AN ALGORITHM FOR AUTOMATIC HMM STRUCTURE GENERATION IN SPEECH RECOGNITION

Tetsuo Kosaka and Shigeki Sagayama

ATR Interpreting Telephony Research Laboratories

ABSTRACT - We discuss the number of mixture components in continuous mixture density HMM phone models (CHMMs) and present the principle of "distribution size uniformity," instead of the "mixture number uniformity" principle applied in conventional approaches. An algorithm is introduced for automatically determining the number of mixture components. The performance of this algorithm is shown through recognition experiments involving all Japanese phonemes.

INTRODUCTION

One of the major problems with a CHMM is how to determine its structure, e.g. the number of states, the transition and the number of mixture components (Takami, 1992. Kamp, 1985). This paper considers the number of mixture components in each state in the structure of CHMMs. In most conventional approaches to speech recognition that involve the CHMM, the number of mixture components, heuristically determined in advance, is uniformly given to all states. This, however, causes the frequent presence of extremely small variances of mixture components in the CHMMs, since different phones may have different distribution characteristics and different amounts of available training data. We discuss automatic rather than heuristic allocation of a given total number of output probability distributions among the CHMM states.

In this paper, we propose an algorithm for automatically determining the number of CHMM mixture components for each state, and the performance of this algorithm is shown through recognition experiments involving all Japanese phonemes.

INVESTIGATION ON THE INFLUENCE OF THE NUMBER OF MIXTURE COMPONENTS

When the number of mixture components is especially large, it is important to investigate the precision of the model estimation using an insufficient amount of training data.

To avoid this problem, the sizes of output distributions were analyzed. Figure 1 shows the relationship between the amount of training data and the distribution size of CHMM output for various numbers of mixture components. Each scatter point represents the value for one phoneme. We define the distribution size as the determinant of the covariance matrix of the output probability distribution. Distribution size s is defined by the following equation;

$$s = (1/Nm_n)\sum_{n=1}^{N}\sum_{m=1}^{m_n} \log|\mathrm{S}_{nm}|$$

where N is the number of states, m_n is the number of mixture components at state n and S_{nm} is the covariance matrix of the m-th mixture at state n. In this figure, the distribution size tends to decrease with the number of mixture components. In particular, the distribution size becomes very small for an insufficient amount of training samples. This is the reason for the occurrence of estimation errors when the training data is insufficient.

We investigate the estimation errors by using the HMM likelihood output for both training data and test data. Figure 2 shows the relationship between the number of mixture components and the likelihood

output for both training data and test data. The likelihood outputs are normalized by dividing by the likelihood output for one mixture. Thus, all lines start at 1.0. Though HMM likelihoods increase with increasing number of mixture components for the training data, not every output increases for the testing data. By comparing these results with the distribution size obtained previously we could make the following observation. If the size of the distribution is extremely small, the output value increases very much with the training data, but decreases with the test data. This indicates the occurrence of over-tuning of the CHMM to the samples in the training data.

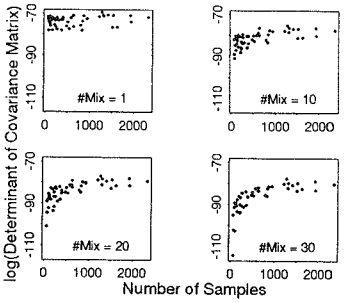

Figure 1: Relationship Between the Number of Training Samples and the Distribution Size

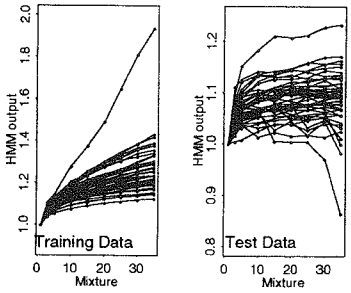

Figure 2: Relationship between the number of mixture components and the HMM log likelihood

PRINCIPLE

When the CHMM parameters are trained with insufficient data, the determinant of every covariance as a distribution size decreases a great deal. On the other hand, the CHMMs having a larger distribution size require a greater number of mixture components to represent them precisely. In consideration of the above reason, the principle we give here is that of uniformity of the output probability density distribution size rather than the conventional principle of the uniformity of the number of mixture components.

ALGORITHM

In this paper, we propose two ways to automatically determine the number of mixture components for a given total number of output probability distributions.

- The number of mixture components is the same within a phoneme but may be different among phonemes (proposed method I)

- Each state has a variable number of mixture components (proposed method II)

In this section, we describe an algorithm for proposed method II. Since the algorithm of proposed method I is very similar, it is omitted. To determine the number of mixture components, we made a number of conventional CHMM phone model sets, each with a different number of mixture components. Figure 3 shows the relationship between the number of mixture components and the distribution size for all states using the above results. The curves that result from this experiment are labelled $f_n^p(m_n^p)$, where p is the phoneme number, n is the state number and m is the number of mixture components. It was found that all these functions monotonously decrease with the number of mixture components. The reason why some lines drop significantly is that there is over-tuning due to training data insufficiency.

The algorithm for automatic determination of the number of mixture components for a given total number of output probability distributions using $f_n^p(m_n^p)$ on the figure is as follows;

The algorithm for automatic determination of the number of mixture components

Defined Symbols

P: the number of phonemes(=49) N: the number of states(=3)
M: the total temporary number of gaussian distributions
M_o: the total final number of gaussian distributions
m_n^p: the number of mixture components k: the number of iteration
s_k: distribution size at k α: learning constant
s_0: initial value for s_k

Algorithm

1. set the initial value

$$s_1 = s_0$$
$$k = 0$$

2. calculate the number of mixture components

$$k = k + 1$$
$$m_n^p = f_n^{p-1}(s_k)$$
$$M = \sum_{p=1}^{P}\sum_{n=1}^{N} m_n^p$$

3. renew s_k

if $M = M_o$ then goto STEP4.
else if $M < M_o$ then $s_{k+1} = s_k - \alpha$
else if $M > M_o$ then $s_{k+1} = s_k + \alpha$
goto STEP2.

4. re-estimation of CHMM with obtained number of mixture components

Since re-estimation of the CHMM parameters only commences after the numbers of mixture components have been finalized, this algorithm can work rapidly.

PHONEME RECOGNITION EXPERIMENTS

The above algorithm for a given total number of output probability distributions was tested in speaker independent phoneme recognition experiments using all 34 Japanese phonemes. The given average numbers of mixture components per state were 1, 3, 5 and 10 corresponding to M_o = 147, 441, 735 and 1470. In consideration of calculation costs, the maximum number of mixture components for a state was set as 35. We tested the algorithm in three different ways:

- All states have the same number of mixture components (conventional method)

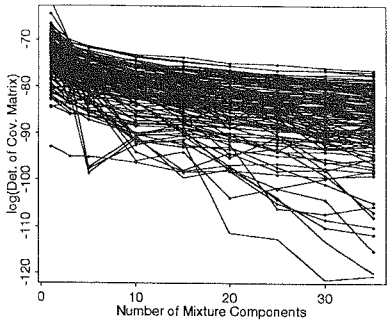

Figure 3: Relationship between the number of mixture components and the distribution size of HMM ($f_n^p(m_n^p)$)

- The number of mixture components is the same within a phoneme but may be different among phonemes (proposed method I)
- Each state has a variable number of mixture components (proposed method II)

Table 1 shows the experimental conditions. Diagonal-covariance 4-state 3-loop CHMMs were provided for these experiments. The structure of the model is shown in Figure 4. The phoneme CHMMs were trained with 736 isolated words uttered by 12 male speakers. The testing data was 216 isolated word utterances from 10 different speakers. The recognition rate was calculated by the following equation;

$$\text{Cor}(\%) = (N - S - D)/N \times 100(\%)$$

where N is the total number of testing samples, S is the number of substitutions and D is the number of deletions.

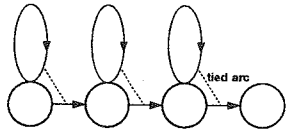

Figure 4: The Structure of HMM

DISCUSSION

The results are shown and compared in Table 2. The experimental results reveal that, for the same total number of output probability density distributions, having two types of variable number of mixture components are better performance-wise than a fixed number of mixture components.

Figure 5 shows the number of mixture components for each phoneme obtained with "proposed method I" where the average number of mixture components is 10. The number of training samples is also indicated. In stop /k/, vowels, nasals and /r/, large numbers of mixture components were chosen. These phonemes exhibit large spectrum variations among speakers. Contrary to these phonemes, the numbers of mixture components are small in the affricates and fricatives apart from /h/.

Table 1: Experimental Conditions

Analysis Conditions	Sampling freqency: 12 kHz Hamming window: 20 ms Pre-emphasis: 0.98 Analysis period: 5 ms 16th-LPC cepstrum + log power + 16th-Δcepstrum + Δlog power	
Speech Data	Training Data	12 male speakers, 736 isolated words 736 × 12 = 8,832 samples
	Testing Data	10 male speakers, 216 isolated words 216 × 10 = 2,160 samples
Phoneme Models	p1 p2 t1 t2 k1 k2 b1 b2 d1 d2 g1 ng m n N r w y s sh h z ch1 ch2 ts1 ts2 sy hy zy cy py ky by gy1 ngy my ny ry aa ii uu eei oou a i u e o silence (1 means word-top)	

There are two reasons why a small number of mixture components results. One is an insufficient number of training samples. The other is that training samples have large spectrum variations. For example, although /s/ and /sh/ are trained on sufficient data, the number of mixture components is still small.

Table 3 shows the number of mixture components for each state obtained with the "proposed method II" where the average number of mixture components is 5. The number of mixture components is large in the 1st and 3rd states for vowels where the spectrum variation is large because of the influence of phoneme context. For unvoiced plosives, the number of mixture components is small in the 1st state which seems to represent the closure segment. These results correspond well with observed speech phenomena.

CONCLUSIONS

We discussed the number of mixture components in CHMMs and presented the principle of "distribution size uniformity," instead of the "mixture number uniformity" principle applied in conventional approaches. An algorithm was introduced for automatically determining the number of mixture components. The performance of this algorithm was shown through recognition experiments involving all Japanese phonemes. We showed that the proposed two methods perform superior to the conventional method of a fixed number of mixture components per state.

Our future research topics include:

- automatic determination of the number of states for CHMMs,
- extension to continuous speech recognition by combination with the LR-parser (Kita, 1989).

REFERENCES

J.Takami et al. (1992): *A Successive State Splitting Algorithm for Efficient Allophone Modeling*, Proc. of ICASSP'92, 66.6.

Kita, K. et al. (1989): *HMM Continuous Speech Recognition Using Predictive LR Parsing*, Proc ICASSP89, pp. 703-706, Albuquerque.

Y.Kamp. (1985): *State Reduction in Hidden Markov Chains Used for Speech Recognition*, IEEE Trans. on ASSP, Vol.33, No.4, pp.1138-1145.

Table 2: Phoneme Recognition Rate (%)

ave. #mixtures = the total number of gaussian distributions/(the number of phonemes × the number of states)

ave. #mixtures	all alike (conventional method)	same within phoneme (proposed method I)	all variable (proposed method II)
1	62.46		
3	72.46	74.58	73.76
5	74.92	76.93	77.47
10	77.84	78.49	79.62

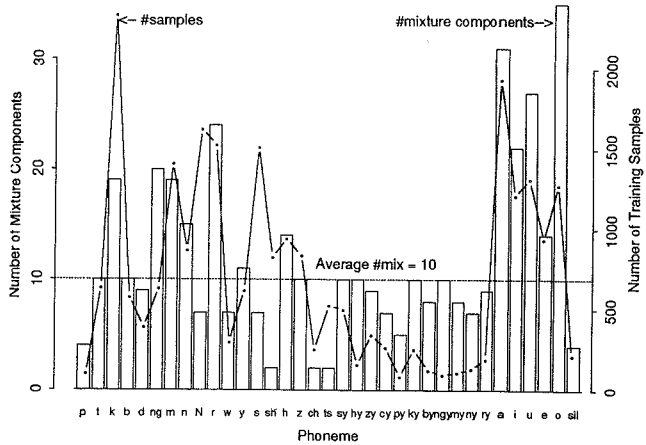

Figure 5: The Number of Mixture Components and The Number of Samples for Each Phoneme

Table 3: The Number of Mixture Components for Each State (mixture average: 5)

Phoneme	The Number of Mixture Components for Each State	Phoneme	The number of Mixture Components for Each State
p2	1 - 2 - 7	y	5 - 5 - 4
t2	1 - 1 - 3	s	2 - 1 - 12
k2	3 - 3 - 10	sh	2 - 1 - 3
b2	3 - 3 - 13	h	5 - 2 - 11
d2	4 - 2 - 4	z	5 - 1 - 5
g1	1 - 4 - 8	ch2	1 - 1 - 2
ng	11 - 7 - 11	ts2	1 - 1 - 3
m	7 - 5 - 14	a	10 - 4 - 23
n	7 - 6 - 13	i	9 - 3 - 17
N	3 - 1 - 1	u	14 - 4 - 19
r	7 - 8 - 11	e	9 - 1 - 15
w	5 - 3 - 2	o	23 - 3 - 31