APPLICATION OF CLUSTERING TECHNIQUES TO MIXTURE DENSITY MODELLING FOR CONTINUOUS-SPEECH RECOGNITION

Christian Dugast, Peter Beyerlein, Reinhold Haeb–Umbach

Philips Research Laboratories, P.O. Box 1980, D-52021 Aachen, Germany
E-mail: {dugast,beyerlei,haeb}@pfa.philips.de

ABSTRACT

Clustering techniques have been integrated at different levels into the training procedure of a continuous-density hidden Markov model (HMM) speech recognizer. These clustering techniques can be used in two ways. First acoustically similar states are tied together. It will help to reduce the number of parameters but also allow to train otherwise rarely seen states together with more robust ones (state-tying). Secondly densities are clustered across states, this reduces the number of densities while at the same time keeping the best performances of our recognizer (density-clustering). We have applied these techniques both to word-based small-vocabulary and phoneme-based large-vocabulary recognition tasks. On the WSJ task, we could achieve a reduction of the word error rate by 7%. On the TI/NIST-connected digit task, the number of parameters was reduced by a factor 2-3 while keeping the same string error rate.

1. INTRODUCTION

Clustering techniques have been integrated at different levels into the acoustic-phonetic training procedure of a continuous-density hidden Markov model (HMM) speech recognizer. The main idea of clustering is to concentrate what is acoustically similar. For a continuous-density HMM system, acoustic similarity can be seen at different levels: At the phoneme level (triphone), at the state (or mixture) level and at the density level.

Clustering at the first two levels (phoneme and state) leads to symbol tying. It answers the question "Which triphones are acoustically similar?" and will help us to define a reduced set of models to be trained. It should give us the possibility to avoid the duplication of models, and therefore reduce the number of parameters of our system. Furthermore it can make more efficient use of training material, for example, while training rarely seen states together with more robust ones. Clustering at this level is also known in the literature as tying. Following the work at CMU [3] and at CUED [2], we decided to concentrate on state-tying rather than triphone tying.

Clustering at the density level reduces the number of densities and at the same time keeps the best performances of our recognizer (density-clustering). Density-clustering is done across symbols and is independent of the previously mentioned tying. We have applied these techniques both to word-based small-vocabulary and phoneme-based large-vocabulary continuous-speech recognition.

The following results have been achieved. Compared to our base models (without state-tying) which gave state-of-the-art results at the Wall Street Journal (WSJ) benchmarking test in November '93 [1], on a reduced training set (WSJ0 with 15 hours of speech material) we can multiply the number of trained triphones by more than a factor of two by using state-tying techniques. With this increased triphone coverage on the test set the word error rate was reduced globally over three different test sets by more than 6%, while at the same time the number of parameters was reduced by 30%. The method has been extended to the WSJ0+1 training set (totalizing 62 hours of speech) to augment the triphone coverage of the test sets now up to 99.6% (with 3 times more triphones). This leads to a decrease of the WER by more than 7% relative to the November 1993 system without increasing the number of parameters of our system.

For word model based small-vocabulary speech recognition, state-tying identifies acoustically similar states within different words. This results in deciding automatically which parts of speech of the recognition vocabulary are similar and therefore can be modelled together. It will avoid duplication of models and thus reduce the number of parameters.

At the density level, the clustering technique allows a complexity reduction and robustness increase both for word-based and phoneme-based systems. For the WSJ large-vocabulary recognition task, on the reduced WSJ0 training set, the total number of densities could be reduced by another 20% with only a very slight increase of the word error rate.

On small-vocabulary recognition tasks, such as the TI/NIST Connected Digits recognition task, a combi-

0-7803-2431-5/95 $4.00 © 1995 IEEE

nation of state and density-clustering led to a considerable complexity reduction: The number of model parameters could be reduced by a factor of two compared to the non-tied system.

2. STATE-TYING

For large-vocabulary continuous-speech recognition, triphones have been identified to be adequate to model co-articulation effects. However, in any realistic training set some triphones will occur very rarely, some even never. There are several methods to approach this problem. One method is to model all triphones present in the training set separately, and then to apply smoothing techniques to overcome the problem of sparse training data. Another method is to pool rarely seen triphones in a monophone model which serves as a backing-off model. The latter has been integrated in our system. The minimum number of observations above which a separate triphone is modelled has been set to 75. It is clear that within the set of chosen triphones, some are acoustically very similar to others, which leads to duplication of models.

Our state-tying technique is very similar to [2]. We have compared different clustering criteria: A furthest-neighbor criterion (1) applied directly to the spectral mean vectors and a maximum-likelihood criterion (2), which combines the spectral distance and the observation counts in one criterion. The furthest-neighbor does not quantify the goodness of a rarely seen model, it takes the spectral mean vectors as they have been observed. The distance measure

$$d(C_i, C_j) = \max_{m_k \in C_i, n_l \in C_j} \left[\sum_c (|m_{k_c} - n_{l_c}|) \right] \qquad (1)$$

calculates the distance between two clusters C_i and C_j, where each cluster is defined by a set of mean vectors m_k and n_l. Two clusters C_i and C_j are clustered together if their distance lies below a certain threshold. The new cluster will be the union of the original clusters. Surprisingly, the maximum likelihood criterion (2) used for density-clustering has not given the best results. It might indicate that the spectral information is enough to decide whether two models may be tied together or not.

Table 1 presents results for different initial numbers of states and different thresholds on the maximum diameter of a cluster ('cluster threshold') using the furthest-neighbor criterion. The training set used for these experiments is the so-called WSJ0 training set summing up to around 15 hours of speech equally balanced between male and female speakers. Our base system (WSJ0_a) is the non-tied system optimized for the 5000 word vocabulary WSJ benchmarking test from November 1993 [1]. The optimal number of triphone-states for our base system had been set to 2208 + 130 triphone and monophone states, respectively. As can be seen from the first two lines of Table 1, applying state-tying led to a reduction of the number of densities by a factor of two without changing the total word error rate over three test sets. These three test sets consist in total of 20774 pronounced words.

Table 1: Bigram word error rates on different WSJ test sets with and without state-tying. Training was done on WSJ0.

No. of states before tying	Cluster threshold	No. of states after tying	No. of densities (Male + Female)	5k test set (*si_evl5, si_dev5, si_dt_05*)	Test set triph. coverage
2208	0	2208	245k	11.90%	75%
2208	16	1336	115k	11.92%	
5565	0	5565	225k	12.02%	90%
5565	16	2435	163k	11.12%	
23508 (all triph.)	15	4621	235k	12.20%	99.7%

State-tying allows to group together triphones which are acoustically similar but not necessarily often seen. The consequence is that more triphones can be modeled: The triphone coverage of the test set lexicon will be higher. We increased the number of triphones to be modelled and found an optimum at 1855 triphones (1855*3 = 5565 states), which makes our second system (WSJ0_b, lines 3 and 4 from Table 1). This leads to word error rate improvements of more than 6% on the same set of test sets, when compared to the WSJ0_a system.

As stated above, the optimal number of triphones modelled on the WSJ0 material was 1855. We give hereafter results while modelling all 7836 triphones seen during training. A drawback in modelling all triphones present in a training set is that there is no observation left to model backing-off monophones. During recognition, a decision has to be taken: To which trained model will be assigned the untrained but essential new triphone? Decision trees are often used at this place.

Our solution was very pragmatic: We took from our WSJ0_b system the monophone backing-off models, properly rescaled and added them to our all-triphone system. As can be seen from Table 1, the word error rate increased significantly with respect to our WSJ0_b system and goes back to the WER level of our WSJ0_a system. To interpret the result, it has to be observed that from the 7836 different triphones occurring in the

training, 3781 occur less than 10 times. Our conclusion is that under a certain occurrence threshold (that is 35) state-tying results in a splitting of rarely seen training material (that would otherwise be globally modelled in a monophone) and leads to less robust modelling.

Table 2: Bigram word error rates on the evaluation test set 93 with and without state-tying for the same number of densities. Training was done on WSJ0+1.

Init. # States	5592	18276
Cluster threshold	0	16
# States after Tying	5722	4166
# Densities(Male+Female)	523k	495k
20k *eval_93*	17.7%	16.4%
Test Set Triphone Coverage	90%	99.6%

The next step was to build our models on the bigger WSJ0+1 training set totalizing 62 hours of speech. To augment the triphone coverage on the recognition vocabulary, we included right context diphones to our triphone list (4087 triphones and 557 diphones). The backing-off monophone models were seen on average 350 times. Table 2 shows that by state-tying an improvement in the WER by more than 7% for about the same amount of densities has been achieved on the evaluation set of November 1993. This is mostly due to the triphone coverage ratio increase on the test set, from 90% to 99.6%.

3. DENSITY-CLUSTERING

HMM states may share some or all component densities of their mixture densities, if they model acoustically similar events. As a result of state-tying (see section 2), complete models will be tied together. Density-clustering on the other hand allows two different models to share common regions of the acoustic space (see Fig. 1). It is done across HMM states and is independent of the previously mentioned state-tying.

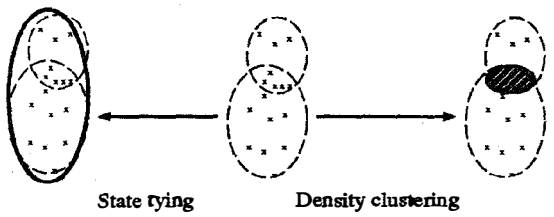

Figure 1: State-tying versus density-clustering.

Nevertheless, the motivation for state-tying or density-clustering is basically the same: To reduce the amount of parameters required to model the recognition vocabulary and thus the computing and memory demands to make the recognition fit on cheap hardware.

Actually, in the implementation there is no difference between state-tying and density-clustering; the same algorithm is applied either to single-density models (state clustering) or to the pool of mixture densities (density clustering). For density-clustering, we used an agglomerative clustering procedure together with a maximum likelihood criterion, which is given by

$$d(C_1, C_2) = \frac{c_1 \cdot c_2}{c_1 + c_2} \cdot ||m_1 - m_2||, \qquad (2)$$

where c_i are the observation counts and m_i the mean-vectors of the density-cluster $C_i, i = 1, 2$.
At the beginning of the cluster-algorithm each cluster is equal to one density. If the two clusters C_i, C_j have a minimum distance $d(C_i, C_j)$,they are merged to a new cluster C_N by adding the counts of the clusters $c_N = c_i + c_j$ and computing a new weighted mean-vector $m_N = \frac{c_i \cdot m_i + c_j \cdot m_j}{c_i + c_j}$. The procedure stops, if the desired number of clusters is obtained.

From the obtained clusters a tied-density inventory is derived. The clustering operation is part of the HMM training process and can be included anywhere and even several times within the iterative re-estimation of the model parameters.

3.1. Small-Vocabulary Speech Recognition

Here, small-vocabulary speech recognition is used as a synonym for an acoustic modeling approach which employs hidden Markov models of words rather than of phonemes. The goal of density-clustering is to identify similar acoustic events across different word models, hence, the clustering algorithm works on the whole acoustic space.

Table 3 shows the effects of clustering on the number of parameters and on the error rate for experiments on the adult speakers' portion of the TI/NIST Connected Digits Recognition Task. The different degrees of tying have been obtained by successive density splitting and clustering steps during training. For details on the non-tied system, see [4]. In table 3 experiments with similar string error rate are grouped together. It can be seen that the number of model parameters could be reduced by a factor of 2 – 3 without increase in error rate. For a medium error rate performance range (1.3% - 3% string error rate), the results can be stated alternately: given the same number of parameters, the tied system achieves a 30% better error rate.

Table 3: String error rate (SER) on TI Digits for various configurations.

SER [%]	configuration
3.37	0.6k non-tied single densities 19200 parameters
2.97	0.3k tied densities, 1k weights 10600 parameters
2.59	1.2k non-tied densities 39600 parameters
2.67	0.3k tied densities, 1.6k weights 11200 parameters
1.91	2.4k non-tied densities 79200 parameters
1.90	0.8k tied densities, 3.2k weights 28800 parameters
1.45	4.8k non-tied densities 158400 parameters
1.30	2k tied densities, 10k weights 74000 parameters
1.16	9.5k non-tied densities 304000 parameters
1.14	5k tied densities, 18.5k weights 178000 parameters
0.95	19k non-tied densities 608000 parameters

3.2. Large-Vocabulary Speech Recognition

Our large-vocabulary recognition system has a large number of densities (more than 80k for each gender). State-tying has been done before training to decide on which model to train. Density-clustering will be done at the end of the training to avoid duplicate modelling of shared acoustic spaces. To manage the large number of densities of our WSJ end-system and to keep its discriminative capability, only densities corresponding to the same context-independent phoneme are clustered. Table 4 presents results on the male-speaker part of the Wallstreet Journal task where the number of densities of our tied-state WSJ0_b system of section 2 has been reduced by another 20% (from 80k to 65k) with only a slight increase of the word error rate. In total, combining state-tying and density-clustering, the number of densities was reduced from 122k to 65k while at the same time the word error rate was slightly reduced from 12.53% to 12.07%.

Table 4: Word error [%] rates without and with density-clustering on different test sets for a male specific tied-state system.

Initial # States	2208	5565	5565
Initial Cluster thresh.	0	16	16
# States after Tying	2208	2435	2435
Density-Clustering	NO	NO	YES
# Densities *(male only)*	122k	80k	65k
5k*male* *(si_evl5,si_dev5,si_dt_05)*	12.53	12.02	12.07

4. SUMMARY

State-tying allows to avoid the duplication of acoustically similar models. A consequence is that rarely seen acoustic events can be modelled together with more robust ones. We have observed that very rare events (seen less than 35 times) will weaken the models they are tied to. Furthermore, it has been noticed that for state-tying, the furthest-neighbor criterion was superior to a maximum-likelihood criterion. Density-clustering based on a maximum-likelihood criterion allows to better model the part of the acoustic space that is shared by two different models. A combination of the two clustering techniques leads to a reduction of the number of parameters by a factor of up to two and to a significant error rate reduction on two tasks, the TI/NIST-connected digit and the WSJ tasks.

5. REFERENCES

[1] X. Aubert, C. Dugast, H. Ney, V. Steinbiss. Large Vocabulary Continuous Speech Recognition of Wall Street Journal Corpus. In *Proc. ICASSP '94*, Vol. II, pp. 129-132, Adelaide, Australia, April 1994.

[2] S.J. Young, P. C. Woodland. The Use of State-Tying in Continuous Speech Recognition. In *Proc. EUROSPEECH '93*, Vol. 3, pp. 2203-2206, Berlin, Germany, Sept. 1993.

[3] M. Hwang, X Huang. Shared-Distribution Hidden Markov Models for Speech Recognition. In *Trans. on Speech and Audio Processing*, vol 1, No. 4, pages 414-420, Oct. 1993.

[4] R. Haeb-Umbach, D. Geller, H. Ney. Improvements in Connected Digit Recognition Using Linear Discriminant Analysis and Mixture Densities. In *Proc. ICASSP '93*, pages 239-292, Minneapolis, MN, April 1993.