REFINING TREE-BASED STATE CLUSTERING BY MEANS OF FORMAL CONCEPT ANALYSIS, BALANCED DECISION TREES AND AUTOMATICALLY GENERATED MODEL-SETS

Daniel Willett, Christoph Neukirchen, Jörg Rottland, Gerhard Rigoll

Department of Computer Science
Faculty of Electrical Engineering
Gerhard-Mercator-University Duisburg, Germany
e-mail: {willett,chn,rottland,rigoll}@fb9-ti.uni-duisburg.de

ABSTRACT

Decision tree-based state clustering has emerged in recent years as the most popular approach for clustering the states of context dependent hidden Markov model based speech recognizers. The application of sets of phones, mainly phonetically motivated, that limit the possible clusters, results in a reasonably good modeling of unseen phones while it still enables to model specific phones very precisely whenever this is necessary and enough training data is available. Formal Concept Analysis, a young mathematical discipline, provides means for the treatment of sets and sets of sets that are well suited for further improving tree-based state clustering. The possible refinements are outlined and evaluated in this paper. The major merit is the proposal of procedures for the adaptation of the number of sets used for clustering to the amount of available training data, and of a method that generates suitable sets automatically without the incorporation of additional knowledge.

1. INTRODUCTION

The great importance of parameter tying in hidden Markov model (HMM) based speech recognition [5] has often been stated. Young [11] has given a good summary on all the parameters in continuous speech recognizers that are possible candidates for being tied. Tying always aims on a reduction of the number of free system parameters while it tries to maintain a suitable acoustical resolution. The tying of HMM states, often referred to as state clustering, is a procedure that is essential for context dependent HMM-based speech recognition systems, as usually lots of phones are never observed in specific contexts and others are observed too sparsely to allow the estimation of individual models.
The idea of decision tree-based state clustering is the gradual separation of the whole set of initially clustered triphone states by some likelihood criterion [1, 3, 12]. In the terminology of set theory that will be used in the following text, the common node-splitting criterion formulates as

$$\underset{\mathcal{P},\mathcal{S}}{\operatorname{argmax}}\left(L_{\mathcal{P}\cap\mathcal{S}} + L_{\mathcal{P}\cap\overline{\mathcal{S}}} - L_{\mathcal{P}}\right) \tag{1}$$

which means, that the cluster $\mathcal{P}$ and question $\mathcal{S}$ are chosen, which lead to maximum (log-)likelihood gain on the training data when splitting $\mathcal{P}$ by intersecting with $\mathcal{S}$. The number of possible questions $\mathcal{S}$ is usually strictly limited in order to keep the computational cost reasonably low. It has to be considered that there are $2^{(n-1)}$–1 possible ways to split a set of n elements into two nonempty parts and in triphone systems n, the number of HMM states, amounts to several thousands. The possible questions $\mathcal{S}$ are most commonly defined by the incorporation of phonetical knowledge [12]. Besides the reduction of the computational complexity, restricting the node-splitting procedure to a wisely chosen set of questions results in a reasonably good modeling of unseen triphones.
The gradual clustering process can be visualized in a tree whose root node represents the whole set of states, while the arcs represent the intersection with one of the predefined sets of models and its complementary set. Fig. 1 shows such a tree. Most commonly, all the states of identical HMM position and identical central phone are clustered in a separate procedure. Paul [8], however, suggested to cluster all states in a single tree, not separating between different central phones and HMM position. In the case of multiple trees, $\mathcal{A}$

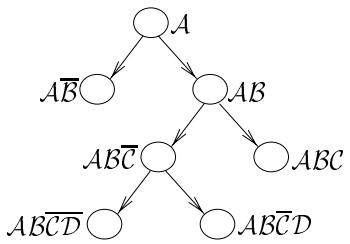


Figure 1: Splitting the set $\mathcal{A}$ of HMM states according to some subsets $\mathcal{B}$,$\mathcal{C}$ and $\mathcal{D}$

represents all the states of a specific position with a specific central phone (e. g. second states of all **/th/**). The sets $\mathcal{B}$,$\mathcal{C}$ and $\mathcal{D}$ are subsets of the whole set of states (e. g. *Vowel/*/** or **/*/Front*). One of the most important issues in the tree-based clustering procedure is the question which sets to allow as possible splits within the tree nodes. In the original approach [1], the specific subsets are set up using phonetic questions concerning the models' left and right context phone. Hence, in the following text, the term question will always be used for those sets that define the possible splits, although these sets will not always be defined by questions concerning the phonetical class of the context phone. In [2] and [6], similarity measures were used to generate suitable questions. It has been shown that the incorporation of additional (phonetical) knowledge is not crucial to the success of tree-based clustering, but that a wisely chosen similarity measure can do just the same or even better.
Formal Concept Analysis was introduced by Wille in 1982 [10]. It is a theory of data analysis which identifies conceptual structures among data sets. The web-page [9] is a good place to find addi-

tional references and introductions that are omitted in this paper. Besides the basic concepts of treating sets of attributes and sets of entities that define attributes, some algorithms [4] that are essential to this theory are interesting for being used within the context of tree-based clustering.

2. A MINIMUM NUMBER OF SETS

Intuitively, Odell [7] used phonetic categories (e. g. Vowel or Front) as well as several intersections of these categories (e. g. Front-Vowel) and single phone questions. This was done, although the tree is capable of producing intersected sets (and most of the single-phone questions as well) by successively applying its question set. In Figure 1 for example, if we think of $\mathcal{B}$ representing all left vowel triphones (*Vowel/*/**) and $\mathcal{C}$ representing all the left front triphones (*Front/*/**), we see that the subset $\mathcal{ABC}$ is producible as a tree-node. However, producing it comes along with a separation of the remaining triphones into the clusters $\mathcal{A}\overline{\mathcal{B}}$ and $\mathcal{A}\overline{\mathcal{BC}}$, that might not be intended. Figure 2 illustrates this difference. The left hand side shows the decision tree with the single question $\mathcal{BC}$ available, the right hand side shows a hypothesized decision tree with this question not available. In order to find out, whether the incorporation

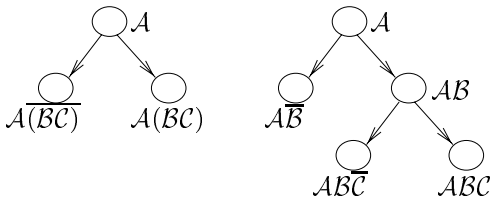


Figure 2: Tree with and without the question based on the intersection $\mathcal{BC}$

of intersected questions, as it is usually done, is really useful, a first line of experiments has been carried out. With a greedy algorithm that removes each question that can be generated by intersecting the others (including their complementary counterparts), a minimal question set was identified. In Formal Concept Analysis this greedy algorithm is used for computing minimal generators for sets of sets of attributes. Its optimality has been proven, although it is not relevant in this context.

We found that 21 phonetic categories are sufficient to generate all the questions proposed in [7], while single phone questions where omitted in these tests as by intersecting them (and the complementary sets), the generation of all the over 10^6 sets is possible which is not wanted in this case. This circumstance will be exploited, however, in Section 5.

Experimental results on only using questions based on the generating phone sets are presented in Section 7. It turns out that including sets that can be generated by intersecting others improves the success of the decision tree-based clustering procedure. This is probably mainly due to the circumstance which has been described above, namely that the intersection of questions within the decision tree leads to separated sets in the remaining nodes.

3. A MAXIMUM NUMBER OF SETS

In the previous section, we found out that the incorporation of intersected questions can be useful, although they can be constructed in the tree anyway. Thus, an approach, or at least an interesting experiment, could be to supply the clustering procedure with all the possible intersections of all the basic (phonetically motivated) questions. In terms of Formal Concept Analysis one would say that all the resulting sets have a certain attribute structure, or that each of them describes an individual concept. A very nice algorithm for computing this lattice of sets efficiently is Ganther's Next-Closure algorithm [4]. This algorithm is based on an artificial lexical ordering of the sets' elements. It generates the lattice of sets according to this order by successively intersecting those sets that contain some minimum elements that are chosen according to the introduced order. The two possibilities for applying this algorithm are outlined in Figure 3. On the one hand, the algorithm can be applied

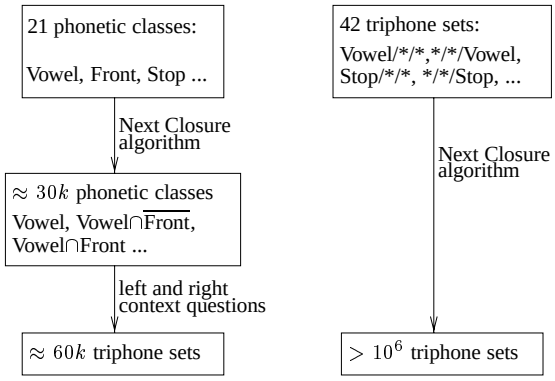


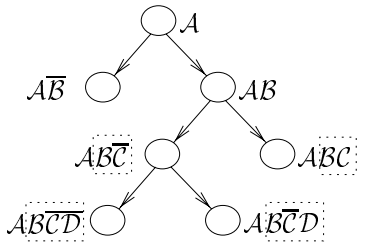
Figure 3: Applying Ganther's algorithm at the phone and at the triphone level

on the 21 phone sets that have been identified to generate all the phonetical categories, i. e. all possible intersections of phonetic classes. Starting with these 21 phone sets the algorithm identifies about 30k different intersections. The application of these 30k phonetic categories as questions concerning the left and the right context leads to about $2 \cdot 30k = 60k$ questions. On the other hand, applying the Next-Closure algorithm directly on the 42 sets of triphones based on the questions concerning the left- and the right context results in a number of sets of over 10^6.

Because of memory limitations, we were only able to run experiments with the 60k questions based on the intersection of phonetical categories. The results are summarized in Section 7. A slight increase in recognition accuracy could be measured, compared to the baseline system that uses Odell's question set.

4. A REASONABLE NUMBER OF REASONABLE SETS

In order to cope with the numerous ($> 10^6$) sets that are the result of intersecting on the triphone set level instead of the level of phonetic categories, we propose to follow a special technique, namely the iterative generation of relevant intersections using the tree-based clustering procedure itself. Started with only the 42 questions based on the generating 21 phonetical categories, the tree-based clustering procedure itself produces those intersections, that are interesting concerning the likelihood criterion, in its tree nodes. These can be added to the question set for a second clustering procedure. The figures 4 and 5 illustrate this iterative process, that can be repeated several times as long as one assumes the number of generated sets too small, or as long as the recognition accuracy of the resulting recognition systems improves on some cross-reference test set. Experimental results using the proposed technique are compiled in Section 7. The enlargement of the number of sets leads to a measurably improved recognition accuracy.



$\Rightarrow$ New questions: $\mathcal{BC}, \mathcal{B}\overline{\mathcal{C}}, \mathcal{B}\overline{\mathcal{CD}}, \mathcal{B}\overline{\mathcal{C}}\mathcal{D}$

Figure 4: First pass with only the sets $\mathcal{B}$,$\mathcal{C}$,$\mathcal{D}$ available for splitting

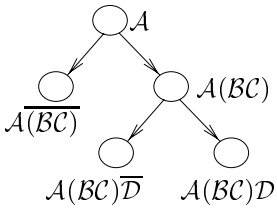


Figure 5: Second pass with the sets $\mathcal{BC}$, $\mathcal{B}\overline{\mathcal{C}}$, $\mathcal{B}\overline{\mathcal{CD}}$ and $\mathcal{B}\overline{\mathcal{C}}\mathcal{D}$ additionally available

5. AUTOMATICALLY GENERATING MODEL-SETS

As a spinoff of the iterative procedure of the previous section, an automatic clustering algorithm is obvious, that allows using the tree-based clustering approach without the incorporation of phonetic or other knowledge sources at all. It is illustrated in Figure 6. Instead of the phonetic categories that were used in the previous section as initial questions, the multi-pass procedure can be initialized with only those questions that are defined by a specific left or right phone (such as */*/*th*, *ng*/*/*, etc.), and thus use no initial sets that are based on phonetic categories. The first pass of the procedure proposed in Section 4 results in several thousands of triphone sets, that can be used to define the questions of the second pass. Again, several passes can be applied in order to generate an even larger amount of questions. In the experiments, the recognition system based on the tree-based clustering procedure of this second pass even outperforms the one that uses those sets based on the phonetic questions as proposed in [7].

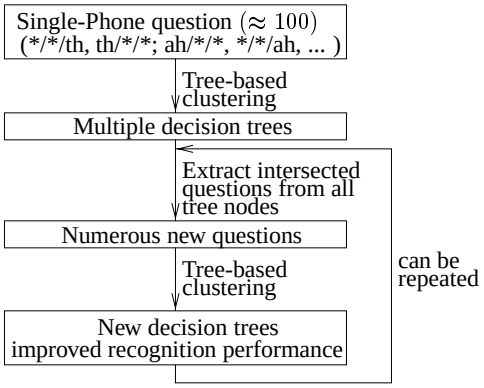


Figure 6: Tree-based clustering with generated model-sets

6. BALANCED DECISION TREES

Observations showed that the decision trees are often very degenerated and unbalanced. Especially when enlarging the question set or generating it automatically using the procedures described in the previous sections, or when applying single-phone context questions as in [7] in addition to the phonetically motivated questions, the splits that occur in the tree nodes are often very unbalanced. They often split off only a few elements and leave the major part of models behind for being split again in the next level of the tree. Fig. 7 shows an unbalanced tree on the left. As far as computational complexity is concerned, this observation does not concern, as there is no need to traverse the tree during training or recognition. However, extracting only a few models and leaving misbalanced clusters behind, is dangerous for the modeling of unseen and sparsely observed triphones. They might remain in the bigger cluster, although separating them more evenly would be better with respect to discrimination. In order to cope with this observation, we ran a couple of experiments in which we tried to prefer those splits that lead to well balanced decision trees. Fig. 7 shows a balanced tree on its right side. More balanced decision trees can be set up by a moderate modification of the node splitting criterion that favors the more even splits. Several modifications are possible. The one we used is the criterion

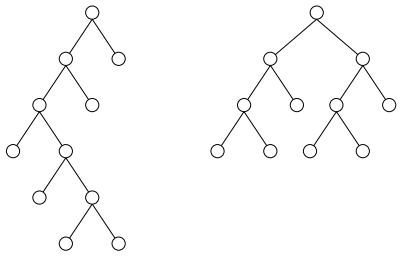


Figure 7: Unbalanced and balanced decision tree

$$\underset{\mathcal{P},\mathcal{S}}{\operatorname{argmax}} \left(L_{\mathcal{P}\cap\mathcal{S}} + L_{\mathcal{P}\cap\overline{\mathcal{S}}} - L_{\mathcal{P}} - \beta \frac{(|\mathcal{P}\cap\mathcal{S}| - |\mathcal{P}\cap\overline{\mathcal{S}}|)^2}{|\mathcal{P}|^2} \right) \quad (2)$$

where $L_{\mathcal{P}}$ is the log-likelihood of the parent node with its set of triphones $\mathcal{P}$, $L_{\mathcal{P}\cap\mathcal{S}}$ and $L_{\mathcal{P}\cap\overline{\mathcal{S}}}$ are the log-likelihoods of the child nodes when splitting $\mathcal{P}$ by intersecting with set $\mathcal{S}$, $||$ is the number of set elements and β is a balancing factor that controls the degree of balancing.
When using the several thousands of generated question of Section 4 or 5, the modified splitting criterion Eq. 2 offered another slight improvement.

7. EXPERIMENTS AND RESULTS

In order to evaluate the proposed procedures, several experimental speech recognition systems were set up on the Wall Street Journal (WSJ0) database. The recognition accuracy was measured on the Nov.'92 evaluation set. The Limsi-phoneset and 5k-dictionary were used, as well as the standard bigram language model of perplexity 110. In each of the clustering experiments the tree depth was adjusted in order to result in recognition systems with approximately the same number of parameters, i. e. the same number of triphone clusters. After the tree-based clustering procedure that is based on single Gaussian mixture models, the number of mixture components of all pdfs in all the experiments was enlarged

to 10 Gaussians per HMM state. The baseline system uses the question set as proposed by Odell in [7]. The first row in Table 1 shows the achieved recognition accuracy when using this question set in ordinary multi tree mode with a separate tree for each state of the models with a specific central phone. Below, the results are given for Paul's [8] single-tree approach. We evaluated it, to find out, whether allowing clusters of different central phones provides an improvement in recognition accuracy for this recognition task. The results suggest that this is not the case. The slightly reduced accuracy is probably due to a decreased discrimination that comes along with the clustering of triphone states of different central phone. Hence, in the following experiments each set of models with a specific central phone was clustered separately. Row 3 lists the error rate achieved with Odell's question set reduced by the single-phone questions. Interesting to notice is that the recognition accuracy only decreases very slightly against the one of Row 1 with all the questions.

experiment number	description	word error [%]
1	baseline (multiple trees)	12.04
2	single-tree [8]	12.10
3	no single phone questions	12.06
4	minimal question set first pass (Section 2)	12.32
5	minimal question set second pass (Section 4)	11.73
6	enlarged question set 60k questions (Section 3)	11.68
7	enlarged question set (60k) balanced according to Eq. 2	11.62
8	single phone questions only first pass (Section 5)	12,84
9	single phone questions only second pass $\Rightarrow$ generated sets	11.75
10	single phone questions only second pass $\Rightarrow$ generated sets balanced according to Eq. 2	11.41

Table 1: Recognition performance achieved in the experiments

The application of the reduced minimal question set that only contains those $2 \cdot 21$ questions that are required to generate all of the original ones (see Section 2) results in an increased error rate. Running the multiple pass procedure, as proposed in Section 4, started with this minimal question set in the first pass, however, results in a remarkably good recognition performance (Row 5). Allowing all the 60k possible splits that are the result of intersecting all the phonetical categories, as outlined in Section 3, leads to the best word error rate measured to this point (Row 6). Balancing the decision tree according to Eq. 2 provides another slight improvement (Row 7).

Not surprisingly, the first pass of the procedure proposed in Section 5 which only uses the single-phone questions degrades against the other experiments that apply more profound model-sets. The second pass (Row 9), however, that uses a question set of all the intersections of sets found in the trees of the first pass, achieves the same performance as the best phonetically-based system so far. Balancing the decision-tree again provides another slight improvement (Row 10). It is the lowest word error rate observed in our experiments on tree-based clustering. It should be noticed that this performance was achieved without the incorporation of phonetical knowledge and without the need for additional clustering procedures and similarity measures.

8. CONCLUSION

The paper has presented several means for efficiently combining the initial triphone sets for an improved tree-based clustering performance. It has been shown that the proposed methods can effectively be used to increase the accuracy of context dependent HMM-based speech recognizers. Additionally, an automatic procedure has been proposed that uses the tree-based clustering procedure itself to construct the set of initial sets of triphones that limit the possible splits within the tree nodes. This procedure not only achieves the same performance as the standard approach that incorporates additional (phonetical) knowledge, but even outperforms it. Furthermore, the idea of balanced decision trees has been presented as well as experiments which showed that a soft balancing of the trees can be useful.

9. REFERENCES

[1] L. R. Bahl et al.: "Context Dependent Modelling of Phones in Continuous Speech Using Decision Trees", DARPA Speech and Natural Language Processing Workshop, Pacific Grove, 1991, pp. 264–270.

[2] K. Beulen, E. Bransch, H. Ney: "State Tying for Context Dependent Phoneme Models", ICASSP '98, Seattle, pp. 1179–1182.

[3] J. Duchateau, K. Demuynck, D. Van Compernolle: "A Novel Node Splitting Criterion in Decision Tree Construction for Semi-Continuous HMMs", Eurospeech'97, Rhodes, pp. 1183–1186.

[4] B. Ganther: "Two basic algorithms in concept analysis", Technical Report, TH Darmstadt, FB4, 1984.

[5] X. D. Huang, Y. Ariki, M. A. Jack: "Hidden Markov Models for Speech Recognition", Edinburgh University Press, 1990.

[6] A. Kosmala, G. Rigoll: "Tree-Based State Clustering Using Self-Organizing Principles for Large Vocabulary On-Line Handwriting Recognition", ICPR '98, Brisbane, pp. 1313–1316.

[7] J. J. Odell: "The Use of Context in Large Vocabulary Speech Recognition", PhD thesis, University of Cambridge, 1996.

[8] D. Paul: "Extensions to Phone-State Decision-Tree Clustering: Single Tree and Tagged Clustering", ICASSP '98, Munich, pp. 1487–1490.

[9] U. Priss: "A Formal Concept Analysis Homepage", http://php.indiana.edu/ upriss/fca/fca.html.

[10] R. Wille: "Restructuring Lattice Theory: An Approach Based on Hierarchies of Concepts". Ordered Sets, D. Reidtel, Dordrecht, 1982, pp. 445–470.

[11] S. J. Young: "The General Use of Tying in Phoneme-Based HMM Speech Recognition", ICASSP '92, Adelaide, pp. I 569–572.

[12] S. J. Young, J. J. Odell, P. C. Woodland: "Tree-Based State Tying for High Accuracy Acoustic Modelling", Human Language Technology Workshop, Plainsboro, 1994, pp. 307–312.