan, vol.35, no.7, pp.33–39, 2004.

[3] W. Zhang and S. Nakagawa, "Continuous speech recognition using an on-line speaker adaptation method based on automatic speaker clustering," IEICE Trans. Inf. & Syst., vol.E86-D, no.3, pp.188–196, March 2003.

[4] T. Kemp and A. Waibel, "Unsupervised training of a speech recognizer: Recent experiments," Proc. EUROSPEECH'99, pp.2725–2728, 1999.

[5] J. Ogata and Y. Ariki, "Unsupervised acoustic model adaptation based on phoneme error minimization," Proc. 7th ICSLP, pp.1429–1432, 2002.

[6] S. Nakagawa and Y. Tsurumi, "An supervised speaker adaptation method for continuous parameter HMM by using maximum a posteriori probability estimation and recognition results," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J78-D-II, no.2, pp.188–196, Feb. 1995.

[7] T. Yokoyama, T. Shinozaki, K. Iwano, and S. Furui, "Unsupervised batch-type topic adaptation for language models," ASJ Spring Conference Record, 3-4-1, pp.129–130, 2003.

[8] T. Kawahara, T. Kobayashi, K. Takeda, N. Minematsu, K. Itoh, M. Yamamoto, A. Yamada, T. Utsuro, and K. Shikano, "Sharable software repository for Japanese large vocabulary continuous speech recognition," Proc. 5th ICSLP, pp.763–766, 1998.

[9] A. Kai, Y. Hirose, and S. Nakagawa, "Dealing with out-of-cocabulary words and speech disfluencies in an N-gram based speech understanding system," Proc. 5th ICSLP, pp.2427–2430, 1998.

[10] N. Kitaoka, N. Takahashi, and S. Nakagawa, "Large vocabulary continuous speech recognition using linear lexicon speaker with n-best approximation and tree lexicon search with 1-best approximation," IEICE Trans. Inf. & Syst. (Japanese Edition), vol.J87-D-II, no.3, pp.799–807, March 2004.

[11] M. Moroto, K. Yamamoto, and H. Matsumoto, "An improvement of syllabic models in large vocabulary continuous speech recognition," ASJ Spring Conference Record, 3-1-3, pp.95–96, 2001.

[12] S. Nakagawa and K. Yamamoto, "Evaluation of segmental unit input HMM," Proc. 21th ICASSP, pp.439–442, 1996.

[13] T. Kawahara, H. Nanjo, T. Shinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of Spontaneous Japanese," Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, pp.135–138, 2003.

[14] K. Itoh, M. Yamamoto, K. Takeda, T. Takezawa, T. Matsuoka, T. Kobayashi, K. Shikano, and S. Itahashi, "JNAS: Japanese speech corpus for large vocabulary continuous speech recognition research," J. Acoust. Soc. Jpn. (E), vol.20, no.3, pp.199–206, March 1999.

[15] O. Segawa, K. Takeda, and F. Itakura, "Continuous speech recognition without end-point detection," IEEJ Trans. EIS, vol.124-C, no.5, pp.1121–1126, 1997.

[16] A. Lee, K. Shikano, and T. Kawahara, "Real-time word confidence scoring using local posterior probabilities on tree trellis search," Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp.I-793–796, May 2004.

[17] S. Nakagawa and T. Koshikawa, "Speaker adaptation for continuous parameter HMM using maximum a posteriori probability estimation," J. Acoust. Soc. Jpn., vol.49, no.10, pp.721–728, March 1993.

Seiichi Nakagawa received his B.E., M.E. degrees from Kyoto Institute of Technology in 1971 and 1973, and D.Eng. degrees from Kyoto University in 1977. He has been a professor in the Department of Information and Computer Sciences at Toyohashi University of Technology since 1990. He was a visiting scientist in the Department of Computer Science at Carnegie-Mellon University in 1985–1986. He received 1997 and 2001 Paper Awards from IEICE and the 1988 JC Bose Memorial Award from the Institution of Electronics Telecommunication Engineers. His major research interests include automatic speech recognition/speech processing, natural language processing, human interface, and artificial intelligence.

Tomohiro Watanabe was born in 1979. He received his B.E. in information and computer sciences from Toyohashi University of Technology in 2002, and is now student of master course in the Department of Information and Computer Sciences at Toyohashi University of Technology. His research interests include spoken language processing.

Hiromitsu Nishizaki was born in 1975. He received his B.E., M.E., and D.Eng. degrees in information and computer sciences from Toyohashi University of Technology in 1998, 2000, and 2003. He is now a research associate in the Interdisciplinary Graduate School of Medicine and Engineering at University of Yamanashi. His research interests include spoken/natural language processing.

Takehito Utsuro received his B.E., M.E., and D.Eng. degrees in electrical engineering from Kyoto University in 1989, 1991, and 1994. After serving at Nara Institute of Science and Technology and Toyohashi University of Technology, he has been a lecturer in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University, since 2003. He was a visiting schoar in the Department of Computer Science at Johns Hopkins University in 1999–2000. His professional interests in natural language processing, spoken language processing, machine learning, and artificial intelligence.