APPROXIMATING THE KULLBACK LEIBLER DIVERGENCE BETWEEN GAUSSIAN MIXTURE MODELS

John R. Hershey and Peder A. Olsen

IBM T. J. Watson Research Center

ABSTRACT

The Kullback Leibler (KL) Divergence is a widely used tool in statistics and pattern recognition. The KL divergence between two Gaussian Mixture Models (GMMs) is frequently needed in the fields of speech and image recognition. Unfortunately the KL divergence between two GMMs is not analytically tractable, nor does any efficient computational algorithm exist. Some techniques cope with this problem by replacing the KL divergence with other functions that can be computed efficiently. We introduce two new methods, the variational approximation and the variational upper bound, and compare them to existing methods. We discuss seven different techniques in total and weigh the benefits of each one against the others. To conclude we evaluate the performance of each one through numerical experiments.

Index Terms— Kullback Leibler divergence, variational methods, gaussian mixture models, unscented transformation.

1. INTRODUCTION

The KL-divergence, [1], also known as the *relative entropy*, between two probability density functions $f(x)$ and $g(x)$,

$$D(f\|g) \stackrel{\text{def}}{=} \int f(x) \log \frac{f(x)}{g(x)} \, \mathrm{dx}, \tag{1}$$

is commonly used in statistics as a measure of similarity between two density distributions. The divergence satisfies three properties, hereafter referred to as the divergence properties:

1. Self similarity: $D(f\|f) = 0$.
2. Self identification: $D(f\|g) = 0$ only if $f = g$.
3. Positivity: $D(f\|g) \geq 0$ for all f, g.

The KL divergence is used in many aspects of speech and image recognition, such as determining if two acoustic models are similar, [2], measuring how confusable two words or HMMs are, [3, 4, 5], computing the best match using histogram image models [6], clustering of models, and optimization by minimizing or maximizing the KL divergence between distributions.

For two gaussians $\hat{f}$ and $\hat{g}$ the KL divergence has a closed formed expression,

$$\begin{aligned} D(\hat{f}\|\hat{g}) &= \frac{1}{2}\big[\log \frac{|\Sigma_{\hat{g}}|}{|\Sigma_{\hat{f}}|} + \mathrm{Tr}[\Sigma_{\hat{g}}^{-1}\Sigma_{\hat{f}}] - d \\ &\quad + (\mu_{\hat{f}} - \mu_{\hat{g}})^T \Sigma_{\hat{g}}^{-1}(\mu_{\hat{f}} - \mu_{\hat{g}})\big] \end{aligned} \tag{2}$$

whereas for two GMMs no such closed form expression exists.

In the rest of this paper we consider f and g to be GMMs. The marginal densities of $x \in \mathbb{R}^d$ under f and g are

$$\begin{aligned} f(x) &= \textstyle\sum_a \pi_a \mathcal{N}(x; \mu_a; \Sigma_a) \\ g(x) &= \textstyle\sum_b \omega_b \mathcal{N}(x; \mu_b; \Sigma_b) \end{aligned} \tag{3}$$

where π_a is the prior probability of each state, and $\mathcal{N}(x; \mu_a; \Sigma_a)$ is a gaussian in x with mean μ_a and variance Σ_a.

We will frequently use the shorthand notation $f_a(x) = \mathcal{N}(x; \mu_a; \Sigma_a)$ and $g_b(x) = \mathcal{N}(x; \mu_b; \Sigma_b)$. Our estimates of $D(f\|g)$ will make use of the KL-divergence between individual components, which we thus write as $D(f_a\|g_b)$.

In the next section we review the mechanics of estimating $D(f\|g)$ using Monte Carlo sampling. Section 3 reviews the related unscented transformation. In section 4, we show two ways of estimating $D(f\|g)$ by approximating f and g with a single gaussian. In section 5 we show how Jensen's inequality leads to an approximation in terms of products of gaussians. In section 6 we review the matched bound approximation and in sections 7 and 8, we introduce two approximations based on variational methods [7]. The final section shows experimental results.

2. MONTE CARLO SAMPLING

The only method that really can estimate $D(f\|g)$ for large values of d with arbitrary accuracy is Monte Carlo simulation. The idea is to draw a sample x_i from the pdf f such that $E_f[\log f(x_i)/g(x_i)] = D(f\|g)$. Using n i.i.d. samples $\{x_i\}_{i=1}^n$ we have

$$D_{\mathrm{MC}}(f\|g) = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i)/g(x_i) \to D(f\|g) \tag{4}$$

as $n \to \infty$. The variance of the estimation error is $\frac{1}{n}\mathrm{Var}_f[\log f/g]$.

To compute $D_{\mathrm{MC}}(f\|g)$, we need to generate the i.i.d. samples $\{x_i\}_{i=1}^n$ from f. To draw a sample x_i from a GMM f we first draw a discrete sample a_i according to the probabilities π_a. Then we draw a continuous sample x_i from the resulting gaussian component $f_{a_i}(x)$.

The Monte Carlo method is the only method we discuss that yields a convergent method. It satisfies the similarity property, but the positivity property does not hold (the identification property will only fail in very artificial circumstances and with probability 0).

3. THE UNSCENTED TRANSFORMATION

The unscented transform, [8], is an approach to estimate $E_{f_a}[h(x)]$ in such a way that the approximation is exact *for all* quadratic functions $h(x)$. It is possible to pick $2d$ "sigma" points $\{x_{a,k}\}_{k=1}^{2d}$ such that

$$\int f_a(x)h(x) \, \mathrm{dx} = \frac{1}{2d}\sum_{k=1}^{2d} h(x_{a,k}). \tag{5}$$

One possible choice of the sigma points is

$$x_{a,k} = \mu_a + \sqrt{d\lambda_{a,k}} \; e_{a,k} \tag{6}$$

$$x_{a,d+k} = \mu_a - \sqrt{d\lambda_{a,k}} \; e_{a,k}, \tag{7}$$

for $k = 1, \ldots, d$ where $\lambda_{a,k}$ and $e_{a,k}$ are the eigenvalues and eigenvectors of the covariance Σ_a of the gaussian f_a. The KL divergence may be written as $D(f\|g) = \sum_a \pi_a E_{f_a}[h]$ for $h = \log(f/g)$, so the unscented estimate is

$$D_{\text{unscented}}(f\|g) = \frac{1}{2d} \sum_a \pi_a \sum_{k=1}^{2d} \log \frac{f(x_{a,k})}{g(x_{a,k})}. \tag{8}$$

The unscented estimate satisfies the similarity property, but the identification or positivity property do not hold in general. The unscented estimator is similar to a Monte Carlo technique except that the samples are chosen deterministically.

4. GAUSSIAN APPROXIMATIONS

A commonly used approximation to $D(f\|g)$ is to replace f and g with gaussians, $\hat{f}$ and $\hat{g}$. In one incarnation, one uses gaussians whose mean and covariance matches that of f and g. The mean and covariance of f are given by

$$\begin{array}{rcl} \mu_{\hat{f}} & = & \sum_a \pi_a \mu_a \\ \Sigma_{\hat{f}} & = & \sum_a \pi_a (\Sigma_a + (\mu_a - \mu_{\hat{f}})(\mu_a - \mu_{\hat{f}})^T). \end{array} \tag{9}$$

The approximation $D_{\text{gaussian}}(f\|g)$ is given by the closed-form expression, $D_{\text{gaussian}}(f\|g) = D(\hat{f}\|\hat{g})$, using equation (2).

Another popular method is to use the nearest pair of gaussians resulting in,

$$D_{\min} = \min_{a,b} D(f_a\|g_b). \tag{10}$$

Both $D_{\text{gaussian}}(f\|g)$ and $D_{\min}(f\|g)$ satisfy the positivity and similarity properties, but the identification property does not hold. Although they are simple to formulate, as we show later, they are both rather poor approximations.

5. THE PRODUCT OF GAUSSIANS APPROXIMATION

The likelihood $L_f(g)$, defined by $L_f(g) = E_{f(x)}[\log g(x)]$ relates to the KL divergence by $D(f\|g) = L_f(f) - L_f(g)$. Thus any estimate of the likelihood can be related to the KL divergence. An upper bound on the likelihood results from using Jensen's inequality to move the log outside the expected value

$$\begin{array}{rcl} L_f(g) & = & \sum_a \pi_a E_{f_a(x)} \log \sum_b \omega_b g_b(x) \\ & \leq & \sum_a \pi_a \log \sum_b \omega_b E_{f_a(x)}[g_b(x)] \\ & = & \sum_a \pi_a \log \sum_b \omega_b \int f_a(x) g_b(x) dx \\ & = & \sum_a \pi_a \log \sum_b \omega_b z_{ab}, \end{array} \tag{11}$$

where $z_{ab} \stackrel{\text{def}}{=} \int f_a(x) g_b(x) dx$ is the normalizing constant for a product of Gaussians, which has a well known closed-form solution. The KL-divergence can now be estimated in a simple closed form:

$$D_{\text{product}}(f\|g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} z_{aa'}}{\sum_b \omega_b z_{ab}}, \tag{12}$$

where $z_{aa'} \stackrel{\text{def}}{=} \int f_a(x) f_{a'}(x) dx$. $D_{\text{product}}(f\|g)$ satisfies the similarity property, but not the identification or positivity property. Also, $D_{\text{product}}(f\|g)$ tends to greatly underestimate $D(f\|g)$.

6. THE MATCHED BOUND APPROXIMATION

If f and g have the same number of components then by the chain rule for relative entropy, [9], we have the following upper bound

$$\begin{array}{rcl} D(f\|g) & \leq & D(\pi\|\omega) + \sum_a \pi_a D(f_a\|g_a) \\ & = & \sum_a \pi_a \left(\log \pi_a/\omega_a + D(f_a\|g_a)\right). \end{array} \tag{13}$$

suggested by Do [10]. Based on this equation, Goldberger et. al, [6], suggest a similar approximate formula to estimate $D(f\|g)$. Define a matching function, $m : \{1, \ldots, n_f\} \to \{1, \ldots, n_g\}$, between the n_f components of f and n_g components of g as follows:

$$m(a) = \arg\min_b D(f_a\|g_b) - \log(\omega_b). \tag{14}$$

Goldberger's approximate formula can then be written

$$D_{\text{goldberger}}(f\|g) = \sum_a \pi_a \left(D(f_a\|g_{m(a)}) + \log \frac{\pi_a}{\omega_{m(a)}} \right). \tag{15}$$

Unlike equation (13), $D_{\text{goldberger}}(f\|g)$ is not an upper bound of $D(f\|g)$. It also satisfies none of the divergence properties. This can be seen by considering the case where f and g are equal to a single gaussian, h but f is formulated as a mixture of identical components. It has also been reported that the method performs poorly with GMMs that have a few low-probability components [5]. However, compared to some of the preceding methods, $D_{\text{goldberger}}$ turns out to work well empirically.

7. THE VARIATIONAL APPROXIMATION

In this section we introduce a variational lower bound to the likelihood. In section 5 we pulled the log outside the integral for an upper bound. Here we will take it inside the sum to obtain a lower bound. We define variational parameters $\phi_{b|a} > 0$ such that $\sum_b \phi_{b|a} = 1$. By Jensen's inequality we have

$$\begin{array}{rcl} L_f(g) & \stackrel{\text{def}}{=} & E_{f(x)} \log g(x) \\ & = & E_{f(x)} \log \sum_b \omega_b g_b(x) \\ & = & E_{f(x)} \log \sum_b \phi_{b|a} \frac{\omega_b g_b(x)}{\phi_{b|a}} \\ & \geq & E_{f(x)} \sum_b \phi_{b|a} \log \frac{\omega_b g_b(x)}{\phi_{b|a}} \\ & \stackrel{\text{def}}{=} & \mathcal{L}_f(g, \phi). \end{array} \tag{16}$$

Since this is a lower bound on $L_f(g)$, we get the best bound by maximizing $\mathcal{L}_f(g, \phi)$ with respect to ϕ. The maximum value is obtained with:

$$\hat{\phi}_{b|a} = \frac{\omega_b e^{-D(f_a\|g_b)}}{\sum_{b'} \omega_{b'} e^{-D(f_a\|g_{b'})}}. \tag{17}$$

Likewise, we define

$$\mathcal{L}_f(f, \psi) \stackrel{\text{def}}{=} E_{f(x)} \sum_{a'} \psi_{a'|a} \log \frac{\pi_{a'} f_{a'}(x)}{\psi_{a'|a}} \tag{18}$$

and find the optimal $\psi_{a'|a}$:

$$\hat{\psi}_{a'|a} = \frac{\pi_{a'} e^{-D(f_a \| f_{a'})}}{\sum_{\hat{a}} \pi_{\hat{a}} e^{-D(f_a \| f_{\hat{a}})}}. \tag{19}$$

If we define $D_{\text{variational}}(f\|g) = \mathcal{L}_f(f, \hat{\psi}) - \mathcal{L}_f(g, \hat{\phi})$ and substitute $\hat{\phi}_{b|a}$ and $\hat{\psi}_{a'|a}$, the result simplifies to

$$D_{\text{variational}}(f\|g) = \sum_a \pi_a \log \frac{\sum_{a'} \pi_{a'} e^{-D(f_a \| f_{a'})}}{\sum_b \omega_b e^{-D(f_a \| g_b)}}. \tag{20}$$

$D_{\text{variational}}(f\|g)$ satisfies the similarity property, but it does not in general satisfy the positivity property. Like D_{gaussian} and D_{product}, $D_{\text{variational}}$ is a simple closed-form expression. In optimization problems, gradients with respect to the parameters of f and g can be readily computed. In its formulation with ϕ and ψ, alternating between optimization of the variational parameters and the parameters of g leads to an EM algorithm. The method can also be extended to the KL-divergence between hidden Markov models.

The methods of Do and Goldberger in the preceding section can be seen as approximations to this formula, where the ϕ and ψ are a generalization of the matching function. For equal numbers of components, if we restrict $\phi_{b|a}$ and $\psi_{a'|a}$ to have only one non-zero element for a given a, the formula reduces exactly to the chain rule upper bound given in equation (13). For unequal numbers of components, we get a formula similar to $D_{\text{goldberger}}$ except that it satisfies the similarity property.

8. THE VARIATIONAL UPPER BOUND

Here we propose a direct upper bound on the divergence again using a variational approach. We introduce the variational parameters $\phi_{b|a} \geq 0$ and $\psi_{a|b} \geq 0$ satisfying the constraints $\sum_b \phi_{b|a} = \pi_a$ and $\sum_a \psi_{a|b} = \omega_b$. Using the variational parameters we may write

$$\begin{array}{rclcl} f & = & \sum_a \pi_a f_a & = & \sum_{ab} \phi_{b|a} f_a \\ g & = & \sum_b \omega_b g_b & = & \sum_{ab} \psi_{a|b} g_b. \end{array} \tag{21}$$

With this notation we use Jensen's inequality to obtain an upper bound of the KL divergence as follows

$$\begin{aligned} D(f\|g) &= \int f \log(f/g) \\ &= -\int f \log \left(\sum_{ab} \frac{\psi_{a|b} g_b}{\phi_{b|a} f_a} \frac{\phi_{b|a} f_a}{f} \right) \mathrm{d}x \\ &\leq -\sum_{ab} \phi_{b|a} \int f_a \log \left(\frac{\psi_{a|b} g_b}{\phi_{b|a} f_a} \right) \mathrm{d}x \\ &= D(\phi\|\psi) + \sum_{ab} \phi_{b|a} D(f_a \| g_b) \\ &\stackrel{\text{def}}{=} D_{\phi,\psi}(f\|g). \end{aligned} \tag{22}$$

The best possible upper bound can be attained by finding the variational parameters $\hat{\phi}$ and $\hat{\psi}$ that minimize $D_{\phi,\psi}(f\|g)$. The problem is convex in ϕ as well as in ψ so we can fix one and optimize for the other. Fixing ϕ the optimal value for ψ is seen to be

$$\psi_{a|b} = \frac{\omega_b \phi_{b|a}}{\sum_{a'} \phi_{b|a'}}. \tag{23}$$

Similarly, fixing ψ the optimal value for ϕ is

$$\phi_{b|a} = \frac{\pi_a \psi_{a|b} e^{-D(f_a \| g_b)}}{\sum_{b'} \psi_{a|b'} e^{-D(f_a \| g_{b'})}}. \tag{24}$$

At each iteration step the upper bound $D_{\phi,\psi}(f\|g)$ is lowered, and we refer to the convergent as $D_{\text{upper}}(f\|g)$. Since any zeros in ϕ and ψ are fixed under the iteration we recommend starting with $\phi_{b|a} = \psi_{a|b} = \pi_a \omega_b$. This iteration scheme is of the same type as the Blahut-Arimoto algorithm for computing the channel capacity [11, 12], and has similar convergence properties.

Other approximations emerge as special cases from $D_{\phi,\psi}(f\|g)$ for various choices of ϕ, ψ. Firstly, the value $\phi_{b|a} = \psi_{a|b} = \pi_a \omega_b$ yields the convexity bound on $D(f\|g)$, [9]:

$$D(f\|g) \leq \sum_{a,b} \pi_a \omega_b D(f_a \| g_b). \tag{25}$$

Secondly, if $n_f = n_g$ the matched pair bound of equation (13) can be obtained using $\phi_{b|a} = \pi_a$ for $b = a$ and 0 otherwise, and $\psi_{a|b} = \omega_b$ for $b = a$ and 0 otherwise.

Because it is an upper bound, the positivity property $D_{\psi,\phi}(f\|g) > 0$ for $f \neq g$ holds for all ϕ, ψ. Furthermore, $D_{\hat{\psi},\hat{\phi}}(f\|f) = 0$ since the special case of the matching pair bound yields a zero for this case. Assuming convergence to the minimum, which requires initialization in the interior of the constraint surface, $D_{\text{upper}}(f\|g)$ will satisfy the three divergence constraints.

9. EXPERIMENTS

In our experiments we used GMMs from an acoustic model used for speech recognition [2]. The features $x \in \mathbb{R}^d$ are 39 dimensional, $d = 39$, and the GMMs all have diagonal covariance. Furthermore the acoustic model consists of a total of 9,998 gaussians belonging to 826 separate GMMs. The number of gaussians per GMM varies from 1 to 76, of which 5 mixtures attained the lower bound of 1. The median number of gaussians per GMM was 9. We used all combinations of these 826 GMMs to test the various approximations to the KL divergence. Each of the methods was compared to the reference approximation, which is the Monte Carlo method with one million samples, denoted $D_{\text{MC(1M)}}$.

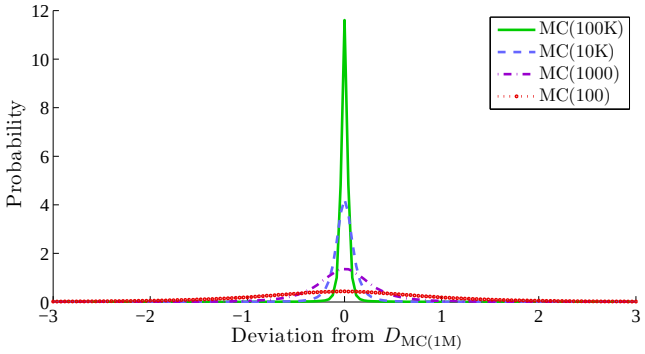

Fig. 1. Distribution of Monte Carlo (MC) approximations, for different numbers of samples, relative to the reference estimate $D_{\text{MC(1M)}}$, computed from all pairs of GMMs.

Figure 1 shows how the accuracy of the Monte Carlo (MC) estimate improves with increasing number of samples. For all the plots,

the horizontal axis represents deviations from $D_{\mathrm{MC(1M)}}$ for each method. The vertical axis represents the probability derived from a histogram of the deviations taken across all pairs of GMMs. Note that even at 100K samples there is still significant deviation from the reference estimate $D_{\mathrm{MC(1M)}}$.

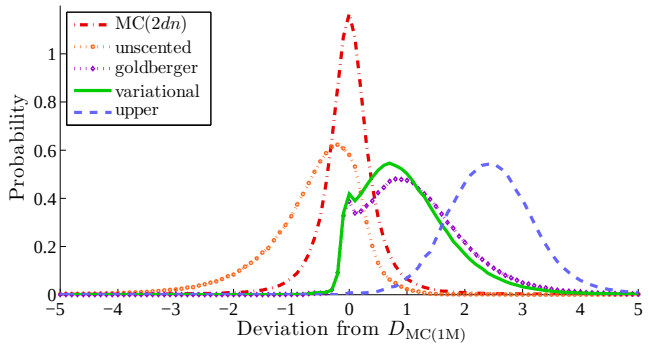

Fig. 2. Distribution of leading approximations to KL divergence relative to the reference estimate $D_{\mathrm{MC(1M)}}$.

Figure 2 shows the corresponding histgrams for $D_{\mathrm{unscented}}$, $D_{\mathrm{variational}}$, $D_{\mathrm{goldberger}}$, D_{upper}, and $D_{\mathrm{MC}(2dn)}$, which is MC with $2dn$ samples, where d is the number of dimensions, and n is the number of gaussians in f. First, note that $D_{\mathrm{unscented}}$ is not as good as $D_{\mathrm{MC}(2dn)}$, despite using the same number of samples, so the MC method seems preferable. Second, notice that $D_{\mathrm{variational}}$ and $D_{\mathrm{goldberger}}$ are similar, but that $D_{\mathrm{variational}}$ is noticeably better. Third, note the small peak at zero, for $D_{\mathrm{variational}}$,$D_{\mathrm{goldberger}}$, and D_{upper}. This stems from certain cases where the approximations become exact, such as with the single-component gaussian mixture models. Fourth note that D_{upper} is an upper bound, and hence has a larger bias; nevertheless it has a small variance.

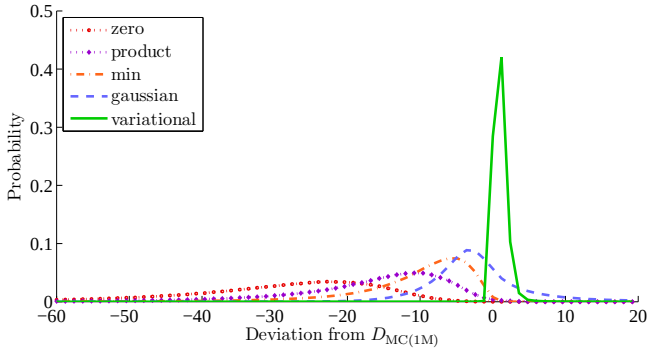

Fig. 3. Distribution of the simple/closed-form approximations to KL divergence relative to the reference estimate $D_{\mathrm{MC(1M)}}$. A trivial lower bound of zero is also included for reference.

Figure 3 plots the distributions of the simple / closed-form approximations, showing that D_{product}, D_{min}, and D_{gaussian} are significantly worse than $D_{\mathrm{variational}}$. The trivial lower-bound of zero is included to illustrate a worst-case scenario. It also indirectly shows the overall distribution of the data.

The simple methods were relatively quick to compute. In our experiments, D_{min}, $D_{\mathrm{goldberger}}$, $D_{\mathrm{variational}}$, and D_{upper}, all took less than 0.1 ms per pair of GMMs. The D_{gaussian}, $D_{\mathrm{unscented}}$, and $D_{\mathrm{MC}(2dn)}$ took around 1 ms per pair. The computation time of Monte Carlo approximations scaled linearly with the sample size, relative to $D_{\mathrm{MC}(2dn)}$, making $D_{\mathrm{MC}(1M)}$ thousands of times more costly than the faster methods.

If accuracy is the primary concern, then MC is clearly best. However, when computation time is an issue, or when gradients need to be evaluated, the proposed methods may be useful. Of the simple, closed-form expressions, the variational approximation, $D_{\mathrm{variational}}$, is the most accurate. The variational upper bound, D_{upper}, is preferable when an upper bound is desired. When bias is not an issue, as when KL-divergences are to be compared with each other, the two variational approximations are equally accurate. Finally, some of the more popular methods, D_{gaussian}, D_{min}, and $D_{\mathrm{unscented}}$, should be avoided, since better alternatives exist.

10. REFERENCES

[1] Solomon Kullback, *Information Theory and Statistics*, Dover Publications Inc., Mineola, New York, 1968.

[2] Peder Olsen and Satya Dharanipragada, "An efficient integrated gender detection scheme and time mediated averaging of gender dependent acoustic models," in *Proceedings of Eurospeech*, Geneva, Switzerland, September 1-4 2003, vol. 4, pp. 2509–2512.

[3] Harry Printz and Peder Olsen, "Theory and practice of acoustic confusability," *Computer, Speech and Language*, vol. 16, pp. 131–164, January 2002.

[4] Jorge Silva and Shrikanth Narayanan, "Average divergence distance as a statistical discrimination measure for hidden Markov models," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 890–906, May 2006.

[5] Qiang Huo and Wei Li, "A DTW-based dissimilarity measure for left-to-right hidden Markov models and its application to word confusability analysis," in *Proceedings of Interspeech 2006 - ICSLP*, Pittsburgh, PA, 2006, pp. 2338–2341.

[6] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, "An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures," in *Proceedings of ICCV 2003*, Nice, October 2003, vol. 1, pp. 487–493.

[7] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul, "An introduction to variational methods for graphical models," *Machine Learning*, vol. 37, no. 2, pp. 183–233, 1999.

[8] Simon Julier and Jeffrey K. Uhlmann, "A general method for approximating nonlinear transformations of probability distributions," Tech. Rep. RRG, Dept. of Engineering Science, University of Oxford, 1996.

[9] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, John Wiley and Sons, New York, NY, 1991.

[10] Minh N. Do, "Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models," *IEEE Signal Processing Letters*, vol. 10, no. 4, pp. 115–118, April 2003.

[11] R. E. Blahut, "Computation of channel capacity and rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-18, pp. 460–473, July 1972.

[12] S. Arimoto, "An algorithm for computing the capacity of arbitrary discrete memoryless channels," *IEEE Trans. Inform. Theory*, vol. 18, pp. 14–20, 1972.