INFORMATION THEORETIC APPROACHES TO MODEL SELECTION

Jonathan Hamaker, Aravind Ganapathiraju, Joseph Picone

Institute for Signal and Information Processing
Department of Electrical and Computer Engineering
Mississippi State University, Mississippi State, Mississippi 39762
{hamaker, ganapath, picone}@isip.msstate.edu

ABSTRACT

The primary problem in large vocabulary conversational speech recognition (LVCSR) is poor acoustic-level matching due to large variability in pronunciations. There is much to explore about the "quality" of states in an HMM and the inter-relationships between inter-state and intra-state Gaussians used to model speech. Of particular interest is the variable discriminating power of the individual states. The fundamental concept addressed in this paper is to investigate means of exploiting such dependencies through model topology optimization based on the Bayesian Information Criterion (BIC) and the Minimum Description Length (MDL) principle.

1. INTRODUCTION

Hidden Markov Models (HMMs), due to their flexible topology and variance modeling properties, provide the statistical framework for most speech recognition research. Recent work suggests that there exists an optimal structure to an HMM which best models the variability in the speech data [1, 2]. Unfortunately, in traditional LVCSR systems, the initial model choice is empirical (e.g. most state-of-the-art systems use a three-state HMM phone model regardless of the task) — yet it bears a major impact on the ability of the trained model to fit the training data and generalize thereafter.

We conjecture that by effectively exploiting the model topology, we can sift out modalities of the data which were previously being lumped into states of an oversimplified model or spread across states of an overly complex model. Lumping such modalities together in models often results in a recognition system that can use sequences of states and mixtures that never occur in the training data. This leads to decreased performance, and compromises the discrimination capability of the model.

In order to test the above hypothesis we ran a series of experiments on a telephone-quality continuous alphadigit recognition task [3]. Our baseline system used syllable models with the number of states proportional to the average duration of the syllable in the training data and gave a WER of 11.1%. We followed this by setting an upper bound of 20 on the number of states in each syllable model, yielding a 1% absolute decrease in WER. Motivated by the paradigm for phone modeling where all models are of a fixed length, we built a system where all syllable models were 10 states long. This system, however, increased the WER to 12.5%. These experiments demonstrate a strong dependence on the model topology and its ability to represent the data. Specifically, in LVCSR, the inherent variability of speech makes arbitrary definitions of model structure difficult.

The final choice of the model length in the above experiments was, however, still empirical. Choosing the "best" HMM topology would require knowledge of the dynamic structure of the data. This is neither practical nor necessary as there exist efficient data-driven approaches [1, 2] that simultaneously optimize model fit and model complexity. Theoretical frameworks which have been explored include heuristic approaches based on successive state splitting [1], the Minimum Description Length (MDL) [4], and the Bayesian Information Criterion (BIC) [5]. We propose new methods which focus on systematic optimization of the HMM model topology based on the use of MDL and BIC. In this paper, we focus on the specific problem of determining the optimum number of states in the model, which we refer to as the model order.

2. THEORETICAL FOUNDATIONS

The model order decision criteria used in this study, BIC and MDL, are loosely based around the same principle: *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. Both BIC and MDL provide data-driven methods 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

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

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 [6]. 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 [7].

Bayesian inference based on posterior probabilities has an alternative formulation in terms of information-theoretic concepts which is expressed as the MDL principle [4]. The dualism between the two formulations is useful both for a deeper understanding of the underlying principles, as well as for the construction of prior distributions. The maximization of the joint likelihood of the data and the model,

$$P(M, X) = P(M)P\langle X|M\rangle \qquad (2)$$

implicit in Bayesian model inference is equivalent to minimizing its counterpart in the log domain,

$$\log P(M, X) = \log P(M) + \log P\langle X|M\rangle . \qquad (3)$$

From coding theory, $-\log P(E)$ is the least number of bits required to transmit an instance of the discrete event E and guarantees a minimum average code length of a representative message. Accordingly, the terms in the above equation can be interpreted as message or description lengths. $-\log P(M)$ is the description length of the model under the prior distribution; $-\log P\langle X|M\rangle$ corresponds to a description of the data X using the model M on which the number of model parameters is based. The negative logarithm of the joint probability can therefore be interpreted as the total description length of model and data. Thus, inference or estimation by MDL is equivalent to and a useful alternative to conceptualization of posterior probability maximization [5]

3. EXPERIMENTS

The experiments conducted in this work are motivated by encouraging results obtained by empirically determining a model order based on durations computed via Viterbi alignments. All experiments used the syllable as the basic unit of recognition owing to its longer temporal context. Results from alphadigit experiments (see Table 1) clearly show that model topology is an important factor in achieving models which best represent the data and highlight a need for a more principled approach. We begin by applying information theoretic model selection techniques to a relatively simple task of alphadigit recognition and then extending these results to experiments on the much more complex Switchboard (SWB) task.

Selection of number of states	WER	Ins	Del	Sub
duration / 2	11.1%	0.3%	0.6%	10.2%
upper limit of 20	**10.1%**	**0.6%**	**1.3%**	**8.3%**
each with 10	12.5%	0.8%	0.6%	11.0%

Table 1: Recognition performance for three syllable Alphadigit systems using heuristic approaches to model order determination.

3.1. Alphadigits

The OGI Alphadigit corpus [8] is a telephone database collected using a T1 interface with over 3000 subjects reading a list of either 19 or 29 alphanumeric strings (e.g., "8 h a 8 b h"). Its acoustic properties are similar to the well-known SWITCHBOARD corpus described below. The 1102 unique strings comprising the prompted utterances were each six words long, and each list was designed to balance the phonetic context of all word pairs.

Since there had been no published results on this data, there existed no standard partitioning of the database for common evaluations. We developed such a partitioning by splitting the data along gender lines. In addition we defined a 3000 utterance evaluation set from the test data, on which all our results are quoted. This test set definition is publicly available [9].

3.2. SWITCHBOARD

The SWITCHBOARD Corpus (SWB) [10] is currently the standard benchmark for telephone-based conversational speech applications. It contains 2430 conversations averaging 6 minutes in length; i.e. over 240 hours of recorded speech, and about 3 million words of text, spoken by over 500 speakers of both sexes from every major dialect of American English.

In this work, we are using a new segmentation of SWB [11] which seeks to balance the trade-off between linguistically and acoustically motivated segmentations. The new segments ensure ample context for acoustic as well as language modeling applications. Experiments using this segmentation have already proven successful. By simply reestimating models produced at WS97 [12] on a small subset of the total training set, a 2% absolute decrease in word-error rate has been achieved.

3.3. Syllable-Based Recognition

While context-dependent (CD) phones have been the dominant method of modeling speech acoustics, the large number of frequently occurring acoustic patterns make them a relatively inefficient decompositional unit. A CD phone spans an extremely short time-interval, and therefore is not amenable to integration of spectral and temporal dependencies. For applications such as SWB, focus has shifted to a larger acoustic context due to poor performance of phone-based approaches. The syllable is a particularly appealing acoustic unit due to its close connection to articulation, its integration of some co-articulation phenomena, and its potential for a compact representation of conversational speech.

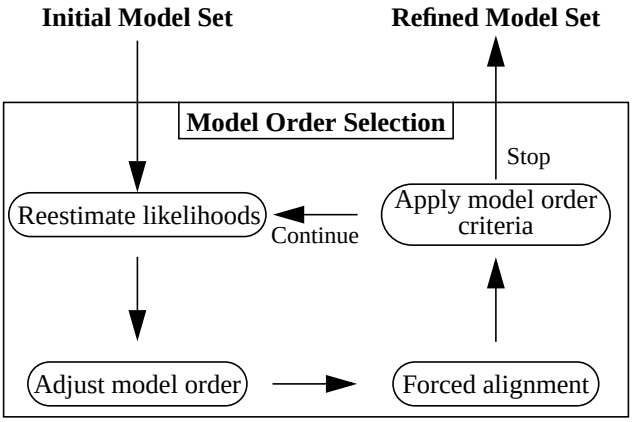

Figure 1. Model order selection scheme.

The use of an acoustic unit with a longer duration also makes it possible to simultaneously exploit temporal and spectral variations. Parameter trajectories and multi-path HMMs are examples of techniques that can exploit the longer acoustic context, but have had marginal impact on CD phone-based systems. Our recent experiments with syllable acoustic models on the SWB corpus show performance comparable to CD phone-based systems [12].

3.4. Experimental Procedure

For both the Alphadigit and SWB tasks, we propose the fairly simple iterative model selection scheme, shown in Figure 1, similar to the commonly used flat-start procedure for training HMMs. We use a small portion of the training set to build model topologies, beginning with a standard left-to-right model topology with the model order proportional to duration based on a forced alignment. After a small number of reestimation passes, we use the model topology decision measures to score each model based on the likelihoods and model complexity and adjust the topology accordingly.

When adjusting each model we have three options —

add a state, remove a state, or remove this model from further consideration. When adding a state, we use perturbed Gaussian values for the new state. When removing a state, we reestimate the probabilities by assuming that all data mapped to the deleted state is now represented by its surrounding states, proportional to the transition probabilities. For each new configuration of the model, we compute the likelihood of the data given the new model using a recognition run on the segments previously aligned with this model. This likelihood is evaluated using the model selection criteria and the process for each model is stopped when a maxima in the model complexity score is reached.

4. CONCLUSIONS

While HMMs have gained vast popularity for speech recognition applications, there is still much to learn about using them to their fullest potential. We have shown through initial experimentation that recognition performance using HMMs is highly sensitive to the model order used, and thus, we should further explore methods for choosing a model set which best represents the data of interest. However, we must also be cognizant of the resources necessary for following an iterative scheme on a dataset of the size of an LVCSR application.

In this work we have presented a method which balances these two needs using the well-founded BIC and MDL principles. In future experiments we plan to examine minimization of global complexity. One might reason that minimization of each individual model implies global minimization, but we believe that interaction between models may have a causal effect of one model's order on the other's.

5. REFERENCES

1. H. Singer, M. Ostendorf, "Maximum Likelihood Successive State Splitting," *Proceedings of IEEE ICASSP*, Atlanta, Georgia, USA, Vol. II, pp. 601-604, May 1996.

2. P. Lockwood, M. Blanchet, "An Algorithm for the Dynamic Inference of Hidden Markov Models (DIHMM)," *Proceedings of IEEE ICASSP*, Minneapolis, Minnesota, USA, Vol. II, pp. 251-254, April 1993.

3. J. Hamaker, A. Ganapathiraju, J. Picone, and J. Godfrey, "Advances in Alphadigit Recognition Using Syllables," *Proceedings of IEEE ICASSP*, Seattle, Washington, USA, Vol. I, pp. 421, May 1998.

4. J. Rissanen, "Minimum Description Length Principle," *Encyclopedia of Statistical Sciences*, Vol. 5, pp. 523-527, 1985.

5. A. Stolke, "Bayesian Learning of Probabilistic Language Models," Ph.D. Dissertation, University of California, Berkeley, 1994.

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

7. 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.

8. http://www.cse.ogi.edu/CSLU/corpora/alphadigit

9. 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/*

10. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proceedings of IEEE ICASSP*, San Francisco, California, USA, Vol. I, pp. 517-520, March 1992.

11. N. Deshmukh, et. al., "Resegmentation of Switchboard", to be presented at Fifth ICSLP, Sydney, Australia, December 1998.

12. 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.