TR-IT-0113

HMnet Evaluation for Phonetic Environment Variation of Training Data

Hoi-Rin Kim

1995.05

Evaluation of the SSS algorithm has usually been performed on the aspects of recognition performance and number of the free parameters to be estimated, which is related to robustness of the model. In this work, I have investigated how the variation of phonetic environments of training data affects on the HMnet generated by SSS algorithm. By analyzing the results, I think we could understand more better the SSS algorithm's behavior and limitation.

目次

表目次

図目次

1 Introduction

In HMM-based acoustic modeling, it has been well known that it is very important to appropriately compromize the degree of precision and robustness of each model for a given training data. When we model a phone (or triphone), if each phone is independently modeled, then the precision of the model would be high, but the robustness of the model become weak. This problem is basically dependent on the amount of training data used in the modeling procedure. Another more basical problem is that any training data cannot cover all phonetic environments which are able to be pronounced in the language. These problems can be solved by using definitely infinite training data, but this is also impossible in realistic sense. However, we can easily guess that the acoustic characteristics of all triphones would not be fully independent to each other, that is, would have certain dependency in some phonetic environments. In order to apply the knowledge (in a sense, assumption) to triphone modeling, we must observe acoustic relations between all phonetic combinations, but this is also a difficult job. As an alternative approach, we can use statistical methodology in obtaining the relationship and compromizing the precision and robustness. Several novel methods to obtain precise and robust phone models have been proposed such as in HMnet [1], Senone [2], and Genone [3], and they showed good performance. HMnet is similar with Senone in that they have state sharing architecture. Main difference between them is that Senone is made through a merging procedure which is controlled by only acoustic characteristics, but HMnet is generated through a splitting procedure which is controlled by both acoustic and phonetic characteristics. On the other hand, Genone has the parameter sharing architecture only in Gaussian mixtures, not in mixture weights, and is conceptually same as Senone in the merging procedure. So, Genone is more adequate than HMnet or Senone for more training data because its architecture consequently increases the degree of freedom in observation parameter sharing constraints.

Among these properties, I especially looked at the phonetic constraints in the splitting procedure in HMnet, called as SSS (successive state splitting) algorithm. The HMnet generation procedure considers the phonetic environments of the previous, present, and the following phones, not same as Senone or Genone. Therefore, I think it is necessary to observe which relations the generation procedure has with respect to variations of phonetic environments of training data. Until now, evaluation of the SSS algorithm has usually been performed on the aspects of recognition performance and number of free parameters to be estimated, which is related to robustness of the model [4]. So, in this work, I observed HMnet's characteristics on phonetic context environment variation in training data. By analyzing the results, I think we could more better understand the SSS algorithm's behavior and limitation.

In chapter 2, I will briefly review the SSS algorithm, in chapter 3, I will describe preliminary experimental results for Korean speech DB, and in chapter 4, I will evaluate the HMnet on different phonetic environments of training data. Lastly, in chapter 5, I will summarize the experimental observations.

2 Brief Review of SSS Algorithm and HMnet

We can see HMnet as a network configuration of allophonic HMMs which have phone context dependency and state sharing architecture. This HMnet is automatically generated by an unified algorithm, that is, SSS algorithm, which can simultaneously determine and estimate an optimal set of allophones, an optimal state sharing architecture, and optimal parameters on maximum likelihood criterion. The processing sequence is briefly as follows. More details are in [1] and [4].

1. Training of an initial model:

   As an initial model, an HMM consisting of one state is trained with all training data containing every phone context.

2. Determination of split state:

   For each state, the distribution size of output probability density is calculated, then the state having the largest distribution size will be split in the next step.

3. Split of the state:

   The determined state is split into two states. At this time, the algorithm examines two split domains, that is, contextual domain and temporal domain. Then, the domain which accomplishes a higher likelihood for all the training samples is selected by comparing the maximum likelihood obtained by split on each domain.

4. Re-estimation of the model parameters:

   The model parameters of all states which were affected by the state split are re-estimated. The steps from 2 to 4 are repeated until a prescribed number of total states is reached.

5. Change and final estimation of output probability density distributions.

The HMnet generated by the SSS has been evaluated in terms of modeling efficiency and recognition performance. Modeling efficiency is ratio of total number of states needed to represent all the allophone models without any state sharing to total number of states in the obtained HMnet itself. Therefore, the modeling efficiency measures the degree of state sharing, or statistical robustness, of the HMnet for the given training data and the state number of HMnet.

3 Preliminary Experiment for the SSS Algorithm with Korean Speech Data

3.1 Speech DB and experiment conditions

The speech data which I used for preliminary experiment were Korean word database which was collected in hotel reservation task domain. The vocabulary consists of 244 words which include some connected digits, 26 English alphabet pronunciations, month, week, date names, and so on. Total 40 male speakers spoke the 244 words one time. The 244 words include all Korean distinct phones except only one phone, so consequently 39 distinct phones including one silence unit.

In original, the speech data were digitized at 16 bit, 16 kHz sampling rate in ETRI, Korea. But, in order to consist with conventional acoustic analysis method in ATR, I downsampled the data to 12 kHz on digital domain. Then, the 12 kHz data were processed as follows.

- 20 msec Hamming window
- 5 msec window shift rate
- 34 dimension feature vector by LPC-based analysis:
  log power, 16 dimension cepstrum, delta log power, 16 dimension delta cepstrum

3.2 Automatic segmentation into phone unit by Viterbi alignment

In order to effectively use the SSS algorithm, phone unit samples are needed, but original word data do not have any phone boundary information. So I segmented the data by Viterbi alignment with 39 context-independent phone models which were trained by concatenative training method for all the 40 speaker data. The speaker-independent context-independent phone models were trained under the conditions as follows.

- Model topology: 4 state simple left-to-right HMM.
- Output probability distribution at each state: 5 mixture Gaussian, diagonal covariance matrix.

After training, all the data used in the training were segmented into the corresponding phone units by Viterbi alignment. I used SSS-ToolKit(Ver 3.0) [5] in all the procedures.

3.3 Context-dependent phone modeling by SSS

At first, I observed the running characteristics of the SSS algorithm with Korean speech DB to confirm the ability of the algorithm. In training mode, I used 10 speakers' data among the automatically segmented data. The training word data include total 21,268 phones. Also, I used another 2 speakers' data as test data set, including total 4,256 phones. For the SSS algorithm's parameter setting, basically I used default values of the algorithm, for example, consideration of only single left and right phone context, maximum 4 state splitting in temporal domain, one mixture Gaussian output distribution with diagonal covariance matrix at each state in determination of HMnet topology, and so on. And, I set maximum number of states splitted to 500, and in final re-estimation of model parameters each model was trained for the mixture number 1, 3, and 5 respestively.

Firstly, I investigated various phonetic characteristics of current training data in order to treat as references of observations followed. Table 1 shows sample number of each phone in training data used, and Table 2 shows phone perplexity, number of distinct triphones, and triphone entropy. Next, in training procedure using SSS, I observed various characteristics on variation of state number of HMnet as follows.

- Number of allophones (see Figure 1)
- Allophone entropy (see Figure 2)

- Modeling efficiency (see Figure 3)
- Mutual information between allophone models and training data (see Figure 4)
- Phone recognition accuracy for training data (see Figure 5)

Where allophone entropy was computed with number of training samples used in parameter estimation of the corresponding allophone model, and mutual information was obtained using the following equation so as to estimate discriminative power of each model [6].

$$I(m, y) = \log P(y \mid m) - \log \sum_{m'} P(y \mid m')P(m'), \tag{1}$$

where m is an allophone model in the HMnet corresponding to the phone speech data, y. Therefore, this value is closely related to recognition rate.

In the Figures 1 and 2, we can see that, even though the number of allophones are abruptly varied, the allophone entropy monotonously increases. This means the SSS algorithm runs so as to distribute samples uniformly to each allophone model. In the modeling efficiency curve of the Figure 3, the increased area means that states are split mainly in contextual domain, on the other hand, the decreased area means that states are split mainly in temporal domain. From this result, we can guess it is appropriate to decide the number of states at between 300 and 350 as a proper compromization of precision and robustness on this task domain. In the mutual information curve of the Figure 4, we can see that, as the mixture number of output probability distribution at each state increases or the number of states increases, the mutual information also increases. That means that the number of free parameters are related to precision of model and this consists also with the result of phone recognition shown in the Figure 5.

3.4 Result and discussion of speaker-independent phone recognition

In order to evaluate the performance of the allophone models in the HMnet generated by SSS, I used another 2 speakers' data with the same vocabulary as training data. The speaker-independent test results are shown in the Figures 6 and 7. From these figures, we can see that it is appropriate to decide the number of state to 250 at mixture number 5, 430 at mixture number 3, and 500 at mixture number 1. But, in the previous section, I described that it was appropriate to decide the number of state between 300 and 350 from the modeling efficiency curve. This does not consist with the results from the recognition test. This means that it is very difficult to decide appropriately the number of state by refering to only the modeling efficiency curve. Therefore, I think that it is necessary to device a proper measure which can decide the number of state to be able to balance the precision and robustness of the HMnet in a given condition.

表 1: Number of each phone in training data (10 speakers).

Category	Index	Phone	Samples	Category	Index	Phone	Samples
Vowel	1	a	1299	Consonant	21	b	90
	2	v	300		22	bv	430
	3	o	1499		23	bs	569
	4	u	630		24	P	0
	5	E	80		25	p	360
	6	e	230		26	s	1448
	7	U	360		27	S	440
	8	i	2818		28	j	50
Semi-vowel	9	y	580		29	jv	90
	10	w	210		30	C	30
Consonant	11	g	200		31	c	269
	12	gv	90		32	h	819
	13	gs	150		33	n	340
	14	K	140		34	ns	170
	15	k	130		35	r	300
	16	d	80		36	l	1249
	17	dv	190		37	m	110
	18	ds	10		38	ms	470
	19	T	10		39	ng	40
	20	t	110	Silence	40	-	4878

表 2: Phonetic characteristics in training data (10 speakers).

Phone perplexity	5.0
Number of distinct triphones	472
Triphone entropy	7.6

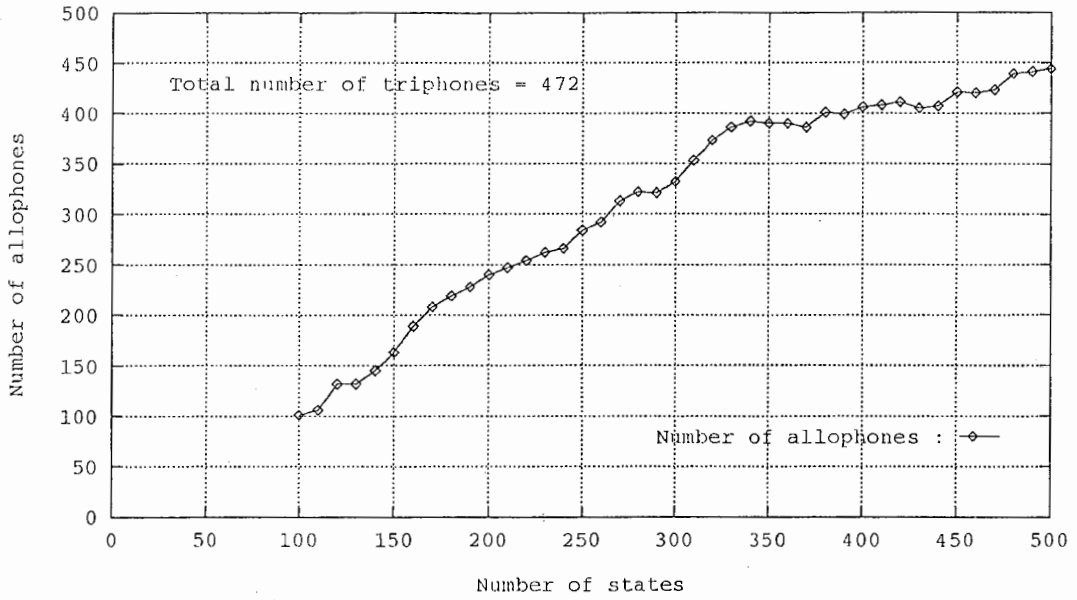

図 1: Number of allophones.

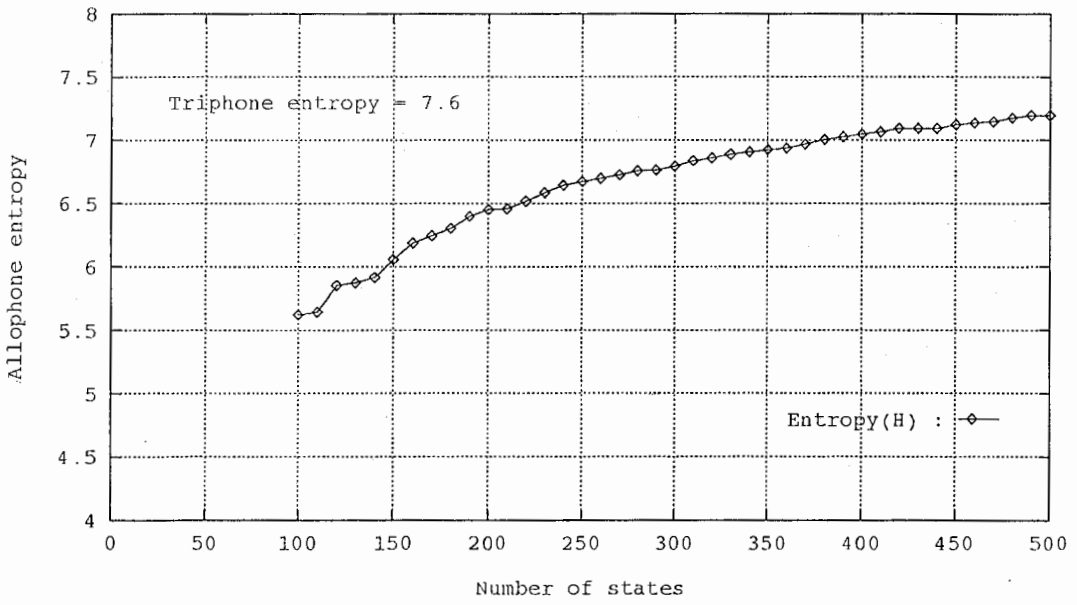

図 2: Allophone entropy.

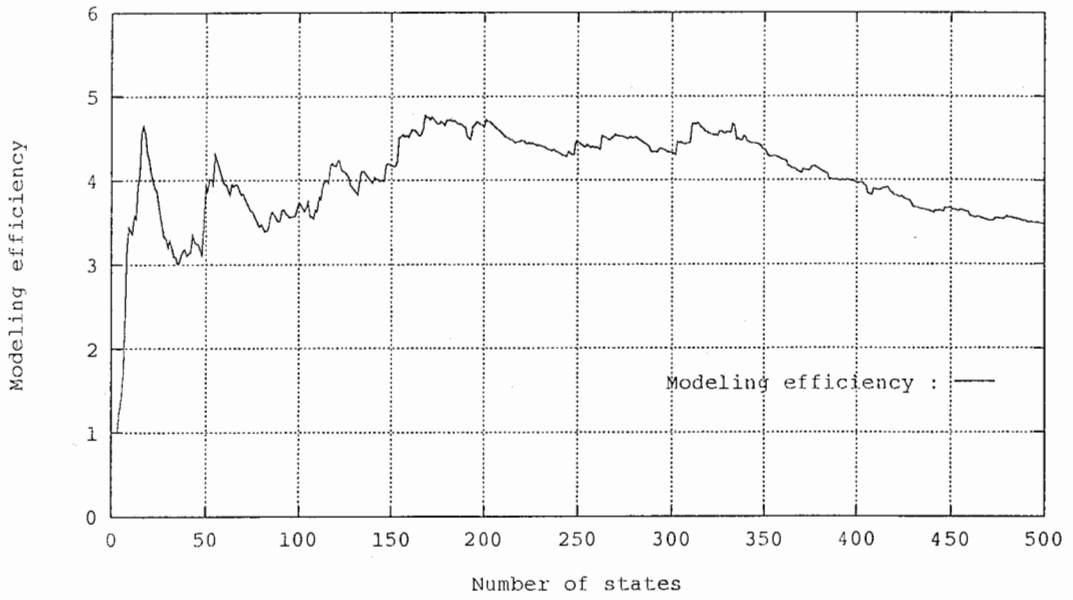

図 3: Modeling efficiency.

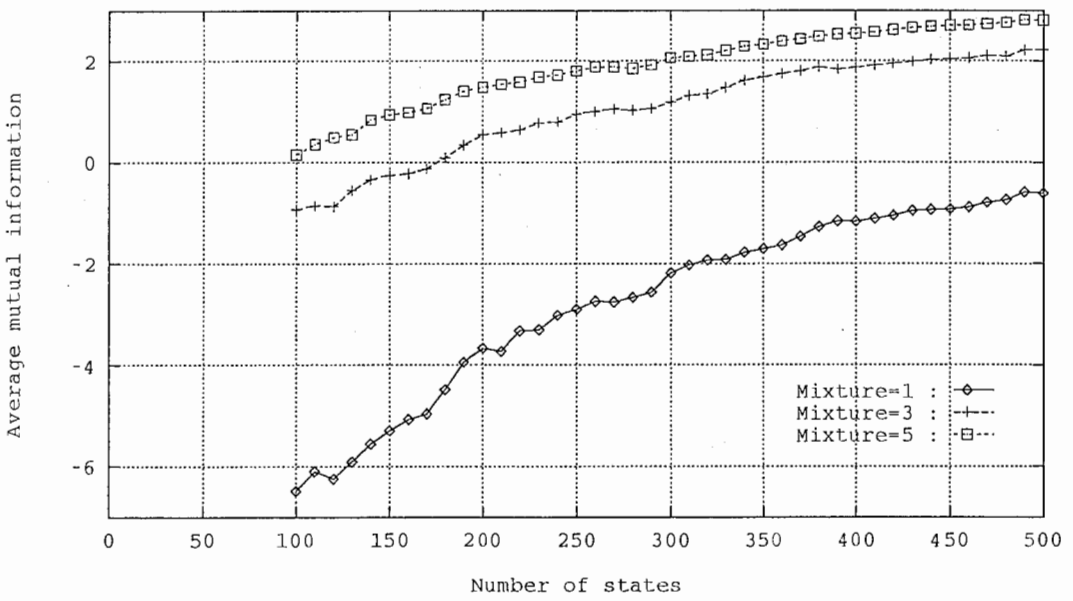

図 4: Mutual information between allophone models and training data.

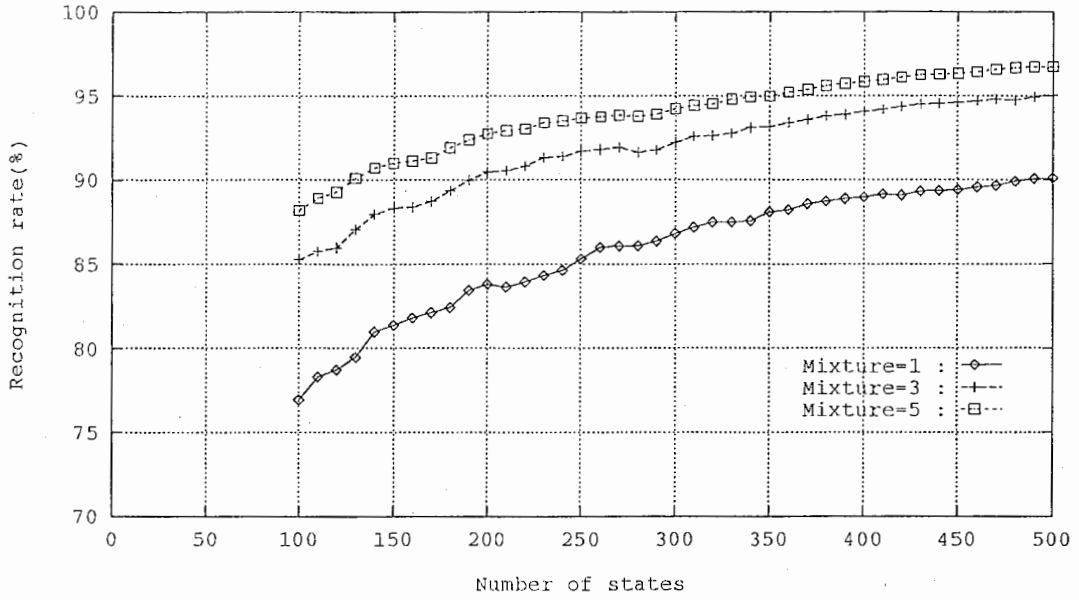

図 5: Phone recognition accuracy for training data.

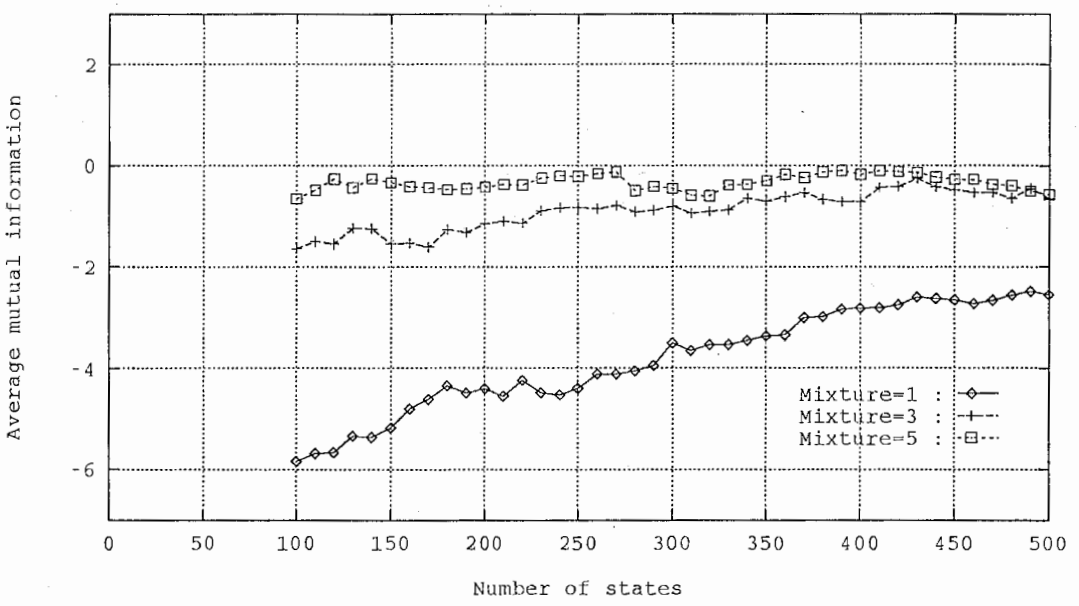

図 6: Mutual information between allophone models and test data.

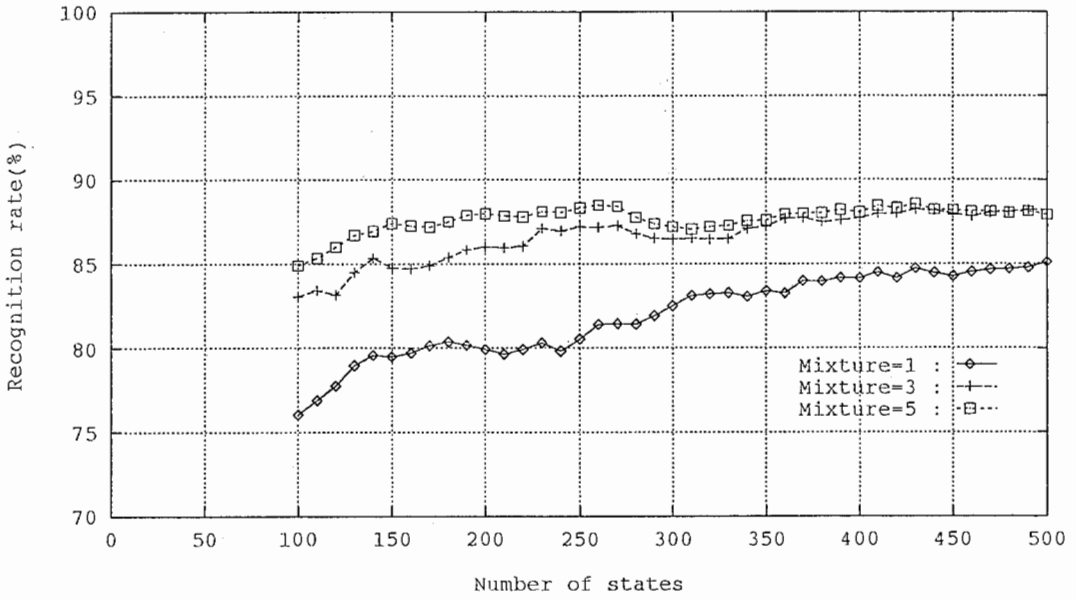

図 7: Phone recognition accuracy for test data.

4 Evaluation for Phonetic Environment Variation of Training Data

4.1 Evaluation sets

In order to investigate the characteristics of HMnet on the variation of phonetic environment of training data, it is necessary to extract from the given data several evaluation sets which have different phonetic contexts. So, I first devided total 2,128 phones in 244 words into two data sets, one for training data and the other for test data, so that each set might satisfy the following conditions.

- Training data set must include all the distinct phones, or 39 phones listed in the Table 1.
- Distribution of phone sample in training data set keeps to be similar with that in original 244 words.
- Ratio of the amount of data between training set and test set is set to about 4:1.

Therefore, the contexts of training data and test data determined by the conditions are basically independent on each other. Using these constraints, I extracted 40 groups randomly that each group consisted of training data set with 1,697 phones and test data set with 431 phones. Then, I computed phone perplexity of training data set for all groups. Finally, I selected 3 groups as evaluation sets by refering to the phone perplexity. The training data sets of the selected 3 groups are listed in Table 3. As we can see in the table, the phonetic variety in the training data set, A is smallest among the 3 sets, on the other hand, the variety in the test data set, A is largest.

4.2 Speaker-dependent experiment for manually segmented data

In order to see more accurately the characteristics of the SSS algorithm for the evaluation sets, first I performed evaluation using manually segmented data in speaker-dependent mode. The manually segmenting data were obtained by correcting the phone boundary information of the Viterbi segmenting data for one speaker. Considering the amount of data and speaker dependency, the mixture number in HMnet was set to 1, and the maximum state number was set to 300. The experiment results for each evaluation set are illustrated in Figures 8 to 14.

From these results, we can know the facts as follows:

- In the Figure 8, the larger the perplexity becomes for same amount of training data, the more the states are split in contextual domain rather than in temporal domain. So, it results in more allophones for complex set.
- In the Figure 9, as number of state increases, allophone entropy is continuously increased for all the 3 sets. It means that SSS splits state of the allophone model having more training data rather than one having less. This tendancy is same as in preliminary experiment.
- In modeling efficiency curve of the Figure 10, the modeling efficiency usually becomes better as perplexity becomes larger. This means that SSS runs so that sharing of training data may become larger in more complex context. Also, the figure shows that increasing and decreasing portion in modeling efficiency rise concentratively. This means SSS splits states concentratively in contextual or temporal domain.
- In the Figures 11 and 12, we can see that the mutual information and the accuracy increase continuously for all the sets. This means that state splitting causes the discriminative power of HMnet for the training data to be improved consistently .
- But, for the test data in the Figures 13 and 14, the diversity in training data greatly affects on the discriminative power for test data. From this, we can know that, when the contexts of training data and test data are independent on each other, diversity of training data becomes an important factor which affects on the performance of model.

表 3: Training data sets.

Set	Phone perplexity	Number of distinct triphones	Triphone entropy
A	4.0	392	7.4
B	4.3	428	7.6
C	4.7	450	7.8

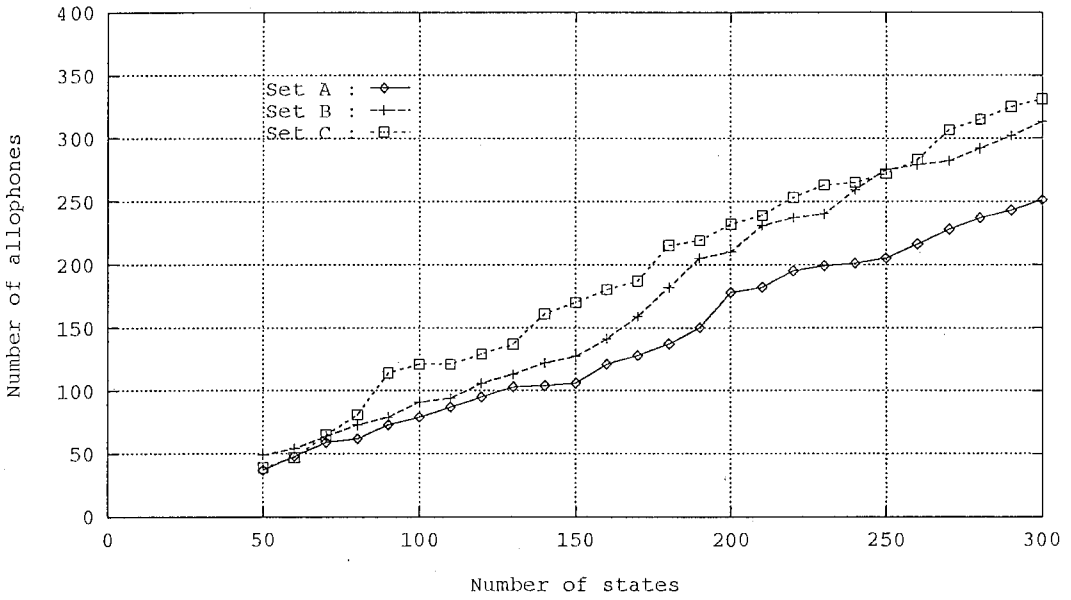

図 8: Number of allophones for each training set with manually segmented data (1 speaker).

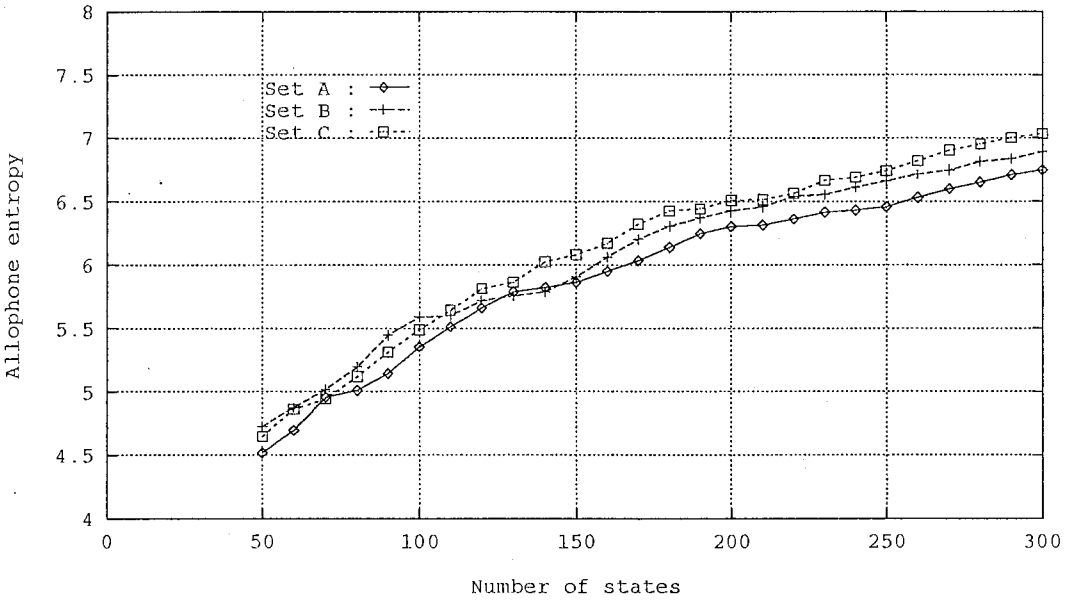

図 9: Allophone entropy for each training set with manually segmented data (1 speaker).

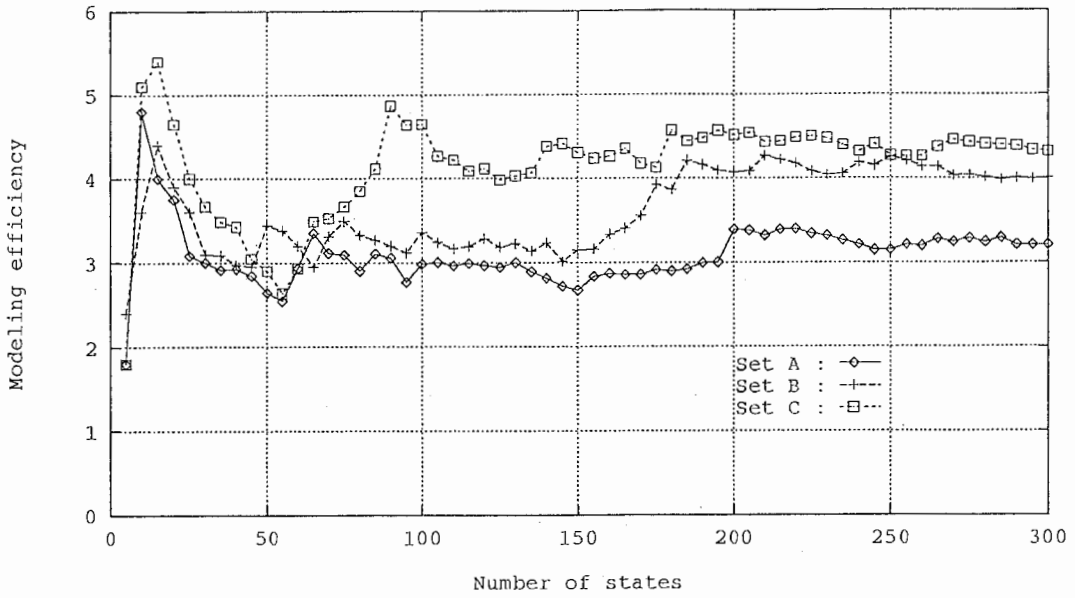

図 10: Modeling efficiency for each training set with manually segmented data (1 speaker).

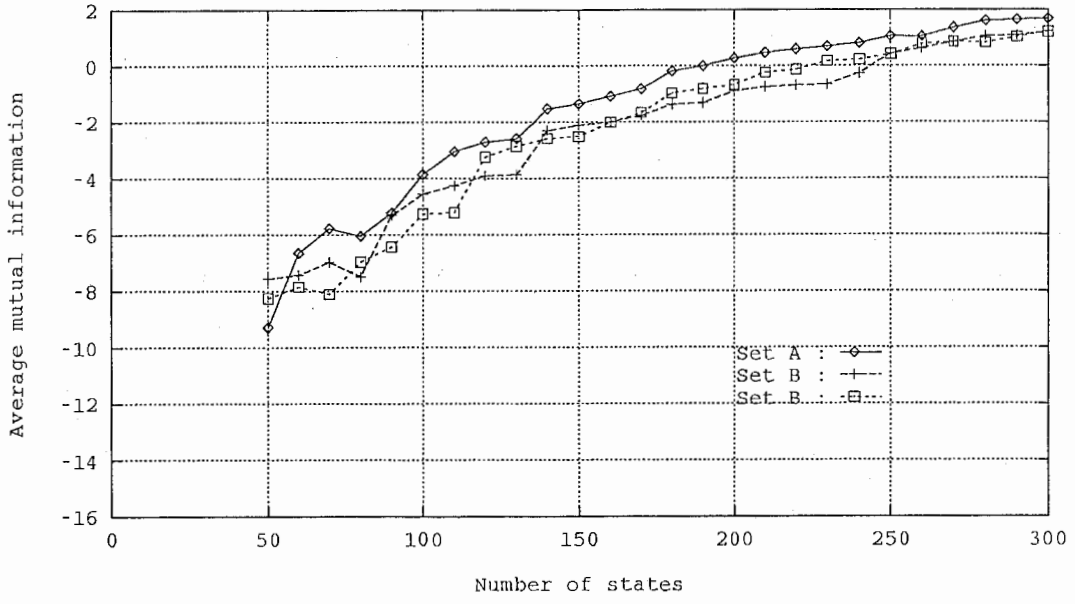

図 11: Mutual information between allophone models and training data with manually segmented data (1 speaker).

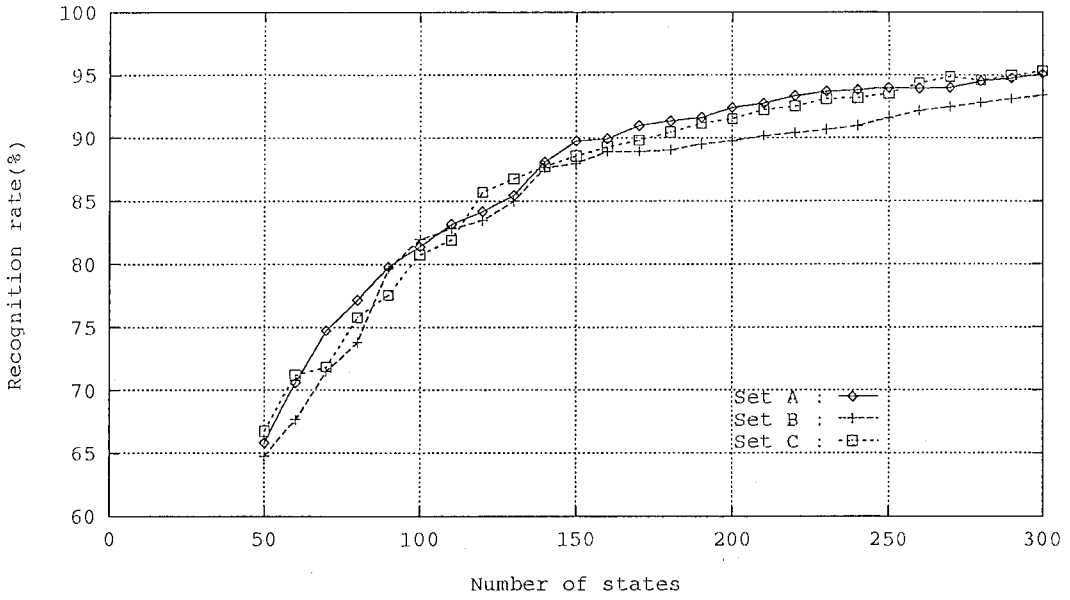

図 12: Phone recognition accuracy for training data with manually segmented data (1 speaker).

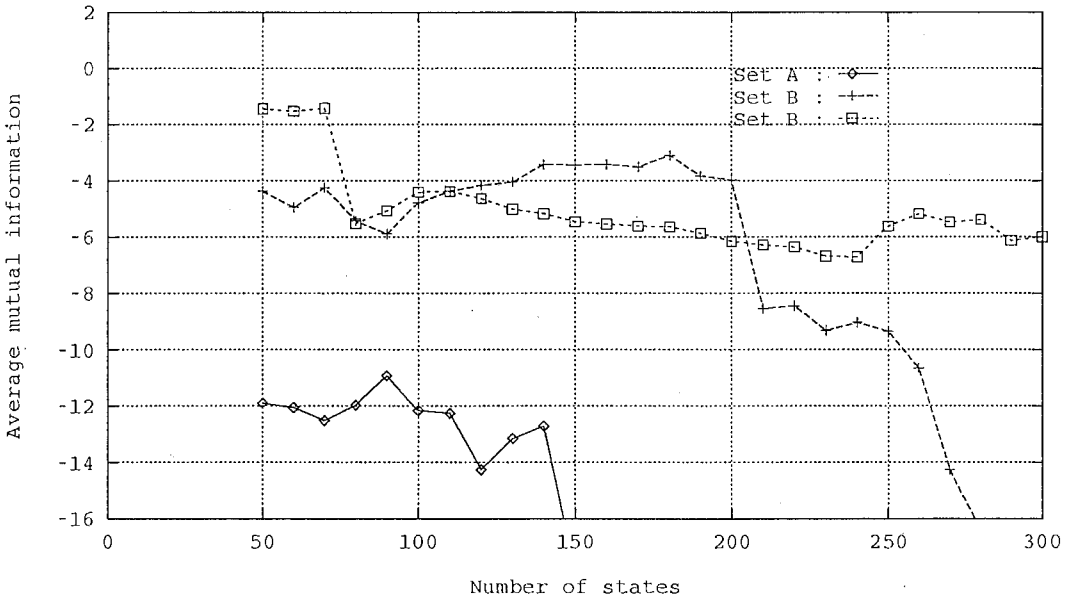

図 13: Mutual information between allophone models and test data with manually segmented data (1 speaker).

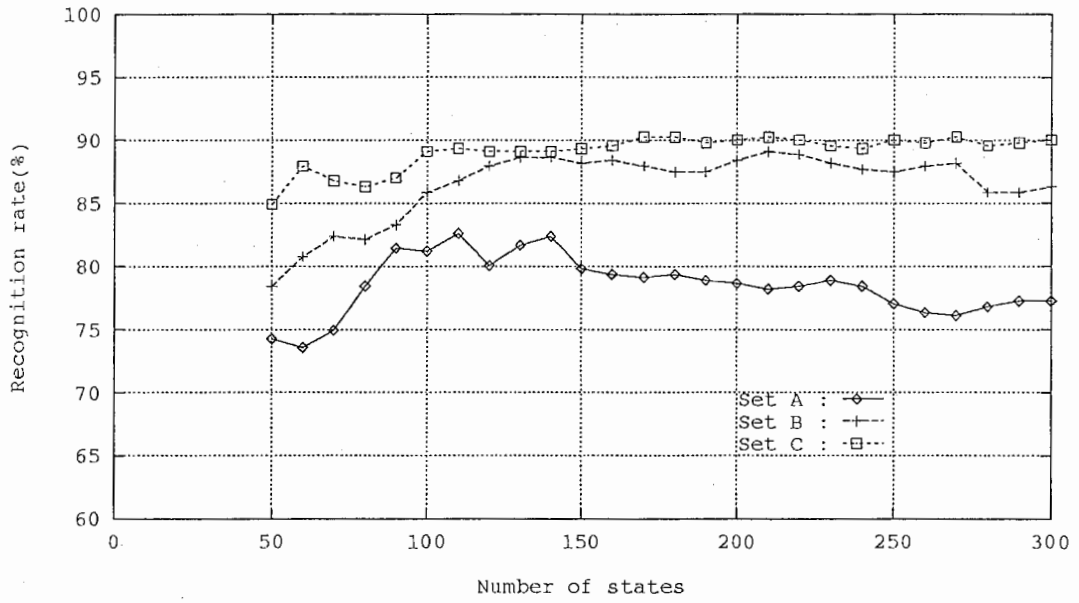

図 14: Phone recognition accuracy for test data with manually segmented data (1 speaker).

4.3 Speaker-dependent experiment for automatically segmented data

To compare the results for manually segmented data with those for automatically segmented data, I performed speaker-dependent experiment on the same conditions as in the previous section, using the automatically segmented data of the same speaker. The results are illustrated in Figures 15 to 21. From the figures, we can see the facts as follows:

- From the comparison of the mutual information curves for the manual and automatic segments, we can confirm that the HMnet generated by using manual segments is more robust than that by automatic segments as we expected.
- Therefore, to obtain more reliable HMnet with automatic segments, it is better to increase the perplexity, or the diversity of training data and decrease the number of state in the HMnet.
- The performance of HMnet is entirely better for manual segments, but the fundamental trend of evaluation results by the two segmentation methods is similar.

With these observations, next I performed speaker-independent evaluation using the automatic segments.

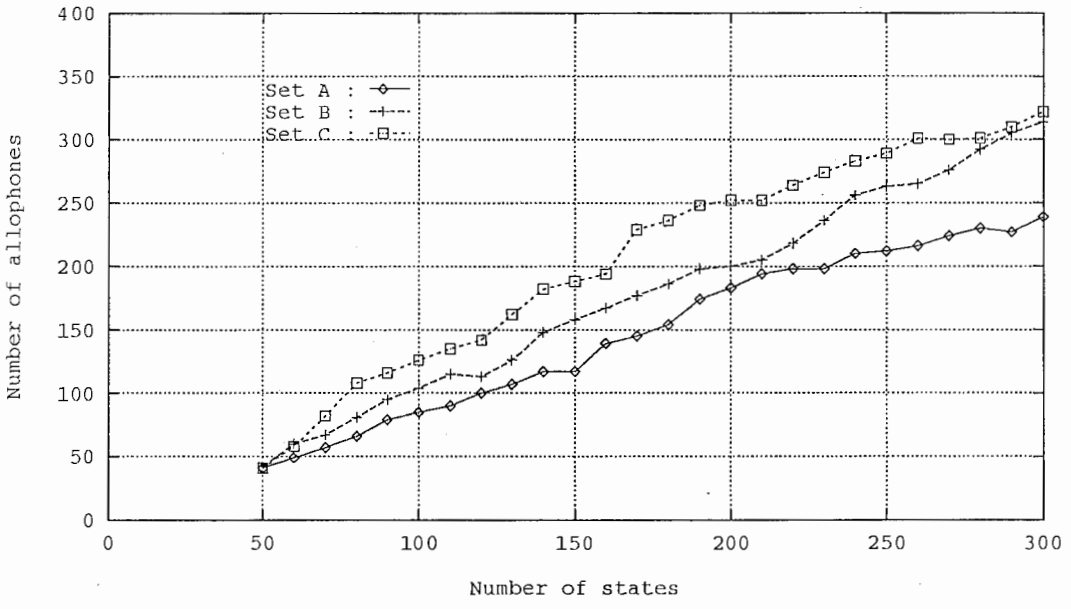

図 15: Number of allophones for each training set with automatically segmented data (1 speaker).

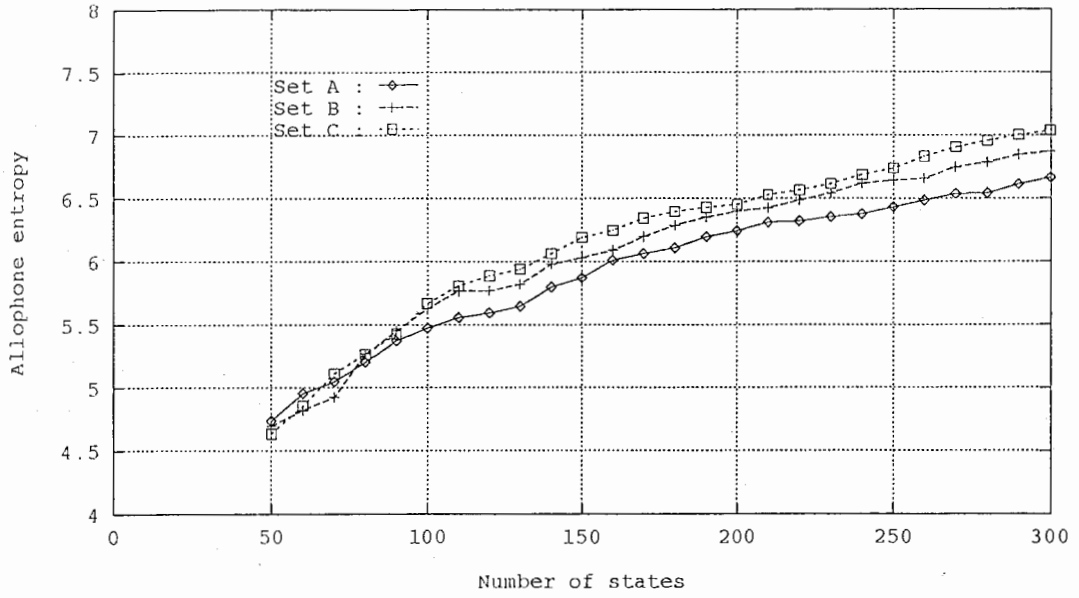

図 16: Allophone entropy for each training set with automatically segmented data (1 speaker).

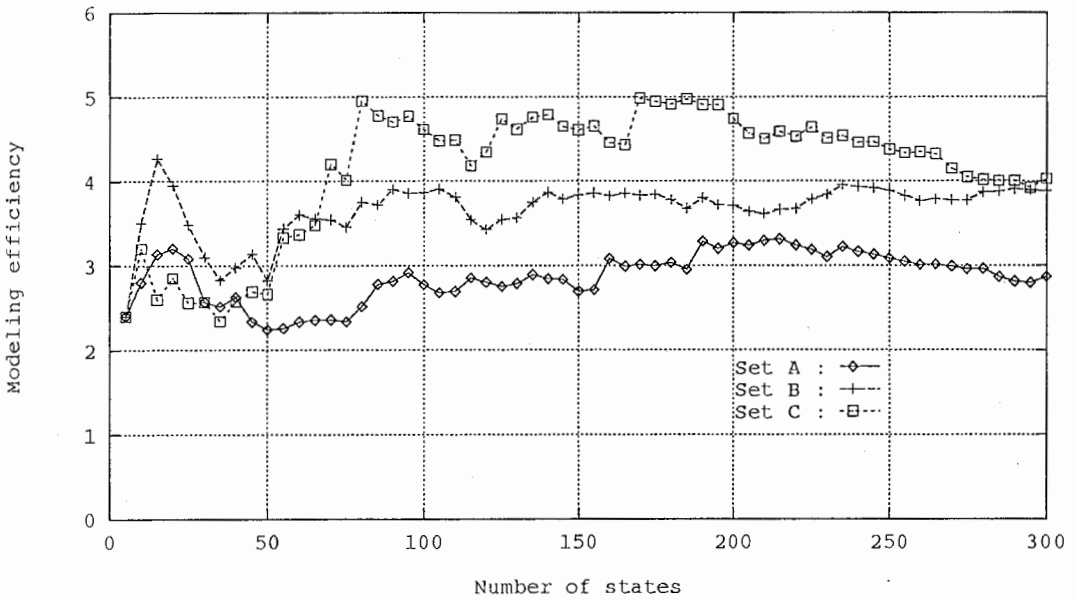

図 17: Modeling efficiency for each training set with automatically segmented data (1 speaker).

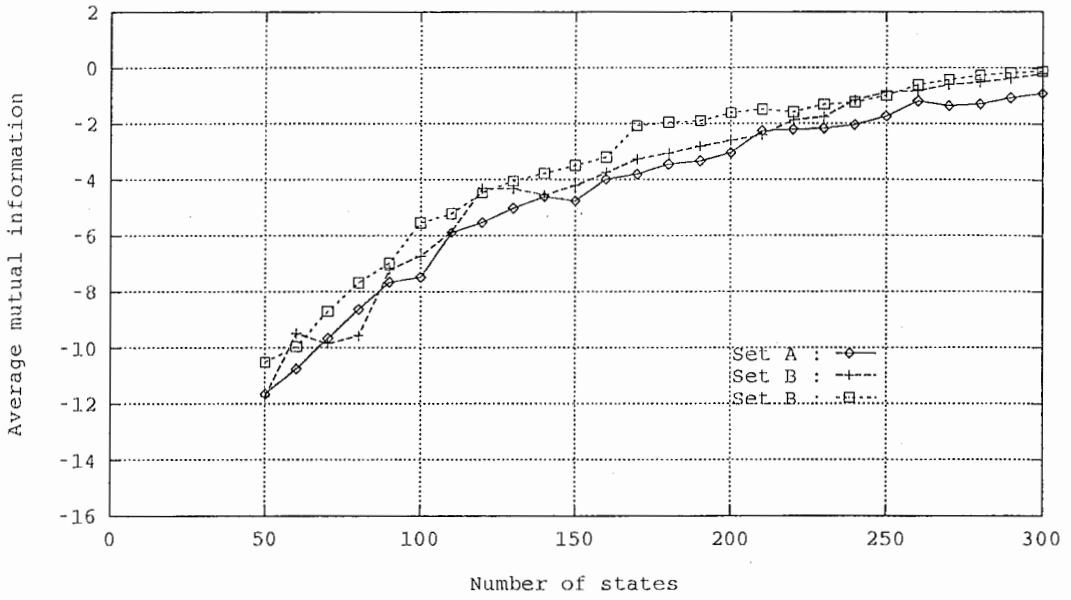

図 18: Mutual information between allophone models and training data with automatically segmented data (1 speaker).

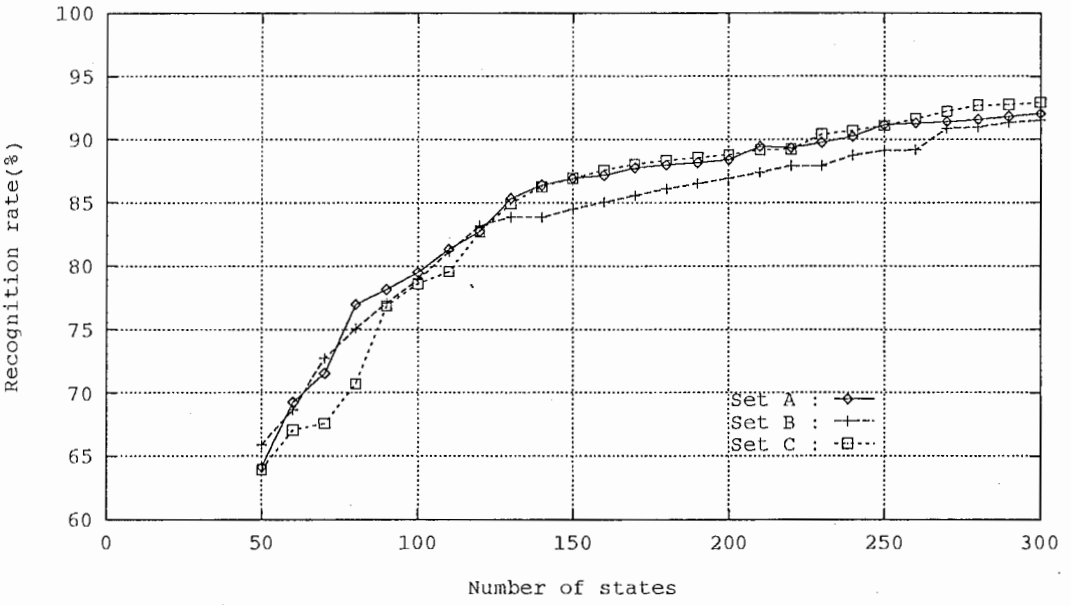

図 19: Phone recognition accuracy for training data with automatically segmented data (1 speaker).

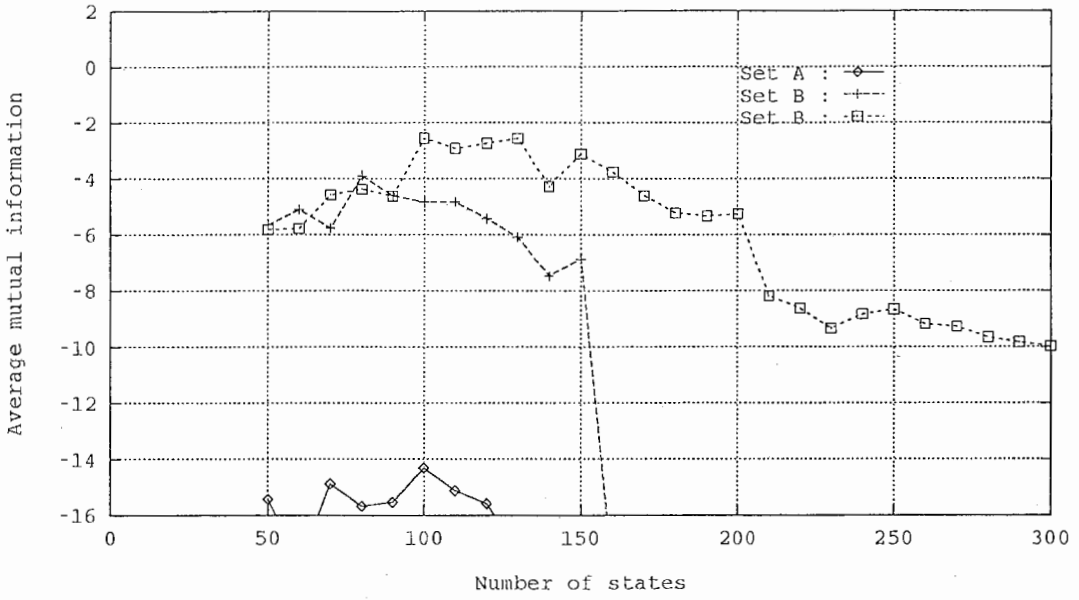

図 20: Mutual information between allophone models and test data with automatically segmented data (1 speaker).

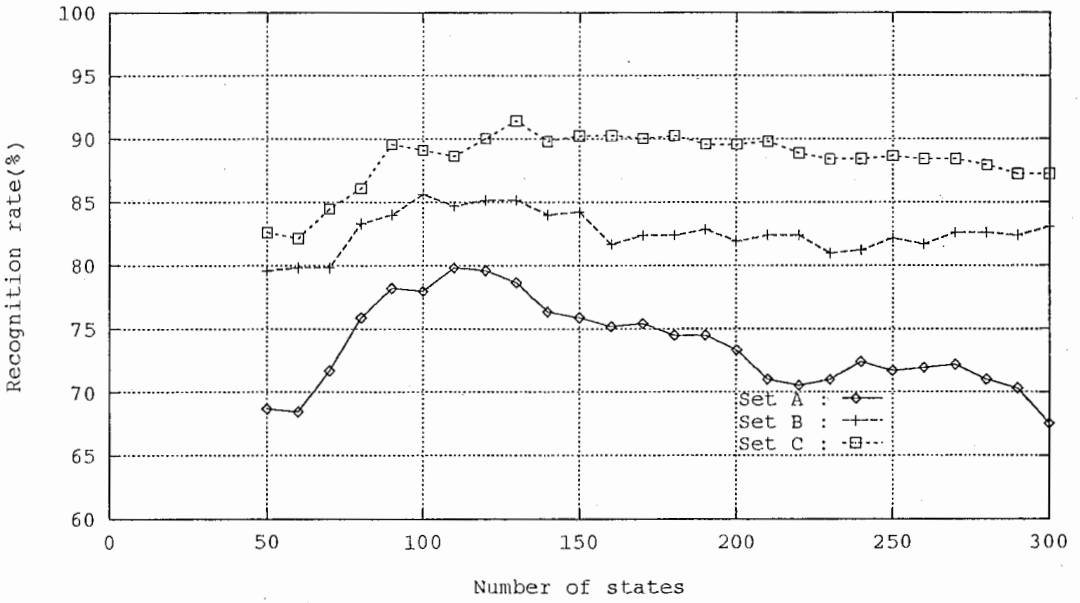

図 21: Phone recognition accuracy for test data with automatically segmented data (1 speaker).

4.4 Speaker-independent experiment for automatically segmented data

Finally, I investigated the HMnet characteristics in speaker-independent case for the phonetic environment variation of training data. In the training for each evaluation data set, the automatically segmented phone data of 9 speakers were used. Each training data set consists of total 15,273 phones. The mixture number in HMnet was set to 3 to reflect the speaker-independency, and the maximum state number was set to 300. Figures 22 to 25 show number of allophone, allophone entropy, modeling efficiency, and mutual information, respectively for each training data set.

In phone recognition test, I performed experiments for 5 cases which were listed in Table 4. In the table, multi-speaker case means the case that the speakers included in the test data set are same as ones in the training data set. 2 speakers used in speaker-independent case are another speakers that are not included in the training data. Also, context-closed case means the case that the phonetic environment in the test data is same as that in the training data, and the test data in context-open case are the remained data excluding each training set from the entire phones in 244 words. Context-mixed case is the case that the test data include the entire phones in 244 words.

From the results of recognition test illustrated in Figures 26 to 30, we can see the facts as follows:

- In the context-open case, the perplexity of training data greatly affects on recognition performance of test data. This is due to the following two reasons. One is that, as the correlation between the phonetic contexts in the training data and the test data becomes weak, the diversity of training data becomes important. The other is that current speech data are too small and biased, so the perplexity of context-open data, or test data, depends on the training data.

- In the context-closed case, the performance variation by training data set is little, but the recognition performance is greatly improved by increasing the number of state, or the number of allophone. On the other hand, increasing the number of state in the context-open case affects little on performance. This means that in context-open case the robustness of model is more important than the precision of model.

- In the speaker-independent case, the performance of context-closed case was worse than that of context-open case. I think this also is due to the characteristics of current data.

表 4: Test data sets for speaker-independent experiment.

Test set	Condition	Amount of data
MS-CC	Multi-speaker, context-closed (or training data itself)	15,273 phones (9 speakers)
MS-CO	Multi-speaker, context-open	3,879 phones (9 speakers)
SI-CC	Speaker-independent, context-closed	3,394 phones (2 speakers)
SI-CM	Speaker-independent, context-mixed (SI-CC + SI-CO)	4,256 phones (2 speakers)
SI-CO	Speaker-independent, context-open)	862 phones (2 speakers)

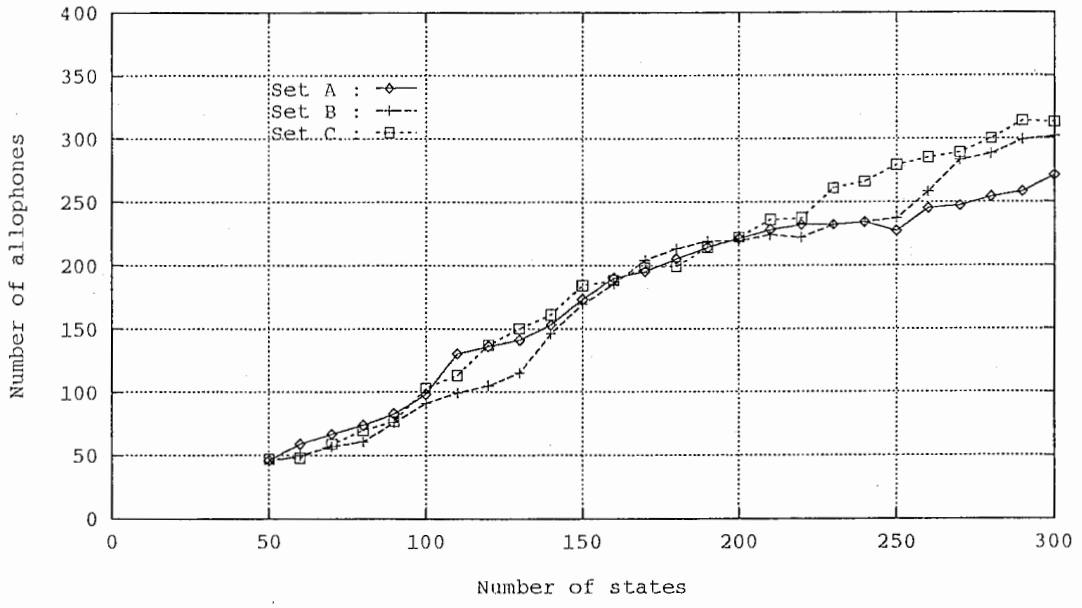

図 22: Number of allophones for each training set with multiple speaker data (9 speakers).

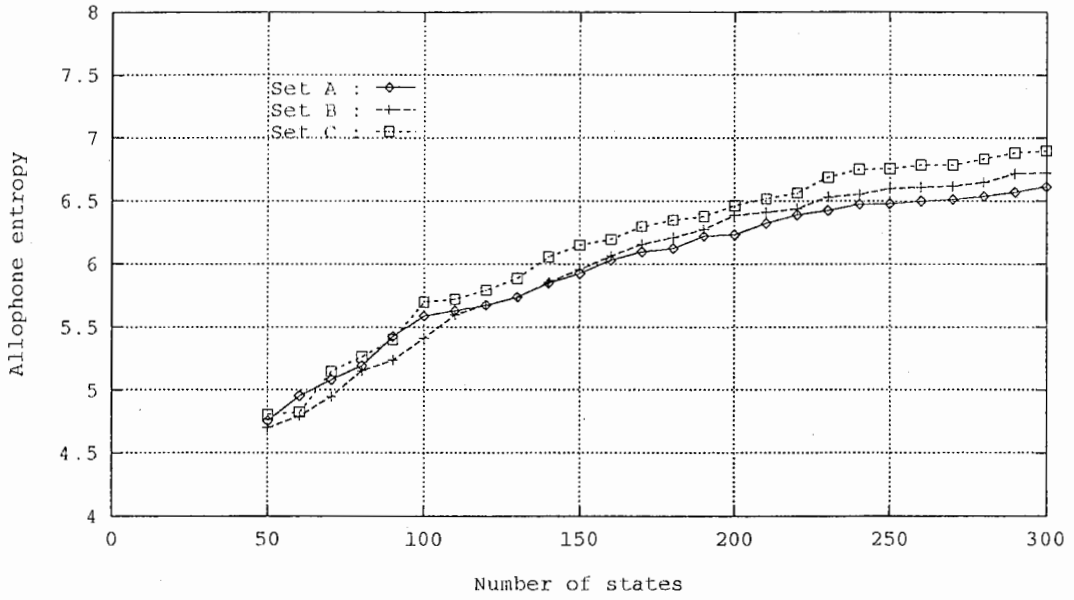

図 23: Allophone entropy for each training set with multiple speaker data (9 speakers).

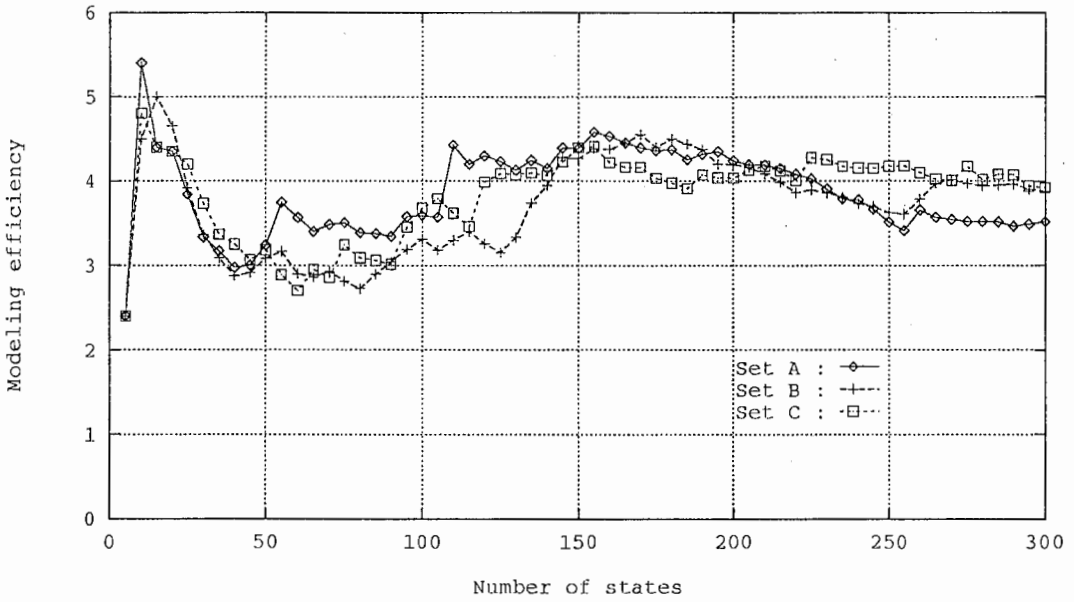

図 24: Modeling efficiency for each training set with multiple speaker data (9 speakers).

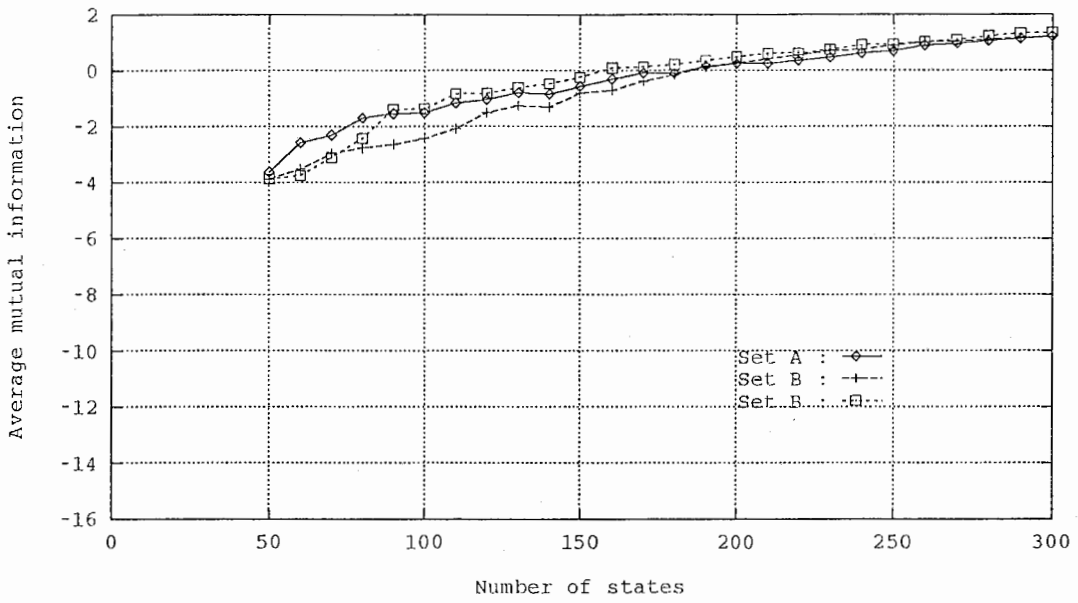

図 25: Mutual information between allophone models and training data with multiple speaker data (9 speakers).

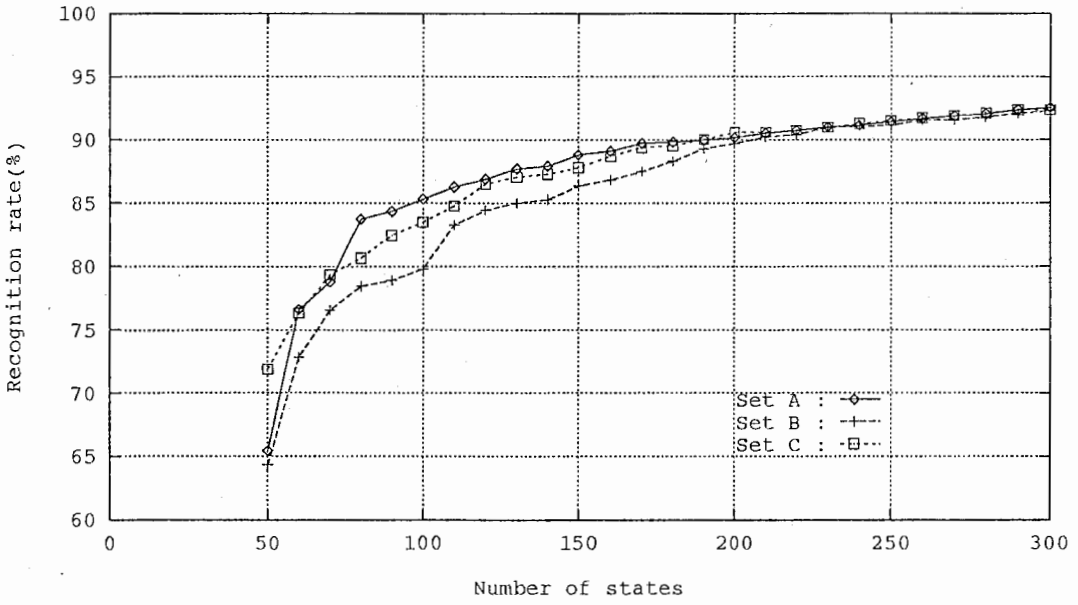

図 26: Phone recognition accuracy for training data (multi-speaker, context-closed, 9 speakers).

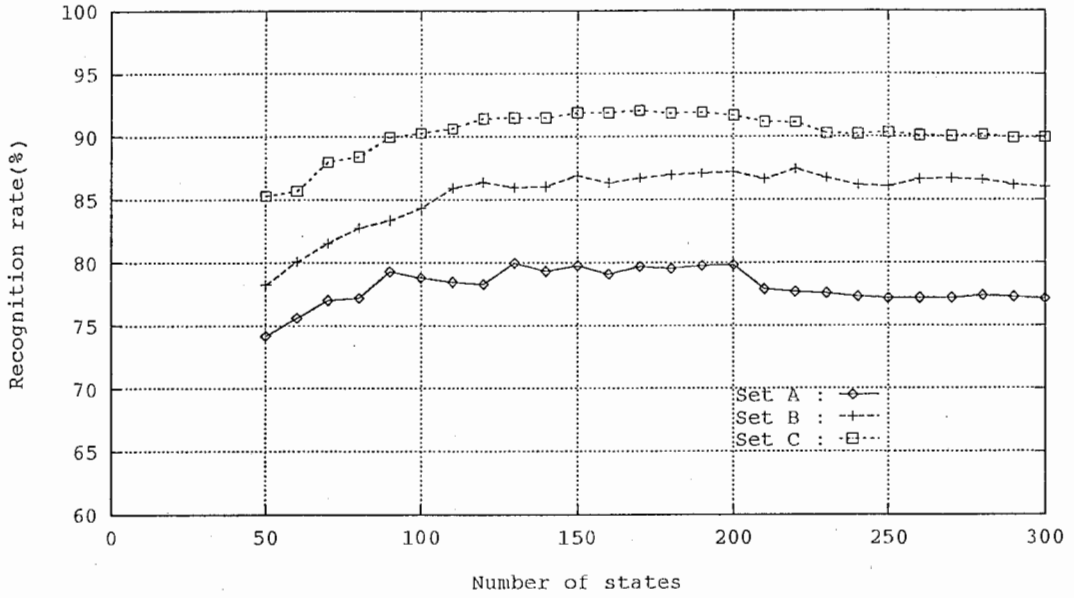

図 27: Phone recognition accuracy for multi-speaker, context-open data (9 speakers).

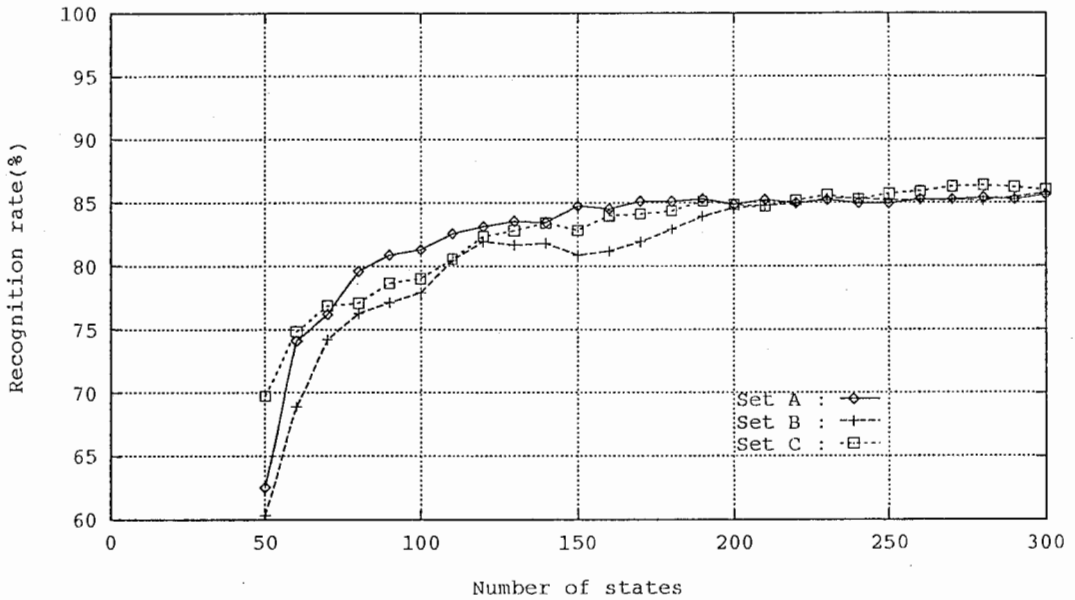

図 28: Phone recognition accuracy for speaker-independent, context-closed data (2 speakers).

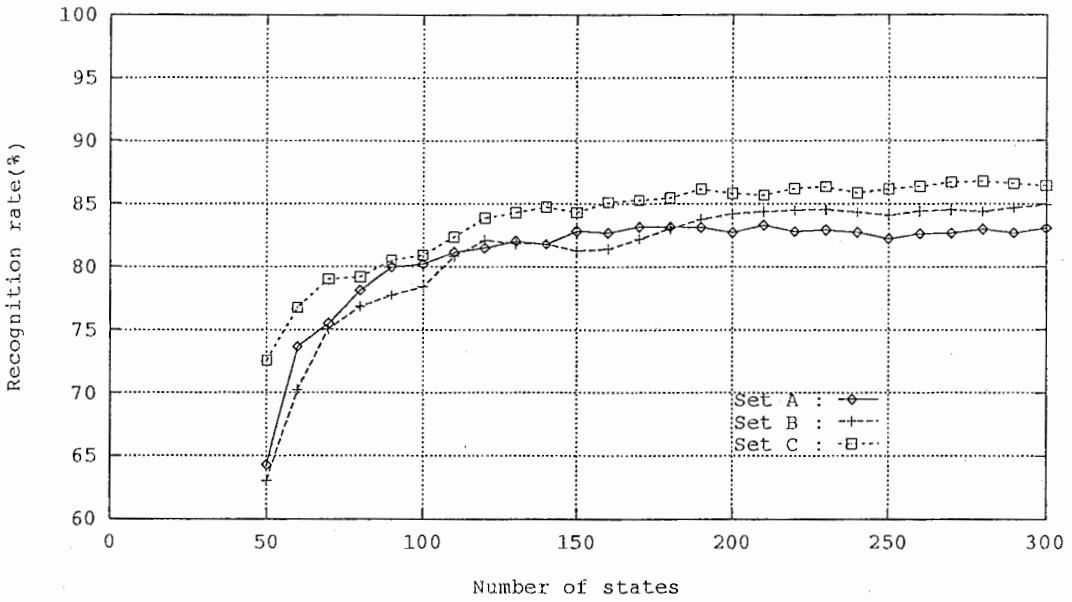

図 29: Phone recognition accuracy for speaker-independent, context-mixed data (2 speakers).

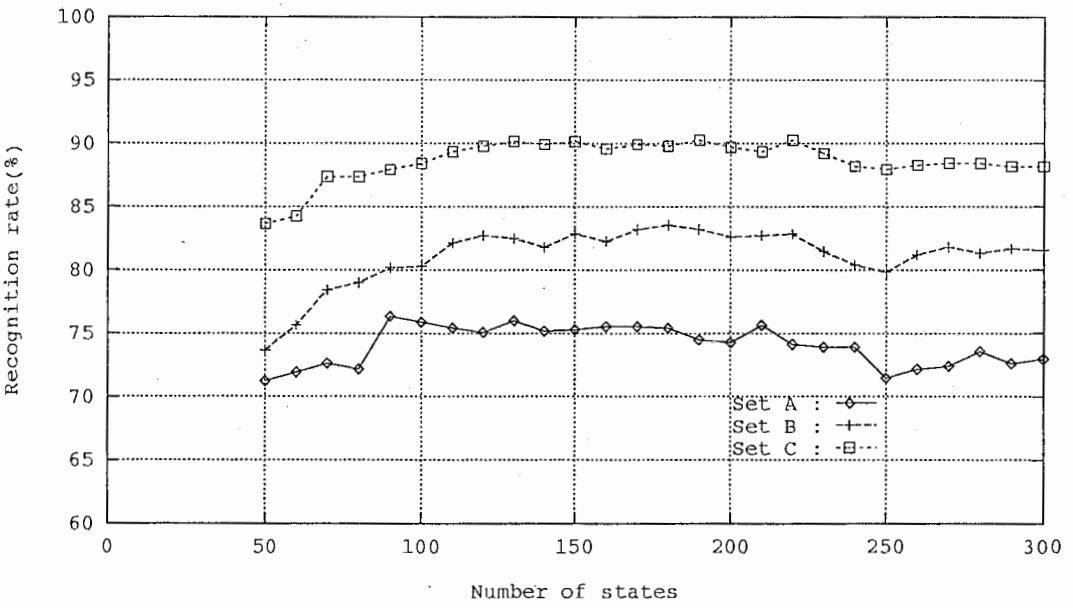

図 30: Phone recognition accuracy for speaker-independent, context-open data (2 speakers).

5 Conclusions

In this work, I have investigated how the variation of phonetic environments of training data affects on the HMnet generated by SSS algorithm. From the experiment results, I could confirm the facts as follows:

- In context-open, or vocabulary-independent, recognition task, phonetic diversity of training data and improvement of robustness by a proper sharing in model parameters are very important.
- In training procedure of HMnet, it is necessary to device a proper measure which can determine the number of state to be able to compromize the robustness and the precision of HMnet more better than the conventional modeling efficiency.
- It is necessary to develop a method to be able to use directly speaker-independent data in determination of HMnet topology. Of cause, some methods called as 3-domain SSS (3D-SSS) and speaker parallel SSS (SP-SSS) [7] have been proposed, but the performance is not so good because those methods basically have so much freedom in state splitting on temporal, contextual, and speaker domains.

参考文献

[1] J. Takami and S. Sagayama, "A successive state splitting algorithm for efficient allophone modeling," *Proc. of ICASSP*, pp. I-573-576, 1992.

[2] M. Hwang and X. Huang, "Subphonetic modeling with Markov states - SENONE," *Proc. of ICASSP*, pp. I-33-36, 1992.

[3] V. Digalakis and H. Murveit, "Genones: Optimizing the degree of mixture tying in a large vocabulary hidden Markov model based speech recognizer," *Proc. of ICASSP*, pp. I-537-540, 1994.

[4] J. Takami and S. Sagayama, "Automatic generation of hidden Markov networks by a successive state splitting algorithm," *Trans. on IEICE*, vol. J76-D-II, no. 10, pp. 2155-2164, 1993.

[5] J. Takami, "SSS-ToolKit (Ver 3.0) user's manual," *Technical Report of ATR-ITL*, TR-IT-0039, 1994.

[6] L. R. Bahl, et al., "Maximum mutual information estimation of hidden Markov model parameters for speech recognition," *Proc. of ICASSP*, pp. 49-52, 1986.

[7] J. Takami, T. Kosaka, and S. Sagayama, "Automatic generation of speaker-common hidden Markov network by adding the speaker splitting domain to the SSS algorithm," *Proc. of ASJ Conference*, pp. 3-1-8, 1992.

Acknowledgements

I would like to thank Dr. Y. Yamazaki, President, ATR-ITL, for giving a chance to research at ATR, and also thank Dr. Sagisaka, Dr. Loken-Kim, Dr. Matsunaga, Mr. Singer, and all other members in ATR-ITL for their support of this work. In addition, I would like to acknowledge Mr. Jaewoo Yang, Head of Human Interfaces Technology Department, ETRI, Dr. Youngjik Lee, Head of Spoken Language Processing Section, ETRI, and all other members of the Section for their deep consideration.