

# Automatic Generation of Non-Uniform Context-Dependent HMM Topologies Based on The MDL Criterion

Takatoshi Jitsuhiro, Tomoko Matsui<sup>†</sup>, Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories,  
2-2-2 Hikaridai, “Keihanna Science City”, Kyoto 619-0288, Japan.

{takatoshi.jitsuhiro, tomoko.matsui, satoshi.nakamura}@atr.co.jp

## Abstract

We propose a new method of automatically creating non-uniform context-dependent HMM topologies by using the Minimum Description Length (MDL) criterion. Phonetic decision tree clustering is widely used, based on the Maximum Likelihood (ML) criterion, and creates only contextual variations. However, it also needs to empirically predetermine control parameters for use as stop criteria, for example, the total number of states. Furthermore, it cannot create topologies with various state lengths automatically. Therefore, we introduce the MDL criterion as split and stop criteria, and use the Successive State Splitting (SSS) algorithm as a method of generating contextual and temporal variations. This proposed method, the MDL-SSS, can automatically create proper topologies without such predetermined parameters. Experimental results show that the MDL-SSS can automatically stop splitting and obtain more appropriate HMM topologies than the original one. Furthermore, we investigated the MDL-SSS combined with phonetic decision tree clustering, and this method can automatically obtain the best performance with any heuristic.

## 1. Introduction

Phonetic decision tree clustering[1] was proposed as a method of generating tied-state structures of acoustic models for speech recognition. Methods based on phonetic decision tree clustering originally used the Maximum Likelihood (ML) criterion to choose the phonetic question with which each state was split. However, owing to the nature of the ML estimation, the likelihood value for training data increases as the number of parameters increases. Consequently, it is impossible to stop splitting with only the ML criterion. The methods based on the ML criterion require heuristic stop criteria, such as the total number of states. Recently, information criteria such as the Minimum Description Length (MDL) have been introduced as splitting and stop criteria in context-dependent Hidden Markov Model (HMM) creation using phonetic decision tree clustering[2]. These methods continue to split states so as to improve the information criteria. The Successive State Splitting (SSS) algorithm was originally proposed to create a network of HMM states of speaker dependent models[3] and was subsequently expanded to the ML-SSS algorithm for speaker independent models[4]. The ML-SSS algorithm has the same problem as phonetic decision tree clustering in that it requires the total number of states as the stop criterion. The ML-SSS algorithm is a bottom-up approach that conducts both contextual clustering and temporal splitting. The maximum number

of temporal states for each phoneme model should be given as another stop criterion for temporal splitting. It is difficult to properly preset these two stop criteria.

We propose an HMM-topology design method using the ML-SSS algorithm in conjunction with the MDL criterion as the splitting and stop criteria. We call the new method the MDL-SSS algorithm. The MDL criterion was successfully introduced in phonetic decision clustering as the criterion for contextual clustering[2]. This paper extensively uses the MDL criterion as the criterion for both contextual and temporal splitting in the ML-SSS algorithm. We define new gain functions based on the MDL criterion and introduce two scaling factors. Furthermore, we investigate the MDL criterion for the methods using both data-driven and phonetic decision tree clustering.

In Section 2, we explain the ML-SSS algorithm and the stop-splitting problem. The MDL criterion is described in Section 3. We define the MDL-SSS algorithm in Section 4. In Section 5, we evaluate the performance of the MDL-SSS and describe the results using the ATR travel arrangement task and short results for lecture speech as more spontaneous speech. Additionally, phonetic decision tree clustering is introduced into the MDL-SSS in Section 6. We summarize our findings in Section 7.

## 2. ML-SSS Algorithm

### 2.1. Problems of ML-SSS

The ML-SSS algorithm has contextual and temporal splitting[4]. First, both the contextual and temporal splitting are performed for all states. Second, the gains of both contextual and temporal splitting are calculated. Finally, these expected gains are compared with each other and the state with the best gain among all states is selected.

The ML-SSS needs the total number of states,  $N_s$ , and the maximum length of state sequences for phoneme models,  $N_p$ . These parameters must be given before starting the splitting. For temporal splitting, the ML-SSS creates one more state and connects it to the original state. The parameters of the two distributions are estimated by the forward-backward algorithm, and the total expected gain of the temporal splitting is also calculated for the temporal split states. Since it is costly to re-estimate the parameters of all states at every splitting, only the parameters for the two candidate states are re-estimated by using probabilities weighted by the statistics of the target state. Since the likelihood value of the temporal split states is an approximate value, it is difficult to use it as a stop criterion. Thereby,  $N_p$  is needed as a stop criterion. However, to find the optimal values of these parameters,  $N_s$  and  $N_p$ , is generally difficult. Experiments should be conducted to find the optimal values.

<sup>†</sup>currently with the Institute of Statistical Mathematics, 4-6-7 Minami-Azabu, Minato-ku, Tokyo 106-8569, Japan.  
e-mail: tmatsui@ism.ac.jp

## 2.2. Gain function by ML-SSS algorithm

We formulate the total expected gains of the splitting states with the ML-SSS algorithm. The total expected gain  $G_{output}(S_i)$  of the output probabilities and  $G_{trans}(S_i)$  of the transition probabilities for state  $S_i$  are

$$G_{output}(S_i) = -\frac{1}{2} \{ \Gamma(S_{i_1}) \log |\Sigma(S_{i_1})| + \Gamma(S_{i_2}) \log |\Sigma(S_{i_2})| - \Gamma(S_i) \log |\Sigma(S_i)| \}, \quad (1)$$

$$G_{trans}(S_i) = \Xi(S_{i_1}, S_{i_1}) \log a_{i_1 i_1} + \{ \Gamma(S_{i_1}) - \Xi(S_{i_1}, S_{i_1}) \} \log(1 - a_{i_1 i_1}) + \Xi(S_{i_2}, S_{i_2}) \log a_{i_2 i_2} - \Xi(S_i, S_i) \log a_{ii}, \quad (2)$$

where  $\Gamma(S_i) = \sum_t \gamma_t(S_i)$  is the expected frequency of the transition from state  $S_i$ .  $\gamma_t(S_i)$  is the probability of staying in  $S_i$  at time  $t$ .  $\Xi(S_i, S_j) = \sum_t \xi_t(S_i, S_j)$  is the expected frequency of the transition from  $S_i$  to  $S_j$ .  $\xi_t(S_i, S_j)$  is the probability of the transition from  $S_i$  to  $S_j$  at  $t$ .  $a_{ii}$  is the self-loop probability.

For contextual splitting, since the transition probabilities are not re-estimated to reduce the amount of calculation, the total expected gain related only to the observation distributions is calculated. For temporal splitting, the transition probabilities are considered because one transition probability is created after temporal splitting. The splitting conditions  $G_c^{(MDL)}(S_i)$  for contextual splitting and  $G_t^{(MDL)}(S_i)$  for temporal splitting are

$$G_c^{(MDL)}(S_i) = G_{output}(S_i), \quad (3)$$

$$G_t^{(MDL)}(S_i) = G_{output}(S_i) + G_{trans}(S_i). \quad (4)$$

Equations (3) and (4) are calculated for each state, and the state with the maximum gain is selected.

## 3. MDL Criterion

The MDL criterion[5] is one of the most popular information criteria and is used for the selection of the optimal model for stochastic models. Generally, when a set of models  $\{\theta^{(i)} | i = 1, \dots, I\}$  is given, the MDL criterion for model  $i$  is

$$L_i(\mathbf{x}) = -\log P(\mathbf{x} | \hat{\theta}^{(i)}) + \frac{\alpha_i}{2} \log N_T + \log I, \quad (5)$$

where  $\mathbf{x} = \{x_1, \dots, x_{N_T}\}$  is the observation data,  $\alpha_i$  is the number of free parameters, and  $\hat{\theta}^{(i)}$  is the ML estimates of model  $i$ .

## 4. SSS Algorithm Using The MDL Criterion

Figure 1 shows the flow chart of the MDL-SSS. The differences of the MDL values for both contextual and temporal splitting are calculated for each state, and the state with the minimum difference value is selected as the split state. Splitting is finished when there is no state that can be split and reduce the criterion by splitting.

We define the criteria for contextual splitting and temporal splitting,  $G_c^{(MDL)}$  and  $G_t^{(MDL)}$ , respectively, as follows:

$$G_c^{(MDL)}(S_i) = -G_c^{(ML)}(S_i) + C_c \frac{\alpha'_c - \alpha_c}{2} \log \Gamma(S), \quad (6)$$

$$G_t^{(MDL)}(S_i) = -G_t^{(ML)}(S_i) + C_t \left\{ \frac{\alpha'_t}{2} \log \Gamma'(S) - \frac{\alpha_t}{2} \log \Gamma(S) \right\}. \quad (7)$$

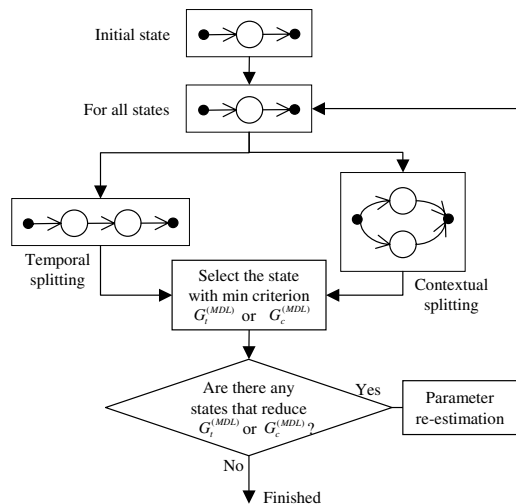


Figure 1: Flow chart of MDL-SSS.

The first terms on the right-hand sides are the negative values of the expected gains in the ML-SSS.  $C_c$  and  $C_t$  are the scaling factors of the second terms, which are not derived from the original definition of Eq. (5). These will be explained later.

$\Gamma(S) = \sum_i \Gamma(S_i)$  represents the expected frequency of the number of samples for all states.  $\Gamma'(S)$  is the value after temporal splitting. Equation (7) compensates the total number of samples because segments that are shorter than the lengths of the state sequences are discarded.  $\alpha_c, \alpha'_c$  are the number of parameters before and after contextual splitting, respectively.  $\alpha_c = 2KM$  and  $\alpha'_c = 2K(M+1)$ , when the order of features is  $K$ , the total number of states is  $M$ , and each state has one Gaussian distribution with a diagonal covariance matrix. For temporal splitting, we suppose that transition probabilities do not depend on both mean vectors and covariances of the Gaussian mixtures. Each state has one Gaussian distribution and one transition probability. Therefore, the number of parameters before and after temporal splitting are  $\alpha_t = (2K+1)M$  and  $\alpha'_t = (2K+1)(M+1)$ , respectively.

The scaling factors,  $C_c$  and  $C_t$ , are not derived from the original MDL criterion. We experimentally found that it is difficult to stop splitting without these factors. This problem appears to be caused by the approximation of the likelihood values of the temporal split states as described in Section 2. In [2], a scaling factor for contextual splitting was also introduced and experimentally found to be effective. However, these factors are expected to be the same regardless of training data. This is confirmed through the experiments. The MDL-SSS algorithm selects the state with the smallest  $G_c^{(MDL)}$  or  $G_t^{(MDL)}$ , and stops splitting when  $G_c^{(MDL)} > 0$  and  $G_t^{(MDL)} > 0$  for all states.

## 5. Experiments

### 5.1. Conditions

For the acoustic training set, we used dialog speech (5 hours in total) from the ATR travel arrangement task (TRA) database and read speech (25 hours) of phonetically balanced sentences (BLA). The same 407 speakers uttered both spontaneous and read speech. For testing, we used dialog speech from the TRA database uttered by a different set of 42 speakers. The sampling frequency was 16 kHz, the frame length was 20 ms, and the frame shift was 10 ms. 12-order MFCC,  $\Delta$ MFCC, and  $\Delta$

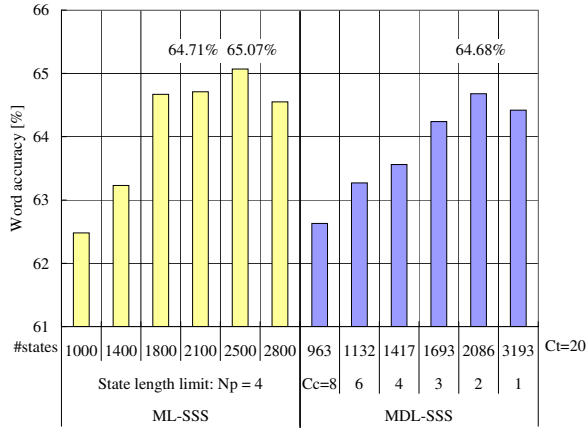


Figure 2: Word accuracy for models with one Gaussian distribution per state.

log power were used as feature parameters. The cepstrum mean subtraction was applied for each utterance. We used 26 kinds of phonemes and one silence. Three states were used as the initial model for each phoneme. One Gaussian distribution for each state was used during topology training. A silence model with three states was built separately from the phoneme models. In Section 5.2, we used speaker independent models with one Gaussian distribution per state. These models could not always produce high performance. Therefore, after we obtained the topology, we increased the number of mixtures and re-estimated the parameters of the HMMs. The final models were gender-dependent models with five Gaussian mixtures for each state. These models were used in the experiments described after Section 5.3. For the language training set, we used 7,195 one-side dialogues that included  $1.6 \times 10^6$  words. Multi-class composite bigram models [6] were used. The full vocabulary size in the set was 27,398.

## 5.2. Comparison of models with one distribution per state

We initially investigated the performance by models with one Gaussian distribution for each state to confirm the adequacy of the model topologies obtained by our proposed method. In this investigation, the lexicon had only 5,100 words including the words in the evaluation data. Figure 2 shows the word accuracy rates by the ML-SSS and MDL-SSS. The MDL-SSS with  $C_c = 2$  and  $C_t = 20$  obtained almost the same performance as the ML-SSS. For the MDL-SSS,  $C_c = 2$  and  $C_t = 20$  performed the best and, for the ML-SSS,  $N_s = 2500$  and  $N_p = 4$  showed the best performance.

## 5.3. Comparison of models with five mixtures per state

Next, we used gender-dependent models with five Gaussian mixtures for each state. Figure 3 shows the word accuracy rates of these models. For the MDL-SSS,  $C_c = 2$  and  $C_t = 20$ , the same values as in the previous section, performed the best. For the ML-SSS,  $N_s = 1400$  and  $N_p = 4$  performed the best. Therefore, for the ML-SSS,  $N_s$  should be carefully adjusted according to the experiments to find the best model.

Figure 4 shows the maximum path length for each phoneme model extracted from both the “ML-SSS (1400 states)”, whose paths were set to a limit of four states, and the “MDL-SSS ( $C_c = 6$ ,  $C_t = 20$ , 1132 states).” All the phoneme models by the ML-SSS had the same maximum path length as the path limit number. On the other hand, each phoneme model by the

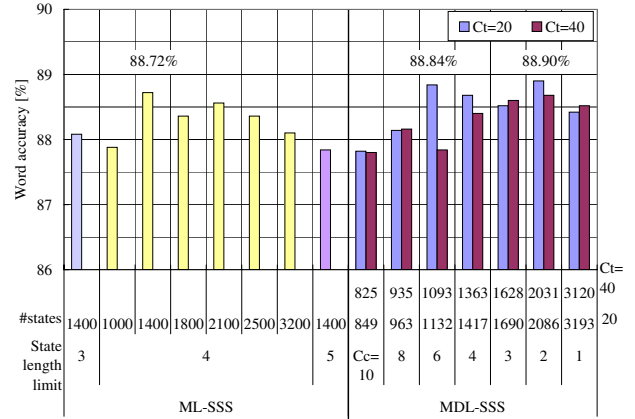


Figure 3: Word accuracy for models with five Gaussian mixtures per state trained by using TRA and BLA.

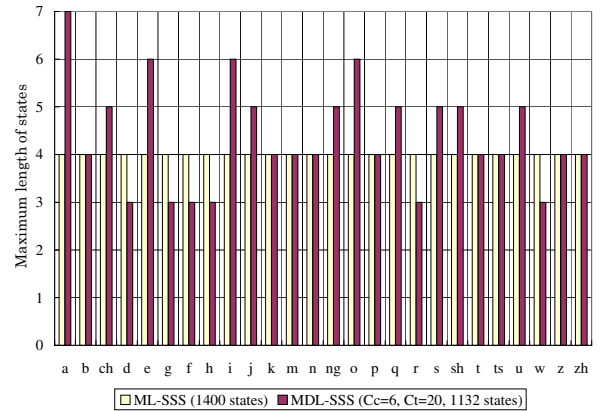


Figure 4: The maximum length of paths for each phoneme.

MDL-SSS had a different maximum path length. This suggests that more adequate path lengths are selected for each allophone by the MDL-SSS.

## 5.4. Effectiveness for different amounts of training data

To confirm whether the scaling factors in the MDL-SSS are independent of the amount of training data, we examined the performance of acoustic models for a smaller amount of training data using only the TRA data. Figure 5 shows the word accuracy for the models generated by the ML-SSS or MDL-SSS. The MDL-SSS obtained the best performance with  $C_c = 2$  and  $C_t = 20$ .

## 5.5. Evaluation using lecture speech

We also evaluated our method by using the lecture speech corpus, “The Corpus of Spontaneous Japanese (CSJ)”[7], which is more spontaneous than the TRA. The training data for the acoustic models was 200 lectures (about 34 hours) by male speakers. The analysis conditions were the same as in Section 5.1. The number of mixtures for each state was 10. We used the word bigram model distributed by Kyoto University. The size of the lexicon was 19 K words. The evaluation experiments were carried out using 253 utterances by one male, “A01M0074”. Figure 6 shows the results. The trend of the results is quite similar to that of the TRA task. This shows that the MDL-SSS can automatically stop splitting and get the best performances by  $C_c = 2$  and  $C_t = 20$  for other tasks.

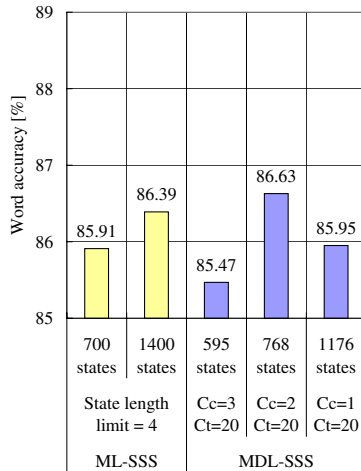


Figure 5: Word accuracy for models trained using TRA.

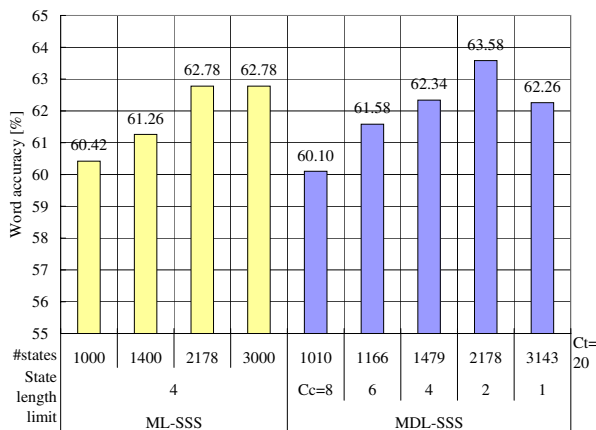


Figure 6: Word accuracy for CSJ.

## 6. MDL-SSS with Decision Tree Clustering

The ML-SSS has another problem in that it cannot deal with unseen contexts because of the data-driven clustering developed by P. A. Chou's algorithm[8]. In [9], phonetic decision tree (PDT) clustering was introduced into the ML-SSS. The authors claim that the combination of decision tree clustering and data-driven clustering is best. For contextual splitting, first, decision tree clustering is used to create two initial distributions from one distribution. Second, Chou's algorithm is used repeatedly as the data-driven clustering. We introduced the MDL criterion to this method.

The number of phonetic categories for the questions was 47. The unseen contexts in the evaluation data were 1.1%. The experimental conditions were the same as in Section 5.3. Figure 7 shows the word accuracy rates. "ML-PDT+Chou-SSS" and "MDL-PDT+Chou-SSS" respectively stand for the ML-SSS and MDL-SSS using both PDT clustering and data-driven clustering. As with the MDL-SSS, the MDL-PDT+Chou-SSS automatically obtained almost the same performance as the ML-PDT+Chou-SSS. The best  $C_c$ ,  $C_t$  of the MDL-PDT+Chou-SSS were different from the best  $C_c$ ,  $C_t$  of the MDL-SSS. This shows that these scaling factors are dependent on the splitting algorithm.

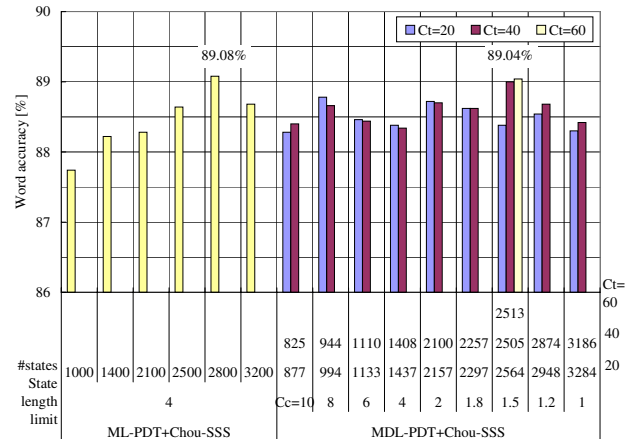


Figure 7: Word accuracy by SSS using decision tree clustering and Chou's algorithm.

## 7. Conclusion

We proposed the Successive State Splitting algorithm in conjunction with the MDL criterion. We introduced the MDL criterion to the ML-SSS algorithm in order to select suitable models automatically. The experimental results show that the MDL criterion can stop both contextual and temporal state splitting by the SSS algorithm. Although we introduced two scaling factors, the best values exist and are robust for training data. Furthermore, we investigated the MDL-SSS combined with phonetic decision tree clustering. This method obtained the best performance automatically.

## 8. Acknowledgments

This research was supported in part by the Telecommunications Advancement Organization of Japan.

## 9. References

- [1] S. J. Young, J. J. Odell and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. of the ARPA Workshop on Human Language Technology*, pp. 307–312, 1994.
- [2] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *The Journal of the Acoustical Society of Japan (E)*, vol. 21, no. 2, pp. 79–86, 2000.
- [3] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," in *Proc. ICASSP'92*, vol. 1, pp. 573–576, 1992.
- [4] M. Ostendorf and H. Singer, "HMM topology design using maximum likelihood successive state splitting," *Computer Speech and Language*, vol. 11, pp. 17–41, 1997.
- [5] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. on IT*, vol. IT-30, no. 4, pp. 629–636, 1984.
- [6] H. Yamamoto and Y. Sagisaka, "Multi-class composite n-gram based on connection direction," in *Proc. of ICASSP'99*, vol. 1, pp. 533–536, 1999.
- [7] S. Furui, K. Maekawa and H. Isahara, "Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –," in *Proc. of ICSLP2000*, vol. 3, pp. 518–521, 2000.
- [8] P. A. Chou, "Optimal partitioning for classification and regression trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, pp. 340–354, 1991.
- [9] H. Singer and A. Nakamura, "Unified framework for acoustic topology modelling: ML-SSS and question-based decision trees," in *Proc. of EUROSPEECH'99*, vol. 3, pp. 1355–1358, 1999.