論文 / 著書情報
Article / Book Information

Title	Dynamic Bayesian network-based acoustic models incorporating speaking rate effects
Authors	Takahiro Shinozaki, Sadaoki Furui
出典 / Citation	IEICE Transactions on Information and Systems, Vol. E87-D, No. 10, pp. 2339-2347
発行日 / Pub. date	2004, 10
URL	http://search.ieice.org/
権利情報 / Copyright	本著作物の著作権は電子情報通信学会に帰属します。 Copyright (c) 2004 Institute of Electronics, Information and Communication Engineers.

PAPER

Dynamic Bayesian Network-Based Acoustic Models Incorporating Speaking Rate Effects

Takahiro SHINOZAKI[†a)], *Nonmember and* Sadaoki FURUI[†b)], *Fellow*

SUMMARY One of the most important issues in spontaneous speech recognition is how to cope with the degradation of recognition accuracy due to speaking rate fluctuation within an utterance. This paper proposes an acoustic model for adjusting mixture weights and transition probabilities of the HMM for each frame according to the local speaking rate. The proposed model is implemented along with variants and conventional models using the Bayesian network framework. The proposed model has a hidden variable representing variation of the "mode" of the speaking rate, and its value controls the parameters of the underlying HMM. Model training and maximum probability assignment of the variables are conducted using the EM/GEM and inference algorithms for the Bayesian networks. Utterances from meetings and lectures are used for evaluation where the Bayesian network-based acoustic models are used to rescore the likelihood of the N-best lists. In the experiments, the proposed model indicated consistently higher performance than conventional HMMs and regression HMMs using the same speaking rate information.

key words: *spontaneous speech recognition, speaking rate, dynamic Bayesian network, acoustic modeling*

1. Introduction

Although conventional HMM-based recognition systems work well for speech in the form of reading a written text, performance is quite poor for spontaneous speech. One of the main factors that makes the recognition of spontaneous utterances difficult is a large variation of the speaking rate. This paper explores several extensions of the HMM to explicitly model the effects of the speaking rate variation. These models are realized by using the dynamic Bayesian network framework, which has an ability to model complex probabilistic dependencies.

Various analyses of spontaneous speech recognition results have revealed that the speaking rate affects recognition performance on many levels. The relationship between individual differences in the speaking rate, defined as an averaged phone rate for a speaker, and the averaged word accuracy was analyzed in [1]. Dependency between the utterance-level speaking rate and the accuracy was reported in [2]. A word-level analysis was conducted in our previous study using the Corpus of Spontaneous Japanese [3], in which the capability of predicting recognition errors of decision trees was used to measure the influence of word attributes including the speaking rate [4]. The results show that the speaking rate, the number of occurrences in the training set, and the number of phones are the most important word attributes for the prediction. Specifically, as shown in Fig. 1, decision trees using these three attributes are as effective as trees using all the attributes listed in Table 1. The trees were trained to predict whether a word could be correctly recognized, using pairs of word attributes and correctness of the recognition result. The performance was evaluated by whether prediction of each recognition result is true or false for words in spontaneous speech that are independent of those used for training the tree. In the figure, AllAtt indicates the correctness of the trees using all the attributes. P, R, and W indicate the number of phonemes in a word, the speaking rate and the word frequency, respectively. It can be seen that omitting any one of the three attributes degrades the prediction performance of the trees.

The reasons for the adverse effect of speaking rate fluctuation include pronunciation variation such as phone deletion, spectral modification, and more directly, the deviation of the speaking rate itself which then causes a mismatch in transition probabilities modeled by the HMM.

A possible strategy to manage this problem is first estimating the speaking rate and then adjusting a recognizer based on the speaking rate. Sentence level acoustic model selection has been described in [5]. The fastest sentences are selected based on the speaking rate calculated by using the 1st pass recognition results, and re-recognized us-

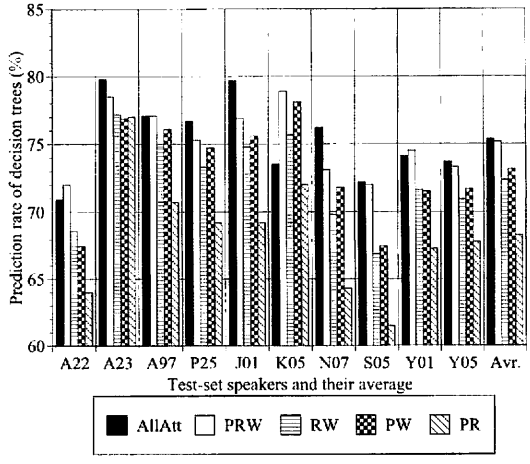


Fig. 1 Contribution of the word attributes to explaining the recognition errors. AllAtt indicates the correctness of the trees using all the attributes. P, R, and W indicate the number of phonemes in a word, the speaking rate and the word frequency, respectively.

Manuscript received December 17, 2003.
Manuscript revised April 22, 2004.

[†]The authors are with the Department of Computer Science, Tokyo Institute of Technology, Tokyo, 152-8552 Japan.
a) E-mail: staka@furui.cs.titech.ac.jp
b) E-mail: furui@furui.cs.titech.ac.jp

Table 1 Word attributes.

Number of phonemes in the word
Word duration (number of frames)
Speaking rate (number of phonemes/number of frames)
Averaged acoustic frame likelihood
Ratio of a certain phoneme class such as vowel or nasal
Part of speech (noun, verb, etc.)
Filled pause or not
Repair or not
Quotation or not
Loan word or not
Word frequency in the training set
Bigram score
Trigram score
Back off class
Word order in the sentence from either beginning or end
Part of speech of the left/right context word
Left/Right context word is filled pause or not
Left/Right context word is repair or not
Left/Right context word is quotation or not
Left/Right context word is loan word or not

ing an acoustic model adapted to those fastest sentences. This method is easy to implement and computationally inexpensive but cannot compensate for speaking rate variation within a sentence. In [6], frame level regulation using regression HMMs has been reported. In this case, acoustic observation density is controlled for each frame, but transition probability is left untouched. A way of modifying pronunciation and acoustic likelihood using a hidden mode is shown in [7]. This modification of pronunciations based on the hidden mode variable has been implemented in [8]. Modification of the acoustic likelihood has been conducted in [9] where speaking rate information is used for each frame. In this method, transition probability was partially controlled in addition to the acoustic observation probability. One difficulty of this modeling was a large increase in the number of model parameters.

Since standard HMM is not powerful enough to model complex dependencies, several extensions have been made. However, such kind of extensions often require large effort for their realization and many other possible extensions are then left untouched. For example, there are many possibilities for how to use the hidden mode variable. The Bayesian network is a flexible statistical framework on which such novel probabilistic models can be rapidly employed [10]–[13]. In [10], an idea of using a Bayesian network for compensating for a changing speaking rate is also mentioned, but experiments using the network were not conducted. This paper explores possibilities of several DBN based acoustic models that have hidden mode variable to deal with the speaking rate variation. These models extend a conventional HMM by modifying the parameters of Gaussian mixtures and/or transition probabilities according to the speaking rate frame by frame. These models are evaluated using utterances from meetings and lectures as test sets by rescoring N-best lists which are generated by a Bigram decoder with a 30 k vocabulary size.

This paper is organized as follows. In Sect. 2, the conventional and proposed models are formulated as a Bayesian network. In Sect. 3, several techniques for measuring speaking rate are reviewed. Experimental results are described and discussed in Sect. 4. It is shown that the proposed models using a hidden mode variable are more effective in improving the recognition rate than a regression HMM using the same speaking rate information. Especially, a hidden mode HMM that adjusts both mixture weights and transition probabilities depending on the speaking rate is the most effective. Finally, the paper is concluded in Sect. 5.

2. DBN Based Acoustic Modeling

In this section, a way of formulating the HMM as a Bayesian network is reviewed and a baseline network for encoding the HMM is defined. Then, several models that extend the HMM are described. Since model complexity and estimation accuracy of the parameters from a training set always pose a trade-off, special attention is payed for the number of parameters of the models.

2.1 Bayesian Network

Bayesian networks are directed graphs in which nodes represent random variables, and edges represent probabilistic dependency relations. A Bayesian network is defined by the graph structure and the Conditional Probability Distribution (CPD) at each node. There are several ways how to define the CPDs. For example, if the variable of the node and those of its parents are both discrete, the CPD can be represented as a Conditional Probability Table (CPT), which lists the probability that the node takes on each of its different values for each combination of values of its parents. When the variable of the node is continuous and the parents are discrete-valued, a set of Gaussian mixtures can be used where each element corresponds to a combination of values of its parents [14].

Since speech recognition is a process for time series of feature vectors, Dynamic Bayesian Networks (DBN) [15] are ideally suited for this purpose. DBNs are Bayesian networks that have directed edges pointing in the direction of time. DBNs have a repeating topology of a common core structure, and the CPDs do not change with time.

2.2 Baseline Model

Figure 2 shows an example of a phone HMM set modeling phones /a/ and /b/. Each phone model consists of three states with a left-to-right topology. Figure 3 shows the DBN structure that models the phone HMM sequence for model training and N-best rescoring [10] where the discrete variable **Phone-Counter** indicates position in the phone sequence and its value is incremented when binary random variable **Phone-Transition** posts it is phone transition. The node **End-of-utterance** is necessary to ensure that the process ends with a transition out of the last phone. In the figure, observed variables are indicated by shading their nodes.

Also, continuous nodes are denoted by circles while discrete nodes are expressed by squares.

In the phone HMM set, a probability density function for acoustic feature vectors is specified by a phone index and the state index of the phone. The Bayesian network has a node **Phone** that represents a phone index and **Phone-State** that represents a state index of the phone. As abbreviated in Fig. 4, the node **Observation** which corresponds to acoustic observation, has incoming arrows from the nodes **Phone** and **Phone-State**. This means that the probability the value of **Observation** takes is dependent on these values since each node in a Bayesian network represents a random variable. Similarly, a phone state transition probability to the next HMM state is modeled by a node **Phone-State-Transition** that has incoming arrows from **Phone** and **Phone-State** indicating probabilistic dependency on these variables. Equation (1) and Eq. (2) show these dependencies of the acoustic observation and the transition probabilities, respectively.

$$P = P(O|P,S), \tag{1}$$

$$P = P(T|P,S). \tag{2}$$

In the equations, O is a single-letter abbreviation of the **Observation** variable for referential convenience, P is **Phone**, S is **Phone-State**, and T is **Phone-State-Transition**.

The node **Phone-State-Transition** represents a binary random variable that indicates either staying at the HMM state or moving to the next state, since the HMM has a left-to-right topology. In this example, cardinalities of the discrete random variables **Phone** and **Phone-State** are two and three, respectively, corresponding to the number of phones and the maximum number of states for each phone. The acoustic observation is a vector of real numbers and **Observation** is a continuous random variable.

A Bayesian network used as a baseline acoustic model has the same structure but larger cardinality for **Phone**. The CPD of the observation node **Observation** is defined using a set of diagonal covariance Gaussian mixtures. Parameters of the network are trained using EM/GEM algorithm on a Bayesian network. Decoding is performed by assigning values for all the hidden variables so as to maximize the joint probability of the entire network. Hereafter, the baseline network is referred to as **BASE**.

2.3 Regression HMM

One possible way of controlling acoustic observation probability density is to use regression models, in which mean values of the Gaussian components are modeled by linear combination of explanation variables. A multiple-regression HMM has been proposed in [16] where F0 information was used as an auxiliary feature for the explanation variables. The mean vector μ of each Gaussian component is expressed as,

$$\mu = R \cdot \xi + \mu_0, \tag{3}$$

where R is the regression coefficient matrix, μ_0 is the constant term, and ξ is the auxiliary vector. Similar models has been proposed and implemented as DBNs in [6], [17] in which F0 and speaking rate are used as auxiliary information.

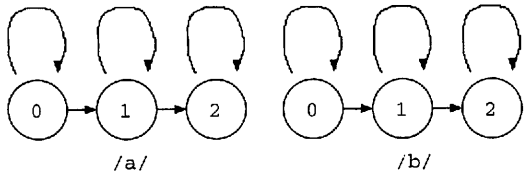


Fig. 2 A phone HMM set consisting of two phones. Each phone is model by a three-state left-to-right HMM.

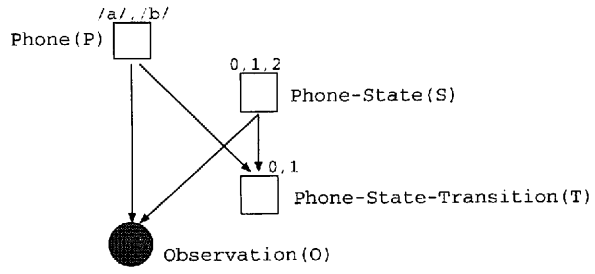


Fig. 4 A portion of a time slice of the DBN in Fig. 3 that encodes the conventional HMM. (BASE)

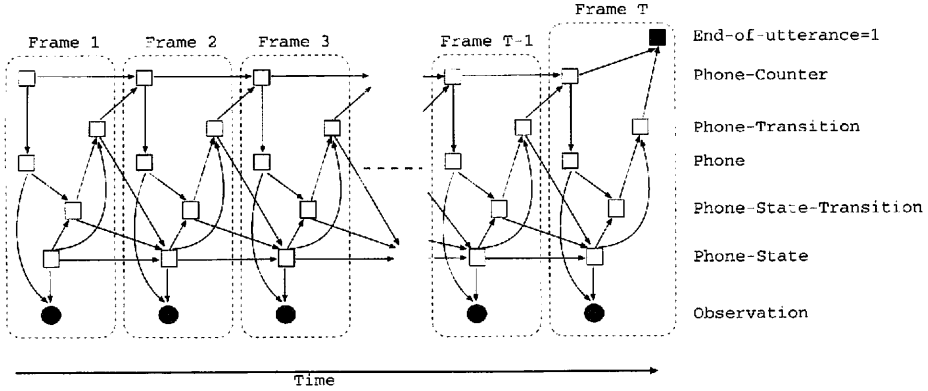


Fig. 3 DBN representation of the phone HMM sequence. Circles denote continuous-value nodes, squares denote discrete nodes, clear means hidden, and shaded symbols indicate observed nodes.

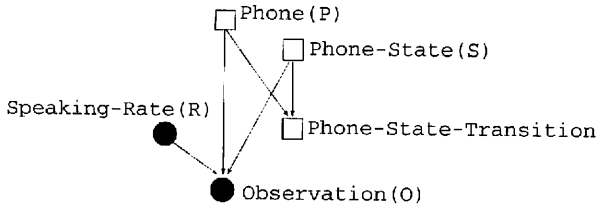


Fig. 5 Regression model. (REG)

In this paper, a DBN version of the multiple-regression HMM is evaluated using a speaking rate and the second and third order terms as explanation variables. The parameters added to the **BASE** model are regression coefficient matrix components that have the same row dimension as the mean vectors and a column dimension of three. The matrices are tied among Gaussian mixture components in each phone to reduce the number of parameters required to define the model. The Bayesian network representation of this model is shown in Fig. 5 where there is an additional node **Speaking-Rate** that represents the speaking rate compared to **BASE**. An arrow directly connecting **Speaking-Rate** and **Observation** expresses the dependency between **Observation** and **Speaking-Rate**. The acoustic observation probability is expressed as shown in Eq. (4), where O is the **Observation** variable, P is **Phone**, S is **Phone-State** and R is **Speaking-Rate**. This model is hereafter called **REG**.

$$P = P(O|P, S, R) \tag{4}$$

2.4 Hidden Mode Mixture Weight Model

Figure 6 shows a Bayesian network of our proposed model in which the acoustic observation node **Observation** has different probability density according to a "mode" of the speaking rate. In this network, two nodes are added to **BASE**; **Mode** and **Speaking-Rate**. **Mode** is a discrete hidden random variable that represents a "mode" of the speaking rate. As indicated by a dotted line in the figure, **Mode** depends on its counterpart in the previous time slice. This dependence is introduced based on an assumption that the speaking rate changes continuously. A CPT is used at this node. **Speaking-Rate** is a one-dimensional continuous random variable of the speaking rate and a set of Gaussian distributions are used for CPD at this node. In this configuration, both the acoustic observation node **Observation** and the speaking rate observation node **Speaking-Rate** have the node **Mode** as their parent.

The CPD at node **Observation** has a different Gaussian mixture for each combination of the values of **Phone**, **Phone-State**, and **Mode**. This means that the CPD has $|Mode|$ times more Gaussian mixtures than **BASE**, where $|Mode|$ is the cardinality of the **Mode** variable. Usually, Gaussian mixtures dominate the number of parameters of an HMM. To reduce the number of parameters for accurate model estimation, the Gaussian components are tied for the different values of **Mode**. That is, different values

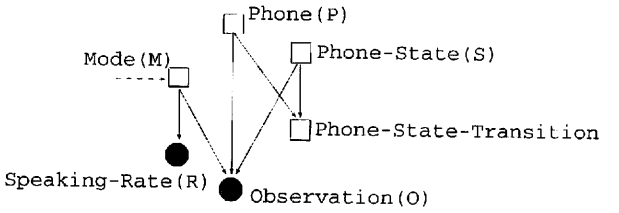


Fig. 6 Hidden mode mixture weight model. The dotted link represents an edge from the previous time frame. (HM-MW)

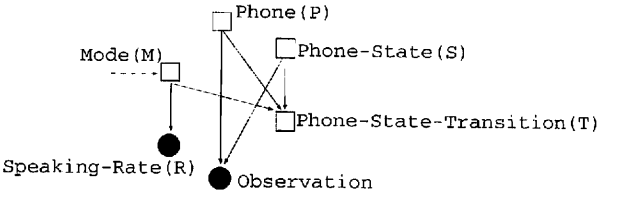


Fig. 7 Hidden mode transition probability model. (HM-TRP)

of **Mode** specify different Gaussian mixture weights for the same Gaussian component.

Speaking-Rate has different distributions of the speaking rate depending on **Mode**, and this is used to detect a mode of the speaking rate. The Gaussian mixtures of **Observation** are modified based on a value of **Mode** by choosing different Gaussian mixture weights, and this is how to compensate for spectral change. Note that the speaking rate mode of each frame is not completely determined simply by the speaking rate but by considering the entire likelihood of the network using an inference algorithm on a Bayesian network. Hereafter, this model adjusting the mixture weights for each time frame by using the hidden mode variable is referred to as **HM-MW**.

Newly introduced parameters in addition to those used in **BASE** are: a CPT of size $|Mode| \times |Mode|$ at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at **Speaking-Rate**, and $|Mode| - 1$ mixture weight vectors for each combination of the values of **Phone** and **Phone-State** at **Observation**. Note that this configuration is applicable not only to the speaking rate but also to any temporal fluctuation that affects speech features.

2.5 Hidden Mode Transition Probability Model

In the model described in the previous subsection, observation probabilities of an underlying HMM are controlled by a hidden mode variable. It is also possible to control transition probabilities by using the hidden mode variable as shown in Fig. 7. The parameterization for the variables **Mode** and **Speaking-Rate** are the same as **HM-MW**. **Mode** is a discrete hidden random variable used to represent the speaking rate mode and **Speaking-Rate** is a one-dimensional continuous random variable modeling the speaking rate.

Additional parameters to those used in **BASE** are: a CPT of size $|Mode| \times |Mode|$ at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at

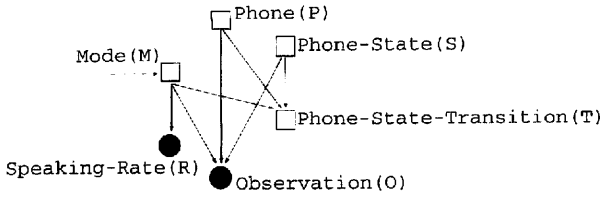


Fig. 8 Hidden mode HMM. (HM-HMM)

Speaking-Rate, and $|Mode| - 1$ transition probabilities for each combination of the values of **Phone** and **Phone-State** at **Observation**. Since the number of parameters required for modeling transition probabilities are fewer than for the Gaussian mixtures, they are separately modeled for each value of **Mode**. This model is hereafter called **HM-TRP**.

2.6 Hidden Mode HMM

The controls of the mixture weights and the transition probabilities can be combined as shown in Fig. 8. The variables introduced to control the underlying HMM parameters are **Mode** and **Speaking-Rate**. **Mode** is a discrete hidden random variable to represent the speaking rate mode and **Speaking-Rate** is a one-dimensional continuous random variable to model the speaking rate, as already explained in the previous subsections.

Additional parameters to those used in **BASE** are union of the additional parameters of **HM-MW** and **HM-TRP**, that is, a CPT of size $|Mode| \times |Mode|$ at **Mode**, a one-dimensional Gaussian distribution for each value of **Mode** for CPD at **Speaking-Rate**, and $|Mode| - 1$ mixture weight vectors and transition probabilities for each combination of the values of **Phone** and **Phone-State** at **Observation**. This model is hereafter called **HM-HMM**.

3. Measurement of Speaking Rate

Many approaches have been reported for calculating/defining the speaking rate. They can be roughly divided into two categories, that is, lexical measures and signal based measures.

Lexical measures count units such as words or phones in a certain period. When correct transcription is available, these measures can be calculated by the forced alignment technique. When the correct transcription is not available, a recognition hypothesis can be used instead. A disadvantage of this method is that the hypothesis is not always correct and the errors degrade the reliability of the estimated speaking rate. Thus when the speaking rate is used to control the recognition system, it is possible that the estimated speaking rate is less accurate for speech segments where the control is more important.

The signal based measures directly estimate speaking rate without relying on the transcription and thus can avoid the problem of the lexical measures. Enrate is one of such measures proposed in [18]. This is defined as the first spectral moment for the wideband energy envelope of the speech signal. The spectral range is approximately restricted between 1 and 16 Hz. The concept of the enrate is based on the fact that the energy envelope of speech rapidly changes when the speaking rate is high. The enrate can be considered as a conversion of TEMAX-gram [19], which was developed to observe the speaking rate as a spectrogram, into a scalar value. Although the correlation between the enrate and the phone or syllable rate is not high, it has been shown in [18] that the enrate is a good predictor of recognition errors.

To improve the correlation with the lexical measures, mrate was proposed in [20]. This is a linear combination of the enrate and peak-counting estimators. The correlation between the syllable rate and the mrate is over 0.6, whereas correlation with the enrate is approximately 0.4 for manually transcribed Switchboard data.

In [21], another way of estimating the speaking rate by detecting vowels has been shown. Modified loudness defined as a difference of higher frequency band loudness and lower frequency band loudness is calculated for every frame. The main part of the energy of a vowel concentrates on lower frequencies, whereas that for the most consonants is located at higher frequencies. Therefore, vowels make peaks in the modified loudness and thus they can be detected by finding maxima of the modified loudness. Speaking rate is obtained by taking an inverse of the vowel frequency.

In the following experiments, lexical measures derived from correct and hypothesized transcriptions and the enrate signal based measure are used. These measures are calculated for each frame of acoustic observation features using significantly overlapped analysis windows.

4. Experiments

4.1 Corpora and Tasks

Two spontaneous speech corpora were used to train and evaluate the DBN based acoustic models. One was a corpus of the Meeting Recorder Project [22] and the other was Corpus of Spontaneous Japanese (CSJ) [3]. Utterances gathered by the Meeting Recorder Project are recorded from meetings with natural settings, and contain background noises and speech overlaps by other speakers. CSJ consists of Japanese academic lecture speech and extemporaneous public speech. Speaker dependent experiments were conducted for the meeting data and speaker independent systems were evaluated using the lecture data. For both of the experiments, utterances recorded using close talking microphones were used.

Speaker dependent models were made using the utterances produced by one male speaker extracted from the meeting corpus. Utterances at nine meetings were used for training, and one meeting was used for testing. Lengths of the utterances for training and testing were 97 and 10 minutes, respectively. Speaker independent experiments were conducted using academic lectures given by male speakers. Ten lectures and five lectures were selected for training and

Table 2 Characteristic of the acoustic models of the tasks.

Task	ICSI meetings	CSJ lectures
Language	English	Japanese
Model type	SD	SI
Feature kind	MFCC_0_D_A	MFCC_E_D_N_Z
Feature dimension	39	25
Window width	25 ms	25 ms
Frame shift	10 ms	10 ms
# of phones	45	42
# of mixtures per state	64	28

testing, respectively. They were subsets of the CSJ official test sets and there was no overlap between training and testing speakers. The amount of the training set was 116 minutes and the test set was 16 minutes. Table 2 shows these conditions.

4.2 Model Training

First a monophone HMM set was made using the training set and HTK. The parameters of the DBN based acoustic models were initialized with the HMM. Then they were trained by the EM/GEM algorithms using GMTK [23] with 10 iterations.

Each phone of the monophone set was modeled by a three state HMM with a left-to-right topology. The number of Gaussian mixtures per monophone state was determined so as to maximize the recognition rate of the task by preliminary experiments; 64 for the meetings and 28 for the lectures. Table 2 shows the characteristic of the acoustic models.

Since the parameters of **Mode** and **Speaking-Rate** do not have corresponding values in the HMM, they were initialized with arbitrary values. For **HM-MW** and **HM-HMM**, the mixture weights were initialized by copying the mixture weights of the monophone HMM. Similarly, for **HM-TRP** and **HM-HMM**, the transition probabilities were initialized by copying those of the monophone HMM. Regression coefficient matrices for **REG** were initialized by giving zeros to all the elements.

After the initialization, most of the trainable parameters, including that of the **Mode**, **Speaking-Rate**, and the regression coefficient matrices, were trained. Only the variances of the Gaussian components in the acoustic observation nodes **Observation** of the networks used for the meeting task were kept constant. For these DBN acoustic models other than **BASE**, speaking rate information was also used in addition to the normal acoustic features. For **REG**, the speaking rate was normalized so that the mean value became zero for the training set. This made it reasonable to initialize the Gaussian components of the model using those of the monophone HMM.

4.3 Experiments Using Oracle Speaking Rate

To investigate the effect and limit of the acoustic models, speaking rate information derived from forced alignment of correct phone state sequences with the utterances were used for both training and testing the acoustic models. The speaking rate was defined as an inverse value of the state holding time. The observed values were smoothed using Eq. (5), where $SR_I(t)$ and $SR_S(t)$ indicate time series of the speaking rate before and after smoothing.

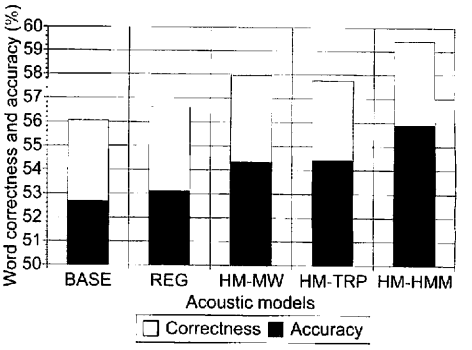


Fig. 9 Word correctness and accuracy of the meeting task given speaking rate measured using true transcript.

$$SR_S(t) = \sum_{s=-20}^{20} SR_I(t+s) \cdot (20 - |s|). \quad (5)$$

The DBN based acoustic models were evaluated by rescoring N-best lists using GMTK with a single pass of max-product inference. The N-best lists were generated using the monophone HMM that was used to initialize the DBN models and a Bigram language model. The Bigram model used for the meeting task was trained on the HUB5E and the one used for the lecture task was trained on 6.7 million words of transcriptions from the CSJ. Their vocabulary sizes were both 30 k. The number of hypotheses generated for each utterance was 50 and 100 for the meeting and the lecture tasks, respectively. The cardinality of the hidden discrete variable **Mode** was set to four.

Figure 9 shows the recognition results of the meeting task. The word accuracy of the baseline model **BASE** was 52.7%, and the absolute improvement of the word accuracy by **REG** and **HM-MW** compared to **BASE** was 0.4% and 1.7% respectively. By controlling the transition probabilities, **HM-TRP** improved the accuracy by 1.7%. The most effective model was **HM-HMM** combining **HM-MW** and **HM-TRP**. This model improved the accuracy by 3.2% for the absolute value by controlling both the mixture weights and the transition probabilities. Similar results were obtained for the lecture task as shown in Fig. 10. The improvement by **HM-HMM** was 2.1% in this case. The sentence level model selection corresponds to fixing the **Mode** variable during an utterance. As an additional experiment, **HM-HMM** was trained and evaluated with this constraint. In this case, no apparent improvement was observed for both of the tasks.

Although both **REG** and **HM-MW** models modify Gaussian mixtures based on the speaking rate, **HM-MW** achieved higher improvement than **REG**. One disadvantage of **REG** might be that it deterministically changes the mean values of the Gaussian components according to the speak-

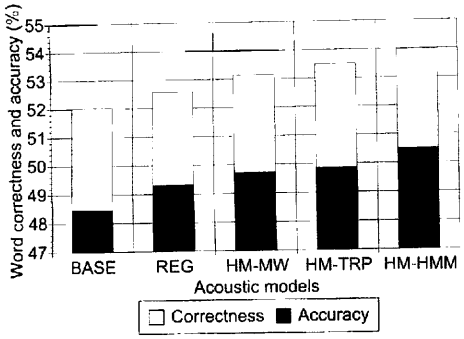


Fig. 10 Word correctness and accuracy of the lecture task given speaking rate measured using true transcript.

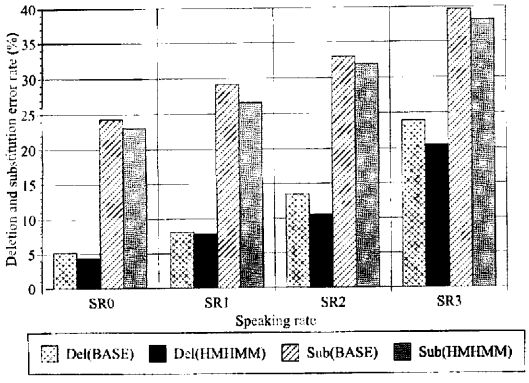


Fig. 11 Error distribution for speaking rate.

ing rate. Even if the true speaking rate information is used, it is possible that at some time frame a given speaking rate does not match the local effects of the speaking rate in terms of the changes of the acoustic characteristics, since it has been smoothed as mentioned above. Moreover, it is possible that the relationship between the speaking rate and the change of speech spectra is essentially probabilistic. **HM-MW**, on the other hand, probabilistically chooses a speaking rate mode considering the entire likelihood of the network and therefore it has a capability to select a mode that does not directly match the speaking rate. This feature was obtained by introducing the hidden variable **Mode** for representing the mode of speaking rate.

Mean deletion and substitution error rates with **BASE** and **HM-HMM** for different speaking rates are shown in Fig. 11. The speaking rate was classified into four classes; SR0 is the slowest and SR3 is the fastest. The speaking rate was calculated for each correct word by averaging phone rates using correct transcription. Therefore, insertion errors are not counted. As can be seen in the figure, both the deletion and substitution error increase for **BASE** as the speaking rate increases. Reduction of the deletion error by **HM-HMM** is higher at faster speaking rates. For substitution errors, **HM-HMM** has relatively uniform effect across different speaking rates. Similar error tendencies are observed for the lecture task, though the result is not shown in the figure.

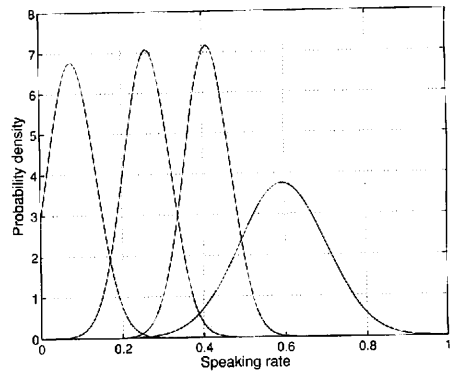


Fig. 12 Gaussian distributions for the variation of the speaking rate mode trained on the ICSI meetings.

Our proposed models, **HM-MW**, **HM-TRP**, and **HM-HMM** have a discrete hidden variable **Mode** that represents a speaking rate mode as explained in Sect. 2. Although the cardinality of the variable is specified beforehand, the correspondence between the value of the variable and the speaking rate is obtained through a training process using a set of Gaussian distributions at **Speaking-Rate**. The distributions are estimated so as to maximize the entire likelihood of the network taking the dependencies on mixture weight and/or transition probability into account. Figure 12 shows the four one-dimensional Gaussian distributions of **HM-HMM** corresponding to each value of the **Mode** estimated using the ICSI meetings. As can be seen in the figure, different values of the **Mode** have different features of the speaking rate.

4.4 Experiments without Using Oracle Speaking Rate

Rescoring experiments without relying on the true transcription were conducted using two different speaking rate measures for **REG** and **HM-HMM**. One measure was **HYP** which was similar to the one used in the oracle experiments with the exception of using the one-best hypothesis in the N-best list as an approximation of the true transcription. For the rescoring, the same acoustic models as the previous experiments were used. The other was **ENRATE** which was the enrate measure. Window width for the enrate calculation was set at 400 ms based on our preliminary experiments. When rescoring, acoustic models trained with enrate were used.

Tables 3 and 4 show the results for the meeting and lecture tasks, respectively. In the table, the results by the baseline model without using the speaking rate information indicated by **BASE** and those by using the speaking rate calculated from true transcription indicated by **ORACLE** are also shown. The cardinality of **Mode** was set to three and four.

As can be seen in Table 3, no improvement was obtained by the regression model **REG** for the meeting task regardless of using **HYP** or **ENRATE** measures. This is probably because the regression model is vulnerable to the decrease of the quality of the speaking rate. Because the

Table 3 Word accuracy of the meeting task.

	REG	HM-HMM |Mode|=3	HM-HMM |Mode|=4
BASE		52.7	
HYP	52.4	53.4	53.0
ENRATE	52.5	53.5	53.1
ORACLE	53.1	55.3	55.9

Table 4 Word accuracy of the lecture task.

	REG	HM-HMM |Mode|=3	HM-HMM |Mode|=4
BASE		48.5	
HYP	49.0	49.3	49.7
ENRATE	48.6	48.8	48.7
ORACLE	49.3	50.0	50.5

one-best hypothesis includes recognition errors, **HYP** is not an accurate approximation of the oracle speaking rate. Although **ENRATE** is free from the recognition errors, it seems to be less effective in explaining the change of acoustic features compared to the oracle speaking rate. **HM-HMM** succeeded in exploiting the speaking rate information to improve the word accuracy. When cardinality of **Mode** was set to three, an absolute improvement of 0.7% and 0.8% was obtained for **HYP** and **ENRATE**, respectively. For the lecture task, as Table 4 indicates, the highest improvement of 1.3% was found for **HM-HMM** with **HYP** measure where the cardinality of **Mode** is set to four. The optimal cardinality of **Mode** probably depends on the underlying HMM complexity such as number of mixtures, amount of training data, and estimation accuracy of the speaking rate.

5. Conclusions

This paper has explored several dynamic Bayesian network based acoustic models for improving recognition accuracy of spontaneous speech using explicitly modeled effect of the speaking rate. Although the DBN based recognition system is slower than conventional systems that are highly tuned for the speech recognition domain, it is beneficial to use the DBN for analyzing underlying principles and prototyping.

When speaking rate information obtained from the true transcription was given, our proposed models, **HM-MW**, **HM-TRP**, and **HM-HMM** indicated higher performances than **BASE**, which encodes conventional HMM, and **REG**, which encodes regression HMM using the same speaking rate information. The absolute improvement achieved by using **HM-HMM** was 3.2% and 2.1% for the meeting and lecture tasks, respectively. These DBN based acoustic models were also evaluated using speaking rate measures without using true transcriptions. Two measures were used for this purpose, best hypothesis-based speaking rate and enrate. Although the regression model **REG** sometimes failed in making use of the these speaking rates, **HM-HMM** showed improvement over the conventional models for the both tasks. In the best condition, **HM-HMM** improved the word accuracy by 0.8% for a meeting task and 1.3% for a lecture task. For both of the experiments with and without oracle speaking rate, our proposed models indicated consistently higher performance than conventional HMMs and regression HMMs using the same speaking rate information.

Future works include investigating more efficient ways of utilizing speaking rate information, finding better methods for speaking rate estimation, incorporating other spontaneous speech features to further improve the recognition accuracy, and implementing computationally efficient systems that can work with more general LVCSR conditions for promising probabilistic models found by using flexible DBN toolkits.

Acknowledgment

The authors would like to thank the members of SSLI laboratory, the University of Washington, USA, for their kind help and fruitful discussion.

References

[1] T. Shinozaki and S. Furui, "Analysis on individual differences in automatic transcription of spontaneous presentations," Proc. ICASSP2002, vol.1, pp.729–732, Orlando, FL, May 2002.
[2] H. Nanjo and T. Kawahara, "Speaking-rate dependent decoding and adaptation for spontaneous lecture speech recognition," Proc. ICASSP, vol.1, pp.725–728, Orlando, FL, May 2002.
[3] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition," Proc. ICSLP, vol.3, pp.518–521, Beijing, China, Oct. 2000.
[4] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," Proc. ASRU, a01ts039, Trento, Italy, Dec. 2001.
[5] N. Mirghafori, E. Fosler, and N. Morgan, "Towards robustness to fast speech in ASR," Proc. ICASSP, vol.1, pp.335–338, Atlanta, GA, May 1996.
[6] T. Stephenson, M. Magimai-Doss, and H. Bourlard, "Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks," Proc. ICASSP, vol.1, pp.20–23, Hong-Kong, May 2003.
[7] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld, "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," Proc. ICSLP, supplement, Philadelphia, PA, Oct. 1996.
[8] M. Finke and A. Waibel, "Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition," Proc. Eurospeech, Rhodes, vol.5, pp.2379–2382, Greece, Sept. 1997.
[9] A. Tuerk and S.J. Young, "Indicator variable dependent output probability modelling via continuous posterior functions," Proc. ICASSP, vol.1, pp.473–476, Salt Lake City, UT, May 2001.
[10] G. Zweig, Speech recognition with dynamic Bayesian networks, Ph.D. Thesis, University of California, Berkeley, 1998.
[11] G. Zweig and S. Russell, "Speech recognition with dynamic Bayesian networks," AAAI, pp.173–180, 1998.
[12] G. Zweig and M. Padmanabhan, "Dependency modeling with Bayesian networks in a voicemail transcription system," Proc. Eurospeech, pp.1135–1138, Budapest, Hungary, Sept. 1999.
[13] J. Bilmes, "Graphical models and automatic speech recognition," Technical Report UWEETR-2001-005, University of Washington, Dept. of EE, Seattle, WA, Nov. 2001.

[14] K. Murphy, "A brief introduction to graphical models and Bayesian networks," http://www.ai.mit.edu/~murphyk/Bayes/bnintro.html [Online], 1998.

[15] T. Dean and K. Kanazawa, "Probabilistic temporal reasoning," AAAI., pp.524–528, 1988.

[16] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden markov model," Proc. ICASSP, pp.513–516, Salt Lake City, UT, May 2001.

[17] T. Stephenson, J. Escofet, M. Magimai-Doss, and H. Bourlard, "Dynamic Bayesian network based speech recognition with pitch and energy as auxiliary variables," Proc. NNSP, pp.637–646, Martigny, Switzerland, Sept. 2002.

[18] N. Morgan, E. Fosler, and N. Mirghafori, "Speech Recognition using on-line estimation of speaking rate," Proc. Eurospeech, vol.4, pp.2079–2082, Rhodes, Greece, Sept. 1997.

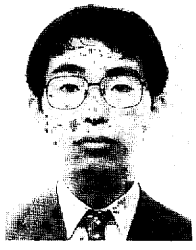
[19] S. Kitazawa, H. Ichikawa, S. Kobayashi, and Y. Nishinuma, "Extraction and representation of rhythmic components of spontaneous speech," Proc. Eurospeech, vol.2, pp.641–644, Rhodes, Greece, Sept. 1997.

[20] N. Morgan and E. Fosler, "Combining multiple estimators of speaking rate," Proc. ICASSP, vol.2, pp.729–732, Seattle, WA, May 1998.

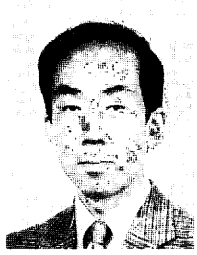
[21] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," Proc. ICASSP, vol.2, pp.945–948, Seattle, WA, May 1998.

[22] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," Proc. Human Language Technology Conference, pp.246–252, San Diego, CA, March 2001.

[23] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," Proc. ICASSP, vol.4, pp.3916–3919, Orlando, FL, May 2002.



Takahiro Shinozaki received B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology, in 1999, 2001, and 2004, respectively. He is currently a Research Scholar in Department of Electrical Engineering, University of Washington. His research interests include acoustic and language modeling for spontaneous speech recognition.



Sadaoki Furui is currently a Professor at Tokyo Institute of Technology, Department of Computer Science. He is engaged in a wide range of research on speech analysis, speech recognition, speaker recognition, speech synthesis, and multimodal human-computer interaction and has authored or coauthored over 400 published articles. From 1978 to 1979, he served on the staff of the Acoustics Research Department of Bell Laboratories, Murray Hill, New Jersey, as a visiting researcher working on speaker verification. He is a Fellow of the IEEE and the Acoustical Society of America. He was President of the Acoustical Society of Japan (ASJ) from 2001 to 2003, and is currently President of the International Speech Communication Association (ISCA) and the Permanent Council for International Conferences on Spoken Language Processing (PC-ICSLP). He was a Board of Governor of the IEEE Signal Processing Society from 2001 to 2003. He has served on the IEEE Technical Committees on Speech and MMSP and on numerous IEEE conference organizing committees. He has served as Editor-in-Chief of both Journal of Speech Communication and the Transaction of the IEICE. He is an Editorial Board member of Speech Communication, the Journal of Computer Speech and Language, and the Journal of Digital Signal Processing. He has received the Yonezawa Prize and the Paper Awards from the IEICE (1975, 88, 93, 2003), and the Sato Paper Award from the ASJ (1985, 87). He has received the Senior Award from the IEEE ASSP Society (1989) and the Achievement Award from the Minister of Science and Technology, Japan (1989). He has received the Technical Achievement Award and the Book Award from the IEICE (2003, 1990). He has also received the Mira Paul Memorial Award from the AFECT, India (2001). In 1993 he served as an IEEE SPS Distinguished Lecturer. He is the author of "Digital Speech Processing, Synthesis, and Recognition" (Marcel Dekker, 1989, revised, 2000) in English, "Digital Speech Processing" (Tokai University Press, 1985) in Japanese, "Acoustics and Speech Processing" (Kindai-Kagaku-Sha, 1992) in Japanese, and "Speech Information Processing" (Morikita, 1998) in Japanese. He edited "Advances in Speech Signal Processing" (Marcel Dekker, 1992) jointly with Dr. M.M. Sondhi. He has translated into Japanese "Fundamentals of Speech Recognition," authored by Drs. L.R. Rabiner and B.-H. Juang (NTT Advanced Technology, 1995) and "Vector Quantization and Signal Compression," authored by Drs. A. Gersho and R.M. Gray (Corona-sha, 1998).