Semi-implicit Variational Inference for Unsupervised Acoustic Unit Discovery

Desh Raj
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218
draj2@jhu.edu

Abstract

We propose a structured variational autoencoder (VAE), which combines the modeling capability of probabilistic graphical models with the training convenience of neural networks, to discover acoustic units in unlabeled speech data. Hidden Markov Models (HMMs) are used to model the temporal states and VAEs are used to model the frame-level emissions. This model is shown to outperform conventional GMM-HMM based methods on two low-resource datasets—Turkish and Dutch, obtained from the Mozilla Common Voice corpus. We make an empirically sound argument for the use of informative priors, learned from high-resource languages such as English, for the HMMs in such models. Finally, we propose a semi-implicit extension to the VAEs to expand the commonly used analytic variational distribution family, by mixing the variational parameter with a flexible distribution.

1 Introduction

In the last few years, automatic speech recognition (ASR) systems have achieved human-level recognition due to the availability of huge training corpora and deep neural networks [Hinton et al., 2012, Veselỳ et al., 2013]. However, for remote languages which do not have such annotated data available, it remains difficult to efficiently train discriminative neural systems. Bayesian methods, on the other hand, provide a convenient alternative since knowledge gleaned from high-resource languages can be transferred to low-resource settings through the use of informative priors [Watanabe et al., 2003].

Acoustic unit discovery (AUD) refers to the task of automatically discovering phoneme-like units in an unlabeled speech dataset [Liu et al., 2017]. It is an important building block for developing ASR systems in the absence of transcribed data, since an accurate AUD system can be used to obtain training data for acoustic modeling, even if we only have access to speech signals.

In such a setting, acoustic units are often modeled as latent variables and learned through EM-like methods. However, full estimation in such approaches is usually intractable, and approximate inference techniques are employed using either sampling or variational Bayes [Ondel et al., 2016]. While the former is usually more accurate, the latter is more suitable for parallel computations. Regardless of the inference policy used, the common thread across all such methods is modeling AUD as a generative problem. Hidden Markov Models (HMMs) [Rabiner and Juang, 1986] have traditionally been used to model speech since they provide the convenience of jointly modeling temporal state shift and frame-level emissions. Gaussian mixture models (GMMs) are typically used to represent the emission probabilities.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

In this work, we use variational autoencoders (VAEs) Kingma and Welling [2013] instead of GMMs to model the emission probabilities, with the stated objective of making learning of complex distributions easier and performing efficient inference. We use this model to perform AUD on two low-resource languages, namely Turkish (6 hours) and Dutch (13 hours), obtained from the recently released Mozilla Common Voice corpora [1]. Furthermore, we emphasize the merits of a Bayesian method for such unsupervised tasks by including an informative prior on the HMM, learned from a high-resource language (the TIMIT English dataset [Garofolo et al., 1993], in this case).

Despite their popularity, VAEs are restricted in the sense that the encoder distribution is required to be reparameterizable and analytically evaluable, which usually restricts it to a small exponential family. To remedy this, we extend our HMM-VAEs to include semi-implicit variational inference, which mixes the variational parameter with a flexible distribution which can assume any density function.

The remainder of the paper is organized as follows. In Section 2, we present our model in some detail. We begin with a background on the AUD generative process using HMMs, and then describe the equations for learning and inference in HMM-VAEs, finally making a case for an informative prior. We then extend this model using semi-implicit variational inference in Section 3. We describe the experiments and datasets in Section 4, while also summarizing some important implementation details. In Section 5, we analyze the results of our models with respect to the accuracy vs. efficiency tradeoff. We present an overview of related work in Section 6, and finally conclude in Section 7.

2 The HMM-VAE model

Following the generative process defined in Ondel et al. [2016], we model AUD using Dirichlet process based HMMs for each acoustic unit. Theoretically, the Dirichlet process prior defined over the acoustic units can be seen as a standard Dirichlet distribution prior for a Bayesian mixture with an infinite number of components. However, we assume that for a dataset containing N examples, we need at most $M \leq N$ units (and therefore, M HMMs) to generate the data. We can the summarize the generative process as follows.

1. Sample $\mathbf{v} = v_1, \ldots, v_M$, with $v_i = \text{Beta}(1, \gamma)$, where γ is the concentration parameter for the Dirichlet process.
2. Sample M HMM parameters $\theta_1, \ldots, \theta_M$ from the base distribution of the Dirichlet process.
3. To sample each segment, choose HMM parameters with probability $\pi_i(\mathbf{v})$, defined by $\pi_i(\mathbf{v}) = v_i \prod_{j=1}^{i-1} (1 - v_j)$, and sample a path $\mathbf{s} = s_1, \ldots, s_n$ from the HMM transition probability distribution.
4. To generate the actual emissions for a GMM-based model, choose a Gaussian component from the mixture and sample a point from the density function.

In this work, we modify the final step of the above generative process so that the emission probability is modeled using a VAE instead of a GMM. The resulting model can be learned using an EM-like procedure described in the next section.

2.1 Inference and Learning in HMM-VAE

Most of the derivation in this section is adapted from similar work in Ebbers et al. [2017]. For using VAEs to model the emission, the latent standard Normal distribution of a VAE is replaced with a state-specific Normal distribution:

$$p\left(\mathbf{x}_{n,t} | z_{n,t} = k; \theta\right) = \mathcal{N}\left(\mathbf{x}_{n,t}; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{1}$$

where θ denotes the set of graphical model parameters, and n, t are the utterance and frame index. For brevity, we will drop the utterance index here on, and only consider a single utterance of length T with observations $\mathbf{Y} = (\mathbf{y}_1, \ldots, \mathbf{y}_T)$, and latent variables $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ and $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_T)$. We can then give the following equations using conditional independence assumptions from the HMM-VAE.

$$\ln p(\mathbf{Y}|\mathbf{X}; \gamma) = \sum_{t=1}^{T} \ln p\left(\mathbf{y}_t | \mathbf{x}_t; \gamma\right)$$

[1] The data has been published under a CC-0 license at https://voice.mozilla.org/en/datasets.

$$\ln p(\mathbf{X}|\mathbf{Z};\theta) = \sum_{t=1}^{T} \ln p\left(\mathbf{x}_t|z_t;\theta\right)$$

$$\ln p(\mathbf{Z};\theta) = \sum_{t=1}^{T} \ln p\left(z_t|z_{t-1};\theta\right)$$

2.1.1 Inference

In AUD, the inference problem is to find the most likely state sequence $\mathbf{Z}$, given some model parameters and an observed sequence. To approximate the posterior distribution of $\mathbf{X}$ and $\mathbf{Z}$, we use variational inference and mean-field approximation, which gives

$$q(\mathbf{X},\mathbf{Z};\theta,\phi) = q(\mathbf{Z};\theta,\phi)\prod_{t=1}^{T} q\left(\mathbf{x}_t;\phi\right). \tag{2}$$

Since we are using a VAE, the terms $q(\mathbf{x}_t;\phi)$ is modeled by the probabilistic encoder network. Again performing mean field inference, we get

$$\begin{aligned}\ln q(\mathbf{Z};\theta,\phi) &= \mathbb{E}_{q(\mathbf{X};\phi)}[\ln p(\mathbf{X}|\mathbf{Z};\theta)] + \ln p(\mathbf{Z};\theta) + \text{ const.} &(3)\\ &= \sum_{t=1}^{T}\left(b\left(z_t\right) + \ln p\left(z_t|z_{t-1};\theta\right)\right) + \text{const.}, &(4)\end{aligned}$$

where const. is a normalizing constant, and

$$b\left(z_t\right) = \mathbb{E}_{q(\mathbf{x}_t;\phi)}\left[\ln p\left(\mathbf{x}_t|z_t;\theta\right)\right]. \tag{5}$$

With this simplification, the inference can be conveniently performed using the well-known Viterbi algorithm.

2.1.2 Learning

Consider a dataset consisting of observed variables $\mathcal{Y} = (\mathbf{Y}_n)$ and hidden variables $\mathcal{H} = (\mathbf{H}_n)$. It is impossible to compute the total data likelihood, since $\mathcal{H}$ is unobserved. Therefore, the objective is to compute the marginal likelihood of $\mathcal{Y}$, which can be written as

$$\ln p(\mathcal{Y}) = \sum_{n=1}^{N} \ln p\left(\mathbf{Y}_n\right). \tag{6}$$

If we approximate the posterior of $\mathbf{H}_n$ using a proposal distribution q, we can use the definition of Kullback-Leibler divergence to write

$$\begin{aligned}\mathcal{L}\left(\mathbf{Y}_n\right) &= \mathbb{E}_{q(\mathbf{H}_n)}\left[\ln p\left(\mathbf{Y}_n|\mathbf{H}_n\right)\right] + \mathbb{E}_{q(\mathbf{H}_n)}\left[\ln\frac{p\left(\mathbf{H}_n\right)}{q\left(\mathbf{H}_n\right)}\right] &(7)\\ &= \mathbb{E}_{q(\mathbf{H}_n)}\left[\ln p\left(\mathbf{Y}_n|\mathbf{H}_n\right)\right] - \mathrm{KL}\left(q\left(\mathbf{H}_n\right)\|p\left(\mathbf{H}_n\right)\right). &(8)\end{aligned}$$

In our HMM-VAE model, the learning task is to estimate the parameters of the HMM and the VAE so as to maximize the likelihood of generating the observed sequence, i.e., given latent variables $\mathbf{H} = (\mathbf{X},\mathbf{Z})$, we have

$$\mathcal{L}(\mathbf{Y};\gamma,\theta,\phi) = \sum_{t=1}^{T}\mathcal{L}\left(\mathbf{y}_t;\gamma,\theta,\phi\right). \tag{9}$$

and

$$\begin{aligned}\mathcal{L}\left(\mathbf{y}_t;\gamma,\theta,\phi\right) = \frac{1}{L}\sum_{l=1}^{L}&\ln p\left(\mathbf{y}_t|\tilde{\mathbf{x}}_t^{(l)};\gamma\right)\\ &- \mathbb{E}_{\tilde{q}(z_t)}\left[\mathrm{KL}\left(q\left(\mathbf{x}_t;\phi\right)\|p\left(\mathbf{x}_t|z_t;\theta\right)\right)\right]\\ &+ \mathbb{E}_{\tilde{q}(z_{t-1},z_t)}\left[\ln p\left(z_t|z_{t-1};\theta\right)\right] + \text{const.}\end{aligned}$$

As such, it does not involve any structure learning since it is assumed that the structure for the generative process is known beforehand. As computation of the expectation in the second term is expensive, the marginal is computed using Viterbi training rather than a Forward-Backward algorithm. The first term in the likelihood is obtained from the VAE and minimized using backpropagation. The overall loss function can also be optimized using minibatch training and gradient descent over all parameters.

2.2 Informative prior

Our treatment of informative priors is inspired from similar work in Ondel et al. [2018]. Let $\mathbf{H}$ be the hidden variables and $\boldsymbol{\eta}$ denote the global parameters of the model (i.e., the parameters of the HMMs and VAEs, in our case). For AUD, $\boldsymbol{\eta}$ can be separated into acoustic parameters $\boldsymbol{\eta}_A$ and "phonotactic" language model parameters $\boldsymbol{\eta}_L$. The posterior of the parameters, given observed data $\mathbf{X}$, can be written as

$$p(\boldsymbol{\eta}, \mathbf{H}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\eta}, \mathbf{H})}{p(\mathbf{X})}. \tag{10}$$

We can ignore the normalization constant and further simplify this using conditional independence assumptions as

$$p(\boldsymbol{\eta}, \mathbf{H}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{H}, \boldsymbol{\eta}_A) p(\mathbf{H}|\boldsymbol{\eta}_L) p(\boldsymbol{\eta}_L) p(\boldsymbol{\eta}_A). \tag{11}$$

Now, assuming we have a high-resource language S, from where we can obtain the complete data $(\mathbf{X}_s, \mathbf{H}_s)$, we can modify the posterior as

$$p(\boldsymbol{\eta}, \mathbf{H}|\mathbf{X}, \mathbf{X}_s, \mathbf{H}_s) \propto p(\mathbf{X}|\mathbf{H}, \boldsymbol{\eta}_A) p(\mathbf{H}|\boldsymbol{\eta}_L) p(\boldsymbol{\eta}_L) p(\boldsymbol{\eta}_A|\mathbf{X}_s, \mathbf{H}_s). \tag{12}$$

As such, the equations are almost the same, with the exception of the distribution of $\boldsymbol{\eta}_A$, which is now informed by $(\mathbf{X}_s, \mathbf{H}_s)$. Practically, this computation of the posterior of $\boldsymbol{\eta}_A$ from the source language data is still intractable, and so we compute an approximate informative prior as

$$q_1^* = \arg\max_{q_1} \mathrm{E}_{q_1(\boldsymbol{\eta}_A)} \left[\ln p\left(\mathbf{X}_p, \boldsymbol{\eta}_A | \mathbf{u}_p, \right)\right] - \mathrm{D}_{\mathrm{KL}}\left(q_1\left(\boldsymbol{\eta}_A\right) \| p\left(\boldsymbol{\eta}_A\right)\right). \tag{13}$$

For implementation purpose, this just means initializing the acoustic model parameters from those learned using language S, and is therefore a simple application of transfer learning.

3 Semi-implicit Variational Inference

In our formulation of the HMM-VAE earlier, the enocoder distribution $q_\phi(\mathbf{x}|\mathbf{y})$ was required to be reparameterizable and analytically evaluable, which restricts it to a small exponential family. In particular, we chose to represent it using a standard Normal, i.e., $q_\phi(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}(\mathbf{y}, \phi), \boldsymbol{\Sigma}(\mathbf{y}, \phi))$. This means that given observation $\mathbf{y}_i$, its corresponding code $\mathbf{x}_i$ is forced to follow a Gaussian distribution, which means it is often too restrictive to model skewness, kurtosis, and multimodality.

To improve the flexibility of modeling the latent code, rather than using a single-stochastic-layer encoder, we use semi-implicit variational inference (SIVI) that can add as many stochastic layers as needed, as long as the first stochastic layer $q_\phi(\mathbf{x}|\mathbf{y})$ is reparameterizable and has an analytical PDF, and the layers added after are reparameterizable and simple to sample from. More specifically, we construct semi-implicit VAE (SI-VAE) by using a hierarchical encoder that injects random noise at M different stochastic layers. Keeping the equation for $q_\phi(\mathbf{x}|\mathbf{y})$ same as earlier, we modify the parameters of the standard Normal as

$$\boldsymbol{\mu}(\mathbf{y}, \phi) = f(\boldsymbol{\ell}_M, \mathbf{y}; \phi) \quad \text{and} \quad \boldsymbol{\Sigma}(\mathbf{y}, \phi) = g(\boldsymbol{\ell}_M, \mathbf{y}; \phi), \tag{14}$$

where ℓ_t's are defined as

$$\boldsymbol{\ell}_t = T_t(\boldsymbol{\ell}_{t-1}, \boldsymbol{\epsilon}_t, \mathbf{y}; \phi), \boldsymbol{\epsilon}_t \sim q_t(\boldsymbol{\epsilon}), t = 1, \ldots, M, \tag{15}$$

where T_t, f, and g are all deterministic neural networks. This parameterization moves the encoder variational distribution beyond a simple Gaussian form.

4 Experiments

We perform experiments on two low-resource language datasets, namely Turkish (tr; 6 hours) and Dutch (nl; 13 hours), obtained from the publicly available Mozilla Common Voice corpus. The informative prior is learned from the English language based TIMIT corpus. Since our model is unsupervised, we only require the speech waveforms to generate a phonetic transcription. However, for purposes of implicit evaluation, we compare our transcribed phonemes with those obtained from a grapheme-to-phoneme conversion of the provided transcriptions.

We extract 13 MFCC (mel-frequency cepstral coefficient) features and their first and second order time derivatives for each frame, where the framing window is of 25 ms, and successive frames are 10 ms apart. This gives a 39-dimensional feature vector for each frame.

The implicit evaluation results are reported using two scoring methods.

1. **Boundary score**: This is a weaker metric which only evaluates the segmentation performance of the model. Phoneme boundaries are compared between the reference and hypothesis transcriptions, and precision, recall, and f-scores are reported.
2. **Normalized mutual information (NMI)**: This is a stronger evaluation metric which takes into account the actual phonemic units generated. It measures the statistical dependency between the reference and hypothesis transcriptions, and outputs a score between 0 and 1, where 0 means no dependence and 1 means the actual phones could be retrieved without error given the sequence of discovered units. Formally, it is defined as

$$\mathrm{NMI} = \frac{I\left(\mathbf{Z}^{(\mathrm{true})}; \mathbf{Z}^{(\mathrm{pred})}\right)}{H\left(\mathbf{Z}^{(\mathrm{true})}\right)} = \frac{H\left(\mathbf{Z}^{(\mathrm{true})}\right) - H\left(\mathbf{Z}^{(\mathrm{true})} | \mathbf{Z}^{(\mathrm{pred})}\right)}{H\left(\mathbf{Z}^{(\mathrm{true})}\right)}, \tag{16}$$

where $H(\cdot)$ is some entropy measure.

4.1 Implementation details

We adapt and extend the publicly available toolkit BEER (Bayesian Speech Recognizer) for our implementation of semi-implicit HMM-VAEs. BEER is written almost entirely in Python and uses a PyTorch backend [Paszke et al., 2017] for the neural network implementations. The feature extraction and dataset creation recipes are based on the open-source ASR toolkit Kaldi [Povey et al., 2011]. For obtaining the phonemic units from the provided graphemic transcriptions, we use CMU's Epitran toolkit [Mortensen et al., 2018]. We have made our implementation available for the wider research community [2].

5 Results and Discussion

Tables 1a and 1b summarize the results on the dev and test sets for Turkish and Dutch, respectively, obtained using the GMM-HMM, HMM-VAE (with and without an informative prior), and the semi-implicit HMM-VAE models.

All models perform better on Turkish than Dutch. On an average, the NMI for Turkish transcription is almost 50% better than that for Dutch. This is surprising, since we have only 6 hours of Turkish data, as opposed to 13 hours of Dutch. We attribute this to two main reasons. First, Dutch has a higher number of phonetic units than Turkish, which makes fine-grained recognition more difficult. Second, we tuned all our models to the Turkish dev set, and used the same configuration for all subsequent experiments. It may be possible to obtain better results on Dutch by more attentive hyperparameter tuning.

The use of an informative prior gives more improvement on Dutch than on Turkish. This may be because we use an English language corpora for learning the prior, and both English and Dutch belong to the Indo-European language family. Turkish, on the other hand, belongs to the Turkic language family, and therefore shares fewer sounds with English than Dutch. We hypothesize that using a more closely related language for the prior would show better improvement in the results.

[2] https://github.com/desh2608/beer

System	dev				test			
	P	R	F	NMI	P	R	F	NMI
GMM-HMM	87.12	92.07	89.53	0.583	82.88	92.04	87.22	0.561
HMM-VAE	91.09	92.11	91.6	0.598	86.16	91.97	88.97	0.577
HMM-VAE (informative prior)	**91.78**	92.01	**91.9**	**0.61**	**86.89**	92.09	**89.41**	**0.588**
SI HMM-VAE	84.49	**93.87**	88.93	0.581	81.17	**94.38**	87.28	0.569

(a)

System	dev				test			
	P	R	F	NMI	P	R	F	NMI
GMM-HMM	81.63	74.55	77.93	0.406	80.81	74.49	77.52	0.399
HMM-VAE	83.07	75.77	79.25	0.422	82.89	75.91	79.25	0.421
HMM-VAE (informative prior)	**84.88**	**76.16**	**80.28**	**0.443**	84.12	76.6	80.18	**0.435**
SI HMM-VAE	83.11	74.98	78.84	0.428	**84.17**	**76.93**	**80.39**	**0.435**

(b)

Table 1: Summary of results of proposed models for AUD on (a) Turkish (tr) and (b) Dutch (nl) datasets. Precision (P), recall (R), and F-score (F) are reported for the boundary metric, along with the normalized mutual information (NMI) between the reference and hypothesis transcriptions.

The semi-implicit HMM-VAE model is found to perform worse than simple HMM-VAE on Turkish, and almost the same as HMM-VAE on Dutch. However, the training ELBO (empirical lower bound) for the SI HMM-VAE was found to be slightly better than HMM-VAE. This would suggest a classic case of overfitting the training set, which is also found to be more severe for the lower resource Turkish language. We believe that the use of an informative prior would help the semi-implicit model more than the simple HMM-VAE model, since then the state transitions could be fixed and the training data could be more efficiently used to train the emission probabilities.

6 Related Work

Bayesian approaches have become important in unsupervised AUD, especially for very low resource languages. In Lee and Glass [2012], a Dirichlet process hidden Markov model (DPHMM) framework is formulated to simultaneously perform three sub-tasks of segmentation, nonparametric clustering and sub-word modeling, and the spoken term detection task is used to evaluate the learned sub-word models. Ondel et al. [2016] uses a similar generative model, but replaces Gibbs sampling with variational Bayes for the inference procedure. This makes the algorithm parallelizable and improves training efficiency. More recently, Ondel et al. [2018] used informative priors in this setting to perform AUD on a low resource language, and similar to our study, they used the TIMIT English corpus as the source language.

Johnson et al. [2016] proposed the structured VAE in the context of modeling video frames. Ebbers et al. [2017] adapted their model and combined it with HMMs to propose an explicit definition of the HMM-VAE. Their paper also experimented with AUD, but it was in the setting of high resource languages. Semi-implicit variational inference was recently introduced in Yin and Zhou [2018] where they employed the method in problems like Bayesian logistic regression, and also proposed the SI-VAE for improving the negative log evidence in modeling image datasets, such as MNIST. To the best of our knowledge, this is the first study to use SI-VAEs coupled with HMMs for acoustic unit discovery.

7 Conclusion

We proposed several HMM based models for unsupervised acoustic unit discovery. We learned forced alignments using a baseline GMM-HMM model and used these alignments to train a HMM-VAE model, which was found to outperform the former owing to the higher expressivity of a VAE. We also demonstrated the effective use of an informative prior, particularly on languages belonging to the same family. Finally, we extended the VAE using semi-implicit variational inference which is arguably more flexible in modeling the latent code of the generated frames. In future work, we would like to investigate the effect of an informative prior on the semi-implicit model.

References

Janek Ebbers, Jahn Heymann, Lukas Drude, Thomas Glarner, Reinhold Haeb-Umbach, and Bhiksha Raj. Hidden markov model variational autoencoder for acoustic unit discovery. In *INTERSPEECH*, pages 488–492, 2017.

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.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

Matthew Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems*, pages 2946–2954, 2016.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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*, pages 40–49. Association for Computational Linguistics, 2012.

Chunxi Liu, Jinyi Yang, Ming Sun, Santosh Kesiraju, Alena Rott, Lucas Ondel, Pegah Ghahremani, Najim Dehak, Lukaš Burget, and Sanjeev Khudanpur. An empirical evaluation of zero resource acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5305–5309. IEEE, 2017.

David R Mortensen, Siddharth Dalmia, and Patrick Littell. Epitran: Precision g2p for many languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*, 2018.

Lucas Ondel, Lukáš Burget, and Jan Černockỳ. Variational inference for acoustic unit discovery. *Procedia Computer Science*, 81:80–86, 2016.

Lucas Ondel, Pierre Godard, Laurent Besacier, Elin Larsen, Mark Hasegawa-Johnson, Odette Scharenborg, Emmanuel Dupoux, Lukas Burget, François Yvon, and Sanjeev Khudanpur. Bayesian models for unit discovery on a very low resource language. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5939–5943. IEEE, 2018.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. The kaldi speech recognition toolkit. Technical report, IEEE Signal Processing Society, 2011.

Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16, 1986.

Karel Veselỳ, Arnab Ghoshal, Lukás Burget, and Daniel Povey. Sequence-discriminative training of deep neural networks. In *Interspeech*, volume 2013, pages 2345–2349, 2013.

Shinji Watanabe, Yasuhiro Minami, Atsushi Nakamura, and Naonori Ueda. Application of variational bayesian approach to speech recognition. In *Advances in Neural Information Processing Systems*, pages 1261–1268, 2003.

Mingzhang Yin and Mingyuan Zhou. Semi-implicit variational inference. *arXiv preprint arXiv:1805.11183*, 2018.