Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 81 (2016) 80 – 86

Procedia Computer Science

5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016, Yogyakarta, Indonesia

Variational Inference for Acoustic Unit Discovery

Lucas Ondel*, Lukaš Burget, Jan Černocký

Brno University of Technology, Czech Republic

Abstract

Recently, several nonparametric Bayesian models have been proposed to automatically discover acoustic units in unlabeled data. Most of them are trained using various versions of the Gibbs Sampling (GS) method. In this work, we consider Variational Bayes (VB) as alternative inference process. Even though VB yields an approximate solution of the posterior distribution it can be easily parallelized which makes it more suitable for large database. Results show that, notwithstanding VB inference is an order of magnitude faster, it outperforms GS in terms of accuracy.

Peer-review under responsibility of the Organizing Committee of SLTU 2016

Keywords:

Bayesian non-parametric, Variational Bayes, acoustic unit discovery

1. Introduction

Whereas Automatic Speech Recognition (ASR) systems are more and more frequently used in daily life applications, the need of labeled data has never been so high. With the ever-growing use of Internet a huge amount of unlabeled audio data coming from many different countries is now available. However, because the labeling process by human expert is expensive this data has still been unexploited. In [1], a nonparametric Bayesian model to automatically segment and label audio data has been proposed. The model has been later extended in [2] to jointly learn the phonetic units and the word pronunciations. An attempt to tackle the problem by mean of neural networks as also been investigated in [3]. In [1] and [2], both models are trained with the Gibbs Sampling (GS) algorithm [4] which can be summarized as follows: sample a new value for each parameter of the model from the probability of the parameter given the data and the others parameters and repeat until convergence. This method has many advantages. First, it allows the optimization of complex Bayesian model without the need of analytical solutions. Second, it can be shown that the dynamic converges toward the optimal distribution independently of the initial value of the parameters. However, despite these powerful features the algorithm carries some severe drawbacks: the parameters of the model cannot be

* Corresponding author.
E-mail address: iondel@fit.vutbr.cz

Peer-review under responsibility of the Organizing Committee of SLTU 2016
doi:10.1016/j.procs.2016.04.033

sampled asynchronously and the rate of convergence may be slow. Hence, even though GS is a popular way to train Bayesian models, its power is limited by its inability to handle large amount of data.

The Variational Bayesian (VB) inference ([5]) is an alternative technique to train Bayesian models which copes with the weaknesses of the GS algorithm. In essence, the application of VB is very similar to the well known Expectation-Maximization (EM) algorithm for latent models, allowing the training to be parallelized and the convergence monitored by a lower bound on the likelihood of the model. However, these benefits have a cost: the procedure may converge toward a local optimum. In this work we compare both training algorithm on a model similar to the one presented in [1]. Results show that for a considerable gain of speed the VB training achieves higher accuracy than the GS training.

2. Model

2.1. Model definition

Our model aims at segmenting and clustering unlabeled speech data into phone-like categories. It is similar to a phone-loop model in which each phone-like unit is modeled by an HMM [1]. This phone-loop model is fully Bayesian in the sense that:

- it incorporates a prior distribution over the parameters of the HMMs
- it has a prior distribution over the units modeled by a Dirichlet process[6].

Informally, the Dirichlet process prior can be seen as a standard Dirichlet distribution prior for a Bayesian mixture with an infinite number of components. However, we assume that our N data samples have been generated with only M components ($M \leq N$) from the infinite mixture. Hence, the model is no longer restricted to have a fixed number of components but instead can learn its complexity (i.e. number of components used M) according to the training data. The generation of a data set with M speech units can be summarized as follows:

1. sample the vector $\mathbf{v} = v_1, ..., v_M$ with
$$v_i \sim \text{Beta}(1, \gamma)$$
where γ is the concentration parameters of the Dirichlet process
2. sample M HMM parameters θ_1,...,θ_M from the base distribution of the Dirichlet process.
3. sample each segment as follows:

transition probabilities and emission probabilities

   (a) choose a HMM parameters with probability $\pi_i(\mathbf{v})$ defined as:
$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$$
   (b) sample a path $\mathbf{s} = s_1, ..., s_n$ from the HMM transition probability distribution
   (c) for each s_i in $\mathbf{s}$:

which mixture model?

      i. choose a Gaussian components from the mixture model
      ii. sample a data point from the Gaussian density function

A similar model have been applied in [1], however, two major differences should be noted: first, we have chosen to consider the stick-breaking construction[7] of the Dirichlet process (step 1 and 2 of the generation) rather than the Chinese Restaurant Process (CRP). See[8] and [1] for training Bayesian models with the CRP. This allow us to use variational methods to infer the distribution over the parameters rather than sampling methods. Secondly, our model

[1] For the sake of readability we write HMM for the complete HMM/GMM model.

does not have any boundary variable. The segmentation of the data is carried out by seeing this mixture of HMMs as a single HMM and using the standard forward-backward algorithm. This modification simplify the inference algorithm and does not require to have some pre-selection process of the boundary frames.

2.2. Model parameters

In absence of information about the prior distribution of the parameters of the model, we use conjugate priors. This choice greatly simplifies the inversion of the model: indeed, due to the conjugacy, the posterior distribution will have the same parametric form of the priors. We denote the hyper-parameters of the priors with the subscript 0 and the hyper-parameters of the posteriors with the subscript n. The distribution of the mean $\boldsymbol{\mu}$ and the diagonal covariance matrix $\boldsymbol{\Sigma}$ with diagonal $\boldsymbol{\lambda}$ is modeled by a Normal-Gamma density: $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, (\kappa_0\boldsymbol{\lambda})^{-1})$ Gamma$(\boldsymbol{\lambda}|\alpha_0, \boldsymbol{\beta_0})$ where $\boldsymbol{\beta}_0$ is the rate parameter of the Gamma distribution. The prior of the weights $\boldsymbol{\pi}$ of a GMM and the row r of the transition matrix of an HMM are modeled by Dirichlet distributions parametrized by the vectors $\boldsymbol{\eta}_0^{(gmm)}$ and $\boldsymbol{\eta}_0^{(hmm,r)}$ respectively. Finally, the prior distribution over the proportions v_i is the Beta$(1, \gamma)$ distribution and its posterior is given by Beta(γ_i^1, γ_i^2)) The model has also 3 set of hidden variables:

- c_i the index of the HMM for the ith segment in the data set
- s_{ij} the HMM state of the jth frame in the ith segment
- m_{ij} the GMM component of the jth frame in the ith segment.

2.3. Inference

We would like to invert the model previously defined to obtain the probability of the parameters given the data. Following variational Bayes (VB) framework, it can be achieved by optimizing a lower-bound on the log-evidence of the data with respect to the distribution over the parameters q:

$$\log p(X) \geq E_q[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0))] - E_q[\log q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta})] \tag{1}$$

where $\mathbf{X}$ is the entire set of features of the N segments, $\mathbf{c} = c_1, ..., c_N$, $\mathbf{S} = s_{11}, ..., s_{NL_N}$, $\mathbf{M} = m_{11}, ..., m_{NL_N}$, $\boldsymbol{\Theta}$ is the set of all the parameters and $\boldsymbol{\Phi}_0$ is the set of the hyper-parameters of the prior distribution over the parameters. The equality is achieved if and only if $q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}) = p(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta})$. Because of the conjugate priors, we have a closed form solution of the posterior distribution when considering the following approximation (This is a typical case of the so-called *mean-field* approximation.):

$$q(\mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}) = q(\mathbf{c}, \mathbf{S}, \mathbf{M})q(\boldsymbol{\Theta}). \tag{2}$$

where we have assumed the statistical independence between the parameters and the hidden variables of the model. Following[7], another approximation is done to cope with the infinite number of component in the mixture; we set $v_T = 1$ to force the weight of any component greater than T to zero. By using the factorization in (2) and variational calculus, one can show that the (log) distributions that maximizes the bound (1) are :

$$\begin{aligned} \log q^*(\mathbf{c}, \mathbf{S}, \mathbf{M}) &= E_{q(\boldsymbol{\Theta})}[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0)] + \text{const} \\ \log q^*(\boldsymbol{\Theta}) &= E_{q(\mathbf{c},\mathbf{S},\mathbf{M})}[\log p(\mathbf{X}, \mathbf{c}, \mathbf{S}, \mathbf{M}, \boldsymbol{\Theta}|\boldsymbol{\Phi}_0)] + \text{const} \end{aligned} \tag{3}$$

A locally optimal posterior distribution is found by evaluating each factor in turn using (3) until convergence. The estimation of the factors is detailed below.

2.4. Estimation of the distribution of the latent variables

From the definition of the model, the approximate posterior distribution of the latent variables can be factorized as:

$$q(\mathbf{c}, \mathbf{S}, \mathbf{M}) = \prod_i^N q(c_i)q(\mathbf{s}_i|c_i) \prod_j^{L_i} q(m_{ij}|c_i, s_{ij}) \tag{4}$$

where L_i is the length of the ith segment. The evaluation of the joint distribution of the latent variables relies on the three factors: $q(c_i)$, $q(\mathbf{s}_i|c_i)$ and $q(m_{ij}|c_i, s_{ij})$. Following[9], the expected value of the log-likelihood for each frame of the ith segment given the parameters of the GMM associated to the state s_{ij} of the HMM corresponding to the cluster c_i is:

$$\log \boldsymbol{\Omega}_i^{(c_i,s_{ij})} = \psi(\boldsymbol{\eta}_n^{(GMM)}) - \psi(\sum_a \eta_{na}^{(GMM)}) + \frac{1}{2}(\psi(\alpha_n) - \log \boldsymbol{\beta}_n) - \frac{1}{2}(\kappa_n + \frac{\alpha_n}{\boldsymbol{\beta}_n}(\mathbf{x}_i - \boldsymbol{\mu}_n)^2) \tag{5}$$

where $\boldsymbol{\Omega}_i^{(c_i,s_{ij})}$ is a matrix of $L_i \times K$ where K is the number of component in the GMM and ψ denotes the gamma function. From Equation 3, the optimal log-distribution of the GMM components at a given frame j is simply computed by:

$$\log q^*(m_{ij}|c_i, s_{ij}) = \log \omega_{i,m_{ij}} + \text{const}\,. \tag{6}$$

where $\omega_{i,m}$ is the mth row of the matrix $\boldsymbol{\Omega}_i^{(c_i,s_{ij})}$. One can now calculate the second (log-)factor in Equation 4:

$$\log q^*(\mathbf{s}_i|c_i) = \sum_{j=2}^{L_i} E[\log A]_{s_{ij-1},s_{ij}} + \sum_{j=1} \log q^*(m_{ij}|c_i, s_{ij}) + \text{const} \tag{7}$$

where the row r of the matrix $E[\log A]$ is given by:

$$E[\log A]_r = \psi(\boldsymbol{\eta}_n^{(hmm,r)}) - \psi(\sum_k \eta_{nk}^{(hmm,r)}). \tag{8}$$

In Equation 7, the first sum does not start from 1 as we assume that the HMM has a left-to-right topology. Finally, from[7], the log-distribution of the assignment variable of the infinite mixture is given by:

$$\log q^*(c_i = t) = \psi(\boldsymbol{\gamma}_t^1) - \psi(\boldsymbol{\gamma}_t^1 + \boldsymbol{\gamma}_t^2) + \sum_{a=1}^{t-1} \psi(\boldsymbol{\gamma}_{na}^2) - \psi(\boldsymbol{\gamma}_a^2 + \boldsymbol{\gamma}_a^2) + \text{const} \tag{9}$$

2.5. Re-estimation of the posterior's parameters

The second step of the iterative training algorithm is to update the hyper-parameters of the model. For the component a of the GMM associated to the state b and previous state b' of the HMM corresponding to the cluster c we define the quantities:

$$\begin{aligned}
\sigma^{(0)} &= \sum_i^N \sum_j^{L_i} q(m_{ij} = a|s_{ij} = b, c_i = c) \\
\boldsymbol{\sigma}^{(1)} &= \sum_i^N \sum_j^{L_i} q(m_{ij} = a|s_{ij} = b, c_i = c)\mathbf{x}_{ij} \\
\boldsymbol{\sigma}^{(2)} &= \sum_i^N \sum_j^{L_i} q(m_{ij} = a|s_{ij} = b, c_i = c)\mathbf{x}_{ij}^2 \\
\sigma^{(3)} &= \sum_i^N \sum_j^{L_i} q(s_{ij-1} = b', s_{ij} = b|c_i = c) \\
\sigma^{(4)} &= \sum_i^N q(c_i = c) \\
\sigma^{(5)} &= \sum_i^N q(c_i < c) = \sum_i \sum_t^{c-1} q(c_i = t)
\end{aligned} \tag{10}$$

where $\mathbf{x}^2$ is the element wise square operation. From these quantities, the hyper-parameters of the posterior distribution become:

$$\begin{aligned}
\kappa_n &= \kappa_0 + \frac{1}{2}\sigma^{(0)} \\
\boldsymbol{\mu}_n &= \kappa_0\boldsymbol{\mu}_0 + \frac{\boldsymbol{\sigma}^{(1)}}{\kappa_n} \\
\alpha_n &= \alpha_0 + \frac{\sigma^{(0)}}{2} \\
\boldsymbol{\beta}_n &= \boldsymbol{\beta}_0 - \frac{1}{2}\frac{(\kappa_0\boldsymbol{\mu}_0 + \boldsymbol{\sigma}^{(1)})^2}{\kappa_n} \\
&\quad + \frac{1}{2}(\boldsymbol{\sigma}^{(2)} + \kappa_0\boldsymbol{\mu}_0^2)
\end{aligned} \tag{11}$$

for the GMM weights:

$$\eta_n^{gmm} = \eta_0^{gmm} + \sigma^{(0)} \tag{12}$$

for the row r of the HMM transition matrix:

$$\eta_n^{hmm,r} = \eta_0^{hmm,r} + \sigma^{(3)} \tag{13}$$

and finally, for the posterior distribution of the infinite mixture proportions:

$$\begin{aligned}
\gamma_c^1 &= 1 + \sigma^{(4)} \\
\gamma_c^2 &= \gamma + \sigma^{(5)}.
\end{aligned} \tag{14}$$

The quantities defined in Equation 10 relies on the expectation with respect to the distribution of the latent variables q^*. For the case of the variables $\sigma^{(0)}$ and $\sigma^{(4)}$ these expectations are straightforward as it is simply the distribution $q^*(m_{ij}|c_i, s_{ij})$ and $q^*(c_i)$ respectively. The quantity $\sigma^{(3)}$ can be computed by using the well known *forward-backward* algorithm using (8) as the transition matrix and (5) for the (log) emission probabilities.

3. Results

5.4 hours (?)

The experiments were conducted on the TIMIT database[10]. The acoustic features were the mean normalized MFCCs + Δ + ΔΔ generated by HTK[11] and each HMM had 3 states with a left-to-right topology and 4 Gaussian per state. The hyper-parameters for the priors were initialized to the same value as in[1] and the truncation T was set to 200. The metric used for comparison was the mutual information[12] between the discovered units and the true phone labels normalized by the entropy of the phones to get the ratio R of learnt information. For the case of unknown boundaries, the discovered units were mapped to the closest phone in time.

why 4

3.1. Fixed boundaries

To compare the VB and GS sampling we considered first a simpler problem where the phone boundaries were known beforehand. This reduced the task in mere clustering of variable length segments. The GS sampler was the same as in[1] with known and fixed boundaries. Figure 1 shows the evolution of the learnt information ratio during the training over 10 iterations [2]. The training of the GS took about 11 hours on a single core (Intel Xeon CPU E5-2670 2.60GHz) as it does not allow easily parallelization. The VB training was finished in less than half an hour using about 300 cores.

[2] By iteration we mean one sweep over the whole database.

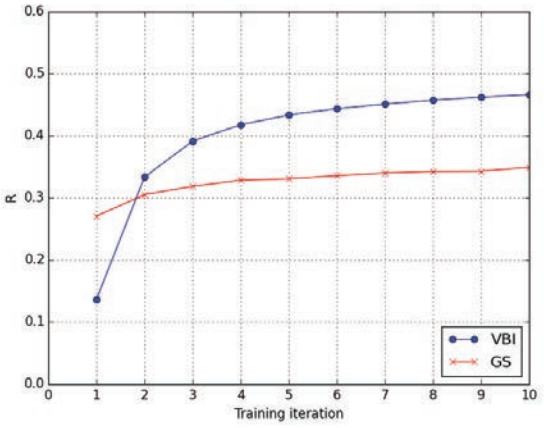

Fig. 1. Evolution of the learnt information R during the training.

As expected the VB training is much faster than the GS one due to the parallelization. Interestingly, the VB algorithm finds a better solution than the GS algorithm whereas both have the same parametrization. Retraining the VB solution with the GS algorithm degrade the performance. Indeed, because the CRP does not allow easily to average the model at different time step of the training, the GS solution is a point estimate of the posterior distribution. On the other hand, the solution from the VB, thanks to the approximation, is an average over the model parameters space. The difference in performance between model averaging and point-estimate shows that the model is still uncertain about its parameters.

3.2. Unknown boundaries

When released from the fixed boundaries constraints, the task becomes harder as the inference process has to jointly cluster and segment the speech data. Results of the training is compared to the fixed boundaries case in Table 3.2. The performance severely drops when the boundaries are unknown. However, the VB training still performs better

	boundaries	R	# units
VB	known	47 %	85
VB	unknown	36 %	197
GS	known	35 %	112

Performance of VBI vs GS

than the GS with fixed boundaries which is a promising results. Note that the number of units drastically increases when the boundaries are unknown. This shows the difficulty of the model to match corresponding sounds with high variability. This problem remains one of the biggest obstacle toward completely unsupervised labeling of speech data.

4. Conclusion

We proposed to train a nonparametric Bayesian model for automatic units discovery within the Variational Bayes framework. Besides simplifying the training scheme, this approach proves to be fast and yields better solution which makes it more suitable for big databases. However, despite the improvement observed, the model still have difficulties with the diversity of speech and tends to learn a large part of unwanted variability. The HMM model for speech segment is convenient but unrealistic and most likely, stronger model will be needed if one wants to achieve accurate

automatic units discovery. We plan to extent the present work by using the VB inference with more complex models, as in [13], and to gain leverage of Bayesian language models [14] to further improve the accuracy of the discovered units.

5. Acknowledgment

This work was supported by the European Union's Horizon 2020 project No. 645523 BISON, and by Technology Agency of the Czech Republic project No. TA04011311 "MINT"

References

1. Lee, C., Glass, J.. A nonparametric bayesian approach to acoustic model discovery. In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*; ACL '12. Stroudsburg, PA, USA: Association for Computational Linguistics; 2012, p. 40–49. URL: http://dl.acm.org/citation.cfm?id=2390524.2390531.
2. Lee, C., Zhang, Y., Glass, J.. Joint learning of phonetic units and word pronunciations for ASR. In: *Proceedings of the 2013 Conference on Empirical Methods on Natural Language Processing (EMNLP)*. 2013, p. 182–192.
3. Renshaw, D., Kamper, H., Jansen, A., Goldwater, S.. A comparison of neural network methods for unsupervised representation learning on the zero resource speech challenge. In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*. 2015, p. 3199–3203. URL: http://www.isca-speech.org/archive/interspeech_2015/i15_3199.html.
4. Bishop, C.M.. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc.; 2006. ISBN 0387310738.
5. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.. An introduction to variational methods for graphical models. *Mach Learn* 1999; **37**(2):183–233. URL: http://dx.doi.org/10.1023/A:1007665907178. doi:10.1023/A:1007665907178.
6. Antoniak, C.E.. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics* 1974;**2**(6).
7. Blei, D.M., Jordan, M.I.. Variational inference for dirichlet process mixtures. *Bayesian Analysis* 2005;**1**:121–144.
8. Rasmussen, C.E.. The infinite gaussian mixture model. In: Solla, S.A., Leen, T.K., Müller, K.R., editors. *NIPS*. The MIT Press. ISBN 0-262-19450-3; 1999, p. 554–560.
9. Murphy, K.P.. Conjugate bayesian analysis of the gaussian distribution. Tech. Rep.; 2007.
10. Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G., Pallett, D.S., Dahlgren, N.L.. DARPA TIMIT acoustic phonetic continuous speech corpus CDROM. 1993. URL: http://www.ldc.upenn.edu/Catalog/LDC93S1.html.
11. Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Moore, G., et al. *The HTK Book, version 3.4*. Cambridge, UK: Cambridge University Engineering Department; 2006.
12. Cover, T.M., Thomas, J.A.. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience; 2006. ISBN 0471241954.
13. Friston, K., Trujillo-Bareto, N., Daunizeau, J.. DEM: A variational treatment of dynamic systems. *NeuroImage* 2008;**41**(3):849–885. doi:10.1016/j.neuroimage.2008.02.054.
14. Teh, Y.W.. A hierarchical bayesian language model based on pitman–yor processes. In: *In Coling/ACL, 2006. 9*. 2006.