DATA-DRIVEN CODEBOOK ADAPTATION IN PHONETICALLY TIED SCHMMS

Thomas Kemp
<kemp@ira.uka.de>

Interactive Systems Laboratories,
Department of Computer Science (ILKD),
University of Karlsruhe,
76128 Karlsruhe, Germany

ABSTRACT

This paper reports the results of our experiments aimed at the automatic optimization of the number of parameters in the semi-continuous phonetically tied HMM based speech recognition system that is part of the speech-to-speech translation system JANUS-2.
We propose different algorithms devised to determine the optimal number of model parameters. In recognition experiments performed on a spontaneous human-to-human dialog database, we show that automatic optimization of the acoustic modeling parameter size with the proposed algorithm improves the recognition performance without increasing the required amount of computing power and memory.

INTRODUCTION

SCHMMs [1] share parameters on the level of centroids and variances of Gaussian densities in the feature space. The distinction of the phonetic events, like senones [2], triphones or phonemes is done with probability distributions which share the common codebook. The parameter sharing leads to robust estimation of the codebook even if only relatively small amounts of training data are available. In addition, it reduces the number of parameters in the system, thus improving the generalization power of the models and reducing both, time and memory requirements.

As training data size has increased in the last few years, the advantages of the strong tying of parameters in the SCHMM are in many cases of lesser concern. More and more attention is focused on systems with less closely tied parameters, such as CDHMMs and phonetically tied SCHMMs. In the latter, codebook parameters are shared on the level of phonemes. This combines the robustness of parameter tying with much finer modeling.

Often, a fixed number of codebook vectors is assigned to each of the phonemes. However, there is more training data available for the more frequent phonemes than for the rarer ones. In addition, the size of the feature space which is covered by each phoneme differs greatly between the phonemes, thus suggesting that some phonemes might better be modeled with more codebook vectors, others with fewer. Taking the same number of codebook vectors for each phoneme, leads on the one hand to overfitting of the training data and hence to a loss in generalization power, accompanied by a waste of computing and memory resources. On the other hand, some phonemes are insufficiently modeled. Some of their subclasses (a subclass might be e.g. a generalized allophone, or a senone) will not be adequately modeled because they are relatively rare, even if they are easy to separate. This will result in a lower recognition rate.
Similar to the results of [3] with a connectionist recognizer, we attempted to adapt the codebook size of each phoneme according to the amount and the distribution of the training data. We designed an algorithm which scans the data and computes the adapted number of codebook vectors without supervision. In the following section we introduce our implementation.

1. BASIC ALGORITHM

The basic algorithm can be summarized as follows:

1. Generate a pool of training samples for each phoneme using forced alignment over the training data.
2. For each phoneme:
   - set codebook size N to 1, train a system
   - While some quality criterion is not yet satisfied:
     (a) increase N
     (b) cluster the data into N clusters with the basic isodata (k-means) algorithm, use the clusters as codebook.
3. Adapt the probability distributions over the new codebooks.

The problem of adapting the codebook size to the training data can thus be reduced to the specification of a quality criterion that can be used to stop the process of increasing the codebook size. Two possible solutions to this problem will be evaluated and compared.

1.1. Variance criterion

The process of adding new reference vectors to the codebook is stopped after the average squared distance of each data point to its nearest codebook vector falls below a given threshold. The algorithm may be viewed as equivalent to the construction of a vector quantizer with a given quantization error, using a squared-distance distortion measure.
For this approach, a threshold must be defined manually. While this necessity is undesirable for an algorithm that should run without supervision, it enables us to control the average number of parameters in the resulting system: a low threshold will lead to more parameters than a higher one. If external constraints on computation time or memory must be met, this flexibility can be a great advantage.

0-7803-2431-5/95 $4.00 © 1995 IEEE

1.2. Prediction criterion

The prediction criterion tries to capture how well the modeling of the recognizer can predict previously unseen data. To compute it, we split our training database into a training set and a crossvalidation set. After computing the codebook on the training set portion of the data, we compute the probability $p(data_{xval}|codebook)$. The (negative log) probability over the codebook size shows a learning curve: after a rapid decrease, it reaches a flat minimum and then begins to rise again. The minimum of the curve defines the optimal number of codebook vectors. Figure 1 shows a typical example.

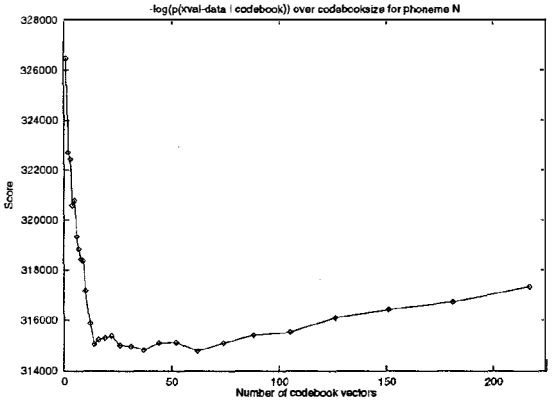

Figure 1. Crossvalidation performance for phoneme 'N'

The prediction criterion is theoretically appealing because it is near to the criterion which one really wants to improve on, namely recognition performance on previously unseen data. It does not need a threshold that has to be set manually.

The number of resulting parameters only depends on the data and on the approximations, which are used in the given speech recognition system to compute the conditional probability $p(data|phoneme)$. Such approximations might be the restriction of covariance matrices to be diagonal, or to take only the top N scoring codebook vectors or the like. If these approximations are held constant, there is no way to control the resulting number of parameters when employing the prediction criterion.

Additionally, the learning curve can have several local minima (cf. figure 1), making the computation of the correct number of codebook vectors an ambiguous task.

2. BASELINE SYSTEM

2.1. Databases

For the described experiments we used the GSST (**G**erman **S**pontaneous **S**cheduling **T**ask) database, which has been collected at the University of Karlsruhe. It consists of human-to-human spontaneous german dialogs in the appointment scheduling domain, i.e. two persons try to schedule a meeting within the next month. The database contains about 3 hours of speech and has an average bigram test set perplexity of 70.

2.2. The JANUS-2 system

The speech-to-speech translation system JANUS-2 [4] is a joint effort of the Carnegie Mellon University, Pittsburgh, and University of Karlsruhe, Germany.

The baseline speech recognition component of JANUS-2 uses a phonetically tied SCHMM with 50 reference vectors per phoneme. Generalized triphones are used to capture contextual information. In the preprocessing stage mel-scale spectra with a frame rate of 10 ms and their first derivatives with respect to time; power, zerocrossing rate and peak-to-peak value are computed. The 37-dimensional input vector is transformed by linear discriminant analysis (LDA, [5]) and split into two 16-component data streams. Stream weights training [6] can be applied. Training can be done with Viterbi alignment or the standard forward-backward training algorithm. The emerging reference vectors can further be trained discriminatively according to the LVQ learning rule.

To speed up computations, usually only the N codebook vectors with the smallest Mahalanobis distances in a codebook are taken into account for the score computation. For all experiments described throughout this paper, N was set to 1.

The decoder computes word lattices with a Viterbi forward pass and a word-dependent n-best [7] backward pass.

Noise models [8] have been recently included to improve the performance of JANUS-2 on spontaneous speech.

In the following table, we give a comparison of recognition results on the spontaneous GSST database with results on a standard read-speech task (ARPA resource management task).

Database	Word accuracy
Resource Management	94.1
GSST	66.9

Table 1. Baseline system recognition results

3. RESULTS OF CODEBOOK ADAPTATION

3.1. Results with the variance criterion

We employed the basic algorithm with the variance criterion using different values for the distance threshold. Out of the thresholds evaluated, we chose one that produced a similar number of codebook vectors (4201) as our baseline system (4600), and another one that produced a significantly lower number (1916). Table 2 shows the resulting codebook sizes of the system with 4201 codebook vectors for a subset of our phoneme set. Results are given for both of the two 16-component data streams which result from the LDA. Note the large differences in the codebook size of the different phonemes.

3.2. Results with the prediction criterion

The results achieved with the prediction criterion were rather surprising. First, the size of the resulting codebooks for stream 1 did not show an obvious correlation to the codebook size that was computed with the variance criterion (cf table 2). Second, the resulting codebook sizes for *all* codebooks defined over stream 2 were one. The resulting system uses 46 codebook vectors for stream 2, and 1631 vectors for stream 1.

Criterion:	var.	pred.	var.	frames
input stream:	1	1	2	available
A-0	80	37	64	41776
AH-0	24	52	16	43164
AI-0	32	62	32	36341
AU-0	32	37	40	17602
CH-0	80	31	32	47453
D-0	128	44	40	34179
EH-0	32	62	48	23279
F-0	8	6	4	32376
G-0	192	52	64	15657
H-0	48	26	24	8628
L-0	128	44	96	24050
M-0	32	26	48	41664
N-0	64	16	80	50000
R-0	128	74	64	50000
S-0	20	8	16	50000
SIL-0	1	1	1	50000
T-0	96	31	24	50000
Sum	2421	1631	1780	1145000

Table 2. Codebook size for some phonemes as computed with different criteria

Model	codebook size	W.A.	error reduction
baseline	4600	66.9%	-
prediction	1677	67.8%	3.9%
variance	1916	65.5%	-6%
variance	4201	69.9%	10%

Table 3. Recognition results

This result suggests that the data in stream 2 can be adequately modeled with a unimodal distribution. An additional experiment in fact showed that the average within-class scatter $|\Sigma_{ik}|$ for stream 1 was 70 times higher than for stream 2, when using a unimodal distribution (i.e. only one codebook vector per phoneme and stream).

This difference in the behaviour of the two data streams might be explained by the properties of LDA. The information content of the coefficients of an LDA output vector is decreasing with the coefficient index. This means, that the first data stream in our recognizer, containing coefficients 1 to 16 of the LDA output vector, carries much more information than the second data stream, which contains the coefficients 17 to 31.

3.3. Recognition results

The recognition results shown in table 3 have been achieved with 2500 context dependent generalized triphones and a bigram language model of perplexity 70. Only first-best results are given. No cross-word triphones were applied, and no reordering of the resulting word lattices took place. All models were trained gender-independent. All covariance matrices were restricted to their main diagonal. For each experiment, the complete recognizer was bootstrapped on pre-segmented data and trained until no further improvement on cross-validation data could be reached.

Codebook size adaptation was capable of decreasing the system error rate by 10 percent with the same number of parameters. The prediction criterion performs better than the variance criterion for a comparable number of parameters (it even outperforms the baseline system with only roughly one third of its parameters), but the best system performance is reached with a large system optimized with the variance criterion.

We performed an additional experiment, where we used only data stream 1 for the system that was adapted with the prediction criterion. This resulted in remarkable 65.6% word accuracy for only one-half of the input available to the system.

4. DISCUSSION

In this paper we have reported about experiments with our codebook adaptation algorithm applied to the task of recognizing a difficult spontaneous human-to-human database. We showed that codebook adaptation leads to significant word error reduction if the same number of parameters is used. We also showed that an adapted system with only 37% of the parameters of the baseline system still performs at least equally well. This allows faster speech recognition with lower requirements of computational resources.

5. ACKNOWLEDGEMENTS

This research was partly funded by grant 413-4001-01IV101S3 from the German Ministry of Science and Technologie (BMFT) as a part of the VERBMOBIL project. The author wishes to thank all members of the Interactive Systems Labs, especially P. Geutner, I. Rogina, T. Schultz, T. Sloboda and M. Woszczyna, for useful discussions and active support. Special thanks also to Alex Waibel.

REFERENCES

[1] X. Huang, K.F. Lee, and H. Hon, *On Semi-Continuous Hidden Markov Modeling*, in Proc. ICASSP 1990, pp. 689-692

[2] M. Hwang, *Subphonetic Acoustic Modeling for Speaker-Independent Continuous Speech Recognition*, Ph.D. thesis, Carnegie Mellon University, 1993

[3] U. Bodenhausen, *Automatic Structuring of Neural Networks for Spatio-Temporal Real-World Applications*, Ph.D. thesis, University of Karlsruhe, June 1994

[4] M. Woszczyna, N.Aoki-Waibel, F.D.Buo, N. Coccaro, K. Horigushi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C.P. Rose, T. Schultz, B. Suhm, M. Tomita, A. Waibel, *Janus 93: Towards Spontaneous Speech Translation*, Proc. ICASSP-94, pp. 345-349, April 1994

[5] G. Yu, R.Schwartz: *Discriminant analysis and supervised vector quantization for continuous speech recognition*, Proc. ICASSP 1990, pp. 685 ff.

[6] I. Rogina and A. Waibel, *Learning State-Dependent Stream Weights for Multi-Codebook HMM Speech Recognition Systems*, Proc. ICASSP 1994

[7] S. Austin and R. Schwartz, *A Comparison of Several Approximate Algorithms for Finding N-best Hypotheses*, in Proc. ICASSP 1991, vol 1, pp 701-704.

[8] T. Schultz and I. Rogina, *Acoustic and Language Modeling of Human and Nonhuman Noises for Human-to-Human Spontaneous Speech Recognition*, elsewhere in these proceedings.