FREE ENERGY FOR SPEECH RECOGNITION

Rita Singh [†], *Kenichi Kumatani* [‡]

[†] Carnegie Mellon University, Pittsburgh, PA, USA
[‡] Spansion Inc., Sunnyvale, CA, USA

ABSTRACT

Traditionally, speech recognizers have used a strictly Bayesian paradigm for finding the best hypothesis from amongst all possible hypotheses for the data to be recognized. The Bayes classification rule has been shown to be optimal when the class distributions represent the true distributions of the data to be classified. In reality, however, this condition is often not satisfied – the classifier itself is trained on some training data and may be deployed to classify data whose statistical characteristics are different from the training data. The Bayes classification rule may result in suboptimal performance under these conditions of *mismatch*. Classification may benefit from the use of modified classification rules in this case.

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 in classification. Furthermore, we show how this novel classification scheme can be used in the framework of existing Bayesian classification schemes implemented in current speech recognizers by simply modifying the class distributions *a priori*. Pilot experiments show that minimization of free energy results in more accurate recognition under conditions of mismatch.

Index Terms— Bayesian classification, Speech recognition, Free energy, Temperature

1. INTRODUCTION

Currently, the best performing state-of-art large vocabulary speech recognition systems are statistical pattern classifiers which model sound units using hidden Markov models (HMMs). Given a sequence of data X derived from the speech signal, the classification problem that is solved is that of finding the class $c(X)$ for which the following expression, given by the Bayes classification rule, is maximized:

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

where C represents any class and $P(C)$ is the *a priori* probability of C. For the speech recognition problem, C represents word-sequence hypotheses. $P(X|C)$ 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. In other words, the distribution of the test data must match those employed by the classification rule.

In HMM-based speech recognition systems, class distributions are represented by HMMs. The parameters of the HMMs are learned from a corpus of training data. Once trained, the system is frequently deployed in varied acoustic environments and used by diverse users, as a result of which the test data are rarely identically distributed to the training data. The requirement for optimal classification – that the distribution of the test data must match the distributions used by the classifier – is violated. Consequently, the classification performance achieved with the Bayes classification rule is suboptimal.

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]) [1, 2]. 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 significant offline computation to obtain the adapted parameters.

An alternative strategy is proposed in [3] 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 $T = 1$. In our previous work [3] classification at elevated temperatures ($T > 1$) is observed to result in large improvements in recognition performance on mismatched test data.

In this paper we explore a third option. It can be shown that elevating the temperature of an HMM in the free energy expression given in [3] 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 recognition accuracy on mismatched data.

The rest of this paper is organized as follows. Free energy based classification is briefly explained in Section 3. Section 4 outlines the heating of HMM parameters. Experimental results are presented in Section 5. Our conclusions are presented in Section 6.

2. RELATED WORK

The relationship between the *thermodynamic* principle of entropy and the *information theoretic* concept of entropy has long been known [4, 5, 6, 7]. 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 [8, 9]. 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 [10, 11]. Entropy has also been used as a measure of the structure in a data set or a model – low entropies implying high predictability and hence high structure [12]. On the other hand, *lack of information* has been characterized as high entropy: the celebrated maximum-entropy method employed to learn models in a variety of fields such as text processing, information retrieval, speech and audio processing [13, 14, 15] and even signal processing [16] 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 [17]. In this and subsequent algorithms of a similar nature [18, 19] 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* [18], 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 [20], Hopfield networks [21], Boltzman machines [22] and the various neural network architectures that build on them [23]. Here the analogy is closer to that in the well known spin-glass effect [24] 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, where evidence is obtained from multiple sources (acoustics and language), 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 [25]. Classification at raised temperatures is generally not performed, and in the case of speech recognition, the only related work we have found is our own prior work on the topic [3].

3. 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 [3]:

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

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 [25].

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

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

Classification with free energy associates data 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:

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

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

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

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 [3], particularly when there is a mismatch between the HMM and the true distribution of the data to be classified.

4. 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:

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

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:

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

where the new mixture weights are given by

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

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

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

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

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

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 a modified *hybrid* definition of the underlying thermodynamic system, where only the *individual states* of the HMM are subject to thermodynamic variations, but the rest of the system are governed strictly by Markovian rules. This results in a modified definition of the free energy:

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

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

where $F(x_t|s_t)$ is the free energy of the state output density of s_t. The term $P(X,s|\Lambda_C)$ 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 by modifying their mixture weights 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 the free energy. $\tilde{P}(X,s|\Lambda_C)$ is now computed as

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

where $\tilde{P}(x_t|s_t)$ is the state output density value of s_t computed using the modified parameters. $F_C(X|\Lambda_C)$ 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 the forward and best-state scores can be computed at elevated temperatures. For *recognition*, we employ the modified state output density values within a conventional Viterbi search as given in Equation 19.

5. EXPERIMENTAL EVALUATION

Experiments were aimed at highlighting the effect of incorporating the free energy term in the HMM state distributions on speech recognition performance under conditions of mismatch. Since there are a vast number of mismatched scenarios that can occur in real life, and that can be simulated, we decided to focus on a simple proof-of-concept mismatch scenario involving non-native speech. Note again that the non-nativity is not the focus of this paper; rather it is the mismatch between the acoustic models and the test data. Experiments were performed with the NATO Non-Native (N4) Speech corpus [26]. 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 TDT2 [27] and TDT3 [28]

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

Table 1. Performance of maximum-likelihood and free-energy based speech recognition. The $T = 1$ column highlighted in yellow corresponds to conventional decoding. The bold numbers are the best results obtained in each row.

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 1 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 1 shows the performance in graphical format to present the trends visually. Each subplot shows the performance for a single accent.

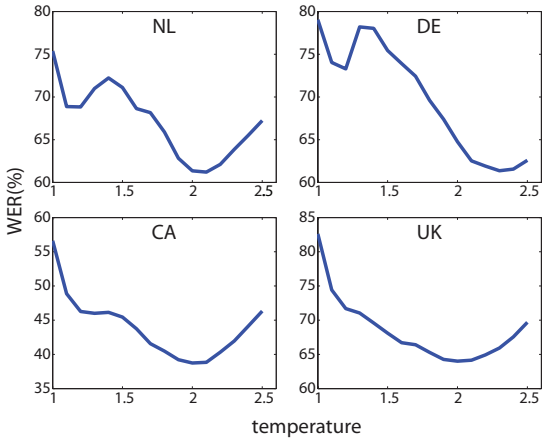

Fig. 1. WERs as a function of temperature for different accents. $T = 1$ represents the conventional MAP decoding strategy.

We note from the results that the optimal recognition performance is *not* obtained at $T = 1$. 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.

Figure 2 shows the performance obtained after one round of unsupervised MLLR adaptation. The left panel shows the gains from adaptation, when conventional Bayesian (Viterbi) decoding is performed in the first pass. The right panel shows results obtained when

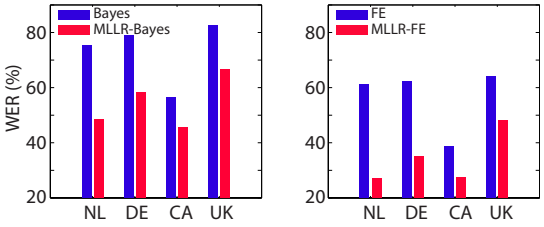

Fig. 2. Effect of unsupervised MLLR adaptation. Left: Original and adapted WERs for conventional decoding. Right: The same for free energy (FE) decoding at elevated temperature.

the first pass of decoding is at an elevated temperature (free-energy based recognition). Note that improvements obtained from elevating the temperature in the first pass of decoding are sustained after MLLR adaptation, showing that the benefits obtained are complementary.

6. CONCLUSIONS

Elevation of temperature is observed to result in significantly improved recognition under conditions of mismatch. Considering that just a simple adjustment has been made to the HMM parameters in the acoustic models to achieve this, the improved classification scheme is promising for use in speech recognizers.

Figure 1 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.

More generally, the notions of "temperature" and "free energy" have often been invoked in the context of annealing for optimization of objective functions defined over a 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 the situations where notions of free energy and temperature may be invoked.

Automatic speech recognition systems, however, present an interesting case. Although they do indeed represent a search over a discrete set, the set itself – representing all possible sentences that may be spoken – 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 recognition in this light in this paper, we have definitely demonstrated that the concept of recognition at elevated temperatures can indeed be cast in formal terms, and furthermore, that even in a single pass of recognition, 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 recognition.

7. REFERENCES

[1] 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.

[2] 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.

[3] R. Singh, M. Warmuth, B. Raj, and P. Lamere, "Classification with free energy at raised temperatures," *Eurospeech*, 2003.

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

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

[6] 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.

[7] 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.

[8] John E. Shore, "On a relation between maximum likelihood classification and minimum relative-entropy classification," in *IEEE Transactions on Information Theory*. 1984, pp. 851–854, Vol. 30, Issue 6.

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

[10] John E. Shore and Robert M. Gray, "Minimum cross-entropy pattern classification and cluster analysis," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 1982, pp. 11–17, PAMI 4, Issue 1.

[11] 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.

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

[13] Kamal Nigam, John Lafferty, and Andrew McCallum, "Using maximum entropy for text classification," in *Workshop on Machine Learning for Information Filtering, IJCAI-99*, 1999, pp. 61–67.

[14] 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.

[15] Andrew McCallum and Dayne Freitag, "Maximum entropy markov models for information extraction and segmentation," in *17th International Conf. on Machine Learning*. 2000, pp. 591–598, Morgan Kaufmann.

[16] J. P. Burg, "Maximum entropy spectral analysis," *37th Annual International Meeting of the Society of Exploration Geophysicists*, 1967.

[17] 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.

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

[19] 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.

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

[21] 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.

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

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

[24] N. Hidetoshi, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, Oxford University Press, Oxford, 2001.

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

[26] 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.

[27] Linguistic Data Consortium, "TDT2 corpus," `https://catalog.ldc.upenn.edu/LDC2000S92`, 2000.

[28] Linguistic Data Consortium, "TDT3 corpus," `https://catalog.ldc.upenn.edu/LDC2001S94`, 2001.