Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition

AUTHOR(S):

Nanjo, H.; Kawahara, T.

Language Model and Speaking Rate Adaptation for Spontaneous Presentation Speech Recognition

Hiroaki Nanjo and Tatsuya Kawahara, *Member, IEEE*

***Abstract*—The paper addresses adaptation methods to language model and speaking rate (SR) of individual speakers which are two major problems in automatic transcription of spontaneous presentation speech. To cope with a large variation in expression and pronunciation of words depending on the speaker, firstly, we investigate the effect of statistical and context-dependent pronunciation modeling. Secondly, we present unsupervised methods of language model adaptation to a specific speaker and a topic by 1) selecting similar texts based on the word perplexity and TF-IDF measure and 2) making direct use of the initial recognition result for generating an enhanced model. We confirm that all proposed adaptation methods and their combinations reduce the perplexity and word error rate. We also present a decoding strategy adapted to the SR. In spontaneous speech, SR is generally fast and may vary a lot. We also observe different error tendencies for portions of presentations where speech is fast or slow. Therefore, we propose a SR-dependent decoding strategy that applies the most appropriate acoustic analysis, phone models, and decoding parameters according to the SR. Several methods are investigated and their selective application leads to improved accuracy. The combined effect of the two proposed adaptation methods is also confirmed in transcription of real academic presentation.**

***Index Terms*—Acoustic modeling, language model adaptation, pronunciation modeling, speaking rate, spontaneous speech recognition.**

I. Introduction

AUTOMATIC speech recognition (ASR) accuracy of read speech exceeds 90% for a dictation system. The system, however, assumes that users clearly utter grammatically correct sentences with orthodox pronunciation for human-to-machine interfaces. On the other hand, the ASR of human-to-human spontaneous speech, which would make possible the automatic transcription or translation of lectures and meetings, is very poor and needs more extensive researches.

To further this end, the five-year project "Spontaneous Speech Corpus and Processing Technology" has been conducted since 1999 [1]. The foremost product of the project is a large-scale spontaneous speech corpus [2]. The Corpus of Spontaneous Speech (CSJ) consists of roughly seven million words. Monologues such as lecture presentations and extemporaneous speeches are mainly recorded. The main goal of this research is the automatic transcription of live talks such as oral presentations for efficient digital archiving.

Manuscript received May 4, 2003; revised February 20, 2004. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Shrikanth Narayanan.

The authors are with Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan (e-mail: nanjo@ar.media.kyoto-u.ac.jp).

Digital Object Identifier 10.1109/TSA.2004.828641

As many previous studies point out, various factors in spontaneous speech affect ASR performance. They include acoustic variation caused by fast speaking and imperfect articulation, and linguistic variation such as colloquial expressions and disfluencies. Using the large-scale database of CSJ, Shinozaki and Furui investigated the correlations of various factors with speech recognition accuracy [3]. They concluded that the speaking rate (SR), out-of-vocabulary (OOV) rate, and the self-repair rate (RR) are directly correlated with accuracy. Other factors are mainly dependent on one of these three. For example, word perplexity (PP) is correlated with the OOV rate. In this work, we are more concerned with PP instead of OOV because perplexity is a widely used measure. Shinozaki and Furui also showed that PP and SR have large correlations with WER and we also confirmed similar tendency using a different test set [4]. Based on the analysis, we address adaptation methods for variations of PP and SR among speakers, which are the two major factors in spontaneous speech. Especially in presentation speeches that have relatively long durations, these two variations are prominent problems, and can be approached by considering the characteristics of the presentation speech.

At first, we examine pronunciation modeling for spontaneous speech, especially the effect of statistical and context-dependent model of pronunciation variation. Then, we propose unsupervised adaptation of the language model, which integrates the pronunciation model. Conventional studies on language model adaptation [5], [6] mainly deal with adaptation to the specific domains or topics. As for lecture presentations, adaptation to each speaker is required because the preference of expressions and their pronunciation are quite different among speakers. Fortunately, lecture presentations and their transcriptions are relatively long, which will make the speaker adaptation possible. Several methods are proposed and their combinations are explored in order to adapt to the speaker's topic, characteristics in expression and pronunciation as a whole.

We also focus on the problem of SR, which is another significant cause of accuracy degradation. In particular, fast speaking often causes poor acoustic matching. Occasionally, some phones themselves may disappear. Thus, it is necessary to cope with fast speech segments. Moreover, we observe that there are frequent changes in SR within a single oral presentation and different error tendencies for fast and slow speech segments. Therefore, handling slow speech as well as fast speech should be considered. There have been studies that consider the factor of SR [7], [8]. Some studies apply a uniform method to all utterances, but they cannot cope with the frequent changes of SR. There have also been studies that adopt dedicated acoustic analysis [9], [10] or pronunciation entries

[11], [12] for fast speech. These methods have effect on some specific portions of utterances. However, any single method is not effective for all varieties of utterances. In this paper, we propose combinations of techniques of acoustic analysis, phone modeling, and decoding parameters and their selective application depending on the SR.

In this paper, this SR-dependent decoding is combined with the language model adaptation as they are expected to give a combined effect in improving the recognition performance of spontaneous presentation speech.

II. Baseline System and Test Set

A. Corpus and Test Set

The CSJ developed by the project consists of academic oral presentations at technical conferences and extemporaneous public speeches on given topics such as hobbies and travels. In this paper, the test-set which consists of 15 academic oral presentations by male speakers is used because the CSJ does not include enough female speakers' speech for acoustic model training as of January 2003. The specifications are shown in Table I. We divide the recorded materials into the utterances based on pause labels by human. They do not necessarily match the linguistic sentences. The total number of utterances is 4603.

TABLE I
Test-Set Presentations

presentation ID	#words	duration (min.)	PP	RR	SR
A01M0097	2592	14.4	39.6	0.62	8.65
A04M0051	2581	14.6	52.5	2.02	8.03
A04M0121	2964	15.4	87.1	3.31	8.77
A03M0156	3243	16.6	86.3	0.90	9.78
A03M0112	3254	15.1	53.8	0.52	9.62
A01M0110	1307	9.6	81.5	2.07	8.25
A05M0011	2791	16.0	72.8	1.33	6.90
A03M0106	3091	13.0	72.8	1.07	10.66
A01M0137	2073	11.1	56.0	1.69	8.64
A04M0123	2619	12.3	52.1	0.84	9.06
A01M0056	2364	11.4	41.6	0.85	8.77
A02M0012	4034	22.2	95.8	0.45	8.45
A06M0064	2399	12.5	81.2	0.33	7.30
A01M0141	2334	15.4	72.4	2.37	8.45
A03M0016	3171	15.7	61.5	1.82	10.20
total	40817	215.3	65.8	1.29	8.77

PP: word perplexity, RR: self repair rate (%),
SR: speaking rate (mora/sec.)

B. Language Model

For language model training, all transcribed data available (as of January 2003) is used. There are 2592 talks excluding the test set and the text size in total is 6.67 M words (= Japanese morphemes).

We trained a backoff word trigram model as a baseline language model using CMU-Cambridge SLM toolkit version 2 [13]. In Japanese texts, words are not delimited with spaces, and transcription of the CSJ was done manually both in an orthographic notation form and a phonetic (*kana*) form for each utterance unit. Thus, automatic alignment of the two by the word unit is needed to obtain the word-pronunciation entries. This was incorporated as a post-processor of the morphological analyzer [14]. Some heuristic thresholding is applied to eliminate erroneous patterns. We selected 30 820 word-pronunciation entries that were found more than three times in the training data. In the 30 820 word-pronunciation entries, there were 24 437 distinct word entries which were defined as the vocabulary. Test set OOV rate is 1.2% with this vocabulary. In the baseline language model, the variation of pronunciation is handled simply by adding multiple entries in the dictionary. The pronunciation entries for each word are extracted from the 30 820 word-pronunciation entries.

C. Acoustic Model

As for acoustic model training, we use only academic presentation speeches by male speakers because the test set was only by male speakers. We use 781 presentations that amount to 106 hours of speech.

Acoustic models are based on continuous density Gaussian-mixture HMM. Speech analysis is performed every 10 ms and a 25-dimensional parameter is computed $(12\ \mathrm{MFCC} + 12\ \Delta\ \mathrm{MFCC} + \Delta\ \mathrm{Power})$. The number of phones used is 43, and all of them are modeled with left-to-right HMM of three states and state-skipping transitions are not allowed in the baseline. We trained context-dependent triphone models. Decision-tree clustering was performed to set up 3000 shared states. We also adopted phonetic tied-mixture (PTM) modeling [15], where triphone states of the same phone share Gaussians but have different weights. Here, 129 codebooks of 192 mixture components were used.

D. Results by the Baseline System

These models are integrated with the large vocabulary speech recognition decoder Julius rev.3.3p3 that was developed at our laboratory [16].

The average word error rate (WER) with the baseline system is 31.4% and the average test-set perplexity is 65.8. For the test-set perplexity computation, OOV words are not counted but pause marks, which are used instead of periods or comma, are included. For the calculation of WER, OOV words are included but pause marks are excluded. The total number of pause marks is 4572.

III. Pronunciation Modeling

First, we investigate the modeling of pronunciation variation. The baseline system coped with the pronunciation variation by simply adding pronunciation variant entries to the dictionary. The method has some significant disadvantages. False matching is increased especially with short functional words which tend to have more pronunciation variations. False matching resulted in a tremendous increase of WER because all pronunciation variants were added to the dictionary; that is, the number of unigram entries of the language model was 24 437 and the number of pronunciation entries was 30 820. Therefore, pronunciation entries whose occurrence probability in each lexical item is lower than a threshold are eliminated. The result is shown in the left portion of Fig. 1. When the threshold is set to a small value (0.01), false matching is increased and so is WER. On the other hand,

when the threshold is set to a large value (0.3), which is similar what occurs when single pronunciation for each word in the lexicon is used ("1-pron. per word" in Fig. 1),[1] WER is also increased (31.6%) because of removal of appropriate pronunciation entries. The optimal threshold is 0.2 (number of pronunciation entries: 25 161), which is used in the baseline system (WER: 31.4%) in this paper.

Next, we introduce an approach in which the pronunciation variation modeling is combined with the statistical language model. Usually, this approach is implemented using a "unigram modeling of pronunciation variation (pron-unigram)" (probability of baseform entries) for each lexical entry [17], [18].

Here we adopt "trigram modeling of pronunciation variation (pron-trigram)" that predicts pronunciation of a word based on the previous words [19]. Specifically, we use 30 820 word-pronunciation entries as a token of trigram language model that enables us to predict a word together with its pronunciation considering a contextual effect. The result is shown in the right-most portion of Fig. 1. We reduced WER with a statistical modeling of pronunciation variation and we achieved the largest WER reduction with the trigram modeling of pronunciation variation (pron-trigram: 30.5%). The WER improvement (0.9%) from the baseline pronunciation modeling is significant of 1% level, where 2-sample test for equality of proportions was performed (sample size $N = 40\,817$). Comparing the unigram and trigram modeling of pronunciation variation (pron-unigram and pron-trigram), we obtained a slight improvement with the pron-trigram although the number of language model entries is increased from 24 437 to 30 820 in the pron-trigram modeling and so is the test-set perplexity from 65.8 to 74.9.

In the following sections, we use the language model with trigram modeling of pronunciation variation, which is the model that is simple and has the highest accuracy.

IV. Unsupervised Language Model Adaptation to Topic and Speaker

Next, we address adaptation of the language model to individual speakers by combining the pronunciation model in order to reduce the PP and WER. We pay close attention to an unsupervised adaptation which requires only test speeches and original training texts for the baseline model.

Conventional studies on language model adaptation have been mainly directed at adapting the model to the specific domains or topics using a trigger model [20] or a cache model [21], [22]. In spontaneous speech recognition, however, adaptation to individual speakers is also required because the preference of expressions and their pronunciation is quite different among speakers. Academic presentations especially have relatively longer speech and their transcriptions, which will enable the adaptation to the speaker.

The proposed adaptation methods are performed in an off-line (not dynamic) manner, thus are more robust against recognition errors than the on-line adaptation methods because the whole recognition result can be used to suppress influences of local recognition errors. The pronunciation variation model is also adapted in the framework.

[1]For each baseform, the most frequent pronunciation in the training data is assigned, thus, the pronunciation may differ from canonical one.

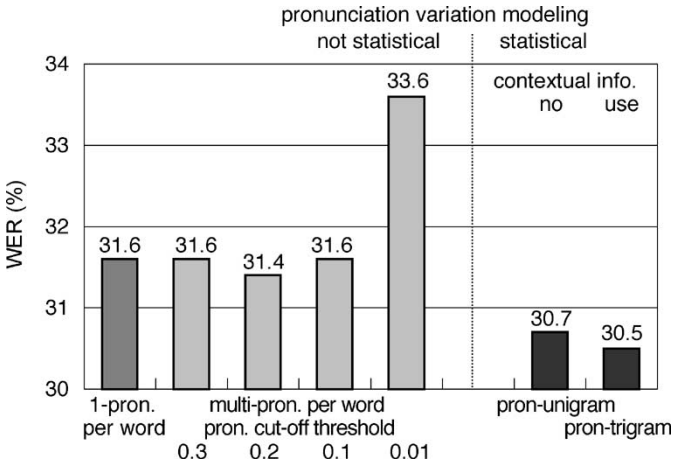

Fig. 1. Comparison of pronunciation modeling.

A. Language Model Adaptation Based on Text Selection

In this section, we present an adaptation method to enhance the language model by weighting texts similar to a test-set presentation based on the initial recognition result. There is a method that selects similar texts based on *a priori* knowledge such as the use of preprints of the corresponding presentation and transcriptions of former presentations by the same speaker. We once tried incorporating preprinted texts for adaptation [23] and improved accuracy by 0.5% to 3.0% absolute for other test presentations of the CSJ. However, we cannot assume that preprints are always available. In this paper, we explore a method without *a priori* knowledge. Specifically, we use PP and statistics of content word occurrences as a similarity measure [24], [25].

1) Language Model Adaptation Scheme: For adaptation, we perform linear interpolation according to

$$P_{SA\_LM}(w) = \lambda \cdot P_{SIMILAR\_LM}(w) + (1-\lambda) \cdot P_{SI\_LM}(w) \quad (1)$$

where $P_{SI\_LM}(w)$ is a probability for word sequence w by the speaker-independent language model (=baseline language model), and $P_{SIMILAR\_LM}(w)$ is a probability by the language model trained with small texts which are similar to the test speaker's presentation and selected based on TF-IDF or PP. As a result of linear interpolation, the speaker-adapted language model SA_LM, which gives a probability $P_{SA\_LM}(w)$ for word sequence w, is generated. The interpolation coefficient λ is estimated using EM algorithm denoted in

$$\hat{\lambda} = \frac{1}{N} \sum_{i=1}^{N} \frac{\lambda \cdot P_{SIMILAR\_LM}(w_i)}{\lambda \cdot P_{SIMILAR\_LM}(w_i) + (1-\lambda) \cdot P_{SI\_LM}(w_i)} \quad (2)$$

where w_i is the ith word of the correct transcription of the corresponding test presentation, which is actually unavailable. In this paper, a development-set is used for the estimation of λ. Fifteen test-set presentations were randomly divided into three groups GA, GB, and GC. Then, for testing with five presentations of

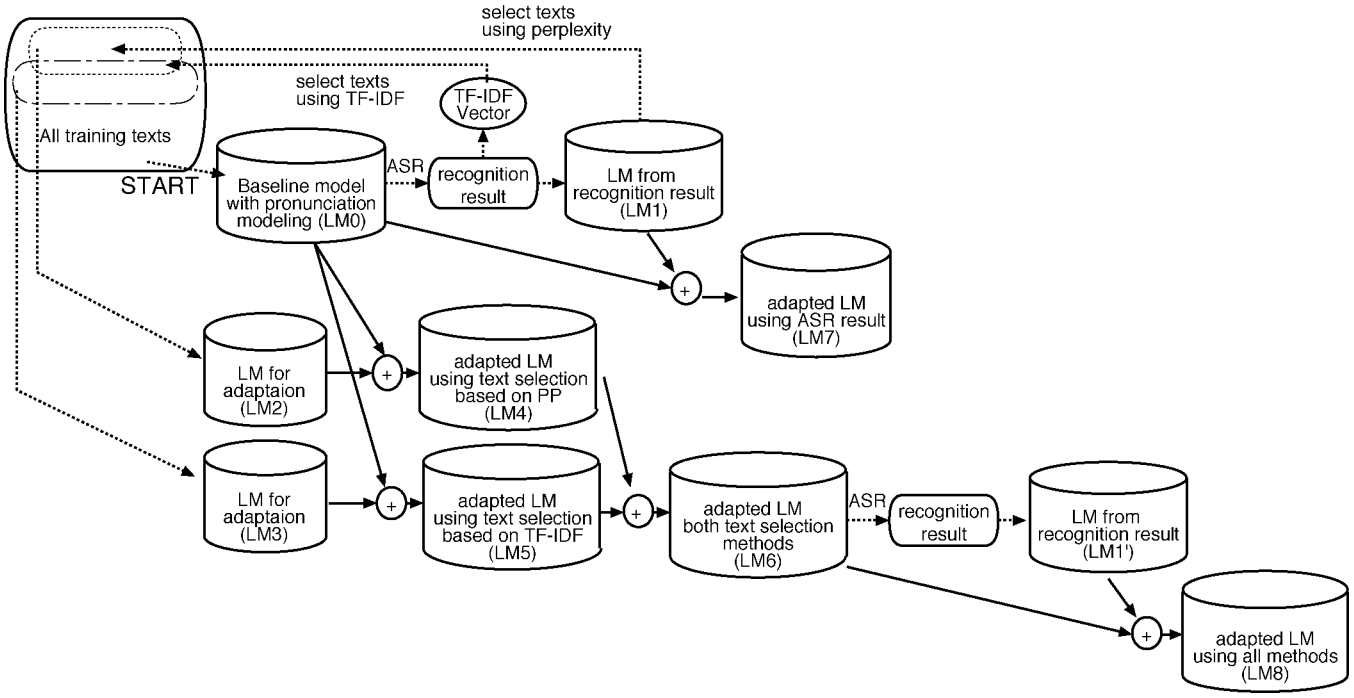

Fig. 2. Flowchart of language model adaptation.

each group, ten presentations of the other two groups are used as a development set so as to estimate the interpolation coefficient λ. For example, when we adapt the language model to a certain presentation X, which belongs to group GA, we used a value of λ which was estimated with ten presentations of GB and GC. The λ estimation is performed according to (2) for each presentation of the development set until convergence, and then their average is set to the final λ value.

2) Text Selection Using Word Perplexity: At first, we defined PP as a measure of similarity and then selected similar texts. The adaptation process is shown in Fig. 2, and described in detail below.

A) Using the speaker-independent language model (*LM*0), ASR is performed to generate an initial recognition result.

B) A language model *LM*1 using the initial recognition result is set up for text selection.

C) PP of each training text (by utterance unit) is computed using *LM*1, and texts that have lower perplexity than a threshold *th* are selected and a language model *LM*2 is generated.

D) Linear interpolation with *LM*0 is performed to generate an adapted model *LM*4. In this process, *LM*0, *LM*2 and *LM*4 correspond to SI_LM, $SIMILAR\_LM$ and SA_LM in (1), respectively.

The values of the threshold *th* and coefficient λ are set up to minimize the development-set perplexity. With the adapted model (*LM*4), the overall test-set perplexity was reduced from 74.9 to 68.7. We also reduced WER by 0.8% absolute, as shown in Table II. We investigated texts selected for adaptation and found that they contained many carrier and filler phrases which often appear at the beginning and end of sentences in Japanese, such as "*desu ne*", "*de ano*:", and "*e: ma*:". They are considered to represent a speaker's characteristic expression. Also, they vary among speakers. These data suggest that the proposed adaptation method properly extracts such features.

3) Text Selection Using TF-IDF: In this section, we explore another text-selection method based on TF-IDF which is widely used as an information retrieval measure. Term frequency (tf_{ij}) is defined as the occurrence counts of a word n_j in a document d_i. Inverse document frequency (idf_j) is defined as the total number of documents divided by the number of documents containing the word n_j. For each noun n_j of document d_i, $a_{ij} = tf_{ij} * \log(idf_j)$ is calculated and a vector $A_i = (a_{i1}, a_{i2}, \ldots, a_{iN})$ for the document d_i is generated. Similarity $S_{x,y}$ between two documents x and y is defined as the cosine of the angle between the corresponding vectors as denoted in (3).

$$S_{x,y} = \frac{(A_x \bullet A_y)}{(||A_x|| \cdot ||A_y||)}. \tag{3}$$

The adaptation process is described below and also shown in Fig. 2.

A) Using the speaker-independent language model (*LM*0), ASR is performed to generate an initial recognition result.

B) The TF-IDF vector for the test-set presentation [A_x in (3)] is calculated.

C) Text (document unit, which is the whole transcription of one presentation) Y that has higher similarity $(S_{x,y})$ than a threshold *th* is selected and a language model *LM*3 is generated.

D) Linear interpolation with *LM*0 is performed to generate an adapted model *LM*5. In this process, *LM*0, *LM*3 and *LM*5 correspond to SI_LM, $SIMILAR\_LM$ and SA_LM in (1), respectively.

Also, the values of the threshold *th* and coefficient λ are set up to minimize the development-set perplexity. With the adapted model (*LM*5), the overall test-set perplexity was reduced from 74.9 to 70.2 and WER was reduced by 1.4% absolute, as shown in Table II. Selected texts with the TF-IDF measure contain a lot of topic-dependent words. This shows that the language model is adapted to the topic. Compared with the text selection based on PP, the text selection based on TF-IDF achieved larger WER reduction although the test-set perplexity reduction is smaller. This suggests that the content words have more effect on overall WER in spite of less frequent occurrence.

4) Combination of Text-Selection Methods: Then, a combination of both text-selection methods is performed. The language model is adapted to the speaker's characteristic expression (*LM*4) using PP as a similarity measure, while it is adapted to the topic (*LM*5) using TF-IDF as a similarity measure. Interpolating these models (*LM*6), the language model is adapted to both speakers' characteristics and the topic. The interpolation coefficient λ is also set up to minimize the development-set perplexity. The result is also shown in Table II. The combination effect is confirmed. With the unsupervised language model adaptation based on the text selection, perplexity, and WER are reduced by 13.1% and 5.6%, respectively, from the *LM*0. A WER of 28.8% was achieved.

TABLE II
RESULT OF LANGUAGE MODEL ADAPTATION USING TEXT SELECTION

text selection method	WER(%)	perplexity
$LM0$	30.5	74.9
perplexity ($LM4$)	29.7	68.7
TF-IDF ($LM5$)	29.1	70.2
both ($LM6$)	28.8	65.1

TABLE III
RESULT OF LANGUAGE MODEL ADAPTATION USING INITIAL RECOGNITION RESULT

	WER(%)	perplexity
$LM0$	30.5	74.9
adapted using ASR result ($LM7$)	28.8	51.8

TABLE IV
EFFECT OF COMBINATIONS OF PROPOSED LANGUAGE MODEL ADAPTATION METHODS

adaptation method	WER(%)	perplexity
baseline	31.4	65.8
pronunciation ($LM0$)	30.5	74.9
text selection ($LM6$)	28.8	65.1
using ASR result ($LM7$)	28.8	51.8
adapted using all methods ($LM8$)	27.6	46.7

B. Language Model Adaptation Using Initial Recognition Result

Next, we introduce another adaptation method by making direct use of the initial recognition result. We do this because presentations are relatively long and their transcriptions contain the speaker's characteristic expressions and topics. The adaptation process is shown in part of Fig. 2.

The backoff word trigram model (*LM*1) trained with the initial recognition result contains several errors. Thus, bigram and trigram entries found only once are discarded. This *LM*1 is the same as the one described in the previous section except there is a cutoff of bigram and trigram entries. The interpolation is performed for adaptation in the same manner as described in the previous section. In this process, *LM*0, *LM*1, and *LM*7 correspond to SI_LM, $SIMILAR\_LM$, and SA_LM in (1), respectively. For this procedure (interpolation of *LM*0 and *LM*1), we use the complementary backoff algorithm [16], which works well when there is a large difference in the N-gram entries between the models. The result is shown in Table III. The simple adaptation method using the ASR result reduced WER by 1.7% absolute.

Finally, all proposed adaptation methods are combined. This process is described below.

(A) Using the adapted model by both text-selection methods, speech recognition is performed again and a more accurate ASR result is generated.
(B) A word trigram model (*LM*1′) is trained with the ASR result.
(C) Linear interpolation with *LM*6 is performed to generate an adapted model *LM*8. In this process, *LM*6, *LM*1′ and *LM*8 correspond to SI_LM, $SIMILAR\_LM$ and SA_LM in (1), respectively.

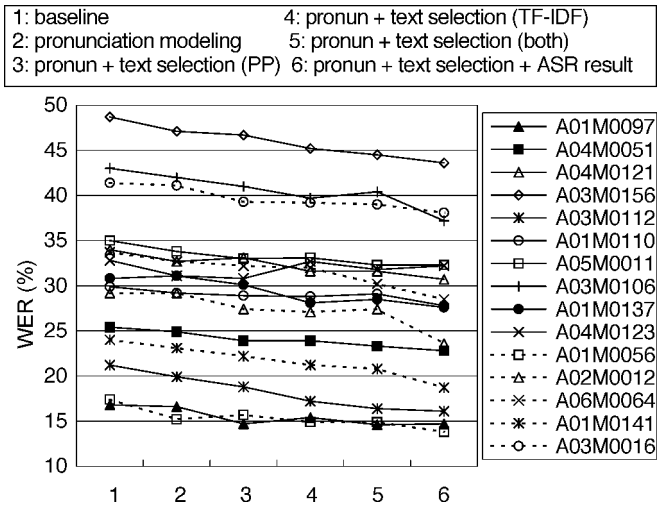

Fig. 3. WER for each test presentation.

The results are summarized in Table IV. The combination effect was confirmed and we got a WER of 27.6%. Fig. 3 shows WER for each test presentation with the proposed methods. A WER reduction of 5.8% to 0.6% absolute is achieved. Even for the test presentations that have higher WERs such as A03M0156, A03M0106, and A03M0016, the proposed methods could reduce WER; namely, it worked robustly with respect to recognition errors.

V. ACOUSTIC MODELING AND DECODING CONSIDERING SPEAKING RATE

In acoustic modeling of spontaneous speech, the SR, especially the fast speech, is considered to be one of the most significant causes of degradation [3]. Fast speaking often causes

incomplete articulation, thus poor acoustic matching. The spectral patterns change and moreover the phone itself may disappear. In this section, we present several acoustic modeling for fast speech and a decoding strategy depending on the current SR. It is also observed that there are frequent changes in SR in a single presentation. These changes cause significant problems when decoding with uniform models and parameters. Actually, it has been found that the tendency of recognition errors differs for fast and slow utterances. Thus, we propose to selectively apply appropriate decoding methods according to the SR.

A. Analysis of Speaking Rate

1) Distribution of Speaking Rate: Distributions of SR in spontaneous speech corpus (CSJ—academic presentation speech: 35 h) and read speech corpus (JNAS—newspaper reading: 40 h) are plotted in Fig. 4. SR is estimated for every utterance that is segmented manually based on long pauses and defined as the mora count divided by the utterance duration (in seconds). For both corpora, phonetic transcription is given manually, thus used for defining morae. An utterance whose SR is x mora/s is classified to N mora/s, where $N \leq x < N+1$.

The mean and standard deviation of the SR of the JNAS corpus are 6.27 and 0.97 and those of the CSJ are 8.70 and 2.10, respectively. It is confirmed that spontaneous speech (CSJ) is faster than read speech (JNAS) and the SR variation of spontaneous speech is larger than that of read speech.

2) Distribution of Phone Duration: Distribution of phone duration in spontaneous and read speech is also plotted in Fig. 5. Phone duration is estimated based on the Viterbi algorithm for given phone transcriptions. As we use three-state phone HMMs without state-skipping, the minimum duration is three frames ($= 30$ ms.). Many segments in the CSJ data may have shorter duration, but are forcedly aligned to three frames. This may have caused a serious mis-match. Moreover, a fast SR suggests that these segments are poorly articulated and cause problems during recognition.

3) Relationship With Recognition Errors: The relationship between the WER and SR is plotted for the test set in Fig. 6. Shinozaki *et al.* [26] and Okuda *et al.* [27] studied this relationship, in which the SR was averaged by the speaker level for the whole presentation talk although the SR changes frequently in one presentation talk. Here, we investigate the relationship in more detail, specifically by the utterance unit.

In Fig. 6, the breakdown of recognition errors is shown for each SR. It is confirmed that faster utterances are generally harder for recognition. Moreover, we observe different tendencies in the errors according to the SR. In fast utterances, substitution errors are increased as well as deletion errors. On the other hand, there are many insertion errors in slower segments. Therefore, we explore methods to deal with these error tendencies according to the SR.

B. Automatic Estimation of Speaking Rate

Several methods to estimate SR have been studied, such as detecting phone boundaries with multilayer perceptorons (MLP) [28], using a Gaussian mixture model (GMM) [29], and detecting vowels using speech features such as loudness, zero-cross counts, and energy envelope [30], [31]. In this paper, we introduce mora decoding [32], [33] to estimate SR. Since Japanese has a regular moraic structure, mora decoding is widely used for initial decoding for fast match [34] or unknown word modeling [35]. Moreover, mora decoding method that uses only a simple phonotactic syllable constraint and an

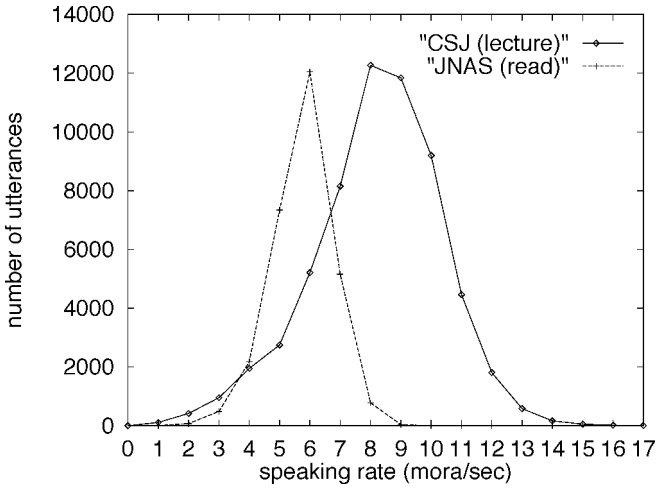

Fig. 4. SR distribution of CSJ and JNAS corpus.

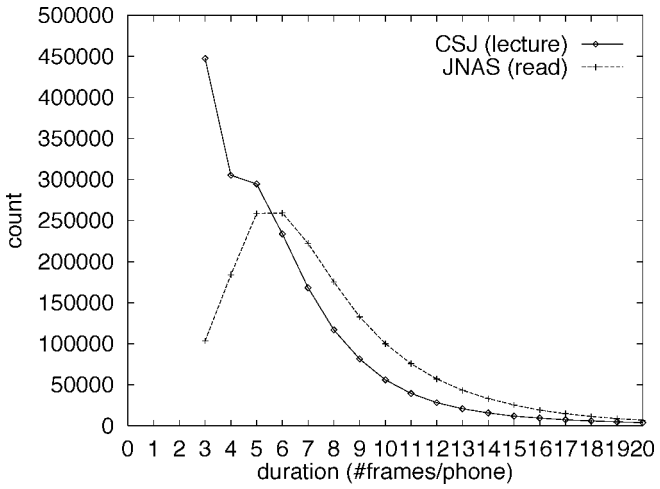

Fig. 5. Phone duration distribution of CSJ and JNAS corpus.

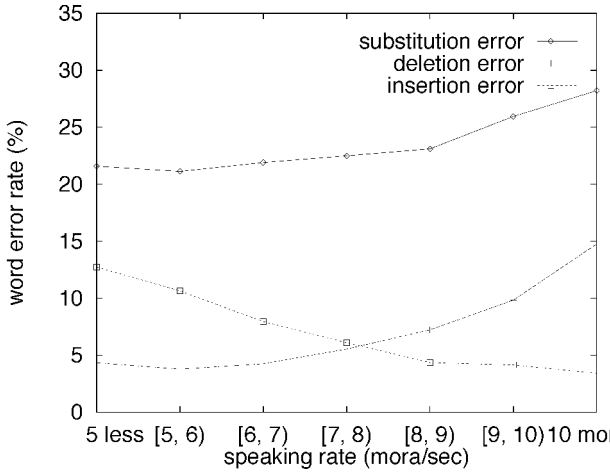

Fig. 6. Ratio of substitution, deletion, and insertion errors for each SR.

acoustic model used in ASR system does not require special training compared with other methods such as MLP and GMM.

We define 292 syllables for Japanese: 10 vowels, 280 consonant+vowel pairs, the double consonant /q/, and the syllabic /N/. These syllables and a short pause /sp/ can be connected freely. The mora count is calculated from the syllable recognition result. In calculating the mora count from the recognition result, we do not count /q/ and regard a long vowel as one mora.

Fig. 7 plots the relationship between the actual and estimated SR. There is high correlation between the two: the correlation coefficient is 0.88. The result verifies the feasibility of SR estimation. The estimated SR is adjusted according to the linear regression equation of estimated and actual SR, which is $y = 1.2 \times x + 0.13$ where x and y are estimated and actual SR. The main reasons why the estimated SR is smaller than the actual SR are that: 1) we do not count /q/ because it causes a lot of false matchings with a pause segment and degrades the SR estimation accuracy and 2) we regard a long vowel as one mora because a long vowel and a short vowel are often falsely matched with each other and this degrades the SR estimation accuracy.

C. Speaking-Rate-Dependent Acoustic Modeling and Decoding

1) Framework: Based on these analyses, we propose applying different decoding methods according to the SR within a multiple-pass recognition framework. The SR in the current speech segment is estimated in the initial recognition with the baseline speaker-independent acoustic model. Then, the most adequate acoustic analysis, phone models and decoding parameters are applied.

Specifically, the following processes are investigated. The first three are intended for fast speech and the last one is for slow speech.

- *Shorter frame length and shift:* To cope with fast speech segments, where the spectral pattern changes rapidly, the frame length and shift for spectral analysis are shortened. After preliminary experiments, we set the frame length of 20 ms and the shift of 8 ms from the baseline of 25 and 10 ms in decoding.
- *State-skipping transitions in phone models*: Another way to cope with fast speech is to add state-skipping transitions in phone models. This allows flexible matching with less than three frames. Here, transition from the first state to the third (last) state is added and all parameters of means, variances and state transition probabilities are re-trained.
- *Syllable models:* Since several phone segments may disappear in spontaneous speech, we model them with syllables of a phone sequence. We select syllables by considering both their duration and training data amount. The following statistic (4) is defined as a criterion for selection:

$$V_s = \sum_i P^{Duration(s_i)} \tag{4}$$

where s_i is a sample i of syllable s, P is an average probability of self-looping transition ($= 0.56$), and $Duration(s_i)$ is the number of frames with which s_i is aligned. The more fast segments that occur, the larger value V_s gets.

We selected 30 syllables which have a larger value of V_s based on the criterion. They are all concerned with functional words. The syllable HMMs are made by concatenating phone HMMs by adding state skipping transitions (Fig. 8). Then, only transition probabilities and weights for Gaussian mixtures are re-estimated. This tied-mixture modeling will ease the problem of data sparseness of the long unit. Actually, simple introduction of the syllable model with separate model training degraded the accuracy [36].

- *Changing insertion penalty:* For slow speech segments, an increase of insertion errors is observed. Thus, a larger word insertion penalty is used in order to suppress insertion errors.

2) Combination of Model Adaptation: Unsupervised acoustic and language model adaptation methods are also combined with SR-dependent decoding framework. For the acoustic model adaptation, the MLLR adaptation [37] is performed. A phone transcription for adaptation is generated from the initial

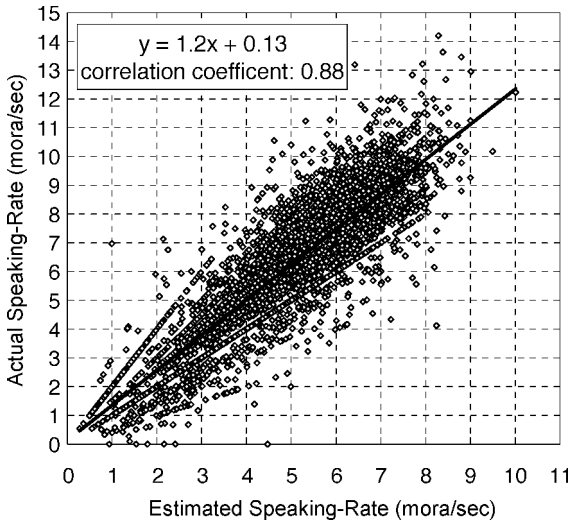

Fig. 7. Relationship of actual and estimated SRs.

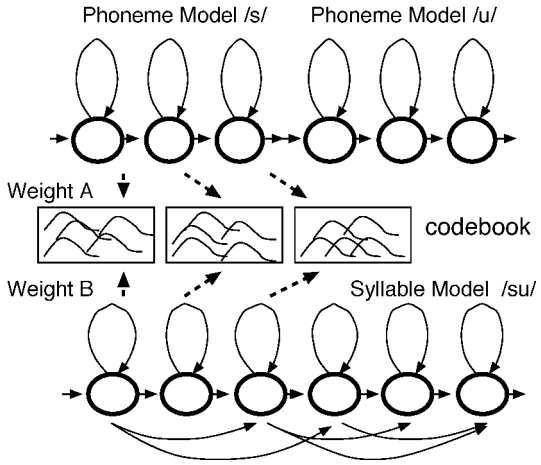

Fig. 8. Tied-mixture syllable modeling.

TABLE V
WER WITH DIFFERENT DECODING ACCORDING TO SR (%)

AM, LM	actual speaking rate (#utterance)	-5 (1113)	5-6 (657)	6-7 (928)	7-8 (963)	8-9 (619)	9-10 (270)	10- (53)	average (4603)
SI, SI	baseline	44.2	29.1	28.2	25.6	28.3	32.5	42.7	30.9
SA, SI		35.3	23.0	23.4	21.6	24.1	29.2	38.7	26.0
SA, SA		34.0	21.7	21.2	19.2	22.0	26.4	35.7	23.9
SA, SA	1. analysis frame	35.0	21.8	21.4	18.5	20.2	24.5	32.7	23.1
SA, SA	2. skipping transition	34.3	20.8	20.4	18.3	20.5	24.8	31.8	22.8
SA, SA	3. syllable model	33.4	21.2	19.9	18.3	20.2	24.0	31.8	22.5
SA, SA	4. insertion penalty (-8)	**30.1**	22.3	22.5	21.4	24.5	30.5	37.7	25.4
SA, SA	1.+2.	37.5	22.2	21.5	18.3	19.8	23.6	29.6	23.1
SA, SA	1.+3.	35.3	22.8	20.9	18.3	19.9	22.9	30.2	22.8
SA, SA	2.+3.	34.1	**20.9**	**20.0**	**18.2**	**19.7**	23.8	30.4	22.3
SA, SA	1.+2.+3.	38.3	22.6	21.2	18.4	19.8	**23.1**	**28.9**	23.1
SA, SA	selected with actual speaking rate [oracle]	30.1	20.9	20.0	18.2	19.7	23.1	28.9	21.7
SA, SA	selected with estimated speaking rate	30.5	21.5	20.1	18.3	19.8	23.7	29.5	22.0

(SI: Speaker Independent model, SA: Speaker Adapted model)

recognition result. For the language model adaptation, we use the method proposed in the previous section.

3) Experimental Results: These techniques and their combinations are evaluated in the test set. They are compared with the baseline system that adopts uniform decoding. The recognition results are listed in Table V.[2]

The unsupervised acoustic model adaptation reduced WER by 4.9% absolute, from 30.9% to 26.0%, and the combination with the language model adaptation methods reduced it further 2.1% absolute to 23.9%.

For fast speech segments, all proposed methods (1, 2, 3) are shown to be effective and improved the overall accuracy. Adding state-skipping transition and syllable modeling improved the accuracy for fast speech, as well as normal speech, except for very slow speech ($\leq$ 5 mora/s). Their combination is more effective. Use of shorter frame length and shift is also effective for fast speech, but less effective for normal speech. The combination of all proposed methods has an effect on the very fast speech (9 mora/s or faster), but results in an increase of errors in slow speech, which cancels this effect. For slow utterances, the use of a severe insertion penalty reduces errors as expected.

Then, selective application of these methods according to the SR is implemented, as indicated in bold font in Table V. The SR is classified into three categories based on the experimental result. If the SR is known and the best techniques are chosen accordingly (oracle case), the overall accuracy could be improved by 2.2% absolute. In actuality, we estimate the SR with the syllable constraint and apply the dedicated decoding methods in the second pass. This strategy enables an improvement of 1.9% absolute from 23.9% to 22.0% (the last row of Table V). This result demonstrates that the automatic estimation brings about comparable performance to the oracle case.

D. Experimental Results With Open Test Set

Finally, we evaluate proposed methods and their combination with a different test set (ten presentations), which is not used for tuning the parameters and deciding the selection algorithm. The specification is listed in Table VI.

[2]Baseline WER is different from the one of the previous section as we change some decoding parameters in this section.

TABLE VI
TEST-SET PRESENTATIONS (OPEN DATA) FOR EVALUATION OF SR-DEPENDENT DECODING

presentation ID	#words	duration (min.)	PP	SR
A01M0007	4591	30	55.9	8.21
A01M0035	6598	28	70.6	9.60
A01M0074	2580	12	58.6	8.82
A02M0117	10375	57	70.8	7.81
A03M0100	2839	15	48.5	8.00
A05M0031	5632	27	83.5	9.41
A06M0134	4794	23	55.5	9.66
KK99DEC005	6956	42	60.6	7.38
YG99JUN001	2954	14	60.4	8.85
YG99MAY005	3282	15	60.8	8.58
total	50601	263	64.1	8.63

For the language model adaptation, we used the interpolation coefficient λ which was estimated with the previous test set (15 presentations). The estimated SR was adjusted according to the linear regression equation which was derived with the previous test set (15 presentations) in Section V-B. For the SR-dependent decoding, the selection of techniques (models and parameters) was performed based on the experimental result of previous section.

The result is shown in Table VII. The same tendencies are observed in this set. Language model adaptation led improvement of WER by 1.5% absolute and the SR-dependent decoding strategy achieved improvement by 1.4% absolute.

The result shows that our proposed language model adaptation method and SR-dependent decoding are effective.

VI. CONCLUSION

We presented adaptation methods for variations of linguistic expressions and speaking rate of individual speakers which are the two major factors affecting the presentation speech recognition accuracy.

First, we presented methods that adapt a language model to speakers' characteristic expression, topic, and pronunciation variation. Several language model adaptation methods combined with pronunciation variation modeling have been investigated. Specifically, we proposed several methods based on the similar text selection and the direct use of the initial recognition result. All proposed methods effectively reduced

TABLE VII
OPEN DATA RESULT (WER WITH DIFFERENT DECODING ACCORDING TO SR (%)

AM, LM	actual speaking rate (#utterance)	-5 (1684)	5-6 (864)	6-7 (1031)	7-8 (855)	8-9 (665)	9-10 (329)	10- (181)	average (5609)
SI, SI	baseline	44.1	29.5	26.8	27.2	27.3	32.1	40.8	30.9
SA, SI		37.6	26.0	23.6	24.3	25.1	29.3	38.5	27.6
SA, SA		36.5	24.3	22.6	22.3	23.7	28.2	36.1	26.1
SA, SA	4. insertion penalty (-8)	**33.3**	24.9	23.2	24.4	25.5	31.3	39.1	27.2
SA, SA	2.+3.	37.7	**24.0**	**21.6**	**21.5**	**21.5**	26.5	32.3	25.0
SA, SA	1.+2.+3.	42.4	25.8	23.2	22.6	22.4	**26.2**	**31.3**	26.4
SA, SA	selected with actual speaking rate [oracle]	33.3	24.0	21.6	21.5	21.5	26.2	31.3	24.4
SA, SA	selected with estimated speaking rate	34.2	24.3	21.7	21.5	21.7	26.4	31.9	24.7

(SI: Speaker Independent model, SA: Speaker Adapted model)

the perplexity and WER, and their combinations reduced WER from 31.4% to 27.6% (a total error reduction rate of 12.1%).

Next, we have proposed a SR-dependent decoding strategy that adaptively applies the most adequate acoustic analysis, phone models, and decoding parameters depending on the current estimated SR. We investigated several techniques and demonstrated that their selective application is effective. This strategy achieved the reduction of WER by 1.9% absolute for the test set and 1.4% absolute for ten other presentations (open set).

As a whole, these adaptation methods have significantly improved accuracy and achieved a WER of 22.0% in automatic transcription of spontaneous lecture presentations. Moreover, the adaptation methods were performed in completely unsupervised manners.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Prof. S. Furui (Tokyo Institute of Technology) for leading this fruitful project of the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology". In this collaborative project, the authors are especially thankful to Dr. A. Yamada and Mr. K. Uchimoto (CRL) for revising the morphological analyzer to our purpose. They are also in debt to Dr. A. Lee (Nara Institute of Science and Technology) for improving the LVCSR engine Julius for spontaneous speech recognition.

REFERENCES

[1] S. Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 1–6.
[2] K. Maekawa, "Corpus of spontaneous Japanese: its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 7–12.
[3] T. Shinozaki and S. Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2002, pp. 729–732.
[4] T. Kawahara, H. Nanjo, T. Sinozaki, and S. Furui, "Benchmark test for speech recognition using the Corpus of spontaneous Japanese," in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 135–138.
[5] S. F. Chen, K. Seymore, and R. Rosenfeld, "Topic adaptation for language modeling using unnormalized exponential models," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1998, pp. 681–684.
[6] M. Mahajan, D. Beeferman, and X. D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, vol. 1, 1999, pp. 541–544.
[7] J. Zheng, H. Franco, and F. Weng, "Word-level rate of speech modeling using rate-specific phones and pronunciations," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1775–1778.
[8] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 2079–2082.
[9] J. Nedel and R. Stern, "Duration normalization for improved recognition of spontaneous and read speech via missing feature methods," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 2001, pp. 313–316.
[10] M. Richardson, M. Hwang, A. Acero, and X. D. Huang, "Improvements on speech recognition for fast talkers," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1999, pp. 411–414.
[11] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 2379–2382.
[12] E. Fosler-Lussier, "Multi-level decision trees for static and dynamic pronunciation models," in *Proc. European Conf. Speech Commun. & Tech. (EUROSPEECH)*, 1999, pp. 463–466.
[13] P. R. Clarkson and R. Rosenfeld, "Statistical language modeling using the CMU-Cambridge toolkit," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 1997, pp. 2707–2710.
[14] K. Uchimoto, C. Nobata, A. Yamada, S. Sekine, and H. Isahara, "Morphological analysis of Corpus of spontaneous Japanese," in *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, 2003, pp. 159–162.
[15] A. Lee, T. Kawahara, K. Takeda, and K. Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1269–1272.
[16] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Itou, and K. Shikano, "Continuous speech recognition consortium—an open repository for CSR tools and models," in *Proc. Int. Conf. Language Resources and Evaluation (LREC2002)*, 2002, pp. 1438–1441.
[17] B. Peskin, M. Newman, D. McAllaster, V. Nagesha, H. Richards, S. Wegmann, M. Hunt, and L. Gillick, "Improvements in recognition of conversational telephone speech," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 1999, pp. 53–56.
[18] H. Schramm and X. Aubert, "Efficient integration of multiple pronunciations in a large vocabulary decoder," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1659–1662.
[19] J. M. Kessens, M. Wester, and H. Strik, "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation," *Speech Commun.*, vol. 29, no. 2–4, pp. 193–207, 1999.
[20] R. R. Sarukkai and D. H. Ballard, "Improved spontaneous dialogue recognition using dialogue and utterance triggers by adaptive probability boosting," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 1, Philadelphia, PA, 1996, pp. 208–211.
[21] R. Iyer and M. Ostendorf, "Modeling long distance dependence in language: Topic mixtures vs. dynamic cache models," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 1, Philadelphia, PA, 1996, pp. 236–239.
[22] P. Clarkson and A. J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *Proc. IEEE Int. Conf. Acoust., Speech & Signal Process. (ICASSP)*, vol. 2, 1997, pp. 799–802.
[23] K. Kato, H. Nanjo, and T. Kawahara, "Automatic transcription of lecture speech using topic-independent language modeling," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 1, 2000, pp. 162–165.

[24] L. Chen, J. L. Gauvain, L. Lamel, G. Adda, and M. Adda, "Using information retrieval methods for language model adaptation," in *Proc. Eur. Conf. Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 255–258.
[25] T. Niesler and D. Willett, "Unsupervised language model adaptation for lecture speech transcription," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, Denver, CO, 2002, pp. 1413–1416.
[26] T. Shinozaki and S. Furui, "Toward automatic transcription of spontaneous presentations," in *Proc. European Conf. Speech Commun. & Tech. (EUROSPEECH)*, 2001, pp. 491–494.
[27] K. Okuda, T. Kawahara, and S. Nakamura, "Speaking rate compensation based on likelihood criterion in acoustic model training and decoding," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 2589–2592.
[28] J. P. Verhasselt and J. P. Martens, "A fast and reliable rate of speech detector," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 4, 1996, pp. 612–615.
[29] R. Faltlhauser, T. Pfau, and G. Ruske, "On-line speaking rate estimation using gaussian mixture models," in *Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing (ICASSP)*, vol. III, 2000, pp. 1355–1358.
[30] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1998, pp. 945–948.
[31] N. Morgan and E. Fosler-Lussier, "Combining multiple estimators of speaking rate," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2, 1998, pp. 729–732.
[32] K. Hirose and K. Iwano, "Detection of prosodic word boundaries by statistical modeling of mora transitions of fundamental frequency contours and its use for continuous speech recognition," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, vol. 3, 2000, pp. 1763–1766.
[33] N. Takahashi and S. Nakagawa, "Syllable recognition using syllable-segment statistics and syllable-based HMM," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, 2002, pp. 2633–2636.
[34] T. Kawahara, T. Munetsugu, N. Kitaoka, and S. Doshita, "Keyword and phrase spotting with heuristic language model," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 2, 1994, pp. 815–818.
[35] A. Kai, Y. Hirose, and S. Nakagawa, "Dealing with out-of-vocabulary words and speech disfluencies in an n-gram based speech understanding system," in *Proc. Int. Conf. Spoken Language Processing (ICSLP)*, vol. 6, 1998, pp. 2427–2430.
[36] H. Nanjo, K. Kato, and T. Kawahara, "Speaking rate dependent acoustic modeling for spontaneous lecture speech recognition," in *Proc. European Conf. Speech Communication and Technology (EUROSPEECH)*, 2001, pp. 2531–2534.
[37] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Comput. Speech and Lang.*, vol. 9, no. 2, pp. 171–185, 1995.

Hiroaki Nanjo received the B.E. degree in 1999 and the M.E. degree in 2001 from Kyoto University, Kyoto, Japan, where he is currently pursuing the Ph.D. degree in the School of Informatics. He has been working on speech recognition and understanding.

Mr. Nanjo is a member of Acoustical Society of Japan (ASJ), the Institute of Electronics, Information and Communication Engineers (IEICE), and the Information Processing Society of Japan (IPSJ).

Tatsuya Kawahara (M'91) received the B.E. degree in 1987, the M.E. degree in 1989, and the Ph.D. degree in 1995, all in information science, from Kyoto University, Kyoto, Japan.

In 1990, he became a Research Associate with Department of Information Science, Kyoto University. From 1995 to 1996, he was a Visiting Researcher at Bell Laboratories, Murray Hill, NJ. Currently, he is a Professor with the Academic Center for Computing and Media Studies, Kyoto University. He is also an Invited Researcher at ATR Spoken Language Translation Research Laboratories. He has published more than 100 technical papers covering speech recognition, confidence measures, and spoken dialogue systems. He has been managing several speech-related projects in Japan, including a free large vocabulary continuous speech recognition software project (http://julius.sourceforge.jp/).

Dr. Kawahara received the 1997 Awaya Memorial Award from the Acoustical Society of Japan and the 2000 Sakai Memorial Award from the Information Processing Society of Japan. Since 2003, he has been a member of the IEEE Signal Processing Society's Speech Technical Committee.