Adaptive HMM Topology for Speech Recognition

Chuan-Wei Ting, Kuo-Yuan Lee and Jen-Tzung Chien

Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan 70101, ROC

Abstract

This paper presents an adaptive algorithm for compensating pronunciation variations in hidden Markov model (HMM) based speech recognition. The proposed method aims to adapt the HMM topology and the corresponding HMM parameters to meet the variations of speaker dialects. In adaptive HMM topology, two hypothesis test schemes are designed to detect whether a new speaking variation occurs in state/phone levels. The test statistics are approximated by the chi-square densities. A new HMM topology is automatically generated by a significance level. Simultaneously, the HMM parameters and their hyperparameters are updated by Bayesian learning of the newly-generated Markov models. The pronunciation variations are cope with by a dialect adaptive HMM topology. We develop the incremental algorithm for corrective training of HMM topology and parameters. Experiments on TIMIT database show that the proposed algorithm is substantially better than the standard HMM with comparable size of parameters.

Index Terms: Pronunciation variation, HMM topology, hypothesis test, Bayesian learning, speech recognition

1. Introduction

There is no doubt that the pronunciation variation is one of the critical issues, which affects the performance of speech recognition system. The variations come from different speakers, accents, genders, etc. A general approach to compensate these variations is to use additional HMMs to characterize different variations. In [5], some triphones were selected to represent possible variations and all triphones were modeled by the multi-path HMMs. Also, the independent component analysis [2] was adopted for compensating pronunciation variations. These methods obtained desirable performance, but the weakness is due to the huge size of multiple HMM parameters. In this study, we intend to maintain the goodness of speaker-independent recognition system, which captures the complete speaker characteristics, and simultaneously explore a corrective training algorithm to adapt HMM topology from new coming data. The model parameters are tuned gradually so that the parameter size is controlled for new coming data. The model parameters are tuned gradually.

very complicated. Also, a maximum likelihood (ML) approach [8] was designed to estimate HMM topology by using the successive state splitting. Two kinds of state splitting including temporal splitting and contextual splitting were considered for adaptive HMM topology. Moreover, a sharing approach [6] was presented for dynamic HMM configuration. The topology did not consider the inter-state characteristics. These two studies presented batch learning rather than incremental learning. In this study, we propose the incremental learning of HMM topology based on Bayesian theory and develop the hypothesis test solutions to automatically determine HMM topology and estimate HMM parameters.

2. HMM and Bayesian Learning

Assuming we are given a N-state continuous-density HMMs with parameters $\Lambda = \{\alpha, \lambda, A\}$, where $\alpha = \{\alpha_t\}_{t=1}^T$ is the initial state probability, $A = \{a_{ij}\}_{i,j=1}^N$ is the transition probability matrix and $\lambda$ contains mixture density parameters $\lambda_i = \{\omega_{ik}, \mu_{ik}, \Sigma_{ik}\}_{k=1}^K$. Let $X = \{x_t\}_{t=1}^T$ denote the set of observed training samples. The state probability density of sample $x_t$ is assumed by a multivariate Gaussian mixture model (GMM)

$$
p(x_t | \lambda_i) = \sum_{k=1}^K \omega_{ik} N(x_t | \mu_{ik}, \Sigma_{ik}) ,
$$

where $\omega_{ik}$ is the mixture weight with constraint $\sum_{k=1}^K \omega_{ik} = 1$ and $N(x_t | \mu_{ik}, \Sigma_{ik})$ denotes the $D$-dimensional Gaussian distribution with mean vector $\mu_{ik}$ and covariance matrix $\Sigma_{ik}$. Using Bayesian inference, the product of likelihood function

$$
p(X | \Lambda) = \prod_{t=1}^T \sum_{i=1}^N \pi_i \omega_{ij} N(x_t | \mu_{jk}, \Sigma_{jk}) ,
$$

and prior density with hyperparameters $\varphi$ [4]

$$
g(\Lambda | \varphi) = \prod_{i=1}^N \prod_{j=1}^{D} \prod_{r=1}^{K} \frac{1}{\Gamma(r-1)} \left[ \sum_{k=1}^K \alpha_{ij} \right]^{r-1} \frac{1}{\Gamma(r)} \left( \alpha_{ij} \right)^{r-1} g(\mu_{jk}, \rho_{jk}) ,
$$

is maximized for parameter estimation. In (2)-(3), $s = (s_1, s_2, \ldots, s_T)$ and $l = (l_1, l_2, \ldots, l_T)$ denote the state and mixture component sequences, respectively, $\eta_s$, $\eta_l$ and $\eta_{s|l}$ denote the hyperparameters of Dirichlet densities and the precision matrix $\rho_{jk} = \Sigma_{jk}^{-1}$ is used. The prior density of Gaussian mean and precision is assumed to be a normal-Wishart density.
where \( \alpha_{\mu} > 1 \), \( \tau_{\mu} > 0 \), \( m_{\mu} \) is the \( D \times 1 \) vector, and \( u_{\mu} \) is a \( D \times D \) positive definite matrix.

### 3. Adaptive HMM Topology and Parameters

Generally, new data collection contains unknown speaking styles/variations that the existing HMMs do not include. The HMM topology should be adapted to compensate these variations for speech recognition. To capture the characteristics of new coming data and preserve the goodness of existing HMMs, an adaptive algorithm of finding HMM topology and parameters are proposed and described in what follows.

#### 3.1 Adaptive HMM Topology

For HMM topology learning, we propose two hypothesis test schemes to decide whether a new Markov state or Gaussian component is introduced in the existing HMMs. The first step in learning procedure is to train HMMs using new data set and evaluate the similarities between original models and new models. The evaluation of similarity is performed either in state level or in Gaussian component level. In this way, the pronunciation variations in a phone caused by different speakers and their dialects are modeled. Figure 1 shows two cases of variations from the original topology.

![Figure 1: Different HMM topologies in phone level](image)

Using HMM, it is assumed that the speech frames in the same state are homogeneous compared to those in different states. From the viewpoint of factor analysis [9], the regularities and similarities in HMM states are sufficiently reflected by the factor loading matrices, or equivalently the sample covariance matrices. It turns out that two HMM states \((s_a, s_b)\) of original and new models are considered to be similar if their covariance matrices are equivalent. The hypothesis test theory is applied to determine the state topology. The null (similar) hypothesis is defined as \(H_0: \Sigma_{s_a} = \Sigma_{s_b}, s_a = s_b\) and the alternative (dissimilar) hypothesis is defined as \(H_1: \Sigma_{s_a} \neq \Sigma_{s_b}, s_a \neq s_b\). Under the null hypothesis \(\Sigma_{s_a} = \Sigma_{s_b}\), the likelihood function is yielded as

\[
L_c(H_0) = \prod_{s=s_a} (2\pi)^{T/2} \sum_{s} \exp \left\{ -\frac{1}{2} T \Sigma_{s}^{-1} x_{s,j} x_{s,j}^T \right\} \tag{4}
\]

where \( s = s_a, s_b \). The likelihood function under \(H_1\) is

\[
L_c(H_1) = \prod_{s=s_a} (2\pi)^{T/2} \sum_{s} \exp \left\{ -\frac{1}{2} T \Sigma_{s}^{-1} x_{s,j} x_{s,j}^T \right\} \tag{5}
\]

where \( s = s_a, s_b \). The likelihood function under \(H_1\) is

\[
L_c(H_1) = \prod_{s=s_a} (2\pi)^{T/2} \sum_{s} \exp \left\{ -\frac{1}{2} T \Sigma_{s}^{-1} x_{s,j} x_{s,j}^T \right\} \tag{6}
\]

Using likelihood ratio test, the variance-based test statistic is generated by

\[
q_v = \frac{L_c(H_0)}{L_c(H_1)} \sum_{s=s_a} \frac{\Sigma_{s}^{-1}}{\Sigma_{s}^{-1}} \tag{7}
\]

This test statistic is derived to be asymptotically approaching to a chi-square distribution with degree of freedom \(D(D+1)/2\) [1]. Given a significance level \(\alpha\), the threshold is obtained by \(X_{\alpha}^{2}(D(D+1)/2)\). Accordingly, we decide whether a new state should be created in current HMM. For the case of three-state HMM shown in Fig. 2, the first and third states in original and new models of a phone are tested to be merged together. As the second state of original and new model has significant difference, the characteristics of \(s_{a,2}\) and \(s_{b,2}\) are separately preserved. A new HMM topology is generated.

![Figure 2: Test of similarity in phone level](image)

Once the above-mentioned null hypothesis is accepted, the next step is to evaluate whether a new Gaussian density is added in current GMM in state level. We test that two Gaussian densities \(g_a\) and \(g_b\) are dissimilar if the means \(\mu_a\) and \(\mu_b\) and covariance matrices \(\Sigma_a\) and \(\Sigma_b\) are different. The null hypothesis is defined by \(H_0: \mu_a = \mu_b, \Sigma_a = \Sigma_b, g_a = g_b\), and the alternative hypothesis is defined by \(H_1: \mu_a \neq \mu_b, \Sigma_a \neq \Sigma_b, g_a \neq g_b\). We calculate the statistic of likelihood ratio \(q\) represented by \(q = q_v \cdot q_m\), where \(q_m\) is given by (7) but the state indices \(s_a\) and \(s_b\) should be replaced by mixture component indices \(g_a\) and \(g_b\). Also, \(q_m\) is the mean-based test statistic of likelihood ratio with \(H_0: \mu_a = \mu_b, g_a \neq g_b\) and \(H_1: \mu_a \neq \mu_b, g_a \neq g_b\). Under \(H_0\), \(\mu_a = \mu_b\), the likelihood function is expressed by

\[
L_m(H_0) = (2\pi)^{T/2} \left| \Sigma_{H_0} \right|^{-T/2} \exp \left\{ -\frac{1}{2} T \Sigma_{H_0}^{-1} x_{s,j} x_{s,j}^T \right\} \tag{8}
\]

where \( \Sigma_{H_0} = \sum_{x_{s,j} \in H_0} \sum_{x_{s,j} \in H_0} (x_{s,j} - \mu)^T (x_{s,j} - \mu) \), \( T = T_{s_a} + T_{s_b} \), and \( s \) denotes \( s_a \) or \( s_b \). The likelihood function under \(H_1\) is

\[
L_m(H_1) = (2\pi)^{T/2} \left| \Sigma_{H_0} \right|^{-T/2} \exp \left\{ -\frac{1}{2} T \Sigma_{H_0}^{-1} x_{s,j} x_{s,j}^T \right\} \tag{9}
\]
\[ L_m(H_1) = (2\pi)^{DT/2} \sum_{\mathbf{h}_l} \frac{e^{-\frac{1}{2} \sum_{i=1}^{T} \sum_{g=1}^{G} (\mathbf{x}_{g,i} - \mu_{g})^T \Sigma_{H_1}^{-1} (\mathbf{x}_{g,i} - \mu_{g})}} {T} \]

where \( \Sigma_{H_1} = (T_g \Sigma_{g_e} + T_f \Sigma_{g_f}) / T_g \). Test statistic is obtained by

\[ q = q_m = \frac{\sum_{i=1}^{T} \sum_{g=1}^{G} (\mathbf{x}_{g,i} - \mu_{g})^T \Sigma_{H_1}^{-1} (\mathbf{x}_{g,i} - \mu_{g})} {T} \]

The learning mechanism is illustrated in Fig. 3. When two HMM states \((s_g, s_f)\) of original and new models that are significantly similar, we further compare the Gaussians in two states. In this figure, solid curve shows the GMMs of original model and dash line shows the Gaussian mixture of new model. There are two Gaussian mixtures in \(s_g\) significantly different from the corresponding nearest Gaussian component in \(s_f\). Hence, we add two Gaussian mixtures in the new state and the optimal number of mixtures of this state is updated to 6.

Figure 3: Test of similarity in state level

### 3.2 Adaptive HMM parameters

A new topology of HMM is obtained by performing two test procedures. After the procedures, Bayesian inference is applied to determine the new topology of HMM. Let \(N'\) and \(K'\) denote the numbers of states and mixture components in updated HMM, where \(N' \geq N\) and \(K' \geq K\). Among \(N'\) states, only the parameters of the states accepting null hypothesis of the first hypothesis test are updated. Those parameters of mixture components are further updated if null hypothesis of the second hypothesis test is accepted. Those hyperparameters unseen in the new data are kept unchanged.

\[ \alpha_{jk} = \kappa_1 (\alpha_{jk} - D) + \kappa_2 (\hat{\alpha}_{jk} + D) \]

\[ \hat{\alpha}_{jk} = \kappa_3 \alpha_{jk} + \kappa_4 \hat{\alpha}_{jk} \]

\[ \beta_{jk} = \frac{\kappa_2 \tau_{jk} m_{jk} + \kappa_5 \tau_{jk} \mathbf{x}_{jk}} {\kappa_1 \tau_{jk} + \kappa_2 \tau_{jk}} \]

\[ \hat{\beta}_{jk} = \kappa_1 (\hat{\beta}_{jk} - D) + \kappa_2 (\hat{\beta}_{jk} + D) \]

\[ \gamma_{ij}(j,k) = p(s_i = j, t_i = k \mid X, \Lambda) \]

\[ \mathbf{x}_{jk} = \sum_{i=1}^{N} \gamma_{ij}(j,k) \mathbf{x}_{i} \]

\[ \hat{\mathbf{x}}_{jk} = \sum_{i=1}^{N} \hat{\gamma}_{ij}(j,k) \mathbf{x}_{i} \]

Using these updated hyperparameters, the parameters of new HMM topology are estimated by

\[ \hat{\alpha}_{jk} = \frac{1 - \hat{\beta}_{jk}} {\sum_{i=1}^{N} (1 - \hat{\beta}_{ij})}, \quad \hat{\beta}_{jk} = \frac{\hat{\beta}_{ij} - 1} {\sum_{j=1}^{K} (\hat{\beta}_{ij} - 1)}, \quad \hat{\gamma}_{ij} = \frac{\gamma_{ij} - 1} {\sum_{j=1}^{K} (\hat{\beta}_{ij} - 1)} \]

\[ \hat{\mu}_{jk} = \frac{\hat{\mu}_{jk} - \hat{\beta}_{jk}} {\sum_{i=1}^{N} (1 - \hat{\beta}_{ij})}, \quad \hat{\mu}_{jk} = \frac{\hat{\mu}_{jk} - \hat{\beta}_{jk}} {\sum_{j=1}^{K} (\hat{\beta}_{ij} - 1)} \]

\[ \hat{\sigma}_{jk} = \frac{\hat{\sigma}_{jk} - \hat{\beta}_{jk}} {\sum_{i=1}^{N} (1 - \hat{\beta}_{ij})}, \quad \hat{\sigma}_{jk} = \frac{\hat{\sigma}_{jk} - \hat{\beta}_{jk}} {\sum_{j=1}^{K} (\hat{\beta}_{ij} - 1)} \]

### 4. Experiments

#### 4.1 Experimental Setup

In the experiments, the proposed method is evaluated by phone recognition using TIMIT database. In this database, there were 6,300 utterances spoken by 630 speakers from eight major dialect divisions of the United States. All utterances were sampled by 16 kHz with 16 bit resolution. The 39-dimensional feature vector was extracted for each frame and composed of 12 MFCCs and one log energy, and their first and second derivatives. 41 models (39 phone models, one silence model, and one short pause model) were constructed in standard HMM. Each model was represented
by a three-state HMM. The number of mixture components in standard HMM was eight. The training data contained 4,620 utterances from 462 speakers and the core test data had 192 utterances from 24 speakers. For two hypothesis tests, the significance level is set to be $\alpha = 0.05$.

### 4.2 Adaptive HMM Topology

In training session, we divided training set and conducted adaptive learning for eight major dialects. The first learning epoch was performed to train initial HMM hyperparameters and the others were spent for incrementally updating hyperparameters and learning new pronunciation variations or topologies. Without loss of generality, we only consider the topology variations caused by the middle state of an HMM. The structures of the first and the last states are unchanged. The cases in Fig. 4 are generated when the states $s_{a,2}$ and $s_{b,2}$ were tested to be dissimilar. We split $s_{a,2}$ into two states ($s_{a,2}^1$, $s_{a,2}^2$) and evaluated the similarity between the split states and $s_{b,2}$. If either $s_{a,2}^1$ or $s_{a,2}^2$ is tested to be significantly similar to $s_{b,2}$, Figs. 4(a) and 4(b) are produced accordingly. Otherwise, the topology degenerates to the case given in Fig. 2. This learning rule attains both contextual and temporal variations of HMM topology. Even though the topology varies only in middle state, the new variations of first and last states are still valid by changing the structure of Gaussian mixtures. Notably, following the proposed learning rule, the HMM topology is adaptive for individual phones.

![Illustration of adaptive HMM topology](image)

### 4.3 Recognition Results

Figure 5 shows the phone accuracies of standard HMM and proposed adaptive HMM topologies using ML and MAP parameter estimation, which are denoted by AHMMT-ML and AHMMT-MAP, respectively. In standard HMM, the cases of different learning epochs are considered by using different amounts of data in ML batch training. In the MAP adaptation (denoted by HMM-MAP) and proposed algorithm, the case of conducting $m$ epochs means that the adaptive HMM topology and its parameters are estimated by using the data from epoch $m$ and the updated hyperparameters from previous $m-1$ epochs. The proposed learning algorithm starts from the baseline HMM with 8 mixture components (denoted by HMM-8). The HMM-MAP adopts 16 mixture components. The recognition accuracy of HMM-8 at the first epoch is 55.3%. After performing eight epochs, or using all training data, the accuracy of HMM-8 is increased to 59.7% and the proposed AHMMT-ML and AHMMT-MAP achieve 61.9% and 63.7%, respectively. AHMMT-MAP outperforms AHMMT-ML at different epochs. However, the adaptive topology learning algorithms greatly increase the size of HMM parameters. For comparative study, we also carry out the standard HMM with 16 mixture components (denoted by HMM-16), which has comparable model size as that in the proposed algorithms. AHMMT-MAP consistently attains higher accuracies than HMM-16. After eight epochs, HMM-16 obtains the phone accuracy of 60.9%, which is worse than those of AHMMT-ML and AHMMT-MAP. Without topology updating, HMM-MAP (59.58%) is not satisfied because new variations such as contextual or temporal variations are not properly characterized in updated HMM.

The proposed adaptive learning algorithm also avoids possible performance degradation by a new learning epoch, e.g. epoch 6, because the new variations have been considered by adapting HMM topology and parameters instead of only adapting HMM parameters.

![Figure 5: Accuracies of different methods at different epochs](image)

### 5. Conclusions

This paper presented an adaptive learning algorithm of HMM topology and parameters. Two hypothesis testing approaches were explored to decide whether a new state or mixture component should be created in current HMM structure. Accordingly, a Bayesian learning algorithm was applied to update the characteristics of new variations and simultaneously preserve the properties of current structure. Compared to the static HMM topology, the proposed dynamic HMM topology achieved the desirable speech recognition performance on using TIMIT database. In future studies, we are developing the pruning algorithm and building compact acoustic models for speech recognition.

### 6. References


