UPPER AND LOWER BOUNDS FOR APPROXIMATION OF THE KULLBACK-LEIBLER DIVERGENCE BETWEEN HIDDEN MARKOV MODELS

Haiyang Li, Jiqing Han, Tieran Zheng, Guibin Zheng

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China

ABSTRACT

The Kullback-Leibler (KL) divergence is often used for a similarity comparison between two Hidden Markov models (HMMs). However, there is no closed form expression for computing the KL divergence between HMMs, and it can only be approximated. In this paper, we propose two novel methods for approximating the KL divergence between the left-to-right transient HMMs. The first method is a product approximation which can be calculated recursively without introducing extra parameters. The second method is based on the upper and lower bounds of KL divergence, and the mean of these bounds provides an available approximation of the divergence. We demonstrate the effectiveness of the proposed methods through experiments including the deviations to the numerical approximation and the task of predicting the confusability of phone pairs. Experimental resuls show that the proposed product approximation is comparable with the current variational approximation, and the proposed approximation based on bounds performs better than current methods in the experiments.

Index Terms— Kullback-Leibler divergence, Hidden Markov model, automatic speech recognition, speech processing.

1. INTRODUCTION

Hidden Markov model (HMM) has been successfully used as a powerful tool in speech recognition and signal processing. The reasons for this success are due to its effectiveness in modeling significant and complex time series with a small set of parameters, and there are efficient estimation techniques to train and evaluate these parameters for the given data set.

In the problems of clustering or classification, it is often necessary to compare different HMMs through a suitable distance or similarity measure. The Kullback-Leibler (KL) divergence [1], also known as the relative entropy between two probability density distributions in statistics, has been used as a measure of similarity between two HMMs. Since there is no closed-form expression of the KL divergence for the HMMs, the Monte Carlo simulation is employed to numerically approximate the KL divergence [2]. Although the Monte Carlo approximation is easy to implement, this method is slow and inefficient. To overcome this problem, several other methods are used to approximate the KL divergence between HMMs. Recent methods include the probabilistic evaluation of the match between every pair of states [3] and the HMM stationary cumulative distribution [4], but these methods are only employed for the HMMs with stationary distribution. For the aspects of automatic speech recognition (ASR), many approximations of the KL divergence are explored between the left-to-right transient HMMs [5, 6, 7]. The method of average divergence distance [5] is based on the transient behavior, but no rigorous relationship with the divergence is stipulated for this method. The work in [6] provides the divergence with only an upper bound. In [7], a variational approximation is derived from the variational methods for mixture models, however, this approximation employs only single Gaussian to model the observation probability at each HMM state for shorthand. Furthermore, the variational approximation depends on a group of variational parameters, which need estimating by a recursive algorithm.

In this paper, we propose two novel methods for approximating the KL divergence between the left-to-right transient HMMs. The first method is a product approximation which can be calculated recursively without extra parameters. The second method is derived from the upper and lower bounds of KL divergence, and the mean of these bounds provides an approximation of KL divergence. While our work formulates the KL divergence with the same HMM definitions as [7], and extends the variational approximation to the case of HMMs with Gaussian mixture models (GMMs) modelling the observation probabilities at states. Moreover, our work can be considered as the expansions of the methods for approximating the KL divergence between GMMs in [8] and [9]. We finally confirm the effectiveness of the proposed methods by experiments.

2. KULLBACK-LEIBLER DIVERGENCE FOR HMMS

The KL divergence can be employed to measure the similarity between two left-to-right transient HMMs used in ASR. To formulate the KL divergence, we follow the definitions of HMMs presented in [7], and this definition method yields a distribution (integrates to one) over all lengths of observation sequence. Suppose that f is a left-to-right transient HMM, and it emits an observation sequence $x_{1:n}$ of length n. The sequence $x_{1:n}$ can be expressed as $x_{1:n} = (x_1, ..., x_n)$, where x_t is an observed vector with $x_t \in \mathbb{R}^d$, and d is the dimension of the vector. The observation probability $f(x_{1:n})$ assigned to a particular observed sequence $x_{1:n}$ can be computed as [7]:

$$\begin{aligned} f(x_{1:n}) &= \sum_{a_{1:n}} \pi_{a_{1:n}} f_{a_{1:n}}(x_{1:n}) \\ &= \sum_{a_{1:n}} \pi_{a_1|a_{\mathcal{I}}} \pi_{a_{\mathcal{F}}|a_n} f_{a_1}(x_1) \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} f_{a_t}(x_t) \end{aligned} \quad (1)$$

where $a_{1:n} = (a_1, ..., a_n)$ is the corresponding hidden state sequence with n emitting states, and a_t takes values in the set of emitting states of f. The non-emitting initial and final states are defined as $a_{\mathcal{I}}$ and $a_{\mathcal{F}}$ respectively. The probability of state sequence is formulated as a Markov chain $\pi_{a_{1:n}} = \pi_{a_1|a_{\mathcal{I}}} \pi_{a_{\mathcal{F}}|a_n} \prod_{t=2}^{n} \pi_{a_t|a_{t-1}}$, where $\pi_{a_1|a_{\mathcal{I}}}$ and $\pi_{a_{\mathcal{F}}|a_n}$ are the initial and final state transitions, and $\pi_{a_t|a_{t-1}}$ is the transition probability. Let $f_{a_{1:n}}(x_{1:n})$ be the

This research is supported by the National Natural Science Foundation of China (No. 91120303) and the Ph.D. Programs Foundation of Ministry of Education of China (No. 20112302110042).

observation probability of $x_{1:n}$ with the hidden state sequence $a_{1:n}$, and it can be calculated as the product of the probabilities given by the states as $f_{a_{1:n}}(x_{1:n}) = \prod_{t=1}^{n} f_{a_t}(x_t)$. For each state a_t, GMM is usually used to model the probability density $f_{a_t}(x)$ of observation x as $f_{a_t}(x) = \sum_i c_i^{a_t} N(x; \mu_i^{a_t}, \Sigma_i^{a_t})$, where $c_i^{a_t}$ ($\sum_i c_i^{a_t} = 1$) is the prior probability of mixture component, $N(x; \mu_i^{a_t}, \Sigma_i^{a_t})$ is a Gaussian with mean $\mu_i^{a_t}$ and variance $\Sigma_i^{a_t}$.

According to the above definitions, the KL divergence between two HMMs can be formulated. Let f and g be two left-to-right transient HMMs, and $\mathbf{X}$ be the set of all possible observed sequences generated by the HMMs. Thus, $\mathbf{X}$ can be expressed as $\mathbf{X} = \cup_{n=1}^{\infty} \mathbb{R}^{n \times d} = \cup_{n=1}^{\infty} \{x_{1:n} \mid x_t \in \mathbb{R}^d, 1 \le t \le n\}$. Furthermore, let $f(\mathbf{x})$ and $g(\mathbf{x})$ be the probability density functions of a particular sequence $\mathbf{x}$ ($\mathbf{x} \in \mathbf{X}$) for the HMMs f and g respectively. The KL divergence between two HMMs is defined as:

$$\begin{aligned} D(f \parallel g) &\stackrel{\text{def}}{=} \int f(\mathbf{x}) \log \frac{f(\mathbf{x})}{g(\mathbf{x})} d\mathbf{x} \\ &= \sum_{n=1}^{\infty} \int f(x_{1:n}) \log \frac{f(x_{1:n})}{g(x_{1:n})} dx_{1:n} \end{aligned} \tag{2}$$

where $f(x_{1:n})$ is given by Eq. (1), and likewise $g(x_{1:n})$ is defined as $g(x_{1:n}) = \sum_{b_{1:n}} \omega_{b_{1:n}} g_{b_{1:n}}(x_{1:n})$ with the probability $\omega_{b_{1:n}}$ of state sequence $b_{1:n}$ for HMM g. Thus $D(f \parallel g)$ can be obtained by separately integrating over sequences of each length and then summing over the individual results. And it can be decomposed as:

$$\begin{aligned} D(f \parallel g) &= \sum_{n=1}^{\infty} \int f(x_{1:n}) \log \frac{f(x_{1:n})}{g(x_{1:n})} dx_{1:n} \\ &= \sum_{n=1}^{\infty} \Big[\int f(x_{1:n}) \log f(x_{1:n}) \, dx_{1:n} \\ &\qquad - \int f(x_{1:n}) \log g(x_{1:n}) \, dx_{1:n} \Big] \\ &= \sum_{n=1}^{\infty} [L_n(f \parallel f) - L_n(f \parallel g)] \end{aligned} \tag{3}$$

where $L_n(f \parallel g) \stackrel{\text{def}}{=} \int f(x_{1:n}) \log g(x_{1:n}) \, dx_{1:n}$. Hence, it shows that the approximation of $D(f \parallel g)$ can be performed by approximating $L_n(f \parallel g)$.

3. APPROXIMATIONS OF THE KULLBACK-LEIBLER DIVERGENCE BETWEEN HMMS

3.1. The variational approximation of the KL divergence

To approximate the KL divergence between two HMMs, a variational approximation is used in [7]. This method introduces a group of variational parameters $\phi_{b_{1:n}|a_{1:n}}$ ($\phi_{b_{1:n}|a_{1:n}} \ge 0$) in the form of conditional Markov chain as $\phi_{b_{1:n}|a_{1:n}} \stackrel{\text{def}}{=} \phi_{b_1|a_1} \prod_{t=2}^{n} \phi_{b_t|a_t b_{t-1}}$, where $\sum_{b_1} \phi_{b_1|a_1} = 1$, $\sum_{b_t} \phi_{b_t|a_t b_{t-1}} = 1$, and $\sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} = 1$. And a lower bound of $L_n(f \parallel g)$ is obtained as [7]:

$$\begin{aligned} &L_n(f \parallel g) \\ &\ge \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} \log \frac{\omega_{b_{1:n}} \exp\left(\sum_{t=1}^{n} L(f_{a_t} \parallel g_{b_t})\right)}{\phi_{b_{1:n}|a_{1:n}}} \end{aligned} \tag{4}$$

where $L(f_{a_t} \parallel g_{b_t}) \stackrel{\text{def}}{=} \int f_{a_t}(x_t) \log g_{b_t}(x_t) \, dx_t$. $f_{a_t}(x_t)$ and $g_{b_t}(x_t)$ are the observation probabilities for the states a_t and b_t in two HMMs f and g respectively. In [7], only single Gaussian is employed to model $f_{a_t}(x_t)$ and $g_{b_t}(x_t)$ for shorthand. Here, we extend the observation probabilities to the case of GMMs. For two GMMs $f_{a_t}(x_t)$ and $g_{b_t}(x_t)$, $L(f_{a_t} \parallel g_{b_t})$ has a lower bound $L^{var}(f_{a_t} \parallel g_{b_t})$ according to the variational approximation between two GMMs [8]. Consequently, we get a lower bound $L_n^{var}(f \parallel g)$ for the right part of Eq. (4):

$$\begin{aligned} &\sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} \log \frac{\omega_{b_{1:n}} \exp\left(\sum_{t=1}^{n} L(f_{a_t} \parallel g_{b_t})\right)}{\phi_{b_{1:n}|a_{1:n}}} \\ &\ge \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \phi_{b_{1:n}|a_{1:n}} \log \frac{\omega_{b_{1:n}} \exp\left(\sum_{t=1}^{n} L^{var}(f_{a_t} \parallel g_{b_t})\right)}{\phi_{b_{1:n}|a_{1:n}}} \\ &\stackrel{\text{def}}{=} L_n^{var}(f \parallel g) \end{aligned} \tag{5}$$

Thus $L_n^{var}(f \parallel g)$ is also a lower bound of $L_n(f \parallel g)$. $L_n^{var}(f \parallel g)$ can be maximized with respect to $\phi_{b_{1:n}|a_{1:n}}$ through a recursive algorithm [7]. Similarly, a lower bound of $L_n(f \parallel f)$ can also be obtained as $L_n^{var}(f \parallel f)$ with the same method. Then, the variational approximation $D^{var}(f \parallel g)$ for HMMs can be computed as:

$$D^{var}(f \parallel g) = \sum_{n=1}^{\infty} [L_n^{var}(f \parallel f) - L_n^{var}(f \parallel g)] \tag{6}$$

And the computation is truncated to a finite series in practice.

3.2. The product approximation of the KL divergence

In this subsection, we propose a product approximation method for the KL divergence between HMMs, and it is also based on the decomposition of KL divergence. $p_f(n)$ is defined as the probability of a particular sequence length n for HMM f:

$$\begin{aligned} p_f(n) &= \int f(x_{1:n}) dx_{1:n} = \int \sum_{a_{1:n}} \pi_{a_{1:n}} f_{a_{1:n}}(x_{1:n}) \, dx_{1:n} \\ &= \sum_{a_{1:n}} \pi_{a_{1:n}} = \sum_{a_{1:n}} \pi_{a_1|a_{\mathcal{I}}} \pi_{a_{\mathcal{F}}|a_n} \prod_{t=2}^{n} \pi_{a_t|a_{t-1}} \end{aligned} \tag{7}$$

It can be calculated with a forward or backward recursion. Note that

$$\int \sum_{a_{1:n}} \frac{\pi_{a_{1:n}}}{p_f(n)} f_{a_{1:n}}(x_{1:n}) \, dx_{1:n} = 1 \tag{8}$$

According to Jensen's inequality and Eq. (8), an upper bound of $L_n(f \parallel g)$ is derived as $L_n^{prod}(f \parallel g)$.

$$\begin{aligned} &L_n(f \parallel g) = \int f(x_{1:n}) \log g(x_{1:n}) \, dx_{1:n} \\ &= \int \sum_{a_{1:n}} \pi_{a_{1:n}} f_{a_{1:n}}(x_{1:n}) \log \sum_{b_{1:n}} \omega_{b_{1:n}} g_{b_{1:n}}(x_{1:n}) \, dx_{1:n} \\ &= p_f(n) \int \sum_{a_{1:n}} \frac{\pi_{a_{1:n}}}{p_f(n)} f_{a_{1:n}}(x_{1:n}) \log \sum_{b_{1:n}} \omega_{b_{1:n}} g_{b_{1:n}}(x_{1:n}) \, dx_{1:n} \\ &\le p_f(n) \log \int \sum_{a_{1:n}} \frac{\pi_{a_{1:n}}}{p_f(n)} f_{a_{1:n}}(x_{1:n}) \sum_{b_{1:n}} \omega_{b_{1:n}} g_{b_{1:n}}(x_{1:n}) \, dx_{1:n} \\ &= p_f(n) \Bigg[\log \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \omega_{b_{1:n}} \int f_{a_{1:n}}(x_{1:n}) \, g_{b_{1:n}}(x_{1:n}) \, dx_{1:n} \\ &\qquad - \log p_f(n) \Bigg] \\ &\stackrel{\text{def}}{=} L_n^{prod}(f \parallel g) \end{aligned} \tag{9}$$

In Eq. (9), the product integration is calculated as:

$$\int f_{a_{1:n}}(x_{1:n})\, g_{b_{1:n}}(x_{1:n})\, dx_{1:n} = \int \prod_{t=1}^{n} f_{a_t}(x_t)\, g_{b_t}(x_t)\, dx_{1:n}$$
$$= \prod_{t=1}^{n} \int f_{a_t}(x)\, g_{b_t}(x)\, dx = \prod_{t=1}^{n} I(a_t, b_t) \quad (10)$$

where $I(a_t, b_t) \stackrel{\text{def}}{=} \int f_{a_t}(x)\, g_{b_t}(x)\, dx$, and it is used to define the integration of product between the two observation probabilities at states a_t and b_t. For two GMMs $f_{a_t}(x)$ and $g_{b_t}(x)$, the integration of product $I(a_t, b_t)$ can be given with a closed form expression described in Appendix.

The accumulation in Eq. (9) can be expressed as $Q(a_{1:n}, b_{1:n})$:

$$\begin{aligned} & Q(a_{1:n}, b_{1:n}) \\ =& \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \omega_{b_{1:n}} \int f_{a_{1:n}}(x_{1:n})\, g_{b_{1:n}}(x_{1:n})\, dx_{1:n} \\ =& \sum_{a_{1:n}} \pi_{a_{1:n}} \sum_{b_{1:n}} \omega_{b_{1:n}} \prod_{t=1}^{n} I(a_t, b_t) \\ =& \sum_{a_1} \pi_{a_1|a_{\mathcal{I}}} \sum_{b_1} \omega_{b_1|b_{\mathcal{I}}} I(a_1, b_1) \\ & \times \sum_{a_2} \pi_{a_2|a_1} \sum_{b_2} \omega_{b_2|b_1} I(a_2, b_2) \times \cdots \\ & \times \sum_{a_n} \pi_{a_n|a_{n-1}} \pi_{a_{\mathcal{F}}|a_n} \sum_{b_n} \omega_{b_n|b_{n-1}} \omega_{b_{\mathcal{F}}|b_n} I(a_n, b_n) \end{aligned} \quad (11)$$

It can be computed recursively by defining $Q_t(a_t, b_t)$ as the contribution from earlier states to the current estimate at states a_t and b_t.

$$Q_1(a_1, b_1) = \pi_{a_1|a_{\mathcal{I}}} \omega_{b_1|b_{\mathcal{I}}} \quad (12)$$

$$Q_t(a_t, b_t) = \sum_{a_{t-1}} \pi_{a_t|a_{t-1}} \sum_{b_{t-1}} \omega_{b_t|b_{t-1}} I(a_{t-1}, b_{t-1})\, Q_t(a_{t-1}, b_{t-1}) \quad (13)$$

The end case is handled as:

$$Q(a_{1:n}, b_{1:n}) = \sum_{a_n} \pi_{a_{\mathcal{F}}|a_n} \sum_{b_n} \omega_{b_{\mathcal{F}}|b_n} I(a_t, b_t)\, Q_n(a_n, b_n) \quad (14)$$

The sum can be computed recursively by saving intermediate results.

Then, an upper bound $L_n^{prod}(f \parallel g)$ is computed from Eq. (9) to Eq. (14) for $L_n(f \parallel g)$ without introducing extra parameters. Similarly, an upper bound of $L_n(f \parallel f)$ can also be obtained as $L_n^{prod}(f \parallel f)$ with the same method. Finally, the product approximation for HMMs can be calculated as:

$$D^{prod}(f \parallel g) = \sum_{n=1}^{\infty} \left[L_n^{prod}(f \parallel f) - L_n^{prod}(f \parallel g) \right] \quad (15)$$

And the computation is also truncated to a finite series in practice.

3.3. Approximation based on bounds for the KL divergence

In [9], an approximation of the KL divergence between GMMs are adopted based on the idea that strict bounds can provide an interval in which the real value of the KL divergence can be found. Motivated by this idea, we also design an approximation for the divergence between HMMs based on bounds.

The upper bound of KL divergence is computed by combining the upper bound of $L_n(f \parallel f)$ and the lower bound of $L_n(f \parallel g)$.

$$\begin{aligned} D(f \parallel g) &= \sum_{n=1}^{\infty} \left[L_n(f \parallel f) - L_n(f \parallel g) \right] \\ &\le \sum_{n=1}^{\infty} \left[L_n^{prod}(f \parallel f) - L_n^{var}(f \parallel g) \right] \stackrel{\text{def}}{=} D^{upper}(f \parallel g) \end{aligned} \quad (16)$$

The lower bound of KL divergence is obtained by combining the lower bound of $L_n(f \parallel f)$ and the upper bound of $L_n(f \parallel g)$.

$$\begin{aligned} D(f \parallel g) &= \sum_{n=1}^{\infty} \left[L_n(f \parallel f) - L_n(f \parallel g) \right] \\ &\ge \sum_{n=1}^{\infty} \left[L_n^{var}(f \parallel f) - L_n^{prod}(f \parallel g) \right] \stackrel{\text{def}}{=} D^{lower}(f \parallel g) \end{aligned} \quad (17)$$

It is reasonable to take the "center" of the interval as the approximation. Therefore, the the mean of the two bounds is computed as the approximation, and it is equal to the mean of $D^{prod}(f \parallel g)$ and $D^{var}(f \parallel g)$.

$$\begin{aligned} D^{mean}(f \parallel g) &= \frac{1}{2} \left[D^{upper}(f \parallel g) + D^{lower}(f \parallel g) \right] \\ &= \frac{1}{2} \left[D^{prod}(f \parallel g) + D^{var}(f \parallel g) \right] \end{aligned} \quad (18)$$

4. EXPERIMENTS

Two parts of experiments are carried out to evaluate the approximation quality of the KL divergence for comparing the similarity between transient HMMs. The first part is conducted to analyze the deviations of the proposed approximations and bounds to the numerical approximation using Monte Carlo simulation, and the second part shows the performance of the proposed approximations for estimating the confusability of phones in speech recognition.

4.1. Experimental data and setup

For experiments, an ASR system is set up for phone recognition task in continuous speech. The training and test data are both from a mandarin speech database provided by Chinese National Hi-Tech Project 863. This reading-style database contains the sentences spoken by 166 different native speakers (83 females, 83 males). The training data set for acoustic model contains 102-hour speech pronounced by 150 speakers (75 females, 75 males). The test set consists of 12-hour speech from other 16 speakers (8 females, 8 males).

The sample rate of the speech data is 16 kHz. In the front-end, the length and shift of analysis frame are 25ms and 10ms respectively. The feature used is 12th-ordered Mel-frequency cepstral coefficients (MFCCs) and the normalized short-time energy, appending their first- and second-order derivatives (39-dimensional feature). The phone set contains 97 phones [10]. The acoustic models are continuous density HMMs for context-independent monophones, and they are trained using the maximum likelihood estimation. Each HMM has 2 non-emission states and 3 emission states with a left-to-right topology, and the number of Gaussian mixture components is 8 for each emision state.

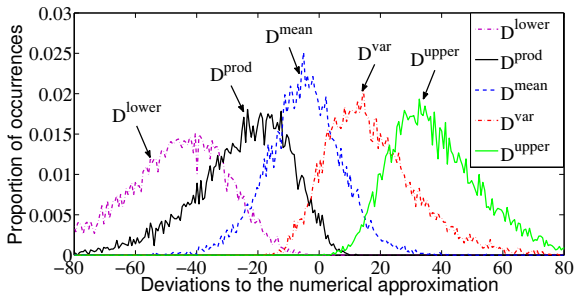

Fig. 1. Histograms of the deviations to the Monte Carlo simulation.

4.2. Deviation analysis

We analyze the deviations of the approximations and bounds to the numerically approximated divergence using Monte Carlo simulation. For the simulation, a method of sample generation is employed [2]. More than 4,000,000 sequence samples are generated in order to get an accurate approximation, and the number of samples generated for each HMM of phone is proportional to the expected frequency of that phone in the training data. With these samples, the numerical approximation of KL divergence as the reference is achieved between different HMMs from set of the acoustic models, and there are $97 \times (97-1) = 9312$ pairs of different HMMs. The deviations to the references are computed for the approximations and bounds respectively, and the histograms of the deviations are shown on Fig. 1. As expected, D^{lower} and D^{upper} are below and above the reference. D^{var} and D^{prod} are closer to the reference. D^{prod} is slightly under the reference, while D^{var} is slightly over-estimates it. Compared with other approximations and bounds, D^{mean} is shown to be closer to the reference, with deviations more concentrated near 0.

4.3. Estimation of confusability for phones

In this subsection, the approximation of KL divergence is used to estimate the confusability for phones, and this evaluation can measure the performance of each approximation employed to predict recognition errors [5, 7]. For this purpose, a phone confusion matrix is constructed for phone recognition error. The entry (i,j) of this matrix is the negative logarithm probability $-\log E(w_i, w_j)$ of substitution error for each phone pair of w_i and w_j, where $E(w_i, w_j) = [P(w_i|w_j) + P(w_j|w_i)]/2$, where $P(w_i|w_j)$ is the proportion of utterances for w_j that are recognized as w_i in the 1-Best result of the test data. The phone error rate of the ASR system is 31.8%, and more than 39,000 substitution errors are detected. The other matrix is generated by the proposed approximation of divergence from acoustic models, and each entry (i,j) of this matrix is the symmetric extension of the approximated divergence between the HMMs of phones w_i and w_j. The symmetric extension is computed with the resistor average symmetrized KL divergence [11]. The absolute row-correlation coefficient is computed between the rows with the same suffix in the two matrices, and this coefficient represents the correlation between the two approximations for a model with respect to all the other models. Finally, the average absolute correlation coefficient is obtained as a function of the phone frequency in the training set. The average absolute coefficient with a higher value indicates a closer correlation. For comparison, the same method is used to calculate the correlation between the confusion matrix and the matrices given by other approximation methods, such as the average divergence distance (ADD) [5] and the upper bound of Kullback-Leibler divergence (UBKLD) [6], in all cases considering the symmetric extensions.

Table 1. *The average absolute row-correlation coefficients.*

Methods of approximation	Average abs. row-corr.	Methods of approximation	Average abs. row-corr.
ADD	0.7372	UBKLD	0.7602
Variational	0.8125	Product	0.8018
Mean of bounds	0.8255		

In Table 1, the average absolute row-correlation coefficients are listed between the confusion matrix and the matrices given by the approximations of divergence. The average absolute row-correlation of the proposed product approximation is slightly lower than the variational approximation, but higher than the methods of ADD and UBKLD. The highest average absolute row-correlation is obtained by the proposed approximation method based on the two bounds. Hence, this method is a more accurate indicator of the acoustic confusability for a recognition task compared with all the other methods.

5. CONCLUSION

In this work, two novel methods of approximation have been proposed for the KL divergence between left-to-right transient HMMs. The first one is the product approximation which can be calculated recursively without introducing extra parameters, and the second one is explored based on the upper and lower bounds of the divergence. To calculate the bounds, the outcomes from both the variational and product approximation are employed. And the variational approximation is extended for the HMMs which use GMM to model the observation probability for each state.

The approximation based on bounds is attractive because it offers an effective approach to estimate the real value of the KL divergence with a clearer theoretical motivation. Compared with current methods, the approximation based on bounds achieves a better performance in both the correlation with respect to the numerical approximation and the prediction of phone error.

6. APPENDIX

If $f_{a_t}(x)$ and $g_{b_t}(x)$ are two distribution functions both given by GMM, $f_{a_t}(x) = \sum_i c_i^{a_t} N(x; \mu_i^{a_t}, \Sigma_i^{a_t})$ and $g_{b_t}(x) = \sum_j c_j^{b_t} N(x; \mu_j^{b_t}, \Sigma_j^{b_t})$, $\int f_{a_t}(x)\, g_{b_t}(x)\, dx$ is given as following:

$$\begin{aligned} I(a_t, b_t) &= \int f_{a_t}(x)\, g_{b_t}(x)\, dx \\ &= \int \sum_i c_i^{a_t} N(x; \mu_i^{a_t}, \Sigma_i^{a_t}) \sum_j c_j^{b_t} N(x; \mu_j^{b_t}, \Sigma_j^{b_t}) dx \\ &= \sum_{i,j} c_i^{a_t} c_j^{b_t} \int N(x; \mu_i^{a_t}, \Sigma_i^{a_t}) N(x; \mu_j^{b_t}, \Sigma_j^{b_t}) dx \end{aligned} \quad (19)$$

The product integration of two normal Gaussian PDFs $N(x; \mu_1, \Sigma_1)$ and $N(x; \mu_2, \Sigma_2)$ can be calculated by [12]:

$$\int N(x; \mu_1, \Sigma_1)\, N(x; \mu_2, \Sigma_2)\, dx =$$

$$|2\pi(\Sigma_1+\Sigma_2)|^{-\frac{1}{2}} \exp\left(-\frac{(\mu_1-\mu_2)^T(\Sigma_1+\Sigma_2)^{-1}(\mu_1-\mu_2)}{2}\right) \quad (20)$$

7. REFERENCES

[1] Solomon Kullback, "Information theory and statistics," Dover Publications Inc., 1968.

[2] Rita Singh, Bhiksha Raj, and Richard M. Stern, "Structured redefinition of sound units by merging and splitting for improved speech recognition," in *Proc. ICSLP*, 2000.

[3] Sayed Mohammad Ebrahim Sahraeian and Byung-Jun Yoon, "A novel low-complexity HMM similarity measure," *IEEE Signal Processing Letters*, vol. 18, no. 2, pp. 87–90, 2011.

[4] Jianping Zeng, Jiangjiao Duan, and Chengrong Wu, "A new distance measure for hidden Markov models," *Expert Systems with Applications*, vol. 37, no. 2, pp. 1550–1555, 2010.

[5] 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.

[6] Jorge Silva and Shrikanth Narayanan, "Upper bound Kullback-Leibler divergence for transient hidden Markov models," *IEEE transactions on Signal Processing*, vol. 56, no. 9, pp. 4176–4188, Sempetember 2008.

[7] John R. Hershey, Peder A. Olsen, and Steven J. Rennie, "Variational Kullback-Leibler divergence for hidden Markov models," in *Proc. ASRU*, 2007, pp. 323–328.

[8] John R. Hershey and Peder A. Olsen, "Approximating the Kullback Leibler divergence between Gaussian mixture models," in *Proc. ICASSP*, 2007, vol. 4, pp. 317–320.

[9] Jean-Louis Durrieu, Jean-Philippe Thiran, and Finnian Kelly, "Lower and upper bounds for approximation of the Kullback-Leibler divergence between Gaussian mixture models," in *Proc. ICASSP*, 2012, pp. 4833–4836.

[10] Chao Huang, Yu shi, Jianlai Zhou, Min Chu, Terry Wang, and Eric Chang, "Segmental tonal modeling for phone set design in mandarin LVCSR," in *Proc. ICASSP*, 2004, vol. 1, pp. 901–904.

[11] Don H. Johnson and Sinan Sinanovic, "Symmetrizing the Kullback-Leibler distance," Technical Report, Rice University, 2000.

[12] Peter Ahrendt, "The multivariate Gaussian probability distribution," Technical Report, Technical University of Denmark, 2005.