IMPROVED DECISION TREES FOR PHONETIC MODELING

Roland Kuhn, Ariane Lazarides, Yves Normandin, and Julie Brousseau

Centre de recherche informatique de Montréal
1801 McGill College Avenue, Suite 800
Montréal, Québec, Canada H3A 2N4
email: kuhn OR lazarids OR norman OR broussea @crim.ca

ABSTRACT

Bahl *et al* employed decision trees to specify the acoustic realization of a phone as a function of its context [Bah91]. By using a computationally cheap Poisson-based evaluation function, they were able to account for a much wider context than previous researchers (five preceding and five following phones). We extend this work in four ways:

1. We employ the Poisson criterion to find quickly the M best questions at a node during tree expansion, then use an HMM-based MLE criterion to make the final choice from these (and for pruning trees - see next point).
2. *Bahl et al* use stopping criteria to halt the growth of a tree, which is then used for recognition. It is preferable to grow an over-large tree and then prune it [Bre84]; we apply the efficient GRD expansion-pruning algorithm [Gel91] to the phonetic modeling problem.
3. Like Bahl *et al*, we allow questions about a large number of preceding and following phones. However, a given search algorithm may make some of these questions difficult to answer. In addition to the "YES" and "NO" children of each question, we grow a "DON'T KNOW" subtree to be used if a question is unanswerable at present.
4. We have experimented with questions based on phonetic features, as well as questions that ask about the presence of specific phones. Our approach permits an arbitrary feature schema to be read in and used in question generation.

1. INTRODUCTION

The phones preceding and following a given phone "ph" affect the acoustic realization of "ph", so that speech recognition performance can be improved by creating several different models for "ph". In the triphone approach, one initially creates a "ph" model for each combination "$< X, ph, Y >$", where X is the preceding and Y the following phone [Lee89]. Similar triphones are clustered until each of the clustered models has sufficient training data.

This approach ignores the strong influence that may be exerted by phones that are further away than the immediately preceding or following one [Sch93]. Bahl *et al* grew decision trees in which the candidate questions at a node concern the phonetic context of "ph" [Bah91]. The question chosen is the one in which the combination of the "YES" model for "ph" (trained on data items yielding "YES" to the question) and the "NO" model (trained on data items yielding "NO") yields the greatest improvement over the model at the node itself. Let the phone preceding "ph" be denoted "−1", the phone before that "−2", and so on; similarly, the phone after "ph" can be denoted "+1", the one after that "+2", and so on. A typical question in the resulting tree might be: "Is −2 an 's'?" Each leaf of the tree corresponds to a context-dependent phone model.

What criterion should be used to evaluate questions? An HMM-based MLE criterion (the obvious choice) is computationally too expensive. Bahl *et al* devised a Poisson model for the acoustic data for "ph" in a given context. This model allows extremely fast question evaluation - these researchers were able to consider questions involving all possible phones at the five preceding and five following positions. The models found at the leaves of their trees are conventional HMMs.

In more recent work, these researchers have modified their original approach in a number of ingenious ways to achieve greater robustness - only the type of questions asked remains the same [Bah94]. They now use a diagonal Gaussian, rather than a Poisson, model to evaluate questions. Furthermore, they merge similar nodes in the tree, turning it into a decision network. Finally, each network now models a given transition in the HMM for a particular phone, rather than the entire HMM for that phone. This last idea also appears in recent work by Young on triphones, in which tree-based clustering is carried out for each state of an HMM [You94].

Moore *et al* have also extended the original Bahl *et al* work, by devising context-adaptive phones (CAPs) [Moo94]. The idea here is to avoid corpus-dependent models by prohibiting two models from being trained on exactly the same data items - in cases of conflict, the model with the wider context wins. For instance, if **all** examples of "ih" with $-1 = s, +1 = k$ occur in the word "six", one does not build an "ih" model for this narrower context but only for

The authors are members of the Institute for Robotics and Intelligent Systems (IRIS) and wish to acknowledge the support of the Networks of Centres of Excellence Program of the Government of Canada, NSERC, and the participation of PRECARN Inc.

0-7803-2431-5/95 $4.00 © 1995 IEEE

$-1 = s, +1 = k, +2 = s$.

Our own extensions of [Bah91] are quite different from [Bah94], [You94], or [Moo94].

2. QUESTION SCORING: POISSON AND HMM MLE CRITERIA

The Poisson criterion is fast but inexact - for instance, it ignores the evolution of the signal for "ph" over time. Rather than abandon it (as did Bahl *et al*), we employ it to find quickly the M best questions at a node during tree expansion. For each of these questions, we then build "YES" and "NO" HMMs and evaluate the MLE performance of these child HMMs; the question whose children perform best is chosen. Tree pruning is carried out with the same HMM-based criterion.

3. PRUNING VS. STOPPING CRITERIA

Bahl *et al* use stopping criteria to halt the growth of a tree; instead of pruning an over-large tree, constraints are applied during tree growth. A node is turned into a leaf node if the score of the best question falls below a threshold, or if the number of data items is too small. According to the decision tree literature, this approach is suboptimal; it is better to grow an over-large tree, then prune it back by examining its performance on new data [Bre84], [Gel91].

Many researchers still use cross-validation [Bre84] to prune decision trees. The iterative GRD expansion-pruning algorithm [Gel91] is guaranteed to perform as well as or better than cross-validation pruning. GRD involves iterative cycles of expansion and pruning on two equal-sized disjoint sets of training data. We were able to extend it to our problem (in which probabilistic models rather than classes are stored in the leaf nodes), using as a pruning criterion either the Poisson or the MLE criterion. We found the GRD algorithm both computationally efficient and easy to implement, and highly recommend it to other decision tree aficionados!

4. TERNARY TREES AND SEARCH ALGORITHMS

Our decision trees are ternary rather than binary - each interior node has "YES", "NO", and "UNKNOWN" children, so that the tree contains models to be used in every conceivable combination of knowledge and ignorance about the phonetic context. This permits the trees to support an almost unlimited range of search algorithms. Like Bahl *et al*, we allow questions about five preceding and five following phones. However, a given search algorithm may make some of these questions difficult to answer.

Suppose that the identity of the +2 phone and following + phones is currently unknown. If the question at the root of the tree happens to concern the +3 phone, we proceed to the subtree at the UNKNOWN child of the root, which is guaranteed to contain no questions concerning the +3 phone or phones further out. If the first question we encounter concerns the +2 phone, we can again enter the UNKNOWN subtree and recurse until we encounter a suitable question (for instance, one concerning the −1 position, if we know about that position). This kind of tree is well-suited to multi-pass search, in which the first few passes score paths in the graph on the basis of incomplete information, while the last pass uses knowledge about the complete context of current phone "ph".

5. FEATURES

Most work on phone context modeling considers the context of a phone as also being defined in terms of phones. However, some articulatory traits in the phones near the phone "ph" being modeled may have a stronger impact on the acoustic realization of "ph" than others. It might be very important whether the −1 phone is a nasal or not, but irrelevant whether it is 'm' or 'n'. We experimented with two feature schemas for generating the set of candidate questions at a position. Schema 1 is based on the Chomsky-Halle feature definitions, schema 2 is based on articulatory feature definitions [Lad82]. Our programs make it possible to read in any phonetic schema based on binary features, and generate questions about the context from that schema.

6. RESULTS

Figure 1 shows a tree with phone-based questions. If the answers to questions about the +1 and −1 positions are both "UNKNOWN", we assume that information further out is unavailable and we arrive at a leaf node. If the current node is descended from a node in which a question about a position was answered "YES" or "NO", it is forbidden to answer "UNKNOWN" to a question about that position or a position further in. Nodes marked "X" are "UNKNOWN" children that should never be reached. Thus, the "UNKNOWN" child of the node whose question is "+1 d?" is "X". Data items in this node are known **not** to have a "k" in the +1 position, so it would be nonsensical to say we cannot answer another question about the same position.

Figure 2 shows a tree for the same phone "uh" grown using the Chomsky-Halle schema. The only question types that are considered for both the phone-based tree of Figure 1 and the Chomsky-Halle tree of Figure 2 are those concerning pauses and silences ("pau" and "sil"); interestingly, the same question "-2 pau?" was chosen twice in each tree.

Table 1 shows recognition results for trees trained on 15796 sentences from the ATIS task and tested on 334 ATIS sentences. For testing purposes, we used a simplified version of our SR system (no cross-word or trigram rescoring, tight beam pruning, discrete models, no $\delta\delta$ parameters). "TYPE" is the type of question in the decision tree. *Phones* questions ask about the identity of neighbouring phones (e.g. "is there an 's' at the −1 position?"); *CH* questions concern Chomsky-Halle features (e.g. "is there a coronal at the +2 position?"); *Artic.* questions concern articulatory features (e.g. "is there a dentialveolar at the +1 position?") Under "CRITERION", *Poisson* means that the tree is expanded and pruned using the iterative GRD algorithm and the Poisson criterion only. *MLE-Poisson* means that the GRD algorithm was used and that during expansion, the Poisson criterion picked the M best questions at each node ($M = 10$ for these experiments), from which the MLE criterion made the final choice; the pruning step used

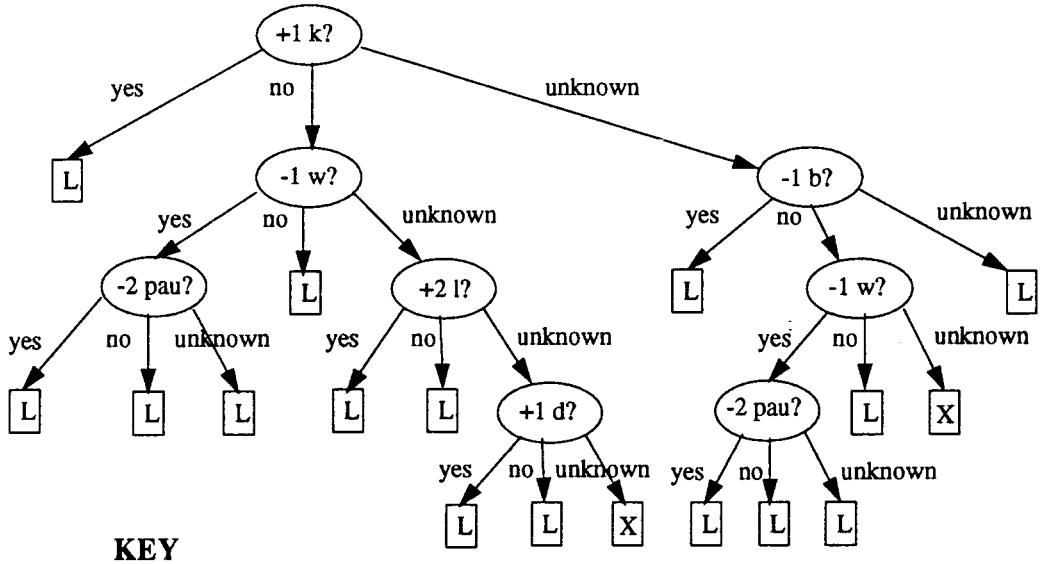

Figure 1: Tree for "uh" Grown with Questions about Phones

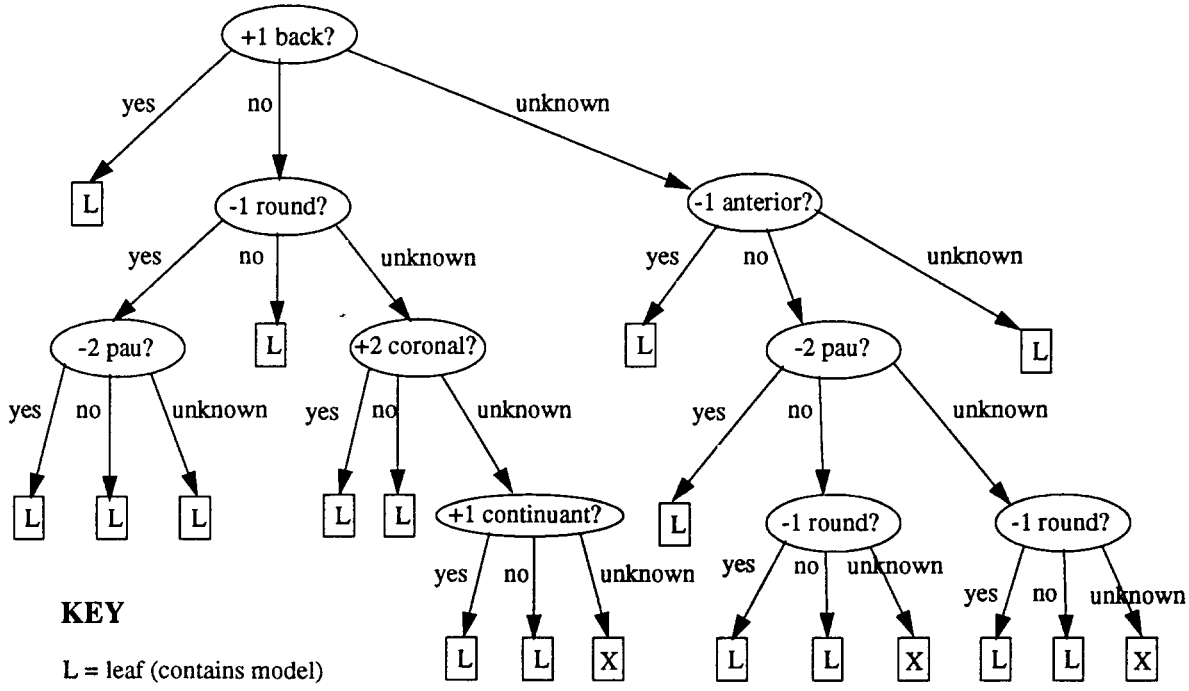

Figure 2: Tree for "uh" Grown with Feature Schema 1 (Chomsky-Halle)

only the MLE criterion. *SPAN* is the width of context for questions, *# HMM* is the number of models in the leaves of all the trees generated, and *% REC.* is the word percent recognition obtained by our simplified system using these models.

Contrary to expectations, our best result for each question type is obtained with Poisson rather than MLE-Poisson training, and with the narrow $+-1$ span. Phone-based and Chomsky-Halle question types appear to yield the best trees. The only advantage of the MLE-Poisson training appears to be that the resulting trees are smaller by about 20% than trees obtained by Poisson training, for a very slight performance deterioration. For comparison with the $+-1$ results shown here, we carried out an experiment with the standard triphone approach; we ended up with 11931 HMMs (7.5 times more models) and attained an inferior performance level (78.3% word recognition).

How often does the MLE criterion agree with the Poisson criterion, i.e. choose the first question on the Poisson M-best list? For the 40 trees obtained on the first MLE-Poisson GRD expansion (span $+-1$), let 0 to 9 be the Poisson rank of the best to worst question. For phone-based questions, the average rank of the question chosen by MLE is 0.59, and MLE picks a new best question (i.e. question other than one with rank 0) 29% of the time. For CH questions, average rank is 0.50 and a new question is picked 27% of the time. For Artic. questions, average rank is 0.68 and a new question is picked 36% of the time. The low average rank indicates that the Poisson criterion picks similar questions to MLE. We can also look at outliers: how often does MLE choose questions ranked 7, 8, or 9 on the Poisson M-best list? This happens 0.7% of the time for phone-based questions, 0.2% for CH questions, 0.2% for Artic. questions.

TYPE	CRITERION	SPAN	# HMM	% REC.
Phones	MLE-Poisson	+- 1	1341	78.80
Phones	Poisson	+- 1	1610	78.95
Phones	Poisson	+- 2	2296	78.30
Phones	Poisson	+- 3	3155	77.91
CH	MLE-Poisson	+-1	1340	78.71
CH	Poisson	+- 1	1600	78.93
CH	Poisson	+- 2	3215	77.87
CH	Poisson	+- 3	3849	77.68
Artic.	MLE-Poisson	+-1	1365	78.70
Artic.	Poisson	+- 1	1642	78.78
Artic.	Poisson	+- 2	3090	77.74
Artic.	Poisson	+- 3	3831	77.66

Table 1: Phonetic Context Tree Recognition Results

7. CONCLUSIONS AND FUTURE WORK

We have applied the GRD algorithm to phonetic trees, and introduced the idea of "DON'T KNOW" children. Apart from this, our most important result is that the Bahl *et al* Poisson criterion for picking questions is a very good one. The more computationally intensive MLE criterion is usually in very close agreement with the Poisson criterion; the only advantage of MLE is that it consistently yields smaller trees (at a very slight performance cost). We attribute the disappointing results for spans wider than $+-1$ to overtraining (i.e., insufficient data); the larger size of trees with wider spans lends support to this hypothesis.

Our plans for future work include:

- Recent work applies context-dependent grouping to each state or transition of an HMM, rather than to the whole HMM [You94], [Bah94]. Our algorithms could be applied here with only minor changes; we plan to carry out the appropriate experiments soon.
- We will experiment with mixed question sets with both phone- and feature-based questions. Perhaps it is important to know the precise identity of the -1 and $+1$ phones, while at more distant positions knowledge about one or two features suffices.
- Bahl *et al*'s criterion could be described as a Poisson MLE criterion. We have devised a Maximum Mutual Information Estimation (MMIE) variant of their criterion that employs the same Poisson model. With this, it would be possible to choose questions in the tree for "ph" in a way that lowers the risk that "ph" will be confused with other phones, which is precisely the criterion desired for speech recognition.

8. REFERENCES

[Bah94] L. R. Bahl, P. V. de Souza, *et al*, "Robust Methods for Using Context-Dependent Features and Models in a Continuous Speech Recognizer", *ICASSP-94*, V. 1, pp. 533-536, April 1994.

[Bah91] L. R. Bahl, P. V. de Souza, *et al*, "Decision Trees for Phonological Rules in Continuous Speech", *ICASSP-91*, V. 1, pp. 185-188, 1991.

[Bre84] L. Breiman, J. Friedman, *et al*, "Classification and Regression Trees", Wadsworth Inc., 1984.

[Gel91] S. Gelfand, C. Ravishankar, and E. Delp, "An Iterative Growing and Pruning Algorithm for Classification Tree Design", *IEEE Trans. PAMI*, V. 13, no. 2, pp. 163-174, Feb. 1991.

[Lad82] P. Ladefoged, "A Course in Phonetics", Harcourt Brace Javanovich, 1982.

[Lee89] K.-F. Lee, "Automatic Speech Recognition - the Development of the SPHINX System", Kluwer Academic Publishers, 1989.

[Moo94] R. Moore, M. Russell, *et al*, "A Comparison of Phoneme Decision Tree (PDT) and Context Adaptive Phone (CAP) Based Approaches to Vocabulary-Independent Speech Recognition", *ICASSP-94*, V. I, pp. 541-544, April 1994.

[Sch93] E.G. Schukat-Talamazzini, H. Niemann, *et al*, "Automatic Speech Recognition without Phonemes", *Eurospeech 93*, V. I, pp. 129–132, Berlin, 1993.

[You94] S. J. Young, J. Odell, and P. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling", *ARPA Workshop on Human Language Technology*, pp. 286-291, Mar. 1994.