BAYESIAN PHONOTACTIC LANGUAGE MODEL FOR ACOUSTIC UNIT DISCOVERY

Lucas Ondel, Lukaš Burget, Jan Černocký

Brno University of Technology
Brno, Czech Republic

Santosh Kesiraju

International Institute of Information Technology
Hyderabad, India

ABSTRACT

Recent work on Acoustic Unit Discovery (AUD) has led to the development of a non-parametric Bayesian phone-loop model where the prior over the probability of the phone-like units is assumed to be sampled from a Dirichlet Process (DP). In this work, we propose to improve this model by incorporating a Hierarchical Pitman-Yor based bigram Language Model on top of the units' transitions. This new model makes use of the phonotactic context information but assumes a fixed number of units. To remedy this limitation we first train a DP phone-loop model to infer the number of units, then, the bigram phone-loop is initialized from the DP phone-loop and trained until convergence of its parameters. Results show an absolute improvement of 1-2 % on the Normalized Mutual Information (NMI) metric. Furthermore, we show that, combined with Multilingual Bottleneck (MBN) features the model yields a same or higher NMI as an English phone recogniser trained on TIMIT.

Index Terms— 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. Semi-supervised training is de facto the current standard technique to cope with unlabeled data. This method is however unsatisfactory as it requires a large amount of untranscribed data for a significant improvement [1] and is not applicable in cases where no transcribed data is available. Alternatively to semi-supervised training, a nonparametric Bayesian model to automatically segment and label audio data has been proposed In [2]. The model was later refined in [3] in order to be trained using the Variational Bayes (VB) method. An attempt to tackle the problem by means of neural networks as also been investigated in [4]. In [3], the Acoustic Unit Discovery (AUD) is done by clustering temporal sequences with a Dirichlet Process (DP) based mixture model where, following the Variational treatment of the DP mixture model [5], the probability of the component of the mixture is approximated by a finite Categorical distribution. This distribution functions as a unigram phonotactic Language Model (LM) over the units. This generative process is quite inaccurate as the probability of a phone (and by extension any phone-like unit) strongly depends on the previous phones. In the present work, we extend the AUD model described in [3] by replacing the naive unigram phonotactic LM by a non-parametric Bayesian bigram phonotactic LM. The article is organized as follows: Section 2 and 3 describes the original model and its extension respectively, Section 4 details the training of the extended model, Section 5 details how we evaluate the AUD task and finally, results are presented in Section 6.

2. INFINITE PHONE-LOOP MODEL

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

$$v_i \sim \text{Beta}(1, \gamma)$$

[1] By abuse of notation we write HMM for the complete HMM/GMM model.

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

$$\theta_i \sim H$$

3. For a sequence of, say, L phone-like units, the sequence of features associated to the unit l is sampled with the following scheme:

   (a) sample the cluster index c_l from the distribution $\boldsymbol{\pi}(\mathbf{v})$ defined as:

$$\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1}(1 - v_j)$$

   (b) From the HMM's parameters θ_{c_l}:

      i. sample a sequence of state $\mathbf{s} = s_1, ..., s_n$

      ii. for each state s_t, sample a Gaussian component m_t and generate data point $\mathbf{x}_t$ from it.

The graphical representation of this model is shown in Figure 1a. The priors over the GMM weights, Gaussian mean and (diagonal) covariance matrix are a Dirichlet and a Normal-Gamma density respectively. A similar model has been applied in [2], however, two major differences should be noted: first, we have chosen to consider the stick-breaking construction [5] of the Dirichlet process (step 1 and 2 of the generation) rather than the Chinese Restaurant Process (CRP). See [7] and [2] for training Bayesian models with the CRP. This allows us to use variational methods to infer the distribution over the parameters rather than sampling methods. Secondly, our 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 Viterbi algorithm. See [3] for the Variational Bayesian treatment of this model.

3. BIGRAM PHONE-LOOP MODEL

The model previously described is able to learn the appropriate number of units for a given data set thanks to the Dirichlet Process prior. The learnt probabilities of each unit to occur can be seen as a simple unigram phonotactic language model. It is well known however, that each language has a specific phone distribution and moreover a specific n-gram phone sequence distribution. Hence, the simple phone-loop model is limited in the sense that it does not make use of the phonotatic context information. To remedy this problem, we can replace the Dirichlet Process prior by a Hierarchical Pitman-Yor process based Language Model (HPYLM) [8] [9]. The HPYLM prior guarantees that the probability of each unit to occur depends on the previous O units, where O is the order of the hierarchy of the HPY. The data generation with a bigram based HPYLM is summarized as follows:

1. sample the HMM parameter sets $\theta_1, ..., \theta_K$ from the prior distribution:

$$\theta_i \sim \phi$$

2. sample a Categorical distribution from the top level Pitman-Yor process (PY)

$$G_1 \sim PY(G_0, \gamma_0, d_0)$$

   where G_0, γ_0 and d_0 are the base distribution, the concentration and the discount parameters of the PY respectively. In our case, we assumed G_0 to be a uniform Categorical distribution

3. sample K context-dependent distributions over the units $G_{2,1}, ..., G_{2,K}$:

$$G_{2,i} \sim PY(G_1, \gamma_1, d_1)$$

   where G_1, γ_1 and d_1 are the base distribution, the concentration and the discount parameters of the second-level PY respectively

4. A sequence of L units $c_1, \ldots, c_L$ and the associated sequence of features is sampled as follow :

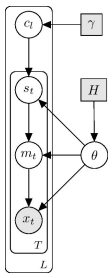
   (a) sample the lth unit index for the c_l:

$$c_l \sim G_{2,c_{l-1}}$$

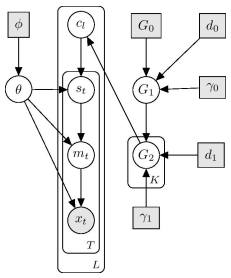
   (b) finally, sample the state path $s_1, \ldots, s_n$, the state's mixture components $m_1, \ldots, m_n$ and the features vectors $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from the HMM with parameters θ_{c_l} as described in section 2

The graphical model corresponding to this generation process is depicted in Figure 1b. We draw the reader's attention to the fact that, contrary to the model presented in Section 2, we assume here a finite number of units. Hence, while the HPY based phone-loop can model context-dependent unit transitions, it is not suitable to infer the number of units. Eventually, this limitation could be resolved by assuming the HMM parameters $\boldsymbol{\theta}$ to be sampled from the top level base distribution G_0 of the HPY. However, because there is no known analytic form for the stick-breaking representation of the HPY [10], and therefore no simple VB inference algorithm adapted to this model, it would require to train the HMM parameters using Gibbs sampling losing the benefits of the VB inference, as discussed in [3].

Fig. 1: Two different AUD models



(a) Phone-Loop model with a Dirichlet Process prior



(b) Phone-Loop model with a bigram phonotactic HPYLM

4. TRAINING

In section 2 and 3 we presented two phone-loop models, the first one learning the complexity (i.e. the number of units) needed to model the data whereas the latter one makes use of the phonotactic context information. Figure 2 shows the evolution of the number of units during the VB training of the DP based phone-loop model. As we can see, the number of units stabilizes very quickly at the beginning of the training. This suggests that we can proceed in two stages: first learning the number of units with the DP based phone-loop model and then refining the HMMs' parameters using the bigram phone-loop model. The DP phone-loop model is trained using VB inference as described in [3]. Once the training of the DP phone-loop model has converged we switch to a 3-steps training procedure that we repeat until convergence:

1. label the data with Viterbi algorithm using the current phone-loop model
2. train the HPY based language model on the labeled data using the Chinese Restaurant Franchise (CRF) [8]
3. set the unit-to-unit transitions according to the trained phonotactic LM and retrain the HMMs' parameters while keeping fixed the aforementioned transitions.

While this algorithm was experimentally proven to be efficient (see Section 6) it is worth mentioning a couple of possible variations. First of all, training the HPYLM on the Viterbi path can be seen as an approximation of the VB training. This approximation could be refined by sampling paths instead of using the most likely one. Sampling several paths for an utterance would account for the uncertainty of the sequence unit. It was found experimentally that doing so considerably slows down the training and yields the same results as the method proposed above. Another important point is that we retrained from scratch the full HPYLM each time we update the HMMs' parameters. Indeed, the CRF assumes a fixed training data whereas in our case the sequences of units possibly change each time we update the acoustic model. This limitation could be tackled by removing all the customers of one utterance and then re-sampling a new sitting arrangement for this utterance. This approximation of the CRF is slightly inaccurate for very small data set but works well for any reasonable size data set. The possible speed up of this approximation is however counterbalanced by some memory overhead as we have to store the utterance corresponding to each customer in the CRF. No performance difference between the two approaches was found experimentally.

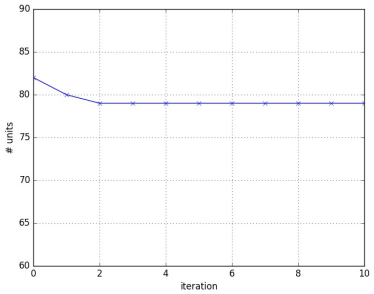


Fig. 2: Evolution of the number of units during the training of the DP model. The number of units was found by labeling (Viterbi decoding) the data and counting how many different units occurred.

5. EVALUATION

The evaluation of the discovered acoustic unit is not as straightforward as it may seem since the usefulness of the discovered units is highly task dependent. In this work, we use the mutual information between the human expert labeling and the discovered units. The mutual information between two random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$

where $H(X)$ is the entropy of X and $H(X|Y)$ is the entropy of X given Y. Note that it is a symmetric measure. Informally, this metric can be understood as a "correlation" measure between the the discovered untis and the true phones.

The mutual information gives a result in bits, however, since the maximum amount of bits to learn depends on the data and the task, we divide by the entropy of the true labels:

$$NMI = \frac{I(X;Y)}{H(X)} \quad (2)$$

where NMI stands for Normalized Mutual Information. This quantity is also known as the *uncertainty coefficient*. Note that the NMI version is not symmetric anymore and range from 0 to 1. Practically, we generate a sequence of units for each utterance of some test data using the Viterbi algorithm and then, we map each unit to its closest label in time. Using this one-to-one mapping the computation of the NMI is straightforward.

6. RESULTS

The experiments were conducted on the TIMIT database [11]. We used two different set of features: the mean normalized MFCC + Δ + $\Delta\Delta$ generated by HTK [12] and the Multilingual BottleNeck (MBN) features [13] trained on the Czech, German, Portuguese, Russian, Spanish, Turkish and Vietnamese data of the Global Phone database. As shown in

model	features	NMI
DP phone loop	MFCC	33.94
Bigram phone loop	MFCC	**34.82**
DP phone loop	GP BN	42.06
Bigram phone loop	GP BN	**42.63**

Table 1: Normalized Mutual Information of the DP phone-loop and the bigram phone-loop for MFCC and MBN features

Table 1, the bigram phone-loop model improves the NMI for both sets of features. The improvement is relatively smaller with the MBN features. This is to be expected as the MBN features are trained and computed using some temporal context which reduces the influence of the bigram LM. Note that the results of the DP phone-loop model are slightly worse than the ones reported in [3] as we have used a separate test set rather than evaluating the NMI on the training data.

In standard ASR systems, it is a common practice to scale down the acoustic scores to alleviate the influence of the wrong assumptions of the HMM. Scaling down the acoustic score (in our case, this corresponds to multiply Equation 5 in [3] by some scaling factor) reduce the dynamic range of the log-likelihood of the emissions' density and thus strengthen the influence of the state transitions and the language model. We found out experimentally that scaling the acoustic scores during the bigram phone-loop model training can significantly improve the final NMI. Figure 3 shows the absolute NMI improvement over the simple DP phone-loop model for various acoustic scale. The optimal scaling differs for the MFCC and the MBN features as the dynamic range of both feature sets are rather different. Final results including

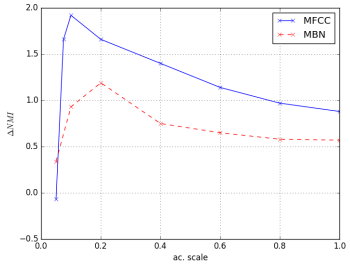


Fig. 3: Absolute improvement of the NMI when scaling down the acoustic scores.

the optimal acoustic scale for MFCC and MBN features are shown in Table 2. For comparison, we computed the NMI

model	features	ac. scale	NMI
DP phone loop	MFCC	-	33.94
Bigram phone loop	MFCC	1.0	34.82
Bigram phone loop	MFCC	0.1	**35.86**
DP phone loop	GP BN	-	42.06
Bigram phone loop	GP BN	1.0	42.63
Bigram phone loop	GP BN	0.2	**43.25**
English phone rec.	-	-	42.21

Table 2: NMI of the DP phone-loop and the bigram phone-loop for MFCC and MBN features with optimal scaling

from the output of a phone recogniser trained with Kaldi [14] using the standard TIMIT recipe. Interestingly, the NMI of this baseline is similar to the MBN DP phone-loop and the bigram MBN phone-loop is about one percent better (see Table 2). Even though care has to be taken as the NMI is not a perfect metric it is a promising results which let us hope that the research field of AUD will soon be mature enough to be applied to low-resource languages that are so far out of reach of speech technologies.

Acknowledgement

The work reported here was carried out during the 2016 Jelinek Memorial Summer Workshop on Speech and Language Technologies, which was supported by Johns Hopkins University via DARPA LORELEI Contract No HR0011-15-2-0027, and gifts from Microsoft, Amazon, Google, Facebook. It was also supported by European Union's Horizon 2020 project No. 645523 BISON and Czech Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPU II) project "IT4Innovations excellence in science - LQ1602"

7. REFERENCES

[1] Scott Novotney and Richard M. Schwartz, "Analysis of low-resource acoustic model self-training," in *INTERSPEECH 2009, 10th Annual Conference of the International Speech Communication Association, Brighton, United Kingdom, September 6-10, 2009*, 2009, pp. 244–247.

[2] Chia-ying Lee and James Glass, "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*, Stroudsburg, PA, USA, 2012, ACL '12, pp. 40–49, Association for Computational Linguistics.

[3] Lucas Ondel, Lukáš Burget, and Jan Černocký, "Variational inference for acoustic unit discovery," in *Procedia Computer Science*. 2016, vol. 2016, pp. 80–86, Elsevier Science.

[4] Daniel Renshaw, Herman Kamper, Aren Jansen, and Sharon Goldwater, "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, pp. 3199–3203.

[5] David M. Blei and Michael I. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Analysis*, vol. 1, pp. 121–144, 2005.

[6] Charles E. Antoniak, "Mixtures of dirichlet processes with applications to bayesian nonparametric problems," *Annals of Statistics*, vol. 2, no. 6, November 1974.

[7] Carl Edward Rasmussen, "The infinite gaussian mixture model," in *NIPS*, Sara A. Solla, Todd K. Leen, and Klaus-Robert Müller, Eds. 1999, pp. 554–560, The MIT Press.

[8] Yee Whye Teh, "A hierarchical bayesian language model based on pitman–yor processes," in *In Coling/ACL, 2006. 9*, 2006.

[9] Mark Johnson, Thomas L. Griffiths, and Sharon Goldwater, "Adaptor grammars: A framework for specifying compositional nonparametric bayesian models," in *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, 2006, pp. 641–648.

[10] Y. W. Teh and M. I. Jordan, "Hierarchical Bayesian nonparametric models with applications," in *Bayesian Nonparametrics: Principles and Practice*, N. Hjort, C. Holmes, P. Müller, and S. Walker, Eds. Cambridge University Press, 2010.

[11] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic phonetic continuous speech corpus CDROM," 1993.

[12] S. J. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. C. Woodland, *The HTK Book, version 3.4*, Cambridge University Engineering Department, Cambridge, UK, 2006.

[13] František Grézl and Martin Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU-2014. St. Petersburg, Russia, 2014*. 2014, pp. 39–45, International Speech Communication Association.

[14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Dec. 2011, IEEE Signal Processing Society, IEEE Catalog No.: CFP11SRW-USB.