Effective Acoustic Model Clustering via Decision Tree with Supervised Decision Tree Learning

Junho Park* · Hanseok Ko*

ABSTRACT

In the acoustic modeling for large vocabulary speech recognition, a sparse data problem caused by a huge number of context-dependent (CD) models usually leads the estimated models to being unreliable. In this paper, we develop a new clustering method based on the C45 decision-tree learning algorithm that effectively encapsulates the CD modeling. The proposed scheme essentially constructs a supervised decision rule and applies over the pre-clustered triphones using the C45 algorithm, which is known to effectively search through the attributes of the training instances and extract the attribute that best separates the given examples. In particular, the data driven method is used as a clustering algorithm while its result is used as the learning target of the C45 algorithm. This scheme has been shown to be effective particularly over the database of low unknown-context ratio in terms of recognition performance. For speaker-independent, task-independent continuous speech recognition task, the proposed method reduced the percent accuracy WER by 3.93% compared to the existing rule-based methods.

Keywords: Acoustic Model, HMM Clustering, Supervised Learning

1. Introduction

In the large vocabulary continuous speech recognition (LVCSR) system, efficient acoustic modeling is one of the most significant issues. Since the prevalent approach for accomplishing speech recognition tasks is by using statistical pattern matching, the challenge is how we can format the exact reference patterns so that classifying observations into discernable classes can be rapidly achieved with consistency. As a result, several acoustic modeling methods involving clustering have been vigorously pursued. However, as we attempt to reach the state-of-the-art in automatic speech recognition, the vocabularies and tasks also become more complex and must accommodate context variations. In such conditions, it becomes a formidable task to represent all of the co-articulation effects from varying contexts by using the context-independent acoustic modeling method alone. This is, in fact, the main reason

* Dept. of Electronics and Computer Engineering, Korea University

that the context-independent acoustic model based system performs poorly in the LVCSR tasks. To cope with this difficulty, the context-dependent acoustic modeling method that represents the variability of context in continuous speech utterances has been actively investigated. The context-dependent model changes into a different model if each acoustic model has different adjacent contexts. That is, context-dependent acoustic models enable more detail modeling and reflect co-articulation effect more effectively than context-independent acoustic models in LVCSR systems[8][9][10][11]. However, in order to train the context-dependent models representing varying contexts reliably, we need a huge amount of training data set, and it should contain all possible contexts in continuous speech utterances. In fact, it is difficult to train context-dependent models reliably because insuring the availability of such a training data set is too difficult to be realized. For that reason, a reliable parameter estimation of context-dependent acoustic models is impractical.

To solve the parameter estimation problem when training data is sparse, the state parameter or mixture tying methods for clustering similar acoustic models were introduced[1][3]. These methods were investigated with the goal of effectively reducing the acoustic variability in the set of context-dependent models to minimize the need for large training data. There are two known clustering methods tasked to achieve the model size reduction: the data driven method and the rule based (decision tree) method[3]. The concept of the data driven method is essentially to tie together the parameters that are acoustically indistinguishable in the training data sets. This allows all the data associated with each individual state to be pooled and to, thereby, enable more robust estimates to be generated. Though the data driven method returns relatively accurate clustering results, its covered context region is limited in training data sets. The decision tree method makes itself structurally easy to expand vocabulary size because it involves building a binary tree for each phone. We can classify unknown contexts by descending the trees. However, it is less accurate and sometimes limited to single mixture Gaussian distributed state for evaluating the node splitting. Since single mixture Gaussian cannot represent the exact properties of any state, it is one of the causes of decision tree learning errors. To resolve this problem, many decision tree construction algorithms are investigated. For example, investigations were focused on rarely seen context problems and alternative tree construction algorithms such as k-means clustering decision tree learning or optimal sub-tree learning algorithms[12]. However, while being somewhat successful in achieving model reduction, these methods all suffer from an increase in algorithm complexities because they need additional tree construction procedures.

This paper introduces a new method that focuses on the desirable attributes of both methods: the achievement of accurate clustering via the data driven method while attaining solutions of unknown contexts via the decision tree method. The proposed scheme essentially constructs a supervised decision rule and applies it over the pre-clustered triphones, employing the C45 algorithm based on the ID3 algorithm, which is known to effectively search through the attributes of the training instances and extract the attribute that best separates the given examples[4]. In particular, the supervised decision tree learning, using both the clustered context-dependent models (as achieved via data driven method) and C45 algorithm, achieves more accurate tied state model and simplifies the adaptation tasking to multiple mixture Gaussian state distributions. Moreover, the scheme also has the inherent ability to effectively append unknown contexts to a context-dependent acoustic model set.

We prescribe, in Section 2, the C45 algorithm used for constructing the decision tree we propose as the supervised decision tree learning. In Section 3 we describe the methodology in applying the proposed algorithm to state clustering task. In Section 4, we show its effectiveness through representative experiments in speech recognition scenarios that include completely known contexts, mixed unknown contexts, and LVCSR task. Concluding remarks are presented in Section 5.

2. C45 Algorithm for Supervised Decision Tree Learning

The C45 algorithm proposed by Quinlan[5] makes a decision tree for classification from symbolic data. The decision tree consists of nodes for testing attributes, edges for branching by values of symbols and leaves for deciding class names to be classified. The C45 algorithm applies to a set of data and generates a decision tree, which minimizes the expected value of the number of tests for classifying the data. Moreover, the C45 algorithm can solve overfit problem using a rule post pruning algorithm while the ID3 algorithm, a previous version of C45, generates that problem.

The most important factor in C45 algorithm is its ability to select the attribute, which is appropriate at each node. The attribute of each node is selected to divide input samples effectively. *Information gain*[7] is used as such a measure of efficiency. In order to define information gain exactly, we first define the degree of complexity about input samples called *entropy*. For the input sample set S has positive and negative samples, the entropy of S is defined as

$$Entropy(S) \equiv -p_p \log_2 p_p - p_n \log_2 p_n \tag{1}$$

where, p_p is the number of positive samples in set S and p_p is the number of negative samples in set S. If all samples in set S are a unique class, the entropy of S is zero. Figure 1. shows an example entropy function for value p_p

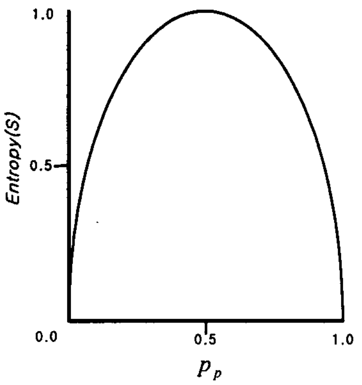

Figure 1. Entropy function of Boolean classification

In the case of existing c types of class in set S, entropy of S is defined as

$$Entropy(S) \equiv \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2}$$

where, p_i is the ratio of class i in set S.

We now define a measuring method to achieve efficient classification of samples for each attribute. For this purpose, we use a method that maximizes the log-likelihood for each node construction decision tree. However, it is not sufficiently a lucid enough procedure to evaluate the log-likelihood of tree for training samples from any node S. Therefore, we use the information gain, which is defined as the expectation of entropy. The log-likelihood for training data samples X at node S of tree essentially represents the conditional probability of samples given node and expressed as

$$L(S) = \log P(\mathbf{X}|S) = \sum_{i} P(X_i|S) \tag{3}$$

The negative entropy $NE(S)$ is defined as

$$NE(S) = \sum_{i} P(X_i|S) \log P(X_i|S) \tag{4}$$

Selecting the attribute that maximizes the log-likelihood at node S procedure has a monotonic relationship with that of maximizing negative entropy. On the other hand, information gain for attribute A, $Gain\,(S, A)$ is obtained by equation

$$Gain(S,A) \equiv Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v) \tag{5}$$

where, *Value* (A) represents the range of attribute A, S_v is a subset of S having V as result of attribute A. Because maximizing the information gain is also maximizing negative entropy, we can construct a decision tree that has a unique target class in each terminal node by applying the optimal attribute using information gain. However, if we increase the depth of the decision tree to classify class of input samples completely, an *overfit* problem can occur. 'Overfit' is defined as:

> *Given a hypothesis space H, a hypothesis $h \in H$ is said to* ***overfit*** *the training data if there exists some alternative hypothesis $h' \in H$, such that h has smaller error than h' over the training examples, but h' has a smaller error than h over the entire distribution of instances[4].*

Namely, an overfit problem leads to important classification errors for the instances of those not contained in the training procedure, although a decision tree is established completely from the initially given training samples. In general, an overfit occurs when noise is present in the data, or when the number of training examples is too small to produce a representative sample of the true target function. To solve this overfit problem, we prune the decision tree employing the rule-post pruning algorithm[4] as follows:

i) Infer the decision tree from the training set, growing the tree until the training data is fit as well as possible and allowing overfitting to occur.

ii) Convert the learned tree into an equivalent set of rules by creating one rule for each path from the root node to a lear node.

iii) Prune each rule by removing any preconditions that result in improving its estimated accuracy.

iv) Sort the pruned rules by their estimated accuracy, and consider them in this sequence when classifying subsequent instances.

This rule-post pruning algorithm prevents the overfit problem and increases clustering reliability for the entire instances.

3. Supervised Decision Tree Learning with the C45 Algorithm for Context-dependent HMMs

As mentioned above, the C45 decision tree learning algorithm is based on the supervised decision tree leaning method. Now, we need to select learning targets for the C45 algorithm in the case of acoustic modeling clustering because we will use the algorithm for clustering context-dependent acoustic models to cope with the previously mentioned disadvantages of the previous decision tree learning algorithms. Additionally, we need to develop a new method that learn decision tree via supervised method using learning target. In this section, we propose a new decision tree construction method that exploits the C45 algorithm to remedy the limitations of each of the two previous methods. In particular, we are motivated to apply the supervised decision tree learning using the C45 algorithm to the context dependent HMM clusters, where pre-clustered triphones on data driven clustering are used as teachers of the supervised learning algorithm. The supervised decision tree learning, using both the clustered context-dependent models (as achieved via data driven method) and the C45 algorithm, is shown to achieve a more accurate tied state model and to simplify the adaptation tasking to multiple mixture Gaussian state distributions. Moreover, the scheme also has the inherent ability to effectively append unknown contexts to context-dependent acoustic model set. In this section, we first explore on building up a training data sample table for the C45 algorithm and then develop a procedure for constructing the decision trees over the context-dependent HMMs using the C45 algorithm.

3.1 Building up the training data sample tables for the C45 algorithm

In order to train decision trees for context-dependent acoustic models, we need to prepare the training data sample sets. Especially, in this paper, the training data sample sets should be appropriate to the ID3 algorithm. Training data sample sets contain data identifications that consist of left, center (itself), and right context phones, answering to each question set Qs, and learning target. These training data sample sets are derived from results of data-driven clustering procedure. Table 1 shows the example of training data sample tables for supervised decision tree learning. Table 1 is constructed via a very simple operation. Firstly, we perform data-driven clustering procedure to determine target classes of triphones. Clustering procedure is done for each phone and each HMM state. The final HMM we use in recognition tasks is trained by the data-driven method. Secondly, we make answering tables for several questions in question set Qs referenced by the left or right context of those triphones. Where Qs contains 'Is its left context a *vowel* ?' or 'Is its right context a nasal ?', for example. In this paper, we select 156 kinds of linguistic questions for node splitting of the decision tree. Q1 to QN, in Table 1, represent each question indexes. Finally, we build up the training data sample tables from target class and all answers of each question sets. The total number of data sample tables is (*total number of phones*)×(*number of HMM states*). The above three steps are said to be preprocessing procedures for supervised decision tree learning.

Table 1. Training examples for the center phone 'ih

Triphone	Q1	Q2	Q3	Q4	....	QN	Target
t-ih+n	No	No	No	Yes	....	No	C1
t-ih+ng	Yes	No	No	Yes	....	No	C1
f-ih+l	No	Yes	No	No	....	No	C2
s-ih+l	No	No	Yes	No	....	Yes	C3
p-ih+ng	No	No	No	No	....	Yes	C3
g-ih+l	Yes	Yes	No	Yes	....	No	C4
t-ih+l	No	No	Yes	Yes	....	No	C5

3.2 Construction of decision trees for context-dependent HMMs using ID3 algorithm

The proposed decision tree learning method is different from previous methods at the point of supervised decision tree learning algorithm cover disadvantage of unsupervised learning rules. Here, the used supervised decision tree learning algorithm is the C45 algorithm.

Input data of C45 algorithm we use are shown in Table 1 which is produced in the above subsection. We essentially construct the required binary decision trees about each center phones using the C45 algorithm. The decision tree construction procedure using C45 algorithm can be summarized as follows:

i) Start with all samples at the root node.
ii) If all examples have a unique target class, then go to stepvi) Otherwise,
iii) Evaluate $Gain(S,A)$s for all questions q, $q \in Q$, and select one question having the largest $Gain(S,A)$ as the attribute of that node.
iv) Split examples into sub-examples of two successive nodes using the selected question in Step iii)
v) For the successive nodes, repeat each step except Step i)
vi) Prune the decision tree using the rule-post pruning algorithm.

In step iii) of the above procedures, $Gain(S,A)$ is described in section 2 and Q is in earlier subsection. The above six steps are identical with general C45 algorithm procedures. That is, we can complete supervised decision tree learning using pre-build learning data samples and previous C45 algorithm, where we are not constrained by the number of mixtures in HMM states and don't need to alternate operations in decision tree learning procedure to reduce learning errors of the tree. Final binary decision trees have nodes applied to the optimal question and terminal nodes have triphones having identical target cluster. The procedure of constructing binary decision trees using the above method not only conserves class from the data driven clustering method, but also makes efficient decision rules about the classes. Figure 2. shows an example binary decision tree in the case of g3 center phone in HTK format. Figure 3. is a flow chart of the proposed decision tree learning algorithm.

After constructing the decision tree, more parameter re-estimation is needed for the reason that more mixture increment or iterative adjustment of HMMs. This mixture increment and parameter re-estimation is accomplished by mean deviation and general Baum-Welch algorithm, respectively. As a consequence of proposed model clustering and tree constructing schemes, we can get more reliable tied state context-dependent acoustic models and decision rule sets for classifying unknown contexts. The final tied state context-dependent acoustic models are equal to ordinary data-driven clustered context-dependent models except that it has appropriate decision rule sets about all of unseen contexts.

```
Center phone = g3
{
   0        'L_Rounded'       -1        -10
  -1           'L_Front'      -2        "g3s21"
  -2      'R_C-Central'       -3         -8
  -3            'R_ss1'       -4         -7
  -4           'R_Stop'       -5        "g3S27"
  -5             'L yo'       -6        "g3S26"
  -6             'L_yu'     "g3S23"     "g3S26"
  -7           'L_Back'     "g3S27"     "g3S25"
  -8      'L_V-Central'     "g3S23"      -9
  -9         'L_V-Back'     "g3S27"     "g3S24"
 -10           'R_Stop'      -11        "g3S26"
 -11      'R_C-Central'     "g3S22"      -12
 -12         'L_V-Back'     "g3S22"     "g3S24"
}
```

Figure 2. Example of decision tree using ID3 algorithm

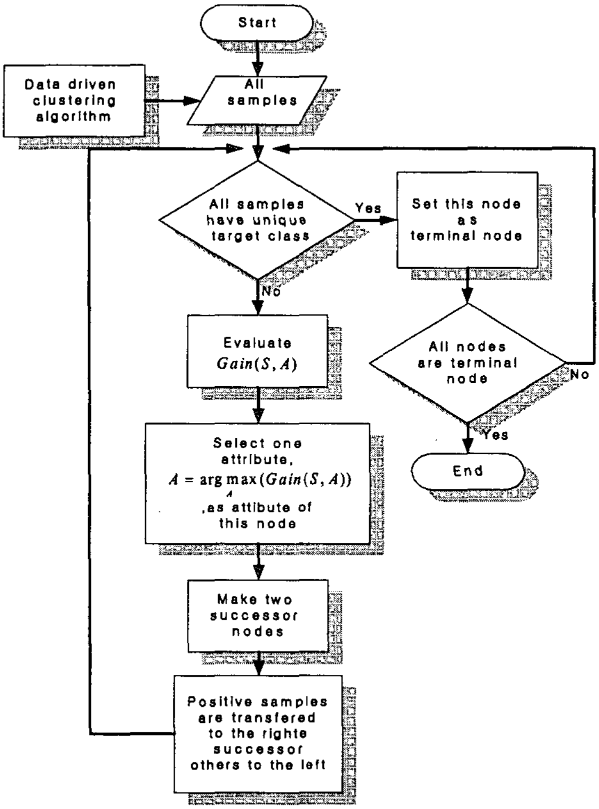
Start
Data driven clustering algorithm
All samples
All samples have unique target class
Yes
Set this node as terminal node
No
Evaluate $Gain(S,A)$
All nodes are terminal node
No
Yes
End
Select one attribute, $A = \arg\max_A(Gain(S,A))$,as attibute of this node
Make two successor nodes
Positive samples are transfered to the righte successor others to the left

Figure 3. Flow chart of supervised decision tree learning algorithm

4. Experiments

In this section, we evaluate the effectiveness of the proposed decision tree construction method described in Section 3 in terms of recognition performance for various unknown contexts ratio scenarios and continuous speech recognition task.

Firstly, we evaluate the recognition performances for various unknown contexts ratios. In context-dependent acoustic modeling, the most important reason of constructing decision trees is to solve the unseen contexts problem. In this paper, we compared the recognition performance of previous decision tree construction method and proposed decision tree construction method for various unknown contexts ratio scenarios. Secondly, we evaluate recognition performance of proposed decision tree learning method on the LVCSR task. Context-dependent acoustic model displays effective performance in LVCSR. Therefore, we need to evaluate the recognition performance of proposed method and compare it with the previous method in a LVCSR task scenario.

In the experiments, we used triphones and trained them through data-driven clustering. Where, we consider beginning and ending silences as word boundaries. Based on the clustered models, a new decision rule was generated using the proposed decision tree method. The performance is measured by the word error rate (WER).There are two kinds of WER evaluation methods. One is percent correct error rate which does not consider word insertion errors. The other is percent accuracy error rate which consider word insertion errors. Percent correct error rate is obtained from the following equation:

$$PercentCorrectErrorRate = \left(1 - \frac{N - D - S}{N}\right) \times 100\% \tag{6}$$

And percent accuracy error rate is obtained from the following equation:

$$PercentAccuracyErrorRate = \left(1 - \frac{N - D - S - I}{N}\right) \times 100\% \tag{7}$$

where, N is the total number of labels in reference transcription, D is the number of deletion errors, S is the number of substitution errors, and I is the number of insertion errors.

In these experiments, we used 36th-order feature vectors consisting of 12-dimensional Mel Frequency Cepstral Coefficient (MFCC) with CMS (Cepstral Mean Subtraction) processed, and their first and second derivatives having delta window of size 2 (± 20 *ms*).The pre-emphasis coefficient is 0.97, and the Hamming window size is 25ms and its frame rate is 10ms. The phonetic models are 3 state left-right continuous HMM which have 8 Gaussian mixtures in each state.

4.1 Performance evaluation on various unknown contexts ratio scenarios

In order to evaluate its effectiveness under various unknown context ratios, we considered the isolate word recognition scenarios. 45,200 speech samples are used for training HMMs, which were recorded by 28 men and 22 women uttering the phonetically-balanced 452 word set. For benchmarking the error rate in completely known contexts task, speech samples of 10 men and 8 women were used, who were not part of the training database. We then made a tree of unknown-contexts ratio tasks for testing by selecting from 240 other words and by combining them with some of the 452 words used in training. Table 2 shows error rate comparing the proposed decision tree construction method to the existing rule-based clustering method for the completely known context task and mixed unknown contexts ratios, 24.2%, 40.5%, and 68.4%, respectively. Here, all errors are substitution error because word network is constructed in the form of parallel finite state automata from uniform language model. As a consequence, the percent correct error rate is identical with the percent accuracy error rate.

The proposed method shows a good performance for completely known contexts included in the training data sets. The result reflects that the parameters of known contexts are obtained by a precise data-driven clustering method in training procedures. However, as the ratio of unknown contexts increases, WER of the proposed method increases more rapidly than the rule-based clustering method. The phenomenon indicates that the constructed decision-rules by the clustered class from data-driven clustering method are less optimal than the linguistically derived decision-rules from the rule-based clustering method.

Table 2. WER (%) on various unknown context ratios

Unknown context ratio	Proposed	Previous
0%	0.78	1.36
24.2%	1.52	1.94
40.5%	5.06	5.22
68.4%	17.73	17.30

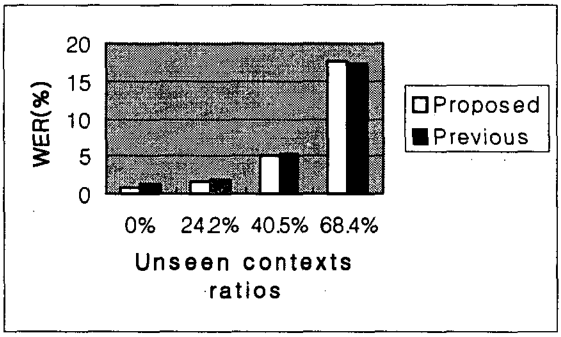

Figure 4. WER for various unseen contexts ratios: WER of proposed method increase more rapidly.

4.2 Performance evaluation on LVCSR task scenario

For the performance evaluations on the continuous speech recognition task, 79,000 continuous speech samples were used for training. They were recorded by 156 men and 130 women uttering 4,000 sentences in a car-navigation task. We then selected another 5.000 speech samples of 1,000 sentences uttered by 90 men and 80 women for a speaker and task independent speech recognition task.

Additionally, in the experiment for continuous speech recognition, we applied a uniform language model to explicitly compare the acoustic modeling performances. Table 3 shows WER comparing the proposed method to the rule-based (Previous) and context independent (CI) acoustic modeling methods for speaker-independent, task-independent continuous speech recognition task. The proposed method is shown to reduce WER by 2.89% of the percent correct error rate and 3.93% of the percent accuracy error rate compared to the previous rule-based method. Especially, insertion error decrease explicitly compared to the previous rule-based method. This result shows that acoustic models obtained by proposed method are less sensitive to noises from breath or lip.

Table 3. WER (%) on continuous speech recognition task.

	WER(%)		Del.	Sub.	Ins.	Tot.
	Percent correct	Percent accuracy				
CI.	35.91	92.34	18	7,692	12,115	21,470
Prev.	12.80	29.52	133	2,615	3,590	21,470
Prop.	12.43	28.36	143	2,525	3,421	21,470

5. Conclusions

In this paper, we developed a new clustering method based on the C45 decision tree learning algorithm that effectively captures the CD modeling. The proposed scheme essentially constructs the decision rule of pre-clustered triphones using the C45 supervised learning algorithm, extracting the positive properties of both the data-driven as well as rule-based methods. The proposed scheme copes the disadvantage of previous unsupervised decision tree learning algorithm by using a pre-clustered triphone from data driven clustering method as learning target. That is, the proposed method does not need alternative processing to solve the clustering problem of multiple mixture Gaussian state distributions. As shown in experimental results, the proposed scheme is shown effective over the database of low unknown-context ratio in terms of recognition performance. However, as the unknown contexts ratio increases in the recognition task, the recognition performance of the proposed method decreases more rapidly than the rule-based method alone. Another positive aspect of the proposed method is its more effective clustering ability on LVCSR tasks. The proposed method reduce error rate by reducing insertion error rate caused almost by breath or lip noises.

Experimental results in this paper show that the proposed method is effective when the speech database used in training contains sufficient contexts along with the reliably trained data-driven clusters enabling the needed binary decision tree. The result also indicate that when a sufficient speech database exists and reliably trained model can be established, our new method has not only the ability of clustering unknown contexts but also achieves improved recognition performance.

References

[1] Young, S. J. 1992. "The general use of tying in phoneme-based HMM speech recognizers." *Proc. of ICASSP*, 569-572.

[2] Bahl, L. R. et al. 1991. "Decision trees for phonological rules in continuous speechs." *Proc. of ICASSP*, 185-188.

[3] Young, S. J. 1997. *The HTK Book*. Cambridge University.

[4] Mitchell, Tom M. 1997. *Machine Learning*. McGraw-Hill.

[5] Quinlan. J. R. 1990. "Decision trees and decision making." *IEEE Trans. on System, Man, and Cybernetics*, Vol. 20, No. 2, 339-346.

[6] Young, S. J. 1996. "Large vocabulary continuous speech recognition: A review." *IEEE Signal Processing Magazine*, Vol. 13. Issue 5, 45-57.

[7] Hamming, R. W. 1986. *Coding and Information Theory*. Prentice-Hall.

[8] Lee, K. 1990. "Context-dependent phonetic hidden Markov models for speaker-independent continuous speech recognition." *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 38, No. 4, 599-609.

[9] Ming, J. & F. Smith. 1999. "A Bayesian triphone model". *Computer Speech and Language*, 13, 195-206.

[10] Digalakis, V. V. et al. 1996. "Genones: generalized mixture tying in continuous hidden Markov model-based speech recognizers." *IEEE Trans. on Speech and Audio Processing*, Vol. 4, No. 4, 281-289.

[11] Hwang, M. et al. 1993. "Shared-distribution hidden Markov models for speech recognition." *IEEE Trans. on Speech and Audio Processing*, Vol. 1, No. 4, 414-420.

[12] Reichl, W. & W. Chou. 2000. "Robust decision tree state tying for continuous speech recognition." *IEEE Trans. on Speech and Audio Processing*, Vol. 8, No. 5, 555-566.

[13] Rabiner, L. & B. H. Juang. 1993. *Fundamentals of Speech Recognition*. Prentice-Hall.

[14] Huang, X. et al. 2001. *Spoken Language Processing*. Prentice-Hall PTR.

Received: January 27, 2003
Accepted: March 5, 2003

▲ Junho Park
Department of Electronics and Computer Engineering, Korea University
5ga-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-922-8997 Fax: +82-2-3291-2450
H/P: 018-255-9553
E-mail: jhpark@ispl.korea.ac.kr

▲ Hanseok Ko
Department of Electronics and Computer Engineering, Korea University
5ga-1, Anam-dong, Sungbuk-ku, Seoul, 136-701, Korea.
Tel: +82-2-3290-3239 (O) Fax: +82-2-3291-2450

H/P: 011-9001-3239
E-mail: hsko@korea.ac.kr