Continuous Speech Recognition Using Tree based State Tying

Sung-Ill Kim , Tetsuro Kitazoe
Department of Computer Science and Systems Engineering
Faculty of Engineering , Miyazaki University
1-1 , Gakuen Kibanadai Nishi , Miyazaki , 889-21 Japan

Abstract For large vocabulary speech recognition using HMMs, context-dependent subword units have been often employed . However, when context-dependent phone models are used, they result in a system which has too many parameters to train . The problem of too many parameters and too little training data is absolutely crucial in the design of a statistical speech recogniser . Furthermore, when building large vocabulary speech recognition systems, unseen triphone problem is unavoidable . In this paper, we analyze the phonetic decision trees which have advantages solving these problems through following two experiments in Japanese contexts . The baseline experimental results show that phonetic decision trees are effective for clustering and reducing the number of states without any degradation in performance . The task experimental results show that the decision tree clustering also have the advantage of providing a mapping for unseen triphones .

key words speech recognition, Hidden Markov Model, context-dependent model, phonetic decision tree

音素決定木を用いた連続音声認識

金星一 北添徹郎
宮崎大学工学部情報工学科
〒889-21、宮崎市学園木花台西1-1

あらまし 隠れマルコフモデル(HMM)を用いた大語彙音声認識にとって、文脈依存サブワードを単位とする方法がしばしば用いられてきた。 しかし、文脈依存音素モデルが用いられる時、学習する場合にはパラメータの数が多すぎる結果を招来する。 多すぎるパラメータと少なすぎる学習データの問題は統計的音声認識システムの実現にとって難しい問題である。 その上、大語彙音声認識システムを構築する時、未出現文脈依存音素モデルの問題は避けない。 本報告では、二つの実験を通じてこういう問題を解決する長所を持つ音素決定木を日本語の文章に適用して分析する。 基本実験は音素決定木がクラスタリングに効果的であり、認識率を低下することなく状態数が減らせることを証明する。 タスク実験では音素決定木がまた未出現文脈依存音素モデルの割り当てにも長所を持つことを証明する。

キーワード 音声認識、隠れマルコフモデル、コンテキスト依存モデル、音素決定木

1. INTRODUCTION

To maximise the performance of a Hidden Markov Model(HMM) based recogniser it is necessary to strike a balance between the level of detail of the models (controlled by the number of parameters in the system) and the ability to accurately estimate those parameters from the data . For large vocabulary speech recognition, we will never have sufficient training data to model all the various acoustic-phonetic phenomena . Since the data is usually unevenly spread, we need to employ some method to balance model complexity against data availability . The use of Gaussian mixture output distributions allows each state distribution to be modeled very accurately . However, when triphones are used they result in a system which has too many parameters to train . The problem of too many parameters and too little training data is crucial in the design of a statistical speech recogniser . Furthermore, when building large vocabulary speech recognition systems unseen triphones are unavoidable . This is vital when producing cross word context dependent system as the majority of contexts appear very few, if any, times . The ability to produce models for unseen contexts makes it easy to produce systems incorporating cross word triphone models .

Traditional methods of dealing with these problems involve sharing models across differing contexts to form so-called generalised triphones and using a posteriori smoothing techniques[1]. However, model -based sharing is limited in that the left and right contexts cannot be treated independently and hence this inevitably leads to sub-optimal use of the available data . A posteriori smoothing is similarly unsatisfactory in that the models used for smoothing triphones are typically biphones and monophones, and these will be rather too broad when large training sets are used . Furthermore, the need to have cross-validation data unnecessarily complicates the training process .

The important aspects of triphone modeling using a limited training data set are how to tie the model parameters and how to handle the unknown contexts . Many research papers have shown that phonetic decision tree provided a similar quality of clustering but offered a solution to the unseen triphone problem[2] . In this paper, we demonstrate its validity in Japanese contexts . Our new system is based on the use of phonetic decision tree which determines contextually equivalent sets of HMM states using classification rules of Japanese phone set[9] . Furthermore, we study this tree-based clustering leads to have the additional advantage of providing a mapping for unseen triphones in Japanese task contexts .

In the next section, the phonetic decision tree based method is described . And experimental results are presented in section 3 for both the baseline CSR(Continuous Speech Recognition) experiments and the paper contribution inquiries task CSR experiments . Finally, section 4 presents our conclusions from this work .

2. TREE-BASED CLUSTERING

A phonetic decision tree[3,4,5,6] is a binary tree in which a question is attached to each node . In the system described here, each of these questions relates to the phonetic context to the immediate left or right . Trees are built using a top-down sequential optimisation process[5,7,8] . Initially, all corresponding HMM states of all allophonic variants of each basic phone are tied to form a single pool . Phonetic questions are then used to partition the pool into subsets in a way which maximises the likelihood of the

training data . The leaf nodes of each tree determine the sets of state tyings for each of the allophonic variants .

For example, in the decision tree shown in figure 1, the root question is answered by checking to see if the immediately preceding phone (the left context) is a vowel (a,aa,i,ii,u,uu,e,ee,o,oo) . If the actual context was aa-t+o in the word 'デパート' the next question to be asked would concern whether the following phone was a plosive (b,p,t) . Since t is not a member of this set and the answer no results in a terminal node, the model labeled C would be used in this context .

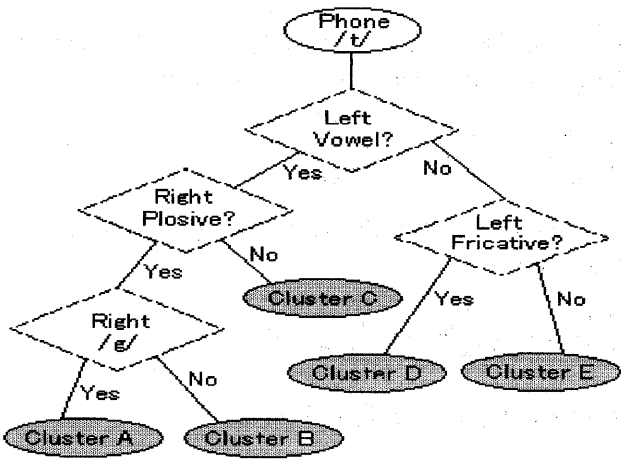

Figure 1 . Example of a phonetic decision tree

Splitting any pool into two will increase the log likelihood since it provides twice as many parameters to model the same amount of data . The increase obtained when each possible question is used can thus be calculated and the question is selected which gives the biggest improvement . This process is repeated until the increase in log likelihood falls below the threshold . As a final stage, the decrease in log likelihood is calculated for merging terminal nodes with differing parents . Any pair of nodes for which this decrease is less than the threshold used to stop splitting are then merged .

The next is the approximate log likelihood of a set models comprising the set of distributions S generating the training data O consisting of E examples .

$$L = \sum_{e=1}^{E} \sum_{t=1}^{T_e} \sum_{s \in S} \ln(\Pr(o_t^e; \mu_s, \Sigma_s)) \gamma_s^e(t) \tag{1}$$

For simple Gaussian distributions

$$L = \sum_{e=1}^{E} \sum_{t=1}^{T_e} \sum_{s \in S} -\frac{1}{2} \left(n \ln(2\pi) + \ln(|\Sigma_s|) + (o_t^e - \mu_s)' \Sigma_s^{-1} (o_t^e - \mu_s) \right) \gamma_s^e(t) \tag{2}$$

And using the parameter reestimation formula of Σ s

$$\sum_{e=1}^{E} \sum_{t=1}^{T_e} (o_t^e - \mu_s)' \Sigma_s^{-1} (o_t^e - \mu_s) \gamma_s^e(t) = n \sum_{e=1}^{E} \sum_{t=1}^{T_e} \gamma_s^e(t) \tag{3}$$

This gives

$$L=\sum_{s\in S}-\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_s|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_s^e(t) \tag{4}$$

Splitting a node changes the set of distributions S by replacing the parent p distribution with a set of descendants D . The total likelihood in this case is given by

$$L=-\sum_{s\in S,s\neq p}\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_s|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_s^e(t)-\sum_{d\in D}\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_d|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t) \tag{5}$$

So the change in overall log likelihood, which is the quantity that needs to be maximised, is just the difference between the likelihood of the parent and its descendants .

$$\delta L=-\sum_{d\in D}\frac{1}{2}\ln(|\Sigma_d|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t)+\frac{1}{2}\ln(|\Sigma_p|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_p^e(t) \tag{6}$$

A similar expression can be used to find the change in likelihood when a set of distributions D are merged to produce a single distribution m .

$$\delta L=-\frac{1}{2}\ln(|\Sigma_m|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_m^e(t)+\sum_{d\in D}\frac{1}{2}\ln(|\Sigma_d|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t) \tag{7}$$

This algorithm is summarised diagrammatically in figure 2 .

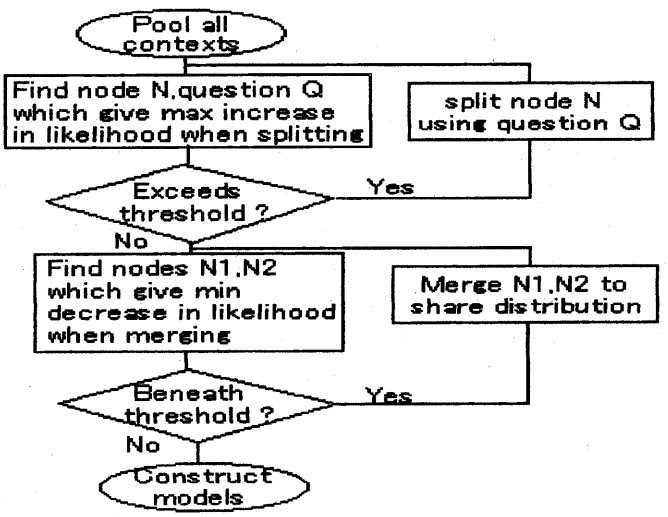

Figure 2 . Algorithm for constructing decision trees

Using a top down clustering procedure based on decision trees avoids the problem of unseen models by using linguistic knowledge together with the training data to decide which contexts (including the unseen ones) are acoustically similar . Once all such trees have been constructed, unseen triphones can

be synthesized by finding the appropriate terminal tree nodes for that triphone's contexts and then using the tied-states associated with those nodes to construct the triphone .

3 . EXPERIMENTS AND RESULTS

3 . 1 SPEECH DATABASE

We trained triphone models in following steps .

1) The initial monophone models are estimated using 5240 labeled word utterances of 10 male speakers in the ATR Japanese speech database(set A) .

2) The monophone models are re-estimated using ATR 503 labeled sentences of 6 male speakers(set B) .

3) The triphone models are re-estimated again using the databases of 2) and ASJ(Acoustical Society of Japan) 150 sentences of 26 male speakers .

In the baseline CSR experiments, 100 sentences of another ASJ 3 male speakers is used for the baseline test . And in the task CSR experiments, 115 sentences of ATR Japanese dialog database which is related with paper contribution inquiries topics is used for the test of this task .

3 . 2 EXPERIMENTAL CONDITIONS

The experimental conditions are listed in Table 1 .

Table 1 . Experimental conditions

sampling rate	16kHz , 16 bit
preemphasis	0.97
window function	25 ms Hamming window
frame period	10 ms
feature parameters	12-order LPC-Cepstrum +△ LPC-Cepstrum + △△ LPC-Cepstrum + log power +△ log power + △△ log power (total 39-order)
model topology	3-state left-right triphone model

3 . 3 BASELINE CSR EXPERIMENTS

Basically, the recognizer was based on the total number of 3158 triphones with the 9474 states of 1 mixtures per state . And the phonetic decision rule was used at several thresholds to generate tied-states from 924 to 3814 states . The tests were done using both the no grammar(nogram) and bigram syntaxes . The word recognition accuracies for these CSR systems are shown in table 2 .

Table 2. Word recognition accuracy(%) dependent on the variation of a threshold

Thresh-old	Number Of Tied-states (Reduction Rate(%))	TSU0001		TSU0003		FUJ0001		TOTAL	
		nogram	bigram	nogram	bigram	nogram	bigram	nogram	bigram
none	9474 (0)	91.9	97.2	88.35	93.3	85.7	89.9	88.7	93.5
300	3814 (59.7)	91.6	93.4	88.2	92.6	79.2	88.8	86.3	91.6
600	2429 (74.4)	91.0	97.4	88.2	92.6	79.2	88.7	86.1	92.9
900	1802 (81.0)	91.8	97.5	88.5	92.4	82.9	88.8	87.7	92.9
1200	1493 (84.2)	95.3	97.5	94.6	93.5	83.7	88.7	91.2	93.2
1500	1270 (86.6)	94.9	96.9	93.6	92.1	83.2	88.5	90.6	92.5
1800	1120 (88.2)	95.0	97.2	92.7	92.2	85.1	89.6	90.9	93.0
2100	1005 (89.4)	92.2	95.8	91.6	91.9	80.0	87.1	87.9	91.6
2400	924 (90.2)	91.0	95.8	91.5	92.4	80.1	86.7	87.5	91.6

3.4 TASK CSR EXPERIMENTS

The experiments on decision tree clustered models were performed on the ATR paper contribution inquiries dialog task. In this experiments, 1493 tied-state distributions were used for the decision tree based system at the threshold 1200 which showed a relatively good recognition accuracy in the baseline experiments. The word recognition accuracies for this context dependent state-tied triphone system are shown in table 3 together with the context independent monophone system. The multiple mixture models for the state-tied triphone systems were built, and at each stage recognition experiments were performed as shown in table 4.

Table 3. Word recognition accuracy(%) for state-tied triphone system with threshold 1200 in comparison with monophone system

System	MAU		MHT		MXM		TOTAL	
	nogram	bigram	nogram	bigram	nogram	bigram	nogram	bigram
Monophone	62.6	70.1	68.3	73.7	70.5	73.2	67.1	72.3
Tied-state triphone	83.6	87.1	88.2	90.1	86.1	89.2	86.0	88.8

Table 4. Word recognition accuracy(%) dependent on the variation of a number of mixture

Number Of Mixture	MAU		MHT		MXM		TOTAL	
	nogram	bigram	nogram	bigram	nogram	bigram	nogram	bigram
1	83.6	87.1	88.2	90.1	86.1	89.2	86.0	88.8
2	86.1	88.7	90.4	94.7	89.0	91.7	88.5	91.7
3	84.8	90.5	90.3	94.9	88.7	92.2	88.0	92.5
4	85.0	90.5	91.1	95.7	88.3	92.2	88.1	92.8

3.5 RESULTS

In the baseline CSR experiments, it is important to tune thretholds because the value of threshold affects the degree of tying and the number of output states in the clustered system as figure 2 shows . Though the performance was relatively flat for a large range of threshold as shown in table 2, the phonetic decision tree remarkably reduces the number of states without any degradation in performance compared with the performance based on 9474 untied states .

In the task CSR experiments, 501 new vocabularies were included among total of 543 vocabulary items in the ATR paper contribution inquiries task . 150 new unseen triphones were also required among total of 1019 triphones in the pronunciation lexicon of task vocabulary item . In these vocabulary independent recognition experiments, it could be seen from table 3 that the decision tree clustering triphone system had 18.9 % improvement using nogram and 16.5 % improvement using bigram in reconition accuracy relative to the monophone system . Furthermore, the results in table 4 also showed that tied-state triphone system ,when the number of mixture was increased to 4, had 21.0 % improvement using nogram and 20.5 % improvement using bigram in reconition accuracy relative to the monophone system . The analysis results of the task experiments are given in Table 5 in comparison with the baseline ones . This shows us that the decision tree clustering triphone system allows previously unseen triphones to be synthesized and has better recognition rates than the monophone system in the task CSR experiments .

Table 5 . Analysis results of comparison between two CSR experiments

Type Of Experiments		Number Of New Vocabulary (Word Coverage(%))	Number Of Unseen Recognition Units (Recognition Units Coverage (%))	Word Recognition Rates Using Bigram
Baseline CSR Experiments		0 (100)	0 (100)	93.2
Task CSR Experiments	Monophone	501 (7.7)	0 (100)	72.3
	Tied-state triphone	501 (7.7)	150 (85.3)	88.8

4. CONCLUSIONS

This paper has described an efficient method of state clustering based on the use of phonetic decision trees when we applied to Japanese contexts . The important triphone modeling using a limited training data set are how to tie the model parameters and how to handle the unknown contexts .

In the baseline CSR experimental results, it has been shown that the decision tree clustering method using classification rules of Japanese phone sets was effective for both clustering and reduction of parameters . This phonetic decision tree remarkably reduced the number of states without any degradation in performance . In the task CSR experimental results, it has been shown that the decision tree clustering also had the advantage of providing a mapping for unseen triphones in Japanese task contexts .

When building a large vocabulary cross-word triphone systems unseen triphones are unavoidable . But the phonetic decision tree offers a solution to the unseen triphone problem . Since cross-word tripho-

nes explicitly model co-articulation effects across word boundaries, they provide a more consistent and accurate representation of speech and thus produce more accurate results . Therefore we will conduct a large vocabulary speech recognition experiments using cross-word triphones in the future .

REFERENCES

1 . Lee K-F (1989) . Automatic Speech Recognition : The Development of the SPHINX system . Kluwer Academic Publishers, Boston .

2 . Hwang M-Y, Huang X, Alleva F (1993) . Prediction Unseen Triphones with Senones . Proc ICASSP'93, Vol , pp. 311-314, Minneapolis .

3 . Bahl LR, de Souza PV, Gopalakrishnan PS , Nahamoo D , Picheny MA (1991) . Context Dependent Modeling of Phones in continuous Speech Using Decision Trees . Proc DARPA Speech and Natural Language Processing Workshop, pp. 264 - 270, Pacific Grove, Calif .

4 . Downey S, Russell MJ (1992) . A Decision Tree Approach to Task Independent Speech Recognition . Proc 6, pp. 181 - 188 .

5 . Odell JJ . (1992) . The Use of Decision Trees with Context Sensitive Phonetic Modeling . MPhil Thesis, Cambridge University Engineering Department .

6 . 堀貴明、加藤正治、伊藤彰則、好田正紀 (1996) . 音素決定木に基づく逐次状態分割法によるHM-Netの性能改善の検討、情処研報Vol.96,No.123, pp83-90

7 . Kannan A, Ostendorf M, Rohlicek JR (1994) . Maximum Likelihood Clustering of Gaussians for Speech Recognition . IEEE Trans on Speech and Audio Processing .

8 . S.Young, J.Jansen, J.Odell, D.Ollason, P.Woodland (1995) . The HTK BOOK .

9 . < Phonetic questions used in the phonetic decision trees >

Features		Phones	Features		Phones
Silence		SIL,sp	Consonant	Plosive	b,by,d,dy,g,gy,k,ky,p,py,t
Voiced		a,aa,i,ii,u,uu,e,ee,o,oo,w,y,r,z,j,b,by,d,dy,g,gy,NG,m,my,n,ny		Nasals	NG,m,my,n,ny
				Front-Consonant	b,by,f,m,my,p,py,w
				Central-Consonant	NG,d,n,ny,r,ry,s,t,ts,z
				Back-Consonant	ch,g,gy,j,k,sh,y
Vowel	Vowel	a,aa,i,ii,u,uu,e,ee,o,oo		Glottis-Consonant	h,hy
	Long-Vowel	aa,ii,uu,ee,oo		Voiced-Fricative	j,z
	Short-Vowel	a,i,u,e,o		Unvoiced-Fricative	f,h,hy,s,sh
	Front_Vowel	e,ee,i,ii		Voiced-Affricative	b,by,d,dy,g,gy
	Back-Vowel	a,aa,o,oo,u,uu		Unvoiced-Affricative	k,p,t
	Narrow-Vowel	i,ii,u,uu		Glides	r,ry,w,y
	Half-Vowel	e,ee,o,oo		Fricative	f,h,hy,j,s,sh,z
	Wide-Vowel	a,aa		Affricates	ch,ts