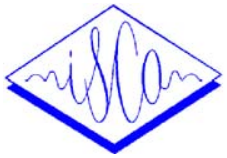

EUROSPEECH '97
5th European Conference on Speech Communication and Technology
Rhodes, Greece, September 22-25, 1997

ACOUSTIC MODELING BASED ON THE MDL PRINCIPLE FOR SPEECH RECOGNITION

Koichi Shinoda and Takao Watanabe
NEC Corporation
4-1-1 Miyazaki, Miyamae-ku, Kawasaki 216, JAPAN
{shinoda,watanabe}@hum.cl.nec.co.jp

ABSTRACT

Recently context-dependent phone units, such as triphones, have been used to model subword units in speech recognition based on Hidden Markov Models (HMMs). While most such methods employ clustering of the HMM parameters(e.g., subword clustering, state clustering, etc.), to control HMM size so as to avoid poor recognition accuracy due to an insufficiency of training data, none of them provide any effective criterion for the optimal degree of clustering that should be performed. This paper proposes a method in which state clustering is accomplished by way of phonetic decision trees and in which the MDL criterion is used to optimize the degree of clustering. Large-vocabulary Japanese recognition experiments show that the models obtained by this method achieved the highest accuracy among the models of various sizes obtained with conventional clustering approaches.

10.21437/Eurospeech.1997-52

1. INTRODUCTION

Over the past few years, extensive studies have been carried out on speaker-independent speech recognition using continuous density Hidden Markov Models (HMMs). It is well known that in most such systems, the use of context-dependent(CD) phone models instead of context-independent(CI) phone models(monophones), improves recognition accuracy[1-7].

Since the number of CD models is usually much larger than that of CI models, using CD models better captures variations in speech data. However, the amount of available training data is likely to be insufficient to support the use of such a large number of CD models. It is often impractical to prepare such a large amount of data. Furthermore, the frequency with which a CD phone appears in training data usually differs substantially in the set of CD phones; in most case, the frequencies for some CD phones are so small that those CD phones do not appear in training data even if a large amount of data is provided. This data insufficiency often causes serious degradation in speech recognition performance. Most recognition systems using CD models employ clustering of model parameters to try to alleviate part of the problem.

Various clustering methods have been developed for this purpose. First, there are several choices for the units to which clustering is carried out; K.F. Lee *et al.*[1], for example, use subword clustering, Hwang *et al.*[2] use state clustering, and Digalakis et al.[3] cluster the mixture components of the HMMs with Gaussian-mixture state observation densities. Second, there are several methods to select the acoustically-similar units to be clustered. Some methods use only the acoustic characteristics of the data and the merging of the units are carried out in a bottom-up manner[4, 2, 3]. The other methods, in addition, utilize *a priori* knowledge about acoustic similarities between the units, which are mostly represented by decision trees[1, 5, 6, 7]. In most of the latter methods, splitting of the units of CI models is carried out in a top-down manner, instead of merging the units of CD models.

In these clustering methods, it is important to properly measure the acoustic similarities between the units utilizing training data, in order to select the units to be clustered. One of the most successful approach is the approach based on the maximum-likelihood(ML) criterion(e.g.,[7]). In the following, for simplicity, the splitting method(top-down clustering) is explained, though the similar explanation is also applicable to the merging method(bottom-up clustering). In this approach, the increase of the likelihood by splitting is calculated for each unit in the unit set, and the unit that has the largest increase is selected and split.

However, this ML approach has one drawback. In most case, the likelihood becomes larger as the number of units becomes larger. In the final stage of the splitting, the model set becomes almost identical to the set of CD models without clustering. Therefore, this approach requires an external parameter to control the degree of clustering. Most methods limit splitting using a threshold on the increase in the likelihood or on the number of units. These thresholds needs to be optimized through a series of recognition experiments using test data or by a cross-validation method. These optimization processes are computationally expensive, need more data, and have no strong theoretical justification.

In this paper we propose a new approach in which a minimum description length(MDL) criterion, instead of the ML criterion, is used for clustering. The MDL approach[9] is based on an information theoretic criterion, which has been used for selecting the probabilistic model with an appropriate complexity for the given amount of data. This MDL criterion is effective not only for selecting the units to be split, but also for deciding whether to stop splitting. Therefore, no other external parameter is needed to control the degree of clustering. We apply this criterion to state splitting using phonetic decision tree.

2. MDL CRITERION

MDL[9] is an information criterion which has been proven to be effective in selecting the optimal model from among various probabilistic models. The MDL criterion selects the model with the minimum description length for the given data as the optimal model from among a set of models. When a set of models $\{1, \ldots, i, \ldots, I\}$ is given, the description length, $l_i(x^N)$, of the data, $\{x^N = x_1, \ldots, x_N\}$, together with an underlying model i is given by,

$$l(i) = -\log P_{\hat{\theta}^{(i)}}(x^N) + \frac{\alpha_i}{2}\log N + \log I \qquad (1)$$

where α_i is the dimensionality (the number of free parameters) of model i, and $\hat{\theta}^{(i)}$ is the maximum likelihood estimates for the parameters $\theta^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{\alpha_i}^{(i)})$ of model i. The first term in (1) is the code length for the data x^N when model i is used as a probabilistic model. This term

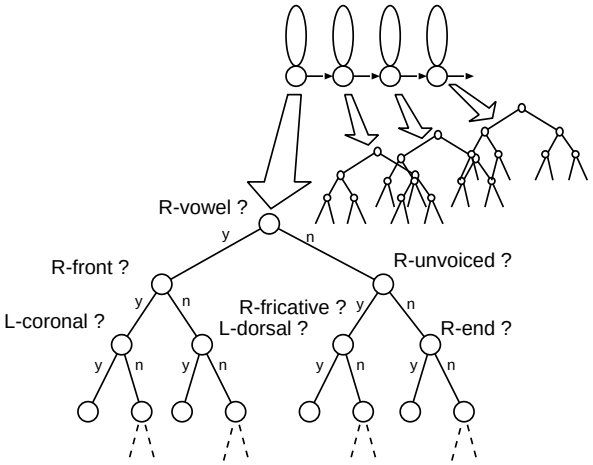

Figure 1. Phonetic decision tree

is identical to the negative of the log likelihood of the data used in the ML criterion. The second term is the encoding length for model i. This term represents for the complexity of model i. The third term is usually referred to as the code length required for choosing model i and is assumed to be constant in this paper. As a model becomes more complex, the value of the first term decreases and that of the second term increases. The second term works as a penalty imposed for adapting a large model size. The description length l has its minimum at a model of an appropriate complexity. As one can see in (1), the MDL criterion does not need any externally given parameters; the optimal model for the data is automatically obtained once a set of models is specified.

When complex models such as those used in speech recognition are used, it is often impractical to calculate the description length for all the possible models, because it requires high computational costs. In our method, therefore, several assumption and approximations are introduced to reduce the computational costs. They are explained in Section 4.

3. TREE-BASED STATE CLUSTERING

In this section, the outline of the proposed method is shown.

For modeling CD phone units, we use triphones[8], in which the left and the right neighboring phones are taken into consideration; two phones that have the same CI identity but with different right or left context are considered different triphones. Each triphone model is a left-to-right HMM with a Gaussian output probability density function (pdf) for which a diagonal covariance is assumed. All HMMs of triphones derived from the same CI phone assumed to have the same number of states.

As a clustering scheme, we employ the state splitting by way of phonetic decision trees[7]. It clusters the triphone states with similar phonetic contexts into one state. One phonetic decision tree is constructed for each state of each CI phone HMM(Figure 1). Each root node of the tree represents a set of all the triphone states corresponding to the CI phone state, and each other node represents one subset of the triphone states. From the top to the bottom, each node is split into two nodes using a question related to phonetic contexts. The examples of the questions are "Is the previous phone unvoiced or not ?" (L-unvoiced ?), and "Is the next phone a fricative ?" (R-fricative ?). In each splitting, one question is selected from among a set of questions, which is prepared beforehand. The MDL criterion is used for the selection of the optimal question and for the decision whether to stop splitting. Finally, when there exist no nodes to be split, the pdf parameters

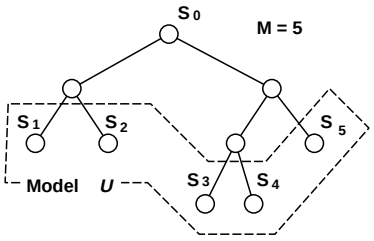

Figure 2. A model (node set) in the decision tree

of each leaf node is copied to the pdf parameters of the triphone states in the corresponding subset and used for recognition.

4. STATE SPLITTING USING MDL CRITERION

In this section, we shall discuss how the MDL criterion used in state splitting. Here a *model* is defined as a node set in a phonetic decision tree. The description length for each model is calculated and the model with the minimum description length is selected from among various models. To reduce the computational cost, we assume that the splitting does not change the frame/state alignment between the data and the model, and that the transition probabilities of HMMs can be neglected. In the following two subsections, how to calculate the description length for a model and how to get the optimal model efficiently are discussed respectively.

4.1. Calculation of Description Length

When a state S_0 of a phone HMM is split into M nodes, $S_1, \ldots, S_M$, as shown in Figure 2, the description length $l(U)$ of the model, $U = \{S_1, \ldots, S_M\}$, is calculated as follows. First, the log likelihood L of node S_m generating a set of training frames, $\mathbf{o}_1, \ldots, \mathbf{o}_T$, is approximately given by,

$$\begin{aligned} L(S_m) &\approx \sum_{t=1}^{T} \log(N(\mathbf{o}_t, \mu_{S_m}, \Sigma_{S_m}))\gamma_t(S_m) \\ &= -\frac{1}{2}(\log((2\pi)^K \mid \Sigma_{S_m} \mid) + K), (S_m), \quad (2) \\ \gamma_t(S_m) &= \frac{\alpha_t(S_m)\beta_t(S_m)}{\sum_s \alpha_t(S_m)\beta_t(S_m)}, \quad (3) \\ , (S_m) &= \sum_{t=1}^{T} \gamma_t(S_m), \quad (4) \end{aligned}$$

where K is the dimensionality of the data vector $\mathbf{o}_t$, $\gamma_t(S_m)$ is the *a posteriori* probability of the observed frame $\mathbf{o}_t$ being generated by state S_m, and , (S_m) is a total state occupancy count for S_m, which is the sum of $\gamma_t(S_m)$ over all the data frames. The forward probability $\alpha_t(S_m)$, the backward probability $\beta_t(S_m)$, the mean vector μ_{S_m}, and the covariance matrix Σ_{S_m} are calculated from the training data. Then, the description length $l(U)$ in (1) is given by,

$$\begin{aligned} l(U) &\approx -\frac{1}{2}\sum_{m=1}^{M} L(S_m) + KM \log \sum_{m=1}^{M} , (S_m) \\ &= \frac{1}{2}\sum_{m=1}^{M} , (S_m) \log(\mid \Sigma(S_m) \mid) \\ &\quad + KM \log V, \quad (5) \\ V &= \sum_{m=1}^{M} , (S_m) = , (S_0), \quad (6) \end{aligned}$$

where the constant terms during the node-splitting process are neglected, and the dimensionality of the model U is $2KM$ (M mean vectors and M diagonal covariances).

4.2. Model Selection

In order to get the optimal model, the calculation of the description length for all the possible model is required. However, it is practically impossible because it needs high computational costs. Therefore, we use an algorithm that needs relatively small computational costs and achieves a suboptimal solution.

Let $\Delta_q(S)$ be the difference between the description length l before splitting and after splitting when node S is split into two by using question q. It is given by the following equation:

$$\begin{aligned}\Delta_q(S) &= \frac{1}{2}(, (S_{qy})\log \mid \Sigma_{S_{qy}} \mid +, (S_{qn})\log \mid \Sigma_{S_{qn}} \mid \\ &\quad -, (S)\log \mid \Sigma_S \mid) + K\log V, \end{aligned} \tag{7}$$

where S_{qy} and S_{qn} is the resulting two nodes after the splitting. First, one state S_0 of a CI phone HMM is set to be the root node. Then the root node is split by each question, and the question q' which minimize $\Delta_q(S_0)$ is selected. If $\Delta_{q'}(S_0) > 0$, then quit splitting. If $\Delta_{q'}(S_0) < 0$, node S_0 is split into two nodes, $S_{q'y}$ and $S_{q'n}$, and the same procedure is repeated for each of these two nodes. This node splitting is recursively carried out until there exist no nodes to be split. This procedure is done for all the states of all the CI phone HMMs.

Compare this MDL approach with the ML approach[7], which is described as follows. Let δ_q(S) be the increase of the log likelihood when node S is split into two by using question q, then,

$$\begin{aligned}\delta_q(S) &= L(S_{qy}) + L(S_{qn}) - L(S) \\ &= -\frac{1}{2}(, (S_{qy})\log \mid \Sigma_{S_{qy}} \mid +, (S_{qn})\log \mid \Sigma_{S_{qn}} \mid \\ &\quad -, (S)\log \mid \Sigma_S \mid). \end{aligned} \tag{8}$$

First, the question q' which maximize δ_q is chosen from among the questions, and then S is split into two nodes $S_{q'y}$ and $S_{q'n}$. This splitting process is recursively carried out. In this ML approach, the splitting process must be stopped by some externally given parameters to control the degree of clustering, since the increase δ_q is positive in almost all the splitting. Most methods use a threshold on the total occupancy count , (S) and/or a threshold on the increase $\delta_q(S)$ as the control parameters. However, the optimization of these parameters requires a series of recognition experiments, which are computationally expensive, need more data, and have no strong theoretical justification. On the contrary, the MDL approach needs no external control parameters; the term $KlogV$ in (7) corresponds to the threshold on the increase δ in (8), and this term is estimated automatically by using the training data.

5. EXPERIMENTS

The proposed method was evaluated by a task of recognizing Japanese 5000 words. Every utterance was digitized at a 16 kHz sampling rate, and analyzed in 10 msec frame periods. The feature used was a vector of 21 components, consisting of a power derivative, 10 mel-scaled cepstral coefficients, and 10 corresponding mel-scaled cepstral time derivatives. The number of Japanese CI phones was set to be 37, and the number of the triphones derived from these phones was 4309. The number of states in each phone HMM was four. A Gaussian output pdf with a diagonal covariance was assumed for each state. The number of questions used in the node splitting was 106. Two databases, Data A and Data B, were prepared for training. Data A consists of data of 46 male speakers, in which each speaker uttered 250 phonetically-balanced words. Data B consists of data of 36 male speakers in which each speaker uttered 2150 phonetically-balanced words. For testing, the speech data of five male speakers, who were not involved in either Data A or Data B, was used. Each of these test speakers uttered 250 words. All the word in the test data are not included in the training vocabulary.

Table 1. Comparison between MDL and ML

	D	V	# of state	Recog. rate(%)
MDL	–	–	2069	80.4
ML 1	60	0	3739	75.4
ML 2	100	0	3000	76.4
ML 3	200	0	2001	76.7
ML 4	300	0	1943	75.4
ML 5	400	0	1200	73.4
ML 6	500	0	1018	71.9
ML 7	1000	0	591	66.6
ML 8	60	200	2777	76.2
ML 9	60	400	2034	77.0
ML 10	60	600	1488	77.8
ML 11	60	800	1248	77.9
ML 12	60	1000	751	77.4

First, the effectiveness of the proposed method was evaluated. Table 1 compares the recognition results obtained from the proposed method(MDL, displaying the average of the 5 speakers) with those obtained in the ML approach when Data A was used for training. In the experiments using ML approach, two thresholds, D and V, were externally given; D for the amount of data was and V for the increase of likelihood. Among the questions with which the state occupancy counts of the resulting two nodes, , (S_{qy}) and , (S_{qn}), were both more than D, the question q' which maximized δ_q was chosen. If $\delta_{q'}$ was larger than V, state S was split. The experiments using the ML approach were carried out for twelve combination of these two thresholds(ML 1–12). As shown in Table 1, the proposed method achieved higher recognition accuracy than any results of the experiments using the ML approach. Although the number of the experiments using the ML approach was small, it is clear that the proposed method is effective in choosing the model with an appropriate size for the amount of training data. The computational cost required for the proposed method is almost the same as each of the experiments using the ML approach. This results indicates that the proposed method in most case needs a much less computational cost than that for the ML approach to get the optimal model for the training data. Table 2 shows the frequency of the questions used in the proposed method. Here, "L-begin" corresponds to a question, "Is the phone located at the beginning of a word ?".

Next, how the model size changes as the amount of data increases was examined. Table 3 shows the results when Data B, which is seven times larger than Data A, was used for training. The increase of the number of states indicated that the control of model size worked well.

Then, the optimality of the model size control was investigated using Data B. For this purpose, a weight coefficient c was added to the second term in (1), as shown in the following:

$$\begin{aligned}l'(U) &= \frac{1}{2}\sum_{m=1}^{M}, (S_m)\log(\mid \Sigma(S_m) \mid) \\ &\quad + cKM\log, (S_0). \end{aligned} \tag{9}$$

Table 2. Distributions of questions asked

vowel		consonant	
L-coronal	67	L-begin	130
L-dorsal	55	L-back	69
L-begin	40	R-a	63
R-coronal	39	R-high	62
L-h	35	L-high	60
L-back	34	L-a	53
L-sonorant	33	L-front	45
R-dorsal	31	R-e	34
L-unvoiced	27	R-back	32
L-n	27	L-e	30
L-fricative	27	L-consonant	28
TOTAL	1110	TOTAL	822

Table 3. Recognition rate (%) using Data A and Data B

Training Set	Data A	Data B
# of nodes	2069	6223
Rec. rate	80.4	86.0

As c becomes large, the penalty for large model size becomes large. The results when c was changed from 0.1 to 10.0 are shown in Table 4. The optimal value of c was 2.0, but the recognition rate of $c = 2.0$ was only 0.7% higher than that of $c = 1$, in which (9) is identical with (1). One can safely state that the model selected by the MDL criterion works well compared to the models tested in Table 4.

Finally, the recognition performance using mixture-Gaussian output pdf was examined using Data B. In this experiment, the number of Gaussian of each state was increased to two, and the model was re-trained using the same training data. The result was shown in Table 5. The error rates was decreased by 21.0% on average. This result indicates that some of the models constructed using the MDL criterion can be split by using some other contexts; i.e., the set of questions prepared beforehand was not sufficient to get the optimal model for recognition.

6. DISCUSSION

This is our first attempt to optimize the model size without any externally given parameters and there remain several problems to be solved. First, some approximation and assumptions are made in the proposed method, and they may affect the performance of model size control. Although the effectiveness of the proposed method justifies the use of them, it should be further examined as how they influence the model size control. Second, the "true model" may not be involved in the set of models provided beforehand. In this case, the selected model by the MDL criterion is not the optimal model. This holds true not only for the proposed method, but also for all the model selection strategies using the MDL criterion in general. Further theoretical research is needed for this problem. Third, the minimization of description length dose not imply the minimization of the recognition error. The conventional ML approach has the same problem; the maximization of the likelihood does not mean the minimization of the recognition error. The MDL criterion has an advantage over the ML criterion in that it has an effective penalty term for model complexity control based on a good theoretical support.

There are also some other criteria developed for model size control. One of those is the widely known Akaike Information Criterion(AIC)[10]. In AIC, the second term in (1) is replaced by $\alpha_i/2$. The comparison between MDL and AIC is not carried out in this paper since the difference between them seems small, and thus the performance of these two is expected to be similar.

Table 4. Recognition rate(%) as a function of coefficient c

c	0.1	0.5	1.0	2.0	4.0	10.0
# of nodes	13927	9798	6223	3949	2418	1341
Rec. rate	85.4	85.9	86.0	86.7	85.9	84.1

Table 5. Recognition rate(%) with mixture-Gaussian output pdf

1 Gauss	2 Gauss
86.0	89.0

7. CONCLUSION

A training method for acoustic modeling that generates the HMMs with appropriate model size is proposed. It achieved as high recognition accuracy as the conventional approach with a large reduction of the overall computational costs in our evaluation experiments.

The MDL criterion can be applied not only to the state splitting using the phonetic decision tree but also to the other clustering methods such as the agglomerative clustering methods. It can also be applied to discriminative training. Studies in these directions seem to be promising.

REFERENCES

[1] K.-F.Lee *et al.*: "Allophone Clustering for Continuous Speech Recognition", *Proc. ICASSP90*, Albuquerque, pp.749-753, 1990.

[2] M.-Y.Hwang *et al.*: "Predicting Unseen Triphones with Senones", *Proc. ICASSP93*, Minneapolis, pp.II-311-314, 1993.

[3] V.Digalakis *et al.*: "Genons: Generalized Mixture Tying in Continuous Hidden Markov Model-Based Speech Recognizers", *IEEE Trans. SAP*, vol.4, No.4, pp.281-289, 1996.

[4] S.J.Young: "The General Use of Tying in Phoneme-Based HMM Speech Recognizers," *Proc. ICASSP92*, San Francisco, pp.569-572, 1992.

[5] L.R.Bahl *et al.*: "Decision Trees for Phonological Rules in Continuous Speech", *Proc. ICASSP91*, Toronto, pp.185-188, 1991.

[6] C.-H.Lee *et al.*: "Improved acoustic modeling for large vocabulary continuous speech recognition", *Computer Speech and Language*, vol.6, No.2, pp.103-207, 1992.

[7] S.J.Young *et al.*: "Tree-Based State Tying for High Accuracy Acoustic Modeling", *Proc. of Human Language Technology*, pp.307-312, 1994.

[8] R.M. Schwartz, *et al*: "Improved Hidden Markov Modeling of Phonemes for Continuous Speech Recognition.", *Proc. ICASSP84*, 35.6, 1984. Algorithm For Efficient Allophone Modeling", *Proc. ICASSP92*, San Francisco, pp.I-573-576, 1992.

[9] J.Rissanen:"Universal Coding, Information, Prediction, and Estimation", *IEEE Trans. IT*, vol.30, No.4, pp.629-636, 1984.

[10] H.Akaike:"A new look at the statistical model identification," *IEEE Trans. AC*, vol.AC-19, pp.716-723, 1974.