MITSUBISHI ELECTRIC RESEARCH LABORATORIES
http://www.merl.com

Classification with Free Energy at Raised Temperatures

Rita Singh, Manfred K. Warmuth, Bhiksha Raj, Paul Lamere

TR2003-22 December 2003

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 an entropy regularization, where the entropy grows monotonically with the temperature. In the resulting generalized classification scheme, teh values of 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 significant improvements in recognition performance. The temperature parameter is far more effective in improving performance on mismatched data than a variance scaling factor, which is another apparent single adjustable parameter than has a very similar analytical form.

Eurospeech 2003

CLASSIFICATION WITH FREE ENERGY AT RAISED TEMPERATURES

[1]Rita Singh, [2]Manfred K. Warmuth, [3]Bhiksha Raj, [4]Paul Lamere

[1]Carnegie Mellon University, [2]University of California Santa Cruz,
[3]Mitsubishi Electric Research Laboratories, [4]Sun Microsystems, U.S.A.

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 an entropy regularization, where the entropy grows monotonically with the temperature. In the resulting generalized classification scheme, the values of $T = 0$ 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 significant improvements in recognition performance. The temperature parameter is far more effective in improving performance on mismatched data than a variance scaling factor, which is another apparent single adjustable parameter that has a very similar analytical form.

1. INTRODUCTION

Speech recognition systems typically model sound classes with continuous density hidden Markov models (HMMs), the parameters of which are learned from some training data. When the test data are similar, or *matched*, to the training data, the models closely approximate the true distribution of the test data, and maximum *a posteriori* probability (MAP) classification using them can be expected to approach true Bayesian classification. Often, however, the test data vary significantly from the training data for reasons such as variations in the degree of spontaneity, environmental noise, recording conditions, *etc*. As a result the HMMs no longer represent the true distributions of the test data, and recognition performance is poorer than that on matched data.

There has been a significant body of work in speech recognition in compensating for mismatch in various ways. Compensation is typically done either by modifying the data, such that they are better represented by the HMMs, or by modifying the HMM parameters so that they better match the test data [1]. Often these methods require explicit models of the phenomena that cause the mismatch, and become ineffective when the model is inappropriate. The methods also require significant computational effort in addition to that needed for recognition.

In this paper we take a completely different approach to the problem of classifying data that have been rendered statistically different from the training data by unknown phenomena. In MAP classification, the discriminant function that is maximized with respect to the classes is the joint probability of the class and the test data. When class distributions are modelled by HMMs, this can be expressed as the sum over all state sequences, of the joint probability of the class, the state sequence, and the data. We observe that if the state sequence were to be interpreted as the *configuration* of the HMM, and the negative log of the joint probability of the class, the state sequence, and the data as the *energy* of the configuration, then the MAP discriminant function is identical to the exponentiated negative free energy [2] of the HMM at a temperature $T = 1$. Consequently, MAP classification is equivalent to minimizing the free energy of the classifier with respect to the class at $T = 1$. In speech recognition systems that must work off complicated language graphs, the classes are word sequences, and recognition is performed by Viterbi decoding, which maximizes the joint probability of the class, the data, *and* the state sequence, with respect to the class. Viterbi decoding now becomes the minimization of the free energy at $T = 0$.

In this paper we discuss classification with free energy at temperatures other than 0 or 1. Now the discriminant function used for classification has no immediate probabilistic interpretation. When the models are representative of the distribution of the test data, the optimal value of T can be expected to be close to 1 (*i.e.* MAP). When the test data and the models are mismatched, the optimal value of T is typically much higher. We further show how minimum free energy classification can be implemented for HMM-based classifiers by modifying the forward algorithm. The conventional forward and Viterbi algorithms are special cases of this modified forward algorithm.

Experiments run on multiple databases show that classification with free energy at increased temperatures can result in large improvements in recognition performance on mismatched data, over conventional MAP recognition or Viterbi decoding. Our experiments also show that this is more effective than the closest other one-parameter adjustment method, which scales the variances of the Gaussians by a constant factor. The rest of the paper is arranged as follows: In the Section 2, we describe minimum free energy classification. In Section 3 we explain how it can be integrated into the search in an HMM-based recognition system. In Section 4 we present experimental results on two different databases, and discuss their implications in Section 5.

2. MINIMUM FREE ENERGY CLASSIFICATION WITH HMMS

In statistical pattern classifiers where the classes are modelled by HMMs, a data sequence $\boldsymbol{x}$ is associated with a class $c(\boldsymbol{x})$ by one of two rules: the maximum *a posteriori* probability (MAP) classification rule, or Viterbi decoding. In MAP classification,

$$c(\boldsymbol{x}) = \arg\max_C \left\{ \log(P(C)) + \log\left(\sum_{\boldsymbol{s}} P(\boldsymbol{x}, \boldsymbol{s} | C) \right) \right\} \quad (1)$$

where $P(C)$ represents the *a priori* probability of class C and $\boldsymbol{s}$ represents a state sequence through the HMM for C. In Viterbi decoding, classification is performed by the rule

$$c(\boldsymbol{x}) = \arg\max_C \{ \log(P(C)) + \log(\max_{\boldsymbol{s}} \{ P(\boldsymbol{x}, \boldsymbol{s} | C) \}) \} \quad (2)$$

While MAP classification considers all state sequences through the HMM, Viterbi decoding is based only on the probability of the most likely state sequence. In our notation we suppress the

parameters of the HMM, since they are fixed during classification.

We now define a third classification rule based on the thermodynamic measure of free energy, that generalizes both above rules. Assume we have a system that has energy $H_{\boldsymbol{s}}$ when it is in a configuration $\boldsymbol{s}$. If P_s is the probability of configuration $\boldsymbol{s}$, and the system is at temperature T, then the Helmholtz free energy $F_T(P)$ is defined as [2]:

$$F_T(P) = \sum_{\boldsymbol{s}} P_{\boldsymbol{s}} H_{\boldsymbol{s}} + T \sum_{\boldsymbol{s}} P_{\boldsymbol{s}} \log(P_{\boldsymbol{s}}) \tag{3}$$

This is the difference between the average energy of the system and T times the entropy of P_s. When the system is at thermal equilibrium at T, the probability values P_s are such that $F_T(P)$ is minimized. Thus the equilibrium free energy F_T is given by

$$F_T = \min_{\{P_s\}} \left\{ \sum_{\boldsymbol{s}} P_{\boldsymbol{s}} H_{\boldsymbol{s}} + T \sum_{\boldsymbol{s}} P_{\boldsymbol{s}} \log(P_{\boldsymbol{s}}) \right\} \tag{4}$$

The optimal values of P_s are given by the Gibbs distribution

$$P_s = \frac{1}{Z} \exp\left(\frac{-H_s}{T}\right) \tag{5}$$

where Z is a normalizing term. Introducing the Gibbs distribution into Eq. 3 gives the following expression for free energy:

$$F_T = -T \log\left(\sum_{\boldsymbol{s}} \exp\left(\frac{-H_{\boldsymbol{s}}}{T}\right) \right) \tag{6}$$

Let $\boldsymbol{s}$ now represent a state sequence through the HMM for class C. Let the energy of the HMM for the state sequence $\boldsymbol{s}$ be

$$H_C(\boldsymbol{x}, \boldsymbol{s}) = -\log(P(C)) - \log(P(\boldsymbol{x}, \boldsymbol{s} | C)) \tag{7}$$

The free energy of the HMM can be defined analogously to Eq 6:

$$F_T(\boldsymbol{x}, C) = -T \log\left(\sum_{\boldsymbol{s}} \exp\left(\frac{-H_C(\boldsymbol{x}, \boldsymbol{s})}{T}\right) \right) \tag{8}$$

$$= -\log(P(C)) - \log\left(\left(\sum_{\boldsymbol{s}} (P(\boldsymbol{x}, \boldsymbol{s} | C))^{\frac{1}{T}} \right)^T \right) \tag{9}$$

The second term in Equation 9 is simply the log of the $1/T$ norm of the $P(\boldsymbol{x}, \boldsymbol{s} | C)$ values. At $T = 1$, it is the log of the sum of all the $P(\boldsymbol{x}, \boldsymbol{s} | C)$ values and equals $\log P(\boldsymbol{x} | C)$. At $T = 0$, it is simply the log of the largest $P(\boldsymbol{x}, \boldsymbol{s} | C)$.

We define the minimum free energy classification rule as

$$c(\boldsymbol{x}) = \arg\min_C \{F_T(\boldsymbol{x}, C)\} \tag{10}$$

It is easy to see that Equation 10 is identical to Equation 1 when $T = 1$, and to Equation 2 when $T = 0$. Thus, both MAP classification and Viterbi decoding are special instances of minimum free energy classification. In these cases the free energy can be related to the probability of the data. More generally however, $F_T(\boldsymbol{x}, C)$ does not have a simple probabilistic interpretation.

We note that free energy has often been used as an optimization criterion in annealing methods for estimating parameters of statistical models [3]. In these methods the temperature of the system is initially set to a high value and slowly decreased to 0, to arrive at estimates of the parameters close to the global optimum. In contrast, in minimum free energy classification the parameters of the HMMs obtained during training remain unchanged. Classification is done at a fixed temperature.

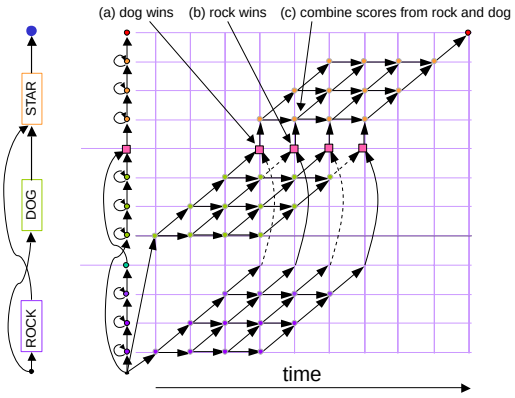

Figure 1. The graph to the left is a collapsed word graph for two word sequences, "ROCK STAR" and "DOG STAR". The grid shows the trellis derived from the HMM for the word graph, and the Bushderby implementation of minimum free energy classification on it. The four entry points into the trellis for STAR are red nodes (shown as boxes). All other nodes are green. At red node (a) the path from ROCK scores higher, and only this score is retained. At red node (b) the score from DOG is higher and is retained. The score at green node (c) combines contributions from both (a) and (b), and thus represents a combination of the scores for DOG and ROCK.

3. IMPLEMENTATION IN HMM-BASED SPEECH RECOGNITION SYSTEMS

In HMMs, direct computation of Equation 9 is impractical, since the $1/T$ norm in the right hand side requires summation over all state sequences through the HMM. However, the free energy can be computed efficiently by a modification of the forward algorithm. We compute the partial free energy for all state sequences terminating at state s of the HMM for class C, at any time t as:

$$\alpha(s, t, C) = -T \log\left(\sum_{s'} (e^{-\alpha(s', t, C)} a(s, s') P(x_t | s))^{\frac{1}{T}} \right) \tag{11}$$

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

where $\pi(s)$ is the initial probability of s, $a(s, s')$ is the probability of transiting from s to s', and $P(x_t | s)$ is the state output probability of generating x_t from s. The overall free energy is given by

$$F_T(\boldsymbol{x}, C) = -T \log\left(\sum_{s} e^{\frac{-\alpha(s, N, C)}{T}} \right) \tag{12}$$

where N is the total number of data points in $\boldsymbol{x}$.

In a speech recognizer, each class is a word sequence. Direct implementation of Equation 10 in a recognizer would require a separate HMM for every word sequence considered as a hypothesis. In practice, recognition is performed on compressed word graphs as shown in the example in Figure 1. Free energy cannot be directly computed on such graphs, unless the underlying word graph is a tree. We propose an engineering approximation to the computation of the free energy. Every state in the HMM for the word graph is given a "colour", that is either green or red. States that have incoming transitions from multiple words are coloured red. All other states are coloured green. The free energy for green states is computed using Equation 11. Red states compute

the free energy of the best incoming transition. All states retain a pointer to the source of the best incoming transition. This is illustrated by Figure 1. We call this specific implementation of free energy classification using coloured nodes as *Bushderby*, since it combines scores of entire bushes of incoming paths at green nodes, instead of retaining only the best incoming score, as in Viterbi decoding. This approximate implementation potentially results in an anomaly in the computation of scores for states within words that can be reached from multiple words: scores for the green nodes within such words do not necessarily represent any single partial word hypothesis, but can actually be a combination of the scores for all potentially competing partial hypothesis that arrive at that word. This is also illustrated in Figure 1. The anomaly is avoided only at $T = 0$ (*i.e.* Viterbi decoding). However, as shall see in the next section, classification at $T > 0$ can still lead to significantly improved error rates.

4. EXPERIMENTAL RESULTS

The performance of minimum free energy classification is potentially related to several factors such as the temperature, the distribution of the test data, and the entropy of the statistical models themselves. A complete experimental analysis of the method is however infeasible in the limited space available to us in this paper. We therefore restrict ourselves to experiments pertaining to the following issues: a) the effect of increasing temperature on recognition of mismatched test data, b) the effect of increasing degrees of mismatch, and c) comparison to the closest one-parameter adjustment method that can be incorporated in conventional Viterbi decoding, *i.e.* variance scaling.

All experiments were performed using the SPHINX-4 open-source system (cmusphinx.sourceforge.net), in which minimum free energy classification has been implemented through Bushderby. In all cases word-list grammars were used. Experiments were performed on AURORA-2 and TID databases. AURORA-2 consists of 8Khz sampled speech, derived from the TIDigits database. The training and test utterances are continuous

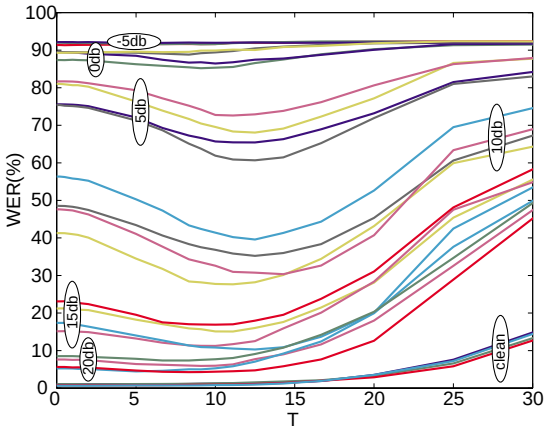

Figure 2. %WER on test data from the Aurora database. The lines are in groups of four. Each group of four curves corresponds to a different SNR, shown within the ellipses. From bottom to top, the SNRs are clean speech, 20db, 15db, 10db, 5db, 0db and -5db. The optimal value of T is around 10 at most SNRs.

sequences of digits. We used 8440 utterances of clean speech to train the models. The test set, labelled *testa* in the database, comprises 28028 utterances (14 hours of speech recordings) that are further subdivided into 7 sets, 6 of which are each corrupted by 4 noise types to an SNR of -5db, 0db, 5db, 10db, 15db or 20db. The 7th set consists of 4 subsets of clean speech. Acoustic models for this experiment were 3-state Bakis topology triphone HMMs with 500 tied states, each modelled by a mixture of 8 Gaussians.

Figure 2 shows a plot of word error rates (WER%) obtained with varying T values. Although it is not clear from the figure, the curves are almost flat between $T = 0$ and $T = 1$. Recall that $T = 0$ represents Viterbi decoding, and $T = 1$ represents MAP classification using the conventional forward algorithm. Figure 2 confirms the common belief that the Viterbi decoding is as effective as recognition using the forward algorithm (while having

	Babble Noise, WER%							**Music, WER%**						
SNR \ T	**0**	**1**	**2**	**5**	**6.67**	**10**	**20**	**0**	**1**	**2**	**5**	**6.67**	**10**	**20**
clean	15.8	15.9	16.0	17.2	17.9	22.7	44.5							
20db	22.0	22.1	22.7	24.5	26.2	41.7	53.7	14.9	15.1	15.0	15.1	16.6	28.1	45.6
15db	29.7	29.8	29.7	29.9	32.1	48.3	61.7	16.3	16.2	16.0	16.0	16.9	30.7	47.9
10db	41.5	41.3	41.0	39.2	41.0	55.8	71.3	19.7	19.4	19.0	18.6	18.9	33.0	50.8
5db	62.5	62.1	61.1	56.0	55.4	67.3	85.1	28.9	28.6	27.8	25.2	25.3	36.6	56.7
0db	85.8	85.3	84.5	79.4	77.5	84.2	98.8	48.0	47.5	46.1	41.8	39.8	48.2	67.3
	Traffic Noise, WER%							**Subway Noise, WER%**						
20db	14.3	14.3	14.1	14.2	14.6	20.7	42.7	16.1	16.2	16.3	17.1	18.3	31.0	46.2
15db	16.1	15.9	16.0	15.9	16.6	22.5	44.1	18.5	18.5	18.4	18.9	20.4	34.2	50.6
10db	24.4	24.4	24.3	24.9	26.1	34.1	47.5	25.1	24.8	24.3	23.2	24.4	37.1	57.0
5db	36.2	36.3	36.3	36.9	37.7	42.1	53.1	39.7	39.2	38.4	35.9	35.7	45.7	67.7
0db	47.4	47.1	46.4	44.9	44.6	48.9	64.0	67.2	66.5	64.8	59.0	57.7	66.0	83.9

Table 1: Bushderby results on TID: Recognition performance on test data corrupted with different noises to different SNRs.

Table 2: WER(%) obtained by scaling Gaussian variances

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

many practical advantages). However, surprisingly, we observe that as SNR decreases, better recognition is obtained at $T > 1$. The optimal T for low SNR data lies in the range 8-12. Finally, we observe that as the mismatch increases, the relative improvement between the performance at the optimal T and the performance of Viterbi decoding initially increases. At SNR 10dB, relative improvements as large as 35% are obtained. But as the mismatch continues to increase the relative improvements decrease again. At very high mismatch, increasing T does not improve performance.

A second set of experiments were performed using a Spanish telephone speech database, consisting of 8KHz sampled speech, provided by Telefonicá Investigación y Desarrollo (TID) to CMU for internal evaluations of robustness algorithms. For experiments with this database, continuous density 8 Gaussian/state HMMs with 500 tied states were trained from 3500 utterances of clean telephone recordings. The test data consisted of 1728 utterances (1 hour) of clean telephone recordings. For each experiment, the entire test set was corrupted to various SNRs by traffic noise, music, babble recorded in a bar, and noise recordings from a subway. These are real noise recordings provided separately as an annexe to the TID database. Table 1 shows the WERs obtained at various T values on test sets corrupted with different noises to different SNRs. Again, we observe that the best performance is for values of T larger than 1, except under matched conditions. Here, the optimal values of T are observed to lie in the range 2-10. On this test set the SNRs were not lowered to the level where Bushderby search became ineffective.

Finally Table 2 shows recognition performance obtained on the TID database for clean speech and speech corrupted by babble noise to 5dB, when the variance of the Gaussians are scaled. Increasing variances does not improve recognition performance.

5. DISCUSSION

As expected, when the test data are matched, the best recognition performance is observed at temperatures close to 1. When the test data are mismatched, the best recognition performance is obtained at raised temperatures, in the range 2-12. The difference between Viterbi decoding and Bushderby search at raised temperatures can be dramatic on mismatched data, where relative improvements in recognition error are observed to be as large as 35% (on the AURORA-2 database). The relative improvements are dependent on the degree of mismatch. As the degree of mismatch increases, relative improvements achieved by raising the temperature initially increase and then decrease again. In other results not reported here for lack of space, it was also observed that relative improvements due to raising the temperature are related to the entropy of the HMMs themselves. Lower improvements were observed on HMMs with higher entropy.

It seems difficult to derive the analytical characterizations of the relationship between T and classification error that might explain the observed behaviour. Nevertheless, there are some theoretical results that support the use of the free energy for classification. For the binary case it is was shown [4] that abstaining from classification when the free energy of both classes are close, increases robustness against overfitting. Regularization with an entropy or a relative entropy (as done in Equation 4) has been used earlier to derive updates for on-line learning algorithm (see *e.g.* [5] and references therein). The updated distribution, which is identical to the Gibbs distribution of Equation 5, is the solution to the minimization problem. The free energy is then used as a potential function for the amortized analyses of the algorithm for the purpose of proving regret bounds that hold for arbitrary sequences of examples.

Increasing the temperature is equivalent to increasing the entropy of the Gibbs distribution, *i.e.* the Gibbs entropy, for the state sequences. An alternative to using temperature as an adjustable parameter would be to use the Gibbs entropy as a parameter. This option remains to be explored.

The Bushderby implementation of minimum free energy classification in speech recognition systems results in the computation of anomalous scores that combine scores from several competing hypotheses as explained in Figure 1. It is unclear whether this anomalous combination of scores hurts performance, as compared to what might be obtained with a theoretically correct implementation where each word sequence has its own HMM and no merging of scores happens.

Finally, we note that free energy has often been used as a criterion for *training* statistical models in annealing methods [3]. In these methods, the model parameters are estimated by gradually cooling the system, to avoid local optima and learn the best parameters for a given training data. Once model parameters are learned, classification usually proceeds using conventional MAP classification rules (*i.e.* $T = 1$). The free energy based classification proposed in this paper presents a counterpart to these approaches. Independently of how the models were trained, we increase the temperature during *classification* to obtain improved error rates on mismatched data.

ACKNOWLEDGEMENTS

Manfred Warmuth acknowledges the support of NSF grant CCR-9821087. Rita Singh was sponsored by the Space and Naval Warfare Systems Center, San Diego, under Grant No. N66001-99-1-8905. The content of the information in this paper does not necessarily reflect the position or the policy of the US Government, and no official endorsement should be inferred.

REFERENCES

1. Singh, R., Raj, B. & Stern, R. M. "Model Compensation and Matched Condition Methods for Robust Speech Recognition," In: *Noise Reduction in Speech Applications (Electrical Engg, series)*, Ed. G. Davis, pp.247-278, CRC Press, USA, 2002.

2. Hertz, J., Krogh, A. & Palmer, R. G., *Introduction to the Theory of Neural Computation*. Lecture Notes, Vol. 1, Santa Fe Institute Studies in the Science of Complexity. Addison Wesley, Redwood City, CA 1991.

3. Rose, K., "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems," *Proc. of the IEEE*, Vol. 86:11, pp.2210-2239, 1998.

4. Freund, Y., Mansour, Y., and Schapire, R. E. "Why Averaging Classifiers Can Protect Against Overfitting". *Proc. Eighth International Workshop on Artificial Intelligence and Statistics*, 2001.

5. Kivinen, J. & Warmuth, M. K. "Averaging Expert Predictions," *Proc. EUROCOLT 1999*. Springer-Verlag, pp.153-167, March 1999.