AUTOMATIC QUESTION GENERATION FOR DECISION TREE BASED STATE TYING

K. Beulen, H. Ney

Lehrstuhl für Informatik VI, RWTH Aachen, University of Technology, D-52056 Aachen

ABSTRACT

Decision tree based state tying uses so-called phonetic questions to assign triphone states to reasonable acoustic models. These phonetic questions are in fact phonetic categories such as vowels, plosives or fricatives. The assumption behind this is that context phonemes which belong to the same phonetic class have a similar influence on the pronunciation of a phoneme. For a new phoneme set, which has to be used e.g. when switching to a different corpus, a phonetic expert is needed to define proper phonetic questions. In this paper a new method is presented which automatically defines good phonetic questions for a phoneme set. This method uses the intermediate clusters from a phoneme clustering algorithm which are reduced to an appropriate number afterwards. Recognition results on the *Wall Street Journal* data for within-word and across-word phoneme models show competitive performance of the automatically generated questions with our best handcrafted question set.

1. INTRODUCTION

In the last years state tying has become a baseline feature of most state-of-the-art speech recognition systems [2, 5, 8, 9, 12]. Most of these systems [2, 5, 8, 12] use a decision tree based framework where the acoustic model of a triphone state is determined by a phonetic decision tree. A phonetic decision tree [1] classifies phonemes using phonetic questions. This approach has the advantage that every possible triphone state can be classified by the tree, so any backing-off models can be avoided. The drawback is that before the decision tree can be constructed, the phonetic questions which categorize the phonetic context of a triphone state must be defined. These questions are quite similar to phonetic categories such as vowels, plosives or fricatives. However, the definition of a phonetic question set for a new corpus is often a time consuming and error-prone process. The assumption behind the choice of phoneme classes as questions is that context phonemes which belong to the same phonetic class have a similar influence on the pronunciation of a phoneme. To define such classes, good phonetic knowledge is needed. But even then it it difficult to define 40-50 reasonable phonetic classes, a significantly smaller number of questions would restrict the tree construction too much. So the problem is twofold, namely the *quantity* and the *quality* of the phonetic question set.

In this paper a new method is presented which automatically defines a sufficient number of good phonetic questions for a given phoneme set. The method first builds up a set of candidate questions which are derived from a phonetic clustering procedure. The phonemes used for this clustering procedure can either be monophones or diphones. Then this set of candidate questions is reduced by pruning away less important questions. The final question set can then be used to grow a normal decision tree.

The tests presented in this paper were performed using the RWTH speech recognition system partly described in [11, 12, 13]. The training corpus was the *Wall Street Journal* WSJ0 corpus, and the recognition tests were carried out on the *Wall Street Journal* November 92 development and evaluation test corpora for the 5 000 word lexicon. The experiments we made show the following results:

- The automatically generated question set performs at least as well as our best handcrafted question set.
- The error rates of the diphone based method as defined in Section 4.2 are minimally better than the error rates of the phoneme based method.
- The question generation method also works for tasks with a significant number of unseen triphones, e.g. across-word models.

The paper is structured in the following way: Section 2 gives a short overview of our baseline state tying approach, in Section 3 the handcrafted question set we are using is presented, Section 4 describes the algorithm which generates the question set, in Section 5 the corpora and the recognition system is presented, Section 6 contains the experiments using the new method, and finally Section 7 gives some conclusions and an outlook on possible modifications of the baseline method.

2. STATE TYING

The aim of state tying is to reduce the number of free parameters of a speech recognition system so the remaining parameters can be estimated more robustly. Therefore, triphone states whose emission probability distributions are very similar according to a distance measure, are tied together. These tied states share the same parameter set which is trained using all the observations assigned to this state set. For calculating the tying of the states, a single Gaussian distribution is estimated for every triphone state using a good segmentation of the data. The states are then grouped into the root node AB of the initial decision tree. This node is then splitted using the phonetic question out of a question set which yields the biggest likelihood improvement $D(A, B)$ for the child nodes A and B:

$$D(A, B) = LL(AB) - (LL(A) + LL(B)) \tag{1}$$

$$= -\frac{1}{2}\left(n_A \sum_{d=1}^{D} \log \left[\frac{\hat{\sigma}_{d,AB}}{\hat{\sigma}_{d,A}} \right]^2 + n_B \sum_{d=1}^{D} \log \left[\frac{\hat{\sigma}_{d,AB}}{\hat{\sigma}_{d,B}} \right]^2 \right) \tag{2}$$

where n_X is the number of observations for node X, D the dimensionality of the feature vector and $\sigma_{d,X}$ the variance of component d of node X. The question is assigned to the tree node, and the triphone states are distributed over the two child nodes according to the question. This procedure is repeated until the desired number of tree leaves is achieved. After this construction process the resulting decision tree is used for training and recognition.

The formula for the log-Likelihood improvement can be easily rewritten to a form which only contains sums and square sums of the observation vector components together with the observation counts:

$$\hat{\mu}_{d,X} = \frac{1}{n_X} \sum_{y \in X} y_d \tag{3}$$

$$\hat{\sigma}^2_{d,X} = \frac{1}{n_X} \sum_{y \in X} y_d^2 - \hat{\mu}^2_{d,X} \tag{4}$$

where y denotes an observation vector with components $d = 1, \ldots, D$, and X a state with the observation count n_X.

This implies to calculate these sums beforehand to avoid a complete training iteration for every tree construction. The obvious drawback is that the segmentation of the data is not specific for the constructed tree.

3. DECISION TREE QUESTIONS

The structure of a binary decision tree is a very simple one. It consists of inner nodes which contain decision rules and outer nodes or leaves which are labelled with the classes. To classify an object according to its features, the feature vector is first classified by the decision rule at the root node. This decision rule assigns the object to the left or right subtree. Then the object is classified using the decision rule of the subtree root node and so on. When the object reaches a leaf, the class label at this leaf is used as the class for the object. The decision rules or questions are generally formulated using propositional logic as e.g.:

$$x_4 \in A \quad \Rightarrow \quad \text{is feature } x_4 \text{ in set } A \text{ ?} \tag{5}$$

$$x_9 < 3.2 \quad \Rightarrow \quad \text{is feature } x_9 \text{ smaller than 3.2} \tag{6}$$

The standard question set we use for english corpora is derived from the question set descibed in [4]. It consists of 44 questions about phonetic properties, e.g. *vowels* or *fricatives*. This question set was enlarged by 43 questions about special phonemes, e.g. *aa* or *uh*, and one question about the word boundary. We also tried a different question set listed in [5], but this set performed sligthly worse.

Table 1 contains as a representative example some statistical information about the 10 most important questions of a specific decision tree clustering of left triphone states.

Table 1: Clustering of the left triphone states

question	occurrences	rel. gain in LL [%]
ORAL-STOP1	14	4.9
VOWEL	26	4.7
SONORANT	72	4.6
LAX-VOWEL	20	3.5
S/Z/SH/ZH	42	3.2
LIQUID	30	2.8
TENSE-VOWEL	22	2.5
R-LABIAL	44	2.2
PALATL	34	2.1
LIQUID-GLIDE	20	2.0

This 'top ten' selection contains one the one hand very broad phonetic classes such as *vowel* and *sonorant* and one the other hand very specific questions such as *s/z/sh/zh* or *palatl*. This shows that a rich and reasonable phonetic question set is likely to perform better than a more or less random question set.

A construction method for decision trees described in [7] does not need any phonetic questions. However, to classify unseen triphones using such a tree some major extensions to the baseline algorithm have to be made which use phonetic questions themselves [10].

4. PROPOSED METHOD

The idea of the automatic question generation method is to automatically find reasonable phoneme classes that can be used as questions in the tree construction process. These classes do not have to have direct correspondants in the handcrafted question set, because for automatic speech recognition the criteria for 'good' phoneme classes can be quite different from phonetic knowledge. In fact, it will be shown later that the 'obvious' method for estimating phoneme classes for decision tree construction is not optimal. Instead, by grouping phonemes according to their similarity as context phonemes, better results can be obtained.

To automatically derive a proper set of questions for a specific task, we start with the same triphone states as for the tree construction process. For this state set, we calculate the sums and square sums as described above. Then the following steps are performed:

1. The triphone states are grouped according to their central phoneme (*phoneme clustering*) or to their central phoneme and left or right context (*diphone clustering*).
2. These groups are then clustered using a bottom-up cluster algorithm and the log-Likelihood distance measure until only one cluster remains.
3. The intermediate clusters during the clustering procedure are recorded.
4. These intermediate clusters are then used as a first question set for the tying procedure.
5. After the first tying the number of occurrences and the log-Likelihood gain for each question is used as a selection criterion to reduce the number of questions.
6. Questions for individual phonemes and for the word boundary are added.
7. This question set is then used to construct the decision tree for training and recognition.

4.1. PHONEME CLUSTERING

For the phoneme clustering method all triphone states with the same segment number are clustered. The intermediate clusters during the clustering procedure are then taken as phoneme classes. This is a reasonable assumption because during the clustering procedure these triphone state clusters are merged whose members, that means the phoneme states, are very similar due to the distance measure.

4.2. DIPHONE CLUSTERING

The idea of the diphone clustering method is to group the phonemes not according to their similarity but to their similarity as a left or right context. Because the main number of phonetic questions in a decision tree is about the context of a triphone state, the grouping of phonemes as context phonemes may be advantageous. To do so, we take into account the tripartite structure (left part, middle part, right part) of our phoneme models and first combine all triphone states of the left part with the same central phoneme and the same left context to left diphone states and all triphone states of the right part with the same central phoneme and the same right context to right diphone states. Then these diphone states are grouped according to their central phoneme. Inside these phoneme groups the same clustering method is used as for the phoneme clustering. After the clustering the *context* phonemes of the intermediate clusters are taken as the baseline phonetic question set. This gives us about 1550 questions in the initial question set. This set is then reduced the following way: A top-down decision tree based clustering is performed and the number of occurrences and the sum of log-Likelihood gains for each question are calculated. Due to these numbers the intermediate question set is reduced. To prevent the algorithm from choosing questions which are too specific for the training corpus, we use a cross validation scheme: The triphone set is split into two sets by chance. During the tree construction every node contains two clusters, one with triphone states from set 1 and one with triphone states from set 2. The log-Likelihood gain is calculated over both sets using the formula

$$\hat{D}(A,B) = (D(A_1,B_1)+D(A_2,B_2)) \cdot \left(\frac{2\sqrt{D(A_1,B_1)\cdot D(A_2,B_2)}}{D(A_1,B_1)+D(A_2,B_2)}\right)^2$$

where X_i is the subcluster i of cluster X, and N is the number of observations in all clusters. The second term is a weighting function for the 'raw' log-Likelihood gain which is equal to one when the log-Likelihood gain for both subclusters is equal and approaches zero when the log-Likelihood gain gets more and more unsymmetric. So those questions should be preferred which gave a relatively homogenous log-Likelihood gain on both subclusters. Due to the result of this decision tree construction the 'best' questions are choosen and then used in a second decision tree construction without any crossvalidation. This final tree is then used for training and recognition.

5. TEST ENVIRONMENT

The speech recognition system which was employed for the tests is also described in [11, 12, 13]. The most important properties are:

- 16 cepstral coefficients together with 16 first and 1 second order derivatives resulting in a 33-component acoustic vector,
- feature reduction by LDA [3],
- continuous HMM with Laplacian mixture densities,
- one single vector of absolute deviations for all distributions,
- Viterbi approximation for training,
- word conditioned search algorithm using a lexical prefix tree in combination with a bigram language model for recognition,
- acoustic rescoring using an n-best algorithm for across-word models.

The training was performed on the *Wall Street Journal* WSJ0 training corpus and the testing on the *Wall Street Journal* November 92 5 000 word development and evaluation test corpora. The test set contains 18 speakers and 12132 spoken words, it's bigram perplexity PP_{bi} is 107.

6. EXPERIMENTS

The first experiment we performed is concerned with the question generation method (see Table 2). In Section 4 we proposed the *phoneme* based and the *diphone* based question generation method. To evaluate both methods we generated two question sets. One question set based on the phoneme method consists of 93 general questions which were not pruned plus 44 phoneme questions including one question for the word boundary (*phonemes*). The other question set based on the diphone method consists of the 1550 questions mentioned above which were pruned down to 100 questions plus the 44 phoneme questions (*diphones*). These are compared to the results for the handcrafted question set (*baseline*). "#quest" means the number of questions in the question set, "ll gain" the log-Likelihood gain per observation vector due to the node splitting, "D-I" the number of deletions and insertions and "WER" the word error rate for the recognition test:

Table 2: Word error rate for phoneme and diphone method on WSJ Nov. 92

question set	#quest	ll gain	D-I [%]	WER [%]
baseline	88	4.20	1.3-0.7	7.1
phonemes	137	4.25	1.4-0.6	7.1
diphones	144	4.25	1.4-0.6	7.0

Obviously there is no big difference in the word error rate between the two automatic methods and the baseline method. However, the absolute difference between the phoneme method and the diphone method is about 20 errors, which means a small advantage for the diphone method.

The second experiment shown in Table 3 was run to find out how many questions should be selected from the diphone baseline question set. Therefore we have varied the threshold of absolute log-Likelihood gain while keeping the minimum occurrence by 3. This gave us three question set of 94, 144 and all questions.

Table 3: Word error rate for diphone method and different question sets on WSJ Nov. 92

#quest	ll gain	D-I [%]	WER [%]
94	4.25	1.4-0.6	7.1
144	4.25	1.4-0.6	7.0
1548	4.30	1.2-0.7	7.0

What can be observed is that the log-Likelihood gain decreases if only a subset of questions is used. However, the error rate is almost the same for all three question sets. This shows that the selection procedure seems to be reasonable for this specific task.

The third experiment shown in Table 4 was performed to test the question generation method in combination with across-word models. One consequence of using across-word models is a significantly higher percentage of unseen triphones. So if the automatically generated questions do not generalize as much as the handcrafted questions, the across-word results are likely to be significantly worse.

Using the same question set as for the within-word models, a new decision tree was constructed which was used for across-word training and recognition.

Table 4: Word error rate for diphone method and different word boundary modelling on WSJ Nov. 92

ac. modelling	quest. set	#quest	D-I [%]	WER [%]
within-word	manual	88	1.3-0.7	7.1
across-word	manual	88	1.0-0.8	6.4
across-word	automatic	94	1.0-0.7	6.5
across-word	automatic	144	1.0-0.6	6.3

The results in Table 4 show, as for the within-word models the automatically generated questions perform as well as the handcrafted questions. Using more efficient selection criteria, we hope to obtain further improvements compared to the actual within-word and across-word results.

7. CONCLUSIONS

We have presented a new method to automatically generate a set of phonetic questions for decision tree based state tying. This method uses only algorithms which are fast and easy to implement. Because the results in this early phase are very encouraging, especially if the very simple question selection scheme is taken into account, the chance of further improvements in recognition accuracy is high. But even if this assumption is not sound, a speech recognition system can profit from the method. Using automatic question generation one is able to switch very quickly to other languages without the need of a phonetic expert.

The main focus for the near future will be

- other question generation schemes,
- better question selection,
- tests on other corpora, especially for other languages.

8. REFERENCES

[1] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, The Wadsworth Statistics/Probability Series, Belmont, CA, 1984.

[2] C. Dugast, R. Kneser, X. Aubert, S. Ortmanns, K. Beulen, H. Ney, "Continuous Speech Recognition Tests and Results for the NAB'94 Corpus," *Proc. ARPA Spoken Language Technology Workshop*, Austin, TX, pp. 156-161, January 1995.

[3] R. Haeb-Umbach, H. Ney, "Linear Discriminant Analysis for Improved Large Vocabulary Continuous Speech Recognition," *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, San Francisco, CA, pp. 13-16, March 1992.

[4] H.-W. Hon, *Vocabulary-Independent Speech Recognition: The VOCIND System*, Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburg, PA, 1992.

[5] J. J. Odell, *The Use of Context in Large Vocabulary Speech Recognition*, Ph.D. Thesis, Cambridge University, Cambridge, March 1995.

[6] S.J. Young, J.J. Odell, P.C. Woodland, "Tree-Based State Tying for High Accuracy Acoustic Modelling," *Proc. ARPA Human Language Technology Workshop*, Plainsboro, NJ, pp. 405-410, Morgan Kaufmann, March 1994.

[7] P. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 13(4), pp. 340-354, 1991.

[8] M. Y. Hwang, "Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition," PhD Thesis CMU-CS-93-230, Carnegie Mellon University, 1993.

[9] L. Lamel, "Issues in Large Vocabulary, Multilingual Speech Recognition," *Proc. Europ. Conf. on Speech Communication and Technology*, Madrid, Spain, pp. 185-188, September 1995.

[10] H. J. Nock, M. J. F. Gales, S. Young, "A Comparative Study of Methods for Phonetic Decision-Tree State Clustering,", *Proc. Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, pp. 111-114, September 1997.

[11] L. Welling, N. Haberland, H. Ney, "Acoustic Front-End Optimization for Large Vocabulary Speech Recognition," *Proc. Europ. Conf. on Speech Communication and Technology*, Rhodes, Greece, pp. 2099-2102, September 1997.

[12] K. Beulen, E. Bransch, H. Ney, "State Tying for Context Dependent Phoneme Models," *Proc. Europ. Conf. on Speech Communication und Technology*, Rhodes, Greece, pp. 1179-1182, September 1997.

[13] S. Ortmanns, H. Ney, F. Seide, I. Lindam, "A Comparison of Time Conditioned and Word Conditioned Search Techniques for Large Vocabulary Speech Recognition," *Proc. Int. Conf. on Spoken Language Processing, Philadelphia*, PA, pp. 2091-2094, October 1996.