Audio Classification with Thermodynamic Criteria

Rita Singh
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213, USA
Email: rsingh@cs.cmu.edu

***Abstract*—Detecting sound events in audio recordings is a challenging problem. A detector must be trained for each sound to be classified. However, the recordings of the examples used to train the detector rarely match the conditions found in the test audio to be classified. If the event detection problem is posed as one of Bayes classification, the problem may be viewed as one of mismatch between the true distribution of the data and that represented by the classifier. The Bayes classification rule results in suboptimal performance under such mismatch, and a modified classification rule is required. Alternately stated, the classification rule must optimize a different objective criterion than the Bayes error rate computed from the training distributions. The use of entropy as an optimization criterion for various classification tasks has been well established in the literature. In this paper we show that free-energy, a thermodynamic concept directly related to entropy, can also be used as an objective criterion for classification in such scenarios. We demonstrate with examples on classification with HMMs that minimization of free-energy is an effective criterion for classification under conditions of mismatch.**

I. Introduction

The relationship between the *thermodynamic* principle of entropy and the *information theoretic* concept of entropy has long been known [1], [2], [3], [4]. In fact, frequently used terms in machine learning and statistics, such as "Gibbs" sampling and "Boltzmann" machines are drawn from Thermodynamics. Not surprisingly, the concept of "Free Energy", originally defined for thermodynamics, has also found its analog in pattern classification and machine learning.

Invocations to the concept of entropy and the related notion of cross-entropy, in particular, are ubiquitous in statistical pattern classification. Entropy can alternately be viewed as the expected log-likelihood of a random variable. Maximum-likelihood estimation, a popular tool for estimation of distributions and models, as well as for classification, effectively minimizes empirical estimates of the relative entropy between the true distribution of a random variable and that specified by the model [5], [6]. Entropy and cross-entropy can be used to characterize both the *compactness* of a data set and the *diversity* of separate data sets. As a result, entropy has been used as a criterion for classification and clustering of data since at least the early eighties [7], [8]. Entropy has also been used as a measure of the structure in a data or a model – low entropies implying high predictability and hence high degree of structure (or organization) [9]. On the other hand, *lack of information* has been characterized as high entropy: the celebrated maximum-entropy methods employed to learn models in a variety of fields such as text processing, information retrieval, speech and audio processing [10], [11], [12] and even signal processing [13] effectively attempt to capture known facts about the data, while assuming maximum ignorance about other facets.

The concept of "free energy" too has found widespread use in various fields of computer science such as statistics, optimization, and machine learning. One of the earliest invocations to free energy was in the now-famous Metropolis Hastings algorithm [14]. In this and subsequent algorithms of a similar nature [15], [16] free energy is employed as a characterization of the randomness in the steps taken by an algorithm in proceeding towards its objective. The "temperature" of the system is used as a control parameter over this randomness. From another perspective, increasing the temperature of a system and thereby its free energy is equivalent to flattening the landscape of an objective function that is being searched for an optimum. This perspective has naturally led to the concept of *annealing* [15], where the temperature of a system (or objective function) is gradually lowered from a high value, to enable an optimization algorithm to escape local optima and increase its likelihood to arrive at a global optimum.

An alternative interpretation is also presented in pattern analysis mechanisms that are based on self organization, such as self-organizing maps [17], Hopfield networks [18], Boltzman machines [19] and the various neural network architectures that build on them [20]. Here the analogy is closer to that in the well known spin-glass effect [21], in which a large number of free-floating magnetic dipoles attempt to align themselves to a local magnetic field, while also affecting the field experienced by their cohorts through their own orientation. The spin glass has a finite number of minimum-free-energy stable configurations into which it can arrive, and the "attraction" of these configurations depends further on the temperature of the system. Analogously, self-organizing network structures attempt to arrive at stable configurations that locally minimize an equivalent of free energy, and their ability to arrive at these configurations is in turn governed by a temperature parameter.

In all cases, (the computational analog of) free energy has eventually been used as a handle to achieve improved optimization over complex, possibly non-convex objective function landscapes.

In this paper we hypothesize that free energy provides a natural objective function to be minimized for *classification* as well. Particularly, in scenarios such as speech recognition and audio labeling, where evidence is obtained from multiple sources. If one of the sources is noisy, recasting classification as a free-energy minimization problem gives us a natural means of flattening the peaks and valleys in the contribution of the noisy component to the overall classification objective. Moreover, expressing this in terms of a "temperature" also provides an intuitive explanation – the noisy information source may be viewed as being at a "higher" temperature.

The literature on the direct use of free energy as an objective function for *classification* is, however, sparse, except in situations where it is used as a mechanism for annealing a solution towards the true optimum [22]. Classification at raised temperatures is generally not performed, and in the case of speech recognition or audio labeling, the only related work we have found is our own prior work on the topic [23].

The rest of this paper is organized as follows. In Section II we discuss the general case of HMM-based classifiers. Free-energy based classification is briefly explained in Section III. Section IV outlines our proposed formulation of *heating* of HMM parameters to emulate free-energy based classification in practical systems. Experimental results on audio retrieval based on audio labeling, and on speech recognition are presented in Section V. Conclusions are presented in Section VI.

II. The Specific Case of HMM-based Classifiers

The formulation of a framework for free-energy based classification depends on the type of classifier under consideration. In this paper, we will focus on the specific case of HMM-based classifiers since these are successfully used in many applications that rely on modeling generative processes, including state-of-art large vocabulary speech recognition systems. We also use the latter for audio event labeling in this paper.

HMM-based classifiers of continuous speech or audio are statistical pattern classifiers which model sound units using hidden Markov models (HMMs). Given a sequence of data X derived from the audio signal, the classification problem that is solved is that of finding the class $c(X)$ for which the following expression, given by Bayes classification rule, is maximized:

$$c(X) = \arg\max_{C} P(C)P(X|C) \tag{1}$$

where C represents any class, $P(C)$ is the a priori probability of C, and $P(X|C)$ is the probability of X given by the HMM for C. This can be equivalently expressed as

$$c(X) = \arg\max_{C} \left\{ \log P(C) + \log \sum_{s} P(X|s, C) \right\} \tag{2}$$

where s is any state sequence through the HMM for C, that might have generated X. This is approximated by the Viterbi algorithm as

$$c(X) = \arg\max_{C} \left\{ \log P(C) + \max_{s}\{\log P(X, s|C)\} \right\} \tag{3}$$

The Bayes classification rule has been shown to be optimal when the class distributions represent the true distributions of the data to be classified. However, in practical systems, the distribution of the test data cannot be guaranteed to match those of the classifier. The parameters of the HMMs are learned from a corpus of training data through a tedious, and often intricate process. Once trained, the system is frequently deployed in varied acoustic environments (or used by diverse users in the case of speech), as a result of which the test data are rarely identically distributed to the training data. Consequently, the classification performance achieved with the Bayes classification rule is far from optimal.

The conventional solution to this problem is to modify the parameters of the HMMs in the classifier to better represent the test data, using one of several methods that have been proposed for the purpose (e.g. MAP/MLLR) [24], [25]. Bayesian classification is then performed using the modified parameters. While these procedures are highly effective, they require adaptation data that are similar to the test data, and also require significant offline computation to obtain the adapted parameters.

An alternative strategy is proposed in [23] where improved recognition of mismatched data is achieved by modifying the classification rule itself. In the modified classification rule, a free-energy term that is governed by a temperature parameter T, is defined for the various classes. The classification rule selects the class with the lowest free energy. The HMM parameters are not modified. Further, the rule itself is computationally no more expensive than the conventional Bayesian classification rule. The modified rule has no Bayesian interpretation except in the specific instance when $T = 1$. Classification at elevated temperatures ($T > 1$) is observed to result in large improvements in recognition performance on mismatched test data.

In the following two sections we propose a third option for integrating free enery criteria into HMM-based classification. It can be shown that elevating the temperature of an HMM in the free-energy expression given in [23] is equivalent to reducing its free energy. We therefore attempt to *modify the parameters* of the state output densities of the HMMs in a manner that reduces their free energy, prior to classification. We refer to this procedure as *heating* the HMMs. As in the case of free-energy based classification, the procedure is based only on the assumption of mismatched test data, without any reference to the specific test data themselves. The resulting HMM remains a probability density with a total probability mass of 1. Bayesian classification rules can now directly be applied to the modified HMM. Experimental results show that this can indeed result in significant improvements in classification performance on mismatched data.

III. Free-Energy based Classification

Free energy is a characteristic of thermodynamic systems. It is the amount of work required to restore the system to a state of equilibrium, implying by definition that when a system is in equilibrium, its free energy is minimum. Consider a system at temperature T that has an energy H_s when it is in some configuration s. Let P_s be the probability that the system is in configuration s, and P be the set of all P_s. The free energy of the system is defined as

$$F(P) = \sum_{s} P_s H_s + T \sum_{s} P_s \log(P_s) \tag{4}$$

The first term represents the average energy in the system and the second term represents the entropy of the system. The minimum free energy is derived by minimizing Equation 4 with respect to P and can be shown to be:

$$F = -T \log \sum_{s} \exp\left(\frac{-H_s}{T}\right) \tag{5}$$

Drawing from this thermodynamic analogy, free energy has been defined for other systems where the notion of a system configuration exists. One such definition is that for parametric statistical models with latent variables, mainly for the purpose of estimation of their parameters [22].

The free energy of an HMM is defined as follows: let Λ_C represent the parameters of the HMM for class C. Let the *a priori* probability of C be $P(C)$. Let X be the data to be classified, and s be any valid state sequence through the HMM, that can generate X. We equate s with the configuration of the HMM and define the *energy* of s, H_s, as

$$H_s = -\log P(C) - \log P(X, s|\Lambda_C) \tag{6}$$

This is the negative of the log of the joint probability of the class, the state sequence, and the data. Using Equation 5, the free energy of the system (i.e. the HMM) is now given by

$$F_C(X|\Lambda_C) = -\log P(C) - T\log\left(\sum_s P^{\frac{1}{T}}(X, s|\Lambda_C)\right) \tag{7}$$

Classification with free energy associates a data sequence X with the class $c(X)$ according to the rule:

$$c(X) = \arg\min_C F_C(X|\Lambda_C) \tag{8}$$

The free energy for an HMM can be efficiently computed using the following variant of the forward algorithm:

$$\alpha(s,t,C) = -T\log\sum_s \left(e^{-\alpha(s',t,C)}a(s',s)P(x_t|s)\right)^{\frac{1}{T}} \tag{9}$$

$$\alpha(s,1,C) = -\log P(C) - \log\pi(s) - \log P(x_1|s) \tag{10}$$

$$F_C(X|\Lambda_C) = -T\log\left(\sum_s e^{\frac{-\alpha(s,N,C)}{T}}\right) \tag{11}$$

where $a(s', s)$ is the transition probability from state s' to state s, $\pi(s)$ is the initial probability of s, and $P(x_t|s)$ is the value of the state output density of s at x_t. The minimum-free-energy classification rule is identical to the Bayes classification rule at $T = 1$. Classification performance has however been empirically observed to be best at higher temperatures, particularly when there is a mismatch between the HMM and the true distribution of the data to be classified.

IV. Modifying HMM Parameters to Decrease Free Energy

The free energy of an HMM as computed using Equation 7 does not represent a probability, and the classification rule in Equation 8 is not the Bayesian rule. Nevertheless it is theoretically possible to redefine the parameters of statistical models in the classifier such that the Bayesian classification rule based on the redefined models is identical to the minimum free-energy classification rule of Equation 8. It can be shown that such redefinition of the statistical parameters requires modification of not only the parameters of the distributions of the classes, but also the *a priori* probabilities of the classes themselves. The modified class parameters must be defined in terms of a *partition function* that cannot be expressed in closed form for HMMs. The modified *a priori* class probabilities are a function of both the temperature and the parameters of the individual classes. It is not clear that the resultant statistical model can still be expressed as an HMM.

On the other hand, conversion of density parameters to simulate minimum-free-energy classification using the Bayesian classification rule is tractable when class distributions are mixture Gaussian densities rather than HMMs. Mixture Gaussian class distributions have the following form:

$$P(X|\Lambda_C) = \sum_k w_{C,k} G(X|\mu_{C,k}, \sigma_{C,k}) \tag{12}$$

where $w_{C,k}$, $\mu_{C,k}$ and $\sigma_{C,k}$ are the mixture weight, mean and variance of the k^{th} Gaussian in the density of class C, and $G(X|\mu, \sigma)$ represents the value of a Gaussian with mean μ and variance σ at a vector X. It can be shown that minimum-free-energy classification at temperature T is identical to Bayesian classification with modified mixture Gaussian densities $P_T(X|\Lambda_C)$ and *a priori* class probabilities $P_T(C)$ that have the following form:

$$P_T(X|\Lambda_C) = \sum_k \tilde{w}_{C,k} G(X|\mu_{C,k}, T\sigma_{C,k}) \tag{13}$$

where the new mixture weights are given by

$$\tilde{w}_{C,k} = \frac{1}{Z_C} w_{C,k}^{\frac{1}{T}} |\sigma_{C,k}|^{\frac{T-1}{2T}} \tag{14}$$

where Z_C is a normalizing constant for the mixture weights of C, and

$$P_T(C) = \frac{Z_C}{Z} P^{\frac{1}{T}}(C) \tag{15}$$

where Z is a normalizing constant.

We note that state output densities in HMMs are usually modeled as mixture Gaussian densities. In Equation 7, which specifies the free energy of an HMM at a temperature T, the individual $P(X, s|\Lambda_C)$ components used within the second term on the right hand side are true probabilities, computed as

$$P(X, s|\Lambda_C) = \pi(s_1)P(x_1|s_1)\prod_{t>1} a(s_{t-1}, s_t)P(x_t|s_t) \tag{16}$$

where s_t is the state at time t in the state sequence s and x_t is the t^{th} observation vector in X. In this paper we propose to use a modified definition of the free energy as follows:

$$\tilde{F}_C(X|\Lambda_C) = -\log P(C) - \log\sum_s \tilde{P}(X, s|\Lambda_C) \tag{17}$$

where

$$\tilde{P}(X, s|\Lambda_C) = \pi(s_1)F(x_1|s_1)\prod_{t>1} a(s_{t-1}, s_t)F(x_t|s_t) \tag{18}$$

where $F(x_t|s_t)$ is the free energy of the state output density of s_t. The term $\tilde{P}(X, s|\Lambda_C)$ does not represent a probability. Since we wish to permit the use of the Bayesian classification rule, we do not use Equation 18 directly. Instead we modify the parameters of the state output densities of the state by modifying the mixture weights of all densities according to Equation 14. We refer to the modification of state density parameters in this manner as *heating* the HMM. The modified densities now result in likelihood values that are approximations to scaled versions of likelihood values that are approximations to scaled versions of the free energy. Instead of using Equation 18, we redefine $\tilde{P}(X, s|\Lambda_C)$ as

$$\tilde{P}(X, s|\Lambda_C) = \pi(s_1)\tilde{P}(x_1|s_1)\prod_{t>1} a(s_{t-1}, s_t)\tilde{P}(x_t|s_t) \tag{19}$$

where $\tilde{P}(x_t|s_t)$ is the state output density value of s_t computed using the modified parameters. $\tilde{F}_C(X|\Lambda_C)$ now still represents a probability and the conventional Bayesian classification rule can be applied. The conventional forward and Viterbi algorithms can be used to compute class probabilities.

The equations above show how forward and best-state scores, computed at elevated temperatures. For *recognition*, we employ the modified state output densities values within a conventional Viterbi search as given in Equation 19.

V. Experimental Evaluation

We performed two sets of experiments: one on an audio-based multimedia event retrieval task, and a second set on a conventional speech recognition task. Experiments were aimed at highlighting the effect of incorporating the free-energy term in the HMM state distributions under conditions of mismatch.

A. Audio Classification for Retrieval

Our first experiment was conducted on the MED11 example-based multimedia-event retrieval task [26]. The task here is as follows: we are given a collection of multimedia recordings. We are also provided a small number of examples of recordings representing a category of events such as "repairing an appliance", "parade", or "skateboarding". The objective is to retrieve all other instances of the same category from the data set, based on what can be learned from the example recordings. In our experiment we only employed the *audio* components of the individual recordings to perform the retrieval; the video was not used.

The approach employed for the retrieval was to compute a vector descriptor for each recording. Vectors obtained from positive and negative examples of the target event were used to train a binary classifier. Subsequently, all recordings in the dataset used for the experiment were classified by this classifier. In our experiments we used a random forest classifier.

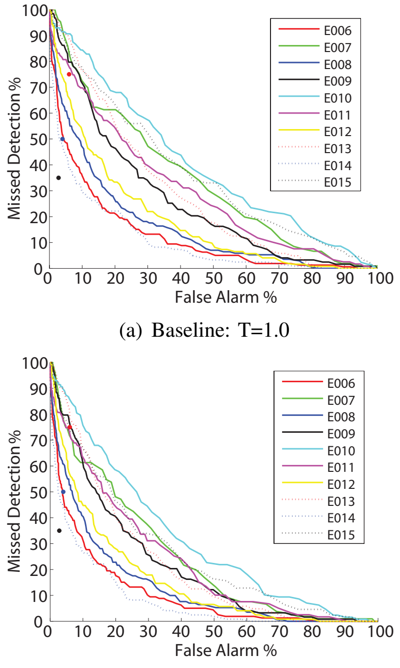
For our experiment we used a data set of 2300 files. The data set included (but was not limited to) approximately 100 examples each of 10 categories of events, labeled "E006"-"E015" in our plots and tables below. Events from any category were considered to be negative for other events. All results were obtained through an 8-way jackknife experiment over the 2300 files.

To obtain the vector descriptors we trained HMMs for 417 different types of sounds, such as "car noise", "glass breaking" etc. using recordings from the Foley database [27]. These sounds may be considered to be the equivalent of "words" that occur in the recordings. Any recording may thus be characterized by the relative frequency with which they occur in it, and it may be expected that the differences between different event categories (E006 etc.) will manifest as differences in these relative frequencies. HMMs were trained on sequences of Mel-frequency cepstral vectors obtained from the recordings. State output distributions were set to be mixtures of 4 Gaussians. Thereafter, each recording was "decoded" into a sequence of audio events using these HMMs. For the purpose of decoding, we assumed that all events were equally probable; this can be expressed as a simple "unigram" grammar over events where all events are equiprobable. The CMU Sphinx-3 speech recognition system was used for the experiment, both to train HMMs and to decode audio.

Once each recording was decoded into a sequence of events, we computed a simple 417-dimensional bag-of-words representation for each recording by counting the number of times each event was found to have occurred in it.

The Foley data have all been recorded in a studio and are clean. On the other hand, the MED11 data are real-life recordings obtained from public sources, and are noisy, where sounds rarely occur in isolation, are corrupted by a variety of background noises, and are recorded over a variety of channels. Frequently music is overlaid on the recordings. Consequently, the data are greatly acoustically mismatched with the Foley recordings.

We compensated for the mismatch by increasing the temperature at which decoding was performed. Figure 1 shows the results obtained in the form of DET curves, which plot the percentage of missed detections as a function of the percentage of false alarms for each of the ten event classes. We note that the retrieval is obtained at $T = 1.75$ is better than that at $T = 1.0$.



(a) Baseline: T=1.0

(b) T=1.75

Fig. 1. Multimedia retrieval with Foley events decoded at various temperatures. The red, blue and black dots in each of the figures represent operating points with MD 75%, 50% and 35% respectively, when MD/FA = 12.5.

Table I shows the percentage missed detection of events at the operating point where the ratio of missed detections to false alarms is 12.5 – the standard operating point specified in IARPA evaluations. We show the results at four temperatures, including the baseline ($T = 1.0$).

As a comparator, we also show results using "AUDs". AUDs [28] are sound units learned directly from the audio in an unsupervised manner. Since the AUDs are learned from the database itself, the HMMs for the AUDs are "matched" to the data. AUDs based descriptions may hence be considered to be descriptions obtained from “matched” models, and can be viewed as establishing an upper bound on the performance to be obtained under the conditions of the experiment.

TABLE I. PERCENT MISSED DETECTION AT MD/FA = 12.5.

Event	T=1	T=1.5	T=1.75	T=2	AUDS
E006	51.5	49.0	48.4	50.3	45.2
E007	82.0	75.4	70.7	77.3	79.2
E008	60.8	59.1	57.9	60.8	50.3
E009	79.6	77.2	74.8	78.0	63.4
E010	87.5	86.0	82.8	86.7	82.0
E011	77.3	74.8	73.1	75.6	71.4
E012	68.2	65.0	61.7	65.8	60.1
E013	85.5	81.7	76.9	81.7	72.1
E014	48.3	46.0	44.2	45.9	43.4
E015	81.65	75.2	73.4	78.9	63.3

We observe from the above results that when we use Foley models, the best results are obtained at elevated temperatures, which effectively account for the mismatch between the Foley data and the MED data. Elevating the temperature improves performance at nearly all operating points. The best results are obtined at T=1.75 in this case. Interestingly, the optimal temperature was the same for all events. We believe this to be an oddity of the database. Increasing the temperature further results in reduced performance. Also, consistently with our original hypotheses that raising the temperature only accounts for mismatch between the distributions of training and test data, we find that the best performance overall is obtained with the *truly* matched AUDs models, although in some cases the difference between the performance obtained with the AUDs and the optimal Foley-based decodes is relatively small.

B. Speech Classification for Transcription

While our first experiment demonstrates the validity of free-energy based classification under conditions of *acoustic* mismatch, our second experiment demonstrates the same when the mismatch is due to *non-acoustic* factors. Our second experiment was conducted on speech signals, where mismatch between training and test audio arose from *accent* mismatch.

Experiments were performed with the NATO Non-Native (N4) Speech corpus [29]. This is a database of non-native speech collected by the NATO Research Study Group and made available to the community from the Linguistic Data Consortium (LDC). The database consists of accented speech from people of four different nationalities: German (DE), Dutch (NL), Canadian (CA) and British (UK). Baseline acoustic models were trained from the US English TDT2 [30] and TDT3 [31] speech corpora, also available from the LDC. We used the CMU Sphinx speech recognition system for our experiments. The system uses several different search strategies for decoding. We used the full flat decoding strategy for continuous speech. The acoustic models used were continuous density 3-state Bakis topology HMMs with no skips permitted between states. The models comprised 6000 tied states, with 8-component Gaussian mixture state output distributions. An ARPA format trigram language model was built using military protocol text collected from the internet. There were no out-of-vocabulary words, but the NATO database did not contribute otherwise to the language model.

The test data were recognized at several temperatures. Table II shows the word error rates obtained for each accent, against the temperature at which the data were decoded. The highlighted (yellow) column in the table corresponding to $T = 1$ is exactly identical to the standard Bayesian decoding, as explained earlier. Columns for $T < 1$ show recognition performance at lowered temperatures, whereas those at $T > 1$ show the same at elevated temperatures. Figure 2 shows the performance in graphical format to present the trends visually. Each subplot shows the performance for a single accent.

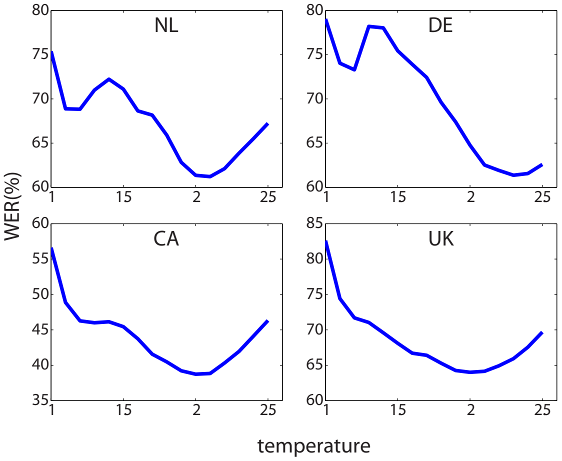


Fig. 2. WERs as a function of temperature for different accents. T=1 represents the conventional MAP decoding strategy.

Once again, we note from the results that the optimal recognition performance is *not* obtained at $T = 1$.

Figure 2 summarizes the effect of raising temperature. We observe that there is a general trend of improved recognition as temperature increases to 2.1; however a “bump” is observed at a temperature of 1.4 or so, and the performance at 1.2 appears to be some form of local optimum.

The best result in all four cases occurs at an elevated temperature in the vicinity of $T = 2$; specifically, if a single elevated temperature were to be chosen as the operating point, it would be $T = 2.1$. The difference between the baseline WER at $T = 1$ and the best result at elevated temperatures is quite large, at nearly 18% absolute in three of four cases.

VI. CONCLUSION

In both of our experiments we find that elevation of temperature results in significantly improved classification results under conditions of mismatch. Morever, the method is not merely effective against *acoustic* mismatch, but seems to generalize to other forms of mismatch as well, *e.g.* accent mismatch in the case of speech signals. The biggest advantage here is that it requires minimal implementation effort: in the experiments here all it takes is a simple adjustment to the HMM parameters in the acoustic models. Similarly simple implementations can be expected for other types of models as well.

More generally, the notion of “temperature” and “free-energy” have often been invoked in the context of annealing for optimization of objective functions defined over continuous support. Classification, on the other hand, is typically a search over a discrete support, and not usually viewed as an optimization problem. This is generally considered to be distinct from

TABLE II. PERFORMANCE OF MAXIMUM-LIKELIHOOD AND FREE-ENERGY BASED SPEECH RECOGNITION, IN TERMS OF PERCENT WORD ERROR RATE. THE $T = 1$ COLUMN HIGHLIGHTED IN YELLOW CORRESPONDS TO CONVENTIONAL BAYESIAN DECODING. THE BOLD NUMBERS ARE THE BEST RESULTS OBTAINED IN EACH ROW.

Temp	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.1	2.2	2.3	2.4	2.5
NL	87.2	75.3	68.8	68.8	70.9	72.2	71.1	68.6	68.1	65.8	62.8	61.3	**61.2**	62.1	63.8	65.5	67.2
DE	90.4	79.0	74.0	73.2	78.1	78.0	75.4	73.9	72.4	69.6	67.3	64.7	62.5	61.9	**61.3**	61.5	62.5
CA	67.0	56.6	48.8	46.2	45.9	46.1	45.4	43.7	41.5	40.4	39.2	**38.7**	38.8	40.3	41.9	44.1	46.3
UK	96.1	82.6	74.4	71.6	71.0	69.5	68.1	66.7	66.4	65.2	64.2	**64.0**	64.1	64.9	65.9	67.5	69.6

the situations where notions of free-energy and temperature may be invoked.

The two problems we have studied here, however, present an interesting case. In the case of the audio data, the set of classes is not merely the subset of 417 Foley sounds, but the set of all possible *event sequences* that may occur in a recording. Similarly, the true set of classes that we search over in the speech recognition example is the set of all possible word sequences. Thus, although the class set is discrete, the set itself can be infinitely large, suggesting that the concept of annealing may be drawn upon if the search space could somehow be ordered and represented over a continuum. However, how this may be done is unclear.

Although we have not actually cast the problem of classification in this light in this paper, we have definitely demonstrated that the concept of classification at elevated temperatures can indeed be cast in formal terms, and furthermore, that even in a single pass of classification, elevation of temperature can result in significantly improved recognition. In future work, we aim to expand this to a fuller formulation of *annealed* search for optimal classification over infinite sets.

This is of particular interest in the case of semantic decoding of audio recordings. Given the potentially infinite set of basic audio events themselves, as well as the fact that it is difficult to obtain large quantities of *matched* annotated data (particularly considering that audio recordings in the wild are rarely pristine, and almost always comprise mixtures of many sounds), mismatches between training and test data will always occur. We speculate that techniques such as the proposed free-energy based decoding at elevated temperature will be key to resolving such problems.

REFERENCES

[1] Edwin T. Jaynes, "Information theory and statistical mechanics," *Phys. Rev.*, vol. 106, pp. 620–630, 1957.

[2] C. H. Bennett, "The thermodynamics of computation – a review," *International Journal of Theoretical Physics*, vol. 21, no. 12, pp. 905–940, 1982.

[3] J. Ladyman, S. Presnell, T. Short, and B. Groisman, "The connection between logical and thermodynamic irreversibility," *Studies in the History and Philosophy of Modern Physics*, vol. 38, pp. 58–79, 2007.

[4] James Ladyman, Stuart Presnell, and Anthony J. Short, "The use of information theoretic entropy in thermodynamics," *Studies in the History and Philosophy of Modern Physics*, vol. 39, pp. 315, 2008.

[5] John E. Shore, "On a relation between maximum likelihood classification and minimum relative-entropy classification," 1984, pp. 851–854, Vol. 30, Issue 6.

[6] Satosi Watanabe, "Pattern recognition as a quest for minimum entropy," in *Pattern Recognition*. 1984, pp. 381–387, Vol. 13, Issue 5.

[7] John E. Shore and Robert M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," 1982, pp. 11–17, PAMI 4, Issue 1.

[8] Joaquim P. Marques de S, Lus M.A. Silva, Jorge M. F. Santos, and Lus A. Alexandre, "Minimum error entropy classification," in *Studies in Computational Intelligence*. 2013, Springer.

[9] Matthew Brand, "An entropic estimator for structure discovery," in *Advances in Neural Information Processing Systems*. 1999, pp. 723–729, MIT Press.

[10] Kamal Nigam, John Lafferty, and Andrew McCallum, "Using maximum entropy for text classification," 1999, pp. 61–67.

[11] Adam L. Berger, Stephen A. Della Pietra, and Vincent J. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, pp. 39–71, 1996.

[12] Andrew Mccallum and Dayne Freitag, "Maximum entropy markov models for information extraction and segmentation," 2000, pp. 591–598, Morgan Kaufmann.

[13] J. P. Burg, "Maximum entropy spectral analysis," *37th Annual International Meeting*, 1967.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, vol. 21, no. 1087, 1953.

[15] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, 1983.

[16] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 6, no. 6, 1984.

[17] T. Kohonen, *Self-organizing maps*, Springer Verlag, Berlin, 2001.

[18] J. J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," *Proceedings of the National Academy of Sciences of the USA*, vol. 79, no. 8, 1982.

[19] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for boltzmann machines," *Cognitive Science*, vol. 9, no. 1, 1985.

[20] Nicolas Le Roux and Yoshua Bengio, "Representational power of restricted boltzmann machines and deep belief networks," *Neural Computation*, vol. 20, no. 6, pp. 1631–1649, 2008.

[21] Nishimori Hidetoshi, *Statistical physics of spin glasses and information processing: An introduction*, Oxford University Press, Oxford, 2001.

[22] Ulrich Paquet, "Bayesian inference for latent variable models," Tech. Rep., Cambridge University, 2008.

[23] Rita Singh, Manfred Warmuth, Bhiksha Raj, and Paul Lamere, "Classification with free energy at raised temperatures," *Eurospeech*, 2003.

[24] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 175–181, 1995.

[25] Daniel Povey and Kaisheng Yao, "A basis representation of constrained MLLR transforms for robust adaptation," *Computer Speech & Language*, vol. 26, no. 1, pp. 35–51, 2012.

[26] DATABASE, "Trecvid 2011," http://www.nist.gov/itl/iad/mig/med11.cfm, 2011.

[27] DATABASE, "The art of foley sound effects library," http://www.sound-ideas.com/what-is-foley.html, Copyright 2000.

[28] Sourish Chaudhury, Rita Singh, and Bhiksha Raj, "Exploiting temporal sequence structure for semantic analysis of multimedia," 2012.

[29] L. Benarousse, E. Geoffrois, J. J. Grieco, R. Series, H. J. M. Steeneken, H. Stumpf, C. Swail, and D. Thiel, "The nato native and non-native (n4) speech corpus," in *Information Systems Technology Panel (IST) Workshop*, Aalborg, Denmark, 2001 (published in 2003), pp. 2210–2239.

[30] DATABASE, "TDT2 corpus," http://projects.ldc.upenn.edu/TDT2/, 2000.

[31] DATABASE, "TDT3 corpus," http://projects.ldc.upenn.edu/TDT3/, 2001.