Time Adjustable Mixture Weights for 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 serious problems in spontaneous speech recognition is the degradation of recognition accuracy due to the speaking rate fluctuation in an utterance. This paper proposes a method for adjusting mixture weights of an HMM frame by frame depending on the local speaking rate. The proposed method is implemented using the Bayesian network framework. A hidden variable representing the variation of the "mode" of the speaking rate is introduced and its value controls the mixture weights of Gaussian mixtures. Model training and maximum probability assignment of the variables are conducted using the EM/GEM and inference algorithms for Bayesian networks. The Bayesian network is used to rescore the acoustic likelihood of the hypotheses in N-best lists. Experimental results show that the proposed method improves word accuracy by 1.6% for the absolute value on meeting speech given the speaking rate information, whereas improvement by a regression HMM is less significant.

1. Introduction

Conventional HMM-based recognizers suffer a lower recognition rate for spontaneous speech. One of the significant factors reducing the recognition rate is the variable nature of spontaneous utterances [1]. For example, the speaking rate can change even within one utterance. A possible strategy to manage this problem is first estimating the speaking rate and then adjusting a recognizer based on the speaking rate. This paper investigates how to use the speaking rate information to control probability density functions in the HMM.

Since estimating the speaking rate is itself a difficult problem and assuming an accurate detection is unrealistic, the method of adjusting the decoder according to the speaking rate should be probabilistic. This paper proposes a method of adjusting mixture weights of the HMM at every frame based on the speaking rate. A hidden variable that represents a "mode" of the speaking rate controls the mixture weights. The proposed method can model complex changes of acoustic characteristics with only a small increase of the number of parameters.

The proposed method is implemented using the Bayesian network framework to realize detailed control of HMM parameters [2]. We use GMTK [3] for model parameter training using the EM/GEM algorithms and for decoding. The network accepts the speaking rate in addition to the usual acoustic features. The proposed method is used to rescore the acoustic likelihood of N-best hypotheses.

This paper is organized as follows. In Section 2, the proposed method is formulated as a Bayesian network. In Section 3, an experimental set up is described. Training processes of the Bayesian network are described in Section 4, and obtained parameters related to the speaking rate are analyzed in Section 5. The proposed method is applied to meeting speech recognition in Section 6 and the results are discussed in Section 7. It is shown that the proposed method is more effective in improving the recognition rate than a regression HMM given speaking rate information. Finally, in Section 8, the paper is concluded.

2. Recognition models

In this section, a way of formulating conventional HMM as a Bayesian network is reviewed and a baseline network encoding the HMM is defined. Then the proposed method and a regression HMM are described.

2.1. Baseline model

Figure 1 shows an example of a phone HMM set modeling phones [a] and [b]. Each phone model consists of three states. Figure 2 shows the corresponding Bayesian network. Note that, in the figure, nodes to encode state transition are omitted. In the phone HMM set, a probability density function for acoustic feature vectors is determined by a phone index and the state index of the phone. The Bayesian network has a node **phone** that represents a phone index and **phoneState** that represents a state index of the phone. Each node in a Bayesian network represents a random variable. In Figure 2, node **obs**, which corresponds to acoustic observations, has only incoming arrows from the nodes **phone** and **phoneState**. This means that a density function of **obs** is determined by these values. In this example, cardinalities of the discrete random variables **phone** and **phoneState** are two and three, respectively, corresponding to the number of phones and the maximum number of states for each phone. A set of diagonal covariance Gaussian mixtures are used for **obs**. In addition to these variables, a network used as the baseline model has additional nodes for representing phone sequences to rescore acoustic likelihood of the hypotheses in the N-best lists. Hereafter, the baseline network is referred to as **BASE**.

The network shown in Figure 2 represents a "time slice" of a dynamic Bayesian network (DBN). When this is used for training and decoding, it is "unrolled" for the frames of input speech feature vectors. The decoding is performed by assigning values for all the hidden variables so as to maximize the likelihood of the entire network.

2.2. Time adjustable mixture weight model

Figure 3 shows a Bayesian network in which the acoustic observation node **obs** has different density functions according to the "mode" of the speaking rate. In this network, two nodes are added to **BASE**: **SRmode** and **SRobs**. **SRmode** is a discrete hidden random variable that represents a "mode" of the speaking rate. **SRobs** is a one-dimensional continuous random

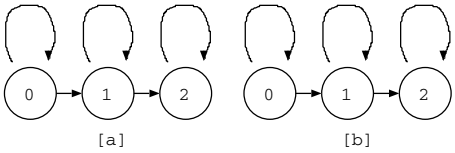

Figure 1: A phone HMM set.

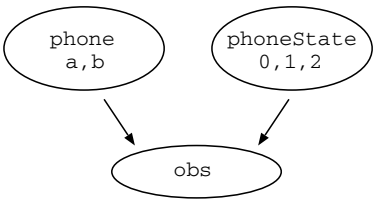

Figure 2: Bayesian network representation of the HMM.

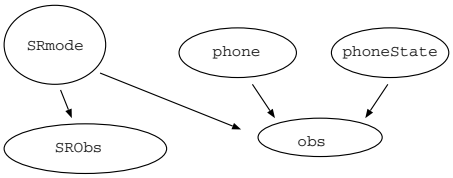

Figure 3: Speaking rate dependent model.

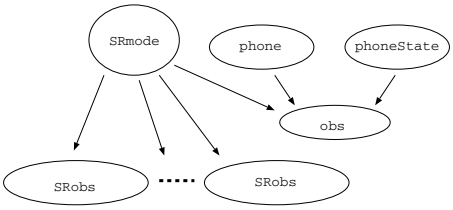

Figure 4: Node multiplication.

variable that accepts the speaking rate as an input. A Gaussian distribution is used to model the density function for each value of **SRmode** at the node. In this configuration, both the acoustic observation node **obs** and the speaking rate observation node **SRobs** have the node **SRmode** as their parent. In this study, only one pair of **SRmode** and **SRobs** is defined and commonly used for all the phone states in the phone set.

The node **obs** has a different Gaussian mixture for each combination of the values of **phone**, **phoneState**, and **SRmode**. This means that **obs** has $|SRmode|$ times more Gaussian mixtures than **BASE**, where $|SRmode|$ is the cardinality of **SRmode**. To reduce the number of parameters for accurate model estimation, the Gaussian components are tied for the different values of **SRmode**. That is, a different value of **SRmode** specifies different Gaussian mixture weights of the same Gaussian components.

SRobs is modeled to have different distributions of the speaking rate depending on **SRmode**, and this is used to detect a mode of the speaking rate. The probability density functions of **obs** are modified based on the value of **SRmode** by choosing different Gaussian mixture weights. 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 that adjusts the mixture weights for each time frame by using the hidden mode variable is referred to as **TAMW**.

Newly introduced parameters in addition to those used in **BASE** are: one conditional probability table (CPT) of size $|SRmode|$ for **SRmode**, one-dimensional Gaussian distributions for each value of **SRmode** for **SRobs**, and $|SRmode| - 1$ mixture weight vectors for each combination of the values of **phone** and **phoneState**. Note that this configuration is applicable not only to the speaking rate but also to any temporal fluctuation that affects speech features.

With the intention of increasing the influence of **SRobs** to the entire likelihood in the network, a slight modification is made, in which **SRobs** node is duplicated as shown in Figure 4. The model parameters of the duplicated nodes are completely shared and therefore the number of parameters is not increased. A duplication factor of three is experimentally chosen.

2.3. Regression HMM

Another possible way of controlling acoustic observation density functions is to use a regression model [4], in which mean values of the Gaussian components are modeled by linear combination of explanation variables. In this paper, a speaking rate and it's second and third order terms are used as the explanation variables. The parameters which have been added to the **BASE** model are regression coefficient matrix components. The matrices are tied among Gaussian components in each phone. The matrices have the same row dimension as the mean vectors and a column dimension of three. This model is also represented as a Bayesian network as shown in Figure 5. In the figure, an arrow directly connects **SRobs** and **obs**. This model is hereafter called **REG**.

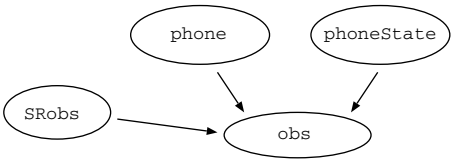

Figure 5: Regression model.

3. Experimental set up

Utterances produced by one male speaker during meetings, extracted from the Meeting Recorder Project [5], have been used for training and evaluating the Bayesian networks. The networks were trained as speaker-dependent acoustic models. Utterances in nine meetings were used for training, and those in a remaining one meeting were used for testing. Lengths of the utterances for training and testing were approximately 97 minutes and 10 minutes, respectively. They were recorded using close talking microphones. The test set consisted of 2,376 words.

By using a monophone HMM and a bigram language model, 50 best hypotheses were generated for each test set utterance. The monophone HMM consisted of 45 phones and

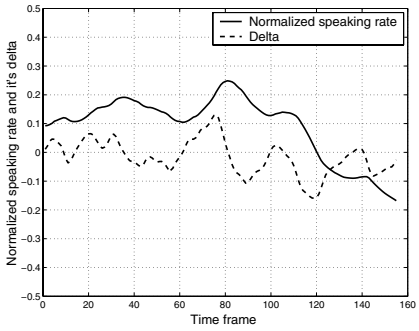

Figure 6: An example of the observed speaking rate.

each phone was modeled by a three state HMM with left-to-right topology having 64 mixtures in each state. The HMM was made using the training set and HTK. The number of mixtures was determined by a preliminary experiment so as to maximize the recognition rate. Acoustic feature vectors have 39 elements consisting of 13 MFCCs including the 0-th term, their deltas and delta deltas. The frame shift was 10 msec. The bigram language model was trained on the Switchboard corpus. The vocabulary size is 30k.

The speaking rate information used by **TAMW** and **REG** is derived from the forced alignment of correct phone state sequences with the utterance. The speaking rate is defined as an inverse value of the state holding time. The observed values are smoothed using Equation (1) and normalized as shown in Equation (2) by subtracting a mean value in the training set. In the equation, $SR_I(t)$ and $SR_S(t)$ indicate time series of the speaking rate before and after smoothing. $SR_N(t)$ is the normalized speaking rate.

$$SR_S(t) = \sum_{s=-20}^{20} SR_I(t+s) \cdot (20 - |s|). \quad (1)$$

$$SR_N(t) = SR_S - MEAN(SR_S). \quad (2)$$

In this paper, true speaking rate information is also used in the evaluation of the recognition rate to investigate an upper bound of the improvement by controlling the recognition systems based on the speaking rate. The speaking rate information for the test set is normalized using the mean value calculated for the training set.

Figure 6 shows an example of the speaking rate for the 1.6 second speech segment. The horizontal axis indicates time frames with a frame shift of 10 msec. The vertical axis is the normalized speaking rate. In the figure, a time function of the delta value of the speaking rate defined by Equation (3) is also shown.

$$SR_D(t) = \sum_{s=1}^{2} s \cdot (SR_N(t+s) - SR_N(t-s)). \quad (3)$$

4. Model training

Model parameters of the Bayesian networks are initialized using the monophone HMM used for the N-best hypotheses generation. Since the parameters of **SRmode** and **SRobs** do not have corresponding values in the HMM, they are initialized with arbitrary values. The cardinality of **SRmode** is set at four. The corresponding four mixture weights are initialized by copying the mixture weights of the monophone HMM. Regression coefficient matrices for **REG** are initialized by giving zeros to all the elements. Since the speaking rate is normalized so that the mean value is 0 using the training set, it is reasonable to initialize the Gaussian components of **REG** with those of the monophone HMM.

After the initialization, the parameters of the Bayesian networks are trained by the EM/GEM algorithms using GMTK with five iterations. During the training, the variances of the Gaussian components in the acoustic observation nodes **obs** of the networks are kept constant because of the limited size of the training data. All other parameters, including that of **SRmode**, **SRobs** and the regression coefficient matrices, are trained.

5. Acquired speaking rate mode

After the training, the model **TAMW** is expected to attain effective speaking rate mode for controlling the mixture weights. Figure 7 shows the four one-dimensional Gaussian distributions of **SRmode** corresponding to each value of the **SRmode**. In the figure, the distributions are weighted by their a priori probabilities as shown in Equation (4),

$$f_i(x) = P(SRmode = i) \cdot N_i(x), \quad i = 0, 1, 2, 3 \quad (4)$$

where $N_i(x)$ denotes a Gaussian distribution specified by **SRmode**.

Figure 8 shows the histogram indicating the distribution of the speaking rate. The lower boundary of the samples corresponds to the zero speaking rate before the mean normalization. By comparing these two figures, it is observed that the four Gaussian distributions are almost uniformly spread across the region where the samples of the speaking rate exist. This distribution is obtained by maximizing the entire likelihood of the network taking the effect of the mixture weight selection into account. From this point of view, our definition of the speaking rate seems to be successful for controlling the mixture weights.

6. Recognition results

Three Bayesian networks, **BASE**, **TAMW**, and **REG**, were used to rescore the acoustic likelihood of the N-best hypotheses. Figure 9 shows the word correctness and the accuracy. In the figure, **HMM** denotes the result of the best hypothesis before the rescoring. **BASE** and the monophone HMM used for the N-best generation are basically the same except that **BASE** is implemented as a Bayesian network and trained five times more than the HMM. Accordingly, recognition rates of these two models are almost the same.

By comparing **BASE**, **TAMW**, and **REG**, it can be seen that **TAMW** improves word correctness by 2.1% and word accuracy by 1.6%, given the speaking rate information, whereas the improvements by **REG** using the same information are only 0.2% and 0.1%, respectively.

As an additional experiment, delta values of the speaking rate were also used for **TAMW** in addition to the instantaneous speaking rate. Since the dimension of the continuous random variable **SRobs** was doubled to two, two dimensional Gaussian distributions with diagonal covariances were used for **SRobs**. The network structure of **TAMW** was invariant in this change. Table 1 shows the recognition results. In the table, TAMW(SR) denotes the result using only the speaking rate and TAMW(SR_D) denotes the result using both speaking rate and

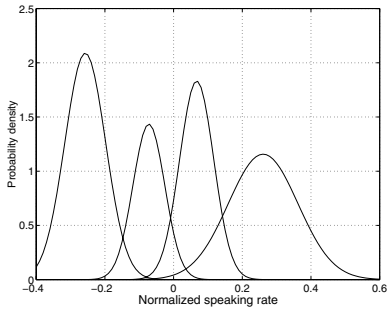

Figure 7: Gaussian distributions for the variation of the speaking rate mode.

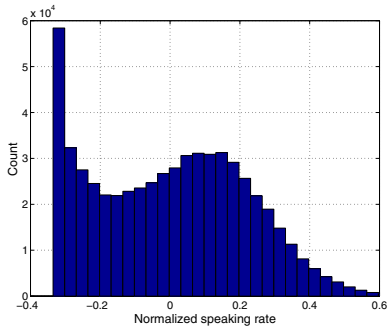

Figure 8: Histogram of the speaking rate.

Table 1: Recognition results

Speaking Rate	TAMW(SR)	TAMW(SR_D)
Word correctness	58.0	57.7
Word accuracy	54.3	54.0

its deltas. No additional improvement is observed in this condition.

7. Discussion

Our proposed model **TAMW** has achieved larger improvement than **REG**. One disadvantage of **REG** might be that it deterministically changes the mean values of the Gaussian components. Even though 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 before. Besides, the relationship between the speaking rate and the change of speech spectra might be essentially probabilistic. **TAMW**, 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 is obtained by introducing the hidden variable **SRmode** for representing the mode of speaking rate.

8. Conclusions

This paper has proposed a new method for improving recognition accuracy of spontaneous speech by modeling the effect of

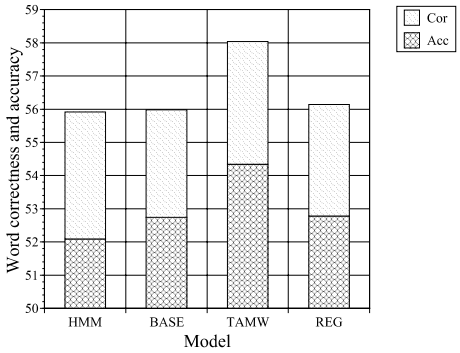

Figure 9: Recognition results.

speaking rate on speech spectra. The proposed method, implemented as a Bayesian network, controls mixture weights at each frame based on a value of a hidden variable that represents the "mode" of the speaking rate. A proposed model **TAMW** was compared with a regression model **REG**, and it was shown that the improvement of the recognition accuracy by **TAMW** was much larger than **REG**. **TAMW** achieved 2.1% and 1.6% improvement in the absolute values of the word correctness and accuracy, respectively, in comparison with the baseline model **BASE**. The use of the delta speaking rate in **TAMW** did not improve the recognition rate.

Future work includes evaluation of the proposed method using speaking rate information estimated without using the true phone state sequences, development of more efficient ways of utilizing speaking rate information, and combination with other spontaneous speech features to further improve recognition accuracy.

9. Acknowledgements

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

10. References

[1] T. Shinozaki and S. Furui, "Error analysis using decision trees in spontaneous presentation speech recognition," *Proc. ASRU*, Madonna di Campiglio, a01ts039, 2001.

[2] J. Bilmes, "Graphical models and automatic speech recognition," Technical Report UWEETR-2001-005, University of Washington, Dept. of EE, 2001.

[3] J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," *Proc. ICASSP*, Orlando, Florida, vol.4, pp. 3916–3919, 2002.

[4] K. Fujinaga, M. Nakai, H. Shimodaira, and S. Sagayama, "Multiple-regression hidden markov model," *Proc. ICASSP*, Salt Lake City, Utah, vol.1, pp. 513–516, May 2001.

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