STATE TYING FOR CONTEXT DEPENDENT PHONEME MODELS

K. Beulen *E. Bransch* *H. Ney*

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

SUMMARY

In this paper several modifications of two methods for parameter reduction of *Hidden Markov Models* by state tying are described. The two methods represent a data driven clustering triphone states with a bottom up algorithm [3, 9], and a top down method growing decision trees for triphone states [2, 10]. We investigate several aspects of state tying as the possible reduction of the word error rate by state tying, the consequences of different distance measures for the data driven approach and modifications of the original decision tree approach such as node merging. The tests were performed on the test corpora for the 5 000 word vocabulary of the WSJ November 92 task and on the evaluation corpora for the 3 000 word VERBMOBIL '95 task. The word error rate by state tying was reduced by 14% for the WSJ task and by 5% for the VERBMOBIL task.

1. INTRODUCTION

Large vocabulary speech recognition systems model the acoustic realization of the words in the vocabulary with phoneme models. Due to the fact that the acoustic realization of the phonemes depends heavily on the phonetic context, it is essential for efficient speech recognition to model this context dependency [6, 8]. The most commonly used context dependent phoneme model is the phoneme model in a triphone context, in practice simply called *triphone*. Although triphones provide an excellent modelling of the context dependency, their exclusive use as acoustic models is prohibitive for vocabulary independent speech recognition because the set of triphones in the recognition vocabulary often contains triphones that cannot be observed in the training. Another serious problem is that many triphones occur very seldom in the training corpus so the estimation of the models may not be reliable. A possible solution of this problem is the so-called *state tying*. To improve the robustness of the parameter estimation the emission probabilities of the triphone states are shared between clusters of states which are similar according to a distance measure. The training data assigned to the states of one cluster is used to estimate the shared emission probability of these states.

In this paper several modifications of two well known methods for parameter reduction of *Hidden Markov Models* by state tying are described. The two methods are a data driven method which clusters triphone states with a bottom up algorithm [3, 9], and a top down method which grows decision trees for triphone states [2, 10]. We investigate the following aspects:

- The possible reduction of the word error rate by state tying,
- the consequences of different distance measures for the data driven approach and
- modifications of the original decision tree approach such as node merging.

The tests were performed on the test corpora for the 5 000 word vocabulary of the WSJ November 92 task and on the evaluation corpus for the 3 000 word vocabulary of the VERBMOBIL '95 task. The reduction of the word error rate by state tying is about 14% for the WSJ task and 5% for the VERBMOBIL task compared to simple triphone models.

2. STATE TYING

The aim of state tying is to reduce the number of parameters of the speech recognition system without a significant degradation in modelling accuracy. The states of the triphones used in training which are similar according to a distance measure are tied together. First, a suitable triphone list is assembled with respect to the training corpus. Because this list has to be quite large to achieve an accurate modelling of the acoustic context, simple models are used for the emission probabilities (one Gaussian density with diagonal or full covariance matrix). Using a segmentation of the training data the mean $\hat{\mu}_X$ and the variance $\hat{\sigma}_X^2$ of the triphone states X are estimated. The triphone states are then subdivided into subsets according to their central phoneme and their position within the phoneme model. Inside these sets the states are tied together according to a distance measure. Additionally it has to be assured that every model contains a sufficient amount of training data. The resulting models are then reestimated.

In this work we investigate two methods for state tying. One data driven method, which clusters triphone states with a bottom up approach, and a method which grows decision trees using a top down algorithm.

3. DATA DRIVEN METHOD

The data driven method [3] works in two steps. In the first step the triphone states being very much alike due to a distance measure are clustered together. In the second step the states which do not contain enough data are clustered together with the nearest neighbour. Then the resulting states are tied and finally reestimated with a higher acoustic resolution.

The main drawback of the data driven method is that for triphones which were not observed in the training

Table 1. Word error rates [%] on WSJ0, Nov.'92, 18 speakers, 12232 spoken words

Tying	Tying Variant	Triphones	Mixtures	Densities	Search Space (# States)	DEL-INS [%]	WER[%]
no tying	–	780	2338	246 000	–	–	8.8
data driven	appr. divergence	1856	2030	150 000	19 000	1.3-1.0	8.6
	log-likelihood	1856	2005	168 000	16 000	1.2-1.0	8.1
CART	one observation (baseline)	7834	2001	192 000	6 000	1.2-0.9	7.6
	50 observations	1833	2001	196 000	6 000	1.2-1.0	8.3
	20 observations	3025	2001	194 000	6 000	1.3-0.9	7.9
	intersection	5220	2001	194 000	7 000	1.2-0.9	7.8
	cross validation	1833	2001	194 000	6 000	1.3-0.9	7.9
	one tree	7834	2001	192 000	6 000	1.3-0.9	7.8
	merge-to-combine	7834	2001	200 000	6 000	1.3-0.9	7.8
	merge-to-reduce	7834	1714	176 000	6 000	1.3-0.9	8.0
	merge-to-reduce	7834	1412	154 000	7 000	1.4-0.8	8.0
	merge-to-reduce	7834	1130	130 000	7 000	1.5-0.8	8.3
	merge-to-reduce	7834	880	102 000	9 000	1.5-0.7	8.4
	GD coupling	1854	2001	198 000	6 000	1.1-1.0	8.0
	GD coupling	5857	2001	210 000	6 000	1.3-0.9	8.1
	smoothed GD coupling	7834	2001	195 000	6 000	1.2-0.8	7.7
	full covariance	3025	2001	196 000	6 000	1.2-1.1	8.0
	smoothed full covariance	3025	2001	212 000	6 000	1.2-0.9	7.9
	smoothed full covariance	7834	2001	199 000	6 000	1.2-0.8	7.7

Table 2. Word error rates [%] on VERBMOBIL '95

corpus	spoken words	CART	Triphones	Mixtures	Densities	Search Space (# States)	DEL-INS [%]	WER[%]
short95	3821	no	707	2122	136 000	10 000	7.7-4.4	34.0
		yes	4712	1501	119 000	11 000	6.3-5.2	32.2
long95	3383	no	707	2122	136 000	16 000	5.9-9.0	39.6
		yes	4712	1501	119 000	10 000	5.8-9.8	37.5

corpus no tied model is available. Thus these unseen triphones are modeled by so-called *backing off models*. Usually these models are simple generalizations of the triphones such as diphones or monophones. The training of the backing off models is performed on the data of the triphones which were not involved in the clustering process. In our tests we have used the triphones which were seen more than 50 times in the training corpus for clustering, and the complementary set for the training of the monophone backing off models.

For the data driven method we tested the following criteria: the approximative divergence [9] of two states and the log-likelihood difference between using only one model or two models for the observations of two clusters.

4. DECISION TREE METHOD

Decision trees are binary trees whose internal nodes are tagged with questions about the data which has to be classified, while the leaves are tagged with class labels. For our purposes we use phonetic questions as "Is the right context a vowel?" and tag the leaves of the tree with mixture labels. To find the appropriate models for a triphone state, one starts at the root of the appropriate tree, ask the questions on the triphone state and, according to the answer, branch to the left (*yes*) or to the right (*no*) until he reaches a leaf. The mixture label at the leaf identifies the mixture model for the triphone state.

The algorithm for tree construction starts with one single node for all the triphone states which have the same central phoneme and the same position within the

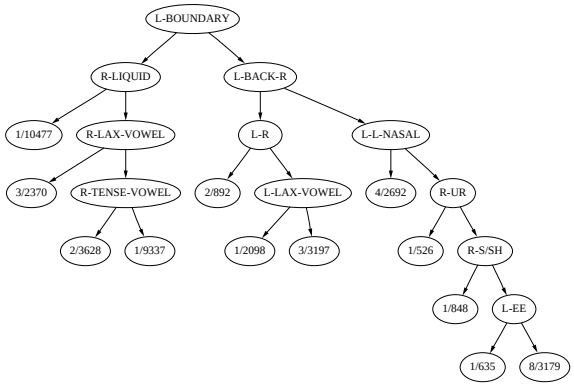

Figure 1. Decision tree for phoneme *th* (state 1).

phoneme model. The observations within every node are modelled with one Gaussian density with diagonal or full covariance matrix. Then the leaves of the tree are consecutively split with the questions which gives the largest local improvement in likelihood. If the improvement falls below a threshold for every possible split of every node, the algorithm stops.

One advantage of the decision tree method compared to the data driven method is that no backing off models are needed because using the decision trees one can find a generalized model for every triphone state in the recognition vocabulary.

Fig. 1 shows a decision tree for phoneme *th* (state 1). The inner nodes are labelled with the proper question, e.g. *L-BOUNDARY* means "Is the left context a word

boundary?". The leaves of the tree are labelled with the number of triphone states and the number of observations which belong to this leaf. One interesting observation is that this tree (and most other trees, too) has a bias to its right. This effect comes from the fact that most questions ask for a very special phoneme property. The consequence is that for these questions most triphone states belong to the right subtree. Another observation is that the rightmost leaf contains a lot of triphone states compared to the other leaves. This is because this node contains triphone states for which all the questions from the root to the leaf were answered by "no". So this triphone set is very heterogenous and new triphones which were mapped onto this state may be modelled not as good as by the other states of the tree.

We also tested the following modifications of the original method: A single tree instead of one distinct tree for every phoneme and state, different triphone lists, a simple cross validation scheme, two distinct models for male and female speakers in every tree node, an additional merging of nodes after the spitting process, and full covariance matrices instead of variance vectors for the Gaussian models in the nodes.

4.1. NO AD-HOC SUBDIVISION

In this method, the subdivision of the states according to their central phoneme and their position within the phoneme model is not used. Instead the algorithm starts with one single root node. The leaves are then splitted by asking questions not only on the context of the triphone state but also on the central phoneme and the position. The problem with this modification is the following: due to the possibility some leaves contain triphone states with different central phonemes, some words in the vocabulary cannot be discriminated any more. There are different solutions to this problem: split every node until it contains only states with one central phoneme or split every node until every word in the vocabulary can be discriminated. In the experiments, we found out that such "heterogenous" nodes are very rare and do not introduce any ambiguities in the lexicon. So we did not use any of the countermeasures listed above.

4.2. DIFFERENT TRIPHONE LISTS

In order to verify the effect of the triphone lists used for the decision tree construction we tested four different triphone lists (see table 1). One list containing all triphones from the WSJ November '92 training corpus (table 1, variant '*one observation*'), two lists containing the triphones from list '*one observation*' which were seen more than 20 or 50 times in the training corpus (table 1, variant '*20 observations, 50 observations*'), and a fourth list containing those triphones from the training corpus which can also be found in the test lexicon (table 1, variant '*intersection*').

4.3. CROSS VALIDATION

The use of cross validation for splitting the nodes was also tested. For cross validation the full triphone list was splitted into the triphones wich were seen more than 50 times in the training text (triphone list 1) and the complementary set (triphone list 2). Triphone list 1 was used to estimate the Gaussian models of the tree nodes while triphone list 2 was used to cross validate the splits. At every node the split with the highest gain in log-likelihood was made which also achieved a positive gain for the triphones of list 2.

4.4. GENDER DEPENDENT COUPLING

Because the training corpus contains two very different groups of speakers, namely male and female speakers, it could be advantageous to construct gender dependent decision trees. This approach has the disadvantage that the training data for the tree construction is being halved. Therefore we used the gender dependent (GD) method described in [7]. Every tree node contains two separate models for male and female data. The log-likelihood of the node data can then be calculated as the sum of the log-likelihoods of the two models.

4.5. NODE MERGING

The baseline algorithm uses only simple questions such as "Has the context X the property Y?" thus the leaves of the tree contain those triphone states for which the conjunction of the answers to the questions from the root to this leaf are true. To allow the construction of disjunctions, we implemented an additional merging of nodes. This merging is performed after the tree growing. The distances of all the leaves are calculated and then the two leaves with the smallest distance are merged. The merged node represents the triphone states for which the disjunction of the conjuncted answers are true. So every possible combination of questions can be constructed.

In our experiments we used two approaches. In the first approach we splitted the decision tree nodes to 3000 leaves and then merged these leaves to 2000 models (*merge-to-combine*). In the second approach we used the tree with 2000 leaves and then merged a significant number of leaves to reduce the number of models in the resulting recognizer(*merge-to-reduce*).

4.6. FULL COVARIANCE MATRIX

To increase the accuracy of the acoustic modelling of the training data, we replaced the diagonal covariance matrices of the Gaussian models by full covariance matrices. This modification results in a large increase in the number of parameters of the decision tree (here: factor of 18). So we implemented a smoothing method which interpolates the covariance matrix Σ_X at a certain node X with the covariance matrix of the parent node $\tilde{X}$ resulting in an interpolated covariance matrix $\hat{\Sigma}_X$:

$$\hat{\Sigma}_X = \lambda \Sigma_X + (1 - \lambda)\Sigma_{\tilde{X}}$$

The interpolation factor λ is calculated by a sigmoid function:

$$\lambda = \frac{1}{1 + \exp(-5(N_x/\delta - 1))}$$

where N_x is the number of observations at node X and δ is the "smoothness" parameter of the sigmoid function. For very small N_x it yields a λ which is approximately zero, and the number of observations where λ is $1/2$ equals δ. For the experiments we used a δ of 500.

5. RESULTS

The system which was used to obtain the results is described in [1, 3]. The most important properties are:

- 30 filter bank outputs together with first and second order derivatives resulting in a 63-component acoustic vector,

- feature reduction down to 35 components by LDA [4],
- 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.

For *Wall Street Journal* the training was done on the WSJ0 training corpus and the tests were performed on the WSJ November 92 test set. For *VERBMOBIL* we used the VERBMOBIL '95 training corpus and the VERBMOBIL '95 test corpus for testing.

Table 1 shows the results for state tying on the WSJ 5 000 word test corpus. State tying with the bottom up method and the log-likelihood measure improves the word error rate (WER) from 8.8% to 8.1% for the log-likelihood measure and to 8.6% for the approximative divergence measure compared to the untied models.

The best result for state tying with decision trees achieved an additional relative improvement of 6% over the bottom up method. This result was obtained by using all triphones for tree construction ('*one observation*'). Triphone subsets ('*50 observations*', '*20 observations*', '*intersection*') for tree construction performed slightly worse. The conclusion is that it is advantageous to use as many triphone states as possible to select the questions at the tree nodes.

The simple cross validation method we have tested reduced the error rate for the triphone list *50 observations* from 8.4% to 7.9%. This result shows the potential improvement that can be achieved by cross validation. The problem with this simple method seems to be the small number of triphone states used for the selection of the questions.

Using only one tree with additional questions on the central phoneme and the state did not improve the error rate. The advantage of this method is its higher flexibility. Since the minimum number of models is equal to 1, trees of any size can be constructed. These trees can then be used in various methods such as state tying (as described here) or speaker adaptation.

The *merge-to-combine* method leads to approximately the same error rate as the baseline method. The *merge-to-reduce* tests show that the number of mixtures can be halved by merging while the error rate increases only by 10% relative.

The plain coupling of two gender dependent models in the tree nodes increases the error rate to approximately 8.0%. By an additional smoothing of the variance vectors the results of the baseline method can also be achieved.

For the tests with a full covariance matrix we first used the triphone list *20 observations*, which results in an error rate of 8.0%. Additional smoothing and the full triphone list reduced this error rate to 7.7%. Compared to the baseline method which employs a variance vector per model, this more accurate method did not improve the recognition results. We think that the reason is the LDA transformation of the feature vector. Because an LDA transformation roughly decorrelates the class dependent covariance matrix, the usage of a variance vector as an approximation for the covariance matrix seems to work quite well.

Table 2 shows the results for decision tree based state tying on the VERBMOBIL corpus. The corpus is subdivided into the *short95* corpus (short sentences) and the *long95* corpus (long sentences). Here the gain due to state tying is lower than on the WSJ task. One possible reason is that the context dependency of phones in the German language is not as high as in the English language. A second possible reason is that the VERBMOBIL corpus contains spontaneous speech while the WSJ corpus contains read speech. Thus we think that by adding across word models to our recognizer, the overall improvement will be comparable for both corpora.

REFERENCES

[1] X. Aubert, C. Dugast, H. Ney, V. Steinbiss, "Large Vocabulary, Continuous Speech Recognition of Wall Street Journal Corpus," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Adelaide, Australia, Vol. II, pp. 129-132, April 1994.

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

[3] 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.

[4] 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.

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

[6] H. Ney, "Acoustic Modelling of Phoneme Units for Continuous Speech Recognition," *Proc. Fifth Europ. Signal Processing Conf.*, Barcelona, pp. 65-72, September 1990.

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

[8] R. Schwartz, Y. Chow, O. Kimball, S. Roucos, U. Krasner, J. Makhoul, "Context-Dependent Modelling for Acoustic-Phonetic Recognition of Continuous Speech," *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Tampa, FL, pp. 1205-1208, March/April 1985.

[9] S.J. Young, P.C. Woodland, "The Use of State Tying in Continuous Speech Recognition," *Proc. Europ. Conf. on Speech Communication and Technology*, Berlin, pp. 2203-2206, September 1993.

[10] 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.