ACCELERATED MONTE CARLO FOR KULLBACK-LEIBLER DIVERGENCE BETWEEN GAUSSIAN MIXTURE MODELS

Jia-Yu Chen, John R. Hershey, Peder A. Olsen and Emmanuel Yashchin

IBM T. J. Watson Research Center

ABSTRACT

Kullback Leibler (KL) divergence is widely used as a measure of dissimilarity between two probability distributions; however, the required integral is not tractable for gaussian mixture models (GMMs), and naive Monte-Carlo sampling methods can be expensive. Our work aims to improve the estimation of KL divergence for GMMs by sampling methods. We show how to accelerate Monte-Carlo sampling using variational approximations of the KL divergence. To this end we employ two different methodologies, control variates, and importance sampling. With control variates we use sampling to estimate the difference between the variational approximation and the the unknown KL divergence. With importance sampling, we estimate the KL divergence directly, using a sampling distribution derived from the variational approximation. We show that with these techniques we can achieve improvements in accuracy equivalent to using a factor of 30 times more samples.

Index Terms— Kullback Leibler divergence, variational methods, gaussian mixture models, control variates, antithetic variates, importance sampling.

1. INTRODUCTION

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

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

is commonly used in statistics as a measure of similarity between two density distributions. 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 word models 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 gaussian mixture models (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 $D(f_a\|g_b)$.

2. MONTE CARLO SAMPLING

In the KL divergence, we can separately estimate each component

$$D_a \stackrel{\text{def}}{=} \int D_a(x) \, \mathrm{d}x \stackrel{\text{def}}{=} \int f_a(x) \left(\log f(x) - \log g(x)\right) \, \mathrm{d}x, \tag{4}$$

so that $D(f\|g) = \sum_a \pi_a D_a$. To estimate $\int D_a(x) \, \mathrm{d}x$ using importance sampling, we define a sampling distribution h, with random variable $X \sim h(x)$ and evaluate the expected value,

$$D_a = \int h(x) \frac{D_a(x)}{h(x)} \, \mathrm{d}x = E_h \frac{D_a(X)}{h(X)} = E_h D_a^h(X). \tag{5}$$

To estimate this expected value we take Monte Carlo (MC) samples X_i from $h(x)$, and evaluate the sample mean;

$$\hat{D_a} = \frac{1}{n} \sum_{i=1}^{n} \frac{D_a(X_i)}{h(X_i)} \to D_a, \tag{6}$$

as $n \to \infty$, by the law of large numbers. The estimation error,

$$\mathrm{Std}(\hat{D}_a) = \frac{1}{\sqrt{n}} \left(\int \frac{D_a^2(x)}{h(x)} dx - D_a^2 \right)^{\frac{1}{2}}, \tag{7}$$

depends on $h(x)$. It is easy to see that, for a positive $D_a(x)$, the optimal sampling distribution would be $h(x) = \frac{D_a(x)}{\int D_a(x) dx}$ which reduces the estimation error to zero. Of course we cannot use this distribution, since it requires computing the integral we wish to evaluate, but it suggests that our sampling distribution should approximate the function that we are integrating. Intuitively this makes sense, since it places more samples where the function is furthest from zero.

It is convenient to use $f_a(x)$ as a proposal distribution, which yields the estimate

$$\hat{D_a} = \frac{1}{n} \sum_{i=1}^{n} \log \frac{f(X_i)}{g(X_i)}. \tag{8}$$

This Monte Carlo estimate serves as our baseline. We can improve convergence if we come up with a sampling distributions close to $\frac{D_a(x)}{\int D_a(x) dx}$. Thus we consider approximations $\tilde{D}_a(x) \approx D_a(x)$ that have a known integral, $\tilde{D}_a \stackrel{\text{def}}{=} \int \tilde{D}_a(x) \, \mathrm{d}x$, and that can be decomposed into a difference of positive functions.

3. CONTROL VARIATES

Another well-known method for variance reduction in Monte Carlo is the use of control variates [7], which also requires integrable approximations $\tilde{D}_a(x)$. Control variates have the advantage that one merely has to evaluate $\tilde{D}_a(x)$ and its integral, rather than sampling from it, and that the approximation $\tilde{D}_a(x)$ need not be composed of positive functions. With control variates, instead of estimating the KL divergence directly we first remove a known quantity given by an approximation to $D_a(x)$, introducing a scaling parameter β_a to obtain the closest approximation.

$$D_a = \beta_a \tilde{D}_a + \int \left(D_a(x) - \beta_a \tilde{D}_a(x) \right) \, \mathrm{d}x \tag{9}$$

If $\tilde{D}_a(x)$ is a good approximation, the integral of the remainder can then be estimated by Monte Carlo with lower variance, as illustrated in Figure 1.

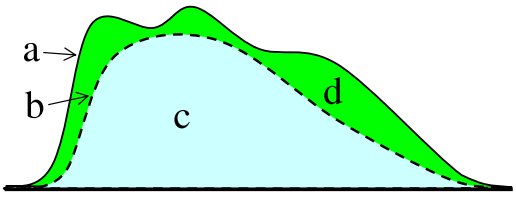

Fig. 1. Integration using control variates: a) (*solid*) the function to be integrated, $D_a(x)$, b) (*dashed*) the control variate approximation, $\beta_a \tilde{D}_a(x)$, c) (*blue*) the known integral of the approximation, $\int \beta_a \tilde{D}_a(x) \, \mathrm{d}x$, and d) (*green*) the remaining area to be estimated, $\int D_a(x) - \beta_a \tilde{D}_a(x) \, \mathrm{d}x$.

We formulate the integral as an expected value with respect to the sampling distribution $h(x)$ of random variable X.

$$D_a = \beta_a \tilde{D}_a + E_h \frac{D_a(X) - \beta_a \tilde{D}_a(X)}{h(X)} \tag{10}$$

A random variate $\tilde{D}_a(X)/h(X)$, with known expected value, $\tilde{D}_a$ is said to be a *control variate* for $D_a(X)/h(X)$ if the two are correlated. Because of this correlation, the second term is smaller than the overall integral, and can be estimated using Monte Carlo:

$$\hat{D}_a = \hat{\beta}_a \tilde{D}_a + \frac{1}{n} \sum_i \frac{D_a(X_i) - \hat{\beta}_a \tilde{D}_a(X_i)}{h(X_i)}. \tag{11}$$

Often we may know the appropriate value of β_a *a priori*. For instance if the approximation is good, as it is with the variational approximations, $\beta_a = 1$ works well. Note that if $\beta_a = 0$, then (11) reduces to the standard Monte Carlo estimation. However we can also empirically optimize β_a to minimize the estimation error. The best value of $\beta_a \tilde{D}_a(x)$ is of course $D_a(x)$, which trivially reduces the estimation error to zero. This motivates using approximations to $D_a(x)$ as control variates.

The relationship between control variates and importance sampling is subtle. Given a sampling distribution $h(x)$, consider using the same function to construct the control variate, so that

$$h(x) = \bar{D}_a(x) \stackrel{\text{def}}{=} \frac{\tilde{D}_a(x)}{\int \tilde{D}_a(x) \, \mathrm{d}x}, \tag{12}$$

then the control variate cancels, and (10) reduces to importance sampling:

$$D_a = \beta_a \tilde{D}_a + E_{\bar{D}_a} \frac{D_a(X) - \beta_a \tilde{D}_a(X)}{\bar{D}_a(X)} = E_h \frac{D_a(X)}{h(X)}$$

However, given a particular control variate, $\beta_a \tilde{D}_a(x)$ the optimal sampling distribution $h(x)$ would be $D_a(x) - \beta_a \tilde{D}_a(x)$, rather than $D_a(x)$. Similarly, given a sampling distribution, $h(x)$, the optimal control variate is $D_a(x) - h(x)$.

4. ANTITHETIC VARIATES

Antithetic variates provide a complementary means of reducing sampling variance [7]. Antithetic variates take advantage of the symmetry of $f_a(x)$ to draw two simultaneous samples. If X is a random variable drawn from $h(x) = f_a(x)$, then $2\mu_a - X$ has the same distribution. By storing some pre-computed constants, we can efficiently evaluate g_b and $f_{a'}$ for both values x and $2\mu_a - x$. Thus we are guaranteed an improvement equivalent to at least doubling the number of sample points. In some cases there may be further gains due to the symmetry.

5. THE VARIATIONAL APPROXIMATION

The variational approximation [8], consists of a difference of lower bounds on the expected likelihoods $L_a^{(f)}(x) \stackrel{\text{def}}{=} f_a(x) \log f(x)$ and $L_a^{(g)}(x) \stackrel{\text{def}}{=} f_a(x) \log g(x)$. We introduce non-negative variational parameters, $\phi_{b|a}$, such that $\sum_b \phi_{b|a} = 1$. By Jensen's inequality we have

$$L_a^{(g)}(x) \geq f_a(x) \sum_{ab} \pi_a \phi_{b|a} \log \frac{\omega_b g_b(x)}{\phi_{b|a}} \stackrel{\text{def}}{=} \tilde{L}_a^{\mathrm{va}(g)}(x)$$

where

$$\hat{\phi}_{b|a} = \frac{\omega_b e^{-D(f_a \| g_b)}}{\sum_{b'} \pi_{b'} e^{-D(f_a \| g_{b'})}} \tag{13}$$

yields the tightest possible bound.

An integrable approximation to $D_a(x)$ is thus

$$\tilde{D}_a^{\mathrm{va}}(x) = \tilde{L}_a^{\mathrm{va}(f)}(x) - \tilde{L}_a^{\mathrm{va}(g)}(x), \tag{14}$$

and the variational KL divergence is

$$D(f \| g) \approx \sum_a \pi_a \left(\int \tilde{L}_a^{\mathrm{va}(f)}(x) \, \mathrm{d}x - \int \tilde{L}_a^{\mathrm{va}(g)}(x) \, \mathrm{d}x \right). \tag{15}$$

6. THE VARIATIONAL UPPER BOUND

An upper bound to the KL divergence is also introduced in [8]. Let the variational parameters $\phi_{b|a} \geq 0$ and $\psi_{a|b} \geq 0$ satisfy the constraints $\sum_b \phi_{b|a} = \pi_a$ and $\sum_a \psi_{a|b} = \omega_b$. Then the following inequality holds:

$$\begin{aligned} D(f \| g) &\leq -\sum_{ab} \phi_{b|a} \left(\log \frac{\psi_{a|b}}{\phi_{b|a}} + D(f_a \| g_b) \right) \\ &\stackrel{\text{def}}{=} D^{vb}(f \| g). \end{aligned} \tag{16}$$

To optimize this bound we iterate until convergence

$$\psi_{a|b} = \frac{\omega_b \phi_{b|a}}{\sum_{a'} \phi_{b|a'}}, \quad \text{and} \quad \phi_{b|a} = \frac{\pi_a \psi_{a|b} e^{-D(f_a \| g_b)}}{\sum_{b'} \psi_{a|b'} e^{-D(f_a \| g_{b'})}}. \quad (17)$$

An integrable upper bound approximation to $D_a(x)$ is

$$\tilde{D}_a^{\mathrm{vb}}(x) = \frac{f_a(x)}{\pi_a} \sum_b \phi_{b|a} \log \frac{\phi_{b|a} f_a(x)}{\psi_{a|b} g_b(x)}. \quad (18)$$

7. VARIATIONAL CONTROL VARIATES

Using (14) and (18) as control variates in (11) allows us to extend the variational methods to achieve arbitrary accuracy via sampling.

8. TAYLOR SERIES

Another straightforward approximation can be made using the Taylor series.

$$D(f\|g) \approx \int \sum_a \pi_a f_a(x) \left(T_{\mu_a}^{(f)}(x) - T_{\mu_a}^{(g)}(x) \right) \mathrm{d}x. \quad (19)$$

where $T_{\mu_a}^{(f)}(x)$ and $T_{\mu_a}^{(g)}(x)$ are second-order Taylor expansions of $\log f(x)$ and $\log g(x)$, around μ_a. We can then define an approximation:

$$\tilde{D}_a^{\mathrm{ts}}(x) = f_a(x) \left(T_{\mu_a}^{(f)}(x) - T_{\mu_a}^{(g)}(x) \right) \quad (20)$$

When using $\tilde{D}_a^{\mathrm{ts}}(x)$ with antithetical variates, errors in the odd-order terms cancel, significantly improving efficiency.

9. VARIATIONAL IMPORTANCE SAMPLING

Although the variational approximation $\tilde{D}_a(x)$ is not positive, it can be decomposed into a difference of functions $D_a(x) = L_a^{(f)}(x) - L_a^{(g)}$, each of which can be approximated using the variational approximations, $\tilde{L}_a^{\mathrm{va}(f)}(x)$ and $\tilde{L}_a^{\mathrm{va}(g)}(x)$. These in turn can be decomposed into positive functions by separating out the constants. We have

$$\tilde{L}_a^{\mathrm{va}(g)}(x) = f_a(x) C_a^{\mathrm{va}(g)} - f_a(x) Q_a^{\mathrm{va}(g)}(x), \quad (21)$$

where we define the constant

$$C_a^{\mathrm{va}(g)} \stackrel{\mathrm{def}}{=} \sum_b \phi_{b|a} \left(\log \left(\frac{\pi_b}{\phi_{b|a}} \right) - \frac{1}{2} \log \left((2\pi)^d |\Sigma_b| \right) \right), \quad (22)$$

and the positive quadratic function,

$$Q_a^{\mathrm{va}(g)}(x) = \frac{1}{2} \sum_b \phi_{b|a} (x - \mu_b)^T \Sigma_b^{-1} (x - \mu_b). \quad (23)$$

The gaussian weighted quadratic function $f_a(x) Q_a^{\mathrm{va}(g)}(x)$ is positive and can be integrated to construct a sampling distribution,

$$h^{\mathrm{va}(g)}(x) = \frac{f_a(x) Q_a^{\mathrm{va}(g)}(x)}{\int f_a(x) Q_a^{\mathrm{va}(g)}(x)}. \quad (24)$$

The desired integral of the likelihood $L_a^{(g)}(x)$ can then put into the form,

$$\begin{aligned} L_a^{(g)} &= \int L_a^{(g)}(x)\, \mathrm{d}x \quad (25) \\ &= C_a^{\mathrm{va}(g)} - \int \left(f_a(x) C_a^{\mathrm{va}(g)} - L_a^{(g)}(x) \right) \mathrm{d}x \\ &= C_a^{\mathrm{va}(g)} - E_{h^{\mathrm{va}(f)}} \frac{f_a(X) C_a^{\mathrm{va}(g)} - L_a^{(g)}(X)}{h^{\mathrm{va}(g)}(X)} \\ &\approx C_a^{\mathrm{va}(g)} - \frac{1}{n} \sum_i \frac{f_a(X_i) C_a^{\mathrm{va}(g)} - L_a^{(g)}(X_i)}{h^{\mathrm{va}(g)}(X_i)} \end{aligned}$$

where the X_i have the distribution $h^{\mathrm{va}(f)}(x)$. Note that $f_a(X) C_a^{\mathrm{va}(g)} / h^{\mathrm{va}(f)}(X)$ can be interpreted as a control variate, with $\beta_a = 1$. The importance sampling distribution is also well matched to the function being estimated,

$$\begin{aligned} h^{\mathrm{va}(f)}(x) = f_a(x) Q_a^{\mathrm{va}(g)}(x) &= f_a(x) C_a^{\mathrm{va}(g)} - \tilde{L}_a^{\mathrm{va}(g)}(x) \\ &\approx f_a(x) C_a^{\mathrm{va}(g)} - L_a^{(g)}(x), \quad (26) \end{aligned}$$

so the estimate will be efficient. A similar derivation applies to $L_a^{(f)}(x)$.

Now we are left with the non-trivial task of sampling from $h^{\mathrm{va}(g)}(x)$. We proceed by using a change of variables to map $f_a(x)$ to a standard normal distribution $\phi(y)$. The cumulative distribution can then be formulated for each term in the quadratic and added together to yield the total cumulative distribution, which is then cast in terms of x, yielding $G_a(x)$. One then samples from the uniform distribution $z_i \sim U(0,1)$ and solves for the value x_i such that $G_a(x_i) = z_i$. To sample from the variational upper bound, one could separate the $\log f$ and $\log g$ components, as in the variational approximation, but the resulting approximation would no longer be a bound. Sampling the upper bound without separating the components poses further difficulties because of the positivity requirement.

10. EXPERIMENTS

In our experiments we used 826 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. GMMs. The number of Gaussians per GMM varies from 1 to 76, with a median of 9. There were 9,998 Gaussians in total. 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 $D_{\mathrm{MC(1M)}}$.

Figure 2 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 $D_{\mathrm{MC(1M)}}$.

Figure 3 shows the results of the closed-form variational approximations. Note that D^{va} performs as well as Monte Carlo using somewhere between 100 and 1000 samples. D^{vb} is comparable to D^{va}, but since it is an upper bound it has a larger bias.

Figure 4 shows histograms for various sampling techniques using 1000 samples. The Taylor series method worked best with the antithetic control variates, and performs about as well as straightforward Monte Carlo with 10K samples. The antithetic technique was

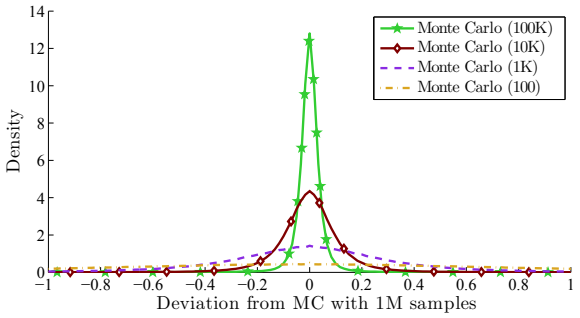

Fig. 2. Distribution of Monte Carlo (MC) approximations, for different numbers of samples, relative to the reference estimate $D_{\mathrm{MC(1M)}}$, computed from all pairs of GMMs.

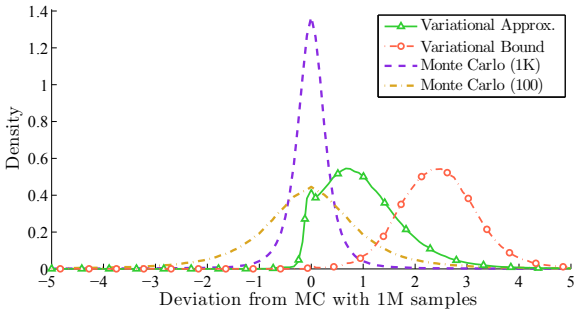

Fig. 3. Distribution of the closed-form approximations to KL divergence relative to the reference estimate $D_{\mathrm{MC(1M)}}$.

necessary here to get a significant improvement over the baseline, Monte Carlo with 1000 samples. The variational control variate, using $\tilde{D}_a^{\mathrm{va}}(x)$, performs significantly better on its own, and with the antithetic technique yields the best of all the methods, giving an accuracy comparable to Monte Carlo with 30K samples. The variational upper bound did not yield a significant improvement relative to the variational approximation. The variational importance sampling, using $\tilde{D}_a^{\mathrm{va}}(x)$, is nearly as good, but since it is more complicated to implement, it fails to earn its keep.

The computation time of the variational methods is quadratic in the number of gaussians in f and g, whereas Monte Carlo sampling is not. But the variational approximations can be aggressively pruned to have the same number of components as $f + g$. The cost of evaluating each sample when using variational methods is then comparable to evaluating one sample for Monte Carlo. Since the total cost of the variational methods still has a fixed quadratic cost, the algorithm with best execution time will depend on the size of the mixtures f and g and on the number of samples. As a rule of thumb, the variational approximations with control variates will generally outperform the other methods if the number of samples is larger than the number of components in $f + g$. Importance sampling is somewhat slower than the other methods as it requires an inversion of a cumulative distribution function at each step.

The antithetic technique gives a performance boost to each control variate method equivalent to doubling the number of samples, or better. However due to the symmetry involved in the extra samples, the additional cost can be almost completely absorbed into a precomputation step.

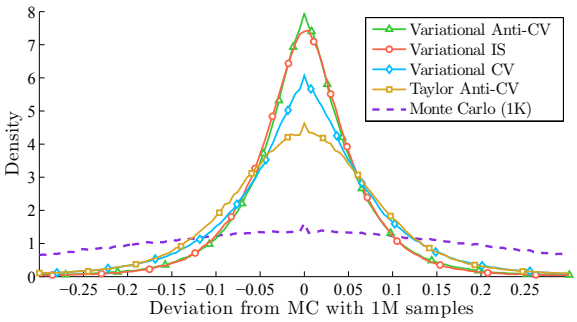

Fig. 4. Distribution of the sampling approximations to KL divergence relative to the reference estimate $D_{\mathrm{MC(1M)}}$.

11. CONCLUSION

In this work we demonstrate how to apply Taylor expansion, variational approximation, and variational upper bound approximation to accelerate sampling methods using control and antithetic variates. These methods allow previously known closed-form approximations to attain arbitrary accuracy given sufficient computational resources. Experimental results shows that these methods substantially outperform the standard Monte Carlo method. An alternative to control variates is variational importance sampling, which performs well but is substantially more complex. We anticipate that the novel principle of combining variational approximation with control and antithetic variates, will have far-reaching applications in probabilistic inference and estimation.

12. 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] R.Y. Rubinstein, *Simulation and the Monte Carlo Method*, Wiley, 1981.

[8] John Hershey and Peder Olsen, "Approximating the Kullback Leibler divergence between gaussian mixture models," in *ICASSP*, Honolulu, Hawaii, April 2007.