WIDE CONTEXT ACOUSTIC MODELING IN READ VS. SPONTANEOUS SPEECH

Michael Finke[1] *Ivica Rogina*[2]

Interactive Systems Laboratories
[1] Carnegie Mellon University, USA
[2] University of Karlsruhe, Germany

ABSTRACT

Context-dependent acoustic models have been applied in speech recognition research for many years, and have been shown to increase the recognition accuracy significantly. The most common approach is to use triphones. Recently, several speech recognition groups have started investigating the use of larger phonetic context windows when building acoustic models. In this paper we discuss some of the computational problems arising from wide context modeling (polyphonic modeling) and present methods to cope with these problems. A two stage decision tree based polyphonic clustering approach is described which implements a more flexible parameter tying scheme. The new clustering approach gave us significant improvement across all tasks - WSJ, SWB, and Spontaneous Scheduling Task - and across all languages involved (German, Spanish, English). We report recognition results based on the JANUS speech recognition toolkit [2, 8] on two tasks comparing acoustic context phenomena in English read versus spontaneous speech. We used our WSJ 60K recognizer and the JANUS SWB 10K polyphonic recognizer.

1. INTRODUCTION

The phonetic context F OW N Z of a given phone OW affects the acoustic realization of that phone. Therefore, using acoustic models which make effective use of the information about the preceding phone (F) and the following phone (N) leads to a significant improvement in terms of speech recognition performance [6]. But this approach ignores the strong influence that may be exerted by phones that are further away than the immediately preceding and following one. There has been some research towards using wider contexts by allowing questions in the decision tree clustering approach to refer to phonetic contexts two or more phones to the left or right of the phone to be modeled (**polyphonic models**) [1, 5].

In this paper we examine the effect of the width of the context on the speech recognition process in the JANUS recognition toolkit [2, 8], especially on computational effort to train and cluster the models and on the resulting error rate. We will see that the error rate can be reduced significantly by increasing the context width. Our main focus will be on comparing the effect of modeling the phonetic context variation in read versus spontaneous speech by referring to the the Wall Street Journal task as benchmark task for English read speech and Switchboard as the spontaneous speech benchmark.

Different methods of clustering have been proposed [6, 4, 7]. The greater the number of models to be clustered, the more infeasible it will become to do agglomerative clustering. Divisive clustering methods are usually implemented as decision trees, using a predefined set of questions for making decisions. In JANUS, we use maximum entropy gain on the mixture weight distributions as the selection measure for dividing a cluster into two subclusters. We have examined the selection of the top-gaining questions during the clustering process and will report the results below.

2. DICTIONARY AND POLYPHONES

In order to explain the number of polyphones observed in the training data of WSJ as well as SWB we have to realize that there are some major differences in size and structure of the transcriptions of the two databases: the WSJ training data consist of 700k words whereas SWB has about twice as many words in the transcription of the acoustic training data. The following figure shows that the frequencies of different word lengths in phones differs very much between Wall Street Journal and Switchboard.

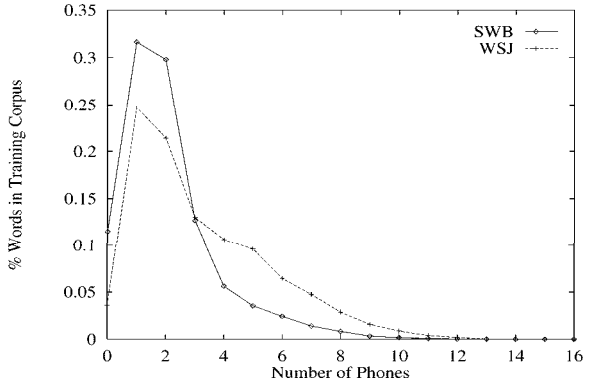

Figure 1. Dictionary word length distribution

While for WSJ the most frequent number of phonemes in the dictionary is 6 phonemes, the most frequent word length in the training data is 2 phonemes. Under these circumstances, triphone modeling gives us word-dependent

models for approximately 50% of the words in the training data. Quintphones cover 80%, and septphones around 90%. Given these figures we expect the benefit from using wide contexts to decrease with the size of the context.

In JANUS, the maximum usable context width is all phones within a word and up to one phoneme into the neighboring word, limited by the current implementation of the decoder. In order to cluster wide context acoustic models in an efficient way, we have to cope with the problem of handling a prohibitively large number of initial acoustic models to start with. Figure 2 shows the number of different models we get when using different context sizes. The one order of magnitude larger numbers for the SWB task are partially due to the greater size of the task in terms of the number of words in the transcription. But another important aspect is much greater diversity expressed in the perplexity of the task as well as the mean number of pronunciation variants per word.

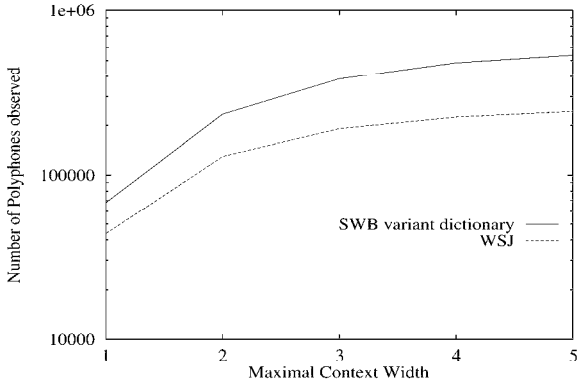

Figure 2. Number of polyphones observed for different context sizes

Many of these polyphones are seen very rarely during training. Figure 3 shows how many polyphonic models are seen a given number of times. We can see that the wider the context, the greater is the part of the polyphones that are seen less often, while the number of very frequent polyphones decreases.

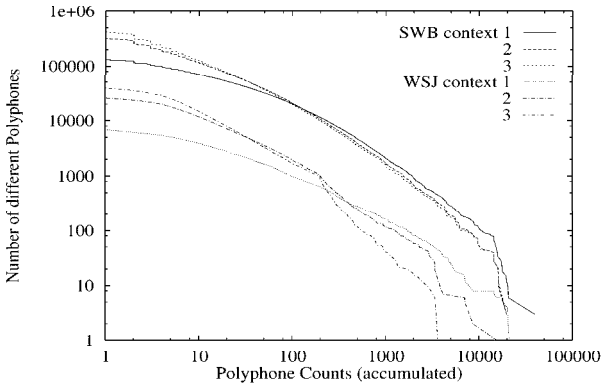

Figure 3. Frequency of polyphone counts

3. THE CLUSTERING ALGORITHM

The polyphonic clustering algorithm first collects all polyphones that occur in the training data. Hereby, the constraints imposed by the decoder which allow cross-word context to contain one phoneme from the neighboring word only are satisfied. Each polyphone is acoustically modeled with three states (subpolyphones), each one modeled as a distribution, i.e. as mixture weights over a codebook.

3.1. Polyphonic Trees

Due to the extremely large number of models we have to handle within the clustering procedures, we had to come up with efficient data structures to organize the polyphones and their associated distributions. One efficient way to represent a set of polyphones are so called polyphonic trees: The root of the tree is the center/mid phone. For each observed immediate context ± 1 there is a child to the root node with the names of the left and the right phone, the count of how often the respective "triphone" was observed, and a pointer to the acoustic model (distribution). Each "triphone" child has a set of children one for each "quintphone" context found around the triphone parent in the training data.

3.2. Initialization

The starting point for the clustering procedure is a decision tree as shown in figure 4 (left). It has one leaf for each phone in the set of phones. Attached to each leaf there is a polyphonic tree containing all the observed polyphones that fall into that leaf. All Polyphones within a polyphonic tree share a single codebook.

3.3. Splitting Criterion

We then develop a decision tree as described in [4, 7], allowing questions about arbitrary contexts. These questions are based on 80 different subsets (e.g. vowels, syllabics, voiced phones...) of our set of phones. For each of these subsets a question is defined with respect to all possible contexts (in this case -2,-1,1,2) and for each question an extended question is added which also asks whether the considered context is tagged as being a word boundary.

The distance metric defining the gain received by splitting a tree node is measured as the loss of entropy as described in [6].

$$p_i^l = \frac{1}{\gamma^l} \sum_{m \in L} \gamma_m \alpha_{mi}, \quad \gamma^l = \sum_{m \in L} \gamma_m$$

$$p_i^r = \frac{1}{\gamma^r} \sum_{m \in R} \gamma_m \alpha_{mi}, \quad \gamma^r = \sum_{m \in R} \gamma_m$$

$$\begin{aligned} D(q) &= \gamma^l H^l + \gamma^r H^r - \gamma H \\ -H^l &= \sum p_i^l \log p_i^l \\ -H^r &= \sum p_i^r \log p_i^r \end{aligned}$$

where γ_m are the counts for model m, α_{mi} counts for component i of model m.

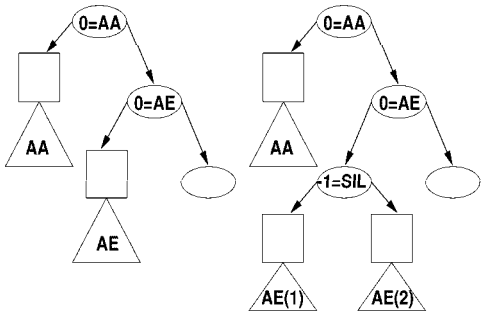

Figure 4. Splitting a decision tree node and its associated polyphonic tree based on a phonetic question

3.4. Training Schedule and Codebook Tying

In our standard training scheme we first grow a decision tree until it reaches the number of desired leaf nodes (typically a few thousand, depending on the size of the available training data; grey nodes in Figure 1). We constraint splits to be only valid as long as both child nodes created still have sufficient training data to train the underlying codebook. Then, a fully continuous Gaussian mixture model is trained for every leaf node. In a second clustering phase, we continue growing the decision tree and eventually train a separate distribution of mixture weights for each of the resulting leafs. This is a new way of optimizing the degree of mixture tying in a large vocabulary hidden markov model based speech recognition system.

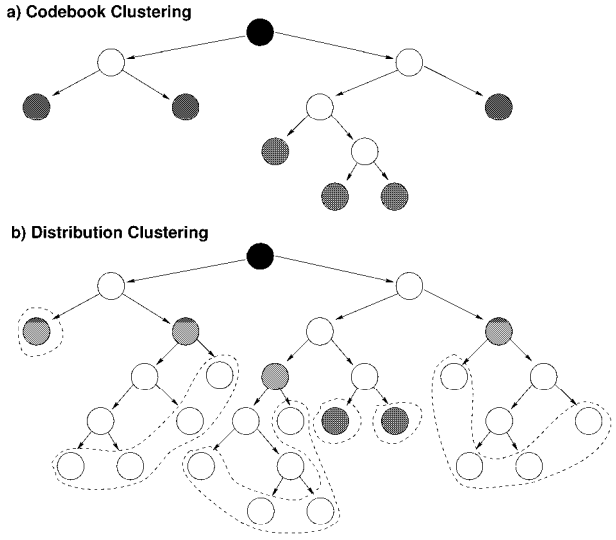

Table 1. Two stage clustering of acoustic models (the distributions in the same dashed area are defined on the same codebook)

Hidden markov models with continuous densities provide a detailed stochastic representation of the acoustic space at the expense of increased computational complexity and lack of robustness. This two level clustering approach addresses the problem of the lack of robustness by having a set of distributions share the same codebook. In particular, the algorithm proposed helps to automatically determine the number of sets of HMM states which share the same codebook and based on that subsets of HMM states which share the same distribution.

4. SELECTION OF QUESTIONS

In our experiments we proved our expectation that questions about far contexts generally get a smaller gain than questions about the close context. So close-context-questions get more frequently used in the decision trees. Wide-context questions become more likely if we look at the deeper levels of the tree. Figure 5 displays the frequencies of the context width in the decision tree questions at different phases of the decision tree growing algorithm. We can clearly see that the part of questions about the wide context 2 in the WSJ task is larger than in the SWB task, which is due to the smaller diversity and greater structural organization of the task

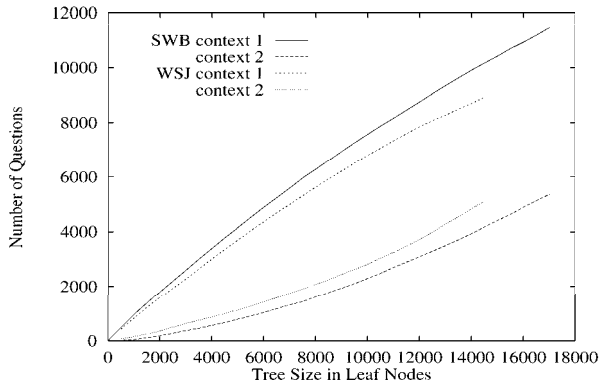

Figure 5. Frequency of questions referring to triphonic vs. quintphonic context

Another interesting observation is the very large number of questions which ask for the word boundary tag, which means that there is a strong focus onto crossword triphones when clustering (see figure 6).

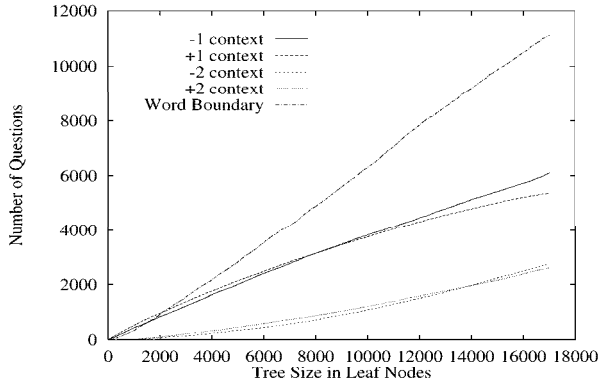

Figure 6. Word boundary related related questions

Figure 7 displays the observed average entropy gain at different stages of the decision tree development:

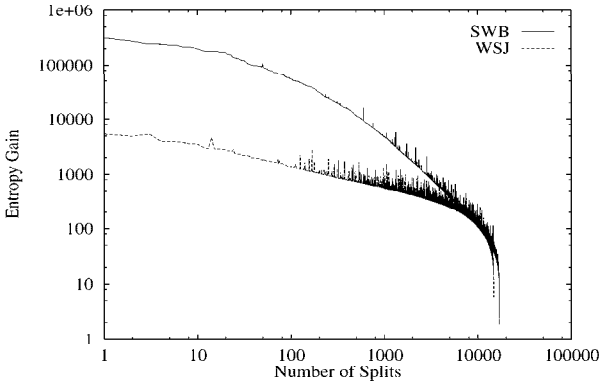

Figure 7. Entropy gain

We have analysed which questions (out of 281 used) are getting the best entropy gains. In both tasks similar questions are among the best scoring questions:

Question	WSJ rank	gain sum	SWB rank	gain sum
+1=SILENCES	1	69108	1	2629568
-1=SILENCES	2	59130	3	2249279
-1=VOICED	10	39089	4	1121313
+1=HIGH_VOWEL	5	49847	9	764036

5. EXPERIMENTS

We have conducted recognition experiments with the JANUS recognizer [2, 8] on three tasks: the Wall Street Journal task (WSJ) the Switchboard LVCSR task, and the German spontaneous scheduling task. The two English tasks use the same phoneme set and the same set of questions, to make them better comparable. All recognizers use approximately the same number of parameters. We have observed a relative error reduction of 5% on the WSJ task, by increasing the context width from 1 to 3. The increase of the context width from 1 to 2 reduced the error on the SWB task by 8%. A similar improvement was achieved on the German, Spanish and English Spontaneous Scheduling tasks. Our currently best performance on the WSJ task (evaluation set Nov. 1994) is at 9.0% errors. The SWB recognizer was top ranking in DARPA's spring 96 LVCSR evaluation [3, 9] and currently has an error rate of 36%.

Task	Context ± 1	Context ± 2	Context ± 3
WSJ	20.9% WE	20.2% WE	19.9% WE
SWB	46.0% WE	43.6% WE	

Table 2. Results on different context width.

6. CONCLUSION

In our experiments so far we have shown that wide context acoustic modeling can reduce the error rates significantly. We presented a new clustering approach which combines clustering and tying in one procedure. The benefit from using wide context is greater for spontaneous speech than for read speech due to the more prominent coarticulation effects when speeking in a spontaneous way.

In the future we intend to examine different distance measures for splitting a decision tree node, and we will work on methods that help to find the optimal context widths and optimal numbers of acoustic models automatically.

7. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technology (BMBF) as a part of the Verbmobil project. The JANUS project was supported in part by the Advanced Research Project Agency and the US Department of Defense.

REFERENCES

[1] L.R. Bahl, P.V. de Souza, P.S. Gopalakrishnan, D. Nahmoo, and M.A. Picheny. Decision Trees for Phonological Rules in Continuous Speech. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, 1991. IEEE.

[2] Michael Finke, Petra Geutner, Hermann Hild, Thomas Kemp, Klaus Ries, and Martin Westphal. The Karlsruhe-Verbmobil Speech Recognition Engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.

[3] Michael Finke, Torsten Zeppenfeld, Martin Maier, Laura Mayfield, Klaus Ries, Puming Zhan, John Lafferty, and Alex Waibel. Switchboard April 1996 Evaluation Report. In *Proceedings of LVCSR Hub 5 Workshop*, April 1996.

[4] M.Y. Hwang. *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition.* PhD thesis, Carnegie Mellon University, 1993.

[5] R. Kuhn, A. Lazadrides, Y. Normandin, and J. Brousseau. Improved Decision Trees for Phonetic Modeling. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 552–555, Detroit, Michigan, 1995. IEEE.

[6] Kai-Fu Lee. *Large-Vocabulary Speaker-Independent Continuous Speech Recognition: The SPHINX System.* PhD thesis, Carnegie Mellon University, April 1988.

[7] Julian James Odell. *The Use of Context in Large Vocabulary Speech Recognition.* PhD thesis, University of Cambridge, March 1995.

[8] A. Waibel, M. Finke, D. Gates, M. Gavalda, T. Kemp, A. Lavie, M. Maier, L. Mayfield, A. McNair, I. Rogina, K. Shima, T. Sloboda, , M. Woszczyna, and P.Zhan. JANUS II - Advances in Spontaneous Speech Translation. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, 1996. IEEE.

[9] Torsten Zeppenfeld, Michael Finke, Klaus Ries, Martin Westphal, and Alex Waibel. Recognition of Conversational Telephone Speech using the JANUS Speech Engine. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Munich, Germany, 1997. IEEE.