Robust Decision Tree State Tying for Continuous Speech Recognition

Wolfgang Reichl and Wu Chou, *Member, IEEE*

***Abstract*—In this paper, methods of improving the robustness and accuracy of acoustic modeling using decision tree based state tying are described. A new two-level segmental clustering approach is devised which combines the decision tree based state tying with agglomerative clustering of rare acoustic phonetic events. In addition, a unified maximum likelihood framework for incorporating both phonetic and nonphonetic features in decision tree based state tying is presented. In contrast to other heuristic data separation methods, which often lead to training data depletion, a tagging scheme is used to attach various features of interest and the selection of these features in the decision tree is data driven. Finally, two methods of using multiple-mixture parameterization to improve the quality of the evaluation function in decision tree state tying are described. One method is based on the approach of k-means fitting and the other method is based on a novel use of a local multilevel optimal subtree. Both methods provide more accurate likelihood evaluation in decision tree clustering and are consistent with the structure of the decision tree. Experimental results on Wall Street Journal corpora demonstrate that the proposed approaches lead to a significant improvement in model quality and recognition performance.**

***Index Terms*—Acoustic modeling, decision tree state tying, speech recognition.**

I. Introduction

DECISION tree state tying based acoustic modeling has become increasingly popular for modeling speech variations in large vocabulary speech recognition [1], [10], [19], [24], [25]. In this approach, the acoustic phonetic knowledge of the target language can be effectively incorporated in the model according to a consistent maximum likelihood framework. The statistical framework of decision tree in acoustic modeling provides two major advantages over the previous rule or bottom up based approaches. First, the classification and prediction power of the decision tree allows to synthesize model units or contexts, which do not occur in the training data. Second, the node splitting procedure of decision tree based state tying is a model selection process. It provides a way of maintaining the balance between model complexity and the number of parameters in order to render a robust model parameter estimation from the limited amount of training data.

Recently, there are many attempts to improve the phonetic decision tree state tying based approach in acoustic modeling [2], [4], [8], [11], [12], [16], [17], [23]. While some authors [4], [12], [16] concentrate on the problem of constructing improved trees, others try to generate optimal sets of questions automatically [2], [23]. Two problems in decision tree state tying are of particular interest. One is the tree growing and node splitting problem and it concerns the issue of how to find an optimal node split, given the particular parametric form of the impurity function (e.g., the likelihood of the training data). Another one is the parametric modeling problem of the data distributions during the process of decision tree node splitting. For phonetic decision tree based acoustic modeling, these two problems are closely related. The problem of optimal node splitting is about finding the best node split, and the parametric modeling is a problem of providing an appropriate metric, which defines the quality of the split. In general, construction of a globally optimal decision tree is a computationally intractable problem. The parametric forms of distributions used in decision tree node splitting are often based on Gaussian distributions, although more accurate multiple-mixture Gaussian distributions are used in the final acoustic model. This disparity is due in part to the computational complexity in the decision tree clustering process. The multiple-mixture Gaussian distribution for each tree node needs to be re-estimated from the data, whereas the parameters of the single-mixture Gaussian distribution can be derived from the cluster members without going back to the training data.

In this paper, we discuss methods for improving the robustness and accuracy in decision tree clustering based acoustic modeling. The novel contributions of this paper are as follows.

- A new segmental two-level clustering algorithm is devised. It combines the phonetic decision tree based state tying with agglomerative clustering to improve model coverage on rarely seen acoustic phonetic events in the training data.
- We present a unified maximum likelihood framework to incorporate generalized phonetic and nonphonetic features in decision tree based state tying. It is data driven and solves the problem of training data depletion in condition dependent acoustic modeling.
- A tagging scheme is introduced in decision tree based state tying and is used to tag various features of interest. In our proposed approach, the tagged information are used in conjunction with phonetic questions during decision tree state tying.
- Two applications of using the above mentioned general tagged features and acoustic phonetic questions in decision tree state tying are given. One application is for gender-dependent acoustic modeling and the other

Manuscript received April 22, 1999; revised September 19, 1999. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Rafid A. Sukkar.

The authors are with Bell Laboratories, Lucent Technologies, Murray Hill, NJ 07974 USA (wuchou@research.bell-labs.com).

Publisher Item Identifier S 1063-6676(00)06920-0.

one is for word boundary dependent acoustic modeling. Comparisons are made and performance advantages are demonstrated.

- We present two approaches of using multiple-mixture Gaussian parameterization to improve the likelihood evaluation function in decision tree node splitting. One is based on a special k-means fitting algorithm and the other one is based on m-level optimal subtree. Both approaches are extensions of the conventional one-level optimal single-mixture Gaussian based approach.
- A short-list based caching scheme is described which significantly reduces the computational complexity in using multiple-mixture Gaussian parameterization in decision tree construction. Experimental evidences are given that multi-level optimal tree building procedure is computationally feasible and advantageous in large vocabulary speech recognition.

The organization of this paper is as follows. We introduce the basic structure of decision tree based acoustic modeling in Section II. The two-level segmental clustering approach for robust decision tree state tying is described in Section III. The unified maximum likelihood framework for incorporating generalized features is presented in Section IV. In Section V, we introduce two approaches of using multiple-mixture Gaussian parameterization in decision tree construction. The theoretical basis and the short-list based caching scheme are described in detail. Section VI is devoted to experimental results and comparisons are made to various known approaches. Finally, we summarize our findings in Section VII.

II. Decision Tree State Tying

One approach to deal with the data sparseness problem in training acoustic models involves sharing of models across different contexts to form the so-called generalized triphones [13]. This model based sharing can be further improved to handle the left and right contexts independently and leads to state based sharing of parameters [14]. However these techniques use only *a priori* phonetic knowledge and are not supported by the actual training data. Although agglomerative clustering procedures are used to automatically determine the tying of states from data and result in high recognition performance [24], the problem of modeling unseen or rarely seen acoustic contexts in the training data remains. Decision tree based state clustering is shown to lead to similar and often better performance in large vocabulary speech recognition [10], [25]. It integrates both *a priori* phonetic knowledge and acoustic similarities derived from data. In decision tree clustering, single mixture Gaussian models are trained first and the phonetic decision tree is used to establish the state tying. One decision tree is constructed for each state of each center phone and all the context dependent states of this phone are clustered into groups by the decision tree algorithm. The resulting clusters of tied states are then retrained and multiple-mixture Gaussian distribution HMMs are estimated. In [12], [17] a joint decision tree for all states of each center phone was introduced, but it was found that the additional questions about the state positions immediately separated the trees and the single tree for a given phone was nothing more than the previous trees joined by additional nodes [12].

Usually, a decision tree is built using a top-down sequential optimization procedure (e.g. classification and regression tree, or CART [3]) starting from the root node of the tree. Each node is split according to the phonetic question which results in the maximum increase in the likelihood on the training data. The gain in likelihood due to a node split can be calculated efficiently from pre-calculated sufficient statistics of the affected states [19], [24]. The process is repeated until the likelihood gain falls below a threshold. A minimum occupation count is often applied to ensure that all terminal nodes have sufficient training data associated with them. Different sets of phonetic questions have been investigated in [11], [12] and good recognition results were obtained using questions about phonetic features and contexts. Methods of improving the quality of the set of questions were also proposed, and additional questions can be added through phoneme or diphone clustering [2] or by a multipass procedure, which adds the intersections of simple questions of a previously generated decision tree to the question set used in the next pass [23]. Additional stopping criteria based on cross-validation have been investigated in [11], [12], [16]. Typically an over-grown tree is constructed first and then the tree is pruned back by merging terminal nodes with different parents, if the likelihood decrease due to node merging is less than a preset stopping threshold. In phonetic decision tree clustering, a set of HMM states is recursively partitioned into subsets according to the phonetic questions at each tree node when traversing the tree from the root to its leaves. States reaching the same leaf node of the decision tree are regarded as similar and tied. The missing triphones are constructed by answering the phonetic questions for the missing triphone and traversing the decision tree from the root node to a final leaf. The most similar leaf node determined by the decision tree is used to synthesis the unseen triphone.

Since the log-likelihood of the training data $L(S) = \log P(X|S)$ generated from a tree node S can not be easily calculated, the common EM auxiliary function [7] is used as objective of the clustering

$$Q(S) = \int_u P(u|X, S) \log P(X, u|S)\, du \tag{1}$$

where u denotes the unobserved data and $X = \{x_1, \cdots, x_T\}$ is the sequence of observation vectors from the training data. Assuming a single mixture Gaussian distribution $N(x|\mu(S), \Sigma(S))$ for the node S, the unobserved data is the sequence of HMM states and the auxiliary function becomes

$$Q(S) = \sum_{x_t} \sum_{s \in S} \gamma_s(x_t) \log N(x_t|\mu(S), \Sigma(S)) \tag{2}$$

where $\gamma_s(x_t)$ is the *a posteriori* probability of the observation x_t at time t generated from state s. Using basic properties of Gaussian distributions, then [25]

$$Q(S) = -\tfrac{1}{2} \sum_{x_t} \sum_{s \in S} \gamma_s(x_t) \left[\log|\Sigma(S)| + D(\log(2\pi) + 1)\right] \tag{3}$$

where D is the dimensionality of the data vector and $|\Sigma(S)|$ denotes the determinant of the covariance matrix $\Sigma(S)$. Because of the monotonic relation between auxiliary function and likelihood: $Q(\hat{S}) \geq Q(S) \Rightarrow L(\hat{S}) \geq L(S)$, the sequential optimization of the auxiliary function in the decision tree clustering also results in the optimization of the likelihood function. Therefore the auxiliary function can be used as objective in the decision tree. By using the single mixture Gaussian assumption for the cluster distribution, the likelihood variation of clustering can be efficiently evaluated for every tree node S based on the already available sufficient statistics of its member states without additional need to access the training data. Each phonetic question splits the states into two subsets S_{yes} and S_{no} and therefore partitions the acoustic space. The question with maximum increase in the auxiliary function $\Delta Q_q(S) = Q(S_{yes}) + Q(S_{no}) - Q(S)$ is selected to split the node. The quality of the decision tree based state tying depends on the parametric form of the distribution used in evaluating $Q(S)$, which should approximate as closely as possible to the multiple-mixture Gaussian distribution used in the final model. The single Gaussian parameterization for cluster distributions in the conventional decision tree based state tying only provides a very limited acoustic resolution and may become inadequate to model the acoustic variability in the training data.

III. Two-Level Segmental Clustering

Previous studies in decision tree tying based acoustic modeling focus mainly on how to incorporate decision tree tying in Baum–Welch based parameter estimation [1], [25]. These approaches are based on the Baum–Welch algorithm to estimate the HMM parameters. However a Viterbi alignment based segmental clustering approach is more consistent with the decoding process used in recognition. The advantages of the segmental k-means training procedure were presented in [18]. For Viterbi training, it is important to note that once the state alignment is given, computations for estimating each individual HMM state become independent from each other. This makes it possible to fully parallel the model training process on multiple CPUs, and on large data set, the training time can be reduced from days to hours.

Mismatch in data alignment is one of the major causes which degrades the robustness and precision of acoustic modeling. Since the decision tree clustering is based on the fundamental assumption that tying of states will not change the alignment of training data, the initial alignment based on untied triphones with Gaussian state observation densities in the standard Baum–Welch training may not represent correctly the final models based on tied states and multiple-mixture distributions. Thus the single-Gaussian model alignment may not be accurate and can provide poor estimates for clustering [16]. In this section, we present a new robust two-level clustering approach for Viterbi based HMM training. It consists of an initial grouping of rare triphones and a subsequent decision tree clustering for state tying.

In addition to the data alignment problem, robust estimation of rarely seen triphones is another critical issue. In the standard method, a single Gaussian, untied triphone system is built first. It is common that only the mean vectors are estimated from data, whereas the variances are smoothed with the mono-phone models. However, the single Gaussian, untied triphone system forms the basis of the decision tree, and estimation errors introduced in the single Gaussian, untied system often have a long term adverse effect to the quality of the decision tree based state clustering.

One of the issues in using Viterbi alignment in decision tree based acoustic modeling is how to make a robust use of the rarely seen triphone samples in the training data. In Baum–Welch based parameter estimation, all possible paths are considered, and it has a much stronger smoothing effect on the parameters of those rarely seen triphones which have only very few training samples in the training data. In Viterbi alignment based segmental clustering approach, only the best path is considered and parameter estimates of these rarely seen triphones can degenerate very quickly with the decrease of training samples. In order to make full use of the training data and improve the robustness of the decision tree based state tying, a two-level segmental clustering scheme is devised in our approach. The first level segmental clustering is performed before forming the single-Gaussian, untied system. It is to cluster those rarely seen triphones into various types of generalized triphones [13] according to their phonetic similarities, so that the number of samples in each of the clustered generalized triphones is above the minimum sample count threshold required for the estimation of the sufficient statistics of the initial, untied states. A low sample count threshold (e.g., five or ten) can be used for a robust estimate of the single state mean and diagonal covariance matrix. The rare triphones are grouped by relaxing the triphone contexts [14]. First, the left contexts of rare triphones are relaxed, and if there are not sufficient samples in the training data to build these models, the right contexts of the rare triphones are disregarded. The second level clustering is a top-down decision tree based clustering of states according to phonetic questions. The phonetic identity of each generalized triphone from the first level clustering is defined to be the intersects of the phonetic properties of all rare triphones in the cluster.

The two-level clustering approach described above takes the advantage of the robustness of generalized triphone at the stage of forming a robust, single-Gaussian, untied system to improve the quality of the subsequent decision tree. The final model is still decision tree tied, in which state tying is determined solely by the likelihood increase on the training data. This is very different from the conventional generalized triphones, where tying is determined purely by the phonetic contexts. In addition, the unseen triphones are always synthesized according to the decision tree without making reference to the generalized triphones.

One advantage of applying a segmental clustering based approach in decision tree state tying is that segmentation of the training data is separated from the model parameter estimation process. Therefore more accurate mixture Gaussian models can be used to provide high quality data alignment, which will lead to more precise estimates of the likelihoods used in the decision tree construction and improve the quality of the final acoustic models. In our approach the decision tree is refined iteratively during the training process, which provides a more precise estimation of state tying. In each iteration, the training data is re-segmented by the Viterbi algorithm using the tied state models generated from the previous iteration. The convergence property of the segmental k-means [18] approach ensures that training data alignment will improve and converge with this iterative process.

For a set of states, S, sharing one common Gaussian distribution and using Viterbi alignment, the auxiliary function is given by

$$Q(S) = \sum_{x_t\,:\,s_t \in S} \log N(x_t|\mu(S), \Sigma(S)) \tag{4}$$

where all observation vectors x_t with a state alignment $s_t \in S$ are considered. In this case the auxiliary function and the log-likelihood are identical and for a single-Gaussian distribution it is

$$Q(S) = L(S) = -\frac{n(S)}{2}[\log|\Sigma(S)| + D(\log(2\pi) + 1)] \tag{5}$$

where $n(S)$ is the number of observation vectors assigned to the states associated with node S.

In the proposed segmental clustering algorithm, a multiple-mixture Gaussian distribution is estimated directly for each tied state before realigning the training data. This differs from many Baum–Welch based approaches, where multiple-mixture distributions are obtained by iterative binary splitting of each Gaussian density function (mixing-up) and data realignment. The block diagram of the proposed two-level segmental clustering training algorithm is illustrated in Fig. 1. The algorithm terminates after a predetermined number of iterations or if the likelihood gain falls below a certain threshold.

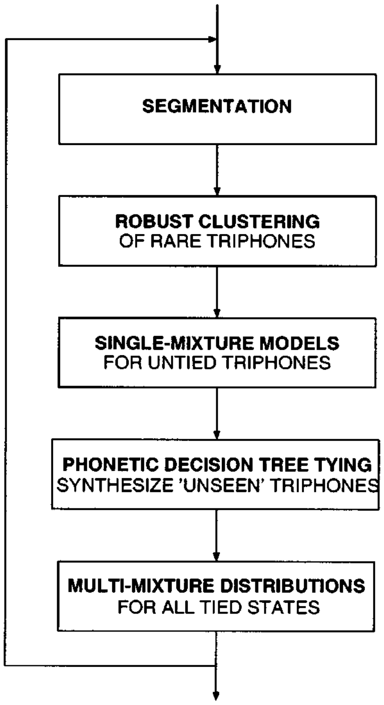

SEGMENTATION
ROBUST CLUSTERING OF RARE TRIPHONES
SINGLE-MIXTURE MODELS FOR UNTIED TRIPHONES
PHONETIC DECISION TREE TYING SYNTHESIZE 'UNSEEN' TRIPHONES
MULTI-MIXTURE DISTRIBUTIONS FOR ALL TIED STATES

Fig. 1. Block diagram of two-level segmental clustering algorithm.

IV. General Features in Decision Tree Based Acoustic Modeling

In speech recognition, many nonphonetic features are used to improve the resolution of the acoustic model and to obtain high recognition performance. Examples of such features include gender, speaker or speaker group identity, speaking rate, channel and environmental conditions, ambient noise level, etc. However, these features are not phonetic features and it has been a problem of how to incorporate them consistently with phonetic features in high-resolution acoustic modeling. The common practice is to manually separate the data according to the specification of the nonphonetic features, such as gender, and retrain a model using only the data which posses these features. This approach has two major problems. First, it depletes the amount of available training data as the number of nonphonetic features increases and puts a limit on the number of nonphonetic features that can be incorporated in the model. In addition, there is no data sharing between various conditions. As a consequence, the model may become poorly estimated and the performance of the model can degrade if data become too sparse after splitting. Secondly, the feature selection process is empirical and heuristic. Some nonphonetic features may influence only certain part of the model. For example, gender difference has more influence on vowels and diphthongs but probably less on stops or fricatives. Usually all model units are retrained based on the selected subset of data and there is no consistent theoretical framework to incorporate nonphonetic features.

Moreover, splitting of training data reduces the available training samples for each acoustic model, which can lead to poorly estimated models. For this reason, other specific models beyond gender conditions are rarely used [15], [21]. One solution to train specific HMMs for different features is through model adaptation techniques such as maximum *a posteriori* (MAP) adaptation [9]. In this approach, generic and condition independent models are estimated first and then adapted to the specific conditions. Although MAP adaptation is useful, it does not change the state tying relations of the generic model. The state tying relations of the generic model may not reflect specific features in the adaptation data. Individual states in the generic model may be separated or tied together according to the likelihood estimation and occupation counts from the complete, unconditioned data set, which may not be optimal for specific conditions. This situation can become acute when the properties of training and adaptation data are substantially different.

For these reasons, it is preferable to use an automatic and unified approach to generate specific acoustic models for different features in a data-driven manner. In [17] several linguistic and phonetic features such as vowel stress were incorporated into the decision tree clustering. In our approach, these features are generalized to include any additional information which may influence the configuration of the model in decision tree state tying. This is achieved by incorporating various features into the decision tree clustering based on a unified maximum likelihood statistical framework. Individual states for different features are only separated if this leads to a significant increase in the likelihood on the training data. The additional information about specific features are provided as tags to the decision tree clustering procedure. In our experiments, gender and word boundary tags are used. The tagging procedure in our approach partitions the training data into specific subsets, based on both the phonetic context and tagged features of interest, upon which initial single-mixture Gaussian models are trained. Consequently, every HMM state is associated with an appropriate label, marking the specific conditions of its training data. The question set of the decision tree is extended to also include questions regarding nonphonetic features. The nonphonetic features, such as gender, channel condition, speaking rate, etc., are those features which can not be derived from their phonetic contexts. During the construction of the decision tree, the best (phonetic or nonphonetic) question is selected to split the tree nodes according to the likelihood criterion [17], [21].

The tagged nonphonetic features are used simultaneously with the regular phonetic features in the decision tree clustering process during model construction. Therefore, the decision tree based model building is according to two types of knowledge sources and there is no manual separation of the training data. As a consequence, the model is built from the same set of training data regardless the number of nonphonetic features which we intend to incorporate. Thus, it solves the training data depletion problem as in prior data separation approaches. Moreover, these generalized features are incorporated in the decision tree based state tying according to a unified and consistent maximum likelihood framework. If separation of data with specific conditions results in a maximum likelihood gain among all other questions, separate HMM states will be constructed for these specific conditions. If no question about a particular feature is used on the path from the root tree node to a particular leaf node, the associated tied state to that leaf node is independent of that feature.

This data-driven approach prevents unnecessary data separation and allows maximum data sharing among various conditions. It constructs a minimum set of states for the given training data and a pre-selected likelihood threshold. As a consequence, the decision tree state tying is extended from tying states with various phonetic contexts to tying states with generalized phonetic and nonphonetic (e.g., gender, position, etc.) features. This leads to a significant increase in the amount of training samples for condition dependent acoustic modeling and the robustness of the condition dependent model is also enhanced. Moreover, there is no hard limit on the number of conditions to which the model can incorporate and the whole process is data driven. The proposed tagging approach is very general and can be used for many other features, such as speaker identity, age group, etc.

V. Decision Tree Clustering Based on Multiple-Mixture Gaussian Distribution

The quality of the decision tree based state tying depends on the parametric form of the distribution used in evaluating the impurity function, which should approximate as closely as possible the multiple-mixture Gaussian distributions used in the final model. The conventional single-mixture Gaussian parameterization for cluster distributions in decision tree based state tying only provides a very limited acoustic resolution and may become inadequate to model the acoustic variability in the training data. Moreover, the use of different likelihood functions in state tying and decoding also introduces a mismatch and violates the assumption that state alignments are unchanged for untied and tied system. This suggests that using multiple-mixture Gaussian distribution, instead of the single-mixture Gaussian, to evaluate the likelihood function should be advantageous.

In this section we will present two methods to improve the quality of decision tree clustering. The quality of the likelihood estimates used during the tree construction is improved by the usage of multiple-mixture Gaussian distributions for each set of states. A k-means based algorithm is described first, and then we present another approach which is based on a novel use of an optimal subtree to partition the acoustic space of a cluster node.

A. K-Means Based Multiple-Mixture Clustering

In this subsection we present an approach using the k-means algorithm to approximate multiple-mixture Gaussian distributions in decision tree state tying. The phonetic questions are further used in the decision tree to partition the initial, untied states into various subsets upon which the likelihood objective function has to be evaluated. A k-means clustering is utilized within each partition to obtain the required M mixture components. The multiple-mixture distributions are consistently used for the likelihood calculation from the initial untied states to the final tied states.

First, we derive the decision tree objective function assuming that all M mixtures are already known. For a multiple-mixture distribution with $m = 1, \cdots, M$ mixture components, the unobserved data of the auxiliary function consists of the state and mixture sequence in the model and the auxiliary function (1) becomes

$$\begin{aligned} Q(S) = &\sum_{x_t}\sum_{s\in S}\sum_{m} P(s, m|x_t) \\ &\cdot \log(c_m N(x_t|\mu_m(S), \Sigma_m(S))) \\ = &\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\sum_{m} P(m|x_t)\log c_m \\ &+\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\sum_{m} P(m|x_t) \\ &\cdot \log N(x_t|\mu_m(S), \Sigma_m(S)) \end{aligned} \tag{6}$$

where

$$P(m|x_t) = \frac{c_m N(x_t|\mu_m(S),\, \Sigma_m(S))}{\sum\limits_{m} c_m N(x_t|\mu_m(S),\, \Sigma_m(S))} \tag{7}$$

denotes the *a posteriori* probability of a mixture component given the observation. Using the mixture weights c_m derived from the Baum–Welch reestimation equations

$$c_m = \frac{\sum\limits_{x_t}\sum\limits_{s\in S} P(m|x_t)\gamma_s(x_t)}{\sum\limits_{x_t}\sum\limits_{s\in S}\gamma_s(x_t)} \tag{8}$$

the auxiliary function becomes

$$\begin{aligned}Q(S) = &\sum_m c_m \log c_m \sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\\ &+\sum_m\sum_{x_t}\sum_{s\in S}P(m|x_t)\gamma_s(x_t)\\ &\cdot\log N(x_t|\mu_m(S),\,\Sigma_m(S))\\ =&\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\left[\sum_m c_m\log c_m-\tfrac{1}{2}\sum_m c_m\right.\\ &\left.\cdot\log|\Sigma_m(S)|-\tfrac{1}{2}\,D\log(2\pi)\right]\\ &-\tfrac{1}{2}\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)\sum_m P(m|x_t)(x_t-\mu_m(S))^T\\ &\cdot\Sigma_m^{-1}(S)(x_t-\mu_m(S)). \end{aligned}\tag{9}$$

Applying basic properties of Gaussian distribution, we derive

$$\begin{aligned}Q(S) = -\tfrac{1}{2}\sum_{x_t}\sum_{s\in S}\gamma_s(x_t)&\left[\sum_m c_m\log|\Sigma_m(S)|-2\sum_m c_m\right.\\ &\left.\cdot\log c_m + D(\log(2\pi)+1)\right].\end{aligned}\tag{10}$$

In case of a segmental approach based on Viterbi alignment, the auxiliary function simplifies to

$$\begin{aligned}Q(S) = -\frac{n(S)}{2}&\left[\sum_m c_m\log|\Sigma_m(S)|-2\sum_m c_m\log c_m\right.\\ &\left.+D(\log(2\pi)+1)\right].\end{aligned}\tag{11}$$

In this formulation, the contribution of each mixture component to $Q(S)$ is the $\log|\Sigma_m(S)|$ term multiplied by the mixture weight c_m plus an additional entropy-like term $\sum_m c_m\log c_m$ which takes into account the contribution of the mixture weights to the log-likelihood.

In our implementation, up to M_0 Gaussian mixture components are estimated for every untied state from the training data. However, the number of mixtures can not be set too large since the available data samples for the untied states vary highly from state to state. During the process of decision tree state tying, the multiple-mixture Gaussian parameterization is carefully maintained and applied to every possible node splitting according to a set of phonetic questions. For this purpose, a k-means clustering is adopted in our approach to fit a M-mixture Gaussian probability density function for each tree node. The objective of this clustering process is to find an approximate multiple-mixture Gaussian distribution based on the multiple-mixture Gaussian parameterization of the untied states in that tree node. The process starts with M seed mixture Gaussians and performs a k-means clustering algorithm to form the required M-mixture Gaussian parameterization for the node. The distance function used is the same log-likelihood metric used in decision tree clustering and it measures the loss in likelihood by merging two Gaussian distributions. This k-means clustering process is sub-optimal but it is very efficient for the purpose of providing a multiple-mixture Gaussian distribution fitting for each tree node encountered in decision tree based state tying. The auxiliary or log-likelihood function of the tree nodes in decision tree based state tying is calculated based on this fitted multiple-mixture Gaussian distribution without coming back to the original acoustic training data.

One important property of this k-means based approach is that the structure of the phonetic decision tree is well maintained and therefore, it has the same prediction capability as the decision tree generated from the single-mixture Gaussian approach. This is different from other data driven k-means based approaches, such as the CPA algorithm [5] used in [16]. In CPA based approach, the locally optimal partitions (splits) are calculated based on the k-means algorithm, but no phonetic questions are used. This leads to a tree without phonetic properties and the prediction power of the phonetic decision tree is therefore lost. As a consequence, in order to construct unseen triphones, an additional "pre-tree" is used in [16]. Our approach is based on phonetic questions to partition the states, and k-means clustering is utilized within each partition to obtain M mixture components. A different approach to evaluate the impurity function for multiple-mixture Gaussian distributions was presented in [8]. There, a criterion to measure the overlap between Gaussian mixture pdfs was developed and applied to semi-continuous HMMs. To avoid extremely unbalanced trees, an additional normalization of the criterion was necessary. Our criteria is based on the likelihood objective function and does not prefer unbalanced trees.

B. Multilevel Optimal Trees for Multiple-Mixture Clustering

In this subsection, we describe another decision tree based state tying algorithm, which is based on a different estimate of multiple-mixture Gaussian distributions in node splitting for phonetic decision tree based acoustic modeling. The key idea in this approach is to use an m-level optimal subtree during the node split, which is an extension of the conventional one-step greedy CART algorithm [3]. In this new approach, the node split is not determined by the improvement of the impurity function evaluated by the one-step splitting of that node as typical in CART but by a multilevel optimal subtree derived from the candidate node. In this paradigm, CART algorithm has become a special case, where the level of algorithm optimality reduces to one. For every question to be evaluated a temporary subtree is constructed. Depending on the depth of the binary subtree, the state set of the evaluated node is splitted recursively in different partitions. These partitions are then used to calculate a new impurity function based on a single Gaussian density per partition [6]. For a two-level optimal splitting, two Gaussian densities are calculated for each state subset S_{yes} and S_{no} from the partitions of the lookahead tree. In a three-level optimal splitting scheme, four Gaussian densities are estimated for each binary outcome of all evaluated questions. The principle of the multilevel optimal subtree approach is illustrated in Fig. 2. It shows the subtree created for the three-level evaluation of a question in the solid node S. The subtree consists of a "yes" and a "no" branch for the primary question with four terminal nodes (shaded) each. These terminals subdivide the state sets S_{yes} and S_{no} into four partitions each, for which a Gaussian density is calculated. A separate subtree is calculated for each of the N_q questions.

Let $\tilde{T}(S, m)$ denote an m-level subtree, with root node S and a maximum level of m. The log-likelihood of the m-level tree $\tilde{T}(S, m)$ is defined to be

$$L(\tilde{T}(S, m)) = \sum_{s' \in \tilde{S}} L(s') \tag{12}$$

which is obtained by summing the log-likelihood over all its leaves.[1]

The proposed m-level optimal subtree based decision tree growing algorithm consists of the following steps.

1) If S is the root node, grow an m-level optimal subtree $\tilde{T}(S, m)$, not necessarily balanced, using the phonetic questions. Split the node S into nodes S_{yes} and S_{no}.

2) Update the log-likelihood of S_{yes} and S_{no} to be

$$L(S_{yes}) = L(\tilde{T}_{yes}(S, m)) \quad \text{and } L(S_{no}) = L(\tilde{T}_{no}(S, m)) \tag{13}$$

where $\tilde{T}_{yes}$ and $\tilde{T}_{no}$ are branches of the m-level optimal subtree $\tilde{T}(S, m)$. The updated log-likelihood of $L(S_{yes})$ and $L(S_{no})$ is modeled by 2^{m-1}-mixture Gaussians because they are the sum of single-mixture Gaussians from the corresponding m-level optimal subtree leaves.

3) For each current terminal tree node S with cluster sample count greater than the minimum sample count threshold, grow an m-level optimal subtree $\tilde{T}(S, m)$ and split the node S into nodes S_{yes} and S_{no} provided that

$$L(\tilde{T}(S, m)) - L(S) > \Delta_m. \tag{14}$$

Update the log-likelihood $L(S_{yes})$, $L(S_{no})$ by performing step 2).

4) The algorithm stops if there is no terminal node that satisfies step 3) and the minimum sample count constraint.

It should be noted that both $L(\tilde{T}(S, m))$ and $L(S)$ are based on multiple-mixture Gaussian distributions. This is because the likelihood of the nodes is updated by its likelihood from the m-level optimal subtree, which is a combination of Gaussians from the corresponding tree leaves. The proposed approach utilizes an m-level optimal subtree to obtain an estimate of the multiple Gaussian distribution for node splitting. Although the m-level optimal subtree $\tilde{T}(S, m)$ is derived from the phonetic questions and using single-mixture Gaussians of the untied states, the leaves of the m-level subtree $\tilde{T}(S, m)$ introduce a multiple-mixture Gaussian parameterization of the log-likelihood of the tree node S. In addition, the multiple-mixture Gaussian parameterization of $L(S)$ obtained from the proposed approach is honest in the sense that all its mixtures are supported on the data partition by the phonetic questions of the decision tree and it will not give an over estimate of $L(S)$. This is different from the previous approach, where the multiple-mixture distributions for the potential splits are derived from a k-means algorithm without any constraints regarding phonetic properties. The conventional one-level greedy tree growing algorithm is again a special case in the proposed approach when optimal subtree level $m = 1$.

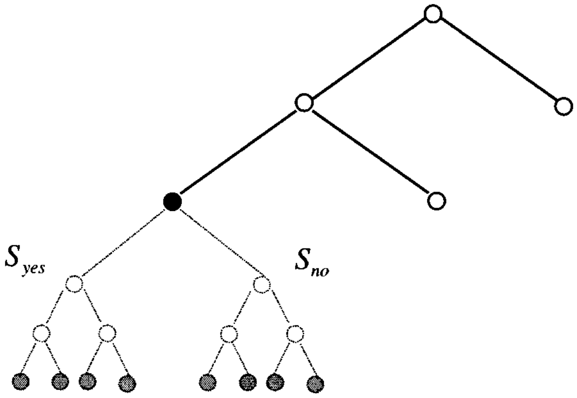


Fig. 2. Subtree construction for the evaluation of a question splitting the solid node.

However, there is a fundamental difference between the proposed approach and the look-ahead search technique used in decision tree based state tying [4], [11], [12]. The look-ahead search is to find a more accurate estimate of the log-likelihood increase when split node S. In other words, it uses a refined estimate of $L(S_{yes})$ and $L(S_{no})$ but does not change the parametric distribution of S nor the value of $L(S)$. In the proposed approach, a m-level subtree is used as a mean to introduce honest multiple-mixture Gaussian parametric distribution for node S, which is used consistently for all related log-likelihood estimates in node splitting, $L(S)$, $L(S_{yes})$ and $L(S_{no})$.

Although the greedy tree splitting algorithm based on single-mixture Gaussian distribution may not be accurate enough, it is nevertheless computationally quite efficient. For single-mixture Gaussian, the log likelihood $L(S)$ of a cluster at tree node S can be calculated by using the already available sufficient statistics from the untied state clusters without additional access of the data. As a consequence, the phonetic decision tree based state tying only constitutes a small portion of the computation in acoustic model building [19]. This may not be the case when multiple-mixture Gaussian distributions are used in node splitting. Although the proposed approach does not make a direct estimation of the multiple-mixture Gaussian distribution in decision tree state tying, growing an m-level optimal subtree can become expensive. Given a set of N_q phonetic questions, finding a two-level optimal subtree $\tilde{T}(S, m)$ involves in an order of $N_q \times N_q$ operations of node splitting. The algorithmic complexity grows exponentially with the subtree level m, making it infeasible for application in large vocabulary speech recognition.

In order to reduce the algorithmic complexity, we propose a scheme that is based on caching the top K best second-level questions of the previous search in a short-list table [6]. The short-list of the best K second-level phonetic questions associated with left and right branches used to construct the m-level optimal subtree $\tilde{T}(S, m)$ is attached to the new children nodes S_{yes} and S_{no}. In the future split, the m-level subtree constructed for S_{yes} and S_{no} will be restricted to questions in the short-list. For two-level optimal subtree, this reduces the algorithmic complexity of doing node splitting from $N_q \times N_q$ to $K \times N_q$, where

[1] We use the log-likelihood instead of the auxiliary function in the following discussion.

K is the depth of the short-list. This approximation is reasonable in two senses. First, the m-level subtree constructed with this caching scheme is always superior than the subtree constructed from one-step greedy algorithm. Second, the top K questions for S_{yes} and S_{no} derived from the m-level optimal subtree construction of their parent node S contain at least the K best $m-1$ level questions. This provides a good coverage of the questions used for the m-level optimal subtrees of S_{yes} and S_{no}. The use of the caching scheme makes it practical to apply the proposed m-level optimal subtree approach for phonetic decision tree based state tying in large vocabulary speech recognition tasks. In addition, other more aggressive caching schemes can also be used which will lead to further complexity reduction. In our speech recognition experiments, we observe a significant speed-up without recognition performance degradation.

VI. Experimental Results

The performance of the proposed decision tree clustering algorithms was evaluated on different experiments for the *Wall Street Journal* (WSJ) task. Twelve mel-cepstral coefficients and the normalized energy plus their first and second order time derivatives were used as acoustic features. The cepstral mean for each sentence was calculated and removed. All HMMs have three emitting states and a left-to-right topology. Training of the acoustic parameters was based on the proposed two-level segmental clustering algorithm for decision tree state tying. Single-mixture models were estimated for all triphones exceeding a sample count threshold in the training data. A minimum count threshold of five and ten examples was used in our experiments, but no significant performance difference was noted for these thresholds. Rare triphones with occurrences below the minimum count threshold were grouped by relaxing first the left context and then the right context. A phonetic decision tree tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones. The final triphone HMMs were built based on the tied states from the clustering. The number of mixtures for each tied state depends on the amount of training data assigned and varies from four to 12. Typically, only two to three iterations of the two-level segmental clustering algorithm were performed to obtain high quality acoustic models. Decoding was done using a one-pass N-gram decoder [26], in which the search was conducted on a layered self-adjusting decoding graph using the cross-word triphone models. The standard SI-84 and SI-284 training data sets were used to train the WSJ models. The pronunciation lexicon was generated automatically using a general English text-to-speech system (41 phones) [22]. The language models used in the experiments are the standard bigram and trigram language models provided by NIST for the WSJ corpus.

A. Two-Level Segmental Clustering

Even for large training data sets the number of actually observed triphones is only a small fraction of the total number of possible triphones. For a phoneme inventory of 41 phones plus two additional silence models, almost 80 000 possible context-dependent phones exist. In Table I, the number of triphones exceeding different minimum frequency thresholds of one, five, ten, and 30 are listed for WSJ SI-84 and SI-284 training data sets.

TABLE I
Number of Triphones Exceeding Different Frequency Thresholds in WSJ Database

Training data	#utterances	min. Frequency			
		1	5	10	30
SI-84	7,200 (15h)	16,500	11,500	8,600	4,600
SI-284	38,700 (60h)	22,500	18,000	14,400	9,600

TABLE II
Word Error Rates for Nov92 WSJ Evaluation (SI-84, Gender-Independent Models)

Language	NOV92	
model	5k-closed	20k-open
bigram	6.7%	14.4%
trigram	5.0%	12.8%

In the SI-84 training data set (7200 sentences, 15 h of speech), only 16 500 triphones occur in the training data. Among them, 8600 triphones have more than ten occurrences, whereas about 7900 triphones have less than ten examples. About 22 500 different triphones are observed in 60 h of speech for the SI-284 data, but only 14 400 have more than ten examples. In the first level clustering of our approach, rare triphones with less than five or ten examples are clustered into groups of generalized triphones to increase the robustness of the estimates for the untied states. The second level decision tree clustering is performed on the first level clusters. After the decision tree based state tying, the triphone model based on SI-84 training data consist of 3447 individual states tied through various contexts. It has a total of 37 000 Gaussian distributions. The average number of mixtures per state is 10.9. Evaluations on the WSJ tasks were performed on the official NOV92 (si_et_05, si_et_20) and NOV93 (si_et_h1) test sets for the closed 5K and open 20K vocabulary. The results are obtained based on a one-pass frame synchronous decoding without adaptation. The word error rates for the NOV92 evaluation of the gender-independent (GI) WSJ system trained on the SI-84 training data are tabulated in Table II.

In the second experiment, the WSJ SI-284 data was used in the training of the acoustic models. About 8100 of the 22 500 observed triphones occur less than ten times in the training data and are grouped into 1029 generalized triphone clusters to ensure robust estimates for the state clustering. After the phonetic decision tree clustering, 8006 individual states with about 99 000 Gaussian distributions (an average of 12.4 mixtures per state) were estimated. The results for the NOV92 and NOV93 tests using this model are listed in Table III. It is shown, that a low 3% word error rate in the 5k vocabulary NOV92 evaluation for gender-independent models is achieved. The error rates for the 20k vocabulary evaluations are between 9.8% for NOV92 and 13.4% for the NOV93 test, based on a trigram language model. The 1.8% out-of-vocabulary words have a significant contribution to the word errors in this open vocabulary test. These low error rates for both the 5k and 20k evaluations indicate the high performance of the proposed two-level segmental clustering approach, and from now on, the model built by two-level segmental clustering algorithm was used as a baseline system for the remaining experiments in this section.

B. Gender-Dependent Models

The experiments in this subsection are concentrated on the use of generalized features in phonetic decision tree state tying. We first present results of the proposed unified maximum likelihood approach to generate gender-dependent acoustic models. We compare the proposed approach with approaches based on MAP adaptation and training of separate gender specific models. Since gender identification is not the issue of this paper, we assume the genders of the test speakers to be known. Table IV tabulates the word error rates of different gender-dependent acoustic models trained on the WSJ-84 dataset using a trigram language model.

The first set of models (sGD) was trained following the conventional practice of splitting the training data into male and female subsets upon which two completely independent HMMs for both genders were built. State tying for male and female models were derived from separate decision trees constructed from gender specific subsets of data. It resulted in two models containing 2633 individual states for male and 2622 individual states for female. Comparing to Table II, a modest word error reduction of about 5% was observed. Adaptation of gender-independent acoustic models to gender-dependent male and female models using MAP adaptation techniques resulted in two model sets (mGD) with 3447 states each. It should be noted that MAP adaptation does not affect the state tying relationship in the generic, gender-independent seed model. MAP adaptation provides a more robust estimates for triphones and results in an error rate reduction comparing to the baseline sGD models. The last row in Table IV is the result of the models obtained from the proposed tagged decision tree clustering (cGD) approach using generalized features. The algorithm decides, based on the data, for every state of all triphones whether the state should be modeled separately for male and female or a joint state for both genders should be used. The total number of states in cGD–HMMs is 5488. About 420 of these states are shared between male and female models, which reduces the total number of individual states slightly below the total of 5255 for the sGD–HMMs. The performance improvement over gender-dependent sGD-models is between 4% and 8%, and the relative error rate reduction over gender-independent models is between 9% and 12% based on the two evaluation test sets. In Table V, results based on HMMs trained from SI-284 data set are tabulated. The automatically clustered gender-dependent cGD-HMMs again perform best among other approaches. The effect of training data fragmentation may become more important for smaller databases, where any data splitting usually results in reduced performance. Data-driven splitting and data sharing between conditions can help to improve the robustness of condition-dependent models in these cases.

An analysis of the decision tree used in constructing cGD–HMMs shows a phonetically reasonable behavior. States for vowels and diphthongs (except for the schwa sound, /aa/) are mostly separated for males and females, while stops and fricatives share up to 34% of their states for both genders. This appears to be consistent with the dependency of phones on vocal tract characteristics. The gender questions are competing in the decision tree with other questions about the phonetic contexts. They are used only if it leads to a maximum increase in the likelihood among all other questions and the minimum sample count constraint is satisfied. A leaf state is shared between genders if no gender specific question separating the male and female data was used in the path from the root tree node to that leaf tree node. Phonemes with more than 10% state sharing between genders in various contexts are listed in Table VI.

Fig. 3 illustrates the frequency of usage of gender questions for different depths of the decision tree. The average depth of the tree in this experiment was 11.4. Most of the gender questions are used in the upper part of the trees. This is an evidence that for many phones strong gender-dependent variations exist and gender-dependent acoustic modeling is useful. The proposed tagging approach to incorporate general features into phonetic decision tree tying detects these dependencies automatically and generates individual states if useful.

TABLE III
WORD ERROR RATES FOR NOV92 AND NOV93 EVALUATION OF THE WSJ TASK (SI-284, GENDER-INDEPENDENT MODELS)

Language	NOV92		NOV93
model	5k-closed	20k-open	20k-open
bigram	5.0%	11.9%	15.4%
trigram	3.0%	9.8%	13.4%

TABLE IV
SI-84 WORD ERROR RATES FOR DIFFERENT GENDER-DEPENDENT ACOUSTIC MODELS WITH TRIGRAM LM (sGD: SPLITTED TRAINING DATA, mGD: MAP ADAPTED MODELS, cGD: TAGGED DECISION TREE CLUSTERING)

Model		NOV92	
SI-84	#states male + female	5k-closed	20k-open
sGD	2,633 + 2,622	4.8%	12.2%
mGD	2 x 3,447	4.5%	12.1%
cGD	2,753 + 2,735	4.4%	11.7%

TABLE V
SI-284 WORD ERROR RATES FOR DIFFERENT GENDER-DEPENDENT ACOUSTIC MODELS WITH TRIGRAM LM (sGD: SPLITTED TRAINING DATA, mGD: MAP ADAPTED MODELS, cGD: TAGGED DECISION TREE CLUSTERING)

Model		NOV92		NOV93
SI-284	#states male + female	5k-closed	20k-open	20k-open
sGD	5,835 + 6,043	3.2%	9.8%	13.3%
mGD	2 x 8,006	3.0%	9.8%	13.2%
cGD	5,984 + 6,181	2.9%	9.5%	13.2%

TABLE VI
RATE OF STATE SHARING BETWEEN GENDERS FOR SOME PHONEMES

/t/	/aa/	/d/	/oy/	/ch/	/zh/	/jh/	/b/	/th/	/uh/	/hh/	/p/
34%	32%	30%	29%	25%	22%	22%	22%	22%	21%	17%	15%

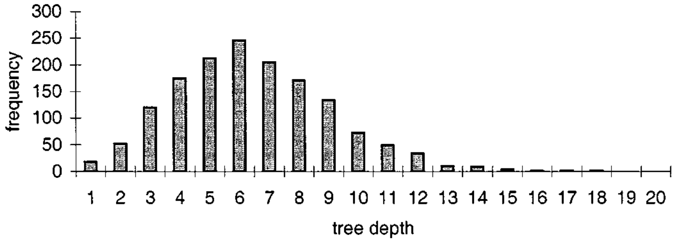


Fig. 3. Frequency of usage for gender questions over the depth of the decision tree.

C. Word-Boundary Dependent Models

The proposed tagging scheme for decision tree clustering was also applied to modeling word-boundary dependent HMMs. While some of the context dependent models (silence and noise models) occur only at word boundaries, most of the triphones appear in both inter- and intra-word positions and exhibit various degrees of dependencies on their positions. Moreover, the number of occurrences of these word-boundary dependent triphones in the training data also varies drastically, and some units may not have enough samples to be modeled separately. Table VII depicts the word error rates for position-dependent (POS–DEP) HMMs trained on WSJ SI-84 with the proposed generalized clustering. Results for position-independent models (POS–IND) are also included for comparison.

The use of position-dependent SI-84 models leads to a 10% word error rate reduction for gender-independent (GI) HMMs and a 5% word error rate reduction for gender-dependent (mGD) models. The SI-84 systems achieve 95.8% word accuracy for WSJ-5k and 88.8% word accuracy for the WSJ-20k task. The total number of individual states for the position-dependent HMMs increased about 30% from approximately 3400 to 4400. For the SI-284 system, a slight error rate reduction was obtained for the gender-independent 20 k task and a more significant 10% word error rate reduction was observed for the gender-dependent models over the baseline results. Table VIII illustrates the average percentage of state sharing between inter- and intra-word models for different phonetic classes.

The number of states shared between inter- and intra-word models varies from 30% for the /dh/ sound and 100% for rare phones like /zh/. These rare phonemes do not have sufficient examples in the training data to allow a state split into position-dependent variants of the same phoneme. Some vowels like /eh/ seems not much affected by word boundaries and share up to 88% of the states. The proposed tagged decision tree clustering approach automatically balances the need to generate separate position-dependent states for improved acoustic resolution and the availability of training data for robust model parameter estimation.

TABLE VII
WORD ERROR RATES POSITION-DEPENDENT HMMS (TRIGRAM LM)

Model		NOV92			
		5k-closed		20k-open	
		POS-IND	POS-DEP	POS-IND	POS-DEP
SI-84	GI	5.0%	4.4%	12.8%	11.6%
SI-84	mGD	4.5%	4.2%	12.1%	11.2%
SI-284	GI	3.0%	3.0%	9.8%	9.5%
SI-284	mGD	3.0%	2.9%	9.8%	8.8%

TABLE VIII
SHARING OF STATES BETWEEN POSITION-DEPENDENT MODELS FOR DIFFERENT PHONEME CLASSES

Fricatives	Stops	Nasals	Vowels
65%	71%	76%	80%

D. Multiple-Mixture Gaussian Based Tree Node Clustering

In this subsection, we present some experimental results of using multiple-mixture Gaussian distributions in decision tree node clustering. Between one and four mixture distributions were estimated for each untied state depending on the amount of available data in the WSJ SI-84 training set. For every examined node in the decision tree, the k-means algorithm was applied to calculate a four mixture distribution. The auxiliary function according to (11) was used as objective in node splitting. A decision tree with 3719 leaves was grown based on this objective function and unseen triphones were constructed in a standard way. The average log-likelihood for the training data increased from -99.69 to -98.01 compared to the standard single-mixture likelihood calculation. Fig. 4 illustrates the relative likelihood gain for some phonemes when multiple-mixture probability density functions were used in node splitting.

The biggest gain by the improved acoustic modeling is noted for vowels (2.3%), while the likelihood for fricatives increases only about 1.0%. The average improvement is 1.7% over all phonemes. This shows that the four component mixture densities fit the data better than the single Gaussian used in the standard decision tree. We expect this improved acoustic modeling in the decision tree clustering to increase the quality of the state tying. HMMs based on the proposed multiple-mixture clustering were constructed using WSJ-SI84 training data, and the model was evaluated on WSJ-92 evaluation test sets. The word error rate for the 5k vocabulary task was slightly reduced from 5.0% to 4.8% and for the 20k evaluation the error dropped from 12.8% to 12.5%. This shows how the improved acoustic modeling during the decision tree clustering leads to better state tying and increases the model accuracy.

E. Multilevel Optimal Subtree Approach

Experiments of using the proposed mutli-level optimal subtree algorithm for state clustering were also performed. A two-level optimal subtree, based on a short-list of top questions, was constructed, and multiple-mixture Gaussian distributions were estimated from the optimal subtree. The average log-likelihood

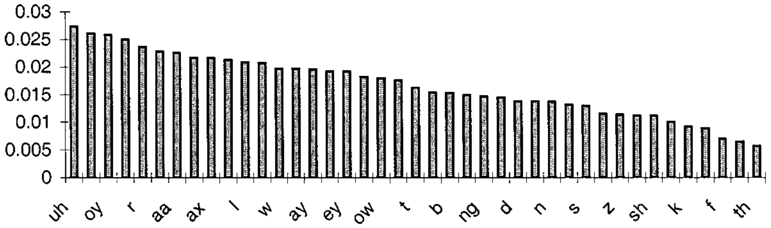


Fig. 4. Likelihood increase in percent for some phonemes.

for the training data increased from -99.69 to -98.65 compared to the standard likelihood calculation. In order to verify the short-list based caching scheme, an experiment was conducted using a short-list based on the top-30 questions and the results were compared with a short-list of size $K = 208$, which is a complete list containing all possible questions. The rank of the best two-level questions during the subtree construction was recorded. First, it was observed that questions used in tree construction are quite different between the proposed multilevel optimal subtree approach and the conventional one-level tree node splitting scheme. In 40% of the cases, the two-level optimal subtree algorithm and the conventional one-level tree splitting algorithm selected identical questions. But for the remaining 60% of cases, the best questions selected by the two approaches differ. Secondly, the top-30 short-list provided a 96% coverage of the best questions used in the subtree algorithm based on the complete list, and only in 4% of the nodes a suboptimal question was chosen because the best two-level question was not in the top-30 short-list. This justifies the usage of the short-list scheme for the two-level optimal subtree construction to reduce the amount of required computation for the likelihood evaluation of the tree nodes. The word error rates for gender-independent and MAP adapted gender-dependent HMMs using optimal subtree based state clustering are given in Table IX.

TABLE IX
WORD ERROR RATES FOR TWO-LEVEL OPTIMAL SUBTREE BASED CLUSTERING (TRIGRAM LM)

Model		NOV92			
		5k-closed		20k-open	
		std.	two-level	std.	two-level
SI-84	GI	5.0%	4.5%	12.8%	12.2%
SI-84	mGD	4.5%	4.4%	12.1%	11.8%
SI-284	GI	3.0%	2.9%	9.8%	9.5%
SI-284	mGD	3.0%	2.9%	9.8%	9.4%

The highest error rate reduction of about 10% based on the two-level optimal subtree based clustering was observed for the SI-84 trained gender-independent HMMs. For gender-dependent models and SI-284 training data the word error reduces about 3%–4%. It is interesting to note that the performance differences between gender-dependent and gender-independent models for the two-level optimal subtree based state clustering is much smaller than those using the standard decision tree clustering method. The two-level optimal subtree splitting algorithm clearly helps to build an improved decision tree and to enhance the quality of the acoustic models. Increasing the levels of the optimal subtree in node splitting beyond two levels did not provide additional performance improvements in our experiments. Our study indicates that multilevel optimal subtree building procedure can be made computationally feasible and it can significantly improve the robustness of the model. The experimental results provide the first experimental evidence that multilevel optimal tree building procedure beyond CART is advantageous in large vocabulary continuous speech recognition.

VII. SUMMARY

In this paper, methods of improving the robustness and quality of acoustic modeling using decision tree based state tying were described. A two-level segmental clustering approach was devised which combines the decision tree based state tying with agglomerative clustering. Under this approach, rarely seen triphones in the training data are first clustered into generalized triphones. These generalized triphone clusters are then used in the second level decision tree based state tying to improve the robustness and coverage of the decision tree based acoustic modeling. In order to incorporate various features in the decision tree clustering, a unified maximum likelihood framework for generalized phonetic and nonphonetic features was proposed. A tagging scheme was used to tag various features of interest and the selection of these features in the state clustering was determined by the log-likelihood increase instead of heuristically separating training data into various conditions. In contrast to the conventional methods which often lead to training data depletion, the proposed approach makes more efficient use of the entire training data and allows training data sharing across various conditions. As a consequence, the decision tree state tying is extended from tying states with various phonetic contexts to tying states with generalized phonetic and nonphonetic (e.g. gender, position, etc.) features. This leads to a significant increase in the amount of training samples for condition dependent acoustic modeling and the robustness of the condition dependent model is also enhanced. Moreover, there is no hard limit on the number of conditions to which the model can incorporate and the whole process is data driven. Finally, two methods based on multiple-mixture Gaussian parameterization were described and applied in large vocabulary speech recognition to improve the evaluation function in decision tree state tying. One method is based on a k-means fitting approach and the other one is based on an application of optimal multilevel subtree. Both methods are consistent with the structure of the decision tree, and therefore, the prediction power of the decision tree is well maintained without the need of a separate tree for unseen triphone generation. The proposed approaches were tested on the Wall Street Journal corporation and compared with other known approaches. The efficacy of the proposed approaches were verified and a significant improvement in model quality and recognition performance was obtained. The application of the generalized decision tree to word-boundary dependent acoustic models for example reduced the word error rate for the 20k-WSJ test data up to 10% and two-level optimal subtree based clustering resulted in about 5% error reduction for the same test data.

References

[1] L. R. Bahl, P. V. de Souza, P. S. Gopalakrishnan, D. Nahamoo, and M. A. Picheny, "Decision trees for phonological rules in continuous speech," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '91*, Toronto, ON, Canada, May 1991, pp. 185–188.

[2] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '98*, Seattle, WA, May 1998, pp. 805–808.

[3] L. Breiman, J. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Belmont, CA: Wadsworth, 1984.

[4] C. Chesta, P. Laface, and F. Ravera, "Bottom-up and top-down state clustering for robust acoustic modeling," in *Proc. Eurospeech '97*, 1997, pp. 11–14.

[5] P. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 340–354, Apr. 1991.

[6] W. Chou and W. Reichl, "High resolution decision tree based acoustic modeling beyond CART," in *Int. Conf. Speech Language Processing '98*, Sydney, Australia, Nov. 1998, pp. 2203–2206.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. R. Statist. Soc.*, vol. 39, pp. 1–83, 1977.

[8] J. Duchateau, K. Demuynck, and D. Van Compernolle, "A novel node splitting criterion in decision tree construction for semi-continuous HMMs," in *Proc. Eurospeech 97*, Rhodes, Greece, 1997, pp. 1183–1186.

[9] 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.

[10] M.-Y. Hwang, X. Huang, and F. Alleva, "Predicting unseen triphones with senones," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '93*, Minneapolis, MN, 1993, pp. 311–314.

[11] R. Kuhn, A. Lazarides, Y. Normandin, and J. Brousseau, "Improved decision trees for phonetic modeling," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '95*, Detroit, MN, 1995, pp. 552–555.

[12] A. Lazarides, Y. Normandin, and R. Kuhn, "Improving decision trees for acoustic modeling," in *Int. Conf. Speech Language Processing '96*, Philadelphia, 1996, pp. 1053–1056.

[13] K.-F. Lee, *Automatic Speech Recognition—The Development of the SPHINX System*. Norwell, MA: Kluwer, 1989.

[14] C.-H. Lee, E. Giachin, L. R. Rabiner, R. Pieraccini, and A. E. Rosenberg, "Improved acoustic modeling for large vocabulary speech recognition," *Comput. Speech Lang.*, vol. 4, no. 2, pp. 103–127, 1992.

[15] C. Neti and S. Roukos, "Phone-context specific gender-dependent acoustic-models for continuous speech recognition," in *IEEE Automatic Speech Recognition Understanding Workshop*, Santa Barbara, CA, Dec. 1997, pp. 192–198.

[16] H. J. Nock, M. J. F. Gales, and S. J. Young, "A comparative study of methods for phonetic decision-tree state clustering," in *Proc. Eurospeech '97*, Rhodes, Greece, 1997, pp. 111–114.

[17] D. B. Paul, "Extensions to phone-state decision-tree clustering: Single tree and tagged clustering," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '97*, Munich, Germany, Apr. 1997, pp. 1487–1490.

[18] L. R. Rabiner, J. G. Wilpon, and B.-H. Juang, "A segmental k-means training procedure for connected word recognition," *AT&T Tech. J.*, vol. 65, pp. 21–31, 1986.

[19] W. Reichl and W. Chou, "Decision tree state tying based on segmental clustering for acoustic modeling," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '98*, Seattle, WA, May 1998, pp. 801–804.

[20] ——, "An approach of decision tree state tying based on multi-mixture Gaussian distributions," Bell Labs Tech. Memo. BL011 333-980 920-6TM, 1998.

[21] ——, "A unified approach of incorporating general features in decision tree based acoustic modeling," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '99*, Phoenix, AZ, Mar. 1999, pp. 573–576.

[22] R. W. Sproat and J. P. Olive, "Text-to-speech synthesis," *AT&T Tech. J.*, vol. 74, pp. 35–44, 1995.

[23] D. Willet, C. Neukirchen, J. Rottland, and G. Rigoll, "Refining tree-based state clustering by means of formal concept analysis, balanced decision treesand automatically generated model-sets," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '99*, Phoenix, AZ, Mar. 1999, pp. 565–568.

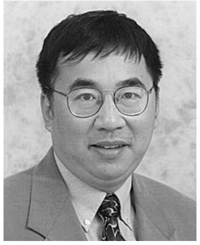
[24] S. J. Young, "The general use of tying in phoneme-based HMM speech recognizers," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '92*, San Francisco, CA, 1992, pp. 569–572.

[25] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree based state tying for high accuracy modeling," in *ARPA Workshop Human Language Technology*, Princeton, NJ, Mar. 1994, pp. 286–291.

[26] Q. Zhou and W. Chou, "An approach to continuous speech recognition based on layered self-adjusting decoding graph," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing '97*, Munich, Germany, Apr. 1997, pp. 1779–1782.

Wolfgang Reichl was born in Germany in 1965. He received the Dipl.-Ing. degree in electrical engineering in 1991, and the Dr.-Ing. degree in electrical engineering in 1996, both from the Technical University of Munich, Germany.

From 1991 to 1996, he was a Research Assistant at the Institute for Human-Machine Communications, Technical University of Munich. He was working in the German Verbmobil project in the area of discriminative training algorithms and neural networks for speech recognition. Since 1996, he is a Member of Technical Staff with the Multimedia Communications Laboratory, Bell Laboratories, Lucent Technologies, Murray Hill, NJ. His research interests include acoustic modeling and language modeling for large vocabulary speech recognition.



Wu Chou (SM'87–M'91) received the M.S. degree in electrical engineering in 1987, the M.S. degree in statistics in 1988, and the Ph.D. degree in electrical engineering in 1990 from Stanford University, Stanford, CA.

He joined the Speech Research Department, AT&T Bell Laboratories, Murray Hill, NJ, in July 1990. Currently, he is with the Multimedia Communication Research Laboratory, Bell Laboratories, Lucent Technologies, Murray Hill. Since joining Bell Labs, he has been involved in various activities of automatic speech recognition and dialog systems research projects. His research interest lies in the field of speech recognition, acoustic modeling, large vocabulary speech recognition, adaptation, vector quantization, acoustic assisted image coding and animation, signal processing, and oversampled sigma-delta modulation. He has authored or coauthored 50 technical papers, two book chapters, and holds eight patents.

Dr. Chou is an active member of IEEE Signal Processing Society. He served as an associate editor of IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING and the Publication Chair of the 1997 IEEE Workshop on Automatic Speech Recognition and Understanding. He is a member of Sigma Xi.