HMnet Evaluation for Phonetic Environment Variations of Training Data in Speech Recognition

Hoi-Rin Kim*

Abstract

In this paper, we propose a new evaluation methodology which can more clearly show the performance of the allophone modeling algorithm generally used in large vocabulary speech recognition. The proposed evaluation method shows the running characteristics and limitations of the modeling algorithm by testing how the variation of phonetic environments of training data affects the recognition performance and the desirable number of free parameters to be estimated. Using the method, we evaluated the hidden Markov network (HMnet) generated by the successive state splitting (SSS) algorithm. From the experiment results, we conclude that, in vocabulary-independent recognition task, the phonetic diversity of training data greatly affects the robustness of model, and it is necessary to develop a proper measure which can determine the number of states compromizing the robustness and the precision of the HMnet better than the conventional modeling efficiency.

Ⅰ. Introduction

In HMM-based acoustic modeling, it has been well known that to appropriately compromize the degree of precision and robustness of each model for a given training data is very important. When we model a triphone, if each triphone is modeled independently from another, then the precision of the model would be high, but the robustness of the model becomes weak. This problem is basically dependent on the amount of training data used in the modeling procedure. Another more basic problem is that any training data cannot cover all phonetic environments which can occur in a language. These problems can be solved theoretically by using infinitely huge training data, but it is impossible in realistic sense. However, we know that the acoustic characteristics of any triphone would not be fully independent from another, that is, would have certain dependency on some phonetic environments. In order to apply the knowledge to the triphone modeling, we must observe acoustic relations among all phonetic combinations, but this also is not easy. As an alternative approach, we can use a statistical methodology to obtain the relationship and to compromise between the precision and the robustness. Several novel methods for realizing precise and robust triphone models have been proposed such as HMnet [1], Senone [2], and Genone [3], and they all showed good performances. HMnet is similar to Senone in the aspect that they have state sharing architectures. Main difference between them is that Senone is made through a merging procedure which is controlled mainly by acoustic characteristics of states, while HMnet is generated through a splitting procedure which is controlled by both acoustic and phonetic characteristics. On the other hand, Genone has a parameter sharing architecture only in Gaussian mixtures, not in mixture weights, and is conceptually same as Senone in the aspect of the merging procedure. So, Genone is more adequate than HMnet or Senone for larger training data because its architecture consequently increases the degree of freedom in observation parameters sharing constraints.

Among these properties, we especially focused on the phonetic constraints in the splitting procedure in HMnet, called the SSS algorithm. In the HMnet generation, the phonetic environments, that is, the previous, present, and the following phones, are considered as constraints for state splitting. Therefore, it is important to observe which relations the generation procedur has with respect to variations of phonetic environments of training data. But, until now, evaluation of the SSS algorithm has usually been performed only on the aspects of recognition performance and the number of free parameters to be estimated, which is related to the robustness of the model [4]. So, in this paper, we observed HMnet's characteristics on phonetic context environment variations in training data.

*Electronics and Telecommunications Research Institute(ETRI)
Manuscript Received 96. 9. 14

By analyzing the results, we could more clearly understand the SSS algorithm's behavior and its own limitation.

In section Ⅱ, we briefly review the SSS algorithm, in section Ⅲ, preliminary experimental results for Korean speech DB are described, and in section Ⅳ, we evaluate the HMnet on different phonetic environments of training data. Lastly, in section Ⅴ, the experimental observations are summarized.

Ⅱ. Brief review of SSS algorithm and HMnet

We can see the HMnet as a network configuration of allophonic HMMs which has phonetic context dependency and state sharing architecture. The HMnet is automatically generated by an unified algorithm, that is, the SSS algorithm, which can simultaneously determine and estimate an optimal set of allophones, an optimal state sharing architecture, and optimal parameters with the maximum likelihood criterion. The processing sequence can be briefly summarized as follows. More details can be found 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 through the 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 the ratio of the total number of states, needed to represent all the allophone models without any state sharing, to that in the obtained HMnet itself. Therefore, the modeling efficiency measures the degree of state sharing, or the statistical robustness, of the HMnet for the given training data and the state number of the obtained HMnet.

Ⅲ. Preliminary experiment for the SSS algorithm with Korean speech data

1. Speech DB and experiment conditions

The speech data for our preliminary experiment were Korean words which were collected for the hotel reservation task. The vocabulary consists of 244 words including some connected digits, 26 English alphabets, months, weeks, date names, and so on. Total 40 male speakers spoke the 244 words once. The whole words are proved to include all Korean distinct phones except only one phone. Consequently, in our data 39 distinct phones including one silence unit exist.

Originally, the speech data were digitized at 16 kHz sampling rate. But, in order to keep consistency with conventional acoustic analysis method, we downsampled the data at 12 kHz. Then, the 12 kHz data were processed as follows.

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

2. Automatic segmentation into phone unit by Viterbi alignment

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

into the corresponding phone units by Viterbi alignment. We used the SSS-ToolKit(Ver 3.0) in all procedures.

3. Context-dependent phone modeling by SSS

At first, we observed the running characteristics of the SSS algorithm with Korean speech DB to confirm the ability of the algorithm. In training mode, we used 10 speakers' data among the automatically segmented data. The training word data include 21,268 phones in total. Also, we used other 2 speakers' data as test data set, including 4,256 phones. For the SSS algorithm's parameter setting, basically default values of the algorithm were utilized, 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, we set the maximum number of states to 500, and in final re-estimation of model parameters, each model was trained for the mixture number 1, 3, and 5, respectively.

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

- Number of allophones (see Fig. 1)
- Allophone entropy (see Fig. 2)
- Modeling efficiency (see Fig. 3)
- Mutual information between allophone models and training data (see Fig. 4)

Where allophone entropy was computed with number of training samples used in the 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 [5].

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

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

In Fig. 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 Fig. 3, the increased area means that the states are split ma-

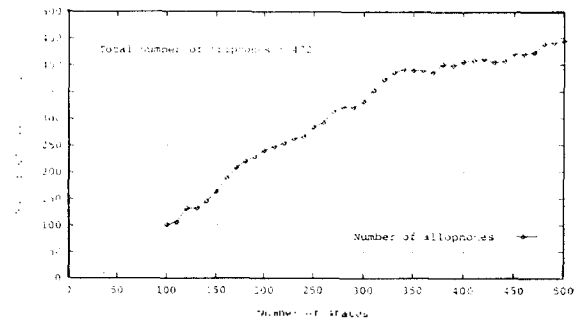

Figure 1. Number of allophones.

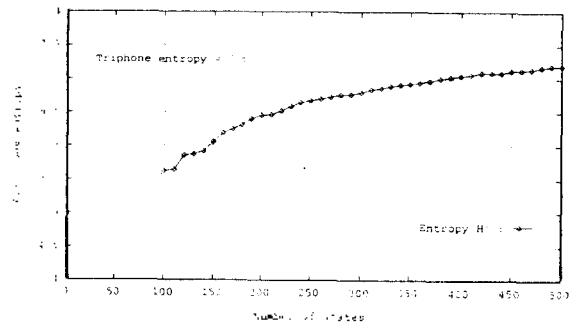

Figure 2. Allophone entropy.

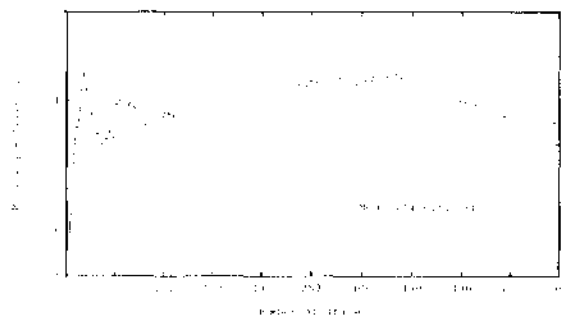

Figure 3. Modeling efficiency.

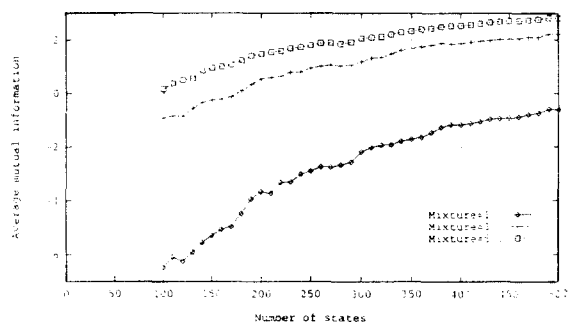

Figure 4. Mutual information between allophone models and training data.

Table 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
Semivowel	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

Table 2. Phonetic characteristics in training data (10 speakers).

Phone perplexity	5.0
Number of distinct triphones	472
Triphone entropy	7.6

inly in contextual domain, on the other hand, the decreased area means that the states are split mainly in temporal domain. From this result, we can say that it is appropriate to decide the number of states at between 300 and 350 as a proper compromization of the precision and the robustness for this task domain In the mutual information curve of Fig. 4, we can see that, as the mixture number of output probability distribution at each state increases and the number of states increases, the mutual information also increases. That means the number of free parameters are related to the precision of the model and this is also consistent with the result of the phone recognition for training data.

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, we used other 2 speakers' data with the same vocabulary as training data. The speaker-independent test results are shown in Fig. 5. From the figure, 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, we asserted that it would be appropriate to decide the number of states between 300 and 350 from the modeling efficiency curve. This is not in harmony with the results from the recognition test. It means to decide appropriately the number of states by referring only to the modeling efficiency curve is not easy. Therefore, it is necessary to develop a proper measure which can guarantee the number of states with the balanced precision and robustness of the HMnet in a given condition.

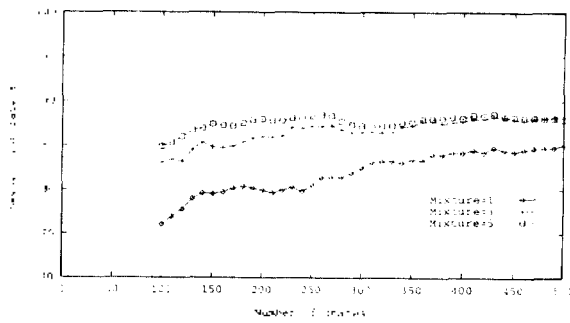

Figure 5. Phone recognition accuracy for test data.

Ⅳ. Evaluation for phonetic environment variations of training data

1. Evaluation sets

In order to see the characteristics of the HMnet with the variation of phonetic environment of training data, it is necessary to extract several evaluation sets having different phonetic contexts from the given data. So, we first divided 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 Table 1.
- Distribution of phones in training data set has to be similar with that in original 244 words.
- Ratio of the amount of data between training set and test set is set to be about 4 : 1.

Therefore, the contexts of training data and test data determined by the conditions are basically independent of each other. Using these constraints, we extracted 40 groups randomly that each group consisted of training data set with 1,697 phones and test data set with 431 phones. Then, we computed phone perplexity of training data set for all groups. Finally, we selected 3 groups as evaluation sets by referring 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 variation in the training data set is smallest for A among the 3 sets, on the other hand, the variation in the test data set is largest for A.

Table 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

2. Speaker-dependent experiment for manually segented data

In order to see more accurately the characteristics of the SSS algorithm for the evaluation sets, firstly we performed evaluation using manually segmented data in a speaker-dependent mode. The manually segmenting data were obtained by correcting the phone boundary information of the Viterbi segmentation data for one speaker. Considering the amount of data and speaker dependency, the mixture number for the HMnet was set to 1, and the maximum state number was limited to 300. The experiment results for each evaluation set are illustrated in Fig. 6 to 9.

From these results, we can find:

- In Fig. 6 the larger the perplexity becomes for same amount of training data, the more the states are split in contextual domain rather in temporal domain. S, it results in more allophones for complex set.
- In the modeling efficiency curve of Fig. 7, the modeling efficiency usually becomes better as perplexity becomes larger. This means that SSSS 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 the states concentratively in contextual or temporal domain.
- In Fig. 8, we can see that the mutual information increases continuously for all sets. This means that the state splitting causes the discriminative power of HMnet for the training data to be improved consistently.

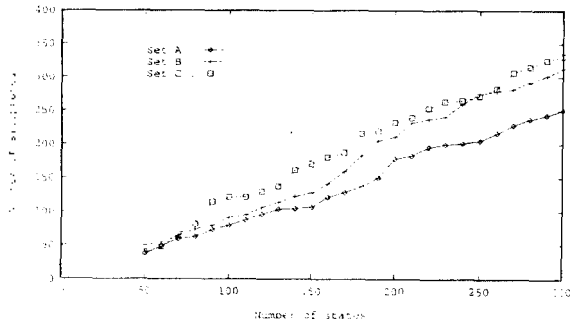

Figure 6. Number of allophones for each training set with manually segmented data (1 speaker).

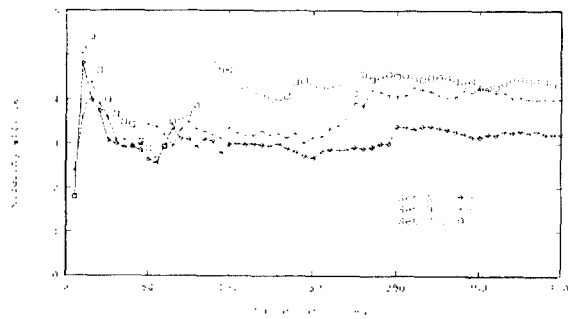

Figure 7. Modeling efficiency for each training set with manually segmented data (1 speaker).

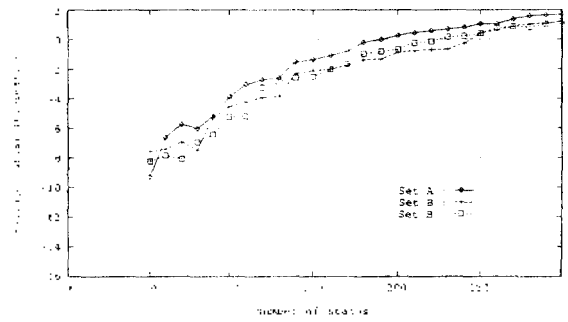

Figure 8. Mutual information between allophone models and training data with manually segmented data (1 speaker).

- But, for the test data in Fig. 9, the diversity in training data greatly affects the discriminative power for test data. From this, we can know that, when the contexts of training data and test data are independent from each other, the diversity of training data becomes an important factor which affects the model performance.

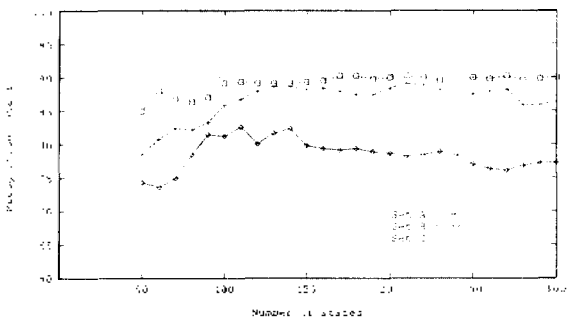

Figure 9. Phone recognition accuracy for test data with manually segmented data (1 speaker).

3. Speaker-dependent experiment for automatically segmented data

To compare the results for manually segmented data with those for automatically segmented data, we performed a 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 Fig. 10 to 13. From the figures, we can see the followings:

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

With these observations, nextly we performed speaker-independent evaluation using the automatic segments.

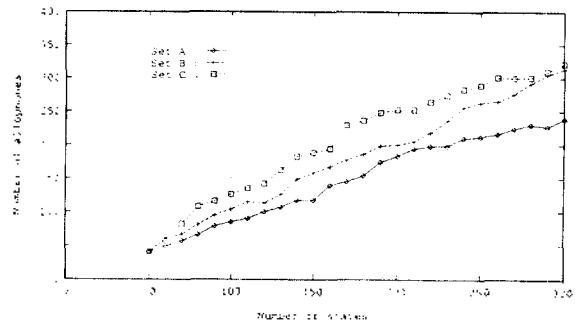

Figure 10. Number of allophones for each training set with automatically segmented data (1 speaker).

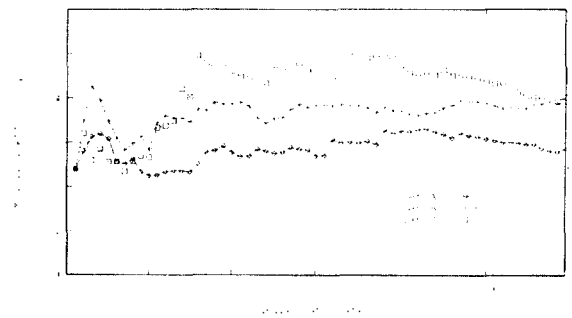

Figure 11. Modeling efficiency for each training set with automatically segmented data (1 speaker).

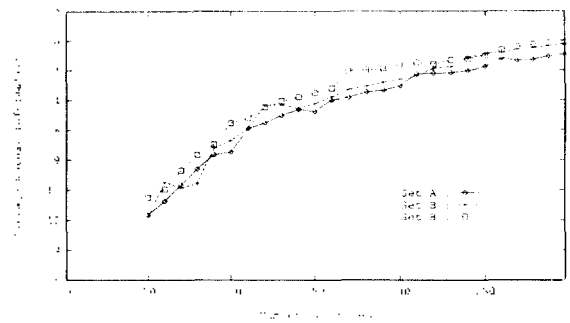

Figure 12. Mutual information between allophone models and training data with automatically segmented data (1 speaker).

Table 4. Test data sets for speaker-independent experiment.

Test set	Condition	Amount of data
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)

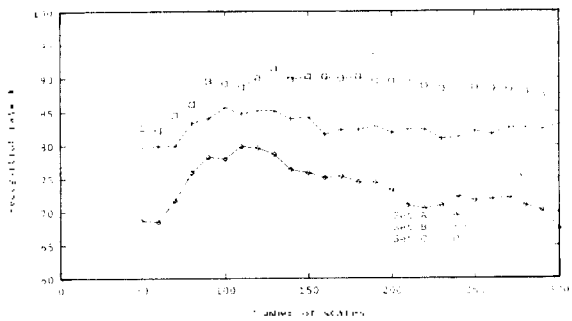

Figure 13. Phone recognition accuracy for test data with automatically segmented data (1 speaker).

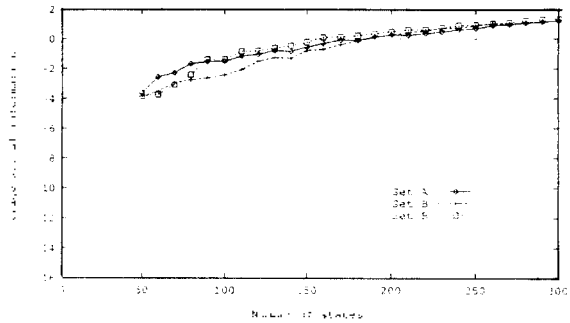

Figure 16. Mutual information between allophone models and training data with multiple speaker data (9 speakers).

4. Speaker-independent experiment for automatically segmented data

Finally, we looked at the HMnet characteristics in speaker-independent case for the phonetic environment variations of training data. In the training with each evaluation data set, the automatically segmented phone data of 9 speakers were utilized. Each training data set consists of 15,273 phones in total. The mixture number in the HMnet was set to 3 to reflect the speaker-independency, and the maximum state number was set to 300. Fig. 14 to 16 show number of allophone, modeling efficiency, and mutual information, respectively for each training data set.

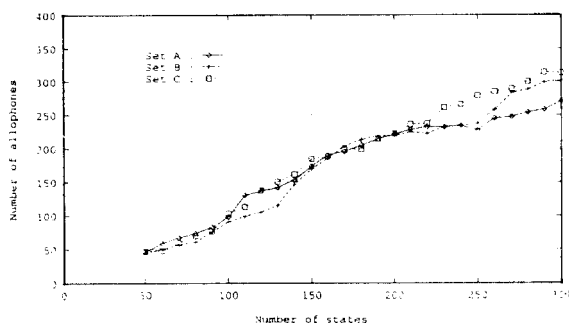

Figure 14. Number of allophones for each training set with multiple speaker data (9 speakers).

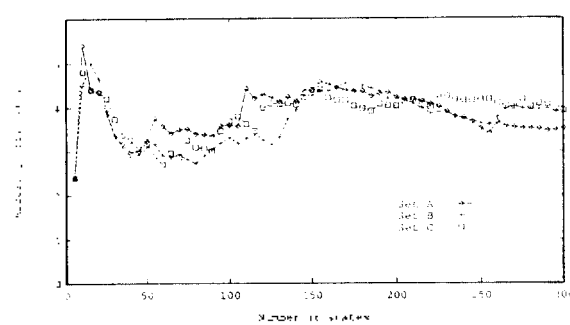

Figure 15. Modeling efficiency for each training set with multiple speaker data (9 speakers).

In phone recognition test, we 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 those in the training data set. 2 speakers used in speaker-independent case are other speakers who 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 remaining data excluding each training set from the entire phones in 244 words. Context-mixed case means that the test data include the entire phones in 244 words.

From the results of our recognition test illustrated in Fig. 17 to 20, we can say the followings:

- In the context-open case, the perplexity of the training data greatly affects the recognition performance of the test data. This is due to the following two reasons. One is, 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, current speech data are too small and considered to be 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 the training data set is negligible, but the recognition performance is greatly improved by increasing the number of states, or the number of allophones. On the other hand, increasing the number of states in the context-open case affects the performance little. This means that in context-open case the robustness of the model is more important than the precision.
- In the speaker-independent case, the performance of context-closed case was worse than that of context-open case. It might occur because of the charact-

eristics of our current data.

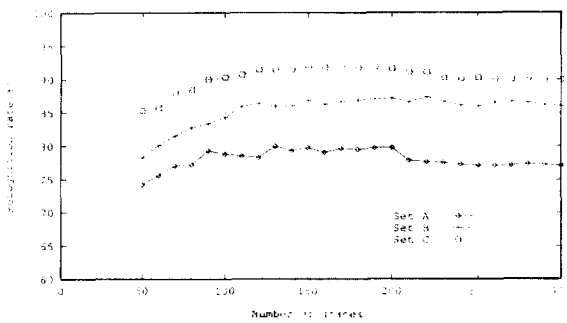

Figure 17. Phone recognition accuracy for multi-speaker, context-open data (9 speakers).

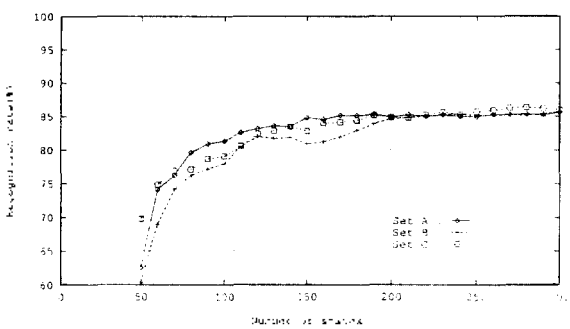

Figure 18. Phone recognition accuracy for speaker-independent, context-closed data (2 speakers).

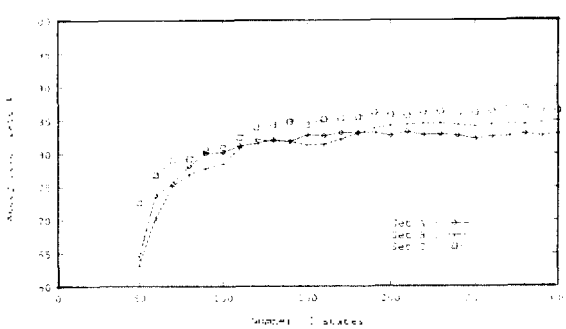

Figure 19. Phone recognition accuracy for speaker-independent, context-mixed data (2 speakers).

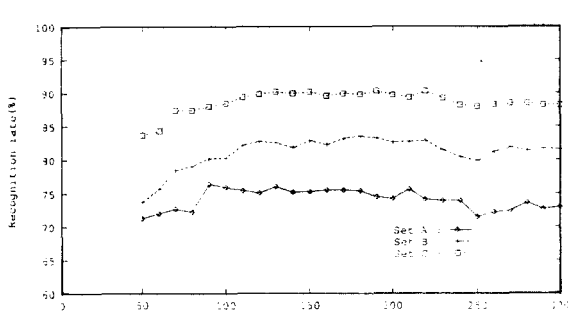

Figure 20. Phone recognition accuracy for speaker-independent, context-open data (2 speakers).

V. Conclusions

In this paper, we have experimented how the variation of phonetic environments of training data affects the HMnet generated by the SSS algorithm. From the experiment results, we say the followings:

- In context-open (or vocabulary-independent) recognition task, phonetic diversity of training data and improvement of robustness through a proper sharing in model parameters greatly affect on the reliability of the system.
- In training procedure of the HMnet, it is necessary to develop a proper measure which can determine the number of states compromizing the robustness and the precision of the HMnet better than the conventional modeling efficiency.
- It is necessary to develop a method which can utilize directly speaker-independent data in determination of the HMnet topology. Some methods such as 3-domain SSS (3D-SSS) and speaker parallel SSS (SP-SSS) [6] have been proposed, but the performance is not so good because those methods basically have too much freedom in state splitting on temporal, contextual, and speaker domains.

References

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

▲Hoi-Rin Kim

Feb. 1984: B.S. in Electronics Engineering, Hanyang University

Feb. 1987: M.S. in Electrical Engineering, KAIST

- Feb. 1992: Ph.D. in Electrical Engineering, KAIST
- From Oct. 1987 to Present: Senior member of technical staff in Spoken Language Processing Section, ETRI
- Main Research Field: Speech Signal Processing, Speech Recognition, etc.
- Member of the Acoustical Society of Korea
- Member of the IEEE