final report for

BAYESIAN INFORMATION CRITERION FOR AUTOMATIC MODEL SELECTION

submitted to:

Dr. Nicholas Younan

May 5, 1999

submitted by:

Jonathan Hamaker and Jie Zhao

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University
Box 9571
413 Simrall, Hardy Rd.
Mississippi State, Mississippi 39762
Tel: 601-325-8335
Fax: 601-325-3149
email: {hamaker, zhao}@isip.msstate.edu

INTRODUCTION

The primary problem in any pattern recognition system is to find a model which provides one with the best chance to recognize the patterns of interest. Many techniques require one to make assumptions about the structure of either the model or the data or both. Ideally one would like to use methods which are able to "learn" this structure. Hidden Markov models (HMMs) along with appropriate training techniques provide the ability to learn the underlying structure of the data. However, most HMM systems continue to assume a model topology rather than using a data-driven approach to automatically learn the model topology.

The attendant problem in conversational speech recognition seems to stem from poor acoustic-level matching as a result of the high degree of variability in pronunciations. There is much to explore about the "quality" of states in a Hidden Markov Model (HMM) and the relationships between inter-state and intra-state Gaussians used to model speech [1]. Of particular interest is the variable discriminating power of the individual states [2]. The variance observed at each state in the model varies significantly. Researchers often refer to this as the variance-reduction problem (and often cite this as the fundamental problem in speech recognition — decreased variance means improved prediction power). In this paper we investigate a data-driven approach for exploiting such dependencies through model topology optimization based on the Bayesian Information Criterion (BIC).

MOTIVATION

A state-of-the-art speech recognition system (as depicted in Figure 1) is a complicated machine which makes use of many concepts from pattern recognition such as supervised learning, structural pattern recognition, statistical confidence limits, and hypothesis-directed search. In such a system, there are many parameters which can be adjusted that have profound effects on the performance of the system. It would be impractical and inefficient for a scientist to attempt to optimize every parameter of the system for every task. Thus, many times assumptions are made to simplify the experimental design. One such assumption in HMM systems is typically the model topology. The standard topology for each model in an HMM speech recognition system is shown in Figure 2.

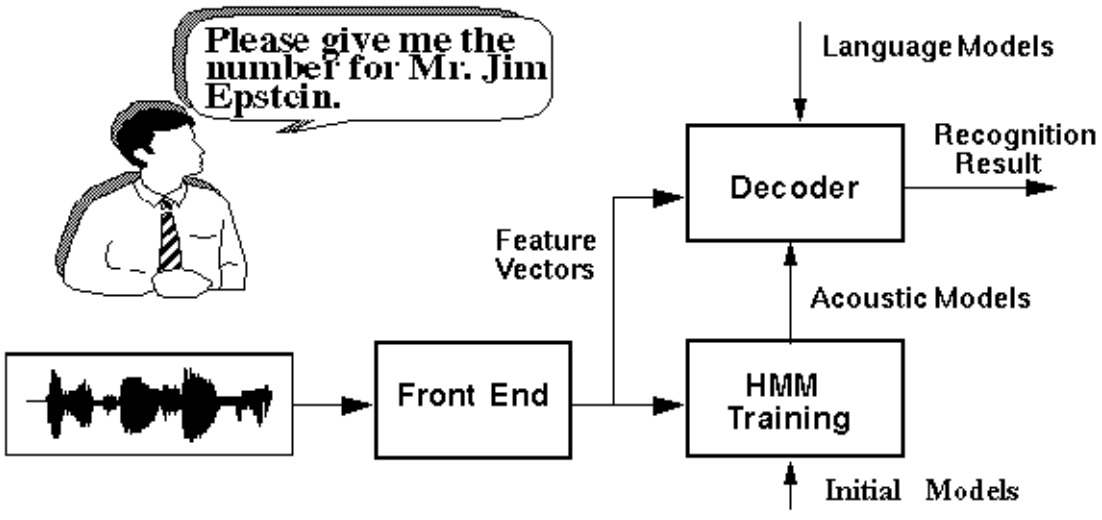

Figure 1. Block diagram of a typical speech recognition system. Note that in this setup, the topology of the initial models does not change during the training process.

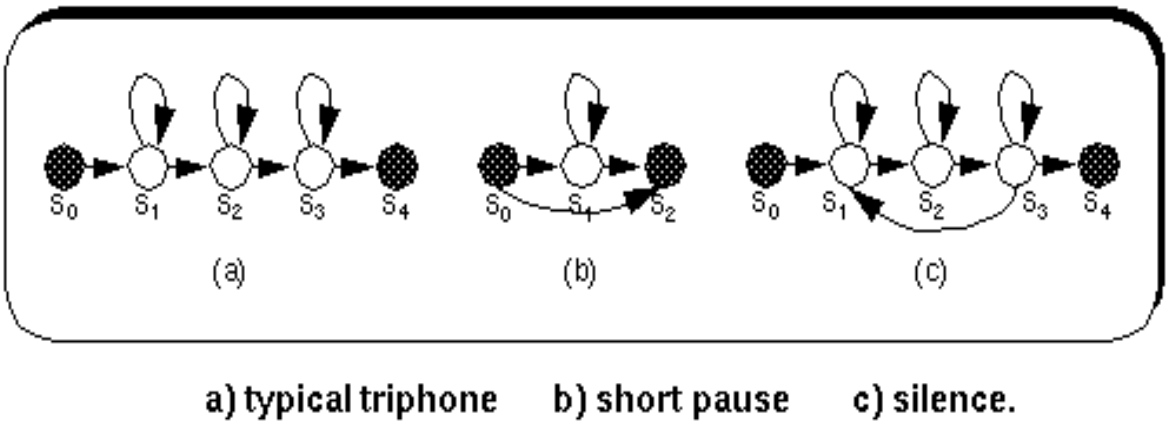

Figure 2. Typical topologies of models in an HMM system. Most models are of the form shown in a). b) and c) show some special exceptions which model long and short silence.

We conjecture that there exists an optimal structure to an HMM which best models the speech data for a given task. Unfortunately, in traditional speech recognition systems, the initial model choice is rather arbitrary — yet it bears a major impact on the ability of the trained model to fit the training data and generalize thereafter. For example, most state-of-the-art triphone based systems use a three-state HMM irrespective of the recognition task while many syllable-based systems use a number of states proportional to the average length of the spoken syllable. In order to test the above hypothesis a number of experiments were run on a telephone-quality continuous alphadigit task [5].

In this previous work we used syllable models where the number of states was made proportional to the average duration of the syllable in the training data. This system gave a WER of 11.1%. We followed this by setting an upper bound on the number of states a syllable model could have to 20 yielding a 1% absolute decrease in WER. Motivated by the paradigm for triphone modeling where all models are of a fixed length, we built a system where all syllable models were six states long. This system, however, increased the WER to 12.5%. As shown in Figure 3, these experiments indicate the probability of an optimal model set for a specific task. These experiments also demonstrate a strong dependence on the model topology and its ability to represent the data. Specifically, in large vocabulary continuous speech recognition, variability in speech makes arbitrary model structure definition a disaster.

To explore this topic, we previously ran a set of model merging experiments to attempt to find the optimal number of states for each model. In these experiments, we started with a number of states equivalent to the average length of the syllable. The models were reestimated on a small set of training data. A Bhattacharyya distance measure was then used to determine the overlap between consecutive Gaussian state distributions. All state pairs that had overlap greater than some constant were merged. Repetitions of merging and reestimating were carried out until no states could be further merged. The constant was empirically chosen. Though this approach produced the best results to date, 9.9% WER, the procedure for determining the model structure was heuristically based. In this work, we attempt to show that using an information-based, data-driven method to determine the model structure will provide a better model set for the task of Alphadigit modeling.

BAYESIAN INFORMATION CRITERION

Typical continuous speech recognition systems contain millions of parameters. Each of these has to be simultaneously optimized according to some measure. Most of these measures involve some optimization of the posterior probability of the data given the model. Many techniques (such as state-tying and clustering) are also employed which attempt to reduce the number of parameters in the system while maintaining a reasonable performance. However, the model topology is rarely a parameter which is considered in this optimization. In fact, the model topology is often static throughout the training process. When it is “optimized”, the process is often heuristic and error-prone. As such, it is useful to determine a measure of the “goodness” of the model topology parameters and to apply a principled approach for trading off the model quality versus the model size.

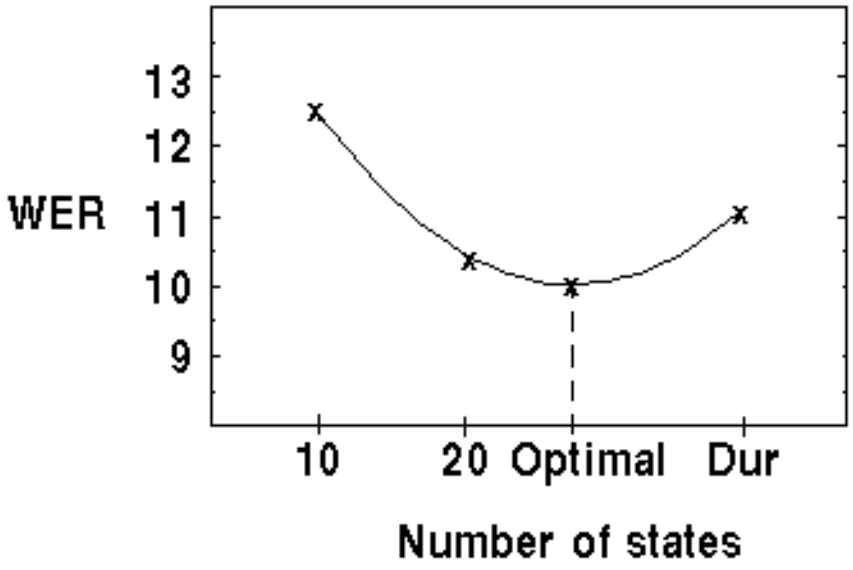

Figure 3. Results of previous work showing a function of the WER versus the model topologies chosen. The number of states chosen for each of these experiments was empirically chosen without taking the data (aside from the length) into consideration.

The model order decision criterion used in this study, Bayesian Information Criterion (BIC), is based on the principle of Occam's Razor: *when given a choice between models that model the data equally well, choose the one with the least complexity.* This is a particularly attractive approach when considering speech recognition since state-of-the-art systems commonly contain millions of parameters and thus require immense resources. BIC provides a data-driven method for determining the optimal trade-off between model complexity and the model's ability to accurately represent the data.

BIC is a likelihood criterion penalized by the model complexity, i.e. the number of parameters in the model. Let $X = \{x_i, i = 1, \ldots, N\}$ be the data set we are modeling and $M = \{M_i, i = 1, \ldots, K\}$ be the candidates for the parametric models. Assuming we maximize the likelihood function $L(X, M_i)$ separately for each model M_i, and if $|M|_i$ is the number of parameters in the model M_i; then the BIC criterion is defined as

$$BIC(M_i) = \log L(X, M_i) - \frac{1}{2}|M|_i \times \log(N)$$

The BIC procedure is to choose the model for which the BIC criterion is maximized. This can be derived as a large-sample version of Bayes procedures for the case of independent, identically distributed observations and linear models [3]. BIC has been widely used for model identification in time series and linear regression. Recently, it has found success in segmentation of speech data and detection of change in speech characteristics [4].

EXPERIMENTAL DESIGN

Note, that the BIC optimization implies an exhaustive search of the model space to find the optimal model. This is impractical for any realistic speech recognition task. Thus, alternative methods must be used for performing a constrained search of the model space. The previous research on model merging [6] provided a greedy framework for model merging which used a heuristic to merge as many states as possible on each iteration. In this work we take a step back from this approach and attempt to iteratively expand only the single most likely merge, using the BIC to determine which merge is further evaluated. This process is shown in Figure 4 and consists of:

1. Create an initial syllable model set with each HMM containing a number of states equivalent to the median syllable duration. This duration was determined by forced alignment of a phone system to a set of development data.
2. Reestimate the model parameters using a standard Viterbi training procedure. This procedure iteratively maximizes the a posteriori probability of the model given the data. During this stage, the model topology is held constant — only the model parameters (means and variances for the Gaussians in the HMM) are adjusted.
3. Perform a state-level forced alignment to determine the likelihood of the model given the data. This is a standard procedure which finds the best supervised path through the models for the given input vectors.
4. Determine which states could be merged according to the Bhattacharyya distance measure described in [6].
5. For each of the possible merges, create a new model set merging only that pair of states.
6. For each of the new model sets, reestimate the model parameters using a standard Viterbi training procedure as described in step 2.
7. For each of the new model sets, perform a state-level alignment as described in step 3.
8. From the forced alignment of each new model set, determine the posterior probability of the model of interest.
9. Apply the BIC measure to each of the new models.
10. Find the set of new models which produce a BIC score higher than the previous iteration and choose the one of that set which maximizes the BIC criteria.
11. Repeat, beginning at step 4. using the chosen model set until no model produces a BIC score higher than the previous iteration. This assumes a convex BIC score surface (i.e. we assume that we are not in a local maxima)

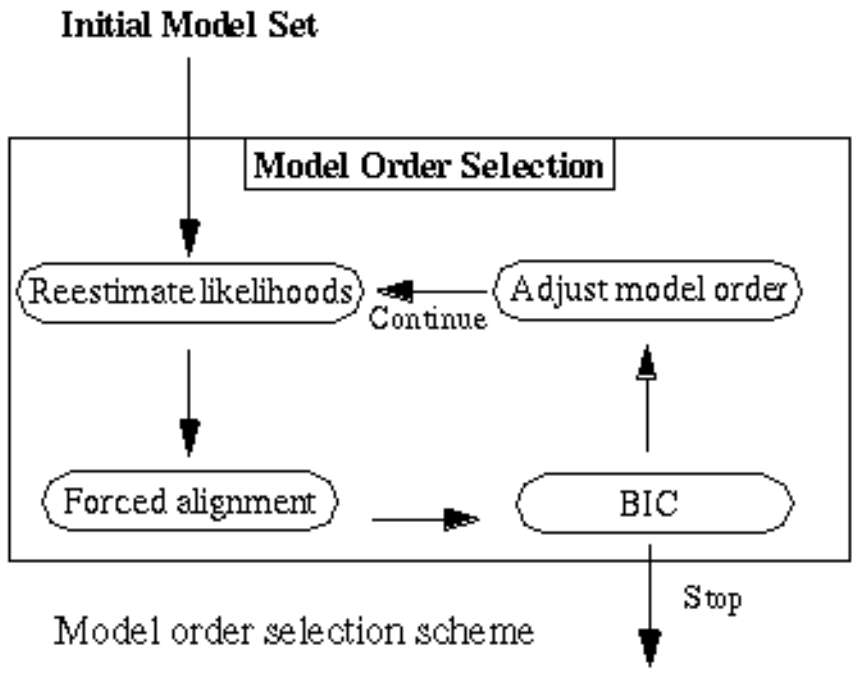

Figure 4. Model selection scheme using the BIC criterion. This process stops when there are no model merges which increase the BIC score.

DATA

A robust and reliable alphadigit system has long been a goal for speech recognition scientists. Recent work on both alphabet and alphadigit systems has taken a focus on resolving the high rates of recognizer confusion for certain word sets. In particular, the E-set (B, C, D, E, G, P, T, V, Z, THREE) and A-set (A, J, K, H, EIGHT). The problems occur mainly because the acoustic differences between the letters of the sets are minimal. For instance, the letters B and D differ primarily in the first 10-20 ms during the consonant portion of the letter. [8]

One reason for the large amount of work on alphabet and alphadigit recognition is the wide availability of high quality corpora dealing with these topics. LDC and OGI, to name a few, have a large repository of corpora suitable for scientific research. The OGI Alphadigit Corpus [9] is a recent release of telephone bandwidth data collected from approximately 3000 volunteers responding to a posting on the USEnet. The subjects were given a list of either 19 or 29 alphanumeric strings to speak. The strings in the lists were each six words long (e.g. E B A 1 Q 2), and there were 1102 separate prompting strings which gave a balanced coverage of vocabulary and contexts.

As mentioned earlier, applying the BIC procedure implies a search process as well as a joint-optimization process across a large number of models. For the work in this project, given that it was somewhat time-limited, we decided to limit the scope of the project to a proof-of-concept framework. Namely, we limited our examination of the model merging to a single model. Previous work [7] showed that, on the alphadigits task, the most frequently confused words were ‘s’ and ‘f’. This is due to the loss of high frequency information in telephone bandwidth data. Thus, we decided to apply the proposed BIC procedures to only the model representing the word ‘f’. We chose 395 utterances from the official training set [10], each having at least one occurrence of the word ‘f’ in it.

EXPERIMENTS

In this project, we were able to run only a single experiment to provide the framework for continuation of the work beyond the course. This experiment intended to find the optimal model size for the model representing the word ‘f’. This was found according to the procedure laid out in an earlier section. Beginning with 30 states, we found that it took only 6 iterations for the model to converge according to the BIC procedure. On average, four possible merges were proposed on each iteration; only one was extended. This left the model with 24 states. Figure 5 shows the movement of the BIC score as the number of iterations increased. Note that the BIC score increases until the 7th iteration where it begins to decrease. Thus, the training procedure could have stopped at that point. We continued to

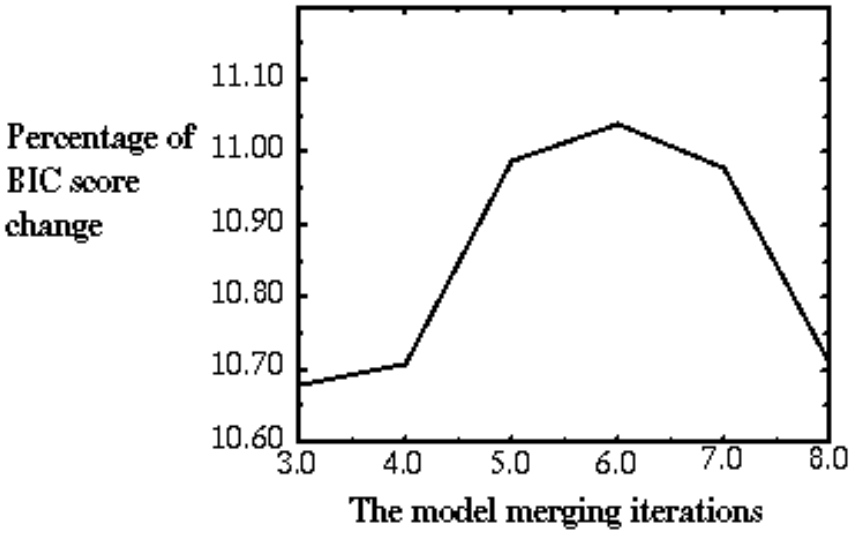

Figure 5. Maximization of the BIC scores in the MERGE_BIC version of the ‘f’ model.

Experiment	log likelihood	BIC score
MERGE_HEUR	-3210.1	-1477516
MERGE_BIC	-2889.5	-1332013

Table 1: Comparison of the results using the heuristically merged model and the model merged according to the BIC scheme. Note that the model derived by the BIC approach has significantly better results in both categories. In fact, it has a better BIC score though it is much more complex than the heuristic model.

an 8th iteration to observe the continued decrease of the BIC score.

With this new model topology in hand, we did a simple comparison to the technique used in previous work [6] which used a heuristic threshold for merging of similar states. In this work, the optimal number of states found for modeling the word ‘f’ was 17 as compared to 24 using the BIC procedure. This model set had been fully trained on a large set of alphadigit data. We then ran two sets of experiments. The first of these (MERGE_HEUR) used the full model set from the previous experiments while the second (MERGE_BIC) replaced the model for ‘f’ in the MERGE_HEUR set with the model for ‘f’ found by the BIC procedure. Both the MERGE_HEUR and MERGE_BIC model sets were given four passes of Viterbi reestimation on the small data set which had been used for BIC training. Both then underwent a forced alignment and a computation of both the average posterior log likelihood and the BIC score. The results for this are summarized in Table 1. Note that the model found by the BIC procedure has a significant increase in the log likelihood as well as in the BIC score. So, while the BIC model is more complex than the heuristically chosen model, the ability of the BIC model to represent the data better than the heuristic model allows it to win out.

SUMMARY AND FUTURE WORK

From the results in this paper, we have shown that there is something to be gained from a structured approach to model selection. In particular, we have shown that using BIC, one can increase the representative power of the HMM for a given task. Though the results given are for only a single model, we believe that application of this approach to a full model set would meet with similar results. However, at this point it is unclear how improvement of one model might effect the performance of the other models.

We believe that moving the BIC procedure described from the single-model optimization to a constrained optimization of the entire model set would have similar results as those shown in this work. We plan to test this hypothesis by carrying out full length experiments where all models in the set are simultaneously optimized. This is a compute intensive process requiring reestimation of literally thousands of possible model sets. Additionally, the metric for success or failure of the procedure should shift from posterior probabilities to the final word error rate. As suggested by one member of the audience, it may be prudent to include the word error rate on a small set as part of the BIC criteria. This would help to ensure that over generalization did not occur.

REFERENCES

1 A. Stolke, S. Omohundro, “Best-first Model Merging for Hidden Markov Model Induction,” *Technical Report TR-94-003*, ICSI, University of California, Berkeley, California, USA, March 1994.

2 G.Doddington, et. al., “Syllable-Based Speech Recognition,” *WS’97 Technical Report*, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland, USA, December 1997.

3 G. Schwarz, “Estimating the Dimension of a Model”, *The Annals of Statistics*, Vol. 6, No. 2, pp 461-464, 1978.

4 S. Chen et. al., “IBM’s LVCSR System for Transcription of Broadcast News used in the 1997 HUB4 English Evaluation,” *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, Feb. 8-11 1998.

5 J. Hamaker, A. Ganapathiraju, J. Picone, and J. Godfrey, “Advances in Alphadigit Recognition Using Syllables,” to appear in the *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, Seattle, Washington, USA, May 1998.

6 J. Hamaker, A. Ganapathiraju and J. Picone, “Information Theoretic Approaches to Model Selection,” *Proceedings of the International Conference on Spoken Language Processing*, vol. 7, pp. 2931-2934, Sydney, Australia, November 1998.

7 J. Hamaker, A. Ganapathiraju and J. Picone, "Syllable-Based Speech Recognition," *Texas Instruments Incorporated*, November 23, 1997.

8 Loizou, Philipos C. and Andreas S. Spanias, “High-Performance Alphabet Recognition,” *IEEE Trans. on Speech and Audio Processing*, vol. 4, no. 6, pp 430-445, Nov. 1996.

9 R.Cole, et. al., “AR.Cole, et. al., “Alphadigit Corpus,” http:// www.cse.ogi.edu/CSLU/corpora/alphadigit, Center for Spoken Language Understanding, Oregon Graduate Institute, 1997.

10. J. Hamaker, et. al., “A Proposal for a Standard Partitioning of the OGI AlphaDigit Corpus,” available at *http://www.isip.msstate.edu/resources/technology/projects/current/speech_recognition/research/syllable/alphadigits/,* Institute for Signal and Information Processing, Mississippi State University, Mississippi State, Mississippi, USA, May 1997.