

HIDDEN MODE HMM USING BAYESIAN NETWORK FOR MODELING SPEAKING RATE FLUCTUATION

Takahiro Shinozaki, Sadaoki Furui

Department of Computer Science, Tokyo Institute of Technology, Japan
{staka, furui}@furui.cs.titech.ac.jp

ABSTRACT

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 Bayesian networks. Utterances from meetings and lectures are used for evaluation where Bayesian network-based acoustic models are used to rescore the utterance hypotheses obtained from a first-pass N-best list. In the experiments, the proposed model shows consistently higher performance than conventional models.

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 speaking rate. This paper explores several extensions of the HMM to explicitly model the effects of speaking rate variation. These models are realized using the Dynamic Bayesian Network (DBN) framework which has an ability to model complex probabilistic dependencies.

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

A possible strategy to manage this problem is to first estimate the speaking rate and then adjust a recognizer based on the speaking rate. A way of modifying acoustic likelihood using a hidden speaking mode variable is described

in [1]. Experiments which have used the hidden speaking mode variable can be found in [2]. However, such kinds of extension of conventional HMMs often require a large effort and many other possible extensions are then left untouched.

The Bayesian network is a flexible statistical framework on which such novel probabilistic models can be rapidly employed [3]. In [3], the possibility of using a Bayesian network for compensating for a changing speaking rate is also mentioned, but the experiments using the network were not conducted. This paper explores several DBN based acoustic models to deal with 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 30k vocabulary size.

This paper is organized as follows. In Section 2, the conventional and proposed models are formulated as Bayesian networks. In Section 3, several techniques for measuring speaking rate are reviewed. Experimental results are described and discussed in Section 4. It is shown that our proposed hidden mode HMM shows consistently higher performance than conventional models for the two tasks. Finally, the paper is concluded in Section 5.

2. DBN BASED ACOUSTIC MODELING

In this section, both conventional and proposed models are formulated as Bayesian networks.

2.1. 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 1 shows the DBN structure that models a sequence of the phone HMM for model training and N-best rescoring [3]. The discrete variable **Phone-Counter** indicates position in the phone sequence and its value is incremented when the binary random variable **Phone-Transition** shows it is a phone transition. The

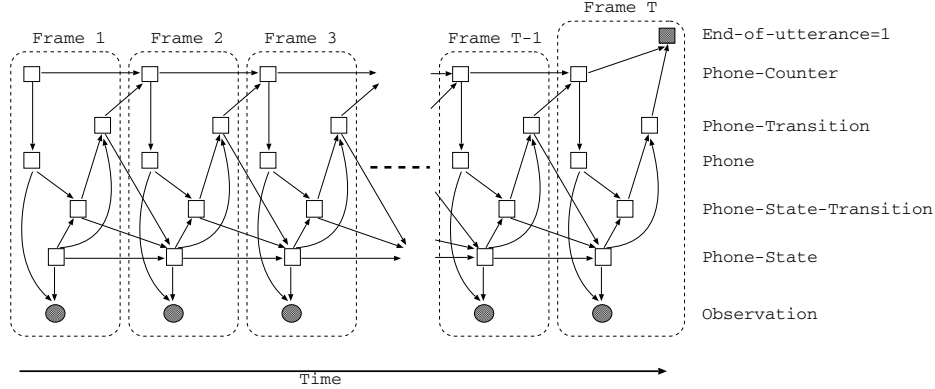


Fig. 1. DBN representation of the phone HMM sequence. Circles denote continuous-value nodes, squares denote discrete nodes, clear means hidden, and shaded symbols indicate observed nodes.

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 distribution for acoustic feature vectors is specified by a phone index and a 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 Figure 3, the node **Observation** which corresponds to an acoustic observation, has incoming arrows from the nodes **Phone** and **Phone-State**. This means that the probability the value of **Observation** takes depends 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.

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 node **Phone-State-Transition** is 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. 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 conditional probability distribution of the observation node **Observation** is defined using a set of diagonal covariance Gaussian mixtures. 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**.

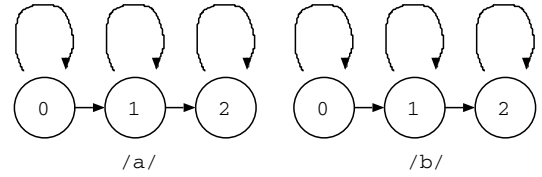


Fig. 2. A phone HMM set consisting of two phones. Each phone is modeled by a three-state left-to-right HMM.

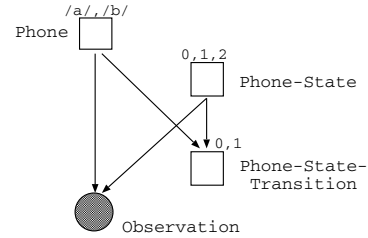


Fig. 3. A portion of a time slice of a DBN that encodes a conventional HMM. (BASE)

2.2. Regression HMM

One possible way of controlling the acoustic observation probability density is to use regression models, in which mean values of the Gaussian components are modeled by linear combination of explanatory variables. A multiple-regression HMM has been proposed in [4] where F0 information was used as an explanatory variable. The mean vector μ of each Gaussian component is expressed as,

$$\mu = R \cdot \xi + \mu_0, \quad (1)$$

where R is the regression coefficient matrix, μ_0 is the constant term, and ξ is a vector of the explanatory variables. Similar models have been proposed and implemented as DBNs in [5] in which F0 and speaking rate are used as auxiliary information. The regression HMM can also be

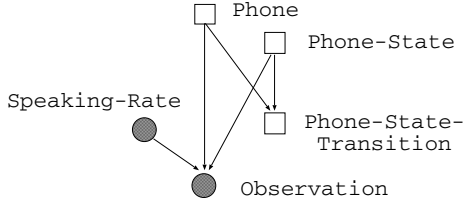


Fig. 4. Regression HMM. (REG)

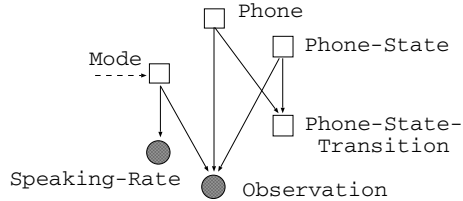


Fig. 5. Hidden mode mixture weight model. The dotted link represents an edge from the previous time frame. (HM-MW)

regarded as an instance of a hand-fixed structured Buried Markov Model (BMM) [6].

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 explanatory variables. The network of this model is similar to the BASE model but has an additional node **Speaking-Rate** as a parent of **Observation** as shown in Figure 4. The regression coefficient matrices are tied among Gaussian mixture components in each phone to reduce the number of parameters required to define the model. This model is denoted as **REG**.

2.3. Hidden mode HMM

To compensate for spectral changes due to speaking rate fluctuation, we propose a “hidden mode mixture weight model”. Figure 5 shows the Bayesian network in which the probability density function of **Observation** depends on the “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. **Speaking-Rate** is a one-dimensional continuous random variable of the speaking rate. In this configuration, both the acoustic observation node **Observation** and the speaking rate observation node **Speaking-Rate** have the node **Mode** as their parent.

A conditional probability table (CPT) is used for the discrete node **Mode**. Conditioned on the **Mode** node, **Speaking-Rate** is a Gaussian. For the **Observation** node, a Gaussian mixture is used for each combination of the values of **Phone**, **Phone-State**, and **Mode**. 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 of **Mode** specify different Gaussian mixture weights for the same Gaussian component. Hereafter, this model adjusting the mixture weights for each time frame by using the hidden mode variable is referred to as **HM-MW**.

Instead of controlling observation probabilities of an underlying HMM, it is also possible to control transition probabilities using the hidden mode variable. We propose a “Hidden mode transition probability model” as shown in Figure 6. The parameterization for the variables **Mode** and **Speaking-Rate** are the same as HM-MW. This model is hereafter called **HM-TRP**.

The controls of the mixture weights and the transition probabilities can be combined. Figure 7 shows our proposed “Hidden mode HMM”. In the network, **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. This model is hereafter called **HM-HMM**.

Note that the HM-MW model is similar to our previously proposed model [7], but the **Mode** node of HM-MW has an additional link from the same node in the previous time frame. Although only the acoustic observation probability was controlled in the previous paper, transition probability is also covered in this paper. The other difference between this paper and the last paper is the expanded experimental conditions.

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 estimated speaking rate is used to control the recognition system, it is possible that the estimate is less accurate for speech segments where the control by speaking rate 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, proposed in [8], is one such measure. 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 16Hz. 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 [9], 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 [8] that the enrate is a good predictor of recognition errors.

To improve the correlation with the lexical measures, *mrate* was proposed in [10]. 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 [11], 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 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 [12] and the other was the Corpus of Spontaneous Japanese (CSJ) [13]. Utterances gathered by the Meeting Recorder Project are recorded from meetings, 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. Experiments for a speaker independent condition were also conducted using academic lectures given by male speakers from the CSJ. Ten lectures were selected as a training set and five lectures were used for testing. The amount of the training set was 116 minutes and the test set was 16 minutes.

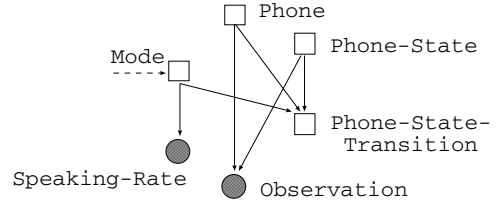


Fig. 6. Hidden mode transition probability model. (HM-TRP)

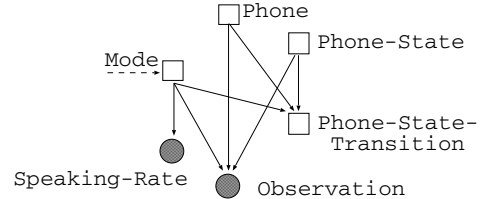


Fig. 7. Hidden mode HMM model. (HM-HMM)

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 [14] with 10 iterations. The GMTK training program determines when either EM or GEM training is appropriate, depending on the degree and type of parameter sharing.

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 1 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. This is because the number of mixtures is large in contrast with the amount of the training data. For these DBN acoustic models other than BASE, speaking rate information was also used in addition to the normal acoustic

Table 1. Characteristic of the acoustic models

	ICSI meetings	CSJ lectures
Language	English	Japanese
Feature kind	MFCC_0.D.A	MFCC_E.D.N.Z
Window width	25ms	25ms
Frame shift	10ms	10ms
# of phones	45	42
# of mixtures per state	64	28

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.

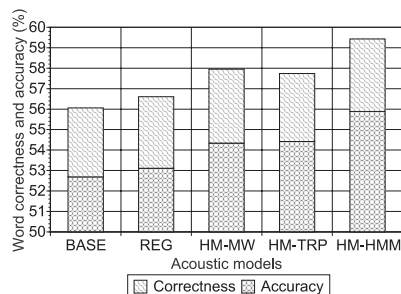
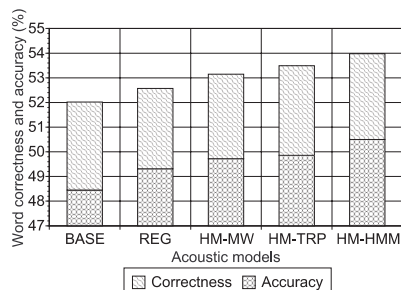
4.3. Experiments using oracle speaking rate

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 was used for both training and testing of the models. The speaking rate was defined as an inverse value of the state holding time. The observed values were smoothed using Equation (2), where $SR_I(t)$ and $SR_S(t)$ indicate time series of the speaking rate before and after smoothing.

$$SR_S(t) = \sum_{s=-20}^{20} SR_I(t+s) \cdot (20 - |s|). \quad (2)$$

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 the CSJ. Their vocabulary sizes were both 30k. 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 8 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 from BASE was 0.4% and 1.7% respectively. Although both models modify Gaussian mixtures based on the speaking rate, HM-MW achieved higher improvement than REG. By controlling the transition probabilities, HM-TRP improved the accuracy by 1.7%. The most effective model was HM-HMM. This model improved the accuracy by 3.2% in absolute terms by controlling both the mixture weights and the transition probabilities. Similar results were obtained for the lecture task as shown in Figure 9. The improvement by HM-HMM was 2.1%.

**Fig. 8.** Word correctness and accuracy of the meeting task given speaking rate measured using true transcript.**Fig. 9.** Word correctness and accuracy of the lecture task given speaking rate measured using true transcript.

4.4. Experiments using estimated 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 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 measure was enrate which directly derived speaking rate from speech signal by signal processing without using transcription. Window width for enrate calculation was set at 400ms based on our preliminary experiments. When rescoring, acoustic models trained with enrate were used.

Tables 2 and 3 show the results for the meeting and lecture tasks, respectively. In the table, **HYP** indicates the results using the speaking rate from one-best hypothesis, and **ENRATE** indicates the results using the enrate measure. The results by the baseline model without using the speaking rate information indicated by **BASE** and those by using the true speaking rate indicated by **ORACLE** are also shown. The cardinality of **Mode** was set to three and four.

As can be seen in Table 2, 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 HMM is vulnerable to the decrease of the quality of the speaking rate. Because the one-best hypothesis includes recognition errors, **HYP** is not an ac-

curate 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 3 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.

Table 2. 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 3. 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

5. CONCLUSIONS

This paper has explored several Dynamic Bayesian Network based acoustic models for improving recognition accuracy of spontaneous speech using the explicitly modeled effect of the speaking rate.

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 a conventional HMM, and REG, which encodes a 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 the oracle transcription. 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 estimated speaking rate, HM-HMM showed consistent 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.

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. Future works include developing 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.

6. REFERENCES

- [1] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, S. Roweis, K. Ross, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld, "Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode," in *Proc. ICSLP*, 1996.
- [2] A. Tuerk and S. J. Young, "Indicator variable dependent output probability modelling via continuous posterior functions," in *Proc. ICASSP*, 2001, vol. 1, pp. 473–476.
- [3] G. G. Zweig, *Speech recognition with dynamic Bayesian networks*, Ph.D. thesis, University of California, Berkeley, 1998.
- [4] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden Markov model," in *Proc. ICASSP*, 2001, vol. 1, pp. 513–516.
- [5] T. Stephenson, M. Magimai-Doss, and H. Bourlard, "Speech recognition of spontaneous, noisy speech using auxiliary information in Bayesian networks," in *Proc. ICASSP*, 2003, vol. 1, pp. 20–23.
- [6] J. Bilmes, "Buried markov models for speech recognition," in *Proc. ICASSP*, 1999, vol. 2, pp. 713–716.
- [7] T. Shinozaki and S. Furui, "Time adjustable mixture weights for speaking rate fluctuation," in *Proc. EUROSPEECH*, 2003, vol. 2, pp. 973–976.
- [8] N. Morgan, E. Fosler, and N. Mirghafori, "Speech recognition using on-line estimation of speaking rate," in *Proc. Eurospeech*, 1997, vol. 4, pp. 2079–2082.
- [9] S. Kitazawa, H. Ichikawa, S. Kobayashi, and Y. Nishinuma, "Extraction and representation of rhythmic components of spontaneous speech," in *Proc. Eurospeech*, 1997, vol. 2, pp. 641–644.
- [10] N. Morgan and E. Fosler, "Combining multiple estimators of speaking rate," in *Proc. ICASSP*, 1998, vol. 2, pp. 729–732.
- [11] T. Pfau and G. Ruske, "Estimating the speaking rate by vowel detection," in *Proc. ICASSP*, 1998, vol. 2, pp. 945–948.
- [12] N. Morgan, D. Baron, J. Edwards, D. Ellis, D. Gelbart, A. Janin, T. Pfau, E. Shriberg, and A. Stolcke, "The meeting project at ICSI," in *Proc. Human Language Technology Conference*, 2001, pp. 246–252.
- [13] S. Furui, K. Maekawa, H. Isahara, T. Shinozaki, and T. Ohdaira, "Toward the realization of spontaneous speech recognition," in *Proc. ICSLP*, 2000, vol. 3, pp. 518–521.
- [14] J. Bilmes and G. Zweig, "The graphical models toolkit: An open source software system for speech and time-series processing," in *Proc. ICASSP*, 2002, vol. 4, pp. 3916–3919.