A Comparative Study of Methods for Phonetic Decision-Tree State Clustering

H.J. Nock *M.J.F. Gales* *S.J. Young*

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, UK.
Tel: [+44] 1223 332800 Fax: [+44] 1223 332662
email : hjn11,mjfg,sjy@eng.cam.ac.uk

ABSTRACT

Phonetic decision trees have been widely used for obtaining robust context-dependent models in HMM-based systems. There are five key issues to consider when constructing phonetic decision trees: the alignment of data with the chosen phone classes; the quality of the modeling of the underlying data; the choice of partitioning method at each node; the goodness-of-split criterion and the method for determining appropriate tree sizes. A popular existing method uses efficient but crude approximate methods for each of these. This paper introduces and evaluates more detailed alternatives to the standard approximations.

1. Introduction

A key problem in building continuous-density Hidden Markov Model (HMM)-based context-dependent acoustic models is maintaining a balance between the desired model complexity and the number of parameters which can be robustly estimated from the available training data. One solution which has proved successful (eg.[8], [1]) is based upon the use of phonetic decision trees.

A phonetic decision tree is a binary tree in which a yes/no question about phonetic context is attached to each node (Figure 1). The tree can be used to recursively partition a set of states into subsets by answering the questions as appropriate for the triphone context in which each state occurs. States reaching the same leaf node are judged to be similar and are then tied. Phonetic decision-trees lead to compact, good quality state clusters with sufficient associated data to allow robust estimation of mixture Gaussian output probability distributions. Decision-tree-based techniques are also attractive because they allow the synthesis of models for contexts which do not occur in the training data, and because implementations such as [8] are considerably more efficient than alternative bottom-up clustering techniques. This paper focusses on the methods used for phonetic decision-tree construction. Although it will concentrate on triphone-based systems, the methods described extend straightforwardly to systems using greater degrees of context-dependency.

Construction of globally optimal decision trees is a computationally intractable problem. In general, trees are constructed using a variant of the following sequential optimization process. Before tree building begins, training observations are associated or *aligned* with states in the untied system. Trees are then grown through recursive partitioning of the set of states at the root of the tree. At each stage, the best of some set of *partitions* of the states at a node is chosen according to a *goodness-of-split criterion*, which will be specified in terms of the data aligned to those states. The process of node splitting continues until the tree reaches some appropriate size.

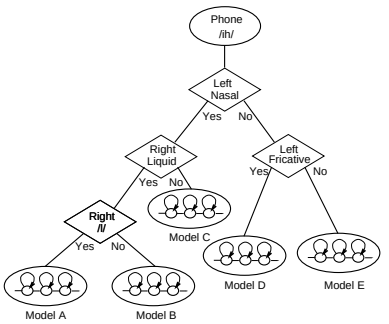

Figure 1: A Phonetic Decision Tree

There are therefore five key issues to consider when constructing phonetic decision trees using this basic method:

- the method for obtaining the alignment of training data with the chosen phone classes;
- the quality of the modelling of the data aligned to each state;
- the set of partitions considered at each node;
- the goodness-of-split criterion for comparing different partitions; and
- the stopping criteria used to determine appropriate tree sizes.

The standard method of [8] uses simple but efficient approximate solutions for each of these. The frame-state alignment is taken from a single-Gaussian, unclustered triphone system. A restricted set of partitions is considered at each node, specified using the phonetic questions. The splitting criterion used is directly related to that used in training: questions are chosen to maximize the likelihood of the data over the resulting partitions. The gain in likelihood due to a split can be calculated efficiently by representing each node using a single Gaussian distribution: the means, variances and state occupation counts (retained during Baum-Welch re-estimation) associated with the underlying states form *sufficient statistics*. Tree growing stops when the log-likelihood gain from a split falls below a threshold or a minimum frame-occupancy threshold is reached.

Although these approximations have been successful for clean, read speech, there is growing evidence that they will not be adequate for modelling more natural spontaneous and large scale speech tasks. This paper will therefore examine each of the five issues in turn and investigate more detailed alternatives. The structure of the paper is as follows. The next section discusses the standard approximations and introduces alternatives closer to the ideal solutions. Section 5. evaluates methods using the SQALE US-English test set and Section 6. presents brief conclusions.

2. Methods Studied

2.1. Alignment of Training Data

Tree construction methods must make the fundamental assumption that tying states will not alter the initial alignment of training observations to states: without this assumption tree-building would be intractable, requiring re-alignment of the training data after every hypothesized partition of a set of states. The standard method obtains the necessary set of frame-state alignment statistics from untied, single Gaussian triphone models, by retaining state occupation counts during Baum-Welch training. Although these statistics can be easily obtained, undertraining of rarely seen contexts means that the resulting state alignments may not be representative of the alignments in a state-tied, Gaussian component-based system. Thus the single Gaussian model set may provide poor sufficient statistics for clustering.

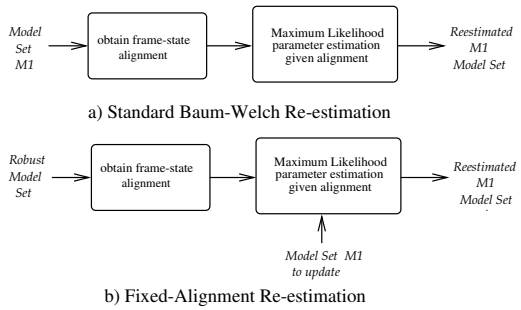

Figure 2: Standard and Fixed-Alignment Reestimation

An alternative method for obtaining more representative statistics for clustering uses *fixed alignment re-estimation*. This is a new variation on Baum-Welch re-estimation in which a previously trained and accurate model set (possibly tied and using Gaussian mixture observation densities) provides the frame-state alignment which is used to estimate the parameters of a second model set. Figure 2 illustrates both standard and fixed alignment reestimation. For this application, an untied single Gaussian system can be estimated based on an alignment taken from a previously tied Gaussian component-based system; the newly updated (untied) model set should then provide a more appropriate statistics for tree construction.

2.2. Base Unit Modelling

The standard method assumes that a single component Gaussian is sufficient to model the data aligned to each untied model state to be clustered (referred to here as a *base unit*). In reality, the distribution may be inadequate to represent the variability which is seen in the data: in general, there will be at least two modes in unnormalised speaker-independent data, corresponding to the data from male and female speakers. In principle, better data modelling can be achieved by using mixture-component Gaussians. However, training mixture component models in an untied system using standard Baum-Welch reestimation would lead to distorted frame-state alignments (Section 2.1.) where data is sparse. With either method, the sufficient statistics obtained from the model set are likely to be poor representatives of the underlying data.

More detailed Gaussian component models of base units can be obtained using the fixed alignment re-estimation method of Section 2.1. Here, the fixed state-alignment from an existing detailed model set is used in the iterative re-estimation of an untied system with mixture-component Gaussian output distributions for each base unit. This method ensures that better data modelling is obtained without introducing the problems of non-representative frame-state alignments which would occur with standard Baum-Welch.

2.3. Partitioning Methods

In principle, all possible partitions of the set of states at each node should be evaluated and the best retained during tree construction. In practice, this is too computationally expensive. Therefore, standard phonetic decision-tree construction methods consider only a restricted subset of (potentially sub-optimal) partitions specified using a small, linguistically-motivated set of questions about context.

Preliminary experiments investigated the use of an extended question set. Although [8] builds separate trees to tie states in corresponding positions within triphone models sharing the same base phone, [2] suggests better use is made of the training data in a tied mixture system when a single tree clusters states across both context and state position. Clustering then begins by pooling the states of all triphones with the same base phone, and the phonetic question set is extended to include questions about a state's position within the associated HMM. However, results in [6] on the limited Resource Management task show that use of this extended question set gave little change in the resulting tied system architecture. The first two questions in the trees constructed almost always partitioned based on state position within the associated HMM, a result consistent with [5] on the Wall Street Journal corpus. We therefore chose to investigate an alternative partitioning method.

A closer approximation to the desired evaluation of all possible partitions can be achieved with Chou's partitioning algorithm ("CPA", [4]), which efficiently finds a locally optimal partition of a set of data. The algorithm is analogous to the iterative k-means algorithm but is applied here with a log-likelihood objective function as in the ML-SSS algorithm of [7]. The CPA method as applied to the problem of partitioning the set of states at each node is illustrated in Figure 3.

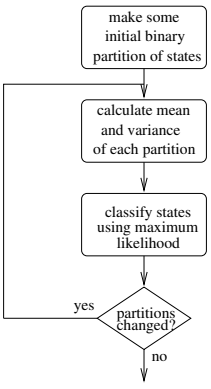

Figure 3: CPA-based partitioning

Preliminary experiments showed that question- and CPA-based clustering of word-internal triphone states lead to similar recognition results on the Resource Management task. For larger-scale recognition tasks, however, state-of-the-art performance requires use of more detailed cross-word triphone models for which it becomes necessary to consider the issue of data sparsity. Many states have only a few associated observations and a large number of triphone contexts will be unobserved in the training data.

Standard phonetic trees offer a straightforward method for syn-

thesising models of unobserved triphones. A state in a model of an unseen triphone can be associated with an output distribution by answering questions in the associated phonetic decision tree. The observation density at the leaf node reached is that used for the unseen state.

This solution is not directly applicable for CPA-derived trees because they do not include phonetic questions. Instead, the following variant of the standard method can be used. Before applying CPA to tree-building, a phonetic tree (*pre-tree*) is grown using standard techniques. By answering questions in this tree, each "unseen" state can be associated with a leaf node and tied to the contextually similar "seen" state or states at that leaf, thus giving it a location in acoustic space. The tied states at pre-tree leaves are then used as the base units for the main CPA-based tree construction phase.

In a preliminary experiment, the pre-trees used for forming CPA base units were grown until each leaf contained a single state. This gave poor results, possibly due to the effects of states with little associated data. Such states may be outliers and thus provide poor acoustic locations for any associated unseen states at the same pre-tree leaf. A further area of concern was that CPA may cluster rarely seen states inappropriately for two reasons. CPA may be grouping states which are contextually very different based on limited and potentially unrepresentative data; in addition, partitions may be chosen due to data sparsity effects unrelated to phonetic context, such as male/female splits of limited data. Although these are also issues for standard methods, they may be exacerbated by the lack of restrictions on groupings in CPA. These problems can all be alleviated by growing pre-trees using a minimum frame occupancy threshold at each leaf, which should give a more robust set of base units with greater associated data for use in the main CPA-based construction process.

3. Goodness-of-Split Criterion

The ideal splitting criterion for building a robust M-component Gaussian system would evaluate the increase in likelihood when the data aligned to each node is modelled by mixture-component Gaussians. However, this criterion would require several iterations of Baum-Welch reestimation for every hypothesized partition and is therefore computationally intractable. Standard methods use a crude but efficient single Gaussian approximation to evaluate partitions. By representing each node with a diagonal-covariance single Gaussian distribution, the log likelihood of the training data associated with any set of states can be calculated without reference to the training data: the state means, variances and occupation counts form sufficient statistics. However, this approximation may lead to the choice of inappropriate partitions: tree construction performs a constrained maximization of the likelihood of the training data over for a single component Gaussian-based system rather than over the target mixture-component Gaussian system.

The severity of this problem can be examined by using more detailed splitting criterion which approximates mixture-component Gaussians at each node. States at a node are clustered into M groups using CPA; each group is then modelled using a single Gaussian when calculating likelihoods. Preliminary experiments on RM [6] with two-component systems showed that performance degraded relative to the standard splitting criterion when the base units were modelled by a single Gaussian. However, improved results were obtained when untied states were modelled by mixture Gaussians as in Section 2.2., with components from the same state allowed to move independently during CPA clustering.

4. Stopping Criteria

One of the most criticial problems in tree construction is determining an appropriate size of tree. An overly large tree will be overspecialised to the training data and generalise poorly; a tree which is too small will model the data badly. Standard methods determine appropriate tree sizes (and therefore the number of states in the tied system) using stopping rules, typically a minimum frame occupancy at each leaf and a minimum gain in likelihood per split. Clustering quality is sensitive to the combination of thresholds used, and determining appropriate values requires time-consuming construction and comparison of several sets of trees. Further, [3] shows that stopping criteria may be a poor method for determining a robust tree size; better results are obtained by growing an overly large tree and pruning back until the tree is robust. Pruning-based methods for state-clustering are appealing because no arbitrary stopping thresholds need be specified: appropriate values are learned automatically during the pruning process.

A V-fold cross-validation method was investigated. A tree is grown to purity[1] using the full data set. A cross-validation estimate of the log-likelihood of held-out data is calculated for each node. Any split which decreases this estimate is potentially overspecialising the tree to the training data and is removed from the tree. Previous work on pruning [5] [6] uses the single diagonal Gaussian log-likelihood criterion of standard tree construction; as discussed in Section 3., this is a coarse approximation and will prune trees to give robust single Gaussian systems. In this work, cross-validation estimates used the approximate mixture-based criterion of the previous section.

5. Experimental Results

Experimental work applied the construction methods above to state clustering in cross-word triphone systems. The baseline (standard) systems used for comparison were all gender-independent, cross-word-triphone, mixture-Gaussian tied state systems; the number of states and mixture components used varied and is specified in the individual experiment descriptions below. One difficulty in evaluating the techniques is that the standard method is sensitive to the particular combination of stopping thresholds in use. Thus, although the baseline systems were optimised over a limited range of thresholds, the figures obtained are still susceptible to some degree of noise. The 39-dimensional acoustic feature vector contained 12 MFCC's and log energy, plus the first and second differentials; per-segment cepstral mean normalization was used. The acoustic training data consisted of 7185 sentences from the SI-284 WSJ-0 subset of the Wall Street Journal database. Recognition results are presented for 200 sentences taken from the US English SQALE evaluation data set. Results were obtained using precomputed lattice rescoring rather than full recognition: all experiments used the same lattice as input and only the acoustic models were changed between experiments. Lattices were produced in two stages. The first used cross-word context-dependent models generated using the clustering method in [8] and a bigram language model; lattices were then expanded using a trigram model.

Table 1 shows the average per-frame log probability after training and resulting absolute word error rate when using a more representative frame-state alignment. Results are for eight-mixture component cross-word triphone systems clustered to reduce the number of distinct states from around 32000 Gaussians to around 2300. The first line of the table gives results for the standard method for which the alignment is taken from an untied single

[1] A pure tree has a single state per leaf.

Gaussian system. The second is for a system trained and clustered using the alignment from a previously tied eight-mixture system. The per-frame likelihood of the training data improves slightly; the small decrease in performance is surprising and suggests that the alignment of data in the single mixture system is reasonably representative for clean Wall Street Journal data. Work in progress suggests this may not be the case for spontaneous speech.

	Av. Log Prob	% Word Error Rate
Single-mix alignment	-65.09	13.79
Eight-mix alignment	-65.04	14.08

Table 1: Results using more representative alignment

A set of experiments compared recognition performance for eight-mixture systems with comparable numbers of parameters built using CPA-based and standard phonetic question-based partitioning. CPA is only locally optimal so two initialisations were investigated:

- calculating the global mean of the data associated with the pool of states, perturbing slightly and using maximum likelihood classification on the states;
- partitioning using the best phonetic question, as in the standard method.

Although the latter gave better results in terms of log-likelihoods, both led to systems with similar recognition performance.

Table 2 shows selected results for CPA-based clustering as the pre-tree minimum frame threshold is varied; as expected, the threshold does affect the quality of clustering with an optimum reached at value 100. A baseline system constructed using the standard method had a word error rate of 13.79%, and it can be seen from the table that even at the best pre-tree thresholds, CPA results gave no gain over the use of phonetic questions.

Pre-Tree Minimimum Frame Threshold	Perturbed Global Mean CPA	Best Phonetic Question CPA
0	14.55	15.05
100	13.76	13.79
150	14.14	14.52

Table 2: CPA-based clustering results

Table 3 shows results for a robust two-component system constructed using the approximate two-component splitting criterion (Two-Mixture GOS). Sufficient statistics were taken from an untied, two-mixture system trained iteratively as in Section 2.2. using a fixed single Gaussian model set alignment. Results are compared with a system constructed using a standard single Gaussian splitting criterion (Single-Mixture GOS). The same minimum frame occupancy threshold was used in both clustering experiments and both systems used around 8k Gaussians. Results show the more detailed splitting criterion degrades performance which suggests, surprisingly, that the single Gaussian approximation is sufficient for modelling the variability in the WSJ-0 data.

Splitting Criterion	% Word Error Rate
Single-Mixture GOS	15.58
Two-Mixture GOS	16.25

Table 3: Two-mixture Splitting Criterion Results

Table 4 compares recognition results for systems constructed using stopping-rule and cross-validated trees. Cross-validated trees were grown and pruned using the approximate two-mixture log-likelihood criterion and pruning estimates used six-fold cross-validation. The baseline for comparison is a standard system with similar numbers of parameters. Such a system can be built using different combinations of tree stopping thresholds; the first line of Table 4 represents performance of the best of several systems investigated. Although performance is lower with the cross-validated system, the benefits of the method are hard to quantify. Whereas the baseline figure required considerable experimentation to find good stopping thresholds, the cross-validated system required just one iteration of the clustering and training process for only a minor performance loss.

	% Word Error Rate
Standard Stopping Criteria	15.84
Cross-Validated	16.25

Table 4: Two mixture component cross-validation-based results

6. Conclusions

This paper has described and evaluated more detailed approaches to five standard issues in phonetic decision-tree construction. Results are comparable with the standard method when CPA-based partitioning methods are used in tree construction and when clustering across both state position and context; performance degrades slightly when the approximate mixture-based criterion is used for splitting and when a more representative frame-state alignment is used. A system constructed using pruning gave performance slightly below the standard method, but required considerably reduced development time. The overall results suggest that when training and testing on read speech, the crude approximations made by the standard method do not seriously affect the quality of clustering. Further work will focus on spontaneous "found" speech where greater data variability may have a more significant effect on clustering performance.

7. Acknowledgements

H. J. Nock is funded by an EPSRC studentship.

8. REFERENCES

1. LR Bahl, PV de Souza, PS Gopalakrishnan, and MA Picheny. Robust methods for using context-dependent features and models in a continuous speech recognizer. In *Proc ICASSP*, pages 533–536, 1994.
2. G Boulianne and P Kenny. Optimal tying of hmm mixture densities using decision trees. In *Proc ICSLP*, pages 350–354, 1996.
3. L Breiman, J Friedman, RA Olshen, and CJ Stone. *Classification and Regression Trees*. Wadsworth Inc, 1994.
4. P Chou. Optimal partitioning for classification and regression trees. *IEEE Trans PAMI*, 13(4):340–354, 1991.
5. A Lazarides, Y Normandin, and R Kuhn. Improving decision trees for acoustic modelling. In *Proc ICSLP*, pages 1053–1057, 1996.
6. H Nock. Contextual clustering for triphone-based speech recognition. Master's thesis, CUED, 1996.
7. M Ostendorf and H Singer. Maximum-likelihood successive state splitting. *Computer Speech and Language*, 1997.
8. SJ Young, JJ Odell, and PC Woodland. Tree-based state tying for high accuracy acoustic modelling. In *ARPA Human Language Technology Workshop*, pages 286–291, 1994.