# 【Invited】 Robust Acoustic Modeling for Speech Recognition

Koichi SHINODA[†]

† Department of Computer Science, Tokyo Institute of Technology
2–12–1, Ookayama, Megoro-ku, Tokyo, 152–8552 Japan

**Abstract** While Hidden Markov Models (HMMs) have been successfully applied to automatic speech recognition, they are not still robust enough against differences in speakers, speaking-styles, and environmental noises. To tackle this problem, we need to study the inner structure of speech by using large corpus and rich computational power. In this direction, the model size tends to be increase and hence the data insufficiency problem becomes more serious. In this paper, we focus on robust modeling against data insufficiency. Approaches based on information criteria such as Minimum Description Length and structural approaches in which models are changed according to the amount of data availabl are discussed. While these techniques have been important for HMM research, it will be more important in the research beyond HMM.

**Key words** acoustic modeling, information criterion, distance measure, MDL, SMAP

## 1. Introduction

Hidden Markov Models (HMMs) has been successfully applied to automatic speech recognition. They can efficiently represent the variety of speech, and have an efficient algorithm called Expectation-Maximization algorithm to estimate its parameters. For read speech, in which each sentence is clearly and rather slowly pronounced, their recognition accuracy is more than 95%. For spontaneous speech used in daily conversation, however, it degrades drastically. In addition, it is largely influenced by speaker characteristics and environmental noises. From these facts, it is clear that HMMs are still not robust enough for our daily use.

One major reason for this lack of robustness is that the recognition process using HMMs are rather superficial. It may be compared to estimating the geological formation of the bottom of the sea by observing the waves on its surface. Its framework is easily applicable for recognizing any time-series data, which means it is not related to the inner structure of human speech. In order to improve the performance of speech recognition, it is necessary to step into the inner structure of speech, to analyze it, and to utilize it.

The following two approaches seem promising for this purpose. One is to simulate human production and perception process (Analysis-by Synthesis approach, e.g., [9]) and the other is to extract useful information from large speech corpus (Data mining approach, e.g., [22]). While these two approaches are both important and deeply related to each other, we focus on the latter in this paper.

Since the search space for mining is vast while computational power available is limited, it is practically important to utilize those methods that need few control parameters, based on information-theory. In this paper we introduce two such methods. One is for acoustic modeling and the other for speaker adaptation. The author believes the investigation in this direction is very important to construct statistical models beyond HMMs.

This paper is organized as follows. In the next section, the acoustic modeling based on the Minimum Description Length (MDL) criterion is explained. In Section 3, the speaker adaptation method based on Structural Maximum A Posteriori (SMAP) estimation is explained. In Section 4, the future direction of these approaches is discussed.

## 2. Model Selection using MDL Criterion

### 2.1 Motivation

It is well known that in most speech recognition systems the use of context-dependent phone units such as triphones rather than context-independent phone units such as monophones provides greater recognition accuracy. While the large number of triphones employed in a typical system can help to capture variations in speech data, the amount of available training data is likely to be insufficient to support the use of such a large number. Such lack of data can seriously degrade speech recognition performance and most recognition systems using triphones cluster the model parameters to try to alleviate the problem. Various clustering methods have been developed for this purpose.

One of the most successful approaches is that based on the maximum-likelihood (ML) criterion [21]. In this approach, state splitting based on phonetic decision trees is used as a clustering scheme for single-Gaussian HMMs. The difficulty with this ML approach, however, is determining when to halt the splitting process, which could be carried on until the model simply consisted of a full set of individual, unclustered parameters. Usually, the splitting process is limited by

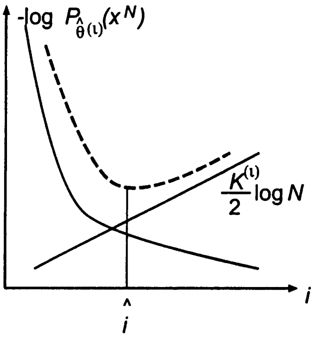

Figure 1 The MDL criterion.

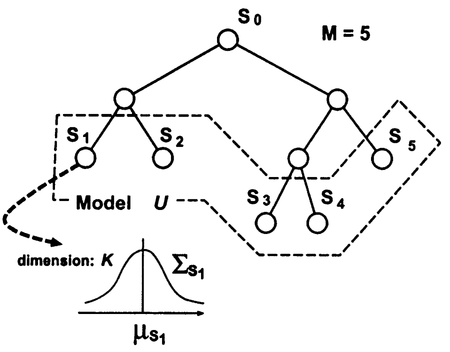

Figure 2 Model (node set) in the decision tree.

imposing a threshold value on the increase in the likelihood or on the number of parameter clusters, but the process required to optimize such thresholds (a series of recognition experiments, cross-validation, etc.) is computationally expensive.

In the following, an approach that uses the *minimum description length*(MDL) criterion for state splitting [16] is explained. This MDL approach is effective for deciding when to stop splitting.

### 2.2 MDL Criterion

The MDL criterion [11] has been proven to be effective in selecting the optimal model from among various probabilistic models. It selects the model with the minimum description length for given data. When a set of models $\{1, \ldots, i, \ldots, I\}$ is given, the description length $l_i(\mathbf{x}^N)$ for data $\{\mathbf{x}^N = x_1, \ldots, \mathbf{x}_N\}$ and an underlying model $i$ is given by

$$l_i(\mathbf{x}^N) = -\log P_{\hat{\boldsymbol{\mu}}^{(i)}}(\mathbf{x}^N) + \frac{K_i}{2}\log N + \log I, \qquad (1)$$

where $K_i$ is the dimensionality (the number of free parameters) of model $i$ and $\hat{\boldsymbol{\mu}}^{(i)}$ represents the maximum likelihood estimates for the parameters $\boldsymbol{\mu}^{(i)} = (\theta_1^{(i)}, \ldots, \theta_{K_i}^{(i)})$ of model $i$. The first term on the right-hand side of (1) represents the code length for data $\mathbf{x}^N$ when model $i$ is used as a probabilistic model. This term is identical to the negative of the log likelihood used in the ML criterion. The second term is related to the complexity of model $i$ and the number of data samples, $N$. The third term is the code length required for choosing model $i$ and is assumed here to be a constant. 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 employing a large model size (see Figure 1). In a comparison among models, the model with the shortest description length $l$ may be considered the one having the most appropriate size and complexity. As may be seen 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 has been specified.

### 2.3 Description length for HMMs

For the use of the MDL criterion, it is first necessary to prepare the model set from which that optimal model is to be selected. For speech recognition using CDHMMs, it is impossible to prepare all the possible models because of the large number of possible structures of CDHMMs. In this study, the focus is on the clustering of the states in CDHMMs and constant values are given to those parameters unrelated to state clustering, such as the number of states in a single unit.

Here a *model* is defined as a node set in a phonetic decision tree in which a Gaussian pdf is assigned for each node. When the root node $S_0$, which represents the whole set of the triphone states in the tree, is split into $M$ nodes, $S_1, \ldots, S_M$, as shown in Figure 2, one model $U(S_1, \ldots, S_M)$ is defined for the node set $\{S_1, \ldots, S_M\}$. Different node sets correspond to different models. The description length for each node set is calculated and the node set with the minimum description length is selected from among various node sets as being the optimum model.

The first term on the right-hand side of Eq. (1) is the negative of the log-likelihood of a probabilistic model with respect to given data. Under some assumptions, the log-likelihood of the data for all the nodes in set $U$ is calculated as follows(for details, see [16]):

$$\begin{aligned} L_{all} &= \sum_{m=1}^{M} L(S_m) \\ &\simeq -\sum_{m=1}^{M} \frac{1}{2}\Gamma_m(K + K\log(2\pi) + \log|\Sigma_m|). \qquad (2) \end{aligned}$$

where $K$ is the dimension of each feature vector, $\Gamma_m$ is the occupancy count for state $m$, $\Sigma_m$ is the covariance of Gaussian distribution for state $m$.

The second term on the right-hand side of Eq. (1) represents the complexity of a model. In the proposed approach, it is assumed that the covariance of each Gaussian pdf is diagonal. The number of parameters to be estimated for model $U$ is $2KM$ (with model $U$ containing $M$ mean vectors and $M$ diagonal covariances). The total number of data samples is the sum of $\Gamma_m$ over $m$. With this total, the second term may be approximated as:

$$R = KM \log W, \tag{3}$$

where $W = \sum_{m=1}^{M} \Gamma_m$. As has been previously noted, the third term on the right-hand side of (1) is fixed at a constant value, $C$, for all possible models.

Finally, using (2) and (3), the description length $l(U)$ for model $U$ is calculated as follows:

$$l(U) \simeq \sum_{m=1}^{M} \frac{1}{2}\Gamma_m(K + K\log(2\pi) + \log|\mathbf{\Sigma}_m|) \\ +KM\log W + C. \tag{4}$$

### 2.4 State splitting using the MDL criterion

In order to get an optimal model, it is needed to calculate description lengths for all possible models, which would involve prohibitively high computational costs. Instead, an algorithm that obtains only a suboptimal solution is used.

Let us first assume that node $S$ of model $U$ splits into two nodes $S_{qy}$ and $S_{qn}$,in response to question $q$. Let $\Delta(q)$ be the difference between the description lengths after the splitting and before it (i.e., $l(U') - l(U)$). Then this $\Delta_m(q)$ will be given by the following equation:

$$\Delta(q) = l(U') - l(U) \\ = \frac{1}{2}(\Gamma_{qy}\log|\mathbf{\Sigma}_{qy}| + \Gamma_{qn}\log|\mathbf{\Sigma}_{qn}| \\ -\Gamma\log|\mathbf{\Sigma}|) + K\log W. \tag{5}$$

where $\Gamma_{qy}$ is the state occupancy count for node $S_{qy}$ and $\Gamma_{qn}$ is that for node $S_{qn}$. In state splitting, the question $q'$ which would minimize $\Delta_0(q')$ when used to split root node $S_0$ is first determined. If $\Delta_0(q') > 0$, then no splitting is conducted. If $\Delta_0(q') < 0$, then 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 carried out until there remain no nodes to be split and is conducted for the root nodes of all the phonetic decision trees in all the HMMs.

Most ML approach apply a threshold value to the total occupancy count and/or to the log-likelihood increase. However, the optimization of these parameters requires a series of recognition experiments which are computationally expensive and require additional data. The MDL approach needs no external control parameters; the term $K\log W$ in (5) corresponds to the threshold for likelihood increase, and this term is estimated automatically on the basis of the training data. Additionally, the threshold term $K\log W$ is specified for each phone in the MDL approach, while the threshold for likelihood increase is shared among all the phones. This indicates that the MDL approach is more robust against the data imbalance among phones than the ML approach.

### 2.5 Discussion

A number of problems remain to be solved, however. First, the degree to which the assumptions implicit in the proposed method affect its performance with regard to the control of model size has to be determined. A second problem is that the set of models provided beforehand may not include the most optimal model ("true model") for the given data. A third problem is that, since it is assumed that the amount of data is sufficiently large in the MDL criterion, it may not apply to the case where the amount of data available is small. These latter two problems are of course true not only for the proposed method but also for other model selection strategies using the MDL criterion, and further theoretical research addressing these problems is needed. A fourth problem is that the minimization of the description length does not necessarily minimize recognition error. It should be noted that conventional ML approaches encounter the same problem.

Two other widely known information criteria used for controlling model size are the Bayesian information criterion (BIC) [12] and the Akaike information criterion (AIC) [1]. The formula for the BIC is

$$l_i^{BIC}(\mathbf{x}^N) = -\log P_{\hat{\boldsymbol{\mu}}^{(i)}}(\mathbf{x}^N) + \frac{K_i}{2}\log N. \tag{6}$$

Comparing this criterion with the MDL criterion (Eq.1), one can easily see that the first and the second terms are identical and that the only difference is that the MDL criterion has a third term. Since throughout this thesis the third term is assumed to be constant, the BIC gives exactly the same results as the MDL criterion here. After the result of our research was first published [15], the approach using the BIC to control the model size in speech recognition was extensively studied. It has been successfully applied to speaker clustering [3], Gaussian mixture modeling [4], modeling of mixture of Gaussian pdf for HMM [5], and segmentation of speech data [19]. Since the BIC gives exactly the same results as the MDL criterion does, the results of these studies strongly support the effectiveness of our approach. They also proved that our approach can be applied to many other data insufficiency problems in speech recognition.

In the AIC, the second term in (1) is replaced by $K_i$ and there is no third term:

$$l_i^{AIC}(\mathbf{x}^N) = -\log P_{\hat{\boldsymbol{\mu}}^{(i)}}(\mathbf{x}^N) + K_i. \tag{7}$$

Practically speaking, it is well known that in many applications the results given by the AIC differ little from those given by the MDL criterion. The MDL criterion and the AIC are therefore not compared here. In theory the difference between the MDL criterion and the AIC is still controversial but it is not discussed here because it is not an important issue here. One typical claim supporting the MDL criterion is that the AIC tends to overestimate the number of parameters needed [13]; while the AIC is likely to select the correct model when the complexity of the true model grows with sample size, such a case is unlikely to happen in actual applications.

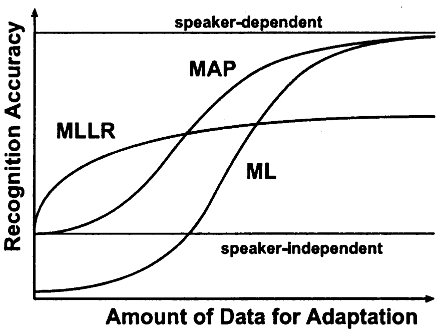

Figure 3 Recognition performance of maximum *a posteriori* (MAP) adaptation, maximum likelihood linear regression (MLLR), and maximum likelihood estimation (ML).

## 3. Structural MAP Approach to Speaker Adaptation

### 3.1 Motivation

*Maximum a posteriori* (MAP) estimation has been successfully applied to speaker adaptation [7]. The MAP estimate of the parameter vector is defined as the mode of the posterior pdf given the adaptation data. It is well known, since MAP estimates are *asymptotically equivalent* to ML estimates, that the resulting recognition performance is similar to that of speaker-dependent (SD) HMMs when the amount of data becomes large. In these conventional MAP estimation methods, HMM parameters of different speech units are often assumed to be independent. Therefore, each model can be adapted only if the corresponding speech unit has been observed in the current set of adaptation data. The improvement is consequently rather small when the amount of adaptation data is extremely limited.

Another category of adaptation techniques, which do not use the MAP framework, are often referred to as *transformation-based* approaches, such as *maximum likelihood linear regression* (MLLR) [10]. This family of techniques limits the number of free parameters by tying the HMM parameters or by applying some constraints on the parameters in order to improve recognition accuracies with a small amount of data. When the amount of adaptation data exceeds a certain value, however, the recognition accuracy often becomes inferior to that obtained with ML estimation of the model parameters. This is because a model with a small number of free parameters could not fully utilize the potential information embedded in the large amount of data. In Figure 3, the difference between the recognition performance of MAP and that of MLLR is shown.

Combinations of MAP and transformation-based approaches have also been studied intensively. Notable examples were in combining MLLR and MAP [6]. The shortcoming of these combined methods is again the use of fixed *structures*, i.e. fixed ways of parameter tying, in the acoustic space. Therefore they have only been shown useful with adaptation data sizes within a narrow range. To alleviate this problem, a *tree structure* has been used in adjusting the number of layers in a tree and the degree of parameter tying according to the amount of available data (e.g., [14]).

*Structural maximum a posteriori adaptation*(SMAP) method takes advantage of both the nice asymptotic property of MAP estimation for large size adaptation and the flexible parameter tying strategy in a tree for small sample adaptation, and achieves the two desired objectives mentioned earlier. In this method, it is assumed that the prior knowledge in a tree node can be used to construct prior density needed for MAP estimation of all the parameters in the successive child nodes.

### 3.2 Tree structure

In SMAP, a tree structure is adopted as the structure to aid MAP estimation, because it offers a natural evolution of prior knowledge embedded in the parent-child relationship between nodes at different tree layers.

Given the set of all the mixture Gaussian components in the set of HMMs, it is needed to first define a distance measure, $d(m,n)$, between Gaussian components, $g_m(\cdot)$ and $g_n(\cdot)$, in order to build a tree. Here the distance is defined as the sum of the *Kullback-Leibler divergence* from $g_m(\cdot)$ to $g_n(\cdot)$ and that from $g_n(\cdot)$ to $g_m(\cdot)$. When diagonal covariance matrices are assumed, the distance $d(m,n)$ is evaluated as follows:

$$\begin{aligned} d(m,n) &= \int g_m(x)\log\frac{g_m(x)}{g_n(x)}dx + \int g_n(x)\log\frac{g_n(x)}{g_m(x)}dx, \\ &= \sum_i [\frac{\sigma_m^2(i) - \sigma_n^2(i) + (\mu_n(i) - \mu_m(i))^2}{\sigma_n^2(i)} \\ &\quad + \frac{\sigma_n^2(i) - \sigma_m^2(i) + (\mu_n(i) - \mu_m(i))^2}{\sigma_m^2(i)}], \qquad (8) \end{aligned}$$

where $\mu_m(i)$ is the $i$-th element of the mean vector $\boldsymbol{\mu}_m$ and $\sigma_m^2(i)$ is the $i$-th diagonal element of the covariance matrix $\boldsymbol{\Sigma}_m$. Next, at each node $k$ in a tree structure, the collection of Gaussian components belonging to node $k$, $\{g_m^{(k)}(\boldsymbol{X}) = \mathcal{N}(\boldsymbol{X}|\boldsymbol{\mu}_m^{(k)}, \boldsymbol{\Sigma}_m^{(k)}) : m = 1, \ldots, M_k\}$, is approximated by a single Gaussian pdf, where $M_k$ is the number of Gaussian components at node $k$. This pdf is called a *node* pdf. When it is assumed that the number of data samples from each mixture components are equal, the parameters for the node pdf are calculated as follows:

$$\mu_k(i) = \frac{1}{M_k}\sum_{m=1}^{M_k} E(x_m^{(k)}(i)) = \frac{1}{M_k}\sum_{m=1}^{M_k} \mu_m^{(k)}(i), \qquad (9)$$

$$\begin{aligned} \sigma_k^2(i) &= \frac{1}{M_k}\sum_{m=1}^{M_k} E((x_m^{(k)}(i) - \mu_k(i))^2) \\ &= \frac{1}{M_k}\left[\sum_{k=1}^{M_k} \sigma_m^{2(k)}(i) + \sum_{m=1}^{M_k} \mu_m^{(k)2}(i) - M_k\mu_k^2(i)\right], \quad (10) \end{aligned}$$

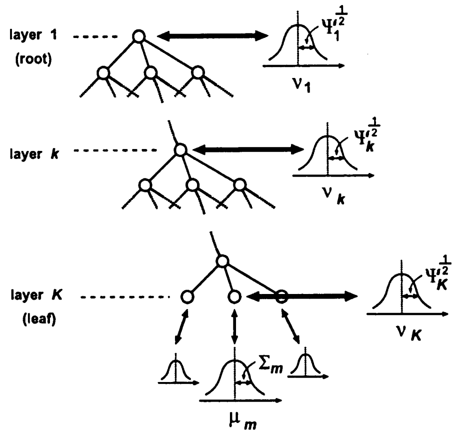

Figure 4 SMAP adaptation for Gaussian pdfs in CDHMMs. For simplicity, the case when the dimension is one (scalar) is shown.

where $\mathbf{x}_m^{(k)}$ is a data vector from Gaussian pdf $g_m^{(k)}$.

Then, the $k$-means clustering algorithm is used to construct a tree structure for the mixture components in $G$.

Finally, we have available a *tree structure* like the one shown in Figure 4, where $K$ is the total number of layers or the depth of the tree. Each node in the $K$-th layer (leaf node) corresponds to one Gaussian mixture component in the set of CDHMMs. The root node (the first layer) corresponds the whole set $G$ of the mixture components in the HMMs. Each intermediate node corresponds to a subset of $G$, and each of its subordinate leaf nodes corresponds to an element of a subset.

### 3.3 SMAP adaptation using hierarchical priors

At each node in the tree, a pdf, which is shared among the mixture components in the corresponding subset of $G$, is assigned. The ML estimates of the pdf parameters for each node in the tree are calculated using the adaptation data. From now on, the focus is on estimation of the parameter set, $\theta_m = (\boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$, for a particular $m$-th mixture component in $G$. The procedure described below is general and can be used to estimate the parameter sets of all the other mixture components in CDHMMs.

Let the node sequence from the root to the leaf corresponding to the $m$-th mixture component be $\{N_1, \ldots, N_k, \ldots, N_K\}$, where $N_1$ is the root node and $N_K$ is the leaf node directly attached to mixture component $m$. We denote $\lambda_k = (\boldsymbol{\nu}_k, \boldsymbol{\Psi}_k)$ as the mean vector and the variance of the Gaussian pdf for node $N_k$.

Here, the pdf for node $N_k$, which has the parameter set, $\lambda_k$, is assumed to have a hyperparameter, $\hat{\lambda}_{k-1}$, directly extended from its immediate parent node, $N_{k-1}$. Then, the MAP estimates, $(\boldsymbol{\nu}_k, \boldsymbol{\Psi}_k) = (\hat{\boldsymbol{\nu}}_k, \hat{\boldsymbol{\Psi}}_k)$, are calculated as follows: for $k = 1, \ldots, K$,

$$\hat{\boldsymbol{\nu}}_k = \frac{\Gamma_k \tilde{\boldsymbol{\nu}}_k + \tau_k \hat{\boldsymbol{\nu}}_{k-1}}{\Gamma_k + \tau_k}, \tag{11}$$

$$\hat{\boldsymbol{\Psi}}_k = \frac{\hat{\boldsymbol{\Psi}}_{k-1} + \Gamma_k \tilde{\boldsymbol{\Psi}}_k + \frac{\tau_k \Gamma_k}{\tau_k + \Gamma_k}(\tilde{\boldsymbol{\nu}}_k - \hat{\boldsymbol{\nu}}_{k-1})(\tilde{\boldsymbol{\nu}}_k - \hat{\boldsymbol{\nu}}_{k-1})^T}{(\xi_k - D) + \Gamma_k}, \tag{12}$$

where $\tilde{\boldsymbol{\nu}}_k$ and $\tilde{\boldsymbol{\Psi}}_k$ are ML estimates for node $k$, $\Gamma_k$ is the occupancy count for node $k$, and $\tau$ and $\xi$ are control parameters. The mean $\hat{\boldsymbol{\nu}}_K$ and the variance $\hat{\boldsymbol{\Psi}}_K$ for the leaf node $N_K$ are obtained by applying Eqs. (11) and (12) successively from the root node to the leaf node.

Equation (11) can be rewritten for the leaf node as follows:

$$\hat{\boldsymbol{\nu}}_K = \sum_{k=1}^{K} w_k \tilde{\boldsymbol{\nu}}_k, \tag{13}$$

where the weighting factor $w_k$ is

$$w_k = \frac{\Gamma_k}{\Gamma_k + \tau_k} \prod_{i=k+1}^{K} \frac{\tau_i}{\Gamma_i + \tau_i}. \tag{14}$$

The mean vector estimated using the SMAP method can be considered as a weighted sum of the ML estimates at the different layers of the tree. Two important characteristics of the weight, $w_k$, are highlighted in the following:

(1) The weight $w_k$ at node $N_k$ becomes larger as the amount of data at that node, $\Gamma_k$, becomes larger.

(2) The weight $w_k$ at node $N_k$ decreases as $k$ becomes smaller.

These properties are desirable for adaptation. When the amount of data is small, the ML-estimated parameters in the upper layers, which represent global transformation, are mainly responsible for the resulting pdf. And when the amount of data is large, the parameters in the lower layers, which represent localized transformation, predominate.

### 3.4 Discussion

This SMAP approach is quite general in its framework and can be easily applied to other adaptation methods. For example, SMAPLR, in which SMAP is applied to maximum *a posteriori* linear regression (MAPLR), was recently proposed and proved to be significantly better than MAPLR when the amount of adaptation data is extremely small [18].

The SMAP method described here uses a tree structure in the model parameter space. While many kinds of tree structures can be used for SMAP estimation, it is important to choose one which represents the similarity of the normalized pdfs of the mixture components well. Good results were obtained when the Kullback-Leibler divergence between mixture components was used as a measure of similarity in constructing the tree structure, but many other similarity measures can be used. Other structures reflecting the relationship between acoustic model parameters are also worth investigating.

## 4. Future Research

We have explained two approaches for robust acoustic modeling in the previous two sections. While they were applied to HMMs, it is clear that these approaches are easily applicable to most statistical models for speech signal.

One of the most promising models may be Dynamic Bayesian Networks (DBNs) (e.g. [2]). It can be regarded as an extension of HMMs and can deal with much more variations in speech. Examples of those variations are asynchronous input of more than one features and long-term dependencies among features in different time. These are thought to be very important to represent the inner structure of speech.

While there is a large possibility that this extension leads to better understanding of the nature of speech and hence more robust speech recognition, the parameter space in which we have to search the structure is much larger and hence much more computational resources are needed. In addition, as the model size increase and accordingly the number of parameters increases, the data insufficiency problem becomes more serious. It is clear that the data insufficiency problem addressed in this paper becomes more serious.

Since the amount of data is limited and often changed, it is important to have a framework which is robust against the change is the amount of data; we have to provide a method that autonomously control model complexity according to the amount of data. While, in this paper, we used MDL criterion, there are other methods that can be used for this purpose. One of the promising method is is Variational Bayes (VB) method [20]. By using the VB method, we can easily introduce related knowledge in the form of prior distribution.

We would like to refer to two other issues which seems to be important to tackle the problem. While in our adaptation method we use Kullback-Leibler Divergence for the distance measure between Gaussian distributions, the distance measures between more complicated probabilistic distributions such as mixture of Gaussians or HMMs should be explored. While the most useful measure in such situations is multi dimensional scaling, its solution is only locally optimal.

The other important issue is "divide and conquer" approach [8]. At present, only MFCCs and their dynamic features are used for speech features, but they may be not good enough to represent all the information needed to represent the inner structure of speech. Multi-stream features of different resolutions in time scale seems to be promising.

## References

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

[2] J. Bilmes, "Buried Markov Models: A Graphical-Modeling approach to Automatic Speech Recognition," *Computer Speech and Language*, vol. 17, No. 2, 2003.

[3] S. S. Chen and P. S. Gopalakrishnan, "Clustering via the Bayesian information criterion with applications in speech recognition," in *Proc. ICASSP-98*, Seattle, pp. 645-648, 1998.

[4] S. S. Chen, E. M. Eide, M. J. F. Gales, R. A. Gopinath, D. Kanevsky, and P. A. Olsen, "Recent improvements to IBM's speech recognition system for automatic transcription of broadcast news," in *Proc. ICASSP-99*, Phoenix, 1999.

[5] S. S. Chen and R. A. Gopinath, "Model selection in acoustic modeling," in *Proc. EuroSpeech-99*, Budapest, pp. 1087-1090, 1999.

[6] V. V. Digalakis and L. G. Neumeyer, "Speaker adaptation using combined transformation and Bayesian methods," *IEEE Trans. Speech Audio Processing*, vol. 4, no. 4, pp. 294-300, 1996.

[7] J. L. Gauvain and C.-H. Lee, "Maximum *a posteriori* estimation for multivariate Gaussian mixture observations of Markov chains," *IEEE Trans. Speech Audio Processing*, vol. 2, no. 2, pp. 291-298, 1994.

[8] C.-H. Lee, "From knowledge-ignorant to knowledge-rich modeling: A new speech research paradigm for next generation automatic speech recognition," in *Abstract Book of Interspeech2004* pp. 109-112, 2004.

[9] S. Hiroya, M. Honda, "Estimation of Articulatory Movements from Speech Acoustics Using an HMM-Based Speech Production Model," *IEEE Transactions on Speech and Audio Processing*, vol. 12, no. 2, pp. 175-185, 2004.

[10] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous-density hidden Markov models," *Computer Speech and Language*, vol. 9, pp. 171-185, 1995.

[11] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Information Technology*, vol. 30, no. 4, pp. 629-636, 1984.

[12] G. Schwartz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, pp. 461-464, 1978.

[13] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrica*, vol. 63, pp. 117-126, 1976.

[14] K. Shinoda and T. Watanabe, "Speaker adaptation with autonomous model complexity control by MDL principle," in *Proc. ICASSP-96*, Atlanta, pp.717-720, 1996.

[15] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech-97*, Rhodes, vol. 1, pp. 99-102, 1997.

[16] K. Shinoda and T. Watanabe, "MDL-based context-dependent subword modeling for speech recognition," *J. Acoust. Soc. Jpn.(E)*, vol. 21, no. 2, pp.79-86, 2000.

[17] K. Shinoda and C.-H. Lee, "A structural Bayes approach to speaker adaptation," *IEEE Trans. Speech Audio Processing*, vol. 9, no. 3, pp. 276-287, 2001.

[18] O. Siohan, T. A. Myrvoll, and C.-H.Lee, "Structural maximum *a posteriori* linear regression for fast HMM adaptation," *Proc. ISCA ITRW ASR2000 Workshop*, Paris, 2000.

[19] A. Tritschler and R. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian information criterion," in *EuroSpeech-99*, Budapest, pp. 679-682, 1999.

[20] S. Watanabe, Y. Minami, A. Nakamura, and N. Ueda, "Variational Bayesian Estimation and Clustering for Speech Recognition,", *IEEE Trans. Speech and Audio Processing*, vol. 12, pp. 365–381, 2004.

[21] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modeling," in *Proc. Human Language Technology*, pp. 307-312, 1994.

[22] G. Zweig, "Speech Recognition with Dynamic Bayesian Networks," PhD Thesis, Univ. of California, Berkeley, 1998.