IEICE TRANS. INF. & SYST., VOL.E91–D, NO.3 MARCH 2008

PAPER Special Section on Robust Speech Processing in Realistic Environments

Language Modeling Using PLSA-Based Topic HMM

Atsushi SAKO[†a)], *Student Member*, Tetsuya TAKIGUCHI[†], *and* Yasuo ARIKI[†], *Members*

SUMMARY In this paper, we propose a PLSA-based language model for sports-related live speech. This model is implemented using a unigram rescaling technique that combines a topic model and an *n*-gram. In the conventional method, unigram rescaling is performed with a topic distribution estimated from a recognized transcription history. This method can improve the performance, but it cannot express topic transition. By incorporating the concept of topic transition, it is expected that the recognition performance will be improved. Thus, the proposed method employs a "Topic HMM" instead of a history to estimate the topic distribution. The Topic HMM is an Ergodic HMM that expresses typical topic distributions as well as topic transition probabilities. Word accuracy results from our experiments confirmed the superiority of the proposed method over a trigram and a PLSA-based conventional method that uses a recognized history.

key words: *language model, adaptation, PLSA, HMM, automatic speech recognition*

1. Introduction

Recently large volumes of multimedia content are broadcast and accessed through digital TV and the Internet. In order to extract exactly what we want to know from them, automatic extraction of meta-information or structuring is very much a necessity. Sophisticated automatic speech recognition (ASR) plays an important role in extracting this kind of information because accurate transcription is inevitable. The purpose of this study is to improve the speech recognition accuracy for automatically transcribing live sports-related speech (especially baseball commentary) in order to produce closed captions and to structure the games for highlight scene retrieval.

As the sports-related live speech in our experiment, we used radio speech instead of TV speech because it has much more information. However radio speech is rather fast and noisy. Furthermore, it is choppy (not smooth-flowing) due to rephrasing, repetition, mistakes and grammatical deviation caused by the spontaneous speaking style. To solve these problems, we proposed adaptation techniques for an acoustic model and language model [1] and a situation-based language model [2].

In order to further improve the speech recognition accuracy, we focus on topic-based language models in this paper. Several models based on Ergodic HMM for considering topics have been reported. [3] clustered utterances into several classes manually, and modeled its output probability as an Ergodic HMM. [4] reported that the sequences of Context Free Grammar (CFG) drive a Ergodic HMM and the CFG rule probabilities are dynamically changed according to topic state distribution. [5] modeled the transitions of utterances as Ergodic HMM and reported that this model reduces the test set perplexity. Moreover several topic-based language models have been studied; stochastic switching-language model (SS N-gram) [6], a language model based on Latent Semantic Analysis (LSA) [7] and a PLSA-based language model using unigram rescaling techniques [8]. SS N-gram requires a large quantity of corpus. In sports tasks, however, it is difficult to create a large corpus. PLSA is a probabilistic model of LSA and more compatible with an N-gram than LSA. Thus, in this paper, we focus on PLSA-based models in particular.

The conventional PLSA-based model estimates a topic distribution using a "history" of recognized transcription. However, it cannot express topic transition. Considering topic transition, the recognition accuracy is improved because it enables the use of the proper language model for each topic. Consequently, we propose a new language model based on PLSA. The model expresses typical distributions of topics and transition probabilities between topics. We implemented this model as an Ergodic HMM that has distribution in each state and transition probabilities between states. We call the HMM a "Topic-HMM". Unigram probabilities are obtained from a distribution of a state through the algorithm described in Sect. 3. Moreover, trigram probabilities are also obtained using a unigram rescaling technique. For each state of the Topic HMM, a trigram is computed as a topic dependent language model.

Conventional models are methods for spoken dialogue system and these are effective because an utterance replying to another person depends on a previous utterance spoken by another. In our task, though it is not a spoken dialogue task, an utterance depends on a previous utterance due to dependency on a process of a baseball game. In comparison with conventional methods, our proposed method has following advantages. First, the proposed method does not need a large corpus due to learning based on latent topic distributions. Next, the proposed method can provide an accurate word prediction capability based on trigram which is computed by the unigram rescaling technique.

The experimental results show that the Topic HMM improves the performance of the word accuracy.

Manuscript received July 4, 2007.
Manuscript revised September 25, 2007.

[†]The authors are with the Graduate School of Science and Technology, Kobe University, Kobe-shi, 657–8501 Japan.
a) E-mail: sakoats@me.cs.scitec.kobe-u.ac.jp

DOI: 10.1093/ietisy/e91–d.3.522

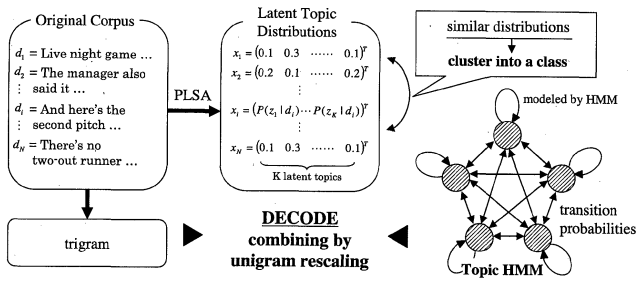


Fig. 1 Overview of proposed method. In an original corpus, each utterance is divided by period from manual transcriptions. The Topic HMM is learned using feature vectors obtained from topic distribution of each utterance. A state of HMM corresponds to a distribution of topics. The decoding is performed with language models constructed by combining the Topic HMM and a trigram using a unigram rescaling technique.

2. Overview of Proposed Method

Baseball commentary during live games tends to repeat similar expressions and word orders due to its nature. For example, "Here's the first pitch" is repeated when pitching is taking place, and "Strike" or "Ball," etc. are spoken after pitching expressions.

The goal of the proposed method is to improve speech recognition performance by incorporating topic-dependent language models and transition probabilities between language models. These are actualized by a *Topic HMM* and a unigram rescaling technique. Figure 1 shows an overview of the proposed method. First, a trigram language model is constructed from the original corpus. This model is the baseline for the adaptation. Next, latent topic distributions are computed from the original corpus based on PLSA. These distributions represent the rates at which an utterance is included in each latent topic. Next, Ergodic HMM is learned using these latent topic distributions instead of word vectors in an utterance. This is because word vectors are too big and too sparse to estimate appropriate states, especially in a small corpus. By considering latent topic distributions as observed feature vectors and classes as hidden states, they are modeled by Ergodic HMM, or "Topic HMM". Each state of the Topic HMM represents a typical distribution of latent topics and there are transition probabilities between states.

In the next section, we briefly describe the basis of PLSA and the adaptation technique for a language model, and Sect. 4 describes our proposed method constructing a Topic HMM.

3. PLSA-Based Language Modeling

Probabilistic Latent Semantic Analysis (PLSA) [9] is a topic decomposition method for documents in a corpus. In this paper, PLSA is performed for utterances and an utterance is employed as a unit of a document. We describe PLSA below with term "utterance" instead of "document". This method analyzes topic probabilities in each utterance and unigram probabilities in each topic. These probabilities are estimated from the co-occurrence probabilities of a word and an utterance, where $u_i (i = 1, \cdots, N)$ denotes an utterance from a text corpus, $w_j (j = 1, \cdots, M)$ denotes a word, and $z_k (k = 1, \cdots, K)$ denotes a latent variable that represents a latent topic. Under the assumption that an utterance and a word are independent of each other given a latent variable, the word probability conditioned on the utterance is estimated using

$$P(w_j|u_i) = \sum_{k=1}^{K} P(w_j|z_k)P(z_k|u_i), \tag{1}$$

where $P(w_j|z_k)$ is a unigram probability conditioned on a latent variable and $P(z_k|u_i)$ is a latent variable probability for each utterance.

Each parameter is estimated by the Expectation Maximization (EM) algorithm. The E-step is

$$P(z_k|u_i, w_j) = \frac{P(w_j|z_k)P(z_k|u_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|u_i)}, \tag{2}$$

and the M-step is

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} N(u_i, w_j)P(z_k|u_i, w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} N(u_i, w_j)P(z_k|u_i, w_j)}, \tag{3}$$

$$P(z_k|u_i) = \frac{\sum_{j=1}^{M} N(u_i, w_j)P(z_k|u_i, w_j)}{N(u_i)} \tag{4}$$

where $N(u_i, w_j)$ is the number of the co-occurrences of the word w_j and the utterance u_i. Table 1 shows a portion of the PLSA results. Characteristic words in each latent topic are listed ($P(w_j|z_k)/P(w_j)$ of these words were high). The summary is subjectively labeled from characteristic words.

Gildea [8] proposed the adaptation technique for language models based on PLSA. The adaptation is approximately performed using a *unigram rescaling* technique because it is difficult to estimate such a large number of N-gram entries. Based on the assumption that the history h_i and the trigram context are independent conditioned on w_i, the trigram probability $P(w_i|w_{i-1}, w_{i-2}, \hat{u})$ is computed as follows:

$$P(w_i|w_{i-1}, w_{i-2}, \hat{u}) \propto \frac{P(w_i|\hat{u})}{P(w_i)} P(w_i|w_{i-1}, w_{i-2}), \tag{5}$$

Table 1 A portion of PLSA results.

	summary	word 1	word 2	word 3
z_1	ball count	tsu: (two)	kaunto (count)	sutoraiku (strike)
z_2	batted ball	uchiage (fly ball)	tsumatta (weakly hit)	guraundo (grounder)
z_3	pitching	dai (1st/2nd/3rd [pitch])	nage (pitch)	piccha: (pitcher)
z_4	pitching result	hizamoto (base of the knee)	sanshin (strike-out)	ka:bu (curve)
z_5	chat	sou (right)	kamo (maybe)	nai (not)

where $\hat{u}$ is an unseen utterance (often, it is a recognized hypothesis), $P(w_i)$ and $P(w_i|w_{i-1}, w_{i-2})$ are a unigram and a trigram probability, respectively, in a baseline language model. Here, normalization is required because the sum of $\hat{u}$ dependent trigram computed by Eq. (5) will not be 1. Since $\hat{u}$ is an unseen utterance, $P(z_k|\hat{u})$ should be estimated in some way. In the conventional method, as $\hat{u}$ is a sequence of recognized words $h_i = (w_1, \cdots, w_i)$, the approximately computing method was proposed in [8] as follows:

$$P(z_k|h_i) = \frac{1}{i+1} \frac{P(w_i|z_k)P(z_k|h_{i-1})}{\sum_{k'=1}^{K} P(w_i|z_{k'})P(z_{k'}|h_{i-1})} + \frac{i}{i+1} P(z_k|h_{i-1}). \quad (6)$$

In this method (called "History"), $P(z_k|\hat{h}_i)$ continues to be updated when a new word is recognized. Note that history consists of words in a current recognizing utterance (namely, w_1 is a first word of an utterance). In the next section, we describe the proposed method to estimate $P(z_k|\hat{u})$.

4. Language Modeling Using Topic HMM

In this section, we describe how to construct a PLSA-based Topic HMM. First, PLSA is performed to estimate topic distribution $P(z_k|u_i)$ for all utterances and unigram distribution $P(w_j|z_k)$. Utterances are divided by a period from manual transcriptions. In the corpus, there were about 8,000 utterances. After carrying out PLSA, each utterance is expressed as a vector that consists of latent topic probabilities as follows:

$$\mathbf{x}_i = \begin{pmatrix} P(z_1|u_i) \\ \vdots \\ P(z_K|u_i) \end{pmatrix}. \quad (7)$$

Namely, word vectors in the original corpus are converted into topic vectors $\mathbf{x}_i$.

Next, an Ergodic HMM (shown in Fig. 2) is trained based on maximum likelihood estimation. In this HMM, probabilistic density function is normal distribution and only diagonal variances (not co-variances) are used. The sequence of topic vectors $\{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ is used as the sequence of observed features, where L is the number of states in HMM and the mean vector μ_n $(n = 1, \cdots, L)$ is estimated as the weighted average of observed topic vectors. It is considered to be the typical latent topic vector in the state s_n. In this paper, an inning is employed as a unit of topic sequence for HMM learning. In addition, transition probabilities between state s_n and $s_{\hat{n}}$ $P(s_{\hat{n}}|s_n)$ are estimated.

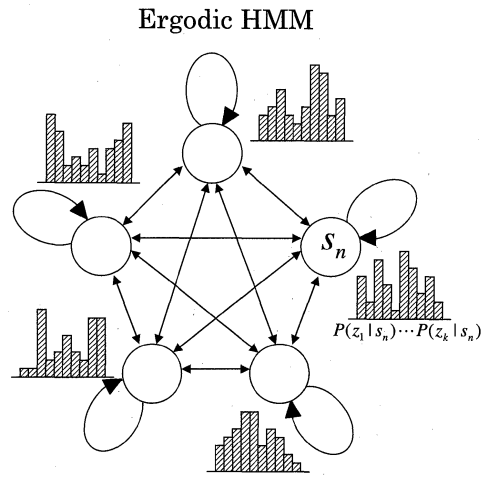


Fig. 2 An Ergodic HMM as a Topic HMM. Each state represents a latent topic distribution as a mean vector.

In History method, history dependent trigram is computed by Eq. (5). Here, we describe how to compute the state dependent trigram in the proposed method. First, state dependent trigram can be derived as follows:

$$\begin{aligned} & P(w_i|w_{i-1}, w_{i-2}, s_n) \\ &= \frac{P(w_{i-1}, w_{i-2})P(w_i|w_{i-1}, w_{i-2})P(s_n|w_i, w_{i-1}, w_{i-2})}{P(w_{i-1}, w_{i-2})P(s_n|w_{i-1}, w_{i-2})} \\ &= \frac{P(w_i|w_{i-1}, w_{i-2})P(s_n|w_i, w_{i-1}, w_{i-2})}{P(s_n|w_{i-1}, w_{i-2})}. \end{aligned} \quad (8)$$

Under the assumption that the state s_n and the trigram context w_{i-1}, w_{i-2} are independent conditioned on w_i, following formula can be derived.

$$\begin{aligned} P(w_i|w_{i-1}, w_{i-2}, s_n) &\approx \frac{P(w_i|w_{i-1}, w_{i-2})P(s_n|w_i)}{P(s_n)} \\ &= \frac{P(w_i|s_n)}{P(w_i)} P(w_i|w_{i-1}, w_{i-2}). \end{aligned} \quad (9)$$

Based on the above assumption,

$$\frac{P(s_n|w_i, w_{i-1}, w_{i-2})}{P(s_n|w_{i-1}, w_{i-2})} \quad (10)$$

is changed into

$$\frac{P(s_n|w_i)}{P(s_n)}. \quad (11)$$

Eqs. (10) and (11) describe the s_n probability ratio between when w_i is known and unknown. Eq. (11) can be expected to play a similar role to Eq. (10).

Next, considering μ_n to be latent topic distribution in the state s_n, word probability depending on the state s_n can

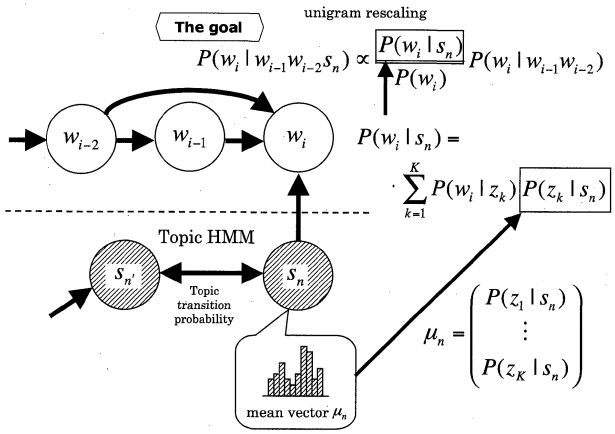


Fig. 3 Summary of building langugae model based on Topic HMM.

be computed as follows:

$$P(w_i|s_n) = \sum_{k=1}^{K} P(w_i|z_k)P(z_k|s_n), \tag{12}$$

where $P(z_k|s_n)$ is the k-th element of the mean vector μ_n. Finally, we can compute the state dependent trigram from Eqs. (9) and (12). Note that, in a similar way to Eq. (5), normalization is required to make the sum of state dependent trigram for all words into 1.

In summary, Fig. 3 shows the building process of a language model based on Topic HMM. The goal is to estimate trigram probability for each state. This is computed by a unigram rescaling technique using mean vector μ_n as the latent topic distribution.

5. Speech Recognition by Estimating the State Sequence of Topic HMM

In this section, we formalize speech recognition to find the most likely word sequence $\mathbf{W} = \{\mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(N)}\}$ as well as the state sequence of Topic HMM $\mathbf{S} = \{s_1, \cdots, s_N\}$, where N is the number of utterances and $\mathbf{w}^{(n)} = \{w_1^{(n)}, \cdots, w_{T^{(n)}}^{(n)}\}$ is the word sequence of an utterance. $T^{(n)}$ is the number of words in an utterance $\mathbf{w}^{(n)}$. Given the sequence of observed feature vectors $\mathbf{O} = \{\mathbf{o}_1, \cdots, \mathbf{o}_N\}$, speech recognition is formalized as follows:

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\text{argmax}}\, P(\mathbf{S}, \mathbf{W}|\mathbf{O}). \tag{13}$$

Eq. (14) can be derived from Eq. (13):

$$(\hat{\mathbf{S}}, \hat{\mathbf{W}}) = \underset{\mathbf{S},\mathbf{W}}{\text{argmax}}\, P(\mathbf{S}, \mathbf{W})P(\mathbf{O}|\mathbf{W}), \tag{14}$$

based on the Bayesian theorem, $P(\mathbf{O})$ is omitted due to independence from $\mathbf{W}$, and the probability $P(\mathbf{O}|\mathbf{W}, \mathbf{S})$ is simplified into $P(\mathbf{O}|\mathbf{W})$ due to independence from $\mathbf{S}$. Moreover, we can derive the following equation;

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(s_1, \cdots, s_N, \mathbf{w}^{(1)}, \cdots, \mathbf{w}^{(N)}) \\ &= P(s_1)P(\mathbf{w}^{(1)}|s_1) \\ &\quad \times \prod_{n=2}^{N} P(s_n|s_1^{n-1}, \mathbf{w}^{(1,\cdots,n-1)}) \\ &\quad \times P(\mathbf{w}^{(n)}|\mathbf{w}^{(1,\cdots,n-1)}, s_1^n). \end{aligned} \tag{15}$$

Based on the following approximation,

- a state depends only on a previous state, and
- an utterance ($\mathbf{w}^{(n)}$) depends only on a present state.
- a word depends only on a present state and the previous two words (state dependent trigram).

We can simplify Eq. (15) as follows:

$$\begin{aligned} P(\mathbf{S}, \mathbf{W}) &= P(s_1)P(\mathbf{w}^{(1)}|s_1) \\ &\quad \times \prod_{n=2}^{N} P(s_n|s_{n-1})P(\mathbf{w}^{(n)}|s_n) \end{aligned} \tag{16}$$

$$= \prod_{n=1}^{N} P(s_n|s_{n-1}) \prod_{i=1}^{T^{(n)}} P(w_i^{(n)}|w_{i-1}^{(n)}, w_{i-2}^{(n)}, s_n), \tag{17}$$

where $P(s_n|s_{n-1})$ and $P(w_i^{(n)}|w_{i-1}^{(n)}, w_{i-2}^{(n)}, s_n)$ are obtained from a language model based on Topic HMM. Note that, in this research, a state transition probability $P(s_i|s_{i-1})$ is assigned to transitions between utterances only, and not words. This is because the Topic HMM is trained using an utterance as a unit. To adjust the effect of a transition probability $P(s_n|s_{n-1})$ of the Topic HMM, scaling factor α is employed. By the same token, scaling factor β is employed to adjust the effect of language model. Finally, the speech recognition that estimates the state sequence of the Topic HMM is formalized as:

$$\begin{aligned} (\hat{\mathbf{S}}, \hat{\mathbf{W}}) &= \underset{\mathbf{S},\mathbf{W}}{\text{argmax}}\, P(\mathbf{O}|\mathbf{W}) \\ &\quad \times \prod_{n=1}^{N} P(s_n|s_{n-1})^{\alpha} \prod_{i=1}^{T^{(n)}} P(w_i^{(n)}|w_{i-1}^{(n)}, w_{i-2}^{(n)}, s_n)^{\beta}. \end{aligned} \tag{18}$$

We implemented this formulation as a 2nd-pass search algorithm over the word-graph obtained from 1st-pass decoding. As the 1st-pass, the conventional — state independent — algorithm is employed. In the 1st-pass, both the state transition probability and the state dependent language model are not considered.

6. Experiments

To evaluate the language model using a PLSA-based Topic HMM, speech recognition was performed. The test set was the commentary speech on a live baseball game. We used the commentaries from the radio instead of TV since it contains much more information. Four commentaries were collected as our corpus. These specification are shown in Table 2. There are about 8000 utterances and about 80000 words. The average length of a utterance is 7.67 words, and 95% of utterances consists of between 1 to 24 sequence of words without punctuation. Many one-word utterances are agreement to another commentator such as "ah", "oh" and "yes". Usually, a process of utterances is as follows:

Table 2 The specification of our corpus.

Date	Speaker	Hours	# of utterances	# of words
2003/09/05	A	1.73	2232	21K
2003/09/06	B	1.81	2210	22K
2003/09/15	B	1.76	2320	21K
2003/09/16	A	1.61	2010	20K

- Describing batter,
- Pitching,
- Pitching results (stike, ball, faul ball, hit, out and so on),
- Chat with commentators.

By way of exception, pitching results were sometime inserted into chat utterances. These characteristics are commonly found in all commentaries in our corpus.

We performed the experiments using three methods: trigram, unigram rescaling from a recognized history as described in Sect. 3 (called "History"), and the proposed method. All of experiments were performed by 4-fold cross validation technique. Namely, testing commentary:2003/09/05, language model, PLSA model and Topic HMM were learned by the other commentaries. Note that, acoustic model was adaptated by another commentary which was spoken by the same speaker. In the next section, we describe the experimental conditions.

6.1 Experimental Conditions

The acoustic model was a syllable-based monophonic HMM (left to right) [11]. The experimental conditions of the acoustic model are summarized in Table 3. The training data consisted of about 200,000 Japanese sentences (200 hours) spoken by 200 males in the Corpus of Spontaneous Japanese (CSJ) [10]. Supervised acoustic model adaptation was performed using MLLR+MAP [12] with about 2 hours of data that matched the test set.

The baseline of the language model was a trigram language model. The training data consisted of 60,000 words collected from manual transcription of baseball games commentary. The number of unique words was about 3,000. The domain and the vocabulary of the training data were matched with the test set.

As the 1st-pass decoder, Julius [13] rev. 3.5. was used. The 2nd-pass decoder was implemented as a wordgraph decoder based on Sect. 5.

Since the performance depends on the number of topics and states, we performed experiments using various parameters.

6.2 Experimental Results

The experimental results are summarized in Table 4. The average accuracy for the "trigram" was 62.2%, for the "History" was 63.6% and for the "proposed method" was 65.3%. The Topic HMM shows the highest performance and this improvement is statistically significant (evaluated by Student's t-test, the 0.05 level). For this reason, two aspects of the advantages of the proposed method are considered. First, in the History method, latent topic distribution $P(z_k|h_i)$ is estimated from recognized word hypothesis. Thus, $P(z_k|h_i)$ will be estimated incorrectly when w_i is incorrect word. Since w_{i+1} is estimated based on unigram rescaling using $P(z_k|h_i)$, incorrect $P(z_k|h_i)$ may cause w_{i+1} bad effect. In the proposed method, latent topic distribution $P(z_k|s_n)$ is estimated beforehand from manual transcriptions. Therefore, by avoiding estimating errors, word recognition errors were reduced, especially since baseball commentary repeats similar expressions many times. Second, the effects of transition probabilities are considered. For example, there were two utterance such as "pitch" and then "KARABURI (meaning strike out)." In the conventional methods, it was recognized as "pitch" and then "TAMURA-RIN (name of player)" due to similar pronunciation. In the proposed method, it was recognized correctly because the Topic HMM indicated that the probability of the player name after pitching is low. This is not an effect of a topic-dependent language model, because each utterance is natural under topic dependency.

Table 3 Experimental conditions of acoustic model.

Sampling rate/Quantization	16 kHz / 16 bit
Feature vector	25 - order MFCC
Window	Hamming
Frame size/shift	20/10 ms
# of phoneme categories	244 syllable
# of mixtures	32
# of states (Vowel)	5 states and 3 loops
# of states (Consonant+Vowel)	7 states and 5 loops

Table 4 Experimental results in word accuracy.

Date	trigram	History	Topic HMM
2003/09/05	67.2%	68.0 (+0.8)%	**70.2 (+3.0)%**
(topology)	-	20 topics	70T and 50S
2003/09/06	69.4%	70.6 (+1.7)%	**72.5 (+3.1)%**
(topology)	-	40 topics	50T and 40S
2003/09/15	49.4%	51.3 (+1.9)%	**52.6 (+3.2)%**
(topology)	-	30 topics	70T and 20S
2003/09/16	62.7%	64.4 (+1.2)%	**66.2 (+3.5)%**
(topology)	-	40 topics	60T and 30S
Average	62.2%	63.6 (+1.4)%	**65.3 (+3.2)%**

Figure 4 shows the average of word accuracy of the proposed method over the number of topics of PLSA from 5 to 80 and the number of states of the Topic HMM from 5 to 60. In this figure, the performances are the average of commentary *for each topology*. The best performance in this figure is less than Table 4 because the average performances shown in Table 4 are the average of *the best performance in each commentary*. In this figure, the best performance was 64.6% with 60 topics and 30 states, and even the worst performance achieved was 63.2%. The topologies of the best performance in each commentary are near by this topology. However, it took much time to seek a topology with the best performance. To seek this topology automatically is a task for future research.

Figure 5 shows the transition probability efficacy of the

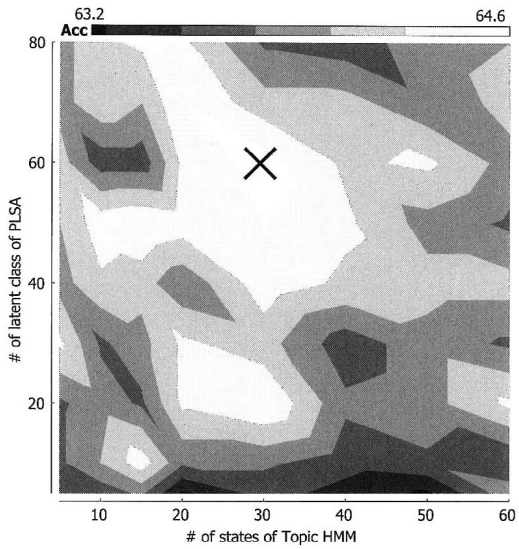


Fig. 4 Average of word accuracy results. Vertical axis shows the number of topics of PLSA. Horizontal axis shows the number of states of the Topic HMM. The best performance was obtained with 60 classes and 30 states of PLSA and the Topic HMM, respectively.

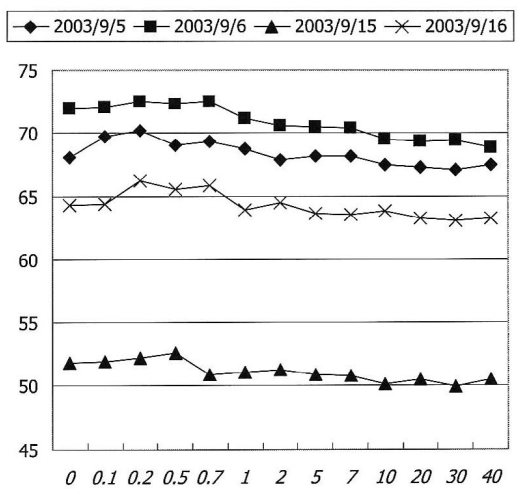


Fig. 5 Effect of transition probability of the Topic HMM on word accuracy. Horizontal axis shows the scaling factor α of transition probability.

Topic HMM. Note that, this is not the average performance but the case of best performance for each case of commentary. In a case where the scaling factor is zero, the transition probability is not used. We can see that too strong effect of transition probability causes decreasing performance, however appropriate effect may provide same level or slight better performance.

7. Conclusions

In this paper, we propose the language modeling method using Topic HMM based on a PLSA framework. The Topic HMM is learned from latent class distributions of each utterance estimated by PLSA. The decoding is performed using a unigram rescaling technique to combine the Topic HMM and trigram models. We performed the experiments on the task of a commentary from a baseball game. The experimental results show that Topic-HMM improves the word accuracy performance.

In the future, we will perform experiments with a general corpus such as CSJ or JNAS. Moreover, we will study automatic determination techniques for the Topic HMM topology, such as the number of topics of PLSA and the number of states of the Topic HMM.

References

[1] Y. Ariki, T. Shigemori, T. Kaneko, J. Ogata, and M. Fujimoto, "Live speech recognition in sports games by adaptation of acoustic model and language model," Eurospeech2003, pp.1453–1456, 2003.

[2] A. Sako and Y. Ariki, "Structuring baseball live games based on speech recognition using task dependent knowledge and emotion state recognition," ICASSP 2005, pp.1049–1052, 2005.

[3] E. Morikawa, M. Yokoyama, Y. Sugita, and K. Shirai, "Spoken dialogue modeling based on statistical approach," Proc. Information Processing Society of Japan, Spoken Language Processing, vol.97, no.16 (19970207), pp.131–137, 1997.

[4] T. Kawabata, "Topic focusing mechanism for speech recognition based on probabilistic grammar and topic Markov model," Proc. ICASSP-95, vol.1, pp.317–320, 1995.

[5] M. Woszczyna and A. Waibel, "Inferring linguistic structure in spoken language," ICSLP 1994, vol.2, pp.847–850, 1994.

[6] T. Nagano, M. Suzuki, A. Ito, and S. Makino, "Language modeling using stochastic switching N-gram," Proc. 18th International Congress on Acoustics, V, pp.3697–3700, 2004.

[7] J.R. Bellegarda, "Exploiting latent semantic information in statistical language modeling," Proc. IEEE, vol.88, no.8, pp.1279–1296, 2000.

[8] D. Gildea and T. Hofmann, "Topic-based language models using EM," Eurospeech'99, pp.2167–2170, 1999.

[9] T. Hofmann, "Probabilistic latent semantic analysis," Proc. Fifteenth Conference on Uncertainty in Artificial Intelligence (UAI'99), 1999.

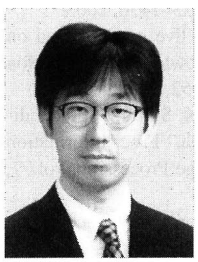
[10] S. Furui, K. Maekawa, and H. Isahara, "Spontaneous speech: Corpus and processing technology," The Corpus of Spontaneous Japanese, pp.1–6, 2002.

[11] J. Ogata and Y. Ariki, "Syllable-based acoustic modeling for Japanese spontaneous speech recognition," Eurospeech 2003, pp.2513–2516, Sept. 2003.

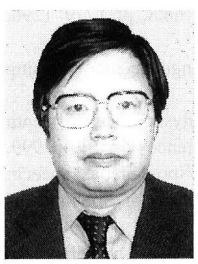
[12] E. Thelen, X. Aubert, and P. Beyerlein, "Speaker adaptation in the philips system for large vocabulary continuous speech recognition," ICASSP 1997, pp.1035–1038, 1997.

[13] A. Lee, T. Kawahara, K. Takeda, M. Mimura, A. Yamada, A. Ito, K. Ito, and K. Shikano, "Continuous speech recognition consortium — An open repository for CSR tools and models," Proc. IEEE Int'l Conf. on Language Resources and Evaluation, 2002.

Atsushi Sako received the B.E. degree in Electronics and Informatics from Ryukoku University, Shiga, Japan, in 2004, and the M.E. degree in Computer Science and Systems Engineering from Kobe University, Hyogo, Japan, in 2006. He is currently pursuing the Ph.D. degree in Informatics and Electronics at Kobe University, Hyogo, Japan. His research interests include situation dependent speech recognition and spoken language understanding. He is a member of Acoustical Society of Japan (ASJ), the Information Processing Society of Japan (IPSJ) and the International Speech Communication Association (ISCA).



Tetsuya Takiguchi received the B.S. degree in applied mathematics from Okayama University of Science, Okayama, Japan, in 1994, and the M.E. and Dr. Eng. degrees in information science from Nara Institute of Science and Technology, Nara, Japan, in 1996 and 1999, respectively. From 1999 to 2004, he was a researcher at IBM Research, Tokyo Research Laboratory, Kanagawa, Japan. He is currently a Lecturer with Kobe University. His research interests include robust speech recognition, auditory scene analysis, and microphone arrays. He received the Awaya Award from the Acoustical Society of Japan in 2002. He is a member of the IEEE, the IPSJ, and the ASJ.



Yasuo Ariki received his B.E., M.E. and Ph.D. in information science from Kyoto University in 1974, 1976 and 1979, respectively. He was an assistant professor at Kyoto University from 1980 to 1990, and stayed at Edinburgh University as visiting academic from 1987 to 1990. From 1990 to 1992 he was an associate professor and from 1992 to 2003 a professor at Ryukoku University. Since 2003 he has been a professor at Kobe University. He is mainly engaged in speech and image recognition and interested in information retrieval and database. He is a member of IEEE, IPSJ, JSAI, ITE, ASJ and IIEEJ.