DECISION TREE STATE TYING BASED ON SEGMENTAL CLUSTERING FOR ACOUSTIC MODELING

W. Reichl and W. Chou

Bell Laboratories, 600 Mountain Ave.,

Murray Hill, NJ 07974, USA

ABSTRACT

In this paper, a fast segmental clustering approach to decision tree tying based acoustic modeling is proposed for large vocabulary speech recognition. It is based on a two level clustering scheme for robust decision tree state clustering. This approach extends the conventional segmental K-means approach to phonetic decision tree state tying based acoustic modeling. It achieves high recognition performances while reducing the model training time from days to hours comparing to the approaches based on Baum-Welch training. Experimental results on standard Resource Management and Wall Street Journal tasks are presented which demonstrate the robustness and efficacy of this approach.

1. INTRODUCTION

One phenomenon in large vocabulary speech recognition is the amount of training data. It introduces a wide range of acoustic variations caused by different speakers and various channel conditions. Decision tree state clustering based acoustic modeling has become increasingly popular for modeling speech spectra variations in large vocabulary speech recognition using Hidden Markov Models (HMMs) [1,2]. In this paper, we discuss methods for improving the robustness and accuracy in decision tree clustering based acoustic modeling, and a fast segmental clustering based approach is described and compared with the standard method. Although we will concentrate on the triphone based acoustic modeling, the approach described in this paper extends straightforwardly to systems using greater degrees of context-dependencies.

In decision tree clustering based acoustic modeling, each node of the decision tree is attached to a question regarding the phonetic context of the triphone units. A set of states can be recursively partitioned into subsets according to the phonetic questions at each tree node when traversing the tree from the root to its leaves. States reaching the same leaf node on the decision tree are regarded as similar and tied. Even in large vocabulary speech recognition, many triphones have only a very few occurrences in the training data, and there are not sufficient data for a robust parameter estimation of these rarely seen triphones. Moreover, it is typical that a significant number of triphones are missing or unseen in the training data. This problem becomes more acute when using cross-word context dependencies, because nearly all possible context combinations are needed during decoding. The missing triphones have to be constructed by using acoustic similar states. The advantage of using decision tree in acoustic modeling is two folds. First, it can lead to compact, high quality state clusters for robust estimation of mixture Gaussian distributions. Secondly, it makes it possible to synthesize unseen triphones by using the acoustic similar states according to the phonetic decision tree.

The standard method [1,2] in decision tree tying based acoustic modeling is based on a Baum-Welch sequential alignment process to estimate parameters of the mixture Gaussian distribution for each state cluster at the leaf node of the decision tree. Although this process can be made parallel on different processors, the data alignment used in model building becomes inconsistent to the data alignment of Viterbi decoding during recognition. In the standard method, this mismatch in data alignment is further compounded by the use of a crude single Gaussian unclustered triphone system in Baum-Welch sequential alignment during training. Another key issue related to decision tree tying based acoustic modeling is the quality of the state clusters for robust parameter estimation. Construction of a globally optimal decision tree is a computationally intractable problem. In practice, a single Gaussian, unclustered triphone system is built first, and training data are segmented into states corresponding to this single Gaussian unclustered system. In order to make use of rarely seen triphones, all seen triphone samples are directly used in the standard method to construct the single Gaussian unclustered triphone system. It is often that only the mean vectors are estimated from data, whereas the

variance vectors are smoothed with the mono-phone models. During the decision tree clustering process, each node of the decision tree is represented by a single Gaussian distribution. The likelihood of the decision tree node on the training data can be derived from the associated single Gaussian unclustered states without touching the training data. In this approach, the estimation error introduced to the single Gaussian unclustered states will have a long term impact on the quality of the decision tree based state clustering.

The segmental clustering based approach described in this paper is based on an Viterbi alignment of training data and is an extension of the conventional segmental K-means approach to decision tree tying based acoustic modeling. It utilizes a two level clustering scheme to improve the robustness of the model estimation. The high recognition performance can be achieved while reducing the model training time from days to hours comparing to the standard method.

2. A FAST SEGMENTAL CLUSTERING APPROACH

Alignment of training speech data for untied states in decision tree based acoustic modeling is a critical issue to the quality of the decision tree state tying based acoustic modeling. In our segmental clustering based approach, the segmentation of the training data is separated from the model parameter estimation process and high resolution multi-mixture Gaussian models are used to provide high quality data alignment in decision tree tying. The training data are aligned through a Viterbi alignment process. The segmentation is according to the best state sequence. Unlike in the standard method, model refinement and parameter estimation are based on a fixed decision tree from the single Gaussian untied system, the decision tree in our approach is part of the iterative training process. It is updated and re-estimated in each iteration during the training of the acoustic models. In each iteration, the training data is re-segmented using the decision tree state tied model generated from the previous iteration. The convergence property of the segmental K-means approach ensures that training data alignment will improve and converge with this iteration process. The high quality training data alignment will lead to more precise estimate of the likelihood variations during the decision tree construction and improve the quality of the decision tree state tying based acoustic modeling.

However, one of the issues in using Viterbi alignment in decision tree state tying based acoustic modeling is how to make robust use of these rarely seen triphone samples in the training data. In Baum-Welch based parameter estimation, all possible paths are considered, and it has a much stronger smoothing effect on the parameters of these rarely seen triphones with very few training samples. In Viterbi alignment based segmental clustering approach, only the best path is considered and parameters of these rarely seen triphones can degenerate very quickly with the decrease of the training samples. In order to make a full use of the training data and improve the robustness of the decision tree based state tying, a two level segmental clustering scheme is used in our approach. The first level segmental clustering is performed before forming the single Gaussian untied system. It is to cluster those rarely seen triphones into various generalized triphones according to their phonetic similarities so that each of the clustered generalized triphones has samples above the minimum sample count threshold for forming single Gaussian states. A single Gaussian untied system is built based on single Gaussian triphones and generalized triphons from clustering phonetically similar rare occurrence triphones. The second level clustering is a decision tree based clustering similar to the standard method. The phonetic identity of each generalized triphone from the first level clustering is defined to be the intersects of the phonetic properties of all rare triphones in the cluster.

One way to group these rare triphones is by relaxing the triphone context [3]. First the left contexts of the rare triphones are skipped and we try to find enough right context dependent bi-phones to build an acoustic model. If there are not sufficient examples in the training data, we skip the right context of the rare triphones and group the tokens to left context dependent bi-phones. Still some very rare context combinations do not have enough examples and may be used to train back-off mono-phones. The two level clustering approach described above takes the advantage of generalized triphone at the stage of forming a robust single Gaussian untied system to improve the quality of the decision tree. The final model is still decision tree tied in which the node splitting is determined solely by the likelihood increase on the training data. In addition, the unseen triphones are always synthesized according to the decision tree without making reference to the generalized triphones. This is very different from the generalized triphone defined in [3] where state tying is determined purely by the context.

In our current approach, for each state of each base phone one decision tree is constructed to cluster all the context dependent states of this phone. The tree uses questions about the phonetic context to find the best set of tied states, which maximize the likelihood of the training data and

have sufficient acoustic data associated. The splitting of sets of states in the tree nodes is controlled by the increase in likelihood for the associated phonetic questions and terminated by a likelihood threshold and a minimum occupation count. The leaves of the decision tree determine the set of context specific states for each phone to be tied together. The log likelihood for a set S of single mixture states sharing one common Gaussian distribution with mean $\mu(S)$ and covariance matrix $\Sigma(S)$ using the segmental clustering based decision tree tying is given by

$$L(S) = \sum_{x_t : s_t \in S} \log\left(p\left(x_t | \mu(S), \Sigma(S)\right)\right)$$

where all frames x_t with a state alignment $s_t \in S$ are considered and we assume the state alignment is not changed by tying the states. For single Gaussian distributions the log likelihood is

$$L(S) = -\frac{n(S)}{2}\left(\log\left|\Sigma(S)\right| + D \cdot \left(\log(2\pi) + 1\right)\right)$$

where $n(S)$ is the number of frames assigned to S and D is the dimensionality of the data vector. This log likelihood can be calculated for every set of by using the already available information from the segmental K-means algorithm without additional accessing the training data. A multi-mixture Gaussian distribution is estimated directly for each tied state in our algorithm. This differs from the Baum-Welch based approach, where multi-mixture distributions are obtained by iterated binary splitting of each Gaussian density function.

Our acoustic model training is based on the segmental K-means algorithm and therefore the estimation of the model parameters is independent for each state and can be effectively parallelized on multiple CPUs. Since all of the time intensive calculations requiring the processing of training data is distributed to multiple CPUs the required time for one segmentation can be reduced drastically even for large data sets.

3. EXPERIMENTAL RESULTS

The performance of the proposed segmental clustering approach to phonetic decision tree tying based acoustic modeling was evaluated on the Resource Management (RM) and the Wall Street Journal (WSJ) tasks. For both databases 12 mel-cepstral coefficients and the normalized energy plus their 1st and 2nd order time derivatives were used as acoustic features. The cepstral mean for each sentence was calculated and removed. All phone models have three emitting states and a left-to-right topology. Training of the acoustic parameters was based on the proposed segmental clustering decision tree tying algorithm. First the parameters of all models with a number of examples exceeding a threshold were estimated. We used a minimum threshold of 10 examples in our experiments. The rare triphones were grouped by skipping first the left context and then the right context. The phonetic decision tree tying was used to cluster equivalent sets of context dependent states and to construct unseen triphones. The final models were built by using the segmental K-means algorithm to estimate the parameters for the tied states. The number of mixtures for each tied state varies from 4 to 12. All systems were gender-independent and used cross-word triphone models. Decoding was done using a one-pass N-gram decoder [5], in which the search was conducted on a layered self-adjusting decoding graph.

The standard SI-109 training data set was used in the RM system. The CMU pronunciations and phone set (47 phones) were used to build the phonetic lexicon. Decoding was based on the standard word-pair grammar. 3500 triphones occur at least 10 times in the training data, and 3000 triphones with a very low frequency count were grouped to 630 triphone clusters. The total number of Gaussian distributions was about 12000. The RM system was tested on the official evaluation test sets (FEB89, OCT89, FEB91, SEP92) using the 991 word vocabulary. The word error rates (100% - word accuracy) for a system with 1911 states and an average of 6.3 mixtures per state are listed in Table 1 for different test sets.

FEB89	**OCT89**	**FEB91**	**SEP92**	**FEB89-SEP92**
2.4%	3.6%	2.3%	6.2%	3.6%

Table 1: Word error rates for the RM task.

The average word error rate of 3.6 % for the RM evaluation tests is one of the best reported results for this task and shows the high performance of the proposed segmental clustering algorithm for acoustic training.

For the WSJ systems, the SI-84 and the SI-284 training data sets were used. The lexicon was generated automatically using a general English text-to-speech system (41 phones) [6]. The language models used in the experiments are the standard bigram and trigram language models provided in the WSJ corpus. The SI-84 training data (7200 sentences) contains about 8600 triphones with more than 10 examples and about 8000 triphones with a frequency count of less than 10. The SI-84 trained models

consist of 3447 tied states with a total of about 37000 Gaussian distributions. The average number of mixtures per state is 10.9. The acoustic models for the WSJ systems are all gender-independent. The evaluations of the WSJ systems were performed on the official NOV92 (si_et_05, si_et_20) and NOV93 (si_et_h1) test sets for the closed 5K and open 20K vocabulary. The results are obtained based on a one-pass frame synchronous decoding without adaptation. The word error rates for the NOV92 evaluation of the WSJ system trained on the SI-84 training data are listed in Table 2.

Model (SI-84)	**NOV92**	
	5k-closed	20k-open
bigram	6.8 %	14.7 %
trigram	5.0 %	13.0 %

Table 2: Word error rates for NOV92 WSJ evaluation.

In the next experiment the full WSJ dataset (SI-284, 38700 sentences) was used in the training of the acoustic models. About 10000 of the 24000 observed triphones occur less than 10 times in the training data. These are grouped into 1029 triphone clusters to ensure the estimated parameters for the state clustering are robust. After the phonetic decision tree clustering 6804 tied states with about 87000 Gaussian distributions and an average of 12.8 mixtures per state were calculated. The results for the NOV92 and NOV93 tests are listed in Table 3.

Model (SI-284)	**NOV92**		**NOV93**
	5k-closed	20k-open	20k-open
bigram	5.4 %	11.9 %	15.4 %
trigram	3.3 %	10.4 %	14.0 %

Table 3: Word error rates for NOV92 and NOV93 evaluation of the WSJ task .

The results of the gender independent system in the 5k vocabulary NOV92 evaluation (column 1 in Table 2) are very close to the best reported word error rates for gender-dependent models [7]. The error rates for the 20k open vocabulary evaluations are between 10.6 % for NOV92 and 14.1 % for the NOV93 H1-C1 test, while the 1.8 % out-of-vocabulary words make a significant contribution to the word errors in this open vocabulary test.

Since the training of the acoustic models is based on the segmental K-means algorithm most of the calculations required can be performed in parallel on different CPUs or computers. Even the phonetic tying for the individual context dependent states can be separated in independent processes. The required time for one training iteration, including the segmentation of the acoustic data, on 6 Pentium-Pro processors for the RM task is about 1.5h and for the WSJ SI-84 dataset less than 3h. This is a reduction in training time from days to hours.

4. SUMMARY

In this paper, a fast segmental clustering approach to decision tree tying based acoustic modeling is proposed for large vocabulary speech recognition. It is based on a two level clustering scheme for robust decision tree based state clustering. This approach was tested on both RM and WSJ tasks. The very low error rates were based on gender independent models and obtained from an one pass decoding without adaptation. These experiments illustrate the robustness and efficacy of the proposed approach. In addition, this algorithm is extremely efficient and the model training time is reduced from days to hours, a much desired feature for large vocabulary speech recognition.

REFERENCES

1. L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahamoo, and M.A. Picheny, “Decision Trees for Phonological Rules in Continuous Speech”, ICASSP 89, Glasgow, May 1989.
2. S.J. Young, J.J. Odell, and P.C. Woodland, “Tree Based State Tying for High Accuracy Modeling”, ARPA Workshop on Human Language Technology, Princeton, NJ, Morgan Kaufmann Publishers, March 1994.
3. C-H. Lee, E. Giachin, L.R. Rabiner, R. Pieraccini, and A.E. Rosenberg, “Improved acoustic modeling for large vocabulary speech recognition”, Computer Speech and Language, Vol. 4, No. 2, pp. 103-127, 1992.
4. L.R. Rabiner, J.G. Wilpon, and B.-H. Juang, “A segmental k-means training procedure for connected word recognition” AT&T Tech. Journal, 65, pp. 21-31, 1986.
5. Q. Zhou, and W. Chou, “An Approach to Continuous Speech Recognition Based on Layered Self-Adjusting Decoding Graph”, ICASSP 97, Munich, April 1997.
6. R.W. Sproat and J.P. Olive, “Text-to-Speech Synthesis”, AT&T Tech. Journal, 74, pp. 35-44, 1995.
7. P.C. Woodland, J.J. Odell, V.Valtchev and S.J. Young, “Large Vocabulary Continuous Speech Recognition Using HTK”, ICASSP 1994, Adelaide, April 1994.