CLASSIFICATION WITH FREE ENERGY AT RAISED TEMPERATURES

Rita Singh
Carnegie Mellon University
Pittsburgh, PA

Manfred Warmuth
University of California
Santa Cruz, CA

Bhiksha Raj
Mitsubishi Electric Research Labs
Cambridge, MA

Paul Lamere
Sun Microsystems Inc
Burlington, MA

ABSTRACT

In this paper we describe a generalized classification method for HMM-based speech recognition systems, that uses free energy as a discriminant function rather than conventional probabilities. The discriminant function incorporates a single adjustable temperature parameter T. The computation of free energy can be motivated using entropy regularization, where entropy grows monotonically with temperature. In the resulting generalized classification scheme, the values T=0 and T=1 give the conventional Viterbi and forward algorithms, respectively, as special cases. We show experimentally that if the test data are mismatched with the classifier, classification at temperatures higher than one can lead to signficant improvements in recognition performance.

INTRODUCTION

HMM-based speech recognition systems model sound classes by HMMs

- Model parameters are learned from training data
- If test data are similar (matched) to training data MAP classification of the test data is effective
- If they are mismatched, it is ineffective

Conventional solution: Modify either data or model parameters to reduce mismatch

This paper: Modify the classification rule

- Use the *free energy* of the HMMs as the objective function to minimize
  - Free energy is a function of temperature T
  - Classification at T=1 is MAP classification
  - Classification at T=0 is Viterbi decoding
- Classification at T>1 results in significant improvements on mismatched test data

MINIMUM FREE ENERGY CLASSIFICATION WITH HMMS

Free Energy: Consider a system at temperature T that

- Has energy H_s when in configuration s
- Has probability P_s of being in configuration s

The Helmholtz free energy $F_T(P)$ of the system is

$$F_T(P) = \sum_s P_s H_s + T \sum_s P_s \log(P_s)$$

The P_s values at thermal equilibrium minimize $F_T(P)$:

$$F_T = \min_{\{P_s\}} \left\{ \sum_s P_s H_s + T \sum_s P_s \log(P_s) \right\}$$

The optimal P_s is given by the Gibbs distribution

$$P_s = \frac{1}{Z} \exp\left(\frac{-H_s}{T} \right)$$

Z is a normalizing term. The equilibrium free energy is

$$F_T = -T \log \sum_s \exp\left(-\frac{H_s}{T} \right)$$

Free Energy in HMMs: Let $\boldsymbol{s}$ be a state sequence through the HMM for class C. $\boldsymbol{s}$ is equivalent to the configuration of the HMM. Let $\boldsymbol{x}$ be a data sequence. The energy of the configuration $\boldsymbol{s}$ can be defined as

$$H_C(x,s) = -\log(P(C)) - \log(P(x, s \mid C))$$

The free energy of the HMM for C can be defined as

$$F_T(x,C) = -T \log\left(\sum_s \exp\left(\frac{-H_C(x,s)}{T} \right) \right)$$

$$= -\log(P(C)) - \log\left(\left(\sum_s P(x, s \mid C)^{\frac{1}{T}} \right)^T \right)$$

At T=1, this is the negative log of the joint probability of $\boldsymbol{x}$ and C. At T=0, this is the negative log of the joint probability of C, $\boldsymbol{x}$ and the best state sequence for C.

Minimum Free Energy Classification Rule:

$$c(x) = \text{argmin}_C \{ F_T(x,C) \}$$

At T=1 this reduces to MAP classification.
At T=0 this reduces to Viterbi decoding.
No probabilistic interpretation at other temperatures.

IMPLEMENTATION IN HMM-BASED SPEECH RECOGNITION SYSTEMS

Minimum free energy classification is implemented in HMMs by a modification of the forward algorithm

Compute partial free energy for all state sequences terminating at s of the HMM for class C, at time t as

$$\alpha(s,t,C) = -T \log\left(\sum_{s'} \left(e^{-\alpha(s',t-1,C)} a(s,s') P(x_t \mid s) \right)^{\frac{1}{T}} \right)$$

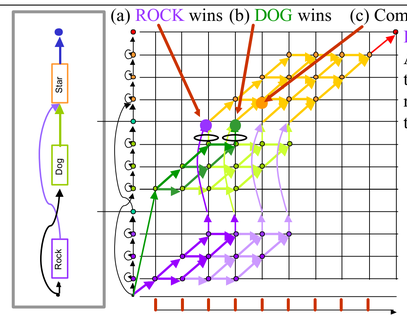
Where x_t is t^{th} observation in x, $a(s,s')$ is the transition probability from s to s'. Initialize recursion as

$$\alpha(s,1,C) = -\log(P(C)) - \log(\pi(s)) - \log(P(x_1 \mid s))$$

$\pi(s)$ is the *a priori* probability of s. The overall free energy of C for an observation of N vectors is

$$F_T(x,C) = -T \log \sum_s e^{\frac{-\alpha(s,N,C)}{T}}$$

Continuous speech recognizers perform recognition on compressed word graphs. Free energy cannot be computed for individual word sequences on such graphs. An engineering solution is proposed:



Bushderby: Every state in the HMM is given a colour

- States with incoming transitions from multiple words are coloured red. Conventional Viterbi logic is applied at these states
- All other states are coloured green. Bushderby scores are computed at these states
- All states maintain pointers to the best incoming transition

Bushderby is an approximate implementation of minimum free energy classification. It potentially results in an anomaly where the score computed for a hypothesis incorporates scores from competing hypotheses. This is illustrated in the above figure.

EXPERIMENTAL RESULTS

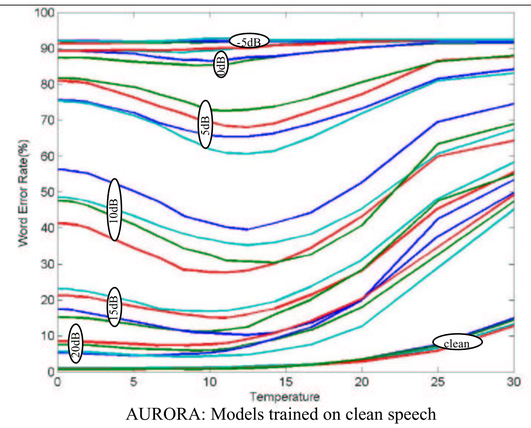
The Bushderby algorithm was implemented in the Sphinx4 speech recognition system. Sphinx4 was used for all the experiments.

Experiments conducted on two databases to evaluate

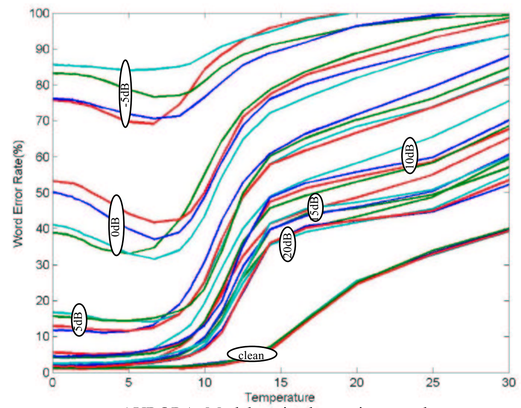
- Effect of increasing temperature on recognition of mismatched test data
- Effect of increasing degrees of mismatch
- Comparison to the closest one-parameter adjustment method that can be incorporated in Viterbi decoding, *i.e.* variance scaling

Databases used:

1. Aurora-2 database: 8Khz speech, consisting of digits sequences, digitally corrupted by 4 types of noises to various SNRs
   - Models trained on clean speech
   - Multi-style models trained on noisy speech
2. TID database provided by Telefonica Investigacion y Desarrollo: 8 Khz speech, consisting of Spanish digits sequences, corrupted by 4 types of noises to various SNRs
   - Models trained on clean speech



AURORA: Models trained on clean speech



AURORA: Models trained on noisy speech

WER(%) obtained by scaling Gaussian Variances on TID data. Results are on clean speech and speech corrupted to 5dB by babble noise

Scaling Factor	1.0	1.1	1.2	1.3
Clean	15.9	15.8	15.8	16.3
Babble 5dB	62.1	62.2	63.0	65.0

Bushderby WER on speech corrupted by babble to 5dB: 55.4%
More extensive results available in paper.

CONCLUSIONS

WER on noisy test data is best when T is between 10 and 12, when the models are trained on clean data.

- Minor degradation on clean data

Even on mixed style trained models, recognition error improves at high T for highly noisy test data.
Similar improvements cannot be achieved by simply increasing the variances of the Gaussians.