Decision Tree State Tying Using Cluster Validity Criteria

Jen-Tzung Chien, Senior Member, IEEE

Abstract—Decision tree state tying aims to perform divisive clustering, which can combine the phonetics and acoustics of speech signal for large vocabulary continuous speech recognition. A tree is built by successively splitting the observation frames of a phonetic unit according to the best phonetic questions. To prevent building over-large tree models, the stopping criterion is required to suppress tree growing. Accordingly, it is crucial to exploit the goodness-of-split criteria to choose the best questions for node splitting and test whether the splitting should be terminated or not. In this paper, we apply the Hubert’s $\Gamma$ statistic as the node splitting criterion and the $T^2$-statistic as the stopping criterion. The Hubert’s $\Gamma$ statistic sufficiently characterizes the clustering structure in the given data. This cluster validity criterion is adopted to select the best questions to unravel tree nodes. Further, we examine the population closeness of two split nodes with a significance level. The $T^2$-statistic expressed by an $F$ distribution is determined to verify whether the mean vectors of two nodes are close together. The splitting is stopped when verified. In the experiments of Mandarin speech recognition, the proposed methods achieve better syllable recognition rates with smaller tree models compared to the conventional maximum likelihood and minimum description length criteria.

Index Terms—Cluster validity, decision tree, $F$ distribution, continuous speech recognition, Hubert’s $\Gamma$ statistic, hypothesis test, $T^2$-statistic.

I. INTRODUCTION

In real-world speech recognition, it is important to deal with the issues of acoustic modeling covering wide ranges of phonetic units, as well as context dependencies to achieve desired performance for large-vocabulary continuous speech recognition (LVCSR). However, we could not collect sufficient training materials spanning huge amount of phonetic and contextual variations. For example, the triphone models are used to simultaneously represent the left and right context dependencies across different phonetic units [23], [37]. With the limited training data, a lot of triphone hidden Markov models (HMMs) are unseen to be estimated. To tackle the problem of data sparseness, it is inevitable to perform HMM state tying [18], [37]. Typically, the HMM state tying is performed either by bottom-up agglomerative clustering or top-down decision trees [24]. The agglomerative clustering ties the distributions of unseen phonetic units to obtain robust HMMs. Its drawback is that it is unable to estimate the parameters of unseen units.

The chi-square measure was used to merge the HMMs [29]. Contrarily, the decision tree based state tying is powerful to estimate the context-dependent HMM parameters for both seen and unseen units [3], [24]. The language-specific phonological rules can be incorporated to build tree models. With the decision trees, the model complexity is configurable according to the size and content of the collected training data [19]. The context dependencies of a decision tree could be grouped to the whole HMM or each state of an HMM [22]. In this paper, we focus on improving the flexibility and robustness in decision tree construction via validating the clustering data in tree nodes.

The construction of decision tree can be stated as follows. First of all, we prepare a set of phonetic questions as the prior knowledge of the phonetics. The specification of question set can be done manually or automatically [5]. Given the phonetic questions, the decision tree is initialized from the root node containing all context-independent observation frames associated with the same phonetic unit. This tree is built in top-down basis by successively splitting the tree nodes using the best phonetic questions. These questions separate the observation frames into two subsets for the answers of “yes” and “no.” The “don’t know” subtree was also grown when a question was difficult to answer [21]. To effectively execute node splitting, we need a meaningful goodness-of-split criterion to select the best questions. After splitting, the observations of two child nodes should be representative and discriminable as two individual clusters. When the tree is continuously growing, it is necessary to develop a stopping criterion to terminate node splitting. The leaf nodes are finally reached to complete the decision tree. Optionally, the tree models could be pruned to prevent over-training in the estimated tree models. The pruning should be performed via evaluating the tree models on new data [13]. The log-likelihood measure was used to evaluate the pruning performance [7], [22]. In general, the construction of decision tree consists of the following four essential components substantially affecting the speech recognition performance [27], [28]:

1) specification of phonetic unit and question set;
2) criterion for selecting the best splitting question;
3) criterion for stopping the node splitting;
4) criterion for pruning the over-large tree.

In this paper, we neglect the first issue since the phonetic units considering different context dependencies have been investigated [26], [33], [37]. The language-dependent phonetic questions have been well defined by the phoneticians. Also, the issue of pruning criterion is not mentioned. Our goal is to develop the cluster validity criteria to automatically select the best questions for splitting and statistically test the hypothesis of splitting as two child nodes. No pruning procedure is performed.
In the literature, the probabilistic models were popular for goodness-of-split evaluation. Bahl et al. [3] measured the improvement due to splitting by the difference of log likelihood between the original data and the split data. The question with the best likelihood improvement was selected for splitting. This method is referred as the maximum likelihood (ML) criterion. A unified ML was presented to incorporate phonetic and non-phonetic features into decision trees [31]. Assuming that the observations are identically and independently distributed, the measure of log likelihood is equivalent to the negative entropy of the output distributions. The increase of log likelihood means the reduction of model uncertainty. In [23], the difference of entropy represented the amount of information loss when merging two HMMs in agglomerative clustering procedure. The Poisson model was adopted to generate the clustering data [3]. This model was fast but inelegant because it ignored the evolution of speech signal over time [21]. Also, the Gaussian model was employed to characterize the continuous observations for tree nodes [19], [22], [38]. Nevertheless, the ML criterion has the drawback of building over-large decision trees because the likelihood is often increased after splitting. It becomes inevitable to prune tree models. On the other hand, the minimum description length (MDL) criterion [32] containing a model complexity penalty was introduced to optimize the degree of clustering [34]. The information-theoretic MDL criterion is identical to the statistical Bayesian information criterion (BIC) [1], which was applied to optimally select the phonetic models as well as determine the number of clusters [6].

ML and MDL are efficient to serve as node splitting criteria. They adopted different thresholds to fulfill stopping criteria during tree growing. Using ML based splitting, the size of decision tree was controlled by the increase of log likelihood and the amount of node data [3], [26], [38]. In MDL (or BIC) based splitting, a penalty factor was incorporated to balance the tradeoff between log likelihood and model complexity [11]. This factor was tunable to adjust tree size. When ML and MDL are employed in node splitting, the clusters of two child nodes are dissimilar in term of log likelihood or description length. However, we do not have a measure of confidence for how significant the clusters should be separate. Accordingly, this paper presents the novel node splitting and stopping algorithms for decision tree state tying based on cluster validity criteria. Different from the parametric ML and MDL methods, we adopt the non-parametric Hubert’s Γ statistic [17] to select the best node-splitting question. Hubert’s Γ statistic measures the degree of agreement between a clustered partition and the proximity matrix of training data. This measure is effective for goodness-of-split validation. To examine whether the splitting should be stopped, or equivalently, two child nodes are close together, we apply the $T^2$-statistic [16] to test the hypothesis of equivalence of two corresponding mean vectors. With a significance level, the node splitting is halted when the $T^2$-statistic of the clustering data is located in the acceptance region. In the experiments, the splitting criterion using Hubert’s Γ statistic and the stopping criterion using $T^2$-statistic are effective for decision tree state tying. The remaining of this paper is organized as follows. In Section II, we address the node splitting methods and explain how the Hubert’s Γ statistic is employed to select the best splitting question. Section III describes the stopping criterion via $T^2$-test of clustered mean vectors. In Section IV, the experimental setup is described. Several sets of evaluation are reported to compare the performance of various combined methods. Finally, the conclusions drawn from this paper are given in Section V.

II. NODE SPLITTING CRITERIA

First of all, the issue of node splitting criterion is addressed. Given a set of training data, we align all utterances via Viterbi decoder and distribute the feature vectors to their associated root nodes of phonetic units. The decision trees are built by successively selecting the best context-dependent question $q$ to split the $N$ observation frames $\{x_1, x_2, \ldots, x_N\}$ of a tree node $S$ into two child nodes corresponding to the answers of “yes” $S_{qy}$ and “no” $S_{qn}$. It is necessary to exploit the goodness-of-split criterion for node splitting.

A. ML and MDL Criteria

Traditionally, the ML criterion is popular to serve as splitting criteria. The use of ML for splitting tree nodes is consistent with that for training HMMs parameters. Let $L(S)$ denote the log likelihood of generating observation frames in node $S$. Assuming the observation frames of tree nodes are Gaussian distributed with shared mean vector and covariance matrix, the increase of log likelihood by splitting $S$ through question $q$ is derived by [38]

$$
\delta_q^{ML}(S) = L(S_{qy}) + L(S_{qn}) - L(S)
= -\frac{1}{2} \left[ C(S_{qy}) \log |\Sigma_{qy}| + C(S_{qn}) \log |\Sigma_{qn}| - C(S) \log |\Sigma(S)| \right]
$$

(1)

where $C(\cdot)$ and $\Sigma(\cdot)$ denote the total state occupancy count and the pooled covariance matrix of the observation frames in tree node, respectively. The question $q_{ML}$ with the maximum increase of log likelihood is chosen for splitting

$$
q_{ML} = \arg \max_q \delta_q^{ML}(S).
$$

(2)

On the other hand, the MDL criterion evaluates the splitting performance according to the description length, which is composed of a log likelihood term and a complexity penalty term [32]. The MDL theory was applied for unsupervised speaker adaptation under different amount of adaptation data [8]. Basically, the description length $D(S)$ is related to log likelihood $L(S)$ by

$$
D(S) = -\log f(S) + \frac{1}{2} m \log C(S) = -L(S) + d \log C(S)
$$

(3)

where $f(\cdot)$ is the likelihood function, $m$ is the number of free parameters and $d$ is the feature dimension. Here, the parameters in tree node $S$ contain a mean vector and a diagonal covariance matrix, i.e., $m = 2d$. The decrease of description length $\delta_q^{MDL}(S)$ due to splitting using question $q$ is calculated by

$$
\delta_q^{MDL}(S) = -\delta_q^{ML}(S) + d \cdot (\log C(S_{qy}) + \log C(S_{qn}) - \log C(S)).
$$

(4)
According to MDL, the best question $\hat{q}_{\text{MDL}}$ for splitting is chosen by maximizing the decrease of description length

$$\hat{q}_{\text{MDL}} = \arg \max_q (-s_q^{\text{MDL}}(S)).$$

The physical meaning of MDL aims to build the tree models with increasing likelihood and reducing model dimensionality. No matter whether we are using ML or MDL, only the difference of log likelihood or description length between the parent node and the child nodes is employed. The clustering structure between two individual child nodes is not considered. Actually, it is important to incorporate the intra-node as well as the inter-node statistics for goodness-of-split evaluation. To this end, the Hubert’s $\Gamma$ statistic is presented hereafter for node splitting.

**B. Hubert’s $\Gamma$ Statistic**

The splitting algorithm of a tree node is intended to choose a phonetic question and impose a clustering structure on tree node S even though S may not have such a structure. The importance of splitting is to discover an indication that the feature vectors of S form two clusters. Generally, the tasks of evaluating the results of a splitting/clustering algorithm are known as the cluster validity. The cluster validity using ML and MDL calculates the log likelihood and description length corresponding to father node and child nodes. These methods specify the parametric form of observation distribution and measure the fitness of observation vectors due to the splitting. Comparatively, we present a nonparametric Hubert’s $\Gamma$ statistic [17], [35] to explore the inherent clustering structure of data set $\{X_1, X_2, \ldots, X_N\}$. Its underlying concept aims at evaluating the degree of agreement between a predetermined partition and the inherent clustering structure. For the example of decision tree node, the partition is determined by a phonetic question $q$. The data partition of “yes” and “no” answers can be characterized via a $N \times N$ proximity matrix (dissimilarity matrix) $Y_q(S) = \{Y_{q}(i,j)\}$. Herein, the component $Y_{q}(i,j)$ is calculated with the Euclidean distance $d(\mu_i, \mu_j) = ||\mu_i - \mu_j||$ of the cluster mean vectors $\mu_i$ and $\mu_j$ corresponding to the observation data pair $(x_i, x_j)$. The cluster mean vectors of two subsets, $\{\mu_qy, \mu_qm\}$, are determined beforehand. In matrix $Y_q(S)$, only the values of 0 and $||\mu_qy - \mu_qm||$ exist. Consistently, the inherent clustering structure of S can be represented using a $N \times N$ proximity matrix X(S) where the component $X_{q}(i,j)$ is directly computed by the distance $d(x_i, x_j) = ||x_i - x_j||$ of data pair $(x_i, x_j)$. Matrix X(S) is independent of question $q$. The Hubert’s $\Gamma$ statistic of S using question $q$, $\Gamma_q(S)$, is obtained by measuring the correlation between two symmetric matrices $X(S)$ and $Y_q(S)$

$$\Gamma_q(S) = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} X_{q}(i,j) Y_{q}(i,j)$$

$$= \frac{2}{N(N-1)} ||\mu_qy - \mu_qm|| \sum_{(x_i, x_j) \in \Omega} ||x_i - x_j||.$$  

This function averages over all possible products of $X_{q}(i,j)$ and $Y_{q}(i,j)$ for $N(N-1)/2$ data pairs $(x_i, x_j)$. It turns out that only the distances of the data pairs associated with different clusters $\Omega$ are summed up. Basically, (6) has twofold meanings. First, the proximity matrices $X(S)$ and $Y_q(S)$ measure the distances of the observation vectors and the cluster mean vectors, respectively. These matrices are generic to cover both inter-class and intra-class statistics. Second, when the product of $X_{q}(i,j)$ and $Y_{q}(i,j)$ is large, it is very likely that the data points $(x_i, x_j)$ are apart and assigned to different clusters with distant mean vectors. Namely, the larger the statistic $\Gamma_q(S)$ is, the stronger the evidence that the tree node contains compact clusters. To this end, we select the best question for node splitting according to the largest Hubert’s $\Gamma$ statistic

$$\hat{q}_\Gamma = \arg \max_q \Gamma_q(S).$$

The selected question $\hat{q}_\Gamma$ is able to finely separate the tree node S into $S_{\hat{q}_\Gamma y}$ and $S_{\hat{q}_\Gamma m}$. In general, the Hubert’s $\Gamma$ statistic is robust because the inter-class and intra-class information is embedded for splitting evaluation. But, the Hubert’s $\Gamma$ statistic suffers from high computation load. In Fig. 1, the system diagram of decision tree construction using Hubert’s $\Gamma$ statistic based node splitting criterion and $T^2$-statistic based stopping criterion is plotted. The node splitting procedure is described as follows. For each nonleaf node S, we calculate the proximity matrices of training frames $X(S)$ and clusters $Y_q(S)$ corresponding to phonetic question $q$. The Hubert’s $\Gamma$ statistics $\Gamma_q(S)$ of all questions are determined so as to select the best question $\hat{q}$ and split S into child nodes $S_{\hat{q}_\Gamma y}$ and $S_{\hat{q}_\Gamma m}$.

**III. STOPPING CRITERIA**

The training samples of a decision tree are continuously split using the node-splitting criteria. The tree models are usually overspecialized and generalize poorly for the new data. It becomes crucial to exploit the stopping criterion to stop tree growing. The robustness of decision trees can be achieved without extra overhead of tree pruning process.

**A. Previous Methods**

To prevent over-training problem, the constraint of ML criterion was modified to restrict the increasing amount of log likelihood above a positive lower bound, i.e., $\hat{q}_{\text{ML}}(S) > T_\delta$ [3], [11], [38]. The threshold $T_\delta$ is critical to control the goodness of ML based decision trees. In MDL (or BIC) based splitting, a penalty factor was incorporated in description length to balance the tradeoff between the log likelihood and the model complexity

$$D(S, p) = -L(S) + p \cdot d \cdot \log C(S).$$

Using the penalty factor, it was feasible to build tree models with scalable tree size. The penalized MDL was reduced to standard MDL in case of $p = 1$. In [11], the penalty factor $p > 1$ obtained better speech recognition performance than $p = 1$. Also, it is popular to specify the floor number of observation frames in tree nodes $T_N$ when using ML and MDL. If the frame number is less than the floor value, the tree growing is terminated. In general, the parameters $T_\delta, T_N$ and $p$ are adjustable to control the model complexity. These parameters are determined empirically. The determination of proper values is sensitive to the changing conditions. We present the $T^2$-statistic based stopping criterion where the tree model complexity is tunable through a statistical significance level.
B. $T^2$-Statistic

More generally, the stopping criterion is herein considered as a statistical hypothesis test problem to investigate whether the data in $S$ are compact or not. If the hypothesis is verified with a significance level, the split should be terminated. Accordingly, we are examining the problem of which the null hypothesis $H_0$ for stopping splitting or the alternative hypothesis $H_1$ for continuing splitting is correct. The hypotheses are defined by

$H_0$: Observation samples in $S$ cannot be divided into two clusters.

$H_1$: Observation samples in $S$ can be divided into two clusters.

Our goal is to derive an appropriate statistical index for hypothesis test. The decision depends on the experimental evidence supporting the acceptance or not of $H_0$. If the observation data in $S$ differ significantly, the data sets in child nodes $S_{y}$ and $S_{n}$ determined by the best question $\hat{q}$ contain the compact clusters. The corresponding mean values should be far away. We may approach the problem by considering the corresponding mean vectors $\mu_y$ and $\mu_n$. For simplicity, we drop the script $\hat{q}$ in the following expressions. The hypotheses turn out to concern two multivariate populations

$H_0$: $\mu_y = \mu_n$ (stop splitting).

$H_1$: $\mu_y \neq \mu_n$ (continue splitting).

Here, the samples in $S_y$ and $S_n$ are drawn from the Gaussian distribution with mean vectors $\mu_y$ and $\mu_n$, respectively. The covariance matrices are assumed unknown and equal, $\Sigma(S_y) = \Sigma(S_n) = \Sigma$. In other words, if the mean values of two child nodes close together (i.e., $\mu_y = \mu_n$), the data samples in $S$ are sufficiently compact. It is plausible to accept the null hypothesis and stop the node splitting. More apparently, we define a new random vector $\mathbf{x}_z$, which is the difference of the independent random vectors in child nodes $S_y$ and $S_n$, i.e., $\mathbf{x}_z = \mathbf{x}_y - \mathbf{x}_n$. Its mean vector and covariance matrix are $\mu_z = \mu_y - \mu_n$ and $\Sigma_z = \Sigma$, respectively. The hypothesis test problem is further interpreted by testing whether the new mean vector $\mu_z$ equals to a specified value $\eta = 0$. The two-sample hypothesis test is rewritten by a one-sample $T^2$-test problem [16]

$H_0$: $\mu_z = \eta = 0$ (stop splitting).

$H_1$: $\mu_z \neq 0$ (continue splitting).

Using this verification approach, we can build the decision trees with cluster validated tree nodes. Similar hypothesis test criterion was employed in a greedy clustering algorithm for speaker identification and speech recognition [15], [20]. Their criterion assumed that the covariance matrices of child nodes were known. In this paper, we use the multivariate hypothesis test criterion for decision tree state tying. A more realistic case of unknown covariance matrix is considered.

Let $N_y$ and $N_n$ denote the numbers of $d$-variate observation vectors in two child nodes and $N = N_y + N_n$. The sample mean vectors of two nodes ($S_y$, $S_n$) are expressed by ($\bar{x}_y$, $\bar{x}_n$). Assuming all data vectors in $S_y$ and $S_n$ are independent, it is straightforward to derive the sample mean of the difference vector $\mathbf{x}_z = \bar{x}_y - \bar{x}_n$, being Gaussian distributed with mean vector $\mu_y - \mu_n$ and covariance matrix $((N_y + N_n)/N_yN_n)\Sigma$. In case of univariate hypothesis test of mean value with unknown variance $\sigma^2$, it is well known to use the test statistic

$$ T^2 = \frac{\eta}{\sqrt{N_y + N_n} \sigma_z}. \quad (9) $$

In (9), $\bar{x}_z$ and $\sigma_z^2$ represent the sample mean and sample variance of univariate random variable $s_z$, respectively. Similarly, in multivariate hypothesis test of mean vector with unknown covariance matrix, it is possible to determine the acceptance region $R_\alpha$ for null hypothesis $\mu_z = 0$ using the test statistic

$$ \frac{N_yN_n}{N_y + N_n} (\mathbf{x}_z - \eta)^T \Sigma_z^{-1} (\mathbf{x}_z - \eta) \quad (10) $$

where $\Sigma_z$ denotes the sample covariance matrix of random vector $\mathbf{x}_z$ given by

$$ \Sigma_z = \frac{1}{N_y + N_n - 2} \left\{ \sum_{i=1}^{N_y} (\mathbf{x}_y(t) - \bar{x}_y)(\mathbf{x}_y(t) - \bar{x}_y)^T + \sum_{i=1}^{N_n} (\mathbf{x}_n(t) - \bar{x}_n)(\mathbf{x}_n(t) - \bar{x}_n)^T \right\}. \quad (11) $$

Here, $(\mathbf{x}_y, \mathbf{x}_n, \Sigma_z)$ serves as complete sufficient statistic for parameters $(\mu_z, \Sigma)$. The $d \times d$ scaled sample covariance matrix $(N_y + N_n - 2)^{-1}\Sigma_z$ has a Wishart distribution $W(N_y + N_n - 2, \Sigma)$ with $N_y + N_n - 2$ degrees of freedom and parametric matrix $\Sigma$ [12]. Under null hypothesis $\eta = 0$, the test statistic is known as the (generalized) $T^2$-statistic [2]

$$ T^2 = \frac{N_yN_n}{N_y + N_n} \frac{\mathbf{x}_z^T \Sigma_z^{-1} \mathbf{x}_z}{\mathbf{x}_z^T \Sigma_z^{-1} \mathbf{x}_z}. \quad (12) $$

Its distribution was named by Hotelling’s $T^2$ distribution with $N_y + N_n - 2$ degrees of freedom [16]. Since the $T^2$-statistic can be expressed by

$$ \frac{N_yN_n}{N_y + N_n} \frac{\mathbf{x}_z^T \Sigma_z^{-1} \mathbf{x}_z}{\mathbf{x}_z^T \Sigma_z^{-1} \mathbf{x}_z}. \quad (13) $$

where the numerator $((N_yN_n)/(N_y + N_n))\Sigma_z^{-1} \mathbf{x}_z$ and the denominator $\Sigma_z^{-1} \mathbf{x}_z$ are independent and respectively distributed by the chi-square distributions with $d$ and $N_y + N_n - d - 1$ degrees of freedom, denoted by $\chi^2(d)$ and $\chi^2(N_y + N_n - d - 1)$ [30], the distribution of $T^2$-statistic is formulated by

$$ \frac{d}{N_y + N_n - d - 1} \sim \chi^2(d) $$

$$ \frac{d}{N_y + N_n - d - 1} \times F(d, N_y + N_n - d - 1). \quad (14) $$

This distribution has an elegant form of Snedecor’s $F$ distribution with $d$ and $N_y + N_n - d - 1$ degrees of freedom. $F(d, N_y + N_n - d - 1)$ [4], [12]. If the significance level is $\alpha$, defined by the probability of error

$$ \alpha = P(T^2 \notin R_\alpha | H_0) \quad (15) $$
for $T^2$-statistic locating in the region of rejecting $H_0$ when $H_0$ is true, the acceptance region of null hypothesis is finally decided by

$$R_{\alpha} : T^2 < \left( \frac{N_y + N_n - 2}{N_y + N_n - d - 1} \right) F_{\alpha}(d, N_y + N_n - d - 1) = \theta_{\alpha}.$$  \hfill (16)

The significance level is also referred as the \textit{Type I error} for hypothesis test. Given a significance level $\alpha$, the proposed stopping criterion is optimally established through calculating $T^2$-statistic and checking if belonging to the resulting acceptance region. The splitting is stopped when $T^2 < \theta_{\alpha}$. We continue node splitting when $T^2 \geq \theta_{\alpha}$. As shown in Fig. 1, the stopping criterion is fulfilled subsequent to the node splitting procedure. We calculate the $T^2$-statistic using the frames in $S_{qy}$ and $S_{qn}$. By computing $F_{\alpha}(d, N_y + N_n - d - 1)$ and comparing with $\theta_{\alpha}$, we make a decision of continuing splitting or terminating splitting. If the continuing splitting hypothesis is verified, we mark $S_{qy}$ and $S_{qn}$ as nonleaf nodes and separately repeat the procedure until the tree node being verified by the split stopping.

Kannan et al. [20] presented a hypothesis test of mean vector and covariance matrix for estimating the parameters of triphone models. Considering the test of mean vector, they used the likelihood ratio as a criterion [15]

$$\lambda_{\text{MEAN}} = \left( 1 + \frac{N_y N_n}{N_y + N_n} x_z^T \Sigma_z^{-1} x_z \right) ^{-(N_y + N_n)/2} \hfill (17)$$

to cluster the distribution mean vectors. Actually, Kannan’s criterion is referred as a test of mean vector \textit{under the assumption of known covariance matrix} $\Sigma$. To know this property, assuming the covariance matrix is known, the test statistic under $H_0$ has the form

$$\frac{N_y N_n}{N_y + N_n} x_z^T \Sigma_z^{-1} x_z < X_{\alpha}^2(d). \hfill (18)$$

The statistic has a chi-square distribution $\chi^2(d)$ with $d$ degrees of freedom. Correspondingly, the acceptance region of $H_0$ for significance level $\alpha$ is obtained by [14]

$$\frac{N_y N_n}{N_y + N_n} x_z^T \Sigma_z^{-1} x_z < X_{\alpha}^2(d). \hfill (19)$$

Notably, when $N$ is large, the test statistic of Kannan’s criterion is written by

$$-2 \log \lambda_{\text{MEAN}} = (N_y + N_n) \times \log \left( 1 + \frac{N_y N_n}{N_y + N_n} x_z^T \Sigma_z^{-1} x_z \right) \approx \frac{N_y N_n}{N_y + N_n} x_z^T \Sigma_z^{-1} x_z. \hfill (20)$$

Kannan et al. [20] aimed to test the equality of mean vectors given the known covariance matrix $\Sigma$. The sample covariance matrix $\hat{\Sigma}_z$ was pretended to be the true covariance matrix $\Sigma$. They applied the criterion $\lambda_{\text{MEAN}}$ and the chi-square statistic to choose similar classes of triphones for shared model parameters. For node splitting and stopping criteria, the increase of
likelihood $\delta_q^M(S)$ and the floor frame number of tree node $T_N$ were adopted, respectively. However, we propose the hypothesis test of mean vector without the assumption of known variance. More intuitively, $T^2$-statistic is formulated to construct the terminating criterion for node splitting. Hubert’s $\Gamma$ statistic is presented to serve as a node splitting criterion.

It is interesting to know the relations between $F$ and other distributions. The $F$ distribution with degrees of freedom $\nu_1$ and $\nu_2$, $F(\nu_1, \nu_2)$, is written by

$$
\Gamma \left[ \frac{(\nu_1+\nu_2)}{2} \right] \Gamma \left[ \frac{\nu_1}{2} \right] \Gamma \left[ \frac{\nu_2}{2} \right] \frac{\nu_1/2}{\nu_2} x^{(\nu_1/2)-1} \left( 1 + \frac{\nu_2}{\nu_1} \frac{x}{\nu_2} \right)^{-(\nu_1+\nu_2)/2}
$$

where $\Gamma(*)$ is gamma function. This distribution[4] has mean $\nu_2/(\nu_2 - 2)$ and variance $(2\nu_2^2/(\nu_2 + 2)) / (\nu_2(\nu_2 - 2)^2(\nu_2 - 4))$. In Fig. 2, we plot F distributions in cases of $\nu_1 = 26$ and different $\nu_2$. The Gaussian distribution with the same mean and variance as $F(26,100)$ is shown. Compared to Gaussian distribution, $F$ distribution is nonsymmetric and left-hand shifting with respect to its mean. The change of distribution shape is relatively small when $\nu_2$ is large. Also, the significance levels of $F$ and chi-square distributions have the property $E_{\nu_1}(\nu_2) \cong \chi^2_{\nu_1}(\nu_2) / \nu_2$ as $\nu_2 \rightarrow \infty$.[4] When we look at $F$ distribution (16) in $T^2$-statistic and chi-square distribution (19) in Kannan’s criterion, it is interesting to see that these two criteria converge asymptotically higher up in the tree where there is more data $N$. The differences of two criteria are present lower in the tree when there is limited data.

IV. EXPERIMENTS

A. Speech Databases, Mandarin Phonology and Implementation Issues

The proposed methods are evaluated by the recognition task of continuous Mandarin speech. The benchmark speech database of MAT-400 in Taiwan was used for training. It contained the Mandarin speech Across Taiwan (MAT) [36] for 400 speakers (216 males and 184 females) talking over the telephone networks. We sampled 20,800 utterances with 12.3 hours covering isolated syllables (MATDB-3), command words (MATDB-4) and continuous speech (MATDB-5), and hence generated the speaker and gender independent HMMs.

**TABLE I**

<table>
<thead>
<tr>
<th>Initials</th>
<th>Finalss</th>
</tr>
</thead>
<tbody>
<tr>
<td>b, p, m, f, d, t, n, l, g, k, h, j, q, x, zh, ch, sh, r, z, c, s, null initial</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>a (3304), ai (3938), au (2835), an (4024), ang (3293)</td>
</tr>
<tr>
<td>o</td>
<td>o (951), ou (2605)</td>
</tr>
<tr>
<td>e</td>
<td>e (56), ê (6798), ei (1725), en (2946), eng (2810)</td>
</tr>
<tr>
<td>er</td>
<td>er (1086)</td>
</tr>
<tr>
<td>i</td>
<td>i (6881), ia (1655), ie (2212), io (2330), iai (75), iau (2478), iou (2760), ian (3541), in (2515), iong (2088), ing (3094)</td>
</tr>
<tr>
<td>u</td>
<td>u (6026), uu (1220), uo (3263), uai (946), uei (3261), uan (2081), uen (1667), uang (1687), ung (3326)</td>
</tr>
<tr>
<td>ü</td>
<td>ü (2555), ũe (1529), ũan (1835), ũn (1030), ũʊng (1050)</td>
</tr>
</tbody>
</table>

The benchmark test speech database (Test500) were recorded via telephones and consisted of 500 sentences uttered by 15 males and 15 females, which were different from those of MAT-400. Test500 contained 4754 syllables and served as the common evaluation set for Mandarin speech recognition [10]. All utterances were sampled at 8 kHz with 16-bit resolution. Each frame was characterized by 12 Mel-frequency cepstral coefficients (MFCCs), 12 delta MFCCs, one delta log energy and one delta-delta log energy (i.e., $d = 26$). The sentence-based cepstral mean subtraction was applied for telephone channel normalization. The syllable language model was used for improving speech recognition performance.

Mandarin is a syllabic and tonal language. Without considering the tones, the number of base syllables is 408. Each base syllable can be divided into an initial part and a final part corresponding to a consonant and a vowel, respectively. When the syllable only has final part, a null initial exists in the beginning. Table I lists 22 initials and 38 finals in Mandarin syllables. Seven groups of finals are seen. Detail Mandarin phonetics was described in [25]. We employed the right context dependent (RCD) sub syllable (initial/final) modeling to construct the HMM units of Mandarin speech. The acoustic modeling of intra and inter syllable contextual dependencies is illustrated in Fig. 3. The intra and inter syllable dependencies were modeled through RCD initials and RCD finals, respectively [26]. RCD initials were group-dependent of 38 finals. RCD finals
were dependent of RCD initials of subsequent syllables. Cumulatively, there were 94 RCD initials, 2400 RCD finals and seven group-dependent null initials produced to build the HMMs of LVCSR system. We arranged the RCD initial, RCD final and null initial by three, six and two HMM states, respectively. Five states were used to characterize the pre-silence, post-silence and between-syllable silence. Due to lack of training data, front three states of RCD finals were context independent. Only rear three states were dependent. The number of mixture components was empirically chosen according to the number of frames in a state. Herein, the HMM modeling of Mandarin speech followed the well-known work done by Lee [25], [26]. We did not investigate the optimal choice of context dependencies.

Without decision tree state tying, we collected 7615 HMM states. To significantly reduce the number of HMM states, we prepared 31 consonant phonetic questions [26] and built 38 decision trees for Mandarin finals instead of using 2400 RCD finals. The state tying was done in HMM level rather than state level. The syllable recognition rates and the number of trained HMM states were reported for evaluation. The computation costs in training as well as recognition phases were evaluated by simulating the algorithms on Sun Workstation with model Ultra 10. For comparison, we realized the baseline acoustic modeling using 22 context independent (CI) initials and 38 CI finals. The resulting syllable recognition rate was 48.9%. The number of trained HMM states was only 375. In CI system, we used eight states to model the syllable with only final part. When we applied 94 RCD initials and 38 CI finals, the syllable recognition rate and the number of HMM states were 51% and 502, respectively. To examine the proposed decision tree state tying, we show the established decision tree for final “a” in Fig. 4. The best question and the numbers of frames and HMM units are displayed at each tree node. The observation frame numbers of 38 decision trees are provided in Table I.

Here, the frame numbers were measured from the MATDB-4 subset. In the following, we compare the performance of node splitting criteria using ML, MDL, and Hubert’s $\Gamma$ statistic and stopping criteria using Kannan’s criterion and $T^2$-statistic. The entire MAT-400 corpus was common to build the decision trees when applying different combined algorithms.

B. Performance Comparison of ML and MDL Using Different Thresholds

First of all, we compare the performance of node splitting criteria using ML and MDL. The differences of log likelihood $\delta_q^{ML}(S)$ and description length $\delta_q^{MDL}(S)$ due to splitting are applied for question selection. Using ML, the splitting is stopped according to the increasing amount of log likelihood.
TABLE II
COMPARISON OF TREE SIZE AND SYLLABLE RECOGNITION RATE WHEN USING ML AND MDL SPLITTING CRITERIA WITH FLOOR FRAME NUMBER PER NODE \( T_N \) AS STOPPING PARAMETER

<table>
<thead>
<tr>
<th>( T_N )</th>
<th>50</th>
<th>150</th>
<th>250</th>
<th>350</th>
<th>450</th>
<th>550</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trained states</td>
<td>5142</td>
<td>3988</td>
<td>2020</td>
<td>1539</td>
<td>1157</td>
<td>996</td>
</tr>
<tr>
<td>Syllable recognition rate (%)</td>
<td>54.6</td>
<td>55.5</td>
<td>56.0</td>
<td>55.2</td>
<td>54.8</td>
<td>53.7</td>
</tr>
<tr>
<td>MDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trained states</td>
<td>4857</td>
<td>3766</td>
<td>1898</td>
<td>1312</td>
<td>1009</td>
<td>887</td>
</tr>
<tr>
<td>Syllable recognition rate (%)</td>
<td>55.1</td>
<td>56.3</td>
<td>57.7</td>
<td>56.4</td>
<td>55.5</td>
<td>54.2</td>
</tr>
</tbody>
</table>

TABLE III
COMPARISON OF TREE SIZE AND SYLLABLE RECOGNITION RATE WHEN USING ML AND MDL SPLITTING CRITERIA WITH THE INCREASING LOG LIKELIHOOD \( T_s \) AND PENALTY \( p \) AS STOPPING PARAMETERS, RESPECTIVELY

<table>
<thead>
<tr>
<th>( T_s )</th>
<th>0</th>
<th>200</th>
<th>400</th>
<th>600</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trained states</td>
<td>2020</td>
<td>1832</td>
<td>1507</td>
<td>1097</td>
<td>688</td>
</tr>
<tr>
<td>Syllable recognition rate (%)</td>
<td>56.0</td>
<td>56.7</td>
<td>55.0</td>
<td>53.5</td>
<td>52.4</td>
</tr>
<tr>
<td>MDL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of trained states</td>
<td>2853</td>
<td>1898</td>
<td>1720</td>
<td>1267</td>
<td>710</td>
</tr>
<tr>
<td>Syllable recognition rate (%)</td>
<td>57.0</td>
<td>57.7</td>
<td>58.0</td>
<td>56.4</td>
<td>54.9</td>
</tr>
</tbody>
</table>

This and the floor number of observation frames \( T_N \). Using MDL, the thresholds of \( T_N \) and penalty factor \( p \) are used for stopping evaluation. When we set \( T_s = 0 \) and \( p = 1 \), the stopping evaluation using ML and MDL depends only on the floor frame number per node \( T_N \). In Table II, we list the syllable recognition rates and the numbers of trained HMM states for ML and MDL based decision trees. Different thresholds \( T_N \) are investigated. It is obvious that the MDL consistently obtains smaller amount of states (lower model complexity) and higher recognition rates than ML for different thresholds \( T_N \). The best recognition rates are obtained in case of \( T_N = 250 \). In this case, ML and MDL methods achieve the syllable recognition rates of 56% and 57.7% and produce 2020 and 1898 HMM states, respectively. Accordingly, we fix \( T_N = 250 \) for subsequent evaluation. When the stopping parameters \( T_s \) and \( p \) are investigated, the performance comparison is illustrated in Table III. We find that the recognition rates are slightly raised to 56.7% and 58% for the cases of \( T_s = 200 \) using ML and \( p = 1.5 \) using MDL, respectively. As compared with those without decision tree state tying, the recognition results using ML and MDL criteria are significantly improved. But, the disadvantage is to induce larger amount of HMM parameters and higher computation costs. Different methods are examined by training time in hours and recognition time in seconds. Training time means the time of executing one iteration of HMM training. The recognition time per sentence is averaged over all test data. We can see that using decision tree state tying consumes much more time cost than that without decision tree. In general, the training overhead using decision tree state tying arises from the larger number of HMM states, the best question selection and the split stopping evaluation. Among them, the question selection for node splitting is the most time-consuming work. The recognition time is mainly affected by the size of trained HMM states.

C. Effects of Hubert’s \( \Gamma \) Statistic and \( T^2 \)-Statistic

Next, we evaluate the effects of using Hubert’s \( \Gamma \) statistic as node splitting criterion and \( T^2 \)-statistic as stopping criterion. For node splitting, the Hubert’s \( \Gamma \) statistic investigates the goodness-of-fit for the partitioned data set using a phonetic question. In Table IV, we compare the number of HMM states and the syllable recognition rate by using Hubert’s \( \Gamma \) statistic for node splitting. The split is terminated according to the floor number of observation frames \( T_N \). Subsequently, the recognition rate is increased to 60% but with 3253 trained states under \( T_N = 150 \). In case of \( T_N = 250 \), the recognition rate slightly changes to 59.3%. The number of HMM states is significantly reduced to 1811. Compared to ML and MDL based node splitting, Hubert’s \( \Gamma \) statistic chooses more suitable question for splitting so as to attain higher recognition performance. As listed in Table VI, the Hubert’s \( \Gamma \) statistic suffers from higher computation burden because of the extensive calculation of proximity matrix \( \mathbf{X}(S) \) and statistic \( \Gamma_q(S) \). Further, the comparison is investigated by simultaneously applying Hubert’s \( \Gamma \) statistic for node splitting and \( T^2 \)-statistic for stopping control. Using \( T^2 \)-statistic, the hypothesis test problem for the closeness of cluster mean vectors is
TABLE IV

<table>
<thead>
<tr>
<th></th>
<th>Number of trained states</th>
<th>Syllable recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hubert’s $\Gamma$ statistic, $T_N = 150$</td>
<td>3253</td>
<td>60%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ statistic, $T_N = 250$</td>
<td>1811</td>
<td>59.3%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ statistic, $T_N = 350$</td>
<td>1291</td>
<td>58.1%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ &amp; $T^2$-statistics, $\alpha = 0.15$</td>
<td>1457</td>
<td>62.9%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ &amp; $T^2$-statistics, $\alpha = 0.1$</td>
<td>1295</td>
<td>64.9%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ &amp; $T^2$-statistics, $\alpha = 0.05$</td>
<td>1013</td>
<td>61.9%</td>
</tr>
<tr>
<td>Hubert’s $\Gamma$ &amp; $T^2$-statistics, $\alpha = 0.01$</td>
<td>752</td>
<td>60.2%</td>
</tr>
</tbody>
</table>

TABLE V

<table>
<thead>
<tr>
<th></th>
<th>Number of trained states</th>
<th>Syllable recognition rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>ML &amp; $T^2$-statistic, $\alpha = 0.1$</td>
<td>1831</td>
<td>57.3%</td>
</tr>
<tr>
<td>ML &amp; $T^2$-statistic, $\alpha = 0.05$</td>
<td>1647</td>
<td>58%</td>
</tr>
<tr>
<td>ML &amp; $T^2$-statistic, $\alpha = 0.01$</td>
<td>1026</td>
<td>56.1%</td>
</tr>
<tr>
<td>MDL &amp; $T^2$-statistic, $\alpha = 0.1$</td>
<td>1692</td>
<td>59.1%</td>
</tr>
<tr>
<td>MDL &amp; $T^2$-statistic, $\alpha = 0.05$</td>
<td>1441</td>
<td>60.1%</td>
</tr>
<tr>
<td>MDL &amp; $T^2$-statistic, $\alpha = 0.01$</td>
<td>917</td>
<td>56.3%</td>
</tr>
</tbody>
</table>

considered. The decision thresholds $\theta_\alpha$ with significance levels $\alpha = 0.15, 0.1, 0.05$ and 0.01 are examined. As demonstrated in Table IV, the best recognition rate 64.9% is achieved in case of $\alpha = 0.1$. The resulting number of trained states is as limited as 1295. This result is better than those of other researchers obtained on the same data sets [10], [33] and our state-of-art system reported in [9]. When reducing the significance level, the complexity of decision trees is reduced as well. Although the recognition performance is altered by different significance levels, we highlight on the reduction of model complexity using $T^2$-statistic. In case of $\alpha = 0.01$, only 752 tied HMM states can attain the recognition rate of 60.2%. The recognition cost is decreased.

To conduct complete evaluation, we also carry out the experiments on using ML and MDL as node splitting criteria and $T^2$-statistic as stopping criterion. The performance comparison with $\alpha = 0.1, 0.05$ and 0.01 is reported in Table V. Using ML, the best recognition rate 58% is obtained for $\alpha = 0.05$ in $T^2$-statistic criterion. For the MDL criterion, we still obtain better recognition rate 60.1% under the case of $\alpha = 0.05$. These results are superior to those of ML using $T_N = 250$, $T_\delta = 200$ and MDL using $T_N = 250$, $p = 1.5$ as stopping control parameters because higher recognition rates and smaller size of acoustic models are obtained. The superiority of using $T^2$-statistic is obvious. Subsequently, we investigate the training time and storage requirements for ML, MDL, and proposed Hubert’s $\Gamma$ and $T^2$-statistics. The results are illustrated in Table VI where the training time breaks down into the processing times for node splitting, stopping evaluation, HMM parameter estimation and other computation. The storage requirements expressed in kilobyte (KB) are estimated when the procedure goes to execute the node splitting for root node of final “a.” We can see that node splitting is the most time-consuming task in the whole procedure. The training time using Hubert’s $\Gamma$ statistic is much higher than those using ML and MDL. The storage costs of three methods are comparable. In
our implementation, we do not need to store the matrices $X(S)$ and $Y_d(S)$ for calculation of Hubert’s $\Gamma$ statistic. Finally, in Tables VII and VIII, we compare the model complexity, training time, recognition time and recognition rate for several combinations of node splitting and stopping criteria. To further investigate the stopping criteria, we report the results of ML, MDL and Hubert’s $\Gamma$ statistic combined with Kannan’s criterion. Using Kannan’s criterion, the chi-square statistic is computed for evaluation of the equality of mean vectors. This method is valid under the assumption of known covariance matrix. The stopping criterion using $T^2$-statistic is built without this assumption. In the table, when MDL is combined with Kannan’s crite-

---

**TABLE VI**

**Comparison of Storage Requirement and Training Time for Node Splitting, Stopping Evaluation, HMM Estimation and Other Computation Using ML, MDL and Hubert’s $\Gamma$ & $T^2$-Statistics**

<table>
<thead>
<tr>
<th></th>
<th>Training Time</th>
<th>Storage Requirement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Node</td>
<td>Stopping</td>
</tr>
<tr>
<td></td>
<td>splitting</td>
<td>evaluation</td>
</tr>
<tr>
<td><strong>ML, $T_N = 250$, $T_0 = 200$</strong></td>
<td>3.2 hr</td>
<td>0.1 hr</td>
</tr>
<tr>
<td><strong>MDL, $T_N = 250$, $p = 1.5$</strong></td>
<td>3.3 hr</td>
<td>0.1 hr</td>
</tr>
<tr>
<td><strong>Hubert’s $\Gamma$ &amp; $T^2$-statistics, $\alpha = 0.1$</strong></td>
<td>5.4 hr</td>
<td>0.8 hr</td>
</tr>
</tbody>
</table>

---

**TABLE VII**

**Comparison of Number of Trained States for CI Initial/CI Final, CD Initial/CI Final and Decision Tree Methods Combined by Various Node Splitting and Stopping Criteria. Number in Bracket Denotes the Corresponding Training Time**

<table>
<thead>
<tr>
<th>Method</th>
<th>$T_N$</th>
<th>Kannan’s criterion</th>
<th>$T^2$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ML</strong></td>
<td>1832 (6.1 hr)</td>
<td>1681 (6.7 hr)</td>
<td>1647 (6.8 hr)</td>
</tr>
<tr>
<td><strong>MDL</strong></td>
<td>1720 (6.1 hr)</td>
<td>1533 (6.6 hr)</td>
<td>1441 (6.8 hr)</td>
</tr>
<tr>
<td><strong>Hubert’s $\Gamma$ statistic</strong></td>
<td>3253 (9.5 hr)</td>
<td>1310 (8.5 hr)</td>
<td>1295 (8.7 hr)</td>
</tr>
<tr>
<td><strong>Simplified Hubert’s $\Gamma$ statistic</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>1423 (7.2 hr)</td>
</tr>
</tbody>
</table>

---

**TABLE VIII**

**Comparison of Syllable Recognition Rates for CI Initial/CI Final, CD Initial/CI Final and Decision Tree Methods Combined by Various Node Splitting and Stopping Criteria. Number in Bracket Denotes the Corresponding Recognition Time**

<table>
<thead>
<tr>
<th>Method</th>
<th>$T_N$</th>
<th>Kannan’s criterion</th>
<th>$T^2$-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>ML</strong></td>
<td>56.7% (1.92 sec)</td>
<td>57.1% (1.54 sec)</td>
<td>58% (1.52 sec)</td>
</tr>
<tr>
<td><strong>MDL</strong></td>
<td>58% (1.77 sec)</td>
<td>57.9% (1.41 sec)</td>
<td>60.1% (1.39 sec)</td>
</tr>
<tr>
<td><strong>Hubert’s $\Gamma$ statistic</strong></td>
<td>60% (2.98 sec)</td>
<td>63.1% (1.09 sec)</td>
<td>64.9% (1.07 sec)</td>
</tr>
<tr>
<td><strong>Simplified Hubert’s $\Gamma$ statistic</strong></td>
<td>N/A</td>
<td>N/A</td>
<td>62.5% (1.32 sec)</td>
</tr>
</tbody>
</table>
rion for $\alpha = 0.05$, we can achieve comparable recognition rate 57.9% using lower model complexity. The number of trained states is reduced from 1720 to 1533 by applying Kannan’s criterion. In addition, the incorporation of $T^2$-statistic is able to attain higher recognition rate 60.1% and smaller number of states 1441 compared to that of Kannan’s criterion. $T^2$-statistic is superior to Kannan’s criterion for stopping tree growing. Although the performance is changed for various significance levels $\alpha$, we think that the proposed $T^2$-statistic adopts statistically reasonable threshold $\theta_T$. This parameter should be less sensitive to training data compared to using $T_N$, $T_\delta$, and $p$. Furthermore, the highest recognition rate 64.9% is achieved by combining Hubert’s $\Gamma$ and $T^2$-statistics. This result is better than 63.1% of the combined Hubert’s $\Gamma$ statistic and Kannan’s criterion with $\alpha = 0.1$. The recognition time is moderate. The larger training time is spent when calculating Hubert’s $\Gamma$ statistic $\Gamma_\mu(S)$. To reduce the training time, we also present a simplified calculation of Hubert’s $\Gamma$ statistic for node splitting. Namely, when using a question to split a large set of observations, we randomly select half set of observations to calculate the Hubert’s $\Gamma$ statistic. The best question is selected according to the simplified statistic. We find that the training is shrunk from 8.7 to 7.2 using this simplification. The syllable recognition rate is degraded to 62.5%, which is still better than ML and MDL methods.

V. CONCLUSION

This paper proposed the cluster validity schemes to build decision trees for continuous speech recognition. Different from conventional ML and MDL methods, the Hubert’s $\Gamma$ statistic was applied to select the best question for node splitting. This approach validated the clustering structure by measuring the degree of agreement between a predetermined partition and the inherent clustering structure. Further, the $T^2$-statistic was presented to test the hypotheses of closeness of the mean vectors of two child nodes. By solving such statistical hypothesis test problem, we exploited the optimal stopping criterion for decision tree construction. Using proposed approaches, we did not perform the tree pruning procedure. The tree models were established without the constraint of floor frame number for tree nodes. In the experiments of Mandarin speech recognition, we modeled the RCD intra and cross syllable coarticulation and shared the HMM states of RCD finals through the decision trees. The decision tree splitting using MDL was better than that using ML. The incorporation of Hubert’s $\Gamma$ statistic further improved the recognition performance. When jointly adopting Hubert’s $\Gamma$ statistic for node splitting and $T^2$-statistic for stopping evaluation, the syllable recognition rates were additionally increased with reduced model complexity. These results demonstrated the superiority of using Hubert’s $\Gamma$ statistic and $T^2$-statistic for construction of phonetic decision trees. Although $T^2$-statistic is herein developed as stopping criterion, it is also meaningful to adopt the statistic for question selection. In the future, we are trying to apply the proposed methods to resolve other clustering problems in general pattern recognition. To validate the performance, we will conduct the tests to evaluate the statistical significance of the results.

ACKNOWLEDGMENT

The author wishes to acknowledge the three anonymous reviewers who provided valuable comments which considerably improved the quality of this paper.

REFERENCES


---

**Jen-Tzung Chien** (S’97–A’98–M’99–SM’04) was born in Taipei, Taiwan, R.O.C., on August 31, 1967. He received the Ph.D. degree in electrical engineering from the National Tsing Hua University (NTHU), Hsinchu, Taiwan, in 1997.

He was an Instructor in the Department of Electrical Engineering, NTHU, in 1995. In 1997, he joined the Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, where he was an Assistant Professor. Since 2001, he has been an Associate Professor in NCKU. He was the Visiting Researchers at the Speech Technology Laboratory, Panasonic Technologies Inc., Santa Barbara, CA, in summer of 1998, the Department of Computer Science, Tokyo Institute of Technology, Tokyo, Japan, in summer of 2002, and the Institute of Information Science, Academia Sinica, Taiwan, in summers of 2003, 2004. His research interests include statistical pattern recognition, speech recognition, speaker adaptation, language modeling, face recognition/verification, biometrics, document classification, multimedia information retrieval and multimodal human-computer interaction. He serves on the board of the International Journal of Speech Technology and was a guest editor of the Special Issue on Chinese Spoken Language Technology.

Dr. Chien is a member of the International Speech Communication Association and the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). He serves on the board of ACLCLP. He was listed in Who’s Who in the World, 2002, and was awarded the Outstanding Young Investigator Award (Ta-You Wu Memorial Award) from National Science Council, Taiwan in 2003, and the Research Award for Junior Research Investigators from Academia Sinica, Taiwan, in 2004.