Modified Phonetic Decision Tree For Continuous Speech Recognition

*Sung-Ill Kim , *Tetsuro Kitazoe and **Hyun-Yeol Chung

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 recognizer. Furthermore, when building large vocabulary speech recognition systems, unseen triphone problem is unavoidable. In this paper, we propose the modified phonetic decision tree algorithm for the automatic prediction of unseen triphones which has advantages solving these problems through following two experiments in Japanese contexts. The baseline experimental results show that the modified tree based clustering algorithm is effective for clustering and reducing the number of states without any degradation in performance. The task experimental results show that our proposed algorithm also has the advantage of providing a automatic prediction of unseen triphones.

I. Introduction

For large vocabulary speech recognition, we will never have sufficient training data to model all the various acoustic-phonetic phenomena. For instance, 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 recognizer. Furthermore, when building large vocabulary speech recognition systems unseen triphones are unavoidable[1]. These new triphones are often encountered during testing due to the limited amount of training data. 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 and to construct a large vocabulary speech recognition system.

There are several ways in which the trainability of a system can be increased including backing-off, smoothing[2],[3] and sharing[4]. Among these methods, a wide variety of sharing techniques using bottom up and top down approaches have been proposed. However one limitation of the bottom up approach is that it does not deal with triphones for which there are no examples in the training data. This problem can be minimized by ensuring that the training data gives adequate coverage of the models needed for recognition. This is possible only for small vocabulary systems. For large vocabulary and cross word context dependent systems, it is, however, virtually impossible to ensure that the training data will include examples of every possible context. Using a top down clustering procedure based on decision trees[5] avoids the problem of unseen triphones by using linguistic knowledge.

In this paper, we propose the modified phonetic decision tree for automatic prediction of unseen triphones[6]-[9]. Our new system is based on the modified phonetic decision tree which determines contextually equivalent sets of HMM states using classification rules of Japanese phone set. Furthermore, we will examine the workability of the modified tree based tied-state triphone system by comparing with the bottom up clustering system.

In the next section, the algorithm for automatic prediction of unseen triphones using phonetic decision tree is described. In section 3, experimental results are presented for both the baseline continuous speech recognition(CSR) experiments and the paper contribution inquiries task CSR experiments. Finally, section 4 presents our conclusions from this work.

*Department of Computer Science and Systems Engineering, Miyazaki University
**Department of information and communication engineering, Yeungnam University.

Manuscript Received : September 3, 1998.

Ⅱ. Prediction of Unseen Triphones

A phonetic decision tree is a binary tree in which a question is attached to each node. Trees are built using a top-down sequential optimization process[10]. In the system described here, each of these questions relates to the phonetic context to the immediate left or right. The system uses separate decision trees for each phone state. Thus, for example, a system with 40 phones and 3 states per phone would have 120 separate trees. 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 maximizes 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 'depato'. 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.

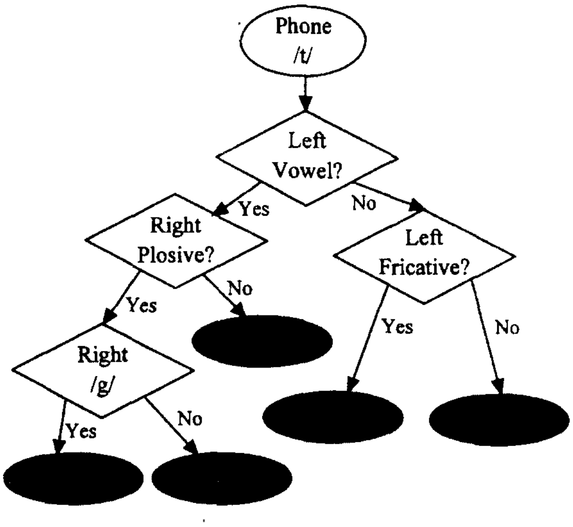

Figure 1. Example of a phonetic decision tree.

The prototype triphone model sets only include those needed to cover the training data. However for a large vocabulary speech recognition system, there are many contexts that we have not seen that can occur in our recognition networks. Therefore rather than actually find out which models are needed, it is easier to generate all possible biphones and triphones and this would also allow us to work with new vocabulary in a task. The effect including all phone lists is to use the decision trees to synthesize all of the new previously unseen triphones.

Splitting any pool into two will increase the log likelihood since it provides twice as many parameters to model the same amount of data. To reduce these parameters, phonetic decision trees is used. First, we can select some question which gives the biggest log likelihood. In the case of incresing log likelihood, the two parts of output probability don' t resemble. Therefore, when the increase of log likelihood show the biggest value, the output probability will be separated. 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.

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

For simple Gaussian distributions

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

And using the parameter reestimation formula of Σ_s

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

So

$$\sum_{e=1}^{E}\sum_{t=1}^{T_e}(o_t^e-\mu_s)'\Sigma_s^{-1}(o_t^e-\mu_s)\gamma_s^e(t)=n\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_s^e(t) \qquad (4)$$

This give

$$L=\sum_{s\in S}-\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_s|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_s^e(t) \qquad (5)$$

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

$$L = \sum_{s\in S, s\neq p} -\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_s|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t)$$
$$-\sum_{d\in D}\frac{1}{2}(n(1+\ln(2\pi))+\ln(|\Sigma_d|))\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t) \quad (6)$$

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

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

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.

$$\delta L = -\frac{1}{2}\ln(|\Sigma_m|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_m^e(t)+\sum_{d\in D}\frac{1}{2}\ln(|\Sigma_d|)\sum_{e=1}^{E}\sum_{t=1}^{T_e}\gamma_d^e(t) \quad (8)$$

This algorithm is summarized diagrammatically in figure 2.

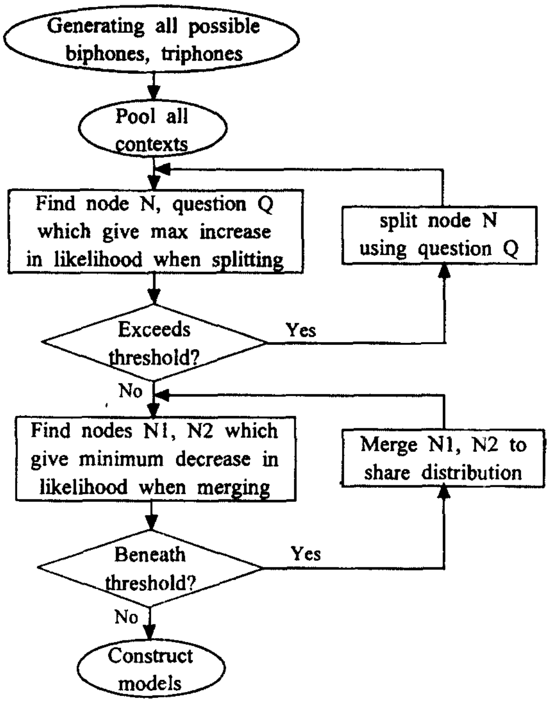
Generating all possible biphones, triphones
Pool all contexts
Find node N, question Q which give max increase in likelihood when splitting
split node N using question Q
Exceeds threshold?
Yes
No
Find nodes N1, N2 which give minimum decrease in likelihood when merging
Merge N1, N2 to share distribution
Beneath threshold?
Yes
No
Construct models

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.

Ⅲ. 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

We obtained a set of 39 dimensional observation sequences for recognition. Table 1 shows the preprocessing analysis condition of the speech data.

Table 1. Analysis of speech signal.

sampling rate	16 kHz, 16 bits
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-to-right triphone model

The questions used to construct the decision tree are chosen to incorporate linguistic knowledge into the clustering procedure by ensuring that unseen contexts are grouped with those which one would expect to be linguistically similar. Table 2 shows the set of phones for each phonetic feature depends upon the Japanese phone set.

Table 2. Phonetic questions used in the phonetic decision tree.

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

3.3 BASELINE CSR EXPERIMENTS

Basically, the system was based on the total number of 3158 triphones with the 9474 states of 1 mixtures per state. The phonetic decision rule was used at several thresholds to generate tied-states from 924 to 3814 states as table 3 showed. And table 4 illustrates the word recognition accuracies in terms of the comparison of three triphone systems in which two tied-state triphone systems have approximately 1400 tied-states. In this experiment, 1493 tree based tied-state distributions were used at the threshold 1200 which showed a relatively good recognition accuracies as indicated in table 3. The bottom up and tree based tied-state triphone system used the same initial set of 9474 untied-state triphones. The cluster thresholds in each case were adjusted to obtain systems with approximately equal numbers of states, 1462 and 1493, respectively. The tests were done using both nogram and bigram syntaxes and the word recognition accuracies were calculated with total average of 3 speakers.

Table 3. Word recognition accuracies(%) in terms of the variation of a threshold.

Threshold	Number of States (Reduction Rate(%))	Word Recognition Rate(%)	
		Nogram	Bigram
none	9474 (0)	88.7	93.5
300	3814(59.7)	86.3	91.6
600	2429(74.4)	86.1	92.9
900	1802(81.0)	87.7	92.9
1200	1493(84.2)	91.2	93.2
1500	1270(86.6)	90.6	92.5
1800	1120(88.2)	90.9	93.0
2100	1005(89.4)	87.9	91.6
2400	924(90.2)	87.5	91.6

Table 4. Comparison of three triphone systems.

Triphone System	Number of States	Word Recognition Rate(%)	
		Nogram	Bigram
Untied-State	9474	88.7	93.5
Bottom Up Tied-State	1462	89.9	91.8
Modified Tree Based Tied-State	1493	91.2	93.2

3.4 TASK CSR EXPERIMENTS

The task CSR experiments were performed on the ATR paper contribution inquiries dialog task. The table 5 illustrates the word recognition accuracies in terms of the comparison of monophone vs. two tied-state triphone systems. The 501 new vocabularies were included among total of 543 vocabulary items in the ATR paper contribution inquiries task. The 150 new unseen triphones were also required among total of 1019 triphones in the pronunciation lexicon of task vocabulary item. In this experiments, since the bottom up tied-state triphone system provides no easy way of handling unseen triphones, we use monophones which overcome this problem and represent unseen triphones. As indicated in table 6, the multiple mixture models for the tree based tied-state triphone system were built, and recognition experiments were performed at each stage.

Table 5. Comparison of monophone vs. two tied-state triphone systems.

Triphone System	Number of States (Number of New Vocabulary)	Word Recognition Rate(%)	
		Nogram	Bigram
Untied-State	0 (501)	67.1	72.3
Bottom Up Tied-State + Monophones	0 (501)	77.3	80.2
Modified Tree Based Tied-State + Unseen Triphones	150(501)	86.0	88.8

Table 6. Word recognition accuracies(%) in terms of the variation of a number of mixture.

Number of States (Number of New Vocabulary)	Word Recognition Rate(%)	
	Nogram	Bigram
1	86.0	88.8
2	88.5	91.7
3	88.0	92.5
4	88.1	92.8

3.5 RESULTS

In the baseline CSR experiments, it is important to tune thresholds because the value of threshold affects the degree of tying and the number of output states in the clustered system. Though the performance is relatively flat for a large range of threshold as shown in table 3, the tree based tied-state triphone system remarkably reduces the number of states without any degradation in performance compared with the performance based on 9474 untied-state triphone system. As indicated in table 4, the performance of the modified tree based triphone system is slightly higher than that of the bottom up system but the modified tree based system has the advantage that they would allow the prediction of unseen triphones automatically.

In the task CSR experiments, it can be seen from table 5 that the tree based tied-state triphone system has 18.9 % improvement using nogram and 16.5 % improvement using bigram in recognition accuracies relative to the monophone system. In comparison with the bottom up clustering system, the tree based clustering system has 8.7 % improvement using nogram and 8.6 % improvement using bigram in recognition accuracies. Furthermore, the results in table 6 also show that the tree based clustering system, when the number of mixture is increased to 4, has 21.0% improvement using nogram and 20.5% improvement using bigram in recognition accuracies relative to the monophone system. This shows us that the modified tree based clustering system allows previously unseen triphones to be synthesized and has better recognition rates than the monophone and bottom up clustering system in the new task.

Ⅳ. Conclusions

The important triphone modeling issues for large vocabulary speech recognition with a limited training data set are how to tie the model parameters and how to handle the unknown contexts. This paper has described an efficient algorithm of automatic prediction of unseen triphones based on the modified phonetic decision tree and classification rules of Japanese phone set when applied to Japanese contexts.

In the baseline CSR experimental results, it is shown that our proposed tree based clustering algorithm is effective for both clustering and reduction of parameters. The tree based clustering system remarkably reduces the number of states without any degradation in performance. In the task CSR experimental results, it is shown that our proposed tree based clustering algorithm also has the advantage of providing a mapping for unseen triphones in Japanese task contexts.

When building a large vocabulary speech recognition systems, unseen triphones are unavoidable. Our experiments suggest that the proposed modified tree based algorithm offers one of the solutions to the unseen triphone problem. This enables us to construct the large vocabulary speech recognition system using context-dependent models easily because the strength of the our proposed tree based clustering algorithm lies in its unseen triphone modeling.

References

1. Hwang M-Y, Huang X., Alleva F., "Prediction unseen triphones with senones," *Proc. ICASSP*, Vol , pp.311-314, Minneapolis, 1993.
2. Lee K-F., "Automatic speech recognition : The development of the SPHINX system," Kluwer Academic Publishers, Boston, 1989.
3. Lee K-F., Hon H-W., "Speaker-independent phone recognition using hidden markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, Vol.37, No 11, pp.1641-1648, 1989.
4. S. J. Young, "The general use of tying in phoneme-based HMM speech recognizers," *Proc. ICASSP*, San Francisco, pp.569-572, 1992.
5. S. J. Young, j. j. Odell, and P. C. Woodland, "Tree based state tying for high accuracy acoustic modeling," *Proc. ARPA Workshop on Human Language Technology*, Princeton, NJ, Morgan Kaufmann Publishers, March, 1994.
6. Bahl L. R., de Souza P. V., Gopalakrishnan P. S., Nahamoo D., Picheny M. A., "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, 1991.

7. Downey S., Russell M. J., "A decision tree approach to task independent speech recognition," *Proc 6*, pp.181-188, 1992.
8. Odell J. J., "The use of decision trees with context sensitive phonetic modeling," *MPhil Thesis*, Cambridge University Engineering Department, 1992.
9. T. Hori, M. Katoh, A. Ito, and M. Kohda, "A study on Improvement of HM-Nets using decision tree-based successive state splitting (in Japanese)," *SLP* Vol. 96, No. 123, pp.83-90, Japan, 1996.
10. Kannan A, Ostendorf M, Rohlicek J. R., "Maximum likelihood clustering of gaussians for speech recognition," *IEEE Trans on Speech and Audio Processing, 1994.*

▲Sung-Ill Kim

Sung-Ill Kim was born in Taegu, Korea, 1968. He received his B.C. and M.S. degrees in the department of electronics engineering from Yeungnam University, in 1997. He currently working on his Dr. degree in Graduate school of Engineering in Miyazaki University.

His reseach interests are speech and image recognition by neural network, and emotion recognition.

▲Tetsuro Kitazoe

Tetsuro Kitazoe was born in Japan, 1937. He received his Dr. science degree in the department of Physics, Osaka University in 1966. He was an assistant professor from 1966 to 1972 and an associate professor from 1972 to 1991 at department of Physics, Kobe University. He moved to Faculty of Engineering, department of Computer Science and Systems Engineering, Miyazaki University as a professor.

His current reseach interests are image recognition by neural nets equation, speech recognition and quantum cosmology at the early stage of universe. He is a member of IPSJ, RSJ and PSJ.

▲Hyun-Yeol Chung

Hyun-Yeol Chung was born in Kyungnam, Korea, on November 26, 1951. He received his B.C. and M.S. degree in the department of electronics engineering from Yeungnam University, in 1975 and 1981, respectively, and the Ph.D of engineering degree in Infromation Sciences from Tohoku University, Japan, in 1989. He was a professor from 1989 to 1997 at school of electrical and electronic engineering, Yeungnam University. Since 1998 he is a professor in the department of information and communication engineering, Yeungnam University. During 1992 to 1993, he was a visiting scientist in the department of computer science, Carnegie Mellon University, Pittsburgh, USA. He was a visiting scientist in the department of information and computer sciences, Toyohashi University of Technology, Toyohashi, Japan in 1994. His research interests are speech analysis, speech recognition, multimedia and digital signal processing application.