EXPERIMENTING WITH A GLOBAL DECISION TREE FOR STATE CLUSTERING IN AUTOMATIC SPEECH RECOGNITION SYSTEMS

Jasha Droppo and Alex Acero

Speech Technology Group
Microsoft Research
jdroppo,alexac@microsoft.com

ABSTRACT

In modern automatic speech recognition systems, it is standard practice to cluster several logical hidden Markov model states into one physical, clustered state. Typically, the clustering is done such that logical states from different phones or different states can not share the same clustered state. In this paper, we present a collection of experiments that lift this restriction. The results show that, for Aurora 2 and Aurora 3, much smaller models perform as least as well as the standard baseline. On a TIMIT phone recognition task, we analyze the tying structures introduced, and discuss the implications for building better acoustic models.

Index Terms— automatic speech recognition, acoustic modeling, phonetic decision tree

1. INTRODUCTION

The basic unit of modern acoustic models for large vocabulary computer speech recognition (LVCSR) systems is the 3-state triphone hidden Markov model (HMM). To represent all possible sequences of phones, there should be one HMM in the acoustic model for every possible center phone label, in every possible left context and right context. Since a typical system may model 44 different phonemes, this means a complete acoustic model would contain a total of over 255 thousand HMM states in 85 thousand triphone models.

Even today, there is not enough transcribed acoustic data to properly train an acoustic model with 255 thousand HMM states. Tree-based state tying [1] was invented to address this issue, and is still in use today. We call the theoretical HMM states that should exist in a complete acoustic model *logical* states, and the concrete shared states that they refer to as *clustered* states.

Typically, the logical states that share the same clustered state all belong to the same state of the same center phone. A clustered state associated with state 2 from one HMM will never be associated with a different state of the same or any other HMM. A clustered state associated with the center phone “ah” will ever be associated with any other center phone. This design feature is based on the assumption that there is no benefit from clustering different phones or different states together, although the literature contains mixed results [2, 3, 4].

This paper explores the result of using conventional tree-based state tying, but allowing cross-center phone and cross-state clustering. The result of using this global decision tree is an acoustic model that better describes the acoustics of the training data, without artificially partitioning the acoustic space. Where a similar system was constructed in [4] to address sloppy conversational speech, the current experiments demonstrate how it can be more generally useful.

This paper is organized into four sections. In Section 2, we discuss the clustering method used for this paper in more detail. Section 3 demonstrates the system behavior on three different tasks: Aurora 2, Aurora 3, and TIMIT. Aurora 2 and Aurora 3 are both connected digit recognition tasks, and cover five different languages. In all five languages, the global decision tree automatically identifies and clusters similar sounds. The resulting models are much smaller, more efficient, and no less accurate. Results on the TIMIT phone decoding task are also presented, including an analysis of the resulting cross-phone and cross-state tying that the tree learns from the data. Finally, our conclusions are discussed in Section 4.

2. METHOD

The standard way of clustering logical HMM states together is to use a set of phonetic decision trees. One tree is built for every state of every center phone. Each of the trees starts with all possible phonetic contexts represented in a root node. Then, a binary question is chosen that best splits the logical states represented by the node into two child nodes. Whichever question creates two new clustered states that maximally increase the log likelihood of the training data is chosen. This process is applied recursively until the log likelihood increase is less than a threshold, at which point a final agglomerative clustering step is performed. The choice of threshold is important, because it directly affects the depth of the tree, and therefore the final size of the acoustic model.

In this work, we do not use a different decision tree for every context independent phone state. Instead, we use a single phonetic decision tree that starts with all logical states in the root node. For each task, we augment the question sets with the extra questions needed to separate all of the logical states for the task. In this way, although the training algorithm is able to cluster the states in the standard way, it is not compelled to.

Another difference between the results in this paper and the usual method is the amount of look-ahead. Usually, each question is evaluated by looking at the likelihood increase induced by its direct children. In this paper, we use a two-step lookahead so that each question is evaluated by calculating the likelihood increase if we use that question, and then make the best possible decision for the next two levels of the tree. As a result, adjacent levels of the tree are able to cooperate and find a slightly better solution.

3. RESULTS

Here we present an analysis of the method on three different standard recognition tasks.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE

	clean train, set A test		
States	Clean	SNR10	Average
180	99.63	72.57	63.28
178	99.65	72.21	62.86
169	99.62	71.31	62.39
159	99.67	71.51	63.40
142	99.63	70.85	62.61
128	99.64	70.79	62.94
109	99.55	68.00	60.68
94	99.54	69.71	63.64

Table 1. *Word accuracy of Aurora 2 system trained on clean data. As the model size decreases, word accuracy drops. Models with much fewer states than the baseline perform almost as well.*

	multi train, set A test		
States	Clean	SNR10	Average
180	99.48	91.96	97.24
175	99.44	91.96	97.27
168	99.49	91.93	97.14
161	99.46	91.90	97.16
148	99.38	91.86	97.07
134	99.31	91.86	97.12
119	99.35	91.67	96.95
105	99.22	91.38	96.65
80	99.13	90.82	96.14
60	98.57	90.07	95.67

Table 2. *Word accuracy of Aurora 2 system trained on multi-style data. Accuracy and model size exhibit the same pattern seen with clean training data.*

3.1. Aurora 2 Results

The Aurora 2 task [5] consists of recognizing strings of English digits embedded in a range of artificial noise conditions. The standard Aurora 2 acoustic model contains eleven sixteen-state whole word HMMs, a three-state silence model, and a one-state short pause model.

This model topology is easy to generate, but likely suboptimal. For instance, the initial sounds of the word pairs (six, seven) and (four, five), although acoustically similar, are modeled separately. Furthermore, although the task contains both short words like "oh" and long words like "seven", all models are sixteen states long.

Whereas digits are a special case where solutions can be hand-optimized, the global decision tree used here is data-driven and can be applied to any system incorporating whole word models. Any acoustic redundancies should be automatically found and incorporated into the acoustic model topology.

Questions for this English digit model consisted of word-id questions and state-id questions. The word-id questions were of the form "is the current word in this class", where the set of all possible one and two word classes was hypothesized. The state-id questions were chosen to be "is the current state greater than or equal to n," where n was able to take on values between 3 and 17.[1]

As mentioned in Section 2, the threshold chosen to control the size of the tree directly affects the final size of the model. Tables 1 and 2 show how recognition accuracy changes with model size (number of clustered states). Word accuracy is shown on three partitions of the data: on clean test data, on test set A with 10dB signal to noise ratio (SNR), and an average accuracy. There is only a slight degradation in accuracy at a level of 140 shared states, which represents a model compression of 25%. Even with as few as 120 shared states (33% reduction) shows only a minimal loss in accuracy over the demonstrated test data.

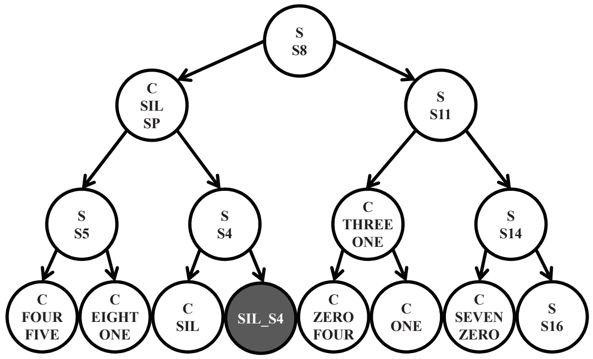

Fig. 1. *The first nodes of the clean model Aurora 2 global state-tying decision tree.*

Figure 1 represents the first few levels of the tree learned from clean training data. This part of the tree was identical for all of the systems reported in Table 1.

The root node asks the question that splits every word's HMM in half: "Is the current state greater than or equal to 8," abbreviated as "S:S8". If the question is false, the data proceeds to the left, and the next question is whether the current word is in the class that contains "SIL" and "SP", abbreviated as "C:SIL SP". If it is not, then another split happens at state S5, which allows the question "C:FOUR FIVE" to cluster the first states of those two words together. All this is done automatically, without linguistic or phonetic knowledge. Notice also that even though the tree is just three levels deep, it has already isolated the final state of the silence model SIL_S4.

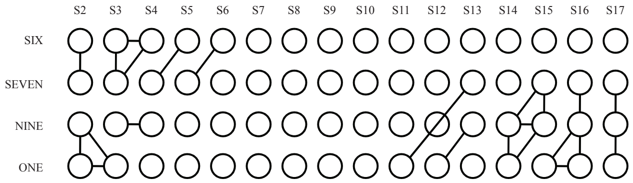

Fig. 2. *An example of logical states clustered across four of the words in the Aurora 2 task.*

Figure 2 shows how the clustered states are shared among the logical phones from four of the acoustic models in a clean trained acoustic model with 142 clustered states. Each row in the figure represents one of the whole-word HMMs, and each circle is one of its logical states. Groups of states joined by lines indicate they all map to the same clustered states.

In the model, there are quite a few HMMs where adjacent states are clustered together. This is probably an indication that the original model had too many states, and that the superfluous parameters have been eliminated automatically by the clustering process.

[1] Valid state numbers for this task range from 2–17 for each of the eleven digit models, 2–4 for the silence model, and 2 for the short pause model.

		Compression Ratio			
Language	Condition	0%	25%	50%	75%
Danish	hm	54.26	55.37	57.34	54.18
Danish	mm	77.47	78.26	77.60	77.21
Danish	wm	91.12	91.27	91.16	90.66
German	hm	84.32	84.64	84.04	83.40
German	mm	82.94	83.31	82.14	81.04
German	wm	93.31	93.33	92.84	90.70
Spanish	hm	73.08	73.56	71.20	67.97
Spanish	mm	89.41	89.87	89.35	89.10
Spanish	wm	94.44	94.39	94.12	94.12
Finnish	hm	65.19	67.67	66.40	64.66
Finnish	mm	83.45	83.58	82.97	80.51
Finnish	wm	93.84	93.93	93.82	94.14

Table 3. *Word recognition accuracy for Aurora 3 systems in all four languages and thee conditions. As the compression ratio increases from 0% to 25%, the accuracy stays the same or increases slightly.*

There are also examples of similar sounds in different words mapping to the same clustered states. The initial sounds of ”SIX” and ”SEVEN” share the same states, as well as the final sounds of the models for ”ONE”, ”NINE”, and ”SEVEN.” Additionally, ”NINE” and ”ONE” share the same initial state. There is no good reason for this cluster, and it can probably be blamed on over-zealous clustering.

3.2. Aurora 3 Results

The Aurora 3 task is similar to the Aurora 2 task in many ways. It consists of recognizing strings of digits recorded in noisy car environments in four European languages (Danish, German, Spanish, and Finnish). Like Aurora 2, the baseline models are sixteen-state whole word models for each digit, in addition to the silence and short pause models.

The question sets for each language of the Aurora 3 task were generated in the same way as the question set for the Aurora 2 task. That is, both word-id and state-id questions were asked. The classes for the word-id questions were automatically generated from all possible subsets of the vocabulary with one or two class members.

Table 3 shows complete Aurora 3 evaluation results for four different model sizes. When training each of these models, a threshold was chosen to create models with up to 75% fewer clustered states than the original, unclustered acoustic model. For each of the four languages, three digit accuracies are reported corresponding to three different training conditions. For high-mismatch (hm), the training data is much less noisy and reverberant than the test data. In the well-matched (wm) condition, both the training and testing data contain the same mix of more or less noisy data. Finally, the medium-mismatch (mm) condition contains different noise conditions for training and testing.

With 25% fewer clustered states than the baseline model, eleven of the systems are actually better than the baseline, with one (Spanish, well-matched) being insignificantly worse.

It is more surprising that, for most cases, even at a 50% compression ratio, the word accuracy is not significantly degraded. In fact, all of the German models are much better than the baseline at this size. This is a good indication that the baseline models had too many states. They were learning patterns and distinctions in the noisy training data that did not generalize well to the testing data.

Class	Members
Anterior	w l el z s dh th v f en n d t m b p dx
EVowel	ey eh ae
Fricat	dh th jh ch v f zh z sh s
iy	iy
Liquid	hh y w r l el
Medium	l el eh ae ow ax ah er ey
UnStrident	w y r l el ng g k en n d t m b p dx
Vowel	oh ax ah ow aw iy ey oy ay er uh uw aa ah ae ao ih ix eh ae ix
Short	oh uh ax ah aa oy ay ih eh ah ey ae ix

Table 4. *A subset of the phone classes used in our TIMIT experiments.*

3.3. TIMIT

The TIMIT corpus [6] provides 3696 utterances for training and 1344 utterances for testing. In addition to the acoustic data, it also includes detailed phonetic transcriptions for each utterance. For our tests, we use use the convention of Hon and Lee [7], training models for 48 different phones and then mapping down to 39 phones for scoring purposes. Recognition was performed using standard Viterbi search with a bigram phonetic language model.

The question set explored during the clustering included questions about the current state, about the current center phone, and about the current left and right context phone class. In contrast, a standard system would only use the context questions. We used a total of 109 phonetic classes to generate the questions, a subset of which are shown in Table 4.

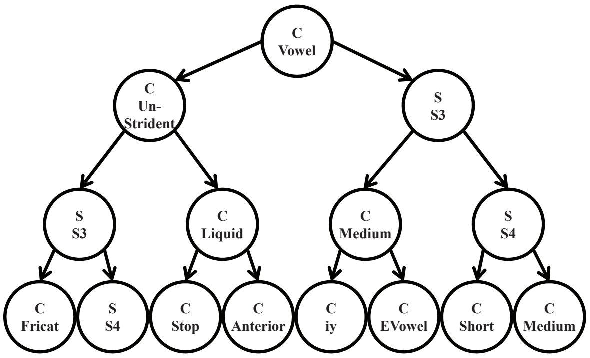

Fig. 3. *The first nodes of the TIMIT global state-tying decision tree.*

Figure 3 shows the initial nodes of the global decision tree generated on TIMIT data. Near the top of the tree, most of the questions are being asked about the class of the center phone, or the state number of the HMM. At deeper levels, more context questions occur alongside the center phone and state questions. And, when the final agglomerative clustering stage occurs, two leaf nodes may be merged that previously had different paths through the tree.

Figure 4 shows how the phone recognition accuracy of the system was affected by varying the model size. We compare to a baseline where each context-independent phone state was clustered independently (HHEd baseline). Although there isn’t a significant difference in accuracy at equivalent model sizes, there are two interesting results from this experiment. We found that the likelihood of the

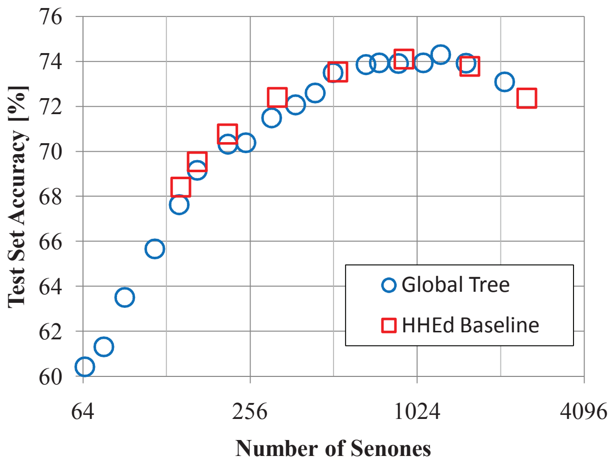

Fig. 4. *Accuracy of the TIMIT system with varying model size*

data under the model was consistently better using the global tree, which indicates it is a better match. Also, notice that the HHEd baseline can not build fewer clustered states than the number of context-independent phone states (144), but the global tree can make much smaller models.

It is also informative to analyze the structure found in the best global decision trees. We found three different classes of tying that would exist in the traditional modeling framework.

3.3.1. Co-articulation

One motivation behind using triphones is that they capture co-articulation effects. That is, the final sound of phone 'a' will change depending on what the next phone will be. But, sometimes, the sound will be similar for two contexts. For instance, the final sound of a+b (the model for 'a' when followed by 'b') might be the same as the final sound of a+c. Standard tree clustering methods will take advantage of this data redundancy and cluster the final state of a+b and a+c together.

But, it could also be the case that the initial state of some 'b' and 'c' models are similar, when they are following the phone 'a'. That is, the initial states of a-b (the model for 'b' when following 'a') and a-c are acoustically similar.

We did find many cases like this in the global tree for TIMIT. For instance, the triphones iy-jh+ah, and iy-ch+ah (along with fifteen other similar physical triphones) share the same clustered state as their first state.

3.3.2. Triphone HMMs with too many states

When aligning a full set of three-state triphones to the training data, one assumes that all of the states are necessary to model the acoustics. It could be the case, as with the whole word models discussed earlier, that this method allocates too many states to some utterances. As with the whole-world models, the expectation is that this would manifest itself as some triphones repeating a clustered state for more than one consecutive state.

Two typical examples are the physical triphones uw-w+ao and uw-w+aw, where the first two states of both models are represented by the same clustered state. This case was common for triphones based on the center phones el, l, and w. Another interesting and unexpected example is the pair of triphones n-m+n and n-ng+n, whose initial and final states are all modeled by the same clustered state.

3.3.3. Acoustically confusable phones

Another common feature found in the globally tied model was that sometimes the center states of triphones with different center phone ids share the same clustered state. This was common with classically confusable center phone pairs such as (n,en), (ix,ih), and (g,k) in similar contexts. What the tree is learning from the data is that, in some contexts, there is not enough acoustic information to distinguish these confusable phones.

4. DISCUSSION

Decision trees are a common tool used to cluster logical states into clustered states for large vocabulary speech recognition systems. They are usually limited to operate independently on every context-independent state of the acoustic model. Using a global decision tree to cluster the logical HMM states generates a better model and allows the decoder to concentrate on real distinctions supported by the acoustic training data.

We have discussed three experiments where a global decision tree was used to jointly cluster every logical state in the acoustic model. For noisy speech, we showed that reducing model size can increase accuracy in some cases, and up to 50% reduction in model parameters can be achieved without any significant accuracy degradation. For TIMIT phonetic decoding, we showed that the global tying structure makes reasonable decisions, complementing the kind of tying that occurs in the baseline model.

5. REFERENCES

[1] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the workshop on Human Language Technology*, Plainsboro, NJ, 1994, pp. 307–312.

[2] D.B. Paul, "Extensions to phone-state decision-tree clustering: single tree and tagged clustering," in *Proc. ICASSP*, 1997, vol. 2, pp. 1487–1490.

[3] A. Lazarides, Y. Normandin, and R. Kuhn, "Improving decision trees for acoustic modeling," in *Proc. ICSLP*, Philadelphia, October 1996.

[4] Hua Yu and Tanja Schultz, "Enhanced tree clustering with single pronunciation dictionary for conversational speech recognition," in *EUROSPEECH 2003*, Geneva, September 2003, pp. 1869–1872.

[5] H. G. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluations of speech recognition systems under noisy condidions," in *ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium"*, Paris, France, September 2000.

[6] J.S. Garofolo, L.F. Lamel, W.M. Fisher, J.G. Fiscus, D.S. Pallett, and N.L.Dahlgren, *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus CDROM*, NIST, 1986.

[7] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 11, pp. 1641–1648, November 1989.