BRNO UNIVERSITY OF TECHNOLOGY
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF INFORMATION TECHNOLOGY
FAKULTA INFORMAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

DISCOVERING ACOUSTIC UNITS FROM SPEECH: A BAYESIAN APPROACH

BAYESOVSKÝ PŘÍSTUP K URČOVÁNÍ AKUSTICKÝCH JEDNOTEK V ŘEČI

PHD THESIS
DISERTAČNÍ PRÁCE

AUTHOR **LUCAS ONDEL**
AUTOR PRÁCE

SUPERVISOR **LUKÁŠ BURGET**
ŠKOLITEL

BRNO 2020

Abstract

From an early age, infants show an innate ability to infer linguistic structures from the speech signal long before they learn to read and write. In contrast, modern speech recognition systems require large collections of transcribed data to achieve a low error rate. The relatively recent field of Unsupervised Speech Learning has been dedicated to endow machines with a similar ability. As a part of this ongoing effort, this thesis focuses on the problem of discovering a set of acoustic units from a language given untranscribed audio recordings. Particularly, we explore the potential of Bayesian inference to address this problem.

First, we revisit the state-of-the-art non-parametric Bayesian model for the task of acoustic unit discovery and derive a fast and efficient Variational Bayes inference algorithm. Our approach relies on the stick-breaking construction of the Dirichlet Process which allows expressing the model as a Hidden Markov Model-based phone-loop. With this model and a suitable mean-field approximation of the variational posterior, the inference is made with an efficient iterative algorithm similar to the Expectation-Maximization scheme. Experiments show that this approach performs a better clustering than the original model while being orders of magnitude faster.

Secondly, we address the problem of defining a meaningful a priori distribution over the potential acoustic units. To do so, we introduce the *Generalized Subspace Model*, a theoretical framework that allows defining distributions over low-dimensional manifolds in high-dimensional parameter space. Using this tool, we learn a phonetic subspace—a continuum of phone embeddings—from several languages with transcribed recordings. Then, this phonetic subspace is used to constrain our system to discover acoustic units that are similar to phones from other languages. Experimental results show that this approach significantly improves the clustering quality as well as the segmentation accuracy of the acoustic unit discovery system.

Finally, we enhance our acoustic units discovery model by using a Hierarchical Dirichlet Process prior instead of the simple Dirichlet Process. By doing so, we introduce a Bayesian bigram phonotactic language model to the acoustic unit discovery system. This approach captures more accurately the phonetic structure of the target language and consequently helps the clustering of the speech signal. Also, to fully exploit the benefits of the phonotactic language model, we derive a modified Variational Bayes algorithm that can balance the preponderance of the role of the acoustic and language model during inference.

Abstrakt

Děti mají již od útlého věku vrozenou schopnost vyvozovat jazykové znalosti z mluvené řeči - dlouho předtím, než se naučí číst a psát. Moderní systémy pro rozpoznávání řeči oproti tomu potřebují k dosažení nízké chybovosti značná množství přepsaných řečových dat. Teprve nedávno založená vědecká oblast "učení řeči bez supervize" se věnuje přenosu popsaných lidských schopností do strojového učení. V rámci této oblasti se naše práce zaměřuje na problém určení sady akustických jednotek z jazyka, kde jsou k disposici pouze nepřepsané zvukové nahrávky. Pro řešení tohoto problému zkoumáme zejména potenciál bayesovské inference.

V práci nejprve pro úlohu určování akustických jednotek revidujeme využití state-of-the-art neparametrického bayesovského modelu, pro který jsme odvodili rychlý a efektivní algoritmus variační bayesovské inference. Náš přístup se opírá o konstrukci Dirichletova procesu pomocí "lámání hůlky" (stick breaking) umožňující vyjádření modelu jako fonémové smyčky založené na skrytém Markovově modelu. S tímto modelem a vhodnou středopolní (mean-field) aproximací variační posteriorní pravděpodobnosti je inference realizována pomocí efektivního iteračního algoritmu, podobného známému schématu Expectation-Maximization (EM). Experimenty ukazují, že tento přístup zajišťuje lepší shlukování než původní model, přičemž je řádově rychlejší.

Druhým přínosem práce je řešení problému definice smysluplného apriorního rozdělení na potenciální akustické jednotky. Za tímto účelem představujeme zobecněný pod-prostorový model (Generalized Subspace Model) - teoretický rámec umožňující definovat pravděpodobnostní rozdělení v nízkodimenzionálních nadplochách (manifoldech) ve vysokorozměrném prostoru parametrů. Pomocí tohoto nástroje učíme fonetický podprostor — kontinuum vektorových reprezentací (embeddingů) fonémů — z několika jazyků s přepsanými nahrávkami. Pak je tento fonetický podprostor použit k omezení našeho systému tak, aby určené akustické jednotky byly podobné fonémům z ostatních jazyků. Experimentální výsledky ukazují, že tento přístup významně zlepšuje kvalitu shlukování i přesnost segmentace systému pro určování akustických jednotek.

Keywords

Unsupervised Speech Learning, Acoustic Unit Discovery, Bayesian inference, Generalized Subspace Model.

Klíčová slova

Učení řeči bez supervize, určování akustických jednotek, bayesovská inference, zobecněný pod-prostorový model.

Reference

ONDEL, Lucas. *Discovering Acoustic Units from Speech: a Bayesian Approach.* Brno, 2020. PhD thesis. Brno University of Technology, Faculty of Information Technology. Supervisor Lukáš Burget

Discovering Acoustic Units from Speech: a Bayesian Approach

Declaration

Hereby I declare that this doctoral thesis was prepared as an original author's work under the supervision of Dr. Lukáš Burget. All the relevant information sources, which were used during preparation of this thesis, are properly cited and included in the list of references.

.....
Lucas Ondel
August 4, 2020

Acknowledgements

A long time ago, in what seems to be another life, I decided to spend a few months in Brno, Czech Republic... Months have turned to years and, to my bewilderment, here I am, submitting a doctoral thesis. It is sometimes difficult to foresee the consequences of small decisions.

First and foremost, I would like to express my sincere gratitude to Lukáš Burget who successfully tame me and led me all along my studies. I can only hope someday of reaching his skills and knowledge. I would like also to thanks Jan "Honza" Černocký, the benevolent dictator of the Brno speech group, who has been a constant support during all these years.

I would like also to deeply thanks my parents, Henri Ondel and Elisabeth Marinier who have raised the little devil I was and, probably, still am. My two brothers, Quentin and Renaud, also deserve some credits for all the joys, adventures, and sometimes fights we had together... As Régis Loisel have written: "Perhaps the purpose of ageing is to remember we were once a child".

These years in the Czech Republic wouldn't have been the same had I not met Michal and Jana Jurka. Their kindness and friendship are invaluable to me and I shall never forget the time I spend with them. Michal, Jana, words are lacking to express my feelings so let me just say: Já Vám děkuji.

Finally, I would like to address a very special thanks to Jinyi Yang for her support and love during this long journey. Jinyi, my next adventure will be with you.

Contents

Mathematical notation **4**

1 Introduction **5**
1.1 Motivations . . . 5
1.2 Related works . . . 6
1.3 Thesis Contributions . . . 8

2 Non-Parametric Bayesian Phone-Loop Model **10**
2.1 Bayesian formulation of the AUD problem . . . 10
2.1.1 Non-parametric Bayesian AUD . . . 11
2.2 Model . . . 14
2.2.1 Acoustic Model . . . 14
2.2.2 Base measure . . . 16
2.2.3 Generative Process . . . 18
2.2.4 Phone-loop interpretation . . . 19
2.2.5 Joint distribution . . . 20
2.3 Inference . . . 22
2.3.1 VB E-step . . . 23
2.3.2 VB M-step . . . 25
2.3.3 Truncation . . . 27
2.4 Experimental Setup . . . 28
2.4.1 Data . . . 28
2.4.2 Features . . . 29
2.4.3 Metrics . . . 29
2.5 Results and analysis . . . 31
2.5.1 Settings . . . 31
2.5.2 Variational Bayes vs Gibbs Sampling . . . 31
2.5.3 Variational Bayes objective for AUD . . . 33
2.5.4 Discriminative features . . . 33
2.5.5 Non-Parametric vs Parametric Phone-Loop . . . 36
2.6 Conclusion . . . 36

3 Generalized Subspace Model for Sound Representation **38**
3.1 Generalized Subspace Model . . . 38
3.1.1 Definition . . . 39
3.1.2 Relation with the i-vector model . . . 40
3.1.3 Inference . . . 41
3.1.4 Example . . . 42

3.2 Subspace Hidden Markov Model . . . 47
3.2.1 Phonetic subspace . . . 49
3.2.2 Encoding the HMM parameters . . . 50
3.2.3 Example: learning the English phonetic space . . . 53
3.3 Dirichlet Process Subspace Hidden Markov Model . . . 56
3.3.1 Revisiting the base measure . . . 56
3.3.2 Approximating the phonetic subspace of the target language . . . 57
3.4 Results . . . 58
3.4.1 Experimental setup . . . 58
3.4.2 Optimal subspace dimension . . . 59
3.4.3 Benefits of the universal phonetic subspace . . . 60
3.4.4 Comparison with the DP-HMM . . . 60
3.5 Conclusion . . . 61

4 Phonotactic Language Model **64**
4.1 Non-Parametric Bigram Phone-Loop Model . . . 64
4.1.1 Hierarchical Dirichlet Process . . . 65
4.1.2 Stick-Breaking constructions . . . 66
4.1.3 Complete Model . . . 67
4.1.4 Joint distribution . . . 67
4.2 Inference . . . 69
4.2.1 VB-M step for the HDP . . . 69
4.3 Improper Variational Bayes Inference . . . 71
4.4 Results . . . 74
4.4.1 Experimental Setup . . . 74
4.4.2 Bigram vs unigram phonotactic language model . . . 74
4.4.3 Effect of the correction factors . . . 74
4.5 Conclusion . . . 76

5 Conclusion **78**
5.1 Future work . . . 78
5.1.1 Acoustic Modeling . . . 78
5.1.2 Language Modeling . . . 79
5.2 Summary of contributions . . . 80

A Variational Bayes **91**
A.1 Variational Bayes objective . . . 91
A.2 Approximating posterior distributions . . . 92
A.2.1 Parametric approximation . . . 93
A.2.2 Mean-Field approximation . . . 93
A.2.3 Structured mean-field approximation . . . 93

B Exponential Family of Distributions **94**
B.1 Exponential family of distribution . . . 94
B.1.1 Partial derivative of the log-normalizer . . . 94
B.1.2 Conjugate Prior . . . 95
B.2 Distributions . . . 95
B.2.1 Categorical . . . 96
B.2.2 Dirichlet . . . 96

B.2.3 Gamma 97
B.2.4 Normal 97
B.2.5 Normal-Wishart 97

Mathematical notation

Vectors are denoted by lower-case bold Roman or Greek letters such as $\mathbf{x}$ or $\boldsymbol{\lambda}$; they are assumed to be column vectors. Uppercase bold Roman or Greek letters, such as $\mathbf{X}$ or $\boldsymbol{\Lambda}$, denote matrices. A superscript $\top$ denote the transpose of a matrix or a vector. The list of mathematical notations used in this thesis are shown in the following table:

Notation	Name	Description
$\mathrm{tr}(\mathbf{M})$	trace	Sum of the diagonal elements of the square matrix $\mathbf{M}$.
$\mathrm{vec}(\mathbf{M})$	vectorize	Returns all the columns of the matrix $\mathbf{M}$ as a vector.
$\mathrm{mat}(\mathbf{m})$	inverse vectorize	Returns a $D \times D$ square matrix $\mathbf{M}$ from a D^2-dimensional vector $\mathbf{m}$. The matrix is created column-wise.
$\mathrm{diag}(\mathbf{M})$	diagonal	Returns the diagonal elements of a square matrix $\mathbf{M}$.
$\mathrm{ltri}(\mathbf{M})$	lower triangular	Returns the lower triangular part of a matrix (not including the diagonal) as a vector.
$\mathbb{1}[\text{condition}]$	indicator	Returns 1 if "condition" is true, 0 otherwise.
$\delta_{\mathbf{x}}(\mathbf{y})$	Dirac delta function	Returns $+\infty$ if $\mathbf{y} = \mathbf{x}$, 0 otherwise, and $\int_y \delta_{\mathbf{x}}(\mathbf{y})\mathrm{d}\mathbf{y} = 1$.
$\langle a \rangle_{p(x)}$	expectation	Expectation of a with respect to $p(x)$: $\int_x a p(x)\mathrm{d}x$.
$\mathrm{D}_{\mathrm{KL}}(q(x)\|\|p(x))$	Kullback-Leibler divergence	Divergence between two distributions defined as: $\int_x q(x) \ln \frac{q(x)}{p(x)}\mathrm{d}x$.
$\mathcal{B}(\alpha, \beta)$	Beta distribution	See appendix B.2.2.
$\mathcal{C}(\boldsymbol{\pi})$	Categorical distribution	See appendix B.2.1.
$\mathcal{D}(\boldsymbol{\alpha})$	Dirichlet distribution	See appendix B.2.2.
$\mathcal{G}(\boldsymbol{\alpha})$	Gamma distribution	See appendix B.2.3.
$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	Normal distribution	See appendix B.2.4.
$\mathcal{NW}(\mathbf{m}, \beta, \mathbf{W}, \nu)$	Normal-Wishart distribution	See appendix B.2.5.

Chapter 1

Introduction

Speech is a highly structured signal which serves as the primary mean of communication among humans. The easiness and apparent simplicity with which we extract information hide the profound complexity of the speech signal and the human hearing apparatus. In acoustically challenging conditions, human listeners effortlessly decode phones, syllables, words composing the message. Remarkably, infants learn to recognize speech long before to know to read or write (Dupoux, 2018). They learn from a very limited set of speakers (mostly their caregivers) and generalizes very well to other speakers and new acoustic conditions. On the contrary, computers use an extremely large amount of data with high diversity in terms of speakers and recording conditions to achieve similar performance to human listeners (Xiong et al., 2016; Stolcke and Droppo, 2017). The difference between humans and machines is particularly striking as the latter requires very strong supervision whereas humans can learn to hear and speak with little guidance. The field of Unsupervised Speech Learning (USL) (Glass, 2012; Goldwater and Johnson, 2007; Lee, 2014; Drexler, 2016; Kamper et al., 2017a) has been dedicated to endow machines with a similar capability: to learn to recognize the speech signal with little or no supervision. This thesis is our contribution to the USL research field and proposes a Bayesian approach to discover a phonological system—the set of basic sounds called acoustic units used to communicate in a language—from a collection of unlabeled audio recordings.

This introductory chapter is organized as follows: first we motivate the research interest of this thesis in section 1.1. Then, we survey related works in section 1.2 and summarize the contributions of this work in section 1.3.

1.1 Motivations

Automatic Speech Recognition (ASR) and related fields have made tremendous progress over the last 50 years. From the single-speaker digit recognition system proposed by Bell's lab (Davis et al., 1952) to recent large vocabulary continuous speech recognition systems (Sak et al., 2014, 2015; Sercu et al., 2016; Bi et al., 2015; Qian et al., 2016; Yu et al., 2016), the ASR technology has matured to the point where, in certain conditions, it shows similar performance to human listeners (Xiong et al., 2016; Stolcke and Droppo, 2017). The growth of computational resources paired with advanced machine learning techniques has yielded an almost continuous reduction of the error rates over time. Whereas early systems relied on expert-designed rules (David and Selfridge, 1962), the field has gradually

moved to statistical methods extracting empirical statistics from large collections of data. The amount of necessary expert knowledge has decreased to the extent that a state-of-the-art system can be built with solely audio recordings and their corresponding textual transcriptions. However, the reduction of expert knowledge has been succeeded by a drastic increase in the amount of data. Nowadays, commercial systems rely on thousands of hours of transcribed data (Saon et al., 2015; Han et al., 2017; Xiong et al., 2018). These algorithms are so data-hungry that the applicability of ASR systems is limited to the very small set of languages in the world for which there is a sufficient amount of transcribed data and commercial interest. Out of the 7000 languages spoken worldwide (Eberhard, David M., Gary F. Simons, and Charles D. Fennig, 2020), only about a hundred of them are covered by ASR with varying degrees of accuracy[1]. This limitation is problematic as language diversity is diminishing worldwide at an alarming pace. Data-driven methods to discover a phonological system would be a strong help for on-field linguists to quickly document endangered languages. Moreover, for languages having low amount of transcribed data, the data-driven phonetic transcription of speech corpus can bootstrap a wide range of downstream applications such as word discovery (Lee et al., 2015), language identification (Shum et al., 2016), topic identification (Liu et al., 2017; Kesiraju et al., 2017) or text-to-speech (Dunbar et al., 2019).

As already mentioned, infants learn to recognize speech long before they learn to read and write (Dupoux, 2018). The inner details of this process remain largely unknown. Yet, a better understanding of the human speech learning mechanism would have a great impact on our knowledge of the brain and how to help children affected by neurological disorders. Investigation on this matter is complicated for ethical and practical reasons. It is impossible to constantly monitor children from their birth in a non-invasive way and designing experiments with toddlers is particularly difficult due to their limited attention and undeveloped verbal communication skills. An unsupervised machine learning model simulating the acquisition of the phonology—and recognizing speech in general—would be a precious tool to psycho-linguists to better understand the cognitive processes underlying speech acquisition by humans.

Finally, the recent success of machine learning in a wide range of areas has heightened the hope and the interest of our modern societies into building more intelligent systems. However, the traditional approach based on training a deep neural network to discriminate an input into a limited number of classes is very restrictive and severely narrows the range of applications. Indeed, the assumption that we can collect a sufficient amount of labeled data in all situations of interest is unrealistic. Conversely, the whole biosphere shows an incredible capacity to learn and to adapt from its sole sensory data. We believe that the development of unsupervised learning of such a complex signal as speech would be a significant breakthrough in direction of a true—or at least a practical—artificial intelligence.

1.2 Related works

The task of discovering a phonological system from only speech data amounts to solve three sub-problems:

- decomposing the speech into variable-length segments

[1] https://cloud.google.com/speech-to-text/docs/languages

- clustering each of these segments, these clusters are often referred to as *acoustic units*
- finding an appropriate model complexity, that is choosing the appropriate number of clusters necessary to describe the language.

These three sub-tasks have been addressed, jointly or independently in numerous works. In the following, we attempt to give a general overview of the prior work on discovering acoustic units.

Early approaches to discovering acoustic units have treated the segmentation and clustering problem separately: (Cohen, 1981) proposes a dynamic programming based speech segmentation algorithm, (Lee et al., 1988) uses two distinct and independent statistical models to segment and cluster the segments respectively, (Černocký, 1998) decompose the speech signal into quasi-stationary sub-signal before quantizing them, (Garcia and Gish, 2006) uses segmental Gaussian Mixture Model to cluster variable-length sequence of features. These approaches have all in common that the number of acoustic units, i.e. clusters, is a user-defined parameter and cannot be inferred from the data.

Another line of work relies on the Segmental Dynamic Time Warping (S-DTW) algorithm (Park and Glass, 2005; Jansen et al., 2010; Jansen and Van Durme, 2011; Kamper et al., 2017b) In these works, the S-DTW algorithm is used to spot re-occurring pattern in a signal. This approach differs from other works as it tries to directly identify words or syllables rather than phone-like units. The rationale is the following: since words last much longer than phones, they are more easily discovered. While this may seem to be a compelling idea, it has, nevertheless, a severe drawback: the number of words in a language being literally infinite, it is clear that we will never have enough data to discover all possible words. Moreover, clustering word-like units is more difficult as they have low occurrence frequency compared to phones.

More recently, various Bayesian Generative Models (BGM) has been proposed to discover acoustic units (Lee and Glass, 2012; Ondel et al., 2016, 2017; Varadarajan et al., 2008; Kamper et al., 2016, 2017a; Kamper, 2017). These models improve over early approaches such as (Lee et al., 1988) by using a single model to segment and cluster speech together. Moreover, the use of non-parametric Bayesian modeling (Orbanz and Teh, 2010; Teh and Jordan, 2010) allows these models to also infer the number of acoustic units from the data itself. Whereas initial models were trained with Gibbs Sampling, the development of variational methods for non-parametric models (Blei, 2004; Blei et al., 2006) has enabled more efficient and scalable training approaches (Ondel et al., 2016). While BGMs have shown to be more efficient than DTW based methods (Ondel et al., 2018), they have relatively weak modeling power—compared to neural network based models—to preserve the tractability of the training.

Neural networks based generative models have been successfully applied to learn a powerful latent representation of speech (Dunbar et al.; Kamper et al., 2015; Hsu and Glass, 2018; Hsu et al., 2017; Milde and Biemann, 2018; Chorowski et al., 2019). While most of these models are trained in an unsupervised fashion, other works replace the traditional transcription with a different modality such as images or videos (Holzenberger et al., 2019; Merkx et al., 2019; Harwath et al., 2016, 2018). While these models have generally more

modeling capability compared to BGMs, they cannot easily cluster the speech signal as the use of discrete latent variables precludes the back-propagation of gradients. Several works have been proposed to incorporate layers with discrete output either by relaxing discrete distributions (Jang et al., 2016; Maddison et al., 2017) or using some gradient approximation (van den Oord et al., 2017), nevertheless, clustering with neural network remains a difficult issue. Finally, recent works have shown than BGMs can be combined in a principled way with neural networks (Johnson et al., 2016). This line of work is particularly interesting as it yields models that can learn jointly continuous and discrete hierarchical representations of the signals.

1.3 Thesis Contributions

This thesis has three major contributions; each of them is presented in a distinct chapter:

Non-Parametric Bayesian Phone-Loop Model In chapter 2, we revisit a non-parametric Bayesian model for acoustic unit discovery proposed in (Lee and Glass, 2012). Whereas the authors originally used the *Chinese Restaurant Process* to sample from the distribution of the model's parameters, we propose to approximate this posterior distribution with the *Variational Bayes* framework. To achieve this, we describe the generative process of the model with the *Sethuraman stick-breaking construction* of the Dirichlet Process. Then, by choosing an adequately structured mean-field factorization of the variational posterior we show that the training of the model is amenable to a Variational Bayes Expectation-Maximization (VB-EM) algorithm. This new inference scheme is beneficial as it considerably speeds up the training and allows us to discover acoustic units from a larger amount of data.

Generalized Subspace Model for Sound Representation Bayesian approaches for acoustic unit discovery rely on, among other components, a prior distribution over sounds. This prior distribution weighs which sounds are likely to be retained as acoustic units when clustering the speech. In general, this distribution is chosen to be non-informative, that is, it allows potentially any possible sounds to be an acoustic unit. In chapter 3, we propose to build a more refined prior which gives higher weights to a subset of sounds similar to phones from other languages. To do so, we introduce a new theoretical framework: the *Generalized Subspace Model* (GSM). The GSM allows learning low-dimensional embeddings representing probability distribution. In our case, we use the GSM in the following manner:

- given a set of phonetically transcribed speech data (from a different language than the target one), we learn a Hidden Markov Model (HMM) model for each phone.
- using the GSM framework we learn a subspace in the total parameter space of the HMM capturing the phonetic variability
- finally, we set the prior distribution over sounds of the acoustic unit discovery model to be non-zero only on the subspace previously learned.

The GSM is a principle way to incorporate prior information into a model. For the task of acoustic units discovery, we use the GSM to teach the model "what is a phone" (by using transcribed data from other languages) before clustering the speech in the target language. In addition to significantly improve the discovery of acoustic units, the GSM is very flexible and can be applied to a wide family of models.

Phonotactic Language Model Most of the Bayesian models for acoustic units discovery rely on the Dirichlet Process prior. While mathematically convenient, this prior assumes the probability of sequence of acoustic units to be given by an unigram distribution. In chapter 4, we propose to address this limitation by developing a model based on the *Hierarchical Dirichlet Process* (HDP). The HDP is a non-parametric prior which defines a probability over an infinite set of conditional distributions. We use a two-level HDP to build a non-parametric AUD model with bigram transition probabilites between acoustic units. By using *Teh's stick-breaking construction* of the HDP, we derive a VB-EM training algorithm almost identical to the one used for the Dirichlet Process based model. Additionally, to reduce the effect of the features; independence assumption of the HMM, we propose a corrected version of the model by introducing language and acoustic scaling factors. We show that these factors can be easily integrated in the VB-EM training and help to control the preponderance of the acoustic and language models for clustering speech data.

Finally, for the sake of reproducibility, a practical implementation of all the models and experiments presented in this thesis can be found at: https://github.com/beer-asr/beer.

Chapter 2

Non-Parametric Bayesian Phone-Loop Model

This chapter describes a non-parametric Bayesian phone-loop model for AUD. It will serve as a basis for more refined models presented in chapters 3 and 4. It is derived from the combination of the Hidden Markov Model (HMM) (Rabiner, 1989) and non-parametric Bayesian methods (Ferguson, 1973; Rasmussen, 2000; Teh, 2010). Whereas the HMM has been used since the early days of statistical speech recognition (Jelinek, 1976), non-parametric Bayesian methods were introduced more recently in the field of speech and language processing. Their capacity to assign probability to infinite sets has found important applications in language modeling (Teh, 2006; Goldwater et al., 2006), unsupervised text segmentation (Mochihashi et al., 2009), and speaker diarization (Fox et al., 2011). Drawing inspiration from (Goldwater et al., 2009; Fox et al., 2011), the first version of the non-parametric phone-loop model for AUD was proposed in (Lee and Glass, 2012) and paved the way to a Bayesian approach to AUD. Our model revisit the model proposed (Lee and Glass, 2012) by replacing the Chinese Restaurant Process with the stick-breaking representation of the Dirichlet Process. This seemingly minor modification has, however, major consequences:

- it allows the use of the Variational Bayes framework as inference instead of Gibbs Sampling. Therefore, it re-formulates the problem of AUD as an optimization of an objective function.
- it allows to reinterpret the model as a phone-loop model making possible, by means of dynamic programming, to consider all possible sequences of units for a given sequence of speech features
- it allows the parallelization of the training allowing use of bigger corpora.

.

2.1 Bayesian formulation of the AUD problem

We now give a formal definition of the AUD problem within the Bayesian framework. Let $\mathbb{E}$ be a vector space, and $\boldsymbol{\eta} \in \mathbb{E}$ a finite dimensional representation of sounds, i.e. $\boldsymbol{\eta}$ is a sound embedding. Given a sequence of N observations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ of forming a speech utterance, we aim to find:

- A collection of P acoustic units $\mathbf{H} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_P)$ best describing the observations. We denote the selected sounds *acoustic units* as they represent the basic elements of speech. For now, we assume P to be known.

- The sequence of indices $\mathbf{u} = u_1, \ldots, u_L,\ L < N$ where $u_i \in \{1, \ldots, P\}$ is the index of an acoustic unit. Thereafter, we will denote $\mathbf{u}$ as the label sequence. Note that, in practice, L is unknown.

Using Bayes' rule, we can formulate the search of the best set of units $\mathbf{H}^*$ and the best label sequence $\mathbf{u}^*$ in probabilistic terms:

$$\mathbf{H}^*, \mathbf{u}^* = \arg\max_{\mathbf{H},\mathbf{u}} p(\mathbf{H}, \mathbf{u}|\mathbf{X}) \tag{2.1}$$

$$p(\mathbf{H}, \mathbf{u}|\mathbf{X}) = \frac{p(\mathbf{X}|\mathbf{H}, \mathbf{u})p(\mathbf{H}, \mathbf{u})}{\int_{\mathbf{H}} \sum_{\mathbf{u}} p(\mathbf{X}|\mathbf{H}, \mathbf{u})p(\mathbf{H}, \mathbf{u})\mathrm{d}\mathbf{H}} \tag{2.2}$$

Because of the complexity of the task and the multiple way of describing a language phonetically (phonetic features, phones, tri-phones, syllables, ...), the notion of „best solution“ is somewhat tedious. We will therefore focus our attention on the quantity $p(\mathbf{H}, \mathbf{u}|\mathbf{X})$ rather than just the most likely solution given by $\mathbf{H}^*$ and $\mathbf{u}^*$.

The Bayesian statement of AUD given in (2.1) and (2.2) is reminiscent of the statistical formulation of ASR advocated by Frederick Jelinek (Jelinek, 1976). However, in the case of AUD, the inventory of units is unknown and needs to be inferred from the data along with the acoustic description of the units encoded in the embeddings $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots$. Conversely, there is no need for these embeddings in ASR since the acoustic description of the words is assumed to be known or is unnecessary for the so-called *end-to-end* approach to ASR (Graves and Jaitly, 2014).

2.1.1 Non-parametric Bayesian AUD

Until now, we have assumed the number of acoustic units P to be fixed. Choosing a good value for P is, however, non-trivial as we don't know beforehand the type of acoustic units which will be chosen by the AUD algorithm. If the units represent phones, then, P might be between 50 or 100 depending on the language. On the other hand, if the units represent phones in context (di-phone, tri-phone, ...), we need to choose a much larger value for P (several thousand at least). We see that any choice of P implies some assumption and, consequently, will affect the type of acoustic units derived from the algorithm. Rather than making a hard decision, we prefer to let the AUD algorithm to choose an adequate P depending on the given data. Practically, this can be achieved by letting $P \to \infty$ and adding a distribution $\mathcal{P}$ over the parameters of $p(\mathbf{u}, \mathbf{H})$[1]. This approach, referred to as *non-parametric Bayesian* (Orbanz and Teh, 2010), does not put any limit on the model complexity *a priori*. Rather, the model complexity is part of the inference process and, therefore, should be chosen in light of the data. In our case, we set $\mathcal{P}$ to be a Dirichlet Process (Orbanz and Teh, 2010).

[1] Loosely speaking, the distribution $\mathcal{P}$ is a hyper-prior, i.e. a prior over the (parameters of the) prior distribution $p(\mathbf{u}, \mathbf{H})$

The Dirichlet process, denoted $\mathcal{DP}(\gamma, G_0)$, is a stochastic process for which each realization $G(\boldsymbol{\eta})$ is a discrete probability distribution over infinitely many outcomes. Informally, it can be seen has an infinite-dimensional Dirichlet distribution. It is parameterized by a probability distribution $G_0(\boldsymbol{\eta})$ called a *base measure* and a concentration parameter γ. The base measure defines the expectation of the Dirichlet process whereas the concentration controls the spread of the probability mass across the dimensions of the sampled probability distributions. When the concentration is close to 0, most of the probability mass is distributed in a few dimensions and conversely, when the concentration is high, the probability mass will be spread in many dimensions.

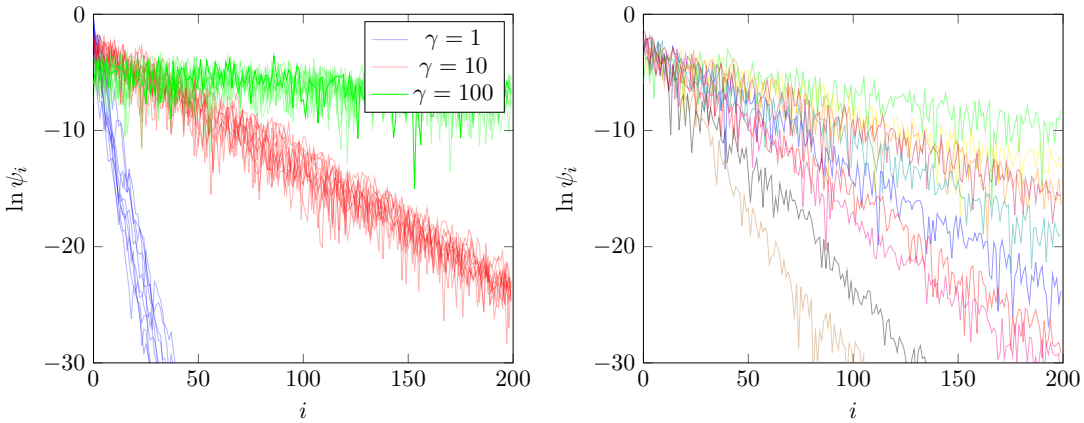
Many Dirichlet process-based models use the Chinese restaurant process as inference scheme (Lee and Glass, 2012; Beal et al., 2002). The Chinese restaurant process is a sampling scheme that draws, in the limit, samples from the posterior distribution over the model's parameters marginalized over all possible distribution G sampled from a Dirichlet process (Rasmussen, 2000). Whereas this approach theoretically guarantees to draw sample from the exact posterior, it also has several issues:

- the theoretical convergence is rarely met in practice as in many cases it involves infinitely long sampling time
- samples are not independent of each other and therefore the training is not easily parallelizable

These drawbacks make the Chinese restaurant process unadapted to speech techonologies which usually require large amounts of data. To address these issues, it is convenient to express the Dirichlet process in terms of the Sethuraman's stick-breaking construction (Sethuraman, 1994):

1. Draw $v_i \sim \mathcal{B}(1, \gamma), \quad i = \{1, 2, \dots\}$
2. Draw $\boldsymbol{\eta}_i \sim G_0, \quad i = \{1, 2, \dots\}$
3. $\psi_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$
4. $G(\boldsymbol{\eta}) = \sum_{i=1}^{\infty} \psi_i \delta_{\boldsymbol{\eta}_i}(\boldsymbol{\eta})$,

where $\mathcal{B}$ is a 2-dimensional Dirichlet distribution (appendix B.2.2) usually called the Beta distribution. The samples from the base measure $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ are referred to as the *atoms* of the sampled probability distribution $G(\boldsymbol{\eta})$. On one hand, this constructive definition of the Dirichlet process introduces the new latent variables $v_1, v_2, \dots$ which are not needed when using the Chinese restaurant process. On the other hand, as it will be described in section 2.3, these new variables make possible to use Variational Bayes to approximate the posterior distribution of the model. The resulting inference algorithm is easily parallelizable and allows to process much larger collection of data.



(a) Stick-Breaking Process

(b) Gamma Stick-Breaking Process

Figure 2.1: Difference between the standard stick-breaking process with various concentration parameters and the stick-breaking process with a Gamma prior. The abscissa represents the indices of the portions of the stick and the ordinate represents the logarithm of these portions (i.e. the log-probabilities of the infinite mixture components). In Fig. 2.1a each line is a draw from the stick-breaking process with a specific concentration; there are 10 draws for each concentration setting (1, 10, 100). In Fig. 2.1b each line is a draw from the stick-breaking process with concentration parameter sampled from the Gamma prior. The Gamma distribution was parameterized by $a_0 = 1$ (shape) and $b_0 = 10$ (rate). The Gamma prior increases the uncertainty of the stick-breaking and let the model choose an adequate value for the concentration γ from the data.

In the context of our AUD model, we use a Dirichlet process to construct the prior $p(\mathbf{u}, \mathbf{H})$ in the following way:

$$G(\boldsymbol{\eta}) \sim \mathcal{DP}(\gamma, G_0) \tag{2.3}$$

$$p(\mathbf{u}, \mathbf{H}) = \underbrace{\left[\prod_{n=1}^{L} \underbrace{G(\boldsymbol{\eta}_{u_n})}_{p(u_n|\mathbf{H})}\right]}_{p(\mathbf{u}|\mathbf{H})} \underbrace{\left[\prod_{k=1}^{\infty} G_0(\boldsymbol{\eta}_k)\right]}_{p(\mathbf{H})} \tag{2.4}$$

where L is the length of the sequence of labels $\mathbf{u}$. Note that since we assume $P \to \infty$, the matrix of embeddings $\mathbf{H} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ has an infinite number of columns. It is important to understand the different roles played by the two terms in (2.4). On one hand, $G_0(\boldsymbol{\eta})$ is a continuous density over the embedding space: it defines which embeddings are likely to be selected as acoustic units. On the other hand, $G(\boldsymbol{\eta}_{u_n})$ is a discrete distribution over an infinite set of atoms and it defines how frequently a unit occurs in speech. In other words, G is a (unigram) language model of the units.

Even though the Dirichlet process assumes a potentially infinite number of classes, it may favour solution with small or large number of units depending on its concentration parameter γ. As can be observed from Figure 2.1a, the concentration parameter γ strongly constrains samples from the Dirichlet process. This constraint can be relaxed by augmenting

the stick-breaking process with a Gamma prior (appendix B.2.3) over the concentration parameter $\gamma \sim \mathcal{G}(a_0, b_0)$[2] leading to a modified stick-breaking process:

1. Draw $\gamma \sim \mathcal{G}(a_0, b_0)$
2. Draw $v_i \sim \mathcal{B}(1, \gamma), \quad i = \{1, 2, \dots\}$
3. ...

As seen from Fig. 2.1b, the Gamma prior increases the variance of the standard stick-breaking process. Therefore, this avoids the issue of choosing a specific concentration parameter as we can infer it from the data directly. Note that the inference is particularly simple as the Gamma distribution is conjugate (appendix B.1.2) to the stick-breaking process.

2.2 Model

The Bayesian formulation of the AUD problem given in section 2.1 does not specify a concrete model. More precisely, one needs to define the acoustic model $p(\mathbf{X}|\mathbf{H}, \mathbf{u})$ and the base measure $G_0(\boldsymbol{\eta})$ in order to estimate the posterior $p(\mathbf{H}, \mathbf{u}|\mathbf{X})$. In this section, we describe both elements and connect them with the stick-breaking representation of the Dirichlet process completing the definition of the non-parametric Bayesian phone-loop AUD model.

2.2.1 Acoustic Model

We define the acoustic model assuming that, given a sequence of N observations $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)$ and a sequence of L units, the likelihood factorizes as:

$$p(\mathbf{X}|\mathbf{H}, \mathbf{u}) = \prod_{l=1}^{L} p(\mathbf{X}^{u_l}|\mathbf{H}, u_l) = \prod_{l=1}^{L} p(\mathbf{X}^{u_l}|\boldsymbol{\eta}_{u_l}), \tag{2.5}$$

where $\mathbf{X}^{u_l}$ is the sequence of observations associated to the lth unit such that $\mathbf{X} = \mathbf{X}^{u_1}, \dots, \mathbf{X}^{u_L}$. We assume this segmentation to be known even though this is not true in practice. This issue will naturally disappear when we reinterpret the full AUD model as a large HMM in section 2.2.4. Following (Lee and Glass, 2012), we set the likelihood $p(\mathbf{X}^{u_l}|\boldsymbol{\eta}_{u_i})$ to be modeled by an HMM with S hidden states and GMM state's emission density with C components:

$$p(\mathbf{X}^{u_l}|\boldsymbol{\eta}_{u_l}) = \sum_{\mathbf{s}^{u_l}} \sum_{\mathbf{c}^{u_l}} p(\mathbf{X}^{u_l}, \mathbf{c}^{u_l}, \mathbf{s}^{u_l}|\boldsymbol{\pi}^1_{u_l}, \dots, \boldsymbol{\pi}^S_{u_l}, \boldsymbol{\mu}^{1,1}_{u_l}, \dots, \boldsymbol{\mu}^{S,C}_{u_l}, \boldsymbol{\Sigma}^{1,1}_{u_l}, \dots, \boldsymbol{\Sigma}^{S,C}_{u_l}) \tag{2.6}$$

$$= \sum_{\mathbf{s}^{u_l}} \sum_{\mathbf{c}^{u_l}} \prod_{n=1}^{N_l} p(\mathbf{x}^{u_l}_n, c^{u_l}_n|\boldsymbol{\pi}^{s_n}_{u_l}, \boldsymbol{\mu}^{s_n,1}_{u_l}, \dots, \boldsymbol{\mu}^{s_n,C}_{u_l}, \boldsymbol{\Sigma}^{s_n,1}_{u_l}, \dots, \boldsymbol{\Sigma}^{s_n,C}_{u_l}) p(s^{u_l}_n|s^{u_l}_{n-1}) \tag{2.7}$$

where N_l is the length of the sequence of observations $\mathbf{X}^{u_l}$ and $p(s^{u_l}_1|s^{u_l}_0) = p(s^{u_l}_1)$ is the probability of the initial state. The parameters and the latent variables introduced in (2.6) correspond to the traditional parameterization of an HMM:

[2] We use the shape/rate parameterization of the Gamma distribution.

- $\mathbf{s}^{u_l} = s_1^{u_l}, \dots, s_{N_l}^{u_l}$ is the sequence of indices of the HMM states for acoustic unit u_l
- $\mathbf{c}^{u_l} = c_1^{u_l}, \dots, c_{N_l}^{u_l}$ is the sequence of indices of the mixture components for the acoustic unit u_l
- $\boldsymbol{\pi}_{u_l}^{i}$ are the mixing weights of the GMM associated to the ith state of the HMM of the acoustic unit u_l
- $\boldsymbol{\mu}_{u_l}^{i,j}$ is the mean of the jth Normal component of the GMM associated to the ith state of the HMM of acoustic unit u_l
- $\boldsymbol{\Sigma}_{u_l}^{i,j}$ is the covariance matrix of the jth component of the GMM associated to the ith state of the HMM of acoustic unit u_l

Notice that we have not included any parameters of the transition probabilities $p(s_n^{u_l}|s_{n-1}^{u_l})$ as it has been empirically observed that they play no significant role when modeling speech (Bourlard, 1996). Consequently, we assume the transition probabilities are fixed parameters such that the probability to go to any state given the current state is the same.

We specify now the relation between the embedding $\boldsymbol{\eta}_{u_l}$ of the acoustic unit with index u_l and the corresponding HMM parameters. First, observe that the joint distribution of $p(\mathbf{x}_n^{u_l}, c_n^{u_l}|s_n^{u_l}, \dots)$ is a product of Normal and Categorical distributions and each of them is a member of the exponential family of distribution (appendix B). Therefore we have:

$$p(\mathbf{x}_n^{u_l}, c_n^{u_l}|s_n^{u_l}, \dots) = p(\mathbf{x}_n^{u_l}|\boldsymbol{\mu}_{u_l}^{s_n,c_n}, \boldsymbol{\Sigma}_{u_l}^{s_n,c_n})p(c_n^{u_l}|\boldsymbol{\pi}_{u_l}^{s_n}) \tag{2.8}$$

$$p(c_n|\boldsymbol{\pi}_{u_l}^{s_n}) = p(c_n|\boldsymbol{\omega}_{u_l}^{s_n}) = \exp\{\boldsymbol{\omega}_{u_l}^{s_n\top} T(c_n^{u_l}) - A(\boldsymbol{\omega}_{u_l}^{s_n})\} \tag{2.9}$$

$$p(\mathbf{x}_n^{u_l}|\boldsymbol{\mu}_u^{s_n,c_n}, \boldsymbol{\Sigma}_{u_l}^{s_n,c_n}) = p(\mathbf{x}_n^{u_l}|\boldsymbol{\theta}_{u_l}^{s_n,c_n}) = \exp\{\boldsymbol{\theta}_{u_l}^{s_n,c_n\top} T(\mathbf{x}_n^{u_l}) - A(\boldsymbol{\theta}_{u_l}^{s_n,c_n})\} \tag{2.10}$$

where $\boldsymbol{\omega}_{u_l}^{s_n}$, $T(c_n^{u_l})$ and $A(\boldsymbol{\omega}_{u_l}^{s_n})$ are the natural parameters, the sufficient statistics and the log-normalizer of the Categorical distribution of the state with index $s_n^{u_l}$. Similarly, $\boldsymbol{\theta}_{u_l}^{s_n,c_n}$, $T(\mathbf{x}_n)$ and $A(\boldsymbol{\theta}_{u_l}^{s_n,c_n})$ are the natural parameters, the sufficient statistics and the log-normalizer of the Normal distribution associated with state $s_n^{u_l}$ and mixture's component $c_n^{u_l}$. Note that to keep the notation uncluttered we write $T(\mathbf{x}), T(c)$, $A(\boldsymbol{\omega})$, $A(\boldsymbol{\theta})$ instead of $T_x(\mathbf{x})$, $T_c(c)$, $A_\omega(\boldsymbol{\omega})$, $A_\theta(\boldsymbol{\theta})$. For both distributions, the natural parameters and the sufficient statistics can be derived from their respective definition (appendices B.2.4 B.2.1):

$$\boldsymbol{\omega}_{u_l}^{s_n} = \begin{bmatrix} \ln\left(\frac{\pi_{u_l,1}^{s_n}}{1-\sum_{k=1}^{C-1}\pi_{u_l,k}^{s_n}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{u_l,C-1}^{s_n}}{1-\sum_{k=1}^{C-1}\pi_{u_l,k}^{s_n}}\right) \end{bmatrix} \qquad T(c_n^{u_l}) = \begin{bmatrix} \mathbb{1}[c_n^{u_l} = 1] \\ \dots \\ \mathbb{1}[c_n^{u_l} = C-1] \end{bmatrix} \tag{2.11}$$

$$\boldsymbol{\theta}_{u_l}^{s_n,c_n} = \begin{bmatrix} \boldsymbol{\theta}_{u_l,1}^{s_n,c_n} \\ \boldsymbol{\theta}_{u_l,2}^{s_n,c_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{i,j}^{-1}\boldsymbol{\mu}_{i,j} \\ -\frac{1}{2}\operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix} \qquad T(\mathbf{x}_n^{u_l}) = \begin{bmatrix} \mathbf{x}_n \\ \operatorname{vec}(\mathbf{x}_n^{u_l}\mathbf{x}_n^{u_l\top}) \end{bmatrix}, \tag{2.12}$$

where „vec“ is the vectorization operation. Note that $\boldsymbol{\omega}$ is a $(C-1)$-dimensional vector whereas $\boldsymbol{\pi}$ is a C-dimensional vector. This difference comes from the fact that the weights $\pi_1, \dots, \pi_C$ are constrained such that $0 < \pi_i < 1$ and $\sum_{i=1}^{C}\pi_i = 1$. Finally, the log-normalizers

are defined as:

$$A(\boldsymbol{\omega}_{u_l}^{s_n}) = \ln\Big(\sum_{c_n} \exp\{\boldsymbol{\omega}_{u_l}^{s_n\top} T(c_n^{u_l})\}\Big) \tag{2.13}$$

$$= \ln(1 + \sum_{k=1}^{C-1} \exp\{\omega_{u_l,k}^{s_n}\}) \tag{2.14}$$

$$A(\boldsymbol{\theta}_{u_l}^{s_n,c_n}) = \ln\Big(\int \exp\{\boldsymbol{\theta}_{u_l}^{s_n,c_n\top}(\mathbf{x}_n^{u_l})\}\mathrm{d}\mathbf{x}_n\Big) \tag{2.15}$$

$$= -\frac{1}{4}\boldsymbol{\theta}_{u_l,1}^{s_n,c_n\top}\,\mathrm{mat}(\boldsymbol{\theta}_{u_l,2}^{s_n,c_n})^{-1}\boldsymbol{\theta}_{u_l,1}^{s_n,c_n} - \frac{1}{2}\ln|-2\,\mathrm{mat}(\boldsymbol{\theta}_{u_l,2}^{s_n,c_n})| + \frac{D}{2}\ln 2\pi, \tag{2.16}$$

where „mat“ is the inverse of the vectorization operator, that is it takes as input a D^2-dimensional vector and returns a $D \times D$ square matrix. We define the embedding $\boldsymbol{\eta}_{u_l}$ to be the concatenation of the natural parameters of the Normal and Categorical distributions of all S states of the HMM modeling the acoustic unit with index u_l. Formally, $\boldsymbol{\eta}_{u_l}$ can be seen as the „super-vector“ of all the parameters of acoustic unit u_l and its layout is defined as:

$$\boldsymbol{\eta}_{u_l} = \begin{bmatrix} \boldsymbol{\eta}_{u_l}^1 \\ \vdots \\ \boldsymbol{\eta}_{u_l}^i \\ \vdots \\ \boldsymbol{\eta}_{u_l}^S \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega}_{u_l}^i \\ \boldsymbol{\theta}_{u_l}^{i,1} \\ \vdots \\ \boldsymbol{\theta}_{u_l}^{i,C} \end{bmatrix}, \tag{2.17}$$

where $\boldsymbol{\eta}_{u_l}^i$ is the concatenation of the natural parameters of the Normal and Categorical distributions for the ith state of the HMM modeling the acoustic unit with index u_l.

2.2.2 Base measure

As discussed previously, the base measure is the distribution describing a priori which sounds (represented as embeddings) are likely to be retained as an acoustic unit. In our case, we have defined an embedding $\boldsymbol{\eta}$ to be the vector of natural parameters of an HMM. We set G_0 to be the conjugate prior (appendix B.1.2) of the conditional HMM likelihood:

$$G_0(\boldsymbol{\eta}) = \prod_{i=1}^{S} p(\boldsymbol{\omega}^i) \prod_{j=1}^{C} p(\boldsymbol{\theta}^{i,j}) \tag{2.18}$$

$$= \exp\{\sum_{i=1}^{S} \boldsymbol{\xi}_0^\top T(\boldsymbol{\omega}^i) - A(\boldsymbol{\xi}_0) + \sum_{j=1}^{C} \boldsymbol{\vartheta}_0^\top T(\boldsymbol{\theta}^{i,j}) - A(\boldsymbol{\vartheta}_0)\}. \tag{2.19}$$

Practically, this implies that the prior over the mixture weights $\boldsymbol{\pi}$ is Dirichlet distribution (appendix B.2.2) and the prior over mean vector $\boldsymbol{\mu}$ and the (inverse) covariance matrix $\boldsymbol{\Sigma}^{-1}$ is a Normal-Wishart distribution (appendix B.2.5). (2.18) can be equivalently expressed as

a prior over the standard parameters as:

$$G_0(\boldsymbol{\eta}) = \prod_{i=1}^{S} p(\boldsymbol{\pi}^i) \prod_{j=1}^{C} p(\boldsymbol{\mu}^{i,j}, \boldsymbol{\Sigma}^{i,j-1}) \tag{2.20}$$

$$p(\boldsymbol{\pi}^i) = \mathcal{D}(\boldsymbol{\alpha}_0) \tag{2.21}$$

$$p(\boldsymbol{\mu}^{i,j}, \boldsymbol{\Sigma}^{i,j-1}) = \mathcal{NW}(\mathbf{m}_0, \beta_0, \mathbf{W}_0, \nu_0) \tag{2.22}$$

Where $\mathcal{D}$ and $\mathcal{NW}$ are the Dirichlet and Normal-Wishart (see the appendices B.2.2 and B.2.5 for details about their parameters). This choice is convenient since, due to the conjugacy, it greatly simplifies the inference, however, it is difficult to control precisely which type of sounds the base measure will emphasize. This issue will be addressed in chapter 3. The natural parameters $\boldsymbol{\xi}_0$, $\boldsymbol{\vartheta}_0$, the sufficient statistics $T(\boldsymbol{\omega}^i)$, $T(\boldsymbol{\theta}^{i,j})$ and the log-normalizing functions $A(\boldsymbol{\xi}_0)$, $A(\boldsymbol{\vartheta}_0)$ of the base measure $G_0(\boldsymbol{\eta})$ can be derived from the definition of the Dirichlet and Normal-Wishart distributions:

$$\boldsymbol{\xi}_0 = \begin{bmatrix} \alpha_{0,1} - 1 \\ \vdots \\ \alpha_{0,C-1} - 1 \\ (\sum_{j=1}^{C} \alpha_{0,j}) - C \end{bmatrix} \tag{2.23}$$

$$T(\boldsymbol{\omega}^i) = \begin{bmatrix} \boldsymbol{\omega}^i \\ -A(\boldsymbol{\omega}^i) \end{bmatrix} \tag{2.24}$$

$$A(\boldsymbol{\xi}_0) = \big(\ln\Gamma(\xi_{0,C} + C) + \sum_{i=1}^{C-1} \ln\Gamma(\xi_{0,i} + 1)\big) - \ln\Gamma(\xi_{0,i} + C) \tag{2.25}$$

$$\boldsymbol{\vartheta}_0 = \begin{bmatrix} \beta_0 \mathbf{m}_0 \\ -\frac{\beta_0}{2} \\ -\frac{1}{2}\operatorname{vec}(\beta_0 \mathbf{m}_0 \mathbf{m}_0^\top + \mathbf{W}_0^{-1}) \\ \frac{\nu_0 - D}{2} \end{bmatrix} \tag{2.26}$$

$$T(\boldsymbol{\theta}^{i,j}) = \begin{bmatrix} \boldsymbol{\theta}^{i,j} \\ -A(\boldsymbol{\theta}^{i,j}) \end{bmatrix} \tag{2.27}$$

$$A(\boldsymbol{\theta}^{i,j}) = -\ln B \tag{2.28}$$

$$B = \beta_0^{\frac{D}{2}} |\mathbf{W}_0|^{-\frac{\nu_0}{2}} \Big(2^{\frac{(\nu_0+1)D}{2}} \pi^{\frac{D(D+1)}{4}} \prod_{d=1}^{D} \Gamma(\frac{\nu_0 + 1 - d}{2})\Big)^{-1}. \tag{2.29}$$

To summarize, an acoustic unit with index u is modeled by an HMM with natural parameters $\boldsymbol{\eta}_u$. The prior probability over each acoustic unit embedding is the conjugate of the HMM likelihood conditioned on its latent variable (s_n and c_n). The relation between the HMM and the base measure is illustrated in Fig. 2.2. All together, the AUD model can be understood as a mixture of HMM with an infinite number of components. Intuitively, inference with such model amounts to cluster segments of the speech signal into temporal patterns.

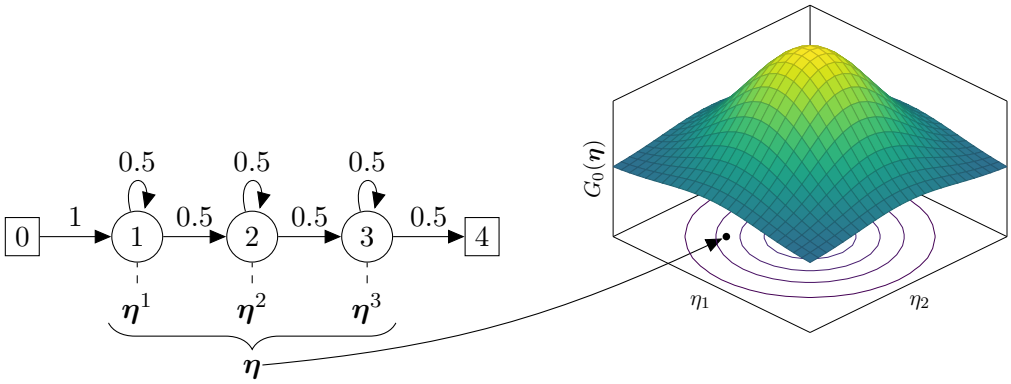


Figure 2.2: Model of an acoustic unit and its relation with the base measure. Each acoustic unit is parameterized by a vector of natural parameters $\boldsymbol{\eta}$ corresponding to the concatenation of all the HMM states' parameters. The base measure, G_0, is a density over the acoustic (natural) parameter space. Therefore, it defines a priori which sounds are likely to be selected as acoustic units. The topology of the HMM and the transition probabilities are the same for each acoustic unit. The square nodes 0 and 4 are the non-emitting start and end states respectively. Here, we have represented the embedding space as a 2-dimensional space (dimensions η_1 and η_2) but in practice, the embeddings live in a much higher dimensional space (several thousands of dimensions at least).

2.2.3 Generative Process

We have introduced the different elements of the AUD model separately. We assemble them now to present the full generative process using the stick-breaking process and a HMM for each acoustic unit:

1. Draw $\gamma \sim \mathcal{G}(a_0, b_0)$
2. Draw $v_i \sim \mathcal{B}(1, \gamma), \quad i = \{1, 2, \dots\}$
3. Draw $\boldsymbol{\eta}_i \sim G_0, \quad i = \{1, 2, \dots\}$
4. $\psi_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$
5. Draw a sequence of units $\mathbf{u}$, $u_j \sim \mathcal{C}(\boldsymbol{\psi})$
6. For each u_j in $\mathbf{u}$
   (a) Draw a state path $\mathbf{s} = s_1, \dots, s_l$ from the HMM transition probability distribution
   (b) for each state s_k in $\mathbf{s}$:
      i. Draw a component $c_k \sim \mathcal{C}(\boldsymbol{\pi}_{u_j}^{s_k})$ from the state's mixture weights
      ii. Draw a data point $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{u_j}^{s_k,c_k}, \boldsymbol{\Sigma}_{u_j}^{s_k,c_k})$

Note that $\boldsymbol{\pi}_{u_j}^{s_k}$, $\boldsymbol{\mu}_{u_j}^{s_k,c_k}$ and $\boldsymbol{\Sigma}_{u_j}^{s_k,c_k}$ are obtained from the natural parameters $\boldsymbol{\eta}_{u_j}$. The graphical representation of the generative process is shown in Figure 2.3. The model is essentially composed of several layers of latent variables, each of them capturing some specific aspect of the speech signal. The first layer ($\mathbf{c}$) quantizes the continuous features space $\mathbf{x}$, the second layer, ($\mathbf{s}$) captures the temporal dynamic of the signal and finally, the

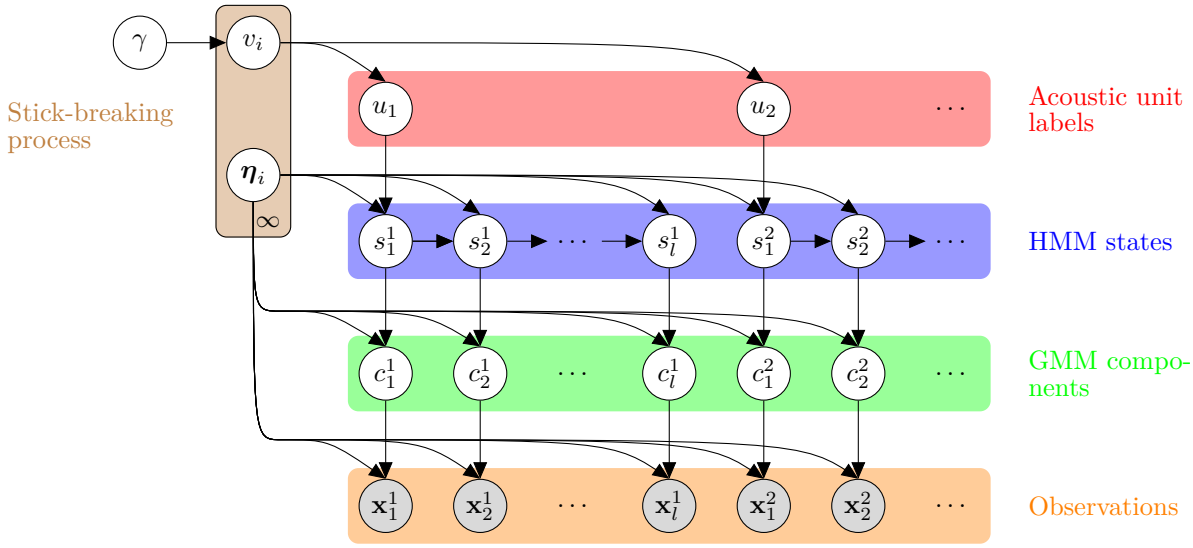


Figure 2.3: Bayesian network of the non-parametric acoustic unit clustering model for a given segmentation. a_i^j refers to the variable a associated to the ith segment of the jth unit. l is the duration of the first unit u_1. Note that in practice the segmentation is unknown and the inference needs to evaluate all possible segmentations.

last layer ($\mathbf{u}$) captures the phonetic information. Finally, despite the fact that the model has many parameters and latent variables, the whole generative process is fully controlled by the following hyper-parameters:

- a_0 and b_0: the parameters of the Gamma distribution control the range of likely values for the concentration of the Dirichlet process.
- $\boldsymbol{\xi}_0$ (or equivalently $\boldsymbol{\alpha}_0$): the parameters of the prior over the GMM mixing weights
- $\boldsymbol{\vartheta}_0$ (or equivalently $\beta_0, \mathbf{m}_0, \mathbf{W}_0, \nu_0$): the parameters of the prior over the mean and precision matrix of each mixture component of the GMMs.

2.2.4 Phone-loop interpretation

The AUD model is a special case of a Hierarchical HMM (Fine et al., 1998) where $p(\mathbf{s}|\mathbf{u})p(\mathbf{u})$ can be interpreted as two nested Markov processes[3]. Estimating the posterior over the latent variable $\mathbf{s}$ and $\mathbf{u}$ given the observations $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_T$ takes $O(T^3)$ time, making the inference impractical. To alleviate this problem, we follow (Murphy and Paskin, 2002) and re-interpret our model as a single level HMM which, consequently, reduces the inference time to $O(T)$.

Converting the 2-level HHM to a flat 1-level HMM requires to merge the sequence of units $\mathbf{u}$ and states $\mathbf{s}$ into a sequence of a single variable $\mathbf{z} = z_1, z_2, \ldots$. Let be $\mathbb{U}$ and $\mathbb{S}$ the sets of the possible units and states such that $\forall i, j\ u_i \in \mathbb{U}$ and $s_j \in \mathbb{S}$. We set $\mathbf{z} = z_1, z_2, \ldots$

[3]In this case, $p(u_t|u_{t-1})$ is simply $p(u_t)$.

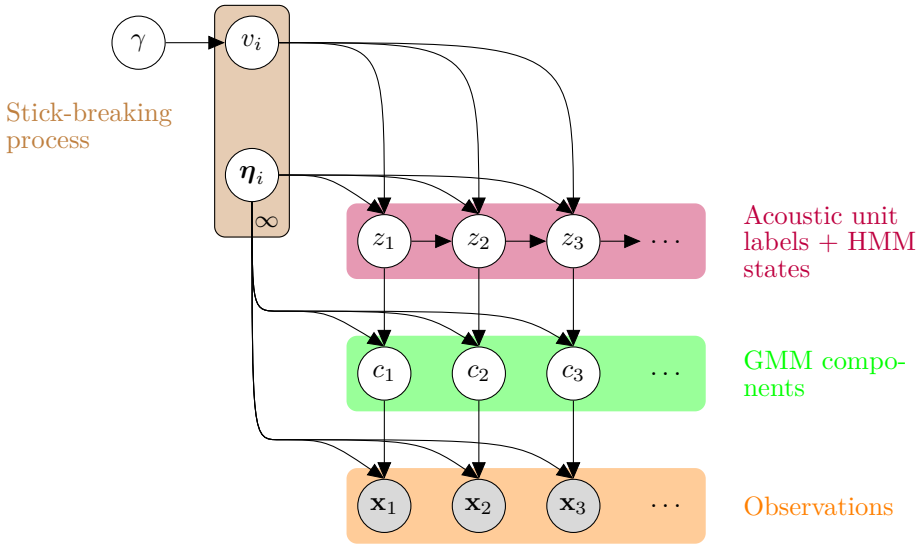


Figure 2.4: Bayesian network of the non-parametric acoustic unit clustering model re-interpreted as a single HMM. Variable z jointly encodes variables u and s. This model is equivalent to the Hierachical HMM model shown in Fig. 2.3 but allows inference of $\mathbf{z}$ given $\mathbf{X}$ in linear time.

such that $z_i \in \mathbb{U} \times \mathbb{S}$ is the concatenation of a particular unit and state. The new generative process induced by the re-parameterization is shown in Fig. 2.4 The probability of sequence $\mathbf{z}$ is given by $p(\mathbf{z}) = p(\mathbf{s}, \mathbf{u})$. as $p(\mathbf{s})$ is a Markov chain, so is $p(\mathbf{z})$ whose graph is represented in Figure 2.5. From this standpoint, the AUD model is equivalent to a non-parametric Bayesian version of the traditional *phone-loop*[4] model which has been applied in several related speech task (Lee and Hon, 1989; Stolcke et al., 2005; Szöke et al., 2010). Also, in the case where we model each acoustic unit by a single state HMM, the infinite phone-loop model reduces to a special case of the infinite HMM (Beal et al., 2002) and of the infinite GMM (Rasmussen, 2000).

Interestingly, merging the two variables has another benefit: it naturally takes into account the segmentation of the observations. Indeed, the variable z_n encodes a unit label for each time step making the state sequence $\mathbf{z}$ to encode the per-frame alignment between the observations and the unit label sequence $\mathbf{u}$. This observation significantly ease up the inference of the model as: (i) it removes the necessity of having an extra boundary variable for segmentation as in (Lee and Glass, 2012), (ii) it allows using dynamic programming to sum over all possible sequences $\mathbf{z}$.

2.2.5 Joint distribution

Finally, to conclude the description of the model, we present the complete joint distribution of a sequence of features $\mathbf{X} = \mathbf{x}_1, \ldots, \mathbf{x}_N$, latent variables $\mathbf{c} = c_1, \ldots, c_N$, $\mathbf{z} = z_1, \ldots, z_N$ and parameters $\mathbf{H} = \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_\infty, \mathbf{v} = v_1, \ldots, v_\infty$, γ. Recall that z_n encodes an acoustic unit index u_n and a particular HMM state s_n. Consequently, we write $\boldsymbol{\eta}_{z_n} = \boldsymbol{\eta}_{u_n}^{s_n}$ which

[4]Obviously, *phone* should be understood as *acoustic unit*

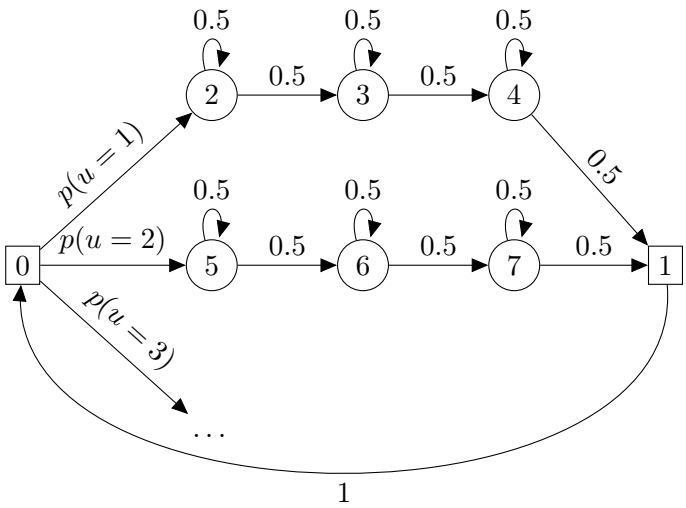


Figure 2.5: Graphical representation of the latent Markov chain of the AUD model reinterpreted as a 1-level HMM. The square nodes 0 and 1 are non-emitting states. The number of phones (i.e. acoustic units), represented by a 3 states left-to-right sub-HMM, is infinite.

corresponds to the natural parameters of the s_nth HMM state of the acoustic unit with index u. Furthermore, the sequence of N units and states $\mathbf{z} = z_1, \ldots, z_N$ can be equivalently defined as a sequence of L acoustic units $\mathbf{u} = u_1, \ldots, u_L$ and L sequences of HMM states $\mathbf{s}^{u_l} = s_1^{u_l}, \ldots, s_{N_l}^{u_l}$. Using these two equivalent formulations, the joint distribution can be written as:

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) = p(\mathbf{H})p(\gamma)p(\mathbf{v}|\gamma)p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{v}) \tag{2.30}$$

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{v}) = p(\mathbf{z}|\mathbf{v})p(\mathbf{X}, \mathbf{c}|\mathbf{z}, \mathbf{H}) \tag{2.31}$$

$$= \prod_{n=1}^{N} p(z_n|z_{n-1}, \mathbf{v})p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) \tag{2.32}$$

$$= \underbrace{\prod_{l=1}^{L} p(u_l|\mathbf{v}) \prod_{n=1}^{N_l} p(s_n^{u_l}|s_{n-1}^{u_l})}_{\prod_{n=1}^{N} p(z_n|z_{n-1},\mathbf{v})} p(\mathbf{x}_n^{u_l}, c_n^{u_l}|\boldsymbol{\eta}_{u_l}^{s_n}), \tag{2.33}$$

where (2.32) is the likelihood expressed as a „flat“ HMM and (2.33) is the likelihood expressed as a two-level hierarchical HMM. Note that we assume z_0 to be a predefined non-emitting starting state as depicted in Fig. 2.5. As explained in section 2.2.1, the per-state emission likelihood (2.32) is a mixture of Normal distributions which is most easily expressed in terms of the natural parameters $\boldsymbol{\omega}_{u_l}^{s_n} = \boldsymbol{\omega}_{z_n}$ and $\boldsymbol{\theta}_{u_l}^{s_n,c_n} = \boldsymbol{\theta}_{z_n}^{c_n}$:

$$p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n}) = p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}^{c_n})p(c_n|\boldsymbol{\omega}_{z_n}) \tag{2.34}$$

$$p(c_n|\boldsymbol{\omega}_{z_n}) = \exp\{\boldsymbol{\omega}_{z_n}^T T(c_n) - A(\boldsymbol{\omega}_{z_n})\} \tag{2.35}$$

$$p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}^{c_n}) = \exp\{\boldsymbol{\theta}_{z_n}^{c_n T}(\mathbf{x}_n) - A(\boldsymbol{\theta}_{z_n}^{c_n})\} \tag{2.36}$$

The transition probability $p(z_n|z_{n-1}, \mathbf{v})$ is more conveniently expressed in terms of variables u_l and $s_n^{u_l}$. The transition probability within a unit's HMM is fixed: $p(s_n|s_{n-1}) =$ const,

so that all states are equiprobables. The probability of the unit index $p(u_l|\mathbf{v})$ is defined by the stick-breaking process as defined in section 2.1.1:

$$p(u_l|\mathbf{v}) = v_{u_l} \prod_{i=1}^{u_l-1} (1 - v_i). \tag{2.37}$$

The prior over the embeddings $\mathbf{H}$ is defined from the base measure:

$$p(\mathbf{H}) = \prod_{u=1}^{\infty} G_0(\boldsymbol{\eta}_u) \tag{2.38}$$

$$G_0(\boldsymbol{\eta}_u) = \prod_{i=1}^{S} p(\boldsymbol{\omega}_u^i) \prod_{j=1}^{C} p(\boldsymbol{\theta}_u^{i,j}) \tag{2.39}$$

$$= \exp\{\sum_{i=1}^{S} \boldsymbol{\xi}_0^\top T(\boldsymbol{\omega}_u^i) - A(\boldsymbol{\xi}_0) + \sum_{j=1}^{C} \boldsymbol{\vartheta}_0^\top T(\boldsymbol{\theta}_u^{i,j}) - A(\boldsymbol{\vartheta}_0)\}. \tag{2.40}$$

Finally, the prior over the stick-breaking process parameters $\mathbf{v}$ and the prior over the concentration parameter γ are given by:

$$p(\mathbf{v}|\gamma) = \prod_{i=1}^{\infty} p(v_i|\gamma) \tag{2.41}$$

$$p(v_i|\gamma) = \mathcal{B}(1, \gamma) \tag{2.42}$$

$$p(\gamma) = \mathcal{G}(a_0, b_0). \tag{2.43}$$

2.3 Inference

As described previously in Section 2.1, given an appropriate model, the AUD task can be cast as inferring the posterior distribution of the model's parameters given a set of data. In the case of the phone-loop model described in this chapter, we aim to estimate the following distribution:

$$p(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma|\mathbf{X}) = \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{p(\mathbf{X})} \tag{2.44}$$

$$p(\mathbf{X}) = \int_\gamma \int_{\mathbf{v}} \int_{\mathbf{H}} \sum_{\mathbf{c},\mathbf{z}} p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) \mathrm{d}\mathbf{H}\mathrm{d}\mathbf{v}\mathrm{d}\gamma. \tag{2.45}$$

As the denominator in (2.44) involves an intractable sum over all possible parameters, estimating $p(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma|\mathbf{X})$ is infeasible. We use the Variational Bayes framework (appendix A) to find an approximate posterior $q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)$. Practically, this amounts to optimize the following lower-bound:

$$\ln p(\mathbf{X}) \geq \langle \ln \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)} \rangle_{q(\mathbf{c},\mathbf{z},\mathbf{H},\mathbf{v},\gamma)} = \mathcal{L}, \tag{2.46}$$

where we write: $\langle f(x)\rangle_{q(x)} = \int_x f(x)q(x)\mathrm{d}x$. To be able to optimize our objective function (2.46), we use the following *structured mean-field* factorization (appendix A.2.3):

$$q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) = q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{H})q(\mathbf{v})q(\gamma) \tag{2.47}$$

$$q(\mathbf{H}) = \prod_{i=1}^{\infty} q(\boldsymbol{\eta}_i) \tag{2.48}$$

$$q(\mathbf{v}) = \prod_{i=1}^{\infty} q(\mathbf{v}_i). \tag{2.49}$$

From (2.46) and (2.47), it directly follows that the optimal factors are given by:

$$\ln q^*(\mathbf{c}|\mathbf{z}) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{H})q(\mathbf{v})q(\gamma)} + \text{const} \tag{2.50}$$

$$\ln q^*(\mathbf{z}) = \langle \ln \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{q(\mathbf{c}|\mathbf{z})}\rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{H})q(\mathbf{v})q(\gamma)} + \text{const} \tag{2.51}$$

$$\ln q^*(\mathbf{H}) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{v})q(\gamma)} + \text{const} \tag{2.52}$$

$$\ln q^*(\mathbf{v}) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{H})q(\gamma)} + \text{const} \tag{2.53}$$

$$\ln q^*(\gamma) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{H})q(\mathbf{v})} + \text{const}\,. \tag{2.54}$$

These equations lead to a training akin to the Expectation-Maximisation (EM) algorithm (Dempster et al., 1977) where we alternately estimate $q(\mathbf{c}|\mathbf{z})$ and $q(\mathbf{z})$ (E-step) and $q(\mathbf{H})$, $q(\mathbf{v})$ and $q(\gamma)$ (M-step). Because of the conjugacy between the prior and the likelihood, we readily see that the optimal factors $q^*(\mathbf{H})$, $q^*(\mathbf{v})$ and $q^*(\gamma)$ will have the same parametric form as their corresponding priors. Also, note that $q(\mathbf{H})$ and $q(\mathbf{v})$ are distributions over an infinite set of random variables and, therefore, cannot be used in any practical implementation. In sections 2.3.1 and 2.3.2, we derive the optimal factors given in (2.47) ignoring this technical issue. In section 2.3.3, we address this issue by truncating the variational posterior, leading to a tractable algorithm.

2.3.1 VB E-step

We assume $q(\mathbf{H})$, $q(\mathbf{v})$ and $q(\gamma)$ are fixed and we estimate the variational posteriors $q^*(\mathbf{c}|\mathbf{z})$ and $q^*(\mathbf{z})$. We start by deriving the optimal variational posterior over the mixture components:

$$\begin{aligned}
\ln q^*(\mathbf{c}|\mathbf{z}) &= \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{H})q(\mathbf{v})q(\gamma)} + \text{const} && (2.55)\\
&= \langle \ln p(\mathbf{X}, \mathbf{c}|\mathbf{z}, \mathbf{H})\rangle_{q(\mathbf{H})} + \text{const} && (2.56)\\
&= \sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})\rangle_{q(\boldsymbol{\eta}_{z_n})} + \text{const} && (2.57)\\
&= \sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}^{c_n})p(c_n|\boldsymbol{\omega}_{z_n})\rangle_{q(\boldsymbol{\theta}_{z_n}^{c_n})q(\boldsymbol{\omega}_{z_n})} + \text{const} && (2.58)\\
\implies q^*(\mathbf{c}|\mathbf{z}) &= \prod_{n=1}^{N} q^*(c_n|z_n) && (2.59)\\
q^*(c_n|z_n) &= \frac{\exp\{\langle \ln p(\mathbf{x}_n, c_n, \mathbf{H}|z_n)\rangle_{q(\mathbf{H})}\}}{\sum_{j=1}^{C} \exp\{\langle \ln p(\mathbf{x}_n, c_n = j, \mathbf{H}|z_n)\rangle_{q(\mathbf{H})}}, && (2.60)
\end{aligned}$$

where C is the number of Normal components per state. The expected likelihood has the following form:

$$\langle \ln p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n}^{c_n})p(c_n|\boldsymbol{\omega}_{z_n})\rangle_{q(\boldsymbol{\theta}_{z_n}^{c_n})q(\boldsymbol{\omega}_{z_n})} = \langle T(\boldsymbol{\omega}_{z_n})\rangle_{q(\boldsymbol{\omega}_{z_n})}^{\top} \begin{bmatrix} T(c_n) \\ 1 \end{bmatrix}\} \tag{2.61}$$

$$+ \langle T(\boldsymbol{\theta}_{z_n}^{c_n})\rangle_{q(\boldsymbol{\theta}_{z_n}^{c_n})}^{\top} \begin{bmatrix} T(\mathbf{x}_n) \\ 1 \end{bmatrix}, \tag{2.62}$$

where $T(\boldsymbol{\omega}_{z_n}) = T(\boldsymbol{\omega}_{u_l}^{c_n})$ and $T(\boldsymbol{\theta}_{z_n}^{c_n}) = T(\boldsymbol{\theta}_{u_l}^{s_n,c_n})$ are defined in (2.24) and (2.27) respectively. The expectations of these functions will be detailed when we derive the optimal variational posterior of the parameters in section 2.3.2. Using (2.60), we can now find the optimal posterior of the global HMM state sequence:

$$\ln q^*(\mathbf{z}) = \langle \ln \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)}{q(\mathbf{c}|\mathbf{z})} \rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{H})q(\mathbf{v})q(\gamma)} + \text{const} \tag{2.63}$$

$$= \sum_{n=1}^{N} \langle \ln \frac{p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})}{q(c_n|z_n)} \rangle_{q(c_n|z_n)q(\boldsymbol{\eta}_{z_n})} + \langle \ln p(z_n|z_{n-1}\mathbf{v})\rangle_{q(\mathbf{v})} + \text{const}\,. \tag{2.64}$$

For the sake of clarity, we define the following variables:

$$\phi_n(z_n) = \langle \ln \frac{p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})}{q(c_n|z_n)} \rangle_{q(c_n|z_n)q(\boldsymbol{\eta}_{z_n})} \tag{2.65}$$

$$A_{z_{n-1},z_n} = \langle \ln p(z_n|z_{n-1}, \mathbf{v})\rangle_{q(\mathbf{v})}, \tag{2.66}$$

which leads to the following formulation of the optimal factor:

$$\ln q^*(\mathbf{z}) = \sum_{n=1}^{N} \phi_n(z_n) + A_{z_{n-1},z_n} + \text{const} \tag{2.67}$$

$$\implies q^*(\mathbf{z}) = \frac{1}{\zeta} \prod_{n=1}^{N} \exp\{\phi_n(z_n) + A_{z_{n-1},z_n}\} \tag{2.68}$$

$$\zeta = \sum_{\mathbf{z}} \prod_{n=1}^{N} \exp\{\phi_n(z_n) + A_{z_{n-1},z_n}\}. \tag{2.69}$$

The normalization constant ζ in (2.69) requires to sum over all possible state sequences $\mathbf{z}$ which is impractical. Nevertheless, this large summation can be computed exactly and efficiently by dynamic programming. Using the associativity and the distributivity properties of the sum and product operations, we have:

$$\zeta = \sum_{z_N} \exp\{\phi_N(z_N)\} \sum_{z_{N-1}} \exp\{A_{z_{N-1},z_N}\} \tag{2.70}$$

$$\times \prod_{n=1}^{N-1} \sum_{z_n} \exp\{\phi_n(z_n) + A_{z_{n-1},z_n}\} \tag{2.71}$$

which can be re-written as a recursive "forward" function $\alpha_n(z_n)$:

$$\zeta = \sum_{z_N} \alpha_N(z_N) \tag{2.72}$$

$$\alpha_n(z_n) = \exp\{\phi_n(z_n)\} \sum_{z_{n-1}} \exp\{A_{z_{n-1},z_n}\}\alpha_{n-1}(z_{n-1}) \tag{2.73}$$

$$\alpha_0(z_0) = 1. \tag{2.74}$$

Alternately, one can derive a recursion flowing backward in time:

$$\zeta = \sum_{z_1} \exp\{\phi_1(z_1) + A_{z_0,z_1}\}\beta_1(z_1) \tag{2.75}$$

$$\beta_n(z_n) = \sum_{z_{n+1}} \exp\{\phi_{n+1}(z_{n+1}) + A_{z_n,z_{n+1}}\}\beta_{n+1}(z_{n+1}) \tag{2.76}$$

$$\beta_N(z_N) = 1. \tag{2.77}$$

In the context of HMM, the computation of (2.74) and (2.77) is known as the *forward-backward* algorithm (Rabiner, 1989), or the *Baum-Welch* algorithm (Baum, 1972). The α_n and β_n recursive functions will proved to be useful to compute the VB M-step.

2.3.2 VB M-step

We now assume that $q(\mathbf{c}|\mathbf{z})$and $q(\mathbf{z})$ are fixed and derive the optimal distribution $q^*(\mathbf{H})$, $q^*(\mathbf{v})$ and $q^*(\gamma)$. Contrary to the VB E-step, the three variational posteriors are assumed to be independent and, therefore, the order is irrelevant. We begin with the posterior over the acoustic unit embeddings:

$$\ln q^*(\mathbf{H}) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma)\rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{v})q(\gamma)} + \text{const} \tag{2.78}$$

$$= \Big[\sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n, c_n|\boldsymbol{\eta}_{z_n})\rangle_{q(c_n|z_n)q(z_n)}\Big] + \sum_{k=1}^{\infty} \ln G_0(\boldsymbol{\eta}_k) + \text{const}\,. \tag{2.79}$$

Using the definition of the base measure in (2.18) and the notation $\boldsymbol{\omega}_{u_l}^{s_n} = \boldsymbol{\omega}_{z_n}$ and $\boldsymbol{\theta}_{u_l}^{s_n,c_n} = \boldsymbol{\theta}_{z_n}^{c_n}$, we write:

$$\ln q^*(\mathbf{H}) = \Big[\sum_{n=1}^{N} \langle \ln p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n})p(c_n|\boldsymbol{\omega}_{z_n})\rangle_{q(c_n|z_n)q(z_n)}\Big] + \sum_{i=1}^{\infty} \ln p(\boldsymbol{\omega}_i) + \sum_{j=1}^{C} \ln p(\boldsymbol{\theta}_i^j) + \text{const} \tag{2.80}$$

$$\implies q^*(\mathbf{H}) = \prod_{i=1}^{\infty} q^*(\boldsymbol{\omega}_i) \prod_{j=1}^{C} q^*(\boldsymbol{\theta}_i^j) \tag{2.81}$$

$$q^*(\boldsymbol{\omega}_i) = \exp\{\boldsymbol{\xi}_i^\top T(\boldsymbol{\omega}_i) - A(\boldsymbol{\xi}_i)\} \tag{2.82}$$

$$\boldsymbol{\xi}_i = \boldsymbol{\xi}_0 + \sum_{n=1}^{N} q(z_n = i) \begin{bmatrix} T(c_n) \\ 1 \end{bmatrix} \tag{2.83}$$

$$q^*(\boldsymbol{\theta}_i^j) = \exp\{\boldsymbol{\vartheta}_i^{j\top} T(\boldsymbol{\theta}_i^j) - A(\boldsymbol{\vartheta}_i^j)\} \tag{2.84}$$

$$\boldsymbol{\vartheta}_i^j = \boldsymbol{\vartheta}_0 + \sum_{n=1}^{N} q(c_n = j|z_n = i)q(z_n = i) \begin{bmatrix} T(\mathbf{x}_n) \\ 1 \end{bmatrix} \tag{2.85}$$

Where the distribution $q(z_n)$ is computed using the forward-backward algorithm:

$$q(z_n) = \frac{\alpha_n(z_n)\beta_n(z_n)}{\zeta}. \tag{2.86}$$

The optimal factors in (2.82) and (2.84) correspond to the natural form of the Dirichlet and Normal-Wishart distributions. The expectations of sufficient statistics $T(\boldsymbol{\omega}_i)$ and $T(\boldsymbol{\theta}_i^j)$ needed in the E-step are given in appendices B.2.2 and B.2.5.

We derive now the optimal variational posterior of the stick-breaking process:

$$\ln q^*(\mathbf{v}) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) \rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{c})q(\mathbf{H})q(\gamma)} + \text{const} \tag{2.87}$$

$$\ln q^*(\mathbf{v}) = \langle \ln p(\mathbf{z}|\mathbf{v}) \rangle_{q(\mathbf{z})} + \ln p(\mathbf{v}) + \text{const}\,. \tag{2.88}$$

$$\tag{2.89}$$

Using the fact that $p(\mathbf{z}|\mathbf{v}) = p(\mathbf{s}|\mathbf{u})p(\mathbf{u}|\mathbf{v})$ and (2.37) we have:

$$\ln q^*(\mathbf{v}) = \langle \ln p(\mathbf{u}|\mathbf{v}) \rangle_{q(\mathbf{u})} + \ln p(\mathbf{v}|\gamma) + \text{const} \tag{2.90}$$

$$= \sum_{k=1}^{\infty} \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i > k] \ln(1 - v_k) + \mathbb{1}[u_i = k] \ln v_k \Big\rangle_{q(\mathbf{u})} \tag{2.91}$$

$$+ (\langle \gamma \rangle_{q(\gamma)} - 1) \ln(1 - v_k) + \text{const}$$

$$\implies q^*(\mathbf{v}) = \prod_{k=1}^{\infty} q^*(v_k) \tag{2.92}$$

$$q^*(v_k) = \mathcal{B}(1 + \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = k] \Big\rangle_{q(\mathbf{u})}, \quad \langle \gamma \rangle_{q(\gamma)} + \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i > k] \Big\rangle_{q(\mathbf{u})}), \tag{2.93}$$

where we have used the indicator operator defined as:

$$\mathbb{1}[\text{condition}] = \begin{cases} 1 & \text{if “condition” is true} \\ 0 & \text{otherwise.} \end{cases} \tag{2.94}$$

The expectations in (2.93) requires summing over all the units of all possible sequences $\mathbf{u}$. Once again, this large summation can be calculated exactly with the forward-backward recursion. Observing that $\sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = k] = \sum_{z_{i-1}, z_i \in \mathbf{z}} \mathbb{1}[z_{i-1} = a_{u_{i-1}}, z_i = b_{u_i}]$, where $a_{u_{i-1}}$ is the index of the last state of the HMM associated with the acoustic unit u_{i-1} and b_k is the index of the first state of the HMM associated with the acoustic unit with index k. Therefore we have:

$$\Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = k] \Big\rangle_{q(\mathbf{u})} = \Big\langle \sum_{n=1}^{N} \mathbb{1}[z_{n-1} = a_{u_{i-1}}, z_n = b_k] \Big\rangle_{q(\mathbf{z})} \tag{2.95}$$

$$= \sum_{n=1}^{N} \sum_{a_{u_{i-1}}} q(z_{n-1} = a_{u_{i-1}}, z_n = b_k) \tag{2.96}$$

$$\Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i > k] \Big\rangle_{q(\mathbf{u})} = \sum_{j=k+1}^{\infty} \langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = j] \rangle_{q(\mathbf{u})} \tag{2.97}$$

$$q(z_{n-1}, z_n) = \frac{1}{\zeta} \sum_{n=1}^{N} \alpha_{n-1}(z_{n-1}) \exp\{\phi_n(z_n) + A_{z_{n-1}, z_n}\} \beta_n(z_n). \tag{2.98}$$

Algorithm 2.1 Training of phone-loop model for acoustic unit discovery

```
function MSTEP(X, q*(c|z), q*(z), q*(γ))
    ▷ Update defined in (2.80)
    q*(H) ← arg max_{q(H)} L
    ▷ Update defined in (2.92)
    q*(v) ← arg max_{q(v)} L
    ▷ Update defined in (2.102)
    q*(γ) ← arg max_{q(γ)} L
    return q*(H), q*(v), q*(γ)

function ESTEP(X, q(H), q(v))
    ▷ Update defined in (2.60)
    q*(c|z) ← arg max_{q(c|z)} L
    ▷ Update defined in (2.68)
    q*(z) ← arg max_{q(z)} L
    return q*(c|z), q*(z)

procedure TRAIN(X, E)
    ▷ E: number of epochs (i.e. E-step + M-step)
    ▷ initialization:
    q*(H) ← random initialization
    q*(v) ← p(v)
    q*(γ) ← p(γ)
    for e ← 1 to E do
        q*(c|z), q*(z) ← ESTEP(X, q*(H), q*(v))
        q*(H), q*(v), q*(γ) ← MSTEP(X, q*(c|z), q*(z), q*(γ))
```

Finally, we estimate the optimal variational posterior over the concentration of the Dirichlet Process:

$$\ln q^*(\gamma) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{v}, \gamma) \rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{H})q(\mathbf{v})} + \text{const} \tag{2.99}$$

$$= \langle \ln p(\mathbf{v}|\gamma) \rangle_{q(\mathbf{v})} + \ln p(\gamma) + \text{const} \tag{2.100}$$

$$= \left[\sum_{k=1}^{\infty} \ln \gamma + \gamma \langle \ln(1 - v_k) \rangle_{q(v_k)} \right] + (a_0 - 1) \ln \gamma - b_0 \gamma \tag{2.101}$$

$$\implies q^*(\gamma) = \mathcal{G}(a_0 + \sum_{k=1}^{\infty} 1, \quad b_0 - \sum_{k=1}^{\infty} \langle \ln(1 - v_k) \rangle_{q(v_k)}) \tag{2.102}$$

2.3.3 Truncation

The optimal variational factors we have derived so far are impractical. Indeed, they involve distributions over infinite set of outcomes ($z_n \in \{1, 2, ..., \infty\}$), infinite-dimensional variables ($\mathbf{H}$, $\mathbf{v}$) and infinite sums in (2.102). Following (Blei et al., 2006), we address this issue by introducing a truncation parameter τ such that $q(v_\tau = 1) = 1, \forall i$. This approximation, motivated by the almost sure truncation of the Dirichlet Process (Ishwaran and James, 2001), ensures that $q(u_i > \tau) = 0$ and, therefore, truncates all infinite sum in the solution

of the optimal factors. Consequently, even if our model theoretically assumes a potentially infinite number of acoustic units, our variational approximation expects at most τ acoustic units. It is important to note that the parameter τ does not define the total number of units derived by the inference. Rather, it is an upper-bound of the maximal number of acoustic units discovered by the model.

The whole training of the model is summarized in Alg. 2.1. Note that this algorithm may converge to a local optimum and therefore needs to be carefully initialized. In practice, we set our initial estimates as:

- $q^*(\mathbf{v}) \triangleq p(\mathbf{v})$
- $q^*(\gamma) \triangleq p(\gamma)$

The posterior over the embeddings $q^*(\mathbf{H})$ is initialized such that the expected value of the prior over the mean parameters of the mixture components is equal to the total data mean plus some noise with small variance.

2.4 Experimental Setup

2.4.1 Data

Our first experimental data set is the TIMIT corpus (Garofolo et al., 1990; Zue et al., 1993). It has a long history and played a key role in the development of acoustic models for speech recognition (Lopes and Perdigao, 2011). Its data is unrealistic and artificial: very clean recordings, no spontaneity, read speech,... However, the controlled quality of the recordings and the manually created phonetic labels make it an ideal data set for developing and testing new speech technologies. TMIT contains phonetically-balanced English read speech recorded at 16kHz. The full corpus has 6300 utterances—about 5.3 hours—and is divided into 438 male speakers and 192 female speakers. There are three groups of sentences: the SA sentences that are read by every speaker to highlight the within-language phonetic variability, and the SX and SI groups which contain phonetically-compact and phonetically-diverse sentences respectively. Since the AUD task is a special case of clustering, there is no need for neither a test set nor a held-out set. Therefore, we trained and evaluated our model on the full corpus without the SA sentences. The choice of removing the SA utterances is common in speech recognition (Lopes and Perdigao, 2011). The overall training data had 4288 utterances (about 3.5 hours). Also, each utterance is provided with a phonetic transcription based on 61 phones. We use these transcriptions as a reference to evaluate the outcome of the AUD model. Contrary to what is usually done in ASR, the 61 phones were not collapsed into a 48 or 39 phone set.

As second experimental data set, we used the MBOSHI corpus (Godard et al., 2017). Contrary to TIMIT, the MOBSHI corpus is closer to a real scenario of documenting an endangered language: it is a set of 16kHz recordings in Mboshi, a Bantu language spoken in Congo-Brazzaville. Similarly to TIMIT, the recommended training and testing sets were merged together forming a set of 5130 utterances from 3 male speakers. The word transcription of the corpus is based on a non-standard graphemic system developed by linguists. In addition, the MBOSHI corpus provides a phonetic time-aligned transcription obtained by forced-alignments of an HMM-GMM based monophone system. This phonetic transcription, based on 68 phones, was used in our evaluation.

Language	Language Family	Hours of training data
Cantonese	Sino-tibetan Chinese	65.0
Pashto	Indo-European	64.7
Turkish	Ural-Altaic	56.6
Tagalog	Austronesian	44.1
Vietnamese	Austroasiatic	53.2
Assamese	Indo-Aryan	46.7
Bengali	Indo-European	53.6
Haitian Creole	French Creole	55.0
Lao	Kra-Dai	71.6
Tamil	Dravidian	72.7
Zulu	Niger-Congo	57.8
Kurdish	Indo-European	69.7
Tok Pisin	English Creole	68.7
Cebuano	Austronesian	70.8
Kazach	Turkic	73.0
Telugu	Dravidian	71.7
Lithuanian	Indo-European	81.4

Table 2.1: List of languages used to train the MBN features extractor.

2.4.2 Features

Because AUD is an unsupervised learning problem, the type of observations given as input to the model is of crucial importance. We considered two types of representation: spectral based features and discriminatively trained features. For the spectral features, the signal is converted into vectors of 12 Mel-Frequency Cepstral Coefficients (MFCCs) (Davis and Mermelstein, 1980) and the signal energy extracted from 25 ms long analysis window at 10 ms rate. These 13-dimensional features are further expanded by adding their first and second derivatives yielding a 39-dimensional feature vector for every 10 ms of speech. To reduce the speaker variability, we applied per-utterance mean normalization. For the discriminative features, we used the Multilingual BottleNeck (MBN) features (Fér et al., 2017). The MBN features are extracted at a 10 ms rate from a 80-dimensional bottleneck layer of a feed-forward neural network trained to classify senones of multiple languages. The neural network was trained on 17 languages listed in Table 2.1; none of them were English. As the neural network is trained on 8kHz recorded speech data, the data was downsampled prior to be presented to the neural network.

2.4.3 Metrics

Evaluating the derived acoustic units is particularly difficult for several reasons. First, the acoustic units, represented as embedding vectors, are not easily interpretable. Second, there is not a single representation of a language's phonology: some representations are compact with coarse level of details, others are more refined but requires more acoustic units. Finally, it is difficult to assess whether the acoustic units capture only the phonetics of the language or if they encode other information such as speaker of channel variability. To cope with these difficulties, we used two metrics to evaluate (i) how closely the data driven

segmentation of speech matches the one of the reference time-aligned transcription (ii) how consistent is the clustering of segmented speech with respect to reference transcription.

Segmentation To evaluate how the phone-loop AUD model segments the speech signal, we compare the segmentation of the most likely sequence of units $\mathbf{u}^*$ (the Viterbi path) with the segmentation of the reference transcription. Practically, we report the Recall, Precision and F-score calculated using the time boundaries of reference labels and the acoustic unit labels. We tolerated boundaries shifted by +- 20 ms. However, a time boundary of an acoustic unit can only match at most one time boundary of the reference transcription.

Normalized Mutual Information To evaluate the quality of the clustering, we computed the Normalized Mutual Information (NMI) between the most likely sequence of units $\mathbf{u}^*$ and the reference transcription $\mathbf{r}$. The NMI is given by:

$$\text{NMI} = 2\frac{H(u) - H[u|r]}{H[u] + H[r]} \tag{2.103}$$

$$H[u] = -\sum_{k=1}^{\tau} \tilde{p}(u = k) \log_2 \tilde{p}(u = k) \tag{2.104}$$

$$H[u|r] = -\sum_{k=1}^{\tau} \sum_{l=1}^{R} \tilde{p}(u = k, r = l) \log_2 \tilde{p}(u = k|r = l) \tag{2.105}$$

$$H[r] = -\sum_{k=1}^{R} \tilde{p}(r = k) \log_2 \tilde{p}(r = k) \tag{2.106}$$

$$\tilde{p}(u = k) = \frac{\sum_{u_i \in \mathbf{u}^*}^{L} \mathbb{1}[u_i = k]}{L} \tag{2.107}$$

$$\tilde{p}(r = k) = \frac{\sum_{r_i \in \mathbf{r}^*}^{M} \mathbb{1}[r_i = k]}{M} \tag{2.108}$$

where R is the number of unique phones in the reference transcription and M is the length of the reference transcription. The conditional probability $\tilde{p}(u|r)$ was estimated by first, mapping each element of sequence $\mathbf{u}^*$ to the one of sequence $\mathbf{r}$ it overlaps the most with, and then, by normalizing the counts of how many times a particular acoustic unit is mapped to a phone of the reference transcription. The NMI is minimal (NMI = 0) when both sequence $\mathbf{u}^*$ and $\mathbf{r}$ are statistically unrelated, on the other hand, the NMI will be maximal (NMI = 1) when there exists a one-to-one mapping between the acoustic units and the reference phones. Importantly, the NMI penalizes AUD systems that uses more units than necessary, or put in another way, when $H[r] < H[u]$. Therefore, we consider solutions that have less acoustic units preferable. When the number of "active units" is lower than the actual number of phones, i.e. $H[r] > H[u]$, the mutual information between the reference transcription and the data-driven transcription will be lower and the NMI will again be less than 1.

Inference	Database	Recall (%)	Precision (%)	F-score (%)	NMI (%)
CRP	TIMIT	74.98	56.43	**64.40**	33.87
VB	TIMIT	68.47	58.36	63.01	**34.81**
CRP	MBOSHI	68.91	38.26	**49.20**	34.41
VB	MBOSHI	55.82	40.43	46.89	**35.98**

Table 2.2: Comparison between the Chinese Restaurant Process (CRP) and the Variational Bayes (VB) inference.

2.5 Results and analysis

2.5.1 Settings

We describe here the configuration of our model used for all our experiments. As explained previously, the base measure is a combination of Dirichlet and Normal-Wishart distributions. The Dirichlet distributions were initialized with all concentration parameters set to 1. The parameters of the Normal-Wishart prior were set as follows:

$$\mathbf{m}_0 = \boldsymbol{\mu} \tag{2.109}$$

$$\beta_0 = 1 \tag{2.110}$$

$$\mathbf{W}_0 = \mathbf{I} \tag{2.111}$$

$$\nu_0 = D + 1 \tag{2.112}$$

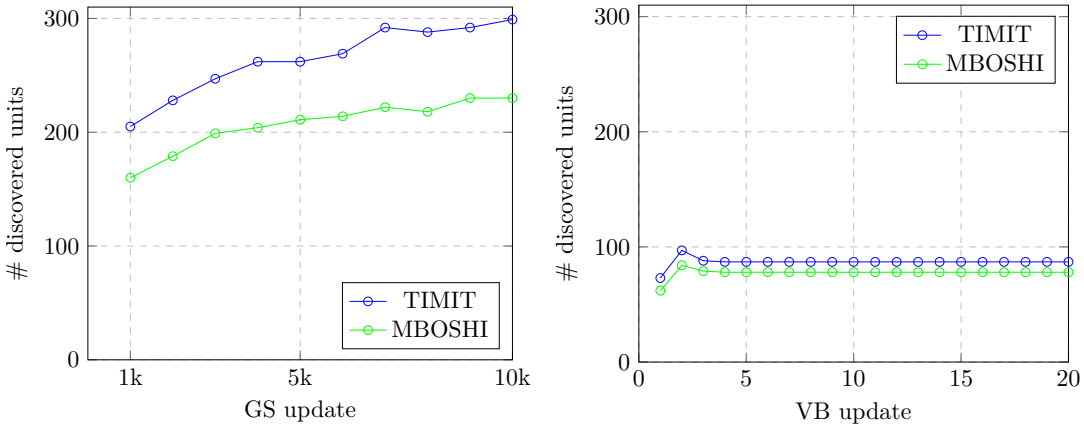
where D is the dimension of the feature vectors and $\boldsymbol{\mu} = \frac{1}{N}\sum_n \mathbf{x}_n$ is sample mean of the whole data set. The Gamma prior over the concentration of the Dirichlet Process was set to have a mean equal to half of the truncation parameter τ. We chose this particular parameterization to encourage the model to use more units at the beginning of the training and let the model prune the number of acoustic units by decreasing the concentration parameter later on. The truncation parameter τ was set to 101. Among these 101 potential units, we reserved one to be the "silence unit". The HMM of the silence unit was configured to have 5 emitting states instead of 3 emitting states for the other units and, furthermore, we constrained the inference graph of the phone-loop to start and end an utterance by this silence unit. Finally, each model was trained for 30 epochs, that is 30 VB E-steps and VB M-steps.

2.5.2 Variational Bayes vs Gibbs Sampling

As a first step, we compare both versions of the AUD phone loop model: the one which uses the Chinese restaurant process as inference scheme (Lee and Glass, 2012) (denoted CRP in further references) and our model which uses the stick-breaking construction and Variational Bayes (VB) inference. The results for the CRP model were obtained by using the publicly available implementation[5]. Results, obtained on the MFCCs features, are shown in Table 2.2. We observe that the results are not so different from each other which is to be expected since the models are almost identical[6] despite using different inference schemes. On one hand, our model does not segment the speech as well as the CRP model

[5]https://github.com/jacquelineCelia/dphmm_silence

[6]The model introduced in (Lee and Glass, 2012) has an extra set of binary variables indicating, for each frame, if it is the beginning of a new acoustic unit.



(a) Chinese Restaurant Process

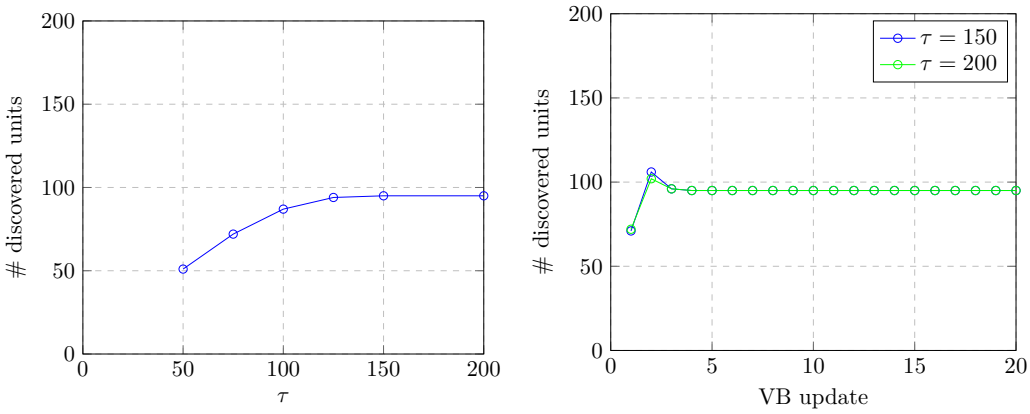
(b) Variational Bayes training

Figure 2.6: Evolution of the number of discovered units during the inference.

but, on the other hand, it achieves a better clustering. However, the biggest different lies in the efficiency of the inference. With our model, the computation of the VB E-step can be parallelized which leads to a very fast training: with 30 parallel cores, the training was finished in roughly 1.5 hours. With the CRP model, such parallelization is not possible, which makes the inference quite slow. In this example, the 10000th update of the Gibbs sampler was reached after 2 days of training.

Another important difference between the two models is the number of discovered acoustic units, i.e. the number of unique labels in the final transcription. Whereas the CRP model tends to use a large number of units, our model is much more parsimonious. Fig. 2.6 shows the evolution of the number of discovered units during the inference for both models. It is not clear why the two algorithms lead to such different number of units. As shown in Fig. 2.7, the truncation parameter does not seem to be a limit as the maximum number of units is never reached. A possible explanation may be with the nature of the mean-field approximation: indeed, it is well known that the mean-field approximation tends to underestimate the variance of the true posterior (Minka et al., 2005) which, in our case, could lead to a solution with less acoustic units. Nevertheless, this feature is advantageous, as, without reducing the quality of the clustering measured in term of NMI, it provides a solution with less parameters.

In Fig. 2.8, we show an example of the output from both AUD systems on one randomly picked utterance. One can see that both models over-segment the speech, especially at the beginning of the utterance, before the actual speech starts. This is a caveat of generative models as they may be sensitive to outliers.



(a) Number of discovered units as a function of the truncation parameter.

(b) Evolution of the number of discovered units during the training.

Figure 2.7: Effect of the truncation parameter on the number of discovered units evaluated on the TIMIT data set.

Features	Corpus	Recall	Precision	F-score	NMI (%)
MFCC	TIMIT	68.47	58.36	**63.01**	34.81
MBN	TIMIT	60.86	55.53	58.07	**37.17**
MFCC	MBOSHI	55.82	40.43	**46.89**	**35.98**
MBN	MBOSHI	55.14	36.73	44.09	32.13

Table 2.3: Comparison between MFCC and MBN features for acoustic unit discovery.

2.5.3 Variational Bayes objective for AUD

Variational Bayes objective is specific for each choice oflikelihood and prior: changing either of those will affect the objective function. Nonetheless,the phone-loop model seems to be a reasonable choice as the Variational Bayes inferenceleads to learning (part of) the phonetic informationOrganizing committeeOrganizing committeeVariational Bayes inference, contrary to Gibbs sampling, optimizes a well defined objective function. This objective function maximizes the expected log-likelihood regularized by a penalty term which forces the posterior distribution to be close to the prior. Fig. 2.9 shows the evolution of the metrics during the VB inference. Interestingly, even though the training is fully unsupervised and does not use the reference transcription, it indirectly optimizes our metrics. This observation is important as it shows that the AUD problem is amenable to an optimization problem. Note that Variational Bayes objective is specific for each choice of likelihood and prior: changing either of those will affect the objective function. Nonetheless, the phone-loop model seems to be a reasonable choice as the Variational Bayes inference leads to learning (part of) the phonetic information.

2.5.4 Discriminative features

The input features to the AUD model are of crucial importance. Indeed, since the AUD is trained to fit the data, if the features carry non-phonetic information, then, the model will use some of its modeling capacity (in our case create more units) to model it. Mainstream ASR, has coped with this issue in several ways:

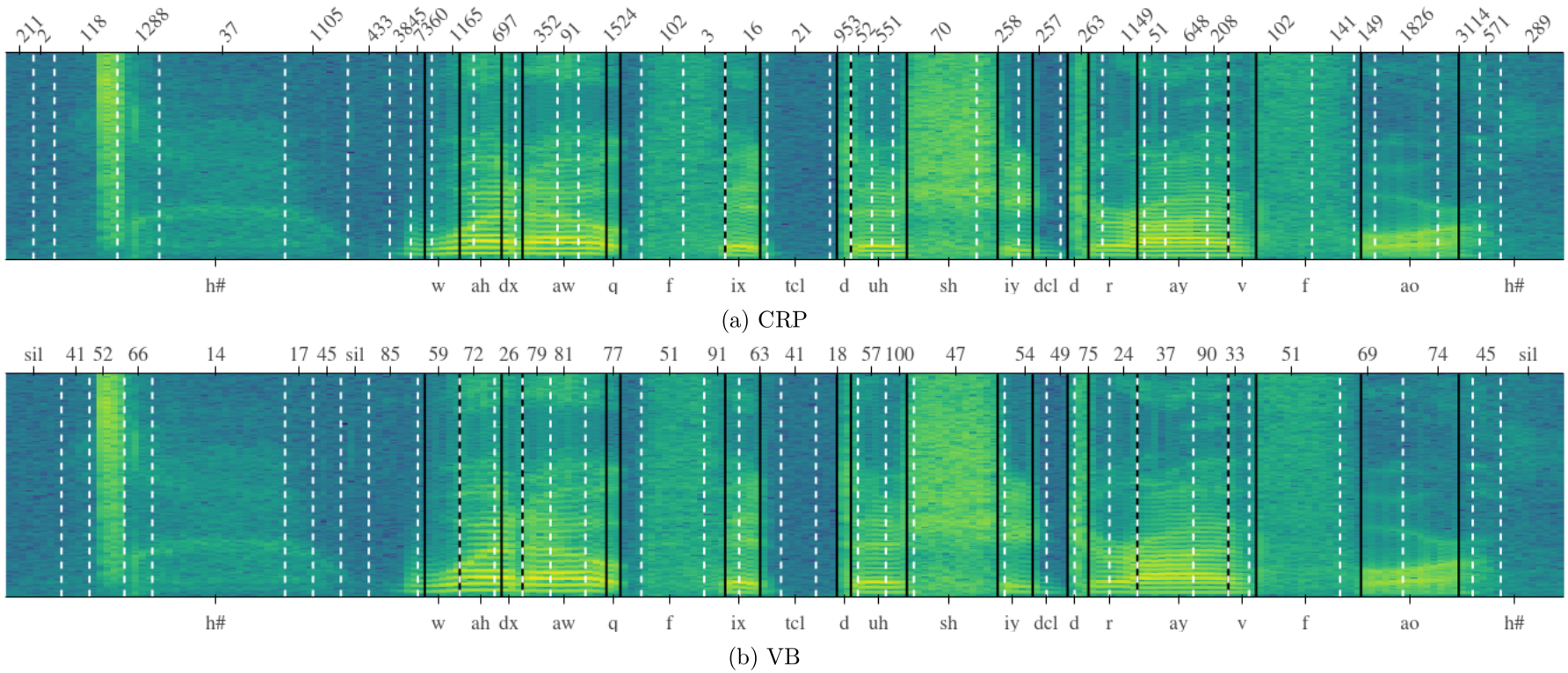


Figure 2.8: Example of segmentation of utterance "What outfit does she drive for?" by the model trained with the Chinese Restaurant Process (CRP) and Variational Bayes (VB). Black lines represent the reference boundaries and white dashed lines are the boundaries of the AUD model. The bottom and top sequences of labels are the time aligned reference and proposed transcription respectively. "sil" corresponds to the silence unit of the model trained with Variational Bayes.

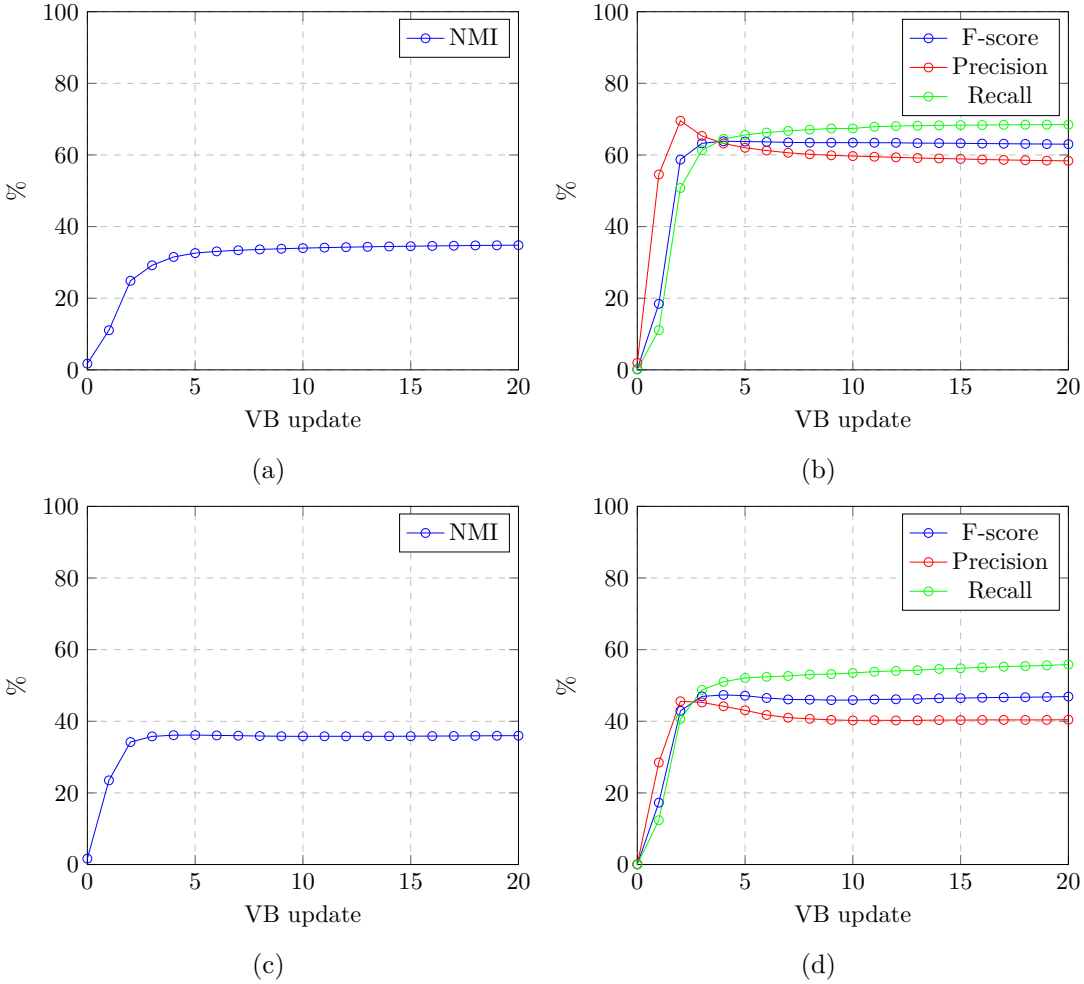


Figure 2.9: Evolution of the clustering and segmentation metrics during the training for the TIMIT database (Fig. 2.9a and Fig. 2.9b) and the MBOSHI database (Fig. 2.9c and Fig. 2.9d).

- using smooth spectral representations such as MFCC, PLP,... (Davis and Mermelstein, 1980; Hermansky, 1990) to remove unwanted variability while preserving the phonetic information. Our choice to utilize the MFCC features follows the same rationale, unfortunately, these representations are far from removing all unnecessary information

- using speaker normalization techniques, that transform the features and/or the model for each speaker (Wegmann et al., 1996; Gouvêa, 1998; Leggetter and Woodland, 1995). Model-based approaches, while effective, assume the identity of the speaker of each utterance to be known. However, this assumption is not always met, especially when dealing with low-resource languages

- using discriminative training to drive the model to ignore information not relevant to the task. Unfortunately, discriminative training is not applicable to our problem since we do not have the labels and try to discover them.

To improve our AUD model without requiring extra annotations, we replace traditional MFCC features by the Multi-Lingual Bottleneck (MBN) one. These features, being trained in a discriminative fashion on several languages, act as a features-based speaker normalization. The results for both features with the stick-breaking process based AUD model are shown in Table 2.3. As one can see, the effect of the MBN features is mitigated. First, we observe that these features are not performing well regarding the segmentation. This is to be expected as discriminative models for speech are known to be inaccurate in the timing of their response (Graves et al., 2006). On the other hand, regarding the clustering quality, the MBN features perform significantly better on TIMIT and much worse on MBOSHI. This illustrates the fact that the MBN features, despite being trained on multiple languages, cannot be considered as robust universal features. In some cases, when the target data is somewhat close to the training data of the MBN extractor, the MBN features provide a good phonetic representation of speech and will help the AUD model. However, when the target data is too different from the training data, the MBN features may provide a poor representation of the speech signal. Unfortunately, it is difficult to know before hand if a particular language will benefit of not from the MBN features, and therefore, the choice of features for the AUD task remains an open problem which depends on the data.

2.5.5 Non-Parametric vs Parametric Phone-Loop

We have defined the phone-loop model using a non-parametric prior leading to an infinite mixture of HMMs. It is possible to define a „parametric version“ of this model by replacing the Stick-Breaking Process prior with a Dirichlet distribution. This is easily done by replacing $p(u_l|\mathbf{v})p(\mathbf{v})$ in (2.33) by:

$$p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\varrho}_0) \tag{2.113}$$

$$p(u_l|\boldsymbol{\pi}) = \mathcal{C}(\boldsymbol{\pi}), \tag{2.114}$$

where $\mathcal{D}$ and $\mathcal{C}$ are the Dirichlet and Categorical distribution respectively. Assuming this new model, it is easy to show that optimal variational posterior $q(\boldsymbol{\pi})$ is given by:

$$q(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\varrho}) \tag{2.115}$$

$$\varrho_k = \varrho_{0,k} + \Big\langle \sum_{u_i \in \mathbf{u}} \mathbb{1}[u_i = k] \Big\rangle_{q(\mathbf{u})}. \tag{2.116}$$

On one hand, the parametric version forces a specific number of element in the mixture and does not let the model learns its complexity. However, the update equation are simpler as one doesn't need to deal with infinite prior/posterior and yet may remain a good approximation of the non-parametric version of the model. From table 2.4, we see that the choice of having a non-parametric model does have a significant positive effect on the clustering quality measured with the NMI. It also leads to a better segmentation on the TIMIT data whereas the segmentation quality is slightly worse on the MBOSHI data.

2.6 Conclusion

In this chapter, we have revisited the model proposed in (Lee and Glass, 2012) by using the Stick-Breaking construction of the Dirichlet process. Consequently, an approximation of the posterior distribution of the model's parameters can be derived using the Variational Bayes. This new algorithm for AUD achieves a better clustering, measured with the NMI, while being much faster and scalable to large database. The model has three main components:

Model	Features	Corpus	F-score	NMI (%)
non-parametric	MFCC	TIMIT	**63.01**	**34.81**
parametric	MFCC	TIMIT	57.03	32.56
non-parametric	MBN	TIMIT	**58.07**	**37.17**
parametric	MBN	TIMIT	54.46	35.52
non-parametric	MFCC	MBOSHI	46.89	**35.98**
parametric	MFCC	MBOSHI	**47.09**	35.5
non-parametric	MBN	MBOSHI	44.09	**32.13**
parametric	MBN	MBOSHI	**44.1**	29.73

Table 2.4: Comparison of the non-parametric and parametric prior for the AUD model.

1. the per-unit likelihood model, which, in our case, is an HMM
2. the stick-breaking process, which is a prior over unigram phonotactic language model
3. the base measure which is a prior over the sounds likely to be chosen as acoustic unit.

A first difficulty is how to define a consistent base measure. Indeed, choosing the right distribution is a non-trivial matter as the support of the base measure is defined over a hardly interpretable high-dimensional space. So far, we have bypassed this problem by using a vague prior which, roughly, allows any sound to be a candidate acoustic unit. While mathematically convenient, this solution is highly unsatisfactory as restricting support of the base measure to a small set of sounds would greatly reduce the searched space and therefore help the algorithm to find better units. This problem will be addressed in chapter 3, where we used *Generalized Subspace Model* to learn a low-dimensional representation of sounds from several languages to help the AUD task.

A second weakness is the assumption of the unigram phonotactic language model. As the n-gram and other sophisticated language models have proven to be essential to achieve accurate ASR, it is reasonable to believe that a more refined language model should be also beneficial for the AUD task. In chapter 4, we extend the non-parametric phone-loop model to incorporate a bigram phonotactic language model using Hierarchical Dirichlet Process.

Finally, the AUD model can also be improved by replacing the HMM by a more refined acoustic model. While we do not explore any other acoustic unit model in this work, an enhanced version of the non-parametric phone-loop based on Variational Auto-Encoder was proposed in (Ebbers et al., 2017; Glarner et al., 2018).

Chapter 3

Generalized Subspace Model for Sound Representation

In chapter 2, we have described a non-parametric phone-loop model to discover acoustic units from speech. This model represents each acoustic unit as a vector of parameters of an HMM. This approach suffers from the fact that the HMM parameter space is high-dimensional—more than a thousand dimensions for common settings—whereas the set of possible acoustic units for a given language is confined to a "small" region of this space. Therefore, a natural question is how we can reformulate our AUD model such that the search space of the acoustic units is restrained to the subset of likely acoustic unit candidates. In this chapter, we develop the theory and the tools to address this problem in a principled way. In section 3.1, we introduce the concept of *Generalized Subspace Model* (GSM): a theoretical framework to embed probabilistic models in arbitrary vector space. Equipped with this new concept, we build in section 3.2 the *Subspace Hidden Markov Model* (SHMM) to represent phones in a low-dimensional space. Finally, in section 3.3, we integrate the SHMM into the non-parametric phone-loop model for acoustic unit discovery. Our integration is done in two steps: first, we use the SHMM to learn the subspace of phone embeddings from several languages. Loosely speaking, the model is learning *what is a phone*. In a second time the AUD system will cluster the speech signal as described in chapter 2 but restraining the search to acoustic unit embeddings living in the subspace of phone learned at the previous step.

3.1 Generalized Subspace Model

A large part of the machine learning field is dedicated to representation of high-dimensional data points using low-dimensional embeddings. The projection from high to low-dimensional space ideally removes unwanted variability and allows for easy manipulation of the data. Techniques to learn this mapping range from simple linear projections such as Principal Component Analysis or Linear Discriminant Analysis (Bishop, 2006) to complex non-linear functions such as t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten and Hinton, 2008). These techniques have also been generalized to build powerful density estimators (Tipping and Bishop, 1999; Prince and Elder, 2007; Ioffe, 2006; Kingma and Welling, 2013; Rezende and Mohamed, 2015). Yet, all these methods have in common that each data point has its own low-dimensional embedding, or put in another way, they project the data onto a low-dimensional manifold. In some cases, we would like the embeddings to represent not

the data itself but rather an ensemble of observations modeled by a density. For instance, one may want to have an embedding to represent a person identity whereas the observations are a set of images of this person. In another example, closer to our application, we would like to learn an embedding representing a phone from several utterances of this particular phone. In this setting, the task is not to learn a manifold in the data space directly, rather, each group of observations is represented by a probabilistic model and we aim to represent the set of models in a low-dimensional space. In speech, joint factor analysis (Kenny et al., 2007), i-vector (Dehak et al., 2009) and Subspace Gaussian Mixture Model (SGMM) (Povey et al., 2011) are typical examples of such model applied to speaker identification and ASR respectively.

Learning a subspace of probabilistic models is, however, quite complex. For instance, an i-vector model only deals with the mean parameters of the mixture components of a GMM to keep a closed form solution of the update equations. On the other hand, the SGMM incorporates the mixture's weights in the subspace but needs to introduce some approximation for the training. Furthermore, subspace models trained in the maximum likelihood fashion are prone to overfit which can significantly hamper the quality of the embeddings. In the following of this section, we introduce the *Generalized Subspace Model* (GSM) which:

- unifies traditional subspace models into a single framework
- is robust against overfitting by having a prior over the susbpace's parameters.

Finally, we describe a stochastic Variational Bayes training which can be applied to any possible subspace model.

3.1.1 Definition

Let's have K sets of observations $\mathbf{X}_1, \ldots, \mathbf{X}_K$ where the ith set has N_i observations: $\mathbf{X}_i = \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}$. Each set is associated to a class (e.g. phone) and has a specific distribution parameterized by vector $\mathbf{h}_i$. We assume that the likelihood of a set of observations is given by a member of the exponential family of distributions (appendix B), eventually conditioned by some latent variable:

$$p(\mathbf{X}_i|\mathbf{Z}_i, \boldsymbol{\eta}_i) = \exp\{\boldsymbol{\eta}_i^\top T(\mathbf{X}_i, \mathbf{Z}_i) - A(\boldsymbol{\eta}_i, \mathbf{Z}_i) + B(\mathbf{X}_i, \mathbf{Z}_i)\}, \tag{3.1}$$

where $\boldsymbol{\eta}_i \in \mathcal{H}$ is the P-dimensional vector of natural parameters of the ith model, $\mathbf{Z}_i$ is a set of latent variables specific to the model[1] and the functions T, A and B are, respectively, the sufficient statistics, the log-normalizer and the base measure[2] specific to the likelihood model. Then, the generative process of the GSM is:

1. $\mathbf{W}, \mathbf{b} \sim p(\mathbf{W}, \mathbf{b})$
2. $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \forall i \in \{1, 2, \ldots, K\}$
3. $\boldsymbol{\eta}_i = f(\mathbf{W}^\top \mathbf{h}_i + \mathbf{b})$

[1] For some models, this set can be empty.

[2] For members of the exponential family, the base measure is the part of the normalization constant that does not depend on the natural parameters and should not be confused with the base measure of the Dirichlet Process.

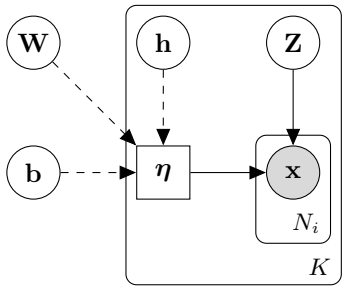


Figure 3.1: Graphical model of the Generalized Subspace Model. Dashed edges pointing to a square node represent a deterministic relation.

4. $\mathbf{Z}_i \sim p(\mathbf{Z})$
5. $\mathbf{X}_i \sim p(\mathbf{X}|\mathbf{Z}_i, \mathbf{W}, \mathbf{b}, \mathbf{h}_i)$,

where:

- $\mathbf{W} \in \mathbb{R}^{P \times D}$ and $\mathbf{b} \in \mathbb{R}^P$ are the subspace parameters
- $\mathbf{h}_i \in \mathbb{R}^D$ is the embedding vector of a model
- $f : \mathbb{R}^P \to \mathcal{H}$ is a differentiable function mapping a real vector into the natural parameter space of the likelihood model.

Note that the set of natural parameters does not necessarily lie in $\mathbb{R}^P$. For instance, the set of natural parameters for the Normal distribution, which is defined by all possible pairs of real vector and positive definite matrix, is only a subset of $\mathbb{R}^P$. The graphical model describing the generative process is shown in Fig 3.1.

3.1.2 Relation with the i-vector model

As its name indicates, GSM generalizes existing subspace models and casts them into a single framework. To illustrate the connection between GSM and other subspace models we show how the i-vector model can be seen as a special instance of a GSM. Let be $\mathbf{X}_i = \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iN_i}$ where $\mathbf{x}_{ij} \in \mathbb{R}^P$ is the jth P-dimensional feature vector of the ith utterance of a speech corpus. The likelihood of the utterance conditioned on the i-vector $\mathbf{h}_i$ is given by:

$$p(\mathbf{X}_i|\mathbf{h}_i) = \prod_{j=1}^{N_i} \Big[\sum_{c=1}^{C} \mathcal{N}(\mathbf{x}_{ij}|\mathbf{W}_c^\top \mathbf{h}_i + \mathbf{b}_c, \mathbf{\Sigma}_c) p(z_{ij} = c) \Big] \tag{3.2}$$

$$p(z_{ij} = c) = \mathcal{C}(z_{ij} = c|\boldsymbol{\pi}) = \boldsymbol{\pi}_c, \tag{3.3}$$

where z_n is a latent variable indicating which mixture's component is assigned to the nth feature vector and C is the number of components in the mixture. From the prior over z_n and the likelihood, the joint distribution of $\mathbf{X}_i$ and z_i is given by:

$$p(\mathbf{X}_i, \mathbf{z}_i|\mathbf{h}_i) = \prod_{j=1}^{N_i} \prod_{c=1}^{C} \Big[\mathcal{N}(\mathbf{x}_{ij}|\mathbf{W}_c^\top \mathbf{h}_i + \mathbf{b}_c, \mathbf{\Sigma}_c)^{\mathbb{1}[[z_{ij}=1]} \pi_c^{\mathbb{1}[[z_{ij}=1]} \Big]. \tag{3.4}$$

For convenience, we use the following placeholder:

$$\boldsymbol{\mu}_{ic} = \mathbf{W}_c^\top \mathbf{h}_i + \mathbf{b}_c \tag{3.5}$$

$$\tag{3.6}$$

and express Eq. (3.4) as an exponential function:

$$\begin{aligned} p(\mathbf{X}_i, \mathbf{z}_i|\mathbf{h}_i) &= \exp\Bigg\{ \sum_{j=1}^{N_i} \sum_{c=1}^{C} \mathbb{1}[z_{ij} = c] \boldsymbol{\mu}_{ic}^\top \boldsymbol{\Sigma}_c^{-1} \mathbf{x}_{ij} - \mathbb{1}[z_{ij} = c] \Big(f \frac{1}{2} \mathbf{x}_{ij}^\top \boldsymbol{\Sigma}_c^{-1} \mathbf{x}_{ij} \\ &\quad + \frac{1}{2} \boldsymbol{\mu}_{ic}^\top \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_{ic} + \frac{P}{2} \ln 2\pi + \ln |\boldsymbol{\Sigma}_c| \Big) + \mathbb{1}[z_{ij} = c] \ln \pi_c \Bigg\} \\ &= \exp\{\boldsymbol{\eta}_i^\top T(\mathbf{X}_i, \mathbf{z}_i) - A(\boldsymbol{\eta}_i, \mathbf{z}_i) + B(\mathbf{X}_i, \mathbf{z}_i)\} \end{aligned} \tag{3.7, 3.8}$$

where we have defined:

$$\boldsymbol{\eta}_i = \begin{bmatrix} \boldsymbol{\mu}_{i1} \\ \vdots \\ \boldsymbol{\mu}_{iC} \end{bmatrix} \tag{3.9}$$

$$T(\mathbf{X}_i, \mathbf{z}_i) = \sum_{j=1}^{N_i} \begin{bmatrix} \boldsymbol{\Sigma}_1^{-1} \mathbf{x}_{ij} \, \mathbb{1}[z_{ij} = 1] \\ \vdots \\ \boldsymbol{\Sigma}_C^{-1} \mathbf{x}_{ij} \, \mathbb{1}[z_{ij} = C] \end{bmatrix} \tag{3.10}$$

$$A(\boldsymbol{\eta}_i, \mathbf{z}_i) = \sum_{j=1}^{N_i} \sum_{c=1}^{C} \mathbb{1}[z_{ij} = c] \big(\frac{1}{2} \boldsymbol{\mu}_{ic}^\top \boldsymbol{\Sigma}_c^{-1} \boldsymbol{\mu}_{ic} + \frac{P}{2} \ln 2\pi + \ln |\boldsymbol{\Sigma}_c| \big) \tag{3.11}$$

$$B(\mathbf{X}_i, \mathbf{z}_i) = \sum_{j=1}^{N_i} \sum_{c=1}^{C} \mathbb{1}[z_{ij} = c] \frac{1}{2} \mathbf{x}_{ij}^\top \boldsymbol{\Sigma}_c^{-1} \mathbf{x}_{ij} + \mathbb{1}[z_{ij} = c] \ln \pi_c. \tag{3.12}$$

From (3.8) and (3.9), we see that the i-vector model is a special instance of the GSM where f is the identity function.

3.1.3 Inference

We now present a generic training algorithm of the GSM which is applicable for a wide class of models. As the exact posterior of the GSM's parameters is not tractable, we use one more time the Variational Bayes objective (appendix A.1):

$$\mathcal{L}[q] = \left[\sum_{i=1}^{K} \langle \ln \frac{p(\mathbf{X}_i, \mathbf{z}_i|\boldsymbol{\Theta})}{q(\mathbf{z}_i)} \rangle_{q(\mathbf{z}_i)q(\boldsymbol{\Theta})} \right] - \mathrm{D_{KL}}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \tag{3.13}$$

$$\boldsymbol{\Theta} = \{\mathbf{W}, \mathbf{b}, \mathbf{h}_1, \ldots, \mathbf{h}_K\} \tag{3.14}$$

where we have grouped the parameters of the GSM into variable $\boldsymbol{\Theta}$ and we have assumed the following parametric mean-field factorization (appendices A.2.1 and A.2.2) of the variational posterior:

$$q(\mathbf{z}_1, \ldots, \mathbf{z}_K, \boldsymbol{\Theta}) = q(\boldsymbol{\Theta}; \mathbf{m}, \boldsymbol{\lambda}) \prod_{i=1}^{K} q(\mathbf{z}_i; \boldsymbol{\phi}_i) \tag{3.15}$$

$$q(\boldsymbol{\Theta}; \mathbf{m}, \boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\Theta}|\mathbf{m}, \mathrm{diag}(\exp\{\boldsymbol{\lambda}\})), \tag{3.16}$$

where $\mathbf{m}$ and $\boldsymbol{\lambda}$ are, respectively, the mean and logarithm of the diagonal of the covariance matrix of the Normal variational posterior. The type of $q(\mathbf{z}_i; \boldsymbol{\phi}_i)$ will depend on the nature of the model. For instance, for the i-vector model, this will be the posterior distribution of the frame-to-component assignment variable. In the general case, we do not consider any particular distribution and we merely assume that the optimal parameter of variational posterior $\boldsymbol{\phi}_i^*$ can be estimated in closed form solution. For conciseness, we will write $q(\mathbf{z}_i)$ and $q(\boldsymbol{\Theta})$ instead of $q(\mathbf{z}_i; \boldsymbol{\phi}_i)$ and $q(\boldsymbol{\Theta}; , \mathbf{m}, \boldsymbol{\lambda})$.

With the factorization assumed in (3.15), the inference becomes an EM-like algorithm where we re-estimate the optimal parameters of the variational posterior $\boldsymbol{\phi}_1^*, \dots, \boldsymbol{\phi}_K^*$ and $\mathbf{m}^*, \boldsymbol{\lambda}^*$ alternately using the following partial objective functions:

$$\boldsymbol{\phi}_i^* = \arg\max_{\boldsymbol{\phi}_i} \langle \ln \frac{p(\mathbf{X}_i, \mathbf{z}_i|\boldsymbol{\Theta})}{q(\mathbf{z}_i)} \rangle_{q(\mathbf{z}_i)q(\boldsymbol{\Theta})} \tag{3.17}$$

$$= \arg\max_{\boldsymbol{\phi}_i} \mathcal{L}_\phi(\boldsymbol{\phi}_i; \mathbf{X}_i, \mathbf{m}, \boldsymbol{\lambda}) \tag{3.18}$$

$$\mathbf{m}^*, \boldsymbol{\lambda}^* = \arg\max_{\mathbf{m}, \boldsymbol{\lambda}} \left[\sum_{i=1}^{K} \langle \ln \frac{p(\mathbf{X}_i, \mathbf{z}_i|\boldsymbol{\Theta})}{q(\mathbf{z}_i)} \rangle_{q(\mathbf{z}_i)q(\boldsymbol{\Theta})} \right] - \mathrm{D_{KL}}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \tag{3.19}$$

$$= \arg\max_{\mathbf{m}, \boldsymbol{\lambda}} \mathcal{L}_{m,\lambda}(\mathbf{m}, \boldsymbol{\lambda}; \mathbf{X}_1, \dots, \mathbf{X}_K, \boldsymbol{\phi}_1, \dots, \boldsymbol{\phi}_K). \tag{3.20}$$

(3.19) has no closed form solution but can be optimized through a stochastic gradient ascent using the „re-parameterization trick“ (Kingma and Welling, 2013):

$$\mathcal{L}_{m,\lambda}(\mathbf{m}, \boldsymbol{\lambda}; \dots) \approx \frac{1}{L} \left[\sum_{l=1}^{L} \sum_{i=1}^{K} \langle \ln \frac{p(\mathbf{X}_i, \mathbf{z}_i|\boldsymbol{\Theta}_l)}{q(\mathbf{z}_i)} \rangle_{q(\mathbf{z}_i)} \right] - \mathrm{D_{KL}}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \tag{3.21}$$

$$\boldsymbol{\Theta}_l = \mathbf{m} + \exp\{\frac{1}{2}\boldsymbol{\lambda}\} \odot \boldsymbol{\epsilon} \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3.22}$$

where $\odot$ is the element-wise multiplication. The complete training algorithm is given in Alg. 3.1. Note that, for simplicity, Alg. 3.1 is presented with a fixed learning rate. In practice, the learning rate changes over time using some adaptive procedure (Duchi et al., 2011; Kingma and Ba, 2014).

3.1.4 Example

To finish our presentation of the GSM, we revisit the Subspace Gaussian Mixture Model (SGMM) on a toy example. The SGMM was originally presented in (Povey et al., 2011) as a mean to improve the acoustic model in an ASR pipeline. We use the GSM framework to:

- give a Bayesian treatment of the model
- include all the parameters of the Gaussian[3] while maintaining a tractable inference thanks to Algorithm 3.1.

[3] In the original version of the SGMM, only the mean vectors and the mixing weights of the mixture's components were included in the subspace.

Algorithm 3.1 Training of the Generalized Subspace Model

function MSTEP($\mathbf{X}_{1:K}, \boldsymbol{\phi}_{1:K}, \mathbf{m}, \boldsymbol{\lambda}, \rho, S$)
  ▷ ρ: learning rate of the stochastic gradient ascent
  ▷ S: number of updates of the stochastic gradient ascent
  ▷ $\boldsymbol{\phi}_{1:K}, \mathbf{m}, \boldsymbol{\lambda}$: current parameters of the variational posteriors
  $\mathbf{m}^{(new)} \leftarrow \mathbf{m}$
  $\boldsymbol{\lambda}^{(new)} \leftarrow \boldsymbol{\lambda}$
  for $s \leftarrow 1$ to S **do**
    $\mathbf{m}^{(new)} \leftarrow \mathbf{m}^{(new)} + \rho \nabla_{\mathbf{m}}^{(new)} \mathcal{L}_{m,\lambda}(\mathbf{m}, \boldsymbol{\lambda}^{(new)}; \mathbf{X}_{1:K}, \boldsymbol{\phi}_{1:K})$
    $\boldsymbol{\lambda}^{(new)} \leftarrow \boldsymbol{\lambda}^{(new)} + \rho \nabla_{\boldsymbol{\lambda}} \mathcal{L}_{m,\lambda}(\mathbf{m}^{(new)}, \boldsymbol{\lambda}^{(new)}; \mathbf{X}_{1:K}, \boldsymbol{\phi}_{1:K})$
  return $\mathbf{m}^{(new)}, \boldsymbol{\lambda}^{(new)}$

function ESTEP($\mathbf{X}_{1:K}, \mathbf{m}, \boldsymbol{\lambda}$)
  ▷ The E-step is model dependent but is identical to the E-step of the unconstrained model (i.e. no subspace).
  for $i \leftarrow 1$ to K **do**
    $\boldsymbol{\phi}_i^{(new)} \leftarrow \arg\max_{\boldsymbol{\phi}_i} \mathcal{L}_{\phi}(\boldsymbol{\phi}_i; \mathbf{X}_i, \mathbf{m}, \boldsymbol{\lambda})$
  return $\boldsymbol{\phi}_{1:K}^{(new)}$

procedure TRAIN_GSM($\mathbf{X}_{1:K}, E, \rho, S$)
  ▷ E: number of epochs (i.e. E-step + M-step)
  $\mathbf{m}^* \leftarrow$ initialization
  $\boldsymbol{\lambda}^* \leftarrow$ initialization
  $\boldsymbol{\phi}_{1:K}^* \leftarrow$ initialization
  for $e \leftarrow 1$ to E **do**
    $\boldsymbol{\phi}_{1:K}^* \leftarrow$ ESTEP($\mathbf{X}_{1:K}, \mathbf{m}^*, \boldsymbol{\lambda}^*$)
    $\mathbf{m}^*, \boldsymbol{\lambda}^* \leftarrow$ MSTEP($\mathbf{X}_{1:K}, \boldsymbol{\phi}_{1:K}^*, \mathbf{m}^*, \boldsymbol{\lambda}^*, \rho, S$)

Let's consider the dataset shown in Fig. 3.2: each point represents task-dependent features and the color represents the class each point belongs to. For instance, the features could be the per-frame MFCC features and the class is the identity of the speaker or, alternately, each point could represent an image of a person and the class is the identity of this person. Since we are concerned with modeling phones, let's assume that each point represents the features of a speech frame and the color indicates the phone associated to this feature vector. Here, our goal is twofold: first, we wish to model the data using some probabilistic model, second, we would like to learn a low-dimensional representation of a phone, i.e. some kind of „phone embedding".

Let K denote the number of phones in our data set. As previously, our data set of N vectors is composed of K sets of sizes $N_1, \ldots, N_K$. We assume the N_i observations of the ith phone to be modeled by a mixture of $C = 2$ Normal densities, parameterized by:

- mixing weights: $\boldsymbol{\pi}_i = \begin{bmatrix} \pi_{i1} \\ \pi_{i2} \end{bmatrix}$ such that $\pi_{i1} + \pi_{i2} = 1$
- mean vectors $\boldsymbol{\mu}_{i1}$ and $\boldsymbol{\mu}_{i2}$

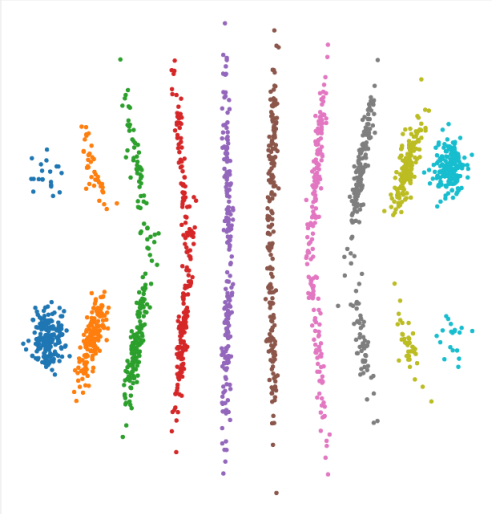


Figure 3.2: Synthetic data for the Subspace Gaussian Mixture Model (SGMM). An artificial representation of speech features where each phone (represented by one particular color) has a bimodal distribution.

- precision matrices $\mathbf{\Sigma}_{i1}$ and $\mathbf{\Sigma}_{i2}$

resulting in a total of 11 free parameters[4]. Let z_{ij} be the latent variable encoding, for the ith phone, to which mixture's component the jth speech frame is assigned. The joint distribution of the model is given by:

$$p(\mathbf{X}_i, \mathbf{z}_i | \dots) = \prod_{j=1}^{N_i} \prod_{c=1}^{C} \mathcal{N}(\mathbf{x}_{ij} | \boldsymbol{\mu}_{ic}, \mathbf{\Sigma}_{ic})^{\mathbb{1}[z_{ij}=c]} \pi_i^{\mathbb{1}[z_{ij}=c]}, \tag{3.23}$$

which can be expressed as an exponential function:

$$p(\mathbf{X}_i, \mathbf{z}_i | \dots) = \exp\{\boldsymbol{\eta}_i^\top T(\mathbf{X}_i, \mathbf{z}_i) - A(\boldsymbol{\eta}_i, \mathbf{z}_i) + B(\mathbf{X}_i, \mathbf{z}_i)\}, \tag{3.24}$$

[4]The mixing weights have 1 free parameter, each of the mean vectors has 2 and each of the precision matrices has 3, therefore: $2 \times (2+3) + 1 = 11$

where we have defined:

$$\boldsymbol{\eta}_i = \begin{bmatrix} \ln \frac{\pi_{i1}}{1-\sum_{c=1}^{C-1}\pi_{ic}} \\ \dots \\ \ln \frac{\pi_{iC}}{1-\sum_{c=1}^{C-1}\pi_{ic}} \\ \boldsymbol{\Sigma}_{i1}^{-1}\boldsymbol{\mu}_{i1} \\ \vdots \\ \boldsymbol{\Sigma}_{iC}^{-1}\boldsymbol{\mu}_{iC} \\ \mathrm{vec}(\boldsymbol{\Sigma}_{i1}^{-1}) \\ \vdots \\ \mathrm{vec}(\boldsymbol{\Sigma}_{iC}^{-1}) \end{bmatrix} \quad T(\mathbf{X}_i, \mathbf{z}_i) = \begin{bmatrix} \mathbb{1}[z_{ij}=1] \\ \dots \\ \mathbb{1}[z_{ij}=C-1] \\ \mathbb{1}[z_{ij}=1]\mathbf{x}_{ij} \\ \dots \\ \mathbb{1}[z_{ij}=C]\mathbf{x}_{ij} \\ -\mathbb{1}[z_{ij}=1]\frac{1}{2}\,\mathrm{vec}(\mathbf{x}_{ij}\mathbf{x}_{ij}^\top) \\ \dots \\ -\mathbb{1}[z_{ij}=C]\frac{1}{2}\,\mathrm{vec}(\mathbf{x}_{ij}\mathbf{x}_{ij}^\top) \end{bmatrix} \tag{3.25}$$

$$A(\boldsymbol{\eta}_i, \mathbf{z}_i) = -\ln(1-\sum_{c=1}^{C-1}\pi_{ic}) + \sum_{c=1}^{C}\mathbb{1}[z_{ij}=c]\left[\boldsymbol{\mu}_{ic}^\top\boldsymbol{\Sigma}_{ic}^{-1}\boldsymbol{\mu}_{ic} + \ln|\boldsymbol{\Sigma}_{ic}|\right] \tag{3.26}$$

$$B(\mathbf{X}_i, \mathbf{z}_i) = -\frac{N_i D}{2}\ln 2\pi. \tag{3.27}$$

Now, we define the prior over the natural parameters $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_K$ as follows:

$$\mathrm{vec}(\mathbf{W}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.28}$$

$$\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.29}$$

$$\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{3.30}$$

$$\boldsymbol{\eta}_i = f(\mathbf{W}^\top\mathbf{h}_i + \mathbf{b}). \tag{3.31}$$

The mapping function f is defined such that:

$$\pi_{ic} = \frac{\exp\{\mathbf{W}_\pi^\top\mathbf{h}_i + \mathbf{b}_\pi\}_c}{1+\sum_{l=1}^{C-1}\exp\{\mathbf{W}_\pi^\top\mathbf{h}_i + \mathbf{b}_\pi\}_l} \tag{3.32}$$

$$\boldsymbol{\Sigma}_{ic} = (\mathbf{L}_{ic}\mathbf{L}_{ic}^\top)^{-1} \tag{3.33}$$

$$\mathrm{diag}(\mathbf{L}_{ic}) = \exp\{\mathbf{W}_L^\top\mathbf{h}_i + \mathbf{b}_L\} \tag{3.34}$$

$$\mathrm{ltri}(\mathbf{L}_{ic}) = \mathbf{W}_{L'}^\top\mathbf{h}_i + \mathbf{b}_{L'} \tag{3.35}$$

$$\boldsymbol{\Sigma}_{ic}^{-1}\boldsymbol{\mu}_{ic} = \mathbf{W}_\mu^\top\mathbf{h}_i + \mathbf{b}_\mu, \tag{3.36}$$

where $\exp\{\dots\}$ is the element-wise exponential function, $\exp\{\dots\}_d$ is the dth dimension of the resulting vector and ltri is a function that returns the lower-triangular part (not including the diagonal) of a square matrix arranged as a vector. Matrices $\mathbf{W}_\pi$, $\mathbf{W}_L$, $\mathbf{W}_{L'}$ and $\mathbf{W}_\mu$ are disjoint parts of the matrix $\mathbf{W}$ ($\mathbf{b}_\pi$, $\mathbf{b}_L$, ... are defined similarly). Importantly, the parameters of the subspace $\mathbf{W}$ and $\mathbf{b}$ are shared across phones and only the embeddings $\mathbf{h}_1, \dots, \mathbf{h}_K$ are phone-specific. In our example we choose $\mathbf{h}_i$ to be a 2-dimensional vector, hence reducing the original 11 free parameters of the GMM to only 2 dimensions.

In the case of the SGMM, $z_{ij} \in \{1, \dots, C\}$ is a discrete variable and therefore the parameters $\boldsymbol{\phi}_i = \boldsymbol{\phi}_{i1}, \dots, \boldsymbol{\phi}_{iN_i}$, of the variational posteriors $q(\mathbf{z}_i) = q(\mathbf{z}_{i1}), \dots, q(\mathbf{z}_{iN_i})$ are simply:

$$q(z_{ij}=c) = \phi_{ij}^c. \tag{3.37}$$

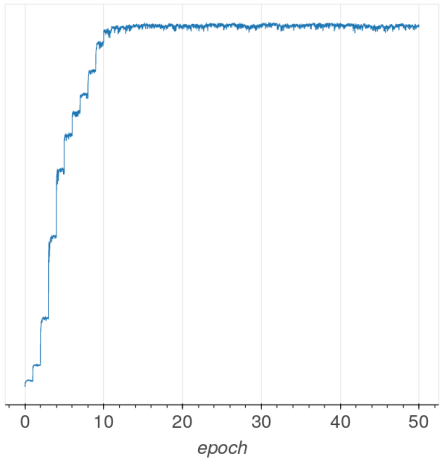


Figure 3.3: Evolution of the variational lower-bound during the training.

Because of the mean-field factorization assumed in (3.15), the variational posterior maximizing (3.18) is given by:

$$\boldsymbol{\phi}_i^* = q^*(\mathbf{z}_i) = \exp\{\langle \boldsymbol{\eta}_i \rangle_{q(\boldsymbol{\Theta})}^\top T(\mathbf{X}_i, \mathbf{z}_i) - \langle A(\boldsymbol{\eta}_i, \mathbf{z}_i) \rangle_{q(\boldsymbol{\Theta})} + B(\mathbf{X}_i, \mathbf{z}_i)\}. \tag{3.38}$$

Because of the non-linear mapping f, the expectations cannot be evaluated in closed form, we approximate them by sampling several values of $\boldsymbol{\eta}_i^l \sim q(\boldsymbol{\Theta})$ and taking the average. Once the optimal parameters of the variational posteriors over the latent variables $\mathbf{z}_i$ have been estimated, we can update the variational posteriors over the subspace's parameters $\mathbf{W}$, $\mathbf{b}$ and the embeddings $\mathbf{h}_1, \ldots, \mathbf{h}_K$. Plugging (3.24) in (3.21), we get the following objective function:

$$\mathcal{L}_{m,\lambda} = \frac{1}{L} \sum_{l=1}^{L} \sum_{i=1}^{K} \left[\boldsymbol{\eta}_i^{l\top} T(\mathbf{X}_i, \mathbf{z}_i) - A(\boldsymbol{\eta}_i^l, \mathbf{z}_i) + B(\mathbf{X}_i, \mathbf{z}_i) \right] - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Theta}) || p(\boldsymbol{\Theta})) \tag{3.39}$$

$$\mathbf{W}_l, \mathbf{b}_l, \mathbf{h}_1^l, \ldots, \mathbf{h}_K^l \sim q(\boldsymbol{\Theta}) \tag{3.40}$$

$$\boldsymbol{\eta}_i^l = f(\mathbf{W}_l^\top \mathbf{h}_i^l + \mathbf{b}_l). \tag{3.41}$$

The optimal parameters $\mathbf{m}^*, \boldsymbol{\lambda}^*$ of the variational posterior $q(\boldsymbol{\Theta})$ are obtained by optimizing (3.39) with a gradient ascent. The gradient of the objective function $\nabla \mathcal{L}_{m,\lambda}$ is easily obtained by any common automatic differentiation software.

As the VB objective is subject to local optima[5], it is important to properly initialize the model, i.e. to provide for an initial guess of the variational posterior's parameters $\boldsymbol{\phi}_i^*$ and $\mathbf{m}^*, \boldsymbol{\lambda}^*$. One may be tempted to train a GMM for each phone independently and then initialize the GSM so that it approximates the learned GMMs. This naive approach is, however, inadequate. Indeed, for mixture models, the ordering of the components is unidentifiable as reordering them will lead to the same exact density. From the standpoint of the parameter space, this model equivalence under reordering implies some kind of symmetry, that is, portions of the space that represent the same model but with different ordering. Therefore, when trained independently, the GMMs will be spread across these equivalent spaces making it hard to find a coherent initialization of the subspace. To avoid this issue, we initialized our SGMM with the following procedure:

[5]More precisely, the local optima are a consequence of our (parametric) mean-field approximation.

1. for each phone, fit the data with a single multivariate Normal density $\mathcal{N}(\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Lambda}}^{-1})$

2. initialize $\boldsymbol{\phi}_i^*$:

$$\boldsymbol{\phi}_i^* = \arg\max_{\boldsymbol{\phi}_i} \langle \ln \frac{p(\mathbf{X}_i, \mathbf{z}_i | \hat{\boldsymbol{\eta}}_i)}{q(\mathbf{z}_i)} \rangle_{q(\mathbf{z}_i)} \tag{3.42}$$

   where $\hat{\boldsymbol{\eta}}_i$ is the vector of natural parameters of the ith phone's GMM such that each component has mean and precision matrix set to $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ respectively. Note that this initialization corresponds to a saddle point of the objective function with respect to the parameters of the variational posteriors. Nevertheless, the noise introduced by the „re-parameterization trick“ in (3.21) will allow the model to escape from this saddle point.

3. set:

$$\mathbf{m}^* = \mathbf{0} \tag{3.43}$$

$$\boldsymbol{\lambda}^* = \frac{1}{D}\mathbf{1} \tag{3.44}$$

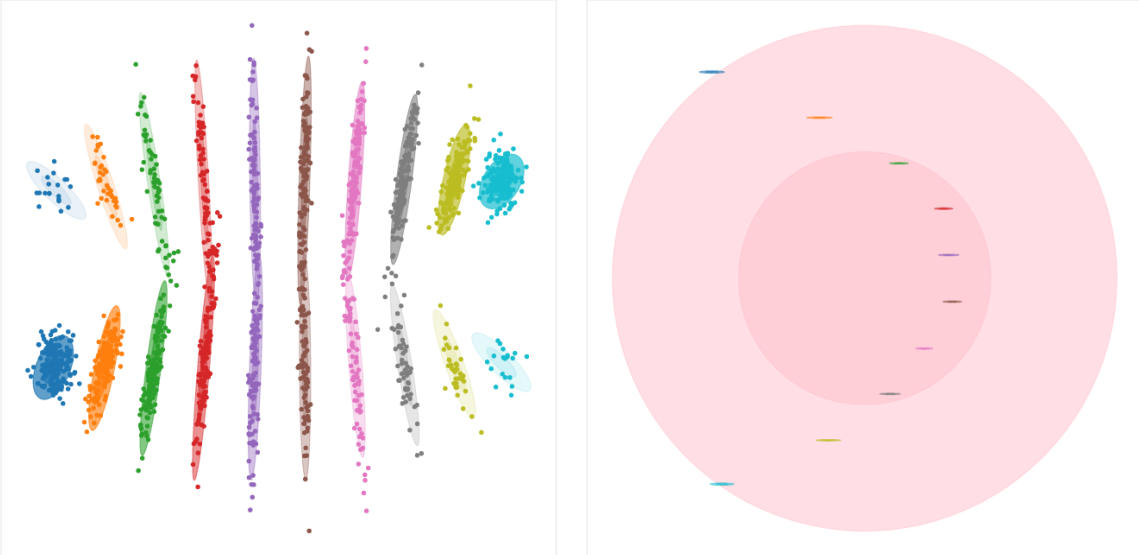
   and, using $q^*(\mathbf{z}_i)$, optimize the objective function described in Eq. (3.19) until convergence. The scaling $\frac{1}{D}$ reduces the noise at the beginning of the training when using relatively large subspace dimensions.

After this initialization procedure, the training can be carried out as described in Section 3.1.3. For this example, we trained the model for 50 epochs. For each VB M-step, we run a stochastic gradient ascent of 100 steps. The learning rate of the gradient ascent was updated following ADAM (Kingma and Ba, 2014) with an initial learning rate of 0.1. The evolution of the variational lower-bound over time is plotted in Fig. 3.3. At the early stage of the training, one can observe big jumps of the lower-bound. This corresponds to the VB E-step which drastically changes the accumulated statistics needed to retrain the subspace. As the training continues, the statistics stabilize and the training converges.

Fig. 3.4 shows the outcome of the training. We see that, each phone's data is properly fit by a 2-components GMM (Fig. 3.4a) whose parameters are constrained to live in a 2-dimensional space. Since we used (approximate) Bayesian inference, we do not learn point estimate of the parameters but a posterior distribution which encodes our uncertainty about the exact value of the parameters. The phones' variational posteriors in the parameter subspace are depicted in Fig. 3.4b. We see that the model has efficiently made use of both of the dimensions of the subspace to extract the phone embeddings. Notice that the posteriors are quite sharp as there is sufficient amount of data for each phone and, therefore, there is little uncertainty about the values of the GMMs' parameters. Finally, Fig. 3.5 shows how the subspace encodes the GMM parameters: the x axis controls mainly the covariance matrices of the GMM's components and the y axis encodes the mean vectors and the mixing weights.

3.2 Subspace Hidden Markov Model

In section 3.1, we have introduced the GSM: a theoretical framework to embed probabilistic models into a low-dimensional subspace. A major benefit of the GSM is that it allows

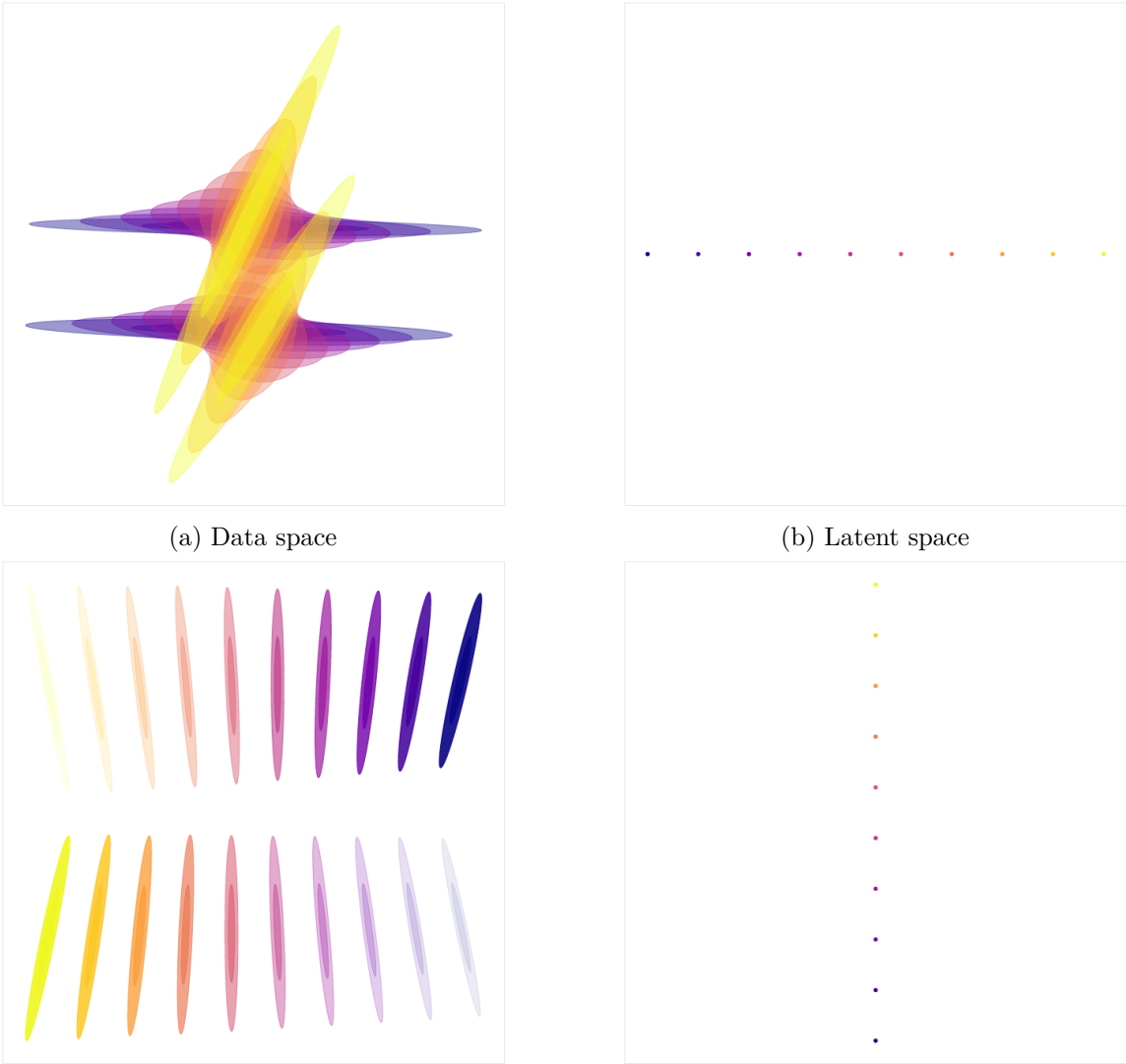


(a) GMM learned for each phone in the data space. The 11 free parameters of the GMM are encoded in the 2-dimensional latent space. The mixing weights are represented by the transparency of the components.

(b) Latent space of the GSM. The pink area shows the Normal prior density and, similarly, the small colored areas represent the posterior distribution over $\mathbf{h}_i$.

Figure 3.4: Outcome of fitting the Subspace Gaussian Mixture Model. Colors indicate a particular „phone“ class.

to build a subspace for a large class of models. For instance, it has been a common practice in ASR to model a phone with an HMM. Using the GSM, it is easy to build an embedding space for the HMM and, consequently, a *phonetic subspace*. We denote the combination of the HMM and the GSM the *Subspace Hidden Markov Model* or SHMM for short. The model closest to the SHMM is the already mentioned SGMM (Povey et al., 2011). Still, it is important to emphasize that, in addition to the technical differences highlighted in Section 3.1.4, both models serve different purposes. The SGMM was introduced to increase the number of Gaussian per HMM state while keeping the number of parameters to tune relatively low. Consequently, the SGMM was providing a more complex phone model compared to the traditional HMM. On the other hand, our SHMM does not increase the model complexity, rather, we use it for the sole purpose to extract a low-dimensional phonetic subspace, having, therefore, a practical representation for phone and acoustic units. To the best of our knowledge, the work closest to our SHMM is (Burget et al., 2010) where the authors used the SGMM to derived low-dimensional embeddings for the senones of an HMM-based ASR system. The SHMM generalizes (Burget et al., 2010) by (i) modeling the whole phone rather than „part-of-the-phone“ (e.g. the senone) (ii) including the covariance matrices as part of the subspace (iii) using Bayesian inference preventing potential overfitting. This section is made of three parts: first, we formally define the concept of *phonetic subspace* as used in this work (Section 3.2.1), second we practically define the SHMM (section 3.2.2) and finally, we demonstrate the potential of the SHMM on the TIMIT dataset to learn an English phonetic subspace in section 3.2.3.



(a) Data space

(b) Latent space

(c) Data space

(d) Latent space

Figure 3.5: How the subspace encodes the GMMs' parameters. The x axis controls mainly the covariance matrices and the y axis controls mainly the mean vectors and the mixing weights.

3.2.1 Phonetic subspace

Traditionally, in speech recognition, a phone is modeled by an HMM with 3 states with a left-to-right topology and each state has a GMM emission density. As seen previously in chapter 2, one can represent an HMM, therefore a phone/acoustic unit, in a vector space by concatenating the states' parameters in a „super vector“ $\boldsymbol{\eta}$. The concept of "super-vector" to represent probabilistic models in a vector space is directly borrowed from (Kenny et al., 2007). Let's consider that we fit an HMM to a set of recordings of the phone /aw/ resulting in the super-vector $\boldsymbol{\eta}_{\text{aw}}$. Moving the vector $\boldsymbol{\eta}_{\text{aw}}$ will change the parameters of the HMM and, consequently, the phone it represents. For instance, a displacement may lead to change the phone from /aw/ to /ow/. Then, moving the vector further will change the original /aw/ phone more profoundly and yield, say the consonant /z/. The key idea is that there

is a continuum between all phones, or expressed in another way, we can smoothly transition from one phone to another. Following this reasoning, we can envision all the phones of a language as vectors in a space, connected by a low-dimensional manifold which represents this continuum. This is depicted by the blue line in Fig. 3.6. This manifold is what we call the *phonetic subspace*. Importantly, this concept of phonetic subspace is independent of the choice of the phone model: GMM, HMM, Linear Dynamical Model... However, the type of model used will influence how well the continuity between phones is represented. We have chosen to use the HMM for convenience and to help to reuse this concept with the AUD task. Yet, it is likely that a more refined model, for instance the recurrent switching linear dynamical systems (Linderman et al., 2017), would lead to a more meaningful phonetic subspace.

When defining the phonetic subspace, we have only considered displacements of the vector $\boldsymbol{\eta}_{\text{aw}}$ leading to a change of phone (/aw/ to /ow/). However, moving the vector away from the phonetic subspace will not change the phone itself but its characteristics. For instance, we can move $\boldsymbol{\eta}_{\text{aw}}$ to make it more adapted to a female or a male speaker. Similarly to the phonetic subspace, by assuming a continuum between different speaker adapted phones, we can define the *speaker subspace*: a low-dimensional manifold, intersecting with the phonetic subspace, which represents the continuum of all possible speaker-adapted versions of a phone. This is shown by the red line if Fig. 3.6. Following the same reasoning, we can imagine a subspace for many other factors: emotion, channel, speaker age... The main advantage of subspace models is to extract from the high-dimensional parameter space of a probabilistic model a low-dimensional manifold capturing only the information relevant for a given task.

Finally, it is important to realize that the phonetic/speaker/... subspace is localized in the parameter space. For instance, in the phonetic subspace, moving the embedding of the phone /aw/ toward a certain direction will end up to a location where the embedding does not represent a phone anymore. More formally, the phonetic subspace is bounded within the parameter space. The GSM handles this localization with the bias parameter $\mathbf{b}$ and the prior over the embeddings $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. These two elements define a "bounded" region of the parameter space which concentrates most of the probability density. The bias vector $\mathbf{b}$ represents the phone centroid which is the average of all the phones of a language.

3.2.2 Encoding the HMM parameters

We have described the phonetic subspace as a manifold in the parameter space of a probabilistic model; in our case an HMM. We now make use of the GSM framework to define the SHMM which will allow us to estimate the phonetic subspace. Similarly, to the AUD model, each phone is modeled by a 3-state HMM with a left-to-right topology. Each state has a GMM emission with K Gaussian components. We limit ourselves to the case where the Gaussian components have a diagonal covariance matrix. The extension to full covariance matrix is straightforward using (3.35).

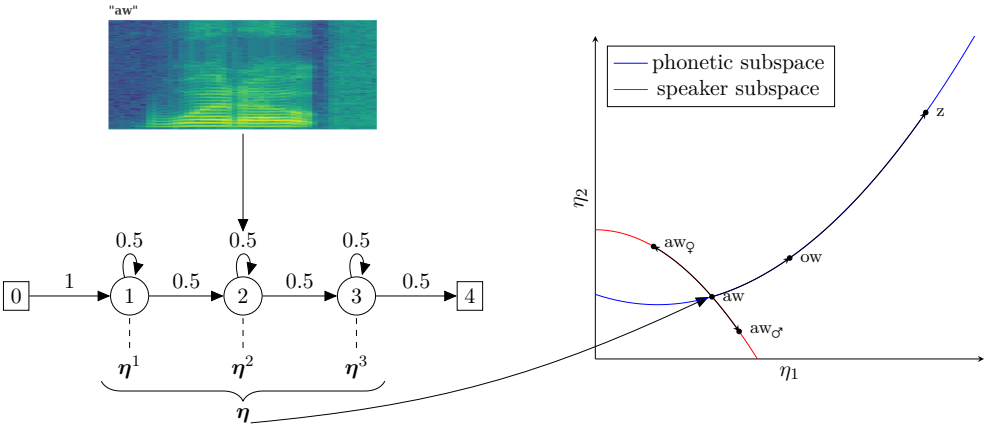


Figure 3.6: Representation of the Subspace Hidden Markov Model (SHMM). Each phone is represented a super-vector $\boldsymbol{\eta}$ which encodes the parameters of an HMM. The model assumes further that all the phones lie on a low-dimensional manifold (1-dimensional in this example) living in the total parameter space. The SHMM can account for multiple subspaces, for instance speaker, phonetic, emotion, ... In our case we are only interested in the phonetic subspace. The parameters space is represented with 2 dimensions for visualization. In practice, however, common settings lead to a parameter space with several thousands of dimensions.

Recall from chapter 2 that the likelihood of the nth speech frame given the parameters of an acoustic unit (or a phone in a supervised learning context) with index u is given by:

$$p(\mathbf{x}_n, c_n | s_n, \dots) = p(\mathbf{x}_n | \boldsymbol{\mu}_u^{s_n,c_n}, \boldsymbol{\Sigma}_u^{s_n,c_n}) p(c_n | \boldsymbol{\pi}_u^{s_n}) \tag{3.45}$$

$$p(c_n | \boldsymbol{\pi}_u^{s_n}) = p(c_n | \boldsymbol{\omega}_u^{s_n}) = \exp\{\boldsymbol{\omega}_u^{s_n\top} T(c_n) - A(\boldsymbol{\omega}_u^{s_n})\} \tag{3.46}$$

$$p(\mathbf{x}_n | \boldsymbol{\mu}_u^{s_n,c_n}, \boldsymbol{\Sigma}_u^{s_n,c_n}) = p(\mathbf{x}_n | \boldsymbol{\theta}_u^{s_n,c_n}) = \exp\{\boldsymbol{\theta}_{u_l}^{s_n,c_n\top} T(\mathbf{x}_n) - A(\boldsymbol{\theta}_u^{s_n,c_n})\}, \tag{3.47}$$

where c_n is the index of the mixture's component and s_n is the index of the HMM state. The natural parameters of the Categorical distribution $\boldsymbol{\omega}_u^{s_n}$, the natural parameters of the Normal distribution $\boldsymbol{\theta}_u^{s_n,c_n}$ and the sufficient statistics $T(c_n)$, $T(\mathbf{x}_n)$ are given by the following equations:

$$\boldsymbol{\omega}_u^{s_n} = \begin{bmatrix} \ln\left(\frac{\pi_{u,1}^{s_n}}{1-\sum_{k=1}^{C-1}\pi_{u,k}^{s_n}}\right) \\ \vdots \\ \ln\left(\frac{\pi_{u,C-1}^{s_n}}{1-\sum_{k=1}^{C-1}\pi_{u,k}^{s_n}}\right) \end{bmatrix} \qquad T(c_n^u) = \begin{bmatrix} \mathbb{1}[c_n = 1] \\ \dots \\ \mathbb{1}[c_n = C-1] \end{bmatrix} \tag{3.48}$$

$$\boldsymbol{\theta}_u^{s_n,c_n} = \begin{bmatrix} \boldsymbol{\theta}_{u,1}^{s_n,c_n} \\ \boldsymbol{\theta}_{u,2}^{s_n,c_n} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_u^{s_n,c_n-1}\boldsymbol{\mu}_u^{s_n,c_n} \\ -\frac{1}{2}\operatorname{vec}(\boldsymbol{\Sigma}_u^{s_n,c_n,-1}) \end{bmatrix} \qquad T(\mathbf{x}_n) = \begin{bmatrix} \mathbf{x}_n \\ \operatorname{vec}(\mathbf{x}_n\mathbf{x}_n^\top) \end{bmatrix}, \tag{3.49}$$

The final super-vector embedding of an acoustic-unit is given by the concatenation of the natural parameters of the Normal and Categorical distributions composing the likelihood

function:

$$\boldsymbol{\eta}_u = \begin{bmatrix} \boldsymbol{\eta}_u^1 \\ \vdots \\ \boldsymbol{\eta}_u^i \\ \vdots \\ \boldsymbol{\eta}_u^S \end{bmatrix} = \begin{bmatrix} \boldsymbol{\omega}_u^i \\ \boldsymbol{\theta}_u^{i,1} \\ \vdots \\ \boldsymbol{\theta}_u^{i,C} \end{bmatrix} \tag{3.50}$$

where i is the index of the HMM state.

From the GSM formalism, the natural parameter vector of the uth phone is given by $\boldsymbol{\eta}_u = f(\mathbf{W}^T\mathbf{h}_u + \mathbf{b})$. We set the mapping function f such that standard parameters are given by the following relation:

$$\pi_{uc}^i = \frac{\exp\{\mathbf{W}_\pi^{i\top}\mathbf{h}_u + \mathbf{b}_\pi^i\}_c}{1 + \sum_{l=1}^{C-1}\exp\{\mathbf{W}_\pi^\top\mathbf{h}_i + \mathbf{b}_\pi^i\}_l} \tag{3.51}$$

$$\mathrm{diag}(\boldsymbol{\Sigma}_u^{i,j}) = \exp\{\mathbf{W}_\Sigma^{i,j\top}\mathbf{h}_u + \mathbf{b}_\Sigma^{i,j}\} \tag{3.52}$$

$$\boldsymbol{\Sigma}_u^{i,j-1}\boldsymbol{\mu}_u^{i,j} = \mathbf{W}_\mu^{i,j\top}\mathbf{h}_i + \mathbf{b}_\mu^{i,j}, \tag{3.53}$$

$$\tag{3.54}$$

where $\exp\{\dots\}$ is the element-wise exponential function, $\exp\{\dots\}_d$ is the dth dimension of the resulting vector, i is the index of the HMM state and j is the index of the mixture's component of the uth acoustic unit. $\mathbf{W}_\pi^i$, $\mathbf{W}_\Sigma^{i,j}$ and $\mathbf{W}_\mu^{i,j}$ are disjoint parts of the matrix $\mathbf{W}$ (and similarly for $\mathbf{b}_\pi^i$, ...).

Our choice for the function f is somewhat arbitrary: we chose f such that the subspace is linear (log-linear for the diagonal of the covariance matrices) in the natural parameter space of the HMM. Yet, beyond convenience, we have no motivation to favor one function over another. In the extreme case, one could possibly define f by a neural network with parameters to learn. This solution, even though appealing, has the major drawback to require a large number of phones to properly estimate the phonetic subspace and the function f. This situation is hardly met in our case as a usual language has around 50 - 100 phones which is by far not enough to learn any reasonable size neural network.

Contrary to the SGMM presented in section 3.1.4, the SHMM has 2 latent variables: the mixture's component index c_n and the HMM state index s_n. Furthermore, the exact alignment between the feature frames and the sequence of acoustic units is unknown, the acoustic unit index u is also a latent variable. Following the same notation as in section 2.2.4, we encode, in a variable z_n, both the state s_n and the acoustic unit index u. Therefore, the parameters $\boldsymbol{\phi}$ of the variational posteriors are given by:

$$q(c_n = i, z_n = j) = \phi_n^{i,j}. \tag{3.55}$$

The optimal variational posterior $q^*(\mathbf{c}, \mathbf{z})$ is obtained by the VB E-step of the HMM training as described in section 2.3.1 where the expectation of the natural parameters of an acoustic

unit/phone $\langle \boldsymbol{\eta}_u \rangle_{q(\boldsymbol{\Theta})}$ is obtained by sampling several values from the variational posterior $\boldsymbol{\eta}_u^l \sim q(\boldsymbol{\Theta})$ and taking the average. Finally, the variational posterior over embeddings and the parameters of the subspace of (3.18) is obtained by optimizing the following objective function:

$$\mathcal{L}_{m,\lambda}(\mathbf{m}, \boldsymbol{\lambda}; \dots) \approx \frac{1}{L}\left[\sum_{l=1}^{L} \left\langle \ln \frac{p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}_l)}{q(\mathbf{z})} \right\rangle_{q(\mathbf{z})}\right] - \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta})) \tag{3.56}$$

$$\mathbf{W}_l, \mathbf{b}_l, \mathbf{h}_1^l, \mathbf{h}_2^l, \dots \sim q(\boldsymbol{\Theta}) \tag{3.57}$$

$$\boldsymbol{\eta}_u^l = f(\mathbf{W}_l^\top \mathbf{h}_u^l + b_l) \tag{3.58}$$

$$\mathbf{H}_l = \left[\boldsymbol{\eta}_1^l, \boldsymbol{\eta}_2^l, \dots\right]. \tag{3.59}$$

The likelihood function $p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H})$ was described in depth in section 2.2.5. In practice, we optimize this objective function with an adaptive gradient ascent (ADAM (Kingma and Ba, 2014)) and we use automatic differentiation software to compute the gradients.

Finally, note that the description of the SHMM we have given can be applied to both supervised or unsupervised tasks. In the supervised setting, the phonetic transcription is given (i.e. the sequence of acoustic units $u_1, u_2, \dots, u_L$ is known) and defines the states' transition probabilities of the global HMM (the HMM composed of the acoustic units/-phones' HMM). In the unsupervised setting, the transcription is unknown and, therefore, the global HMM state transition probabilities are set to form a phone-loop as described in chapter 2.

3.2.3 Example: learning the English phonetic space

We demonstrate now the potential of the SHMM by learning a phonetic subspace for the English language. For this example, we used the TIMIT database as we did for the AUD experiment (see section 2.4.1). However, since we are now dealing with a supervised learning problem, we used the traditional training set (3696 utterances) and test set (412 utterances) (Lopes and Perdigao, 2011). We experimented with the MFCC and MBN features as described in section 2.4.2.

Similarly to the SGMM, the SHMM has some symmetries in its parameter space and requires, therefore, a careful initialization prior training. We used the following scheme:

1. we trained a standard HMM with GMM emissions for each phone using the Baum-Welch training and the provided phonetic transcription[6]. Each GMM has K components.

2. for each state of each phone's HMM

   (a) set the mixing weights $\boldsymbol{\pi}$ such that $\pi_k = \frac{1}{K}$

   (b) compute the per-state global mean $\hat{\boldsymbol{\mu}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\mu}_k$ and global diagonal covariance matrix $\hat{\boldsymbol{\Sigma}} = \frac{1}{K}\sum_{k=1}^{K} \boldsymbol{\Sigma}_k$

   (c) set each Gaussian component to have mean $\hat{\boldsymbol{\mu}}$ and covariance matrix $\hat{\boldsymbol{\Sigma}}$.

[6]Practically speaking, this is equivalent to train an HMM based phone recognizer with a flat phonotactic language model.

3. using the HMM estimated in step 1, we initialize $q^*(z_n = j)$ using the Baum-Welch algorithm and we set $q^*(c_n = i|z_n = j) = \text{const}$

4. we set $\mathbf{m}^*$ and $\boldsymbol{\lambda}^*$ as in (3.43) and (3.44), then, using $q^*(\mathbf{c}, \mathbf{z})$, we optimize the objective function defined in (3.56) for 10000 updates using ADAM with an initial learning rate of 0.001.

After this initialization, we trained the SHMM for 30 epochs as described in section 3.1.3 with the standard VB E-step for HMM as detailed in section 2.3.1. During the training, for each VB M-step we run a stochastic gradient ascent of 1000 steps. Once again, the learning rate of the gradient ascent was updated following ADAM. The state of the ADAM optimizer was preserved from the initialization till the end of the training.

For our first experiment, we trained an SHMM with 4 Normal components per state and a 2-dimensional subspace for visualization purposes. The learned phone embeddings $\mathbf{h}_{\text{aa}}, \mathbf{h}_{\text{m}}, \ldots$ are shown in Fig. 3.7. We observe that phones belonging to the same broad phonetic group tend to be closer to each other than phones from different groups. It confirms that the SHMM is able to learn a consistent phonetic subspace in the sense that distance between phone embeddings correlate with the phone clustering as done by linguists. The embeddings extracted with the MBN features (Fig. 3.7b) are a bit more noisy than the ones extracted with the MFCC features (Fig. 3.7a): closure, weak fricative and stop phones overlap each other. This observation confirms that discriminatively trained features, even though efficient for classification or related tasks, are not ideal for modeling the data.

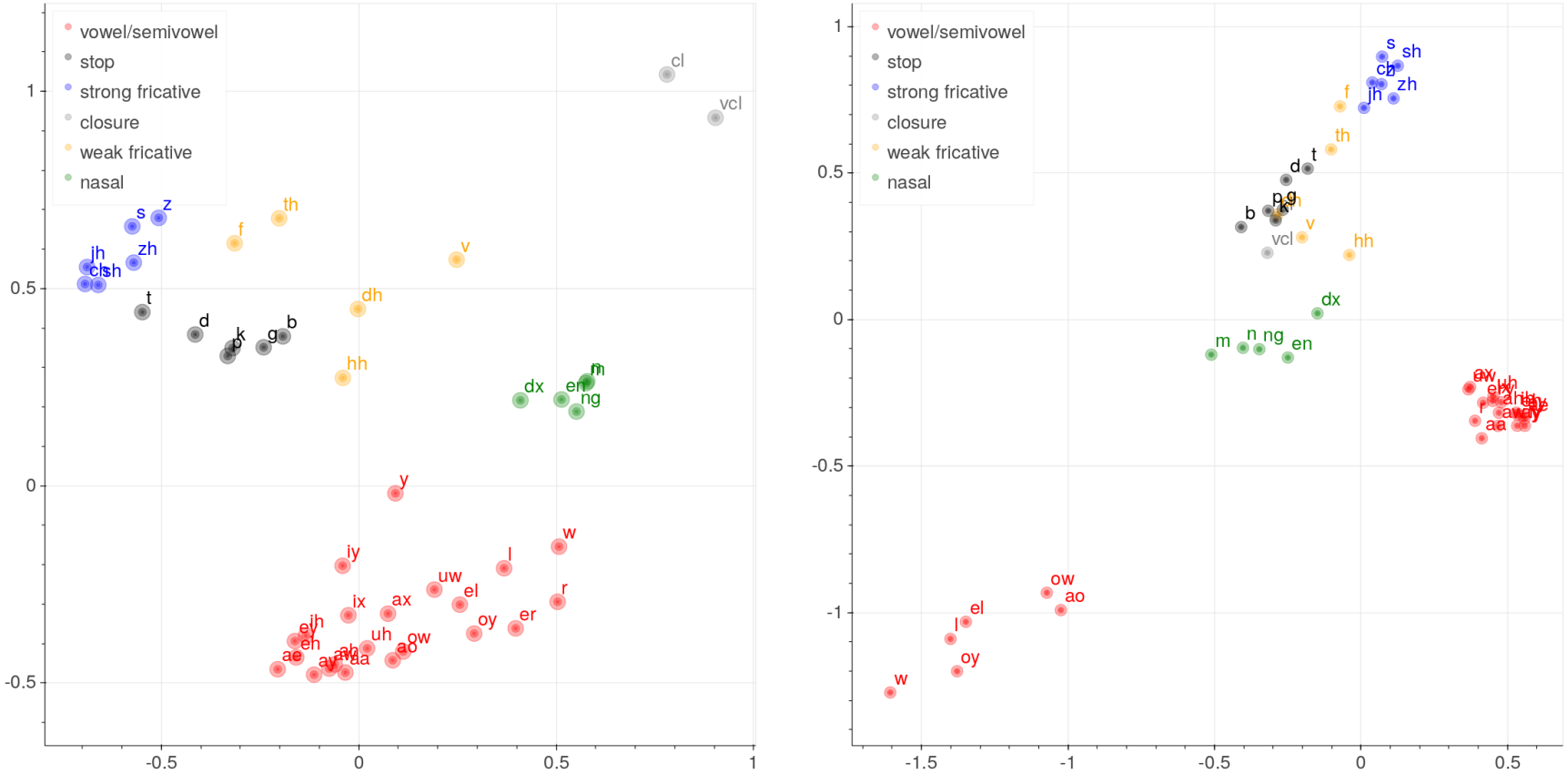


Figure 3.7: Posteriors of the phone embeddings learned by the SHMM on TIMIT using (a) MFCC features and (b) MBN features. Colors indicate the broad phonetic groups defined in (Lopes and Perdigao, 2011).

3.3 Dirichlet Process Subspace Hidden Markov Model

In section 3.2, we have defined the SHMM which, among other benefits, allows us to extract a low-dimensional subspace representing the phonetic continuum of a language. Now, we show how the SHMM and the Dirichlet Process can be combined to form the Dirichlet Process Subspace Hidden Markov Model (DP-SHMM). This new model is very similar to the phone-loop AUD model defined in section 2.2, however, by incorporating the phonetic subspace, it allows for significantly more accurate clustering of the acoustic units.

3.3.1 Revisiting the base measure

The base measure of the non-parametric phone-loop model defines a priori which sound is likely to be an acoustic unit. Practically, the base measure is a multivariate density over a HMM parameter vector $\boldsymbol{\eta}$ denoted $G_0(\boldsymbol{\eta})$. However, as the parameter space is high-dimensional and hardly interpretable, we have so far set the base measure to be a 11vague prior" which allows virtually any sound to become an acoustic unit. This choice has negative consequences as it allows the model to discover units that may not be relevant, for instance, the model may learn strongly speaker-dependent units. This problem can be resolved if we assume that we are given the phonetic subspace of the target language. Remember, from section 3.2.1, that the phonetic subspace describe a region in the total parameter space containing the phones of the language. With this piece of information, the AUD problem is easier as we only have to search for the low-dimensional embeddings $\mathbf{h}_1, \mathbf{h}_2, \dots$ in the phonetic subspace rather than the high-dimensional embeddings $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ in the full parameter space. This approach can be implemented by setting the base measure over the low-dimensional embeddings: $G_0 \equiv p(\mathbf{h})$. By doing so, we limit the prior over the acoustic units to the set of HMM parameters that are phonetically relevant. The modified base measure of the Dirichlet Process of the AUD model is depicted in Fig. 3.8.

Constraining the base measure also changes the generative process which can now be described in the following way:

1. draw $\gamma \sim \mathcal{G}(a_0, b_0)$
2. draw $v_i \sim \mathcal{B}(1, \gamma), \quad i = \{1, 2, \dots\}$
3. draw $\mathbf{h}_i \sim G_0 \quad i \in \{1, 2, \dots\}$

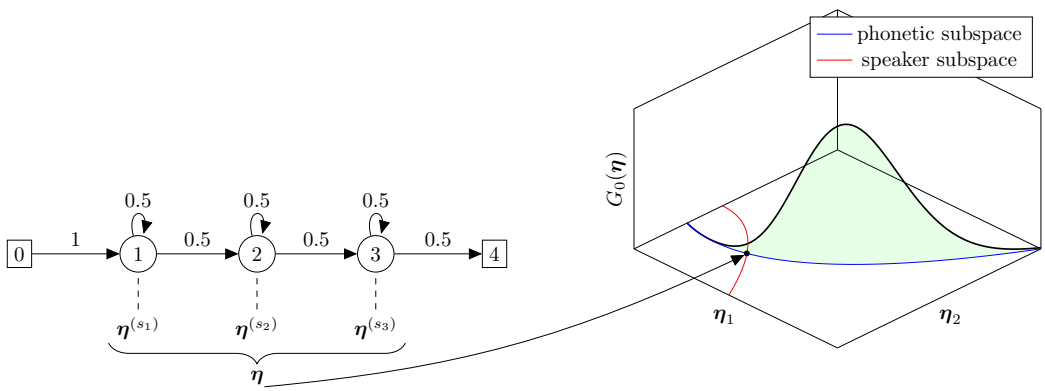


Figure 3.8: Base measure of the SHMM Dirichlet Process Mixture model.

4. map the unit embedding to the HMM parameter space $\boldsymbol{\eta}_i = f(\mathbf{W}^T\mathbf{h}_i + \mathbf{b})$
5. $\psi_i = v_i \prod_{j=1}^{i-1}(1 - v_j)$
6. Draw a sequence of units $\mathbf{u}$, $u_j \sim \mathcal{C}(\boldsymbol{\psi})$
7. For each u_j in $\mathbf{u}$
   (a) Draw a state path $\mathbf{s} = s_1, \ldots, s_l$ from the HMM transition probability distribution
   (b) for each state s_k in $\mathbf{s}$:
      i. Draw a component $c_k \sim \mathcal{C}(\boldsymbol{\pi}_{u_j}^{s_k})$ from the state's mixture weights
      ii. Draw a data point $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{u_j}^{s_k,c_k}, \boldsymbol{\Sigma}_{u_j}^{s_k,c_k})$

From step 5., the generative process is the same as the original AUD model described in section 2.2.3 and the function f is the SHMM mapping function defined in (3.51), (3.52) and (3.53). We call this new model the *Dirichlet Process Hidden Markov Model* (DP-SHMM) and its graphical representation is shown in Fig. 3.9. Interestingly, the base measure is not a proper density function in the $\boldsymbol{\eta}$ space, however, a sample from the Dirichlet Process, $G \sim G_0$, is indeed a discrete probability distribution over the atoms $\mathbf{h}_1, \mathbf{h}_2, \ldots$:

$$G(\mathbf{h}) = \sum_{i=1}^{\infty} \psi_i \delta_{\mathbf{h}_i}(\mathbf{h}). \tag{3.60}$$

The training of the DP-SHMM is the same as the SHMM with the two following modifications:

- the VB E-step is replaced with the one of the standard AUD phone-loop model
- during the VB M-step, the parameters of the susbpace $\mathbf{W}$ and $\mathbf{b}$ are assumed to be known, therefore, we only optimize the variational posteriors $q(\mathbf{h}_1), q(\mathbf{h}_2), \ldots$

3.3.2 Approximating the phonetic subspace of the target language

We have assumed that we had at our disposal the phonetic subspace of the language on which we would like to discover the acoustic units. Of course, this is not true in practice since to learn a phonetic subspace with an SHMM, one needs to have phonetic transcriptions of the audio recordings. Even though the actual phonetic subspace is unavailable, we can still approximate it using other languages. For instance, consider we wish to discover acoustic units from the Czech language. Czech has similar phonetics as other Slavic languages plus some extra typical phones such as the one denoted by the grapheme /ř/. In practice, /ř/ is well approximated by the combination of /r/ and /ž/ and, therefore, any phonetic subspace learned on a language having both /r/ and /ž/ would help to discover the /ř/ sound. From a more general perspective, despite the fact that each language has its own unique set of phones, there is a large overlap among languages of the same family. Consequently, a phonetic subspace from a given language can still be used to help discovering units from another language. Furthermore, we can also build a „universal“ phonetic subspace by learning the subspace on several languages together. This approach allows the subspace to cover a broader phonetic range, giving more flexibility to the AUD model to fit typical phones of the target language.

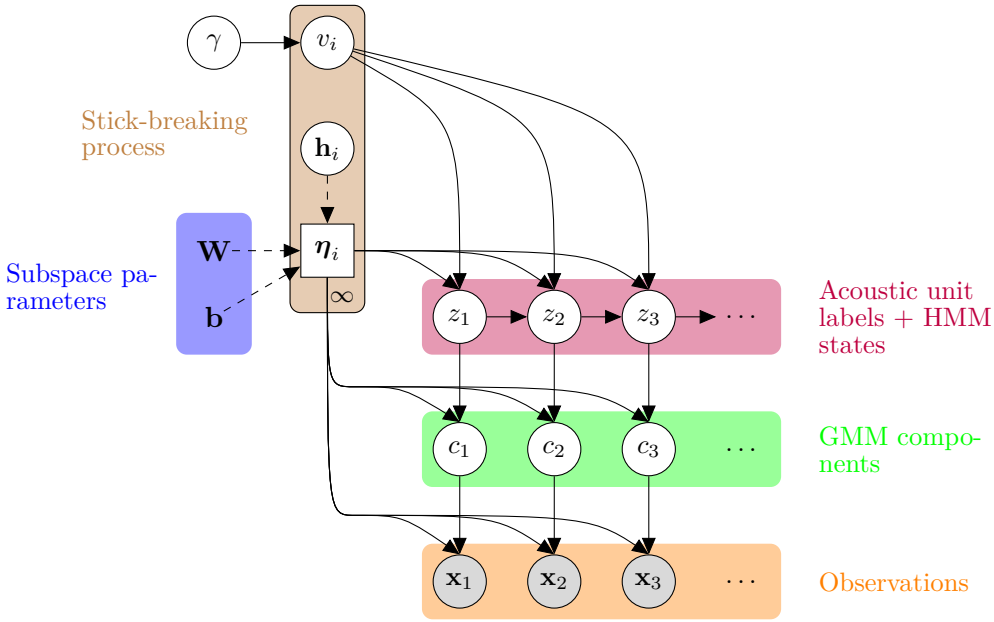


Figure 3.9: Bayesian network of the Dirichlet Process Subspace Hidden Markov Model (DP-SHMM). The atoms of the Dirichlet process are constrained to live in a low-dimensional subspace parameterized by $\mathbf{W}$ and $\mathbf{b}$.

3.4 Results

We now evaluate the DP-SHMM model on the AUD task. Our analysis focuses on the effect of the subspace dimension (section 3.4.2), the „goodness“ of the approximate phonetic subspace (section 3.4.3) and the comparison with the AUD phone-loop model (section 3.4.4).

3.4.1 Experimental setup

We used the same experimental setup as described in chapter 2: we attempt to discover acoustic units from the TIMIT and MBOSHI database using either MFCC or MBN features. To learn the approximate phonetic subspace for the DP-SHMM, which requires phonetically transcribed data (from languages different from the target one), we used a subset of GLOBALPHONE (Schultz, 2002). The GLOBALPHONE corpus is made of 16kHz recordings of read speech utterances of the most widespread languages in the world. Practically, we used 1500 utterances from the French (FR), Spanish (SP), German (GE) and Polish (PO) subsets of the corpus. Altogether, the 6000 utterances amount to roughly 14.6 hours of data. The exact duration for each language subset is shown in Table 3.1. For convenience we refer to this combination of the French, German, Polish and Spanish subset as the „Combined“ set, or CB for short. Note that, for the CB set, similar phones present in different languages are considered to be different. For example, the French /a/ and the Polish /a/ are assumed to be two different phones when training the phonetic subspace.

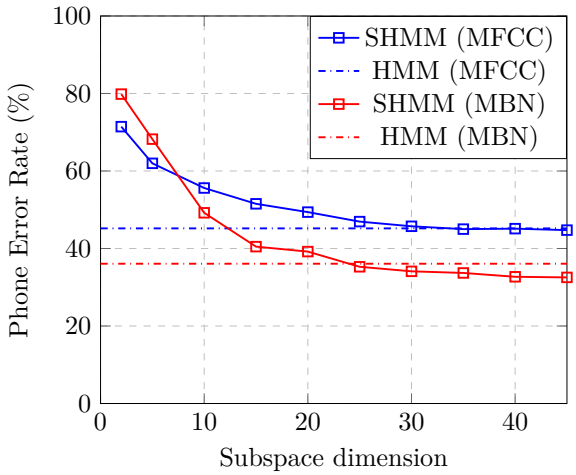


Figure 3.10: Performance of SHMM phone recognizer on TIMIT as a function the subspace dimension. When the dimension of the subspace reaches the number of phones, the model becomes unconstrained and have similar performance to a HMM phone recognizer.

3.4.2 Optimal subspace dimension

With the AUD phone-loop model introduced in chapter 2, the size of an acoustic unit embedding $\boldsymbol{\eta}$ is defined by number of parameters of the corresponding probabilistic model. However, for the DP-SHMM, the dimension of an acoustic unit embedding $\mathbf{h}$ depends on the dimension of the subspace D which is a meta-parameter. From Figure 3.10, we see that, when the phonetic subspace is learned using the actual phones of the language, a 40-dimensional subspace is sufficient to encode all the phonetic variability of the language. However, in the AUD task, we cannot learn the exact phonetic subspace and therefore, the optimal subspace dimension may be radically different. A low-dimensional subspace heavily constrains the AUD search whereas a large number of dimensions allows fine-grained acoustic units modeling potentially non-phonetic information.

For our first experiment, we trained DP-SHMM based AUD models with 50-, 75- and 100-dimensional subspace. The phonetic subspace was learned on the CB set. The results in terms of NMI are shown in Table 3.2. We see that, for TIMIT and MBOSHI, the higher the dimension of the subspace, the better the NMI. From this result, we fixed the subspace dimension to 100 for all subsequent experiments. It is of course possible to set the dimension of the subspace to a higher value but due to limited computational resources, we did not investigate further.

GLOBALPHONE subset	GE	PO	FR	SP	CB
# utterances	1500	1500	1500	1500	6000
amount of data (hours)	2.72	3.41	3.83	4.67	14.63
# phones	41	45	38	40	164

Table 3.1: Statistics of the data to estimate the universal phonetic subspace of the DP-SHMM.

Corpus	Subspace dimension	NMI (%)
TIMIT	50	38.38
TIMIT	75	39.00
TIMIT	100	**39.94**
MBOSHI	50	38.38
MBOSHI	75	39.55
MBOSHI	100	**39.98**

Table 3.2: Results in terms of NMI of the DP-SHMM model on the AUD task with different subspace dimensions and using the MFCC features. In all the cases, the phonetic subspace was estimated with the CB set.

3.4.3 Benefits of the universal phonetic subspace

In Section 3.3.2, we have proposed to approximate the phonetic subspace of the target language using labeled data from one or several languages. We now assess experimentally the benefits of this approach. In this experiment, we have trained and evaluated the DP-SHMM on the AUD task with the phonetic subspace estimated from:

- each individual language from our GLOBALPHONE subset, that is French (FR set), Spanish (SP set), German (GE set) and Polish (PO set)
- all the languages together (CB set)
- the same data set as the target data for the AUD (TIMIT or MBOSHI).

When using the same data as for the AUD task, this is of course a "cheating" experiment as we use the actual labels of the corpus. Nonetheless, it provides an upper bound on the best achievable results with the DP-SHMM. Results, measured with the NMI metric, are shown in Table 3.3. On TIMIT, learning the phonetic subspace from the combination of the 4 languages (CB set) yields a significant improvement to using only one language to estimate the subspace. However, for the MBOSHI corpus, the subspace learned from the CB set is as good as the subspace learn from the FR set or the GE set. In both cases, there is a large difference when using the optimal subspace (learned from the target data) and any other subspace.

3.4.4 Comparison with the DP-HMM

Finally, we compare the DP-SHMM against the previous phone-loop model presented in chapter 2 on the AUD task with no subspace modeling. For the sake of brevity, we refer to this model as the *Dirichlet Process Hidden Markov Model* (DP-HMM). For this experiment, the 100 dimensional phonetic subspace of the DP-SHMM was estimated on the 14.6 hours of the CB set. We ran our experiments on the TIMIT and MBOSHI corpora with both the MFCC and MBN features. The results are shown in Table 3.4. Results show that the DP-SHMM gives a significant improvement over the DP-HMM both in terms of segmentation (F-score metric) and clustering (NMI metric). Also, it experimentally confirms that the base measure is a key element of the non-parametric phone-loop model. An important observation is that the DP-SHMM trained with MFCC features performs better than the DP-HMM trained with MBN features on both corpora. This suggests that the GSM framework offers a more efficient way to implement knowledge transfer across languages.

Features	Subspace Corpus	Target Corpus	NMI (%)
MFCC	GE	TIMIT	36.35
MFCC	PO	TIMIT	38.00
MFCC	FR	TIMIT	37.66
MFCC	SP	TIMIT	37.54
MFCC	CB	TIMIT	39.94
MFCC	same	TIMIT	**43.52**
MFCC	GE	MBOSHI	40.01
MFCC	PO	MBOSHI	39.29
MFCC	FR	MBOSHI	40.29
MFCC	SP	MBOSHI	39.00
MFCC	CB	MBOSHI	39.98
MFCC	same	MBOSHI	**50.87**

Table 3.3: Results in terms of NMI of the DP-SHMM model, on the AUD task using MFCC features with phonetic subspace estimated on various data sets. The category „same“ indicates that the phonetic subspace and the AUD task were run on the same data set.

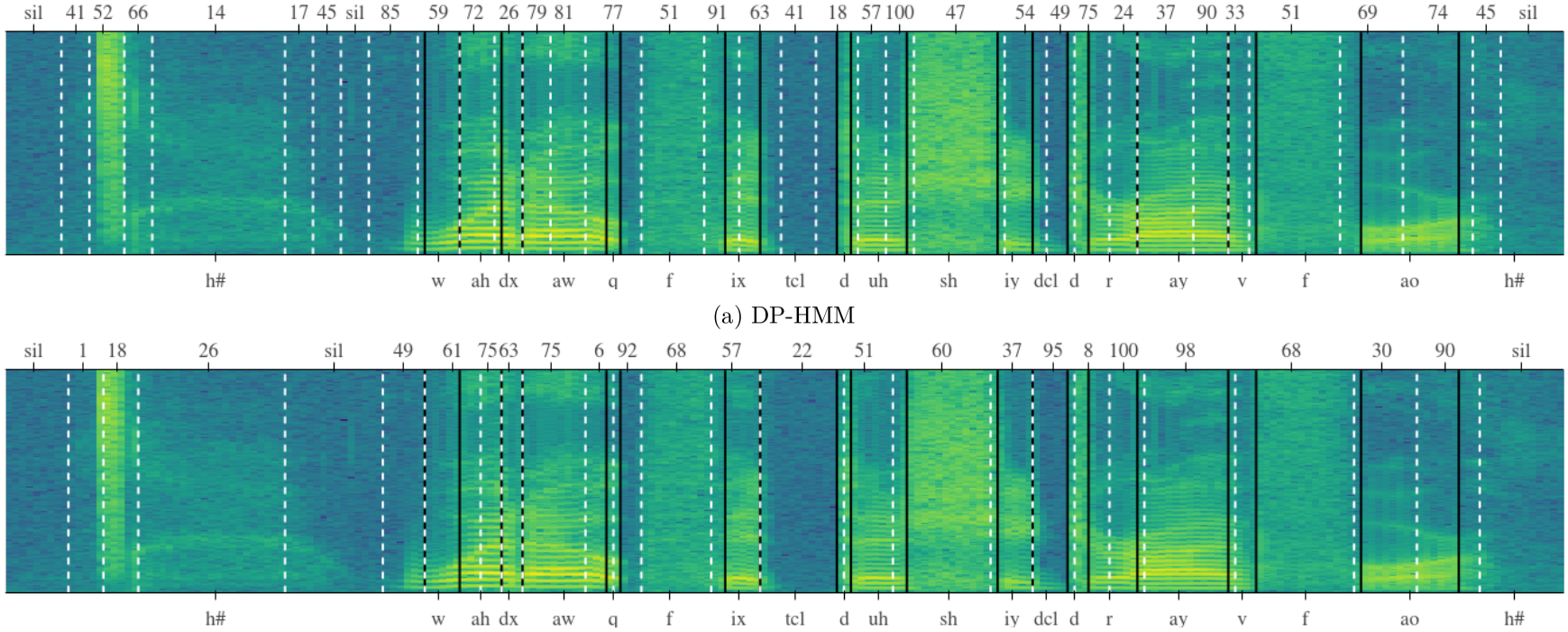
Finally we plotted in Fig. 3.11 the data-driven segmentation for one utterance given the DP-HMM and the DP-SHMM. We observed that, as shown by the F-Score metric in table 3.4, the DP-SHMM provides a much more accurate segmentation and drastically reduces the number of spurious boundaries.

3.5 Conclusion

In chapter 2, we have introduced non-parametric HMM-based model to discover acoustic units from unlabeled audio recordings. This model depends on a base measure: a probability density function setting *a priori* which sound is likely to be an acoustic unit candidate. A common setting for this base measure is a vague prior letting, therefore, all the sounds as possible acoustic units. In this chapter, we have proposed a new method to design a more accurate base measure. First, we have introduced the *Generalized Subspace Model* (GSM), a unified framework to derive embeddings representing probabilistic models. Then, we have applied the GSM to a set of HMMs representing the phones of a language in order to learn a *phonetic subspace*: a smooth low-dimensional manifold in the HMM parameters space cap-

Model	Features	Corpus	F-score	NMI (%)
DP-HMM	MFCC	TIMIT	63.01	34.81
DP-SHMM	MFCC	TIMIT	**77.24**	**39.94**
DP-HMM	MBN	TIMIT	58.07	37.17
DP-SHMM	MBN	TIMIT	**66.40**	**40.17**
DP-HMM	MFCC	MBOSHI	46.89	35.98
DP-SHMM	MFCC	MBOSHI	**57.65**	**39.98**
DP-HMM	MBN	MBOSHI	44.09	32.13
DP-SHMM	MBN	MBOSHI	**56.24**	**36.52**

Table 3.4: Comparison of the DP-HMM and the DP-SHMM on the AUD task.



(a) DP-HMM

(b) DP-SHMM

Figure 3.11: Example of segmentation of utterance "What outfit does she drive for?" by the DP-HMM and the DP-SHMM models. Black lines represent the reference boundaries and white dashed lines are the boundaries of the AUD model. The bottom and top sequences of labels are the time aligned reference and proposed transcription respectively. 'yes, put it in read and write „assigned". 'sil" corresponds to the special silence unit.

turing the phonetic variability of the language. Finally, we used this phonetic subspace to constrain the base measure of the AUD phone-loop model giving rise to a new AUD model: the *Dirichlet Process Subspace Hidden Markov Model* (DP-SHMM). This new model requires labeled data from languages (other than the target one) to estimate a „universal phonetic subspace“. Then, the new AUD model discovers acoustic units constrained to live in this phonetic subspace. Experimental results have shown that this approach provides a significant gain in terms of both NMI and F-score. Also, we have observed that our „universal phonetic subspace“ is by far not optimal compared to the „true“ phonetic subspace of the target language. A better approximation of the phonetic subspace remains an open problem and could lead to significant improvement on the AUD task.

In addition to defining a better base measure, this approach also proposes a formal way to include knowledge extracted form other languages. This can be viewed from a Bayesian perspective where the learned phonetic subspace is used to define an „educated prior“. Importantly, this approach is not limited to the HMM model. Indeed, since it relies on the newly introduced GSM framework, it can be applied to a vast collection of models and to other tasks than AUD.

As a concluding remark, note that the final acoustic unit embeddings $\mathbf{h}_1, \mathbf{h}_2, \ldots$ live in the same space as the phone embeddings of the languages used to estimate the phonetic subspace. From this observation, it is relatively straightforward to interpret the derived acoustic units by comparing their distance to other known phones. For instance, if an acoustic unit emebedding lives close to several nasal phones, it is reasonable to believe that this unit is also a nasal sounds itself. By repeating this process for each acoustic unit, one could obtain a data-driven human-interpretable phone set.

Chapter 4

Phonotactic Language Model

As established in chapter 2, our Bayesian formulation of the AUD problem relies upon three major components: the acoustic unit model, the base measure and the prior over the acoustic unit language model (the phonotactic language model). The designs of the two first elements—the acoustic unit model and the base measure—were addressed in chapter 2 and chapter 3 respectively. We now focus our attention on the prior over the phonotactic language model. So far, we have used the Dirichlet Process Mixture Model as the back-bone of our AUD model. Implicitly, this assumes that each unit label is independent of the other labels from the sequence. This assumption is, however, very unrealistic as each language has very specific phonotactic constraints. To overcome this issue, we revisit the phone-loop AUD to incorporate a bigram phonotactic language model to capture these phonotactic constraints. In section 4.1, we define this new model through the use of a hierarchical non-parametric prior: the Hierarchical Dirichlet Process. The corresponding Variational Bayes inference algorithm is described in section 4.2. In section 4.3, we propose a „corrected“ version of the bigram AUD model to control how the acoustic and language model affects the learning. Finally, results are shown and commented in section 4.4.

4.1 Non-Parametric Bigram Phone-Loop Model

Our Bayesian approach to the AUD task depends on the definition of the prior distribution $p(\mathbf{u}, \mathbf{H})$ where $\mathbf{u} = u_1, \dots, u_L$ is a sequence of L labels and $\mathbf{H} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots)$ is a countably infinite set of acoustic unit embeddings. Recall from chapter 2 that setting $\mathcal{P}$ to be a Dirichlet Process leads to the following construction of the prior:

$$G(\boldsymbol{\eta}) \sim \mathcal{DP}(\gamma, G_0) \tag{4.1}$$

$$p(\mathbf{u}, \mathbf{H}) = \underbrace{\left[\prod_{n=1}^{L} \underbrace{G(\boldsymbol{\eta}_{u_n})}_{p(u_n|\mathbf{H})}\right]}_{p(\mathbf{u}|\mathbf{H})} \underbrace{\left[\prod_{k=1}^{\infty} G_0(\boldsymbol{\eta}_k)\right]}_{p(\mathbf{H})}, \tag{4.2}$$

where $\mathcal{DP}(\gamma, G_0)$ is a Dirichlet Process with concentration γ and base measure G_0. Importantly, we assume G_0 to be a continuous density function. The sampled measure $G(\boldsymbol{\eta})$ is a discrete distribution given by:

$$G(\boldsymbol{\eta}) = \sum_{i=1}^{\infty} \psi_i \delta_{\boldsymbol{\eta}_i}(\boldsymbol{\eta}), \tag{4.3}$$

where ψ_i is defined by step 3 of the stick-breaking process described in section 2.1.1. From (4.2), we see that, regardless of the sampled measure G, the probability of the label sequence is always given by an unigram language model. To overcome this limitation, one has to consider a non-parametric prior which can sample more complex probability distributions. In this work, we shall focus on the Hierarchical Dirichlet Process (HDP) that will allow us to construct a prior over bigram phonotactic language model. The HDP was introduced in (Teh et al., 2004) and applied to language modeling and word segmentation in (Goldwater et al., 2009). These works can be seen as the non-parametric extensions of the Hierarchical Dirichlet distribution for language model introduced in (MacKay and Peto, 1995). Note that the HDP is not the only choice of non-parametric prior able to capture phonotactic constraints, for instance, the Hierarchical Pitman-Yor Process (Teh, 2006) is another non-parametric prior best suited for long tail distributions.

4.1.1 Hierarchical Dirichlet Process

A HDP of order M is a sequence of M Dirichlet Processes where the base measure of the nth process is given by a sample of the $n-1$ process in the sequence. Formally, it is defined as:

$$G_1 \sim \mathcal{DP}(\gamma_0, G_0) \tag{4.4}$$
$$G_2 \sim \mathcal{DP}(\gamma_1, G_1) \tag{4.5}$$
$$\dots \tag{4.6}$$
$$G_M \sim \mathcal{DP}(\gamma_M, G_{M-1}). \tag{4.7}$$

The HDP is fully defined by the M concentration parameters $\gamma_1, \dots, \gamma_M$ and the initial base measure $G_0(\boldsymbol{\eta})$. Note that $G_1, G_2, \dots$ are discrete distributions over the atoms generated from the base measure G_0 at the first step of the process. Using this definition, we can extend the DP mixture model to an HDP mixture model to build a infinite phone-loop AUD model having n-gram phonotactic language model. In this work, we will limit ourselves to bigram language model (using a HDP with order $M = 2$) but the extension to arbitrary n-grams is straightforward. The construction of phone-loop prior $p(\mathbf{u}, \mathbf{H})$ is given by:

$$G_1 \sim \mathcal{DP}(\gamma_0, G_0) \tag{4.8}$$
$$G_{2,i} \sim \mathcal{DP}(\gamma_1, G_1) \quad \forall i \in \{0, 1, 2, ...\} \tag{4.9}$$
$$p(\mathbf{u}, \mathbf{H}) = \underbrace{\left[\prod_{n=1}^{L} \underbrace{G_{2,u_{n-1}}(\boldsymbol{\eta}_{u_n})}_{p(u_n|u_{n-1},\mathbf{H})}\right]}_{p(\mathbf{u}|\mathbf{H})} \underbrace{\left[\prod_{k=1}^{\infty} G_0(\boldsymbol{\eta}_k)\right]}_{p(\mathbf{H})} \tag{4.10}$$

In (4.10), the probability of the sequence of labels $\mathbf{u}$ is defined through an infinite set of distributions $G_{2,1}, G_{2,2}, \dots G_{2,\infty}$ where the ith distribution $G_{2,i}$ is the probability over the labels $1, 2, ...$ given that the previous label of the sequence was i. For convenience, we set $G_{2,0}$ to be the probability over the first label of the sequence. We see that it differs from the DP mixture model which uses a single distribution G to define the probability of a sequence of units. Inference for the HPD mixture model can be done by sampling using an extension of the Chinese Restaurant Process: the Chinese Restaurant Franchise (Teh et al., 2004). However since we have observed in chapter 2 that Variational Bayes inference is more suited to our problem, we will focus on a variational treatement of this model. Similarly to the

DP mixture model, we will first derive a stick-breaking construction (section 4.1.2) of the HDP and then apply the mean-field approximation (section 4.2).

4.1.2 Stick-Breaking constructions

A simple stick-breaking construction of the HDP is to iterate the Sethuraman's stick-breaking construction of the Dirichlet Process:

1. Draw $v_{1,i} \sim \mathcal{B}(1, \gamma_0), \quad i = \{1, 2, \dots\}$
2. Draw $\boldsymbol{\eta}_{1,i} \sim G_0, \quad i = \{1, 2, \dots\}$
3. $\psi_{1,i} = v_{1,i} \prod_{k=1}^{i-1}(1 - v_{1,k}) \quad i = \{1, 2, \dots\}$
4. $G_1(\boldsymbol{\eta}) = \sum_{i=1}^{\infty} \psi_{1,i} \delta_{\boldsymbol{\eta}_{1,i}}(\boldsymbol{\eta})$
5. For i in $\{1, 2, ...\}$:
   (a) Draw $v_{2,i,j} \sim \mathcal{B}(1, \gamma_1), \quad j = \{1, 2, \dots\}$
   (b) Draw $\boldsymbol{\eta}_{2,i,j} \sim G_1, \quad j = \{1, 2, \dots\}$
   (c) $\psi_{2,i,j} = v_{2,i,j} \prod_{k=1}^{j-1}(1 - v_{2,i,k}) \quad j = \{1, 2, \dots\}$
   (d) $G_{2,i}(\boldsymbol{\eta}) = \sum_{j=1}^{\infty} \psi_{2,i,j} \delta_{\boldsymbol{\eta}_{2,i,j}}(\boldsymbol{\eta})$.

The use of this stick-breaking construction for variational inference of the HDP was proposed in (Wang et al., 2011) and applied to the task of topic modeling. However, in our setting, this approach is not ideal because of its high memory requirements. To understand this, observe that the sample distribution G_1 is discrete and, therefore, there is a non-zero probability that $\boldsymbol{\eta}_{2,i,j} = \boldsymbol{\eta}_{2,i,l}$ for $j \neq l$. Consequently, the conditional probability of observing the label u_n given the previous label u_{n-1} is:

$$p(u_n | u_{n-1}, \mathbf{H}) = \sum_{j=1}^{\infty} \psi_{2,u_{n-1},j} \delta_{\boldsymbol{\eta}_{2,u_{n-1},j}}(\boldsymbol{\eta}_{u_n}). \tag{4.11}$$

Since the value of $\boldsymbol{\eta}_{2,i,j}$ is hidden, it implies that the inference needs to learn a (probabilistic) mapping between $\boldsymbol{\eta}_{2,i,j}$ and $\boldsymbol{\eta}_{1,i}$. If we assume that we truncate the stick-breaking process at each level at the index τ, then posterior distribution of the bigram language model will need $O(\tau^3)$ parameters (τ root-level atoms which can be mapped to τ bottom-level atoms and each of them has a distribution over τ possible next units).

An alternative to the Sethurman's stick-breaking construction is the Teh's stick-breaking construction (Teh et al., 2004) of the HDP given by:

1. Draw $v_{1,i} \sim \mathcal{B}(1, \gamma_0), \quad i = \{1, 2, \dots\}$
2. Draw $\boldsymbol{\eta}_{1,i} \sim G_0, \quad i = \{1, 2, \dots\}$
3. $\psi_{1,i} = v_{1,i} \prod_{k=1}^{i-1}(1 - v_{1,k})$
4. $G_1(\boldsymbol{\eta}) = \sum_{i=1}^{\infty} \psi_{1,i} \delta_{\boldsymbol{\eta}_{1,i}}(\boldsymbol{\eta}), \quad i = \{1, 2, \dots\}$
5. For j in $\{1, 2, ...\}$:

(a) Draw $v_{2,i,j} \sim \mathcal{B}(\gamma_1 \psi_{1j}, \gamma_1(1 - \sum_{k=1}^{j} \psi_{1,k}), \quad l = \{1, 2, \dots\}$

(b) $\psi_{2,i,j} = v_{2,i,j} \prod_{k=11}^{i-1}(1 - v_{2,i,k}), \quad l = \{1, 2, \dots\}$

(c) $G_{2,i}(\boldsymbol{\eta}) = \sum_{j=1}^{\infty} \psi_{2,i,j} \delta_{\boldsymbol{\eta}_{1,j}}(\boldsymbol{\eta})$.

We see that the final measure $G_{2,c}(\boldsymbol{\eta})$ is directly defined with the atoms sampled from the root measure G_0. As G_0 is a density, each atom is unique and therefore the bigram language model is given by:

$$p(u_n | u_{n-1}, \mathbf{v}_{2,u_{n-1}}) = \psi_{2,u_{n-1},u_n}. \tag{4.12}$$

In this case, since we do not need to keep some extra mapping, the (approximate) posterior will need only $O(\tau^2)$ parameters. Teh's construction is more parsimonious and allows faster inference but it has also a drawback: by using the samples $v_{1,1}, v_{1,2}, \dots$ as parameters of the Beta distribution (step 5a) the factors $p(\mathbf{v}_2|\mathbf{v}_1)p(\mathbf{v}_1)$ are not any more conjugate (appendix B.1.2) which makes inference of distribution over $\mathbf{v}_1$ more difficult. Nevertheless, we chose to use the Teh's stick-breaking construction of the HDP in our model for practical reasons.

4.1.3 Complete Model

So far, we have only focused on the construction of the non-parametric prior $p(\mathbf{u}, \mathbf{H})$ using a 2-level HDP. We now connect the HDP prior with the remaining part of the AUD model described in chapter 2. Assuming the Teh's stick-breaking construction, the graphical representation of the complete model is shown in Fig. 4.1. We call this new model the HDP-HMM model. Note that extending this model to use Subspace HMM (leading to the HDP-SHMM model) is trivial. A few important observations from Fig. 4.1 can be made: first, the introduction of the HDP prior leads to bigram connections between the label sequence $u_1, u_2, \dots$ as desired. Despite these new connections, the model can still be easily turned into a 1-level HMM (see chapter 2). Therefore, inferring the most likely sequence of labels given some observations will have the same time complexity as in our previous model. Another important observation is that, thanks to the Teh's stick-breaking construction, the relation between the atoms of the HDP $\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots$ and the acoustic model (the HMM states, the GMM components and the observations) remains unchanged. Finally, notice that, for simplicity reasons, we have a prior over the concentration of the top level stick-breaking process (γ_0) but not on the one of the second level stick-breaking process. Therefore, the concentration parameter γ_1 is considered fixed in this work and is set to half of the truncation parameter: $\gamma_1 = \frac{\tau}{2}$.

4.1.4 Joint distribution

Finally, to conclude the presentation of the model, we present the complete joint distribution of a sequence of features $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_N$, the latent variables $\mathbf{c} = c_1, \dots, c_N$, $\mathbf{z} = z_1, \dots, z_N$ and the parameters $\mathbf{H} = \boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_\infty$, $\mathbf{v}_1 = v_{1,1}, \dots, v_{1,\infty}$, γ_0, $\mathbf{V}_2 = \mathbf{v}_{2,1}, \dots, \mathbf{v}_{2,\infty}$. The joint distribution can be written as:

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{V}_2, \mathbf{v}_1, \gamma_0) = p(\mathbf{H})p(\gamma_0)p(\mathbf{v}_1|\gamma_0)p(\mathbf{V}_2|\mathbf{v}_1)p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{V}_2). \tag{4.13}$$

Note that the variable γ_1 is not included as it is considered as a fixed constant. The terms $p(\mathbf{H})$, $p(\gamma_0)$, $p(\mathbf{v}_1|\gamma_0)$ and $p(\mathbf{X}, \mathbf{c}|\mathbf{z}, \mathbf{H})$ are the same as defined in section 2.2.5. The prior

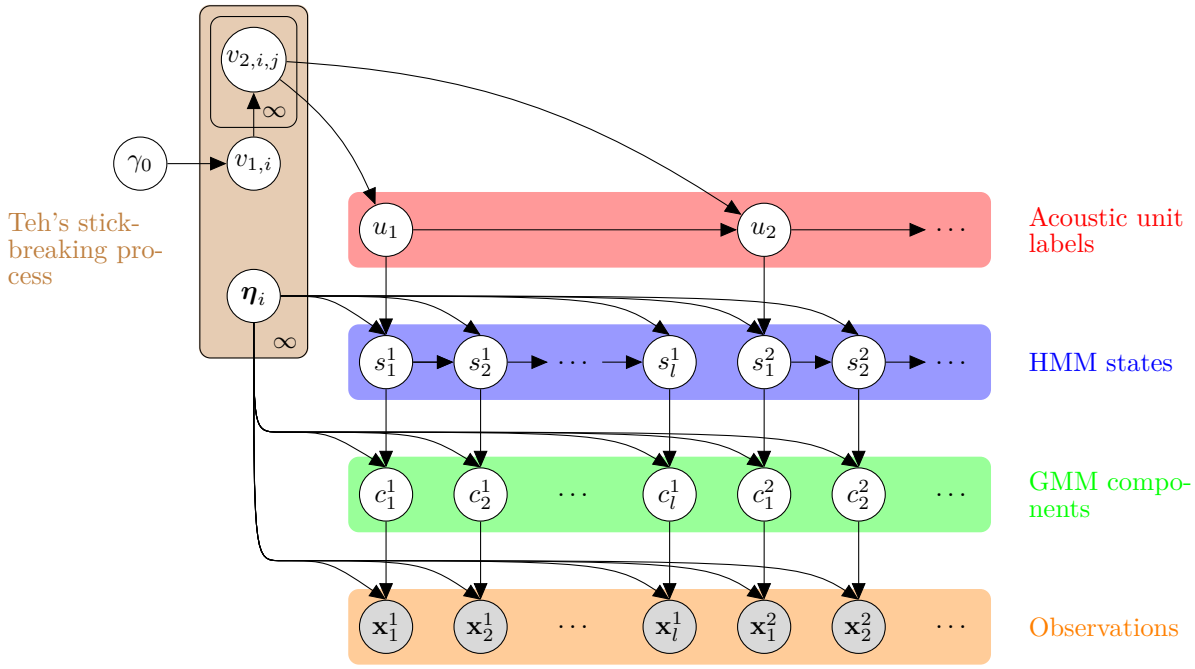


Figure 4.1: Graphical representation of the HDP-HMM with the Teh's stick-breaking construction of the HDP. Contrary to the DP-HMM model, the nth unit label only depends on the $n-1$st labels.

over the second level stick-breaking parameters is given by:

$$p(\mathbf{V}_2|\mathbf{v}_1) = \prod_{i=1}^{\infty}\prod_{j=1}^{\infty} p(v_{2,i,j}|\mathbf{v}_1) \tag{4.14}$$

$$p(v_{2,i,j}|\mathbf{v}_1) \equiv \mathcal{B}\big(\gamma_1\psi_{1,j}, \quad \gamma_1(1-\sum_{k=1}^{j}\psi_{1,k})\big) \tag{4.15}$$

$$\psi_{1,i} = v_{1,i}\prod_{k=1}^{i-1}(1-v_{1,k}), \tag{4.16}$$

and the likelihood of the data and the latent variables is given by:

$$p(\mathbf{X},\mathbf{c},\mathbf{z}|\mathbf{H},\mathbf{V}_2) = p(\mathbf{z}|\mathbf{V}_2)p(\mathbf{X},\mathbf{c}|\mathbf{z},\mathbf{H}) \tag{4.17}$$

$$= \prod_{n=1}^{N} p(z_n|z_{n-1},\mathbf{V}_2)p(\mathbf{x}_n,c_n|\boldsymbol{\eta}_{z_n}). \tag{4.18}$$

Recall that z_n encodes an acoustic unit index u_n and a particular HMM state s_n, therefore, the sequence of N units and states $\mathbf{z} = z_1,\ldots,z_N$ can be equivalently defined as a sequence of L acoustic units $\mathbf{u} = u_1,\ldots,u_L$ and L sequences of HMM states $\mathbf{s}^{u_l} = s_1^{u_l},\ldots,s_{N_l}^{u_l}$. Using

this notation, (4.18) can be equivalently expressed as:

$$p(\mathbf{X}, \mathbf{c}, \mathbf{z}|\mathbf{H}, \mathbf{V}_2) = \underbrace{\prod_{l=1}^{L} p(u_l|\mathbf{v}_{2,u_{l-1}}) \prod_{n=1}^{N_l} p(s_n^{u_l}|s_{n-1}^{u_l})}_{\prod_{n=1}^{N} p(z_n|z_{n-1},\mathbf{V}_2)} p(\mathbf{x}_n^{u_l}, c_n^{u_l}|\boldsymbol{\eta}_{u_l}^{s_n}). \tag{4.19}$$

$$p(u_l|u_{l-1}, \mathbf{v}_{2,u_{l-1}}) = v_{2,u_{l-1},u_l} \prod_{k=1}^{u_l-1} (1 - v_{2,u_{l-1},k}), \tag{4.20}$$

where N_l is the length of the lth segment of the unit with index u_l.

4.2 Inference

Previously, we have described the HDP-HMM from the generative point of view. We now address the problem of estimating the posterior over the latent variables given some observations:

$$p(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{V}_2, \mathbf{v}_1, \gamma_0|\mathbf{X}) \propto p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{V}_2, \mathbf{v}_1, \gamma_0), \tag{4.21}$$

where $\mathbf{V}_2 = (\mathbf{v}_{2,1}, \mathbf{v}_{2,2}, \dots)$. Recall from section 2.2.4 that variable $\mathbf{z}$ encodes the sequence of unit labels $\mathbf{u}$ and HMM states $\mathbf{s}$. Since normalizing (4.21) is intractable, we aim to find an approximate variational posterior (appendix A) with the following structured mean-field factorization (appendix A.2.3):

$$q(\mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{V}_2, \mathbf{v}_1, \gamma_0) = q(\mathbf{c}|\mathbf{z})q(\mathbf{z})q(\mathbf{H})q(\mathbf{V}_2)q(\mathbf{v}_1)q(\gamma_0), \tag{4.22}$$

Our factorized posterior is identical to (2.47) with the addition of factor $q(\mathbf{V}_2)$, therefore, the training of the HDP-HMM model will also be a VB-EM algorithm. Notice that for tractability reasons, it is necessary to assume the posterior over the parameters of the stick-breaking processes at each level of the hierarchy are independent. The optimal factors $q^*(\mathbf{c}|\mathbf{z})$, $q^*(\mathbf{z})$, $q^*(\mathbf{H})$ are easily found using (2.60), (2.68), (2.82) and (2.84) and replacing the prior over the phone-loop state $p(\mathbf{z}|\mathbf{v})$ by the one obtained from the HDP $p(\mathbf{z}|\mathbf{V}_2)$.

4.2.1 VB-M step for the HDP

As already mentioned, $p(\mathbf{v}_1)$ and $p(\mathbf{V}_2|\mathbf{v}_1)$ are not conjugate and, therefore, the factorization in (4.22) is not sufficient to get analytical solutions for the optimal factors $q^*(\mathbf{V}_2)$ and $q^*(\mathbf{v}_1)$. We tackle this issue by first training a DP-HMM phone-loop AUD model and then setting $q^*(\mathbf{v}_1) \equiv q^*_{\text{DP-HMM}}(\mathbf{v})$ and $q(\gamma_0) \equiv q^*_{\text{DP-HMM}}(\gamma)$[1]. In (Hughes et al., 2015), the author proposed to learn the parameters of the root stick-breaking process with numerical optimization but we didn't observe any benefit in our case and consequently opted for a more straightforward solution.

We derive now the optimal variational posterior of the 2-level stick-breaking process of the HDP:

$$\ln q^*(\mathbf{V}_2) = \langle \ln p(\mathbf{X}, \mathbf{c}, \mathbf{z}, \mathbf{H}, \mathbf{V}_2, \mathbf{v}_1, \gamma_0) \rangle_{q(\mathbf{c}|\mathbf{z})q(\mathbf{c})q(\mathbf{H})q(\mathbf{v}_1)q(\gamma_0)} + \text{const} \tag{4.23}$$

$$\ln q^*(\mathbf{V}_2) = \langle \ln p(\mathbf{z}|\mathbf{V}_2) \rangle_{q(\mathbf{z})} + \ln \langle p(\mathbf{V}_2|\mathbf{v}_1) \rangle_{q(\mathbf{v}_1))} + \text{const} \tag{4.24}$$

[1] For consistency reasons, we also initialize variational posteriors of the other parameters with the optimal variational posteriors of the DP-HMM model.

Using the fact that $p(\mathbf{z}|\mathbf{V}_2) = p(\mathbf{s}|\mathbf{u})p(\mathbf{u}|\mathbf{V}_2)$ (and consequently $q(\mathbf{z}) = q(\mathbf{s}|\mathbf{u})q(\mathbf{u})$), we have:

$$\ln q^*(\mathbf{V}_2) = \langle \ln p(\mathbf{u}|\mathbf{V}_2)\rangle_{q(\mathbf{u})} + \langle \ln p(\mathbf{V}_2|\mathbf{v}_1)\rangle_{q(\mathbf{v}_1)} + \text{const} \tag{4.25}$$

$$= \langle \ln \prod_{n=1}^{|\mathbf{u}|} p(u_n|u_{n-1}, \mathbf{v}_{2,u_{n-1}})\rangle_{q(\mathbf{u})} + \langle \ln \prod_{k=1}^{\infty}\prod_{l=1}^{\infty} p(v_{2,k,l}|\mathbf{v}_1, \gamma_0)\rangle_{q(\mathbf{v}_1)} + \text{const} \tag{4.26}$$

$$= \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\langle \sum_{n=1}^{L} \mathbb{1}[u_n = l, u_{n-1} = k] \ln p(u_n|u_{n-1}, \mathbf{v}_{2,u_{n-1}})\rangle_{q(\mathbf{u})} + \langle \ln p(v_{2,k,l}|\mathbf{v}_1)\rangle_{q(\mathbf{v}_1)} + \text{const}\,. \tag{4.27}$$

where L is the length of sequence $\mathbf{u}$. Using (4.12) we have:

$$\ln q^*(\mathbf{V}_2) = \sum_{k=1}^{\infty}\sum_{l=1}^{\infty}\Big\langle \sum_{n=1}^{L} \mathbb{1}[u_n > l, u_{n-1} = k] \ln(1 - v_{2,k,l}) + \mathbb{1}[u_n = l, u_{n-1} = k] \ln v_{2,k,l}\Big\rangle_{q(\mathbf{u})} + \langle \ln p(v_{2,k,l}|\mathbf{v}_1)\rangle. \tag{4.28}$$

From (4.28), we observe that the variational posterior factorizes in the following way:

$$q^*(\mathbf{V}_2) = \prod_{k=1}^{\infty}\prod_{l=1}^{\infty} q(v_{2,k,l}). \tag{4.29}$$

To cope with the infinite product, we truncate the variational posterior by keeping only the τ first posteriors such that $q(v_{2,k,l} = 1) = 1$ if $k > \tau$ or $l > \tau$. Notice that this factorization is not assumed explicitly but is a consequence of our original mean-field factorization in (4.22). Replacing $\ln p(v_{2,k,l}|\mathbf{v}_1)$ by the Beta distribution formula, we express each factor as:

$$\ln q^*(v_{2,k,l}) = (\langle b_l\rangle_{q(\mathbf{v}_1)} + \sigma_1 - 1)\ln(1 - v_{2,k,l}) + (\langle a_l\rangle_{q(\mathbf{v}_1)} + \sigma_2 - 1)\ln v_{2,k,l} + \text{const}, \tag{4.30}$$

$$\tag{4.31}$$

where:

$$a_l = \gamma_1\psi_{1,l} \tag{4.32}$$

$$b_l = \gamma_1(1 - \sum_{i=1}^{l}\psi_{1,i}) \tag{4.33}$$

$$\sigma_1 = \langle \sum_{n=1}^{L} \mathbb{1}[u_n > l, u_{n-1} = k]\rangle_{q(\mathbf{u})} \tag{4.34}$$

$$\sigma_2 = \langle \sum_{n=1}^{L} \mathbb{1}[u_n = l, u_{n-1} = k]\rangle_{q(\mathbf{u})}. \tag{4.35}$$

(4.30) is the parametric form of a unnormalized log Beta distribution which leads to the final solution of our variational posterior:

$$q^*(v_{2,k,l}) = \mathcal{B}(a'_{k,l}, b'_{k,l}) \tag{4.36}$$

$$a'_{k,l} = \langle a_l\rangle_{q(\mathbf{v}_1)} + \sigma_2 \tag{4.37}$$

$$b'_{k,l} = \langle b_l\rangle_{q(\mathbf{v}_1)} + \sigma_1. \tag{4.38}$$

Algorithm 4.1 Training of bigram phone-loop model for acoustic unit discovery

function MSTEP($\mathbf{X}, q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}), q^*(\mathbf{v}_1)$)
  ▷ Update defined in (2.80)
  $q^*(\mathbf{H}) \leftarrow \arg\max_{q(\mathbf{H})} \mathcal{L}$
  ▷ Update defined in (4.29) and (4.30)
  $q^*(\mathbf{V}_2) \leftarrow \arg\max_{q(\mathbf{V}_2)} \mathcal{L}$

function ESTEP($\mathbf{X}, q(\mathbf{H}), q(\mathbf{V}_2)$)
  ▷ For both updates, we use $q(\mathbf{V}_2)$ to estimate the transition probability of the HMM:
  ▷ Update defined in (2.60)
  $q^*(\mathbf{c}|\mathbf{z}) \leftarrow \arg\max_{q(\mathbf{c}|\mathbf{z})} \mathcal{L}$
  ▷ Update defined in (2.68)
  $q^*(\mathbf{z}) \leftarrow \arg\max_{q(\mathbf{z})} \mathcal{L}$
  return $q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z})$

procedure TRAIN($\mathbf{X}, E$)
  ▷ E: number of epochs (i.e. E-step + M-step)
  ▷ initialization: we assume that we have already trained a DP-HMM
  $q^*(\mathbf{H}) \leftarrow q^*_{\text{DP-HMM}}(\mathbf{H})$
  ▷ $q^*(\mathbf{v}_1)$ will remain fixed throughout the training
  $q^*(\mathbf{v}_1) \leftarrow q^*_{\text{DP-HMM}}(\mathbf{v})$
  $q^*(\mathbf{V}_2) \leftarrow p(\mathbf{V}_2|\mathbf{v}_1)$
  for $e \leftarrow 1$ to E **do**
    $q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}) \leftarrow$ ESTEP($\mathbf{X}, q^*(\mathbf{H}), q^*(\mathbf{V}_2)$)
    $q^*(\mathbf{H}), q^*(\mathbf{V}_2) \leftarrow$ MSTEP($\mathbf{X}, q^*(\mathbf{c}|\mathbf{z}), q^*(\mathbf{z}), q^*(\mathbf{v}_1)$)

Observing that $q^*(\mathbf{v}_1) = q^*_{\text{DP-HMM}}(\mathbf{v}) = \prod_{l=1}^{\tau} \mathcal{B}(\phi_{1,l}, \phi_{2,l})$ as defined in (2.93), the expectations are thus given by:

$$\langle a_l \rangle_{q(\mathbf{v}_1)} = \gamma_1 \frac{\phi_{1,l}}{\phi_{1,l} + \phi_{2,l}} \prod_{i=1}^{l-1} (1 - \frac{\phi_{1,i}}{\phi_{1,i} + \phi_{2,i}}) \tag{4.39}$$

$$\langle b_l \rangle_{q(\mathbf{v}_1)} = \gamma_1 \big(1 - \sum_{i=1}^{l} \frac{\phi_{1,i}}{\phi_{1,i} + \phi_{2,i}} \prod_{j=1}^{i-1} (1 - \frac{\phi_{1,j}}{\phi_{1,j} + \phi_{2,j}})\big). \tag{4.40}$$

$$\tag{4.41}$$

The complete training of the HDP-HMM is presented in Algorithm 4.1.

4.3 Improper Variational Bayes Inference

The bigram phone-loop model we have described in this chapter is a non-parametric Bayesian treatment of the traditional HMM-GMM used in speech recognition before the advent of deep learning techniques. The HMM-GMM for ASR, although theoretically convenient, has been known to be a rather crude generative model of speech. Particularly, the assumption that the features are independent given the sequence of HMM states is very

unrealistic. Consequently, it was common during the decoding to correct the model with a balancing factor α, $\alpha \in \mathbb{R}^+$, which controls the preponderance of the acoustic model:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})^\alpha p(\mathbf{w}), \tag{4.42}$$

where $\mathbf{w} = w_1, w_2, \ldots$ is a sequence of words. If $\alpha = 0$, the acoustic model $p(\mathbf{X}|\mathbf{w})$ becomes constant and the most likely decoded sequence of words $\mathbf{w}^*$ is simply the most probable sequence of words from the language model $p(\mathbf{w})$. On the other hand, for α sufficiently large, the role of the language model $p(\mathbf{w})$ becomes negligible. In practice, α is usually set below one as the HMM acoustic model, due to the independence assumption, tends to be underestimated the probability of the sequence of features.

Unfortunately, our infinite HMM phone-loop makes the same assumption and, therefore, suffers from the same caveat. We will now focus on how to balance the acoustic and phonotactic language in our model. Note however that our problem is slightly different: in (4.42), the problem is to correct the model at decoding time whereas we wish to take into account the correction while doing the inference in order to achieve better clustering. Note that a similar technique has been applied for an HMM-based speaker diarization system (Diez et al., 2019).

To begin, we consider the following simplified version of our model:

$$p(\mathbf{X}, \mathbf{z}, \mathbf{H}) = p(\mathbf{X}|\mathbf{z}, \mathbf{H})p(\mathbf{z}|\mathbf{H})p(\mathbf{H}) \tag{4.43}$$

where we have omitted the parameters of the model which will not be affected by any correction factors. $p(\mathbf{X}|\mathbf{z}, \mathbf{H})$ is the acoustic model and $p(\mathbf{z}|\mathbf{H})$ is the phonotactic language model. We now define a "corrected" version of this model by introducing two balancing factors, α and β, controlling respectively the roles of the acoustic and the language models:

$$p(\mathbf{X}, \mathbf{z}, \mathbf{H}) \propto p(\mathbf{X}|\mathbf{H}, \mathbf{z})^\alpha p(\mathbf{z})^\beta p(\mathbf{H}) = f(\mathbf{X}, \mathbf{z}, \mathbf{H}) \tag{4.44}$$

where $\alpha, \beta \in \mathbb{R}^+$. When $\alpha \neq 1$ and $\beta \neq 1$, the corrected model $f(\mathbf{X}, \mathbf{z}, \mathbf{H})$ is an *improper* or *energy based* model (LeCun et al., 2006) in the sense that it does not define a normalized distribution. An important consequence of working with an improper model is that the variational objective function is not a lower-bound any more of the log marignal $\ln p(\mathbf{X})$:

$$\ln p(\mathbf{X}) \ngeq \mathcal{L} = \langle \ln \frac{f(\mathbf{X}, \mathbf{z}, \mathbf{H})}{q(\mathbf{z}, \mathbf{H})} \rangle_{q(\mathbf{z}, \mathbf{H})}. \tag{4.45}$$

However, this is a minor concern as optimizing the right-hand side of (4.45) with respect to q leads to finding the variational posterior $q(\mathbf{z}, \mathbf{H})$ the closest (in KL divergence sense) to the true posterior $p(\mathbf{z}, \mathbf{H}|\mathbf{X})$ defined as

$$p(\mathbf{z}, \mathbf{H}|\mathbf{X}) = \frac{f(\mathbf{X}, \mathbf{z}, \mathbf{H})}{\sum_{\mathbf{z}} \int_{\mathbf{H}} f(\mathbf{X}, \mathbf{z}, \mathbf{H}) \mathrm{d}\mathbf{H}}. \tag{4.46}$$

To see this, let's consider the KL divergence between the variational and the true posteriors:

$$\mathrm{D}_{\mathrm{KL}}(q(\mathbf{z}, \mathbf{H}) \mid\mid p(\mathbf{z}, \mathbf{H}|\mathbf{X})) = \langle \ln q(\mathbf{z}, \mathbf{H}) - \ln p(\mathbf{z}, \mathbf{H}|\mathbf{X}) \rangle_{q(\mathbf{z}, \mathbf{H})}. \tag{4.47}$$

Injecting (4.46) in (4.47) and using the non-negativity of the KL divergence, we obtain the following lower-bound:

$$\ln\Big(\sum_{\mathbf{z}}\int_{\mathbf{H}} f(\mathbf{X},\mathbf{z},\mathbf{H})\mathrm{d}\mathbf{H}\Big) \geq \mathcal{L} = \langle \ln \frac{f(\mathbf{X},\mathbf{z},\mathbf{H})}{q(\mathbf{z},\mathbf{H})}\rangle_{q(\mathbf{z},\mathbf{H})}, \tag{4.48}$$

Where the right-hand side is identical to that of (4.45). Therefore, optimizing $\mathcal{L}$ minimizes the KL divergence in (4.47) which lead to a consistent estimate of the posterior of the energy based model.

If we now restrain our variational posterior to the following mean-field factorization (appendix A.2.2): $q(\mathbf{z},\mathbf{H}) = q(\mathbf{z})q(\mathbf{H})$ it is trivial to show that the optimal variational factors of the corrected model are given by:

$$\ln q^*(\mathbf{z}) = \frac{\alpha}{\beta}\langle \ln p(\mathbf{X}|\mathbf{z},\mathbf{H})\rangle_{q(\mathbf{H})} + \ln p(\mathbf{z}|\mathbf{H}) \tag{4.49}$$

$$\ln q^*(\mathbf{H}) = \alpha\langle \ln p(\mathbf{X}|\mathbf{z},\mathbf{H})\rangle_{q(\mathbf{z})} + \ln p(\mathbf{H}) \tag{4.50}$$

The exact update equations are easily obtained by scaling the sufficient statistics in (2.68), (2.84) and (2.82) yielding:

$$\ln q^*(\mathbf{z}) = \sum_{n=1}^{N} \frac{\alpha}{\beta}\phi_n(z_n) + A_{z_{n-1},z_n} + \text{const} \tag{4.51}$$

$$\implies q^*(\mathbf{z}) = \frac{1}{\zeta}\prod_{n=1}^{N} \exp\{\frac{\alpha}{\beta}\phi_n(z_n) + A_{z_{n-1},z_n}\} \tag{4.52}$$

$$\zeta = \sum_{\mathbf{z}}\prod_{n=1}^{N} \exp\{\frac{\alpha}{\beta}\phi_n(z_n) + A_{z_{n-1},z_n}\}, \tag{4.53}$$

and

$$\begin{aligned}\ln q^*(\mathbf{H}) = &\Big[\sum_{n=1}^{N}\langle \alpha \ln p(\mathbf{x}_n|\boldsymbol{\theta}_{z_n})p(c_n|\boldsymbol{\omega}_{z_n})\rangle_{q(c_n|z_n)q(z_n)}\Big] \\ &+ \sum_{i=1}^{\infty}\ln p(\boldsymbol{\omega}_i) + \sum_{j=1}^{C}\ln p(\boldsymbol{\theta}_i^j) + \text{const}\end{aligned} \tag{4.54}$$

$$\implies q^*(\mathbf{H}) = \prod_{i=1}^{\infty} q^*(\boldsymbol{\omega}_i)\prod_{j=1}^{C} q^*(\boldsymbol{\theta}_i^j) \tag{4.55}$$

$$q^*(\boldsymbol{\omega}_i) = \exp\{\boldsymbol{\xi}_i^\top T(\boldsymbol{\omega}_i) - A(\boldsymbol{\xi}_i)\} \tag{4.56}$$

$$\boldsymbol{\xi}_i = \boldsymbol{\xi}_0 + \sum_{n=1}^{N}\alpha q(z_n = i)\begin{bmatrix}T(c_n)\\ 1\end{bmatrix} \tag{4.57}$$

$$q^*(\boldsymbol{\theta}_i^j) = \exp\{\boldsymbol{\vartheta}_i^{jT} T(\boldsymbol{\theta}_i^j) - A(\boldsymbol{\vartheta}_i^j)\} \tag{4.58}$$

$$\boldsymbol{\vartheta}_i^j = \boldsymbol{\vartheta}_0 + \sum_{n=1}^{N}\alpha q(c_n = j|z_n = i)q(z_n = i)\begin{bmatrix}T(\mathbf{x}_n)\\ 1\end{bmatrix}. \tag{4.59}$$

Informally, the coefficients α and $\frac{\alpha}{\beta}$ weigh how much each observation should be "trusted". When the coefficients are lower than 1, the model will need more data to converge to the same posterior as the uncorrected model and vice versa.

Model	Features	Corpus	F-score	NMI (%)
DP-HMM	MFCC	TIMIT	63.25	35.11
HDP-HMM	MFCC	TIMIT	**64.08**	**35.82**
DP-SHMM	MFCC	TIMIT	**75.56**	39.14
HDP-SHMM	MFCC	TIMIT	75.42	**39.62**
DP-HMM	MFCC	MBOSHI	64.14	36.21
HDP-HMM	MFCC	MBOSHI	**65.47**	**36.53**
DP-SHMM	MFCC	MBOSHI	57.65	39.98
HDP-SHMM	MFCC	MBOSHI	**58.01**	**40.67**

Table 4.1: Comparison of the DP-(S)HMM and the HDP-(S)HMM on the AUD task.

4.4 Results

We now evaluate the HDP-HMM on the AUD task. In section 4.4.2, we measure the benefit of introducing a bigram phonotactic language model using the „natural“, i.e. uncorrected, model and in section 4.4.3, we analyze the effect of the correction factors using the corrected model.

4.4.1 Experimental Setup

Our experimental setup is similar to the one used in the previous chapters: we experimented on both the TIMIT and MBOSHI data sets using the MFCC features. Since we have shown in chapter 3 that the MBN features bring little to no improvement over the MFCCs, we did not use them in these experiments. The HDP based AUD system can be used either with HMM or SHMM as acoustic unit model. We refer to these variants as the HDP-HMM and HDP-SHMM respectively. As described in section 4.2, the HDP-(S)HMM requires a DP-(S)HMM to approximate the variational posterior of the root stick-breaking process and to initialize the variational posterior of the other parameters. For the HDP-SHMM, we used the DP-SHMM system pre-trained on the combination of 4 GLOBALPHONE languages (the „CB“ set described in chapter 3) and with a phonetic subspace of 100 dimensions. For both variants, we truncated the Dirichlet process to 100 acoustic units plus one extra „silence unit“.

4.4.2 Bigram vs unigram phonotactic language model

For our first experiment, we compared the performance of the DP-(S)HMM against the HDP-(S)HMM based AUD system. Results on the TIMIT and MBOSHI corpora are reported in Table 4.1. We observe the HDP prior provides a small but consistent improvement over the DP-(S)HMM in terms of clustering quality (measured with the NMI). The quality of the segmentation (F-score) slightly improves as well except for the case of the HDP-SHMM on TIMIT where we observe a slight degradation of the F-score. Overall, we see that the HDP prior improves the AUD task even without any correction factors.

4.4.3 Effect of the correction factors

We now analyze the effect of the acoustic and language model correction factors α and β. We have restricted our analysis on the HDP-(S)HMM on the TIMIT corpus. The NMI as

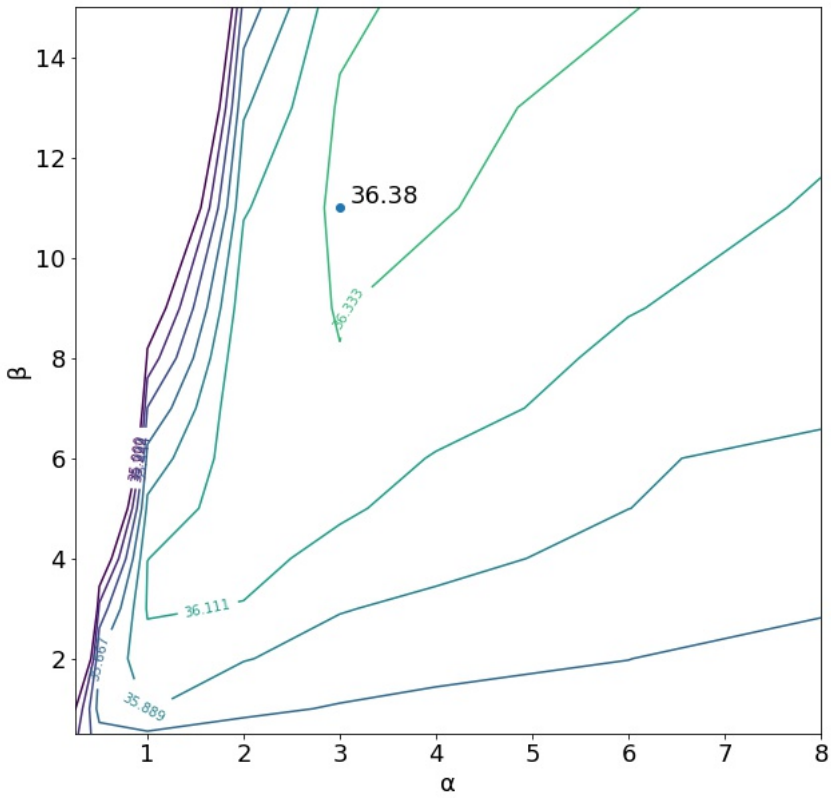


Figure 4.2: NMI as a function of the acoustic (α) and language model (β) scaling factors for the corrected HDP-HMM on the TIMIT corpus. The blue dot indicates the maximum value.

a function of α and β is depicted in Fig. 4.2 for the DP-HMM and in Fig. 4.3 for the DP-SHMM. For both models, we observe the same behavior: increasing both the acoustic and language model scaling factors improves the clustering quality. Note that to achieve better results, the language model scaling factor should remain greater than the acoustic scaling factor. This behavior is rather expected: when β is greater than α, the ratio in (4.49) will be lower than one and, consequently, will decrease the effect of the acoustic model and give more importance to the language model.

Finally, the comparison between the corrected and uncorrected DP-(S)HMM models is shown in Table 4.2. The correction factors provide a significant gain and help to fully benefit from the bigram phonotactic language model. This experiment somewhat biased as we have used the reference transcription to tune the scaling factors to achieve better results. Nevertheless, we see that the optimal factors are the same for both models and—as was observed in ASR—once these factors are tuned on a data set, they generalize well on other corpora.

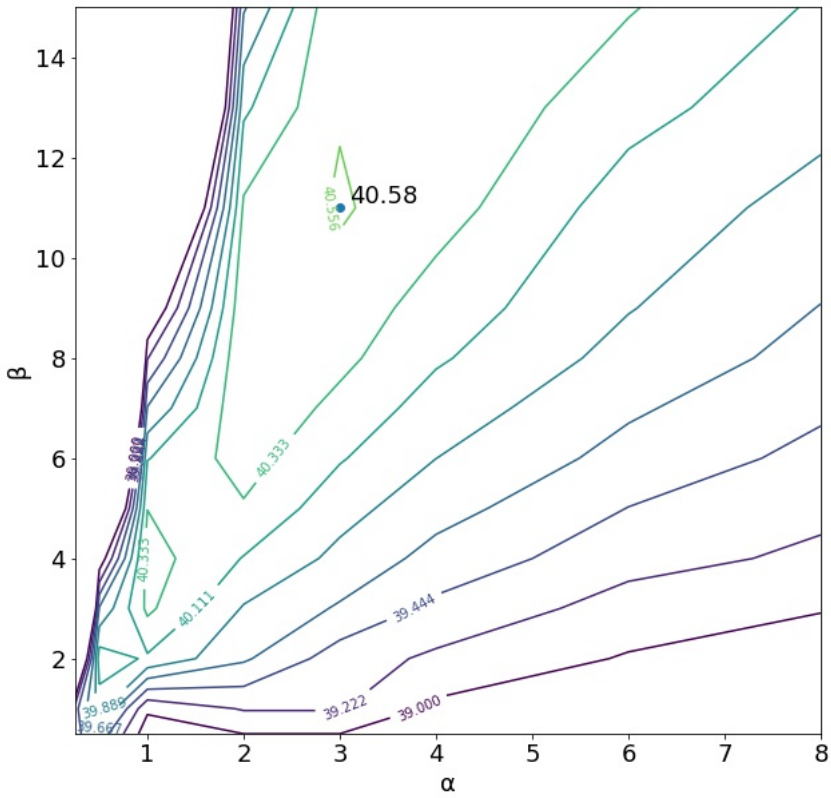


Figure 4.3: NMI as a function of the acoustic (α) and language model (β) scaling factors for the corrected HDP-SHMM on the TIMIT corpus. The blue dot indicates the maximum value.

4.5 Conclusion

In this chapter, we have empowered our AUD system with a bigram phonotactic language model. Our approach relies on the Hierarchical Dirichlet Process: a non-parametric prior over conditional distributions. Replacing the Dirichlet Process by a Hierarchical Dirichlet Process only affects the language model and, therefore, the HDP prior can be used with either the HMM or SHMM based AUD system. We have studied the case of a bigram language model but it is theoretically possible to extend this work to arbitrary n-gram language models. Similarly to the original DP-HMM, this model is trained with a VB-EM algorithm. This is possible thanks to the Teh's construction of the HDP, a hierarchical stick-breaking process. Unfortunately, the Teh's stick-breaking process is not fully conjugate and, therefore, it is difficult to derive the optimal posterior of the parameters of HDP's root level. We bypass this issue by approximating this posterior with the posterior of an unigram DP-HMM. This approximation is very convenient but can also trap our model in a local optimum. This issue could be solved using the Sethuraman stick-breaking process but would considerably increase the computational cost. Experimental results show that

Model	Features	Corpus	α	β	NMI (%)
DP-HMM	MFCC	TIMIT	-	-	35.11
HDP-HMM	MFCC	TIMIT	1.0	1.0	35.82
HDP-HMM	MFCC	TIMIT	3.0	11.0	**36.38**
DP-SHMM	MFCC	TIMIT	-	-	39.14
HDP-SHMM	MFCC	TIMIT	1.0	1.0	39.62
HDP-SHMM	MFCC	TIMIT	3.0	11.0	**40.58**

Table 4.2: Comparison of the best corrected HDP-(S)HMM model against the uncorrected HDP-(S)HMM ($\alpha = 1$ and $\beta = 1$) and the DP-(S)HMM.

the HDP prior gives a small but consistent improvement for the HMM and SHMM based AUD system on both TIMIT and MBOSHI corpora.

Furthermore, we have shown that the HDP-HMM model can be augmented with acoustic and language model factors that weigh the importance of acoustic and language model in the likelihood function. These factors turn the AUD phone-loop model into an energy based model. Nevertheless, we show that optimizing the variational lower-bound of this energy-based model still leads to a consistent estimate of the variational posterior. Our experiments show that, for suitable choice of correction factors, the „corrected“ HDP-HMM achieves better clustering measured in terms of NMI. The segmentation quality however does not seem to benefit from such model correction.

Chapter 5

Conclusion

In the previous chapters, we have proposed several models to address the problem of learning a phonological system from speech. All these models rely upon a Bayesian formulation of the task. With the use of Variational Bayes framework, we have seen that learning the acoustic units, i.e. the phonological system, can be achieved through the optimization of a well-defined objective function. Before summarizing the contributions of this thesis, we briefly discuss potential extensions and promising trends for the unsupervised speech learning research, including new phonetic acoustic model and non-parametric Bayesian neural network.

5.1 Future work

Let us discuss what are, in my opinion, the promising research directions emerging from this thesis. We have seen that the Bayesian formulation of the AUD task leads to the definition of four essential elements:

- acoustic model
- language model
- prior over the language model
- prior over the acoustic model parameter (the base measure in the context of the Dirichlet Process)

Importantly, this formulation is very generic and does not imply any specific model. The choice to use the HMM and the Dirichlet Process was mostly driven by historical reasons and mathematical convenience rather than by a strong belief that they are ideal tools for the task. I believe that significant progress can be made in the field of unsupervised learning of speech by revisiting these "old" models in light of the recent development of the research on Bayesian generative models. In the following, I propose alternative models which could lead to significant improvements.

5.1.1 Acoustic Modeling

The 3-state HMM model remains *de facto* the state-of-the-art generative model for a phonetic unit in speech technologies. Yet, it is widely accepted that the observations independence assumption following from this model is unrealistic and leads to poor modeling

capability. This issue is not dramatic in speech recognition since the language model can compensate for an inaccurate acoustic model. However, in the case of AUD, proper segmentation and clustering of the speech largely depends on the quality of the acoustic model.

A simple way to improve the HMM is by making an observation to depend on the hidden state and on previous observations. This model, called an autoregressive HMM, was recently introduced in Bryan and Levinson (2015). The time dependency between observations does have a cost: the inference requires to compute the autocorrelation function of the input signal. Nevertheless, modern hardware largely allows to perform this computation. Note that in Bryan and Levinson (2015), the authors model raw speech signal which is perhaps unsuited for tAUD. Applying the ARHMM directly on the short term (Mel) spectrum would be, in my opinion, more practical. Interestingly, doing so would lead to model the amplitude and frequency modulations of the speech signal which would be consistent with psychoacoutics studies Elhilali et al. (2003).

Alternatively, rather than changing the HMM, one could transform the features such that they fit better the HMM assumption. This paradigm was the core idea of a recent model: the VAE-HMM Ebbers et al. (2017); Glarner et al. (2018). It is a promising approach as it makes use of neural network to define the generative model. However, the introduction of arbitrarily complex model comes with a downside: whereas it is fairly easy to use gradient ascent to train such a model, it is much more difficult to prevent the model from falling in a local optimum. Also, increasing the model's complexity increases the necessary amount of data which may be problematic when dealing with low-resources languages. Having a neural network-based AUD system is a compelling idea but it remains currently an open problem.

This work has also shown the importance of the acoustic model prior for the outcome of the AUD sytem. The GSM defined in chapter 3 is general enough to accommodate a large family of acoustic models, including the ones mentioned above, but can be extended in several ways. For instance, the SHMM is based on an affine and non-linear transformation. We can envision a deep SHMM where the non-linearity would be learned by a neural network. Another potential improvement of the GSM is the introduction of multiple subspaces. These extra subspaces could either:

- include non-phonetic factors such as speaker variability
- decompose the phonetic subspace to better model linguistic features (for instance there could be separated subspaces for vowels and consonants).

Lastly, let me mention a recent work on the factorization of subspace model Novotny et al. (2019). This line or work is particularly interesting as it could be used in the SHMM to model the language variability.

5.1.2 Language Modeling

A large par of the progress in unsupervised speech learning, including this thesis, is due to the development of Bayesian non-parametric priors. The Dirichlet process and its natural extension the Pitman-Yor process offer a well-grounded framework to define probability distributions over countably infinite sets. But after almost two decades of research, these

tools have also shown their limits. Even though the construction of hierarchical Dirichlet or Pitman-Yor processes is theoretically straightforward, variational inference in such models is nearly intractable for any hierarchy having more than 2 levels. On the other hand, samplers like the Chinese restaurant process can work with arbitrary deep models at the cost of very slow inference and exponential growth of the parameters. Finally, empirical experience has shown that neural network-based language models are far superior to n-gram based language models. All these issues, clearly call for an extension of the non-parametric priors to a much broader class of models.

Defining non-parametric Bayesian priors for neural network based language model may seem a rather difficult task but recent advances in machine learning lean toward this direction. A promising step is the newly introduced Logistic Stick-Breaking Process Ren et al. (2011). This non-parametric prior is defined a spatial stick-breaking process whose parameters are Euclidean embeddings. This is particularly interesting as such embeddings could be the output of a neural network. Another work worth mentioning is Gal (2016) where the authors show how the dropout technique can be reinterpreted as an approximate Bayesian inference. Importantly, they also show how one could get an uncertainty estimate without any significant change in the neural network. Combining both the Logistic Stick-Breaking Process with a Bayesian neural network is a very compelling idea and could pave the way to more powerful non-parametric priors useful for unsupervised speech learning and many other fields.

5.2 Summary of contributions

The aim of this thesis has been to develop a Bayesian approach to the problem of learning a phonological system, i.e. an ensemble of acoustic units, used to communicate in a language, from unlabeled speech recordings. This work can be seen as the extension and the continuation of previous works on non-parametric Bayesian learning applied to language modeling Goldwater and Johnson (2007) and acoustic unit discovery Lee (2014).

In Lee and Glass (2012) the authors proposed a non-parametric Bayesian HMM to cluster unlabeled speech into phone-like units; they used the Chinese Restaurant Process to sample the parameters from the posterior distribution. In chapter 2, we derived a new inference scheme based on the Variational Bayes framework. It allows to cast the problem of discovering acoustic units into an optimization problem with a well-defined objective function. Our approach relies upon Sethuraman stick-breaking construction of the Dirichlet Process which, combined with a suitable structured mean-field factorization of the variational posterior, leads to an analytical VB-EM algorithm. Moreover, this new approach allows for the reinterpretation of the original model as an infinite phone-loop model capable of fast and parallelized inference. The computational benefits from this approach are important as they allow learning phonological units from a large speech corpus. We found experimentally that Variational Bayes training leads to sparser solution, i.e. the model uses less acoustic units to explain the data, and yet achieves better clustering quality in terms of NMI.

In chapter 3, we addressed the issue of how to design a proper prior distribution over the possible acoustic unit embeddings. We first introduced the *Generalized Subspace Model* (GSM): a theoretical framework which allows learning low-dimensional embeddings rep-

resenting probability distributions. The GSM is a natural extension of several existing models, such as the i-vector model or the Subspace Multinomial model, to any conditionally conjugate exponential models (GMM, HMM, PCA, ...). In a controlled experiment, we have shown that the GSM is able to learn a coherent phonetic subspace where the phones, modeled by an HMM, are encoded as 100-dimensional embeddings. Finally, we used the GSM framework to learn a universal phonetic subspace from a multilingual labeled speech corpus. This universal phonetic subspace is then used as the base measure of the Dirichlet Process of our acoustic unit discovery system. By estimating the prior over acoustic units from other languages, we are effectively changing the learning procedure: informally, instead of directly clustering unlabeled speech, we first use supervision from other languages to teach the model the notion of "phone" and then, the model clusters speech from a target language into patterns similar to the phones from other languages. Experimental results have shown the merit of this new approach: the GSM based AUD model achieved much better segmentation and clustering quality than the original non-parametric HMM model. The results also show that the GSM approach is more robust than using multilingual features as an input to the AUD system. This is a strong indication that the GSM is a more principled way to transfer phonetic knowledge from a language to another.

In chapter 4, we developed a new AUD model based on the *Hierarchical Dirichlet Process* (HDP). We coined this new model the HDP-HMM. The HDP is a non-parametric prior which defines a probability over an infinite set of conditional distributions. Thanks to this feature, we built an AUD model based on a bigram phonotactic language model. This is a substantial change compared to the DP-HMM, which can have only a unigram phonotactic language model. To infer the parameters of this new model we derived a VB-EM algorithm based on the Teh's stick-breaking construction of the HDP. As the HDP prior only affects the distribution of the units' labels, the training of the acoustic model is nearly identical to the VB-EM of the DP-HMM model. This key feature allows us to use the HDP prior seamlessly with the HMM or SHMM acoustic models. Teh's stick-breaking construction is particularly convenient since it expresses the sampled conditional distributions directly with the atoms generated by the root base measure and therefore avoids any ordering issue. However, it has the downside that it is not fully conditionally conjugate. Consequently, our training requires first to train a DP-HMM AUD model to estimate the variational posterior of the root stick-breaking process. Experimental results show that the HDP-HMM model applied to the AUD task provides a small but consistent gain over the DP-HMM in terms of clustering quality and segmentation. Moreover, we show that the model can be corrected using two factors weighing the contrinution of the acoustic and language models in the joint probability distribution of the model. We observed empirically that giving more importance to the language model (increasing the language model factor) results in a better NMI.

To conclude, we hope that this thesis has provided an accessible study of Bayesian approaches to the problem of learning a phonological system from speech. We have developed a probabilistic formulation of the task and proposed several models to fulfill it. Altogether, this forms a well-grounded framework, which paves the way to many more models than the ones investigated in the previous chapters. We hope that this thesis will stimulate future research on the challenging problem of unsupervised speech learning.

Bibliography

L. Baum. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a markov process. *Inequalities*, 3:1–8, 1972.

M. J. Beal, Z. Ghahramani, and C. E. Rasmussen. The infinite hidden markov model. In *Advances in neural information processing systems*, pages 577–584, 2002.

M. Bi, Y. Qian, and K. Yu. Very deep convolutional neural networks for lvcsr. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

D. M. Blei. Variational methods for the dirichlet process. In *In Proceedings of the Twenty-First International Conference on Machine Learning (ICML 2004*, 2004.

D. M. Blei, M. I. Jordan, et al. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143, 2006.

H. Bourlard. Reconnaissance automatique de la parole: modélisation ou description. *Journées Etude Parole'96*, pages 263–272, 1996.

J. D. Bryan and S. E. Levinson. Autoregressive hidden markov model and the speech signal. *Procedia Computer Science*, 61:328–333, 2015.

L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, et al. Multilingual acoustic modeling for speech recognition based on subspace gaussian mixture models. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4334–4337. IEEE, 2010.

J. Chorowski, R. J. Weiss, S. Bengio, and A. van den Oord. Unsupervised speech representation learning using wavenet autoencoders. *IEEE/ACM transactions on audio, speech, and language processing*, 27(12):2041–2053, 2019.

J. R. Cohen. Segmenting speech using dynamic programming. *The Journal of the Acoustical Society of America*, 69(5):1430–1438, 1981.

E. David and O. Selfridge. Eyes and ears for computers. *Proceedings of the IRE*, 50(5): 1093–1101, 1962.

K. H. Davis, R. Biddulph, and S. Balashek. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, 24(6):637–642, 1952.

S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

N. Dehak, R. Dehak, P. Kenny, N. Brümmer, P. Ouellet, and P. Dumouchel. Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification. In *Tenth Annual conference of the international speech communication association*, 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

M. Diez, L. Burget, F. Landini, and J. Černockỳ. Analysis of speaker diarization based on bayesian hmm with eigenvoice priors. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:355–368, 2019.

J. F. Drexler. *Deep unsupervised learning from speech.* PhD thesis, Massachusetts Institute of Technology, 2016.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(Jul):2121–2159, 2011.

E. Dunbar, G. Synnaeve, and M. V. Emmanuel. A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling.

E. Dunbar, R. Algayres, J. Karadayi, M. Bernard, J. Benjumea, X.-N. Cao, L. Miskic, C. Dugrain, L. Ondel, A. W. Black, et al. The zero resource speech challenge 2019: Tts without t. *arXiv preprint arXiv:1904.11469*, 2019.

E. Dupoux. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59, 2018.

J. Ebbers, J. Heymann, L. Drude, T. Glarner, R. Haeb-Umbach, and B. Raj. Hidden markov model variational autoencoder for acoustic unit discovery. In *INTERSPEECH*, pages 488–492, 2017.

Eberhard, David M., Gary F. Simons, and Charles D. Fennig. Ethnologue: Languages of the world. twenty-third edition., 2020. URL http://www.ethnologue.com.

M. Elhilali, T. Chi, and S. A. Shamma. A spectro-temporal modulation index (stmi) for assessment of speech intelligibility. *Speech communication*, 41(2-3):331–348, 2003.

T. S. Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

S. Fine, Y. Singer, and N. Tishby. The hierarchical hidden markov model: Analysis and applications. *Machine learning*, 32(1):41–62, 1998.

E. B. Fox, E. B. Sudderth, M. I. Jordan, A. S. Willsky, et al. A sticky hdp-hmm with application to speaker diarization. *The Annals of Applied Statistics*, 5(2A):1020–1056, 2011.

R. Fér, P. Matějka, F. Grézl, O. Plchot, K. Veselỳ, and J. H. Černockỳ. Multilingually trained bottleneck features in spoken language recognition. *Computer Speech & Language*, 46:252–267, 2017.

Y. Gal. Uncertainty in deep learning. *University of Cambridge*, 1:3, 2016.

A. Garcia and H. Gish. Keyword spotting of arbitrary words using minimal speech resources. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 1, pages I–I. IEEE, 2006.

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, 1990.

T. Glarner, P. Hanebrink, J. Ebbers, and R. Haeb-Umbach. Full bayesian hidden markov model variational autoencoder for acoustic unit discovery. In *Interspeech*, pages 2688–2692, 2018.

J. Glass. Towards unsupervised speech processing. In *2012 11th International Conference on Information Science, Signal Processing and their Applications (ISSPA)*, pages 1–4. IEEE, 2012.

P. Godard, G. Adda, M. Adda-Decker, J. Benjumea, L. Besacier, J. Cooper-Leavitt, G.-N. Kouarata, L. Lamel, H. Maynard, M. Müller, et al. A very low resource language speech corpus for computational language documentation experiments. *arXiv preprint arXiv:1710.03501*, 2017.

S. Goldwater, M. Johnson, and T. L. Griffiths. Interpolating between types and tokens by estimating power-law generators. In *Advances in neural information processing systems*, pages 459–466, 2006.

S. Goldwater, T. L. Griffiths, and M. Johnson. A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54, 2009.

S. J. Goldwater and M. Johnson. *Nonparametric Bayesian Models of Lexical Acquisition*. Brown University, 2007.

E. B. Gouvêa. Acoustic-feature-based frequency warping for speaker normalization. *P. Hd. dessertation Carnegie Mellon University*, 1998.

A. Graves and N. Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772, 2014.

A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376, 2006.

K. J. Han, A. Chandrashekaran, J. Kim, and I. Lane. The capio 2017 conversational speech recognition system. *arXiv preprint arXiv:1801.00059*, 2017.

D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

D. Harwath, G. Chuang, and J. Glass. Vision as an interlingua: Learning multilingual semantic embeddings of untranscribed speech. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4969–4973. IEEE, 2018.

H. Hermansky. Perceptual linear predictive (plp) analysis of speech. *the Journal of the Acoustical Society of America*, 87(4):1738–1752, 1990.

N. Holzenberger, S. Palaskar, P. Madhyastha, F. Metze, and R. Arora. Learning from multiview correlations in open-domain videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8628–8632. IEEE, 2019.

W.-N. Hsu and J. Glass. Scalable factorized hierarchical variational autoencoder training. *arXiv preprint arXiv:1804.03201*, 2018.

W.-N. Hsu, Y. Zhang, and J. Glass. Learning latent representations for speech generation and transformation. *arXiv preprint arXiv:1704.04222*, 2017.

M. Hughes, D. I. Kim, and E. Sudderth. Reliable and scalable variational inference for the hierarchical dirichlet process. In *Artificial Intelligence and Statistics*, pages 370–378, 2015.

S. Ioffe. Probabilistic linear discriminant analysis. In *European Conference on Computer Vision*, pages 531–542. Springer, 2006.

H. Ishwaran and L. F. James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

A. Jansen and B. Van Durme. Efficient spoken term discovery using randomized algorithms. In *2011 IEEE Workshop on Automatic Speech Recognition & Understanding*, pages 401–406. IEEE, 2011.

A. Jansen, K. Church, and H. Hermansky. Towards spoken term discovery at scale with zero resources. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

F. Jelinek. Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4):532–556, 1976.

M. J. Johnson, D. K. Duvenaud, A. Wiltschko, R. P. Adams, and S. 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.

H. Kamper. Unsupervised neural and bayesian models for zero-resource speech processing. *arXiv preprint arXiv:1701.00851*, 2017.

H. Kamper, A. Jansen, and S. Goldwater. Fully unsupervised small-vocabulary speech recognition using a segmental bayesian model. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

H. Kamper, A. Jansen, and S. Goldwater. Unsupervised word segmentation and lexicon discovery using acoustic word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(4):669–679, 2016.

H. Kamper, A. Jansen, and S. Goldwater. A segmental framework for fully-unsupervised large-vocabulary speech recognition. *Computer Speech & Language*, 46:154–174, 2017a.

H. Kamper, K. Livescu, and S. Goldwater. An embedded segmental k-means model for unsupervised segmentation and clustering of speech. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 719–726. IEEE, 2017b.

P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(4):1435–1447, 2007.

S. Kesiraju, R. Pappagari, L. Ondel, L. Burget, N. Dehak, S. Khudanpur, J. Černockỳ, and S. V. Gangashetty. Topic identification of spoken documents using unsupervised acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5745–5749. IEEE, 2017.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

C.-H. Lee, F. K. Soong, and B.-H. Juang. A segment model based approach to speech recognition. In *ICASSP-88., International Conference on Acoustics, Speech, and Signal Processing*, pages 501–541. IEEE, 1988.

C.-y. Lee. *Discovering linguistic structures in speech: Models and applications*. PhD thesis, Massachusetts Institute of Technology, 2014.

C.-y. Lee and J. 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.

C.-y. Lee, T. J. O'donnell, and J. Glass. Unsupervised lexicon discovery from acoustic input. *Transactions of the Association for Computational Linguistics*, 3:389–403, 2015.

K.-F. Lee and H.-W. Hon. Speaker-independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 37(11): 1641–1648, 1989.

C. J. Leggetter and P. C. Woodland. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer speech & language*, 9(2):171–185, 1995.

S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pages 914–922, 2017.

C. Liu, J. Yang, M. Sun, S. Kesiraju, A. Rott, L. Ondel, P. Ghahremani, N. Dehak, L. Burget, and S. 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.

C. Lopes and F. Perdigao. Phoneme recognition on the timit database. In *Speech technologies*. IntechOpen, 2011.

L. v. d. Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

D. J. MacKay and L. C. B. Peto. A hierarchical dirichlet language model. *Natural language engineering*, 1(3):289–308, 1995.

C. Maddison, A. Mnih, and Y. Teh. The concrete distribution: A continuous relaxation of discrete random variables. International Conference on Learning Representations, 2017.

D. Merkx, S. L. Frank, and M. Ernestus. Language learning using speech to image retrieval. *arXiv preprint arXiv:1909.03795*, 2019.

B. Milde and C. Biemann. Unspeech: Unsupervised speech context embeddings. *arXiv preprint arXiv:1804.06775*, 2018.

T. Minka et al. Divergence measures and message passing. Technical report, Technical report, Microsoft Research, 2005.

D. Mochihashi, T. Yamada, and N. Ueda. Bayesian unsupervised word segmentation with nested pitman-yor language modeling. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 100–108. Association for Computational Linguistics, 2009.

K. P. Murphy and M. A. Paskin. Linear-time inference in hierarchical hmms. In *Advances in neural information processing systems*, pages 833–840, 2002.

O. Novotny, O. Plchot, O. Glembek, and L. Burget. Factorization of discriminatively trained i-vector extractor for speaker recognition. *arXiv preprint arXiv:1904.04235*, 2019.

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

L. Ondel, L. Burget, J. Černockỳ, and S. Kesiraju. Bayesian phonotactic language model for acoustic unit discovery. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5750–5754. IEEE, 2017.

L. Ondel, P. Godard, L. Besacier, E. Larsen, M. Hasegawa-Johnson, O. Scharenborg, E. Dupoux, L. Burget, F. Yvon, and S. 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.

P. Orbanz and Y. W. Teh. Bayesian nonparametric models. *Encyclopedia of machine learning*, (1), 2010.

A. Park and J. R. Glass. Towards unsupervised pattern discovery in speech. In *IEEE Workshop on Automatic Speech Recognition and Understanding, 2005.*, pages 53–58. IEEE, 2005.

D. Povey, L. Burget, M. Agarwal, P. Akyazi, F. Kai, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, A. Rastrow, et al. The subspace gaussian mixture model—a structured model for speech recognition. *Computer Speech & Language*, 25(2):404–439, 2011.

S. J. Prince and J. H. Elder. Probabilistic linear discriminant analysis for inferences about identity. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

Y. Qian, M. Bi, T. Tan, and K. Yu. Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(12):2263–2276, 2016.

L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

C. E. Rasmussen. The infinite gaussian mixture model. In *Advances in neural information processing systems*, pages 554–560, 2000.

L. Ren, L. Du, L. Carin, and D. Dunson. Logistic stick-breaking process. *Journal of Machine Learning Research*, 12(Jan):203–239, 2011.

D. J. Rezende and S. Mohamed. Variational inference with normalizing flows. *arXiv preprint arXiv:1505.05770*, 2015.

H. Sak, A. W. Senior, and F. Beaufays. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. 2014.

H. Sak, A. Senior, K. Rao, and F. Beaufays. Fast and accurate recurrent neural network acoustic models for speech recognition. *arXiv preprint arXiv:1507.06947*, 2015.

G. Saon, H.-K. J. Kuo, S. Rennie, and M. Picheny. The ibm 2015 english conversational telephone speech recognition system. *arXiv preprint arXiv:1505.05899*, 2015.

T. Schultz. Globalphone: a multilingual speech and text database developed at karlsruhe university. In *Seventh International Conference on Spoken Language Processing*, 2002.

T. Sercu, C. Puhrsch, B. Kingsbury, and Y. LeCun. Very deep multilingual convolutional neural networks for lvcsr. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4955–4959. IEEE, 2016.

J. Sethuraman. A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650, 1994.

S. H. Shum, D. F. Harwath, N. Dehak, and J. R. Glass. On the use of acoustic unit discovery for language recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(9):1665–1676, 2016.

A. Stolcke and J. Droppo. Comparing human and machine errors in conversational speech transcription. *arXiv preprint arXiv:1708.08615*, 2017.

A. Stolcke, L. Ferrer, S. Kajarekar, E. Shriberg, and A. Venkataraman. Mllr transforms as features in speaker recognition. In *Ninth European Conference on Speech Communication and Technology*, 2005.

I. Szöke, F. Grézl, J. Cernockỳ, M. Fapšo, and T. Cipr. Acoustic keyword spotter-optimization from end-user perspective. In *2010 IEEE Spoken Language Technology Workshop*, pages 189–193. IEEE, 2010.

Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes (technical report 653). *UC Berkeley Statistics*, 2004.

Y. W. Teh. A hierarchical bayesian language model based on pitman-yor processes. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 985–992. Association for Computational Linguistics, 2006.

Y. W. Teh. Dirichlet process. *Encyclopedia of machine learning*, pages 280–287, 2010.

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

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

A. van den Oord, O. Vinyals, et al. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315, 2017.

B. Varadarajan, S. Khudanpur, and E. Dupoux. Unsupervised learning of acoustic sub-word units. In *Proceedings of ACL-08: HLT, Short Papers*, pages 165–168, 2008.

J. Černocký. *Speech Processing Using Automatically Derived Segmental Units: Applications to Very Low Rate Coding and Speaker Verification*. PhD thesis, Université Paris XI Orsay, Dec. 1998.

C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical dirichlet process. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 752–760, 2011.

S. Wegmann, D. McAllaster, J. Orloff, and B. Peskin. Speaker normalization on conversational telephone speech. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pages 339–341. IEEE, 1996.

W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. The microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5934–5938. IEEE, 2018.

D. Yu, W. Xiong, J. Droppo, A. Stolcke, G. Ye, J. Li, and G. Zweig. Deep convolutional neural networks with layer-wise context expansion and attention. In *Interspeech*, pages 17–21, 2016.

V. Zue, S. Seneff, and J. Glass. Speech database development at mit: Timit and beyond. *Speech communication*, 9(4):351–356, 1993.

Appendix A

Variational Bayes

In this appendix, we give a brief introduction to the Variational Bayes framework for inference. First, we define the variational objective function and then we summarize the main approaches to optimize this objective.

A.1 Variational Bayes objective

Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ be observed data, $p(\mathbf{X}|\boldsymbol{\theta})$ a parametric likelihood distribution and $p(\boldsymbol{\theta})$ a prior. We aim to find a variational objective function whose optimum is given by the posterior distribution:

$$p(\boldsymbol{\theta}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{X})}, \tag{A.1}$$

where $p(\mathbf{X}) = \int_{\boldsymbol{\theta}} p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. Let $q(\boldsymbol{\theta})$ be any distribution over $\boldsymbol{\theta}$. Estimating the posterior in (A.1) amounts to solve the following minimization problem:

$$p(\mathbf{X}|\boldsymbol{\theta}) = q^*(\boldsymbol{\theta}) = \arg\min_q \mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})). \tag{A.2}$$

Expanding the right hand side of (A.2) and using the fact that $\mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})) \geq 0$, we have:

$$\mathrm{D}_{\mathrm{KL}}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\mathbf{X})) = \langle \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{X})} \rangle_q \tag{A.3}$$

$$= \langle \ln q(\boldsymbol{\theta}) - \ln p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \rangle_q + \ln p(\mathbf{X}) \tag{A.4}$$

$$\implies \ln p(X) \geq \langle \ln \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \rangle_q = \mathcal{L}[q]. \tag{A.5}$$

$\mathcal{L}[q]$ is the variational objective function and it is often referred to as Evidence Lower-BOund (ELBO).

To conclude this brief definition of the variational objective, we show that the distribution $q^*(\boldsymbol{\theta})$ which maximizes $\mathcal{L}[q]$ is the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X})$.

Proof. Our proof is done in 2 steps: first we show that $\mathcal{L}[q]$ is concave in q, then we show that $p(\boldsymbol{\theta}|\mathbf{X})$ is a critical point of the objective function.

$\mathcal{L}[q]$ is concave in q: $\forall \lambda \in (0,1)$, $\forall f(\boldsymbol{\theta})$ we have:

$$\mathcal{L}[\lambda q + (1-\lambda) f] = \big\langle \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \big\rangle_{\lambda q + (1-\lambda) f} \tag{A.6}$$

$$- \big\langle \ln \big(\lambda q(\boldsymbol{\theta}) + (1-\lambda) f(\boldsymbol{\theta})\big) \big\rangle_{\lambda q + (1-\lambda) f} \tag{A.7}$$

$$\lambda \mathcal{L}[q] + (1-\lambda)\mathcal{L}[f] = \big\langle \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) \big\rangle_{\lambda q + (1-\lambda) f} \tag{A.8}$$

$$- \Big(\big\langle \lambda \ln q(\boldsymbol{\theta}) \big\rangle_q + \big\langle (1-\lambda) \ln f(\boldsymbol{\theta}) \big\rangle_f \Big). \tag{A.9}$$

Since $x \ln x$ is a convex function, we have:

$$\big\langle \lambda \ln q(\boldsymbol{\theta}) \big\rangle_q + \big\langle (1-\lambda) \ln f(\boldsymbol{\theta}) \big\rangle_f \geq \big\langle \ln \big(\lambda q(\boldsymbol{\theta}) + (1-\lambda) f(\boldsymbol{\theta})\big) \big\rangle_{\lambda q + (1-\lambda) f} \tag{A.10}$$

$$\implies \mathcal{L}[\lambda q + (1-\lambda) f] \geq \lambda \mathcal{L}[q] + (1-\lambda)\mathcal{L}[f], \tag{A.11}$$

which proves that $\mathcal{L}$ is concave in q.

Critical point of $\mathcal{L}[q]$: We now find the critical points of $\mathcal{L}[q]$ subject to the constraint $\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 = 0$. To do so, we define the Lagrangian:

$$\mathcal{L}'[q] = \mathcal{L}[q] + \nu \Big(\int q(\boldsymbol{\theta}) d\boldsymbol{\theta} - 1 \Big), \tag{A.12}$$

where ν is the Lagrange multiplier. Its functional derivative is given by:

$$\delta \mathcal{L}'[q] = \int \delta(\boldsymbol{\theta}) \big(\ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - \ln q(\boldsymbol{\theta}) - 1 + \nu \big) d\boldsymbol{\theta} \tag{A.13}$$

Using the fundamental lemma of calculus:

$$\int \delta(x) f(x) dx = 0 \tag{A.14}$$

$$\implies f(x) = 0 \quad \forall x \tag{A.15}$$

we get:

$$\delta \mathcal{L}'[q] = 0 \tag{A.16}$$

$$\implies \ln q^*(\boldsymbol{\theta}) = \ln p(\mathbf{X}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}) - 1 - \nu \tag{A.17}$$

$$\implies q^*(\boldsymbol{\theta}) = \frac{p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{Z} \tag{A.18}$$

$$Z = \exp\{1 + \nu\} = \int p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{A.19}$$

which, together with (A.11), proves that $p(\boldsymbol{\theta}|\mathbf{X})$ is the distribution which maximizes the variational lower-bound. □

A.2 Approximating posterior distributions

In most applications, the optimal $q^*(\boldsymbol{\theta})$ cannot be calculated as the integral in (A.19) is intractable. Nevertheless, the Variational Bayes can be used to find an approximate posterior distribution which is the „best approximation“ of the true posterior distribution in the KL divergence sense. To do so, one performs a constrained optimization of the variational lower-bound $\mathcal{L}[q]$ where the constraints are chosen to allow for a close form estimation of the approximate posterior. We review the main strategies used with Variational Bayes.

A.2.1 Parametric approximation

The simplest approximation is to constrain the variational posterior to be of a known parametric type. For instance, we can set:

$$q(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{A.20}$$

Then, maximizing the variational lower-bound reduces to standard calculus $\mathcal{L}[q] = \mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and can be done with (stochastic) gradient ascent.

A.2.2 Mean-Field approximation

An alternative to the parametric approximation is to assume a specific factorization of the variational posterior. The mean-field approximation corresponds to a fully factorized posterior. Let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$, under the mean-field approximation the variational posterior can be expressed as:

$$q(\boldsymbol{\theta}) = \prod_{k=1}^{K} q(\boldsymbol{\theta}_k). \tag{A.21}$$

In this case, optimizing $\mathcal{L}[q]$ amounts to iteratively solve K sub-objective functions given by:

$$\mathcal{L}_k[q] = \langle \ln \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_k)} \rangle_{q(\boldsymbol{\theta}_k)} \tag{A.22}$$

Using calculus of variations, the optimal variational posterior q_k^* is given by:

$$q^*(\boldsymbol{\theta}) = \arg\max_q \mathcal{L}_k[q] \tag{A.23}$$

$$\implies q^*(\boldsymbol{\theta}_k) \propto \exp\{\langle \ln p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta}_{\setminus k})}\} \tag{A.24}$$

where $q(\boldsymbol{\theta}_{\setminus k})$ is the product of all the variational factor but $q(\boldsymbol{\theta}_k)$.

A.2.3 Structured mean-field approximation

The mean-field approximation considerably simplifies the optimization of the variational lower-bound but it fails to capture any correlations between the sets of parameters. The structured mean-field approximation is a variant of the mean-field approximation which preserves some dependencies in the variational posterior. For instance, if $\boldsymbol{\theta} = \{\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3\}$, a possible structured mean-field factorization is:

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\theta}_1)q(\boldsymbol{\theta}_2|\boldsymbol{\theta}_3)q(\boldsymbol{\theta}_3), \tag{A.25}$$

where the dependency between $\boldsymbol{\theta}_2$ and $\boldsymbol{\theta}_3$ is preserved. In the general case, when we have s dependency $q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)q(\boldsymbol{\theta}_j)$, the optimal variational posteriors are given by:

$$q^*(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j) \propto \exp\{\langle \ln p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \rangle_{q(\boldsymbol{\theta}_{\setminus i,j})}\} \tag{A.26}$$

$$q^*(\boldsymbol{\theta}_j) \propto \exp\{\langle \ln \frac{p(\mathbf{X}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)} \rangle_{q(\boldsymbol{\theta}_{\setminus i,j})}\} \tag{A.27}$$

$$\tag{A.28}$$

where $q(\boldsymbol{\theta}_{\setminus i,j})$ is the product of all the variational posterior but $q(\boldsymbol{\theta}_i|\boldsymbol{\theta}_j)$ and $q(\boldsymbol{\theta}_j)$.

Appendix B

Exponential Family of Distributions

In this appendix, we present the main probability distributions used throughout this thesis. We first start to give a brief introduction of the exponential family of distribution and then we provide the list of its principal members.

B.1 Exponential family of distribution

The exponential family of distribution is a set of parametric distributions which can be expressed as follows:

$$p(\mathbf{x}) = \exp\{\boldsymbol{\eta}^\top T(\mathbf{x}) - A(\boldsymbol{\eta}) + B(\mathbf{x})\}, \tag{B.1}$$

where $\boldsymbol{\eta}$ is the vector of natural (or canonical) parameters, $T(\mathbf{x})$ is the vector of sufficient statistics and $A(\boldsymbol{\eta}) = \ln \int_{\mathbf{x}} \exp\{\boldsymbol{\eta}^T T(\mathbf{x})\}\mathrm{d}\mathbf{x}$ is the log-normalizer. The type of the distribution (Normal, Dirichlet, ...) depends on the domain of the natural parameters and on how the vector of sufficient statistics is computed. $B(\mathbf{x})$ is the base measure of the distribution and does not depends on the parameters $\boldsymbol{\eta}$.

B.1.1 Partial derivative of the log-normalizer

An important property which is heavily used in Variational Bayes inference, is the relation between the log-normalizer and the expectation of the sufficient statistics, namely:

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \langle T(\mathbf{x}) \rangle_{p(\mathbf{x})}. \tag{B.2}$$

This is easily verified by taking the partial derivative of the log-normalizer:

$$\begin{aligned}
\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} &= \frac{\partial}{\partial \boldsymbol{\eta}} \ln \int_{\mathbf{x}} \exp\{\boldsymbol{\eta}^\top T(\mathbf{x})\}\mathrm{d}\mathbf{x} && \text{(B.3)}\\
&= \frac{1}{\int_{\mathbf{x}} \exp\{\boldsymbol{\eta}^\top T(\mathbf{x})\}\mathrm{d}\mathbf{x}} \int_{\mathbf{x}} T(\mathbf{x}) \exp\{\boldsymbol{\eta}^\top T(\mathbf{x})\}\mathrm{d}\mathbf{x} && \text{(B.4)}\\
&= \int_{\mathbf{x}} T(\mathbf{x}) \exp\{\boldsymbol{\eta}^\top T(\mathbf{x}) - A(\boldsymbol{\eta})\}\mathrm{d}\mathbf{x} = \langle T(\mathbf{x}) \rangle_{p(\mathbf{x})}. && \text{(B.5)}
\end{aligned}$$

B.1.2 Conjugate Prior

In Bayesian inference, we say that the prior $p(\boldsymbol{\eta})$ is conjugate to the likelihood distribution $p(\mathbf{x}|\boldsymbol{\eta})$ if the posterior $p(\boldsymbol{\eta}|\mathbf{x})$ is of the same type as the prior. For instance, if both $p(\boldsymbol{\eta})$ and $p(\boldsymbol{\eta}|\mathbf{x})$ are Dirichlet distribution then $p(\boldsymbol{\eta})$ is conjugate to the likelihood $p(\mathbf{x}|\boldsymbol{\eta})$. When the likelihood is a member of the exponential distribution:

$$p(\mathbf{x}|\boldsymbol{\eta}) = \exp\{\boldsymbol{\eta}^\top T_x(\mathbf{x}) - A_\eta(\boldsymbol{\eta})\}, \tag{B.6}$$

where $\boldsymbol{\xi}_0$ and $A_\xi(\boldsymbol{\xi_0})$ are, respectively, the natural parameters and the log-normalizer of the distribution $p(\boldsymbol{\eta})$. there exists a conjugate prior member of the exponential family defined as:

$$p(\boldsymbol{\eta}) = \exp\{\boldsymbol{\xi}_0^\top T_\eta(\boldsymbol{\eta}) - A_\xi(\boldsymbol{\xi}_0)\} \tag{B.7}$$

$$T_\eta(\boldsymbol{\eta}) = \begin{bmatrix} \boldsymbol{\eta} \\ -A_\eta(\boldsymbol{\eta}) \end{bmatrix} \tag{B.8}$$

We can verify that $p(\mathbf{x}|\boldsymbol{\eta})$ and $p(\boldsymbol{\eta})$ are indeed conjugate by taking the product of the likelihood and the prior:

$$p(\boldsymbol{\eta}|\mathbf{x}) \propto p(\mathbf{x}|\boldsymbol{\eta})p(\boldsymbol{\eta}) \tag{B.9}$$

$$\propto \exp\{\boldsymbol{\eta}^\top T_x(x) - A_\eta(\boldsymbol{\eta}) + \boldsymbol{\xi}_0^\top T_\eta(\boldsymbol{\eta})\} \tag{B.10}$$

$$\propto \exp\{(\boldsymbol{\xi}_0 + \begin{bmatrix} T_x(\mathbf{x}) \\ 1 \end{bmatrix})^\top T_\eta(\boldsymbol{\eta}).\} \tag{B.11}$$

Re-normalizing (B.11) to get the proper posterior distribution:

$$p(\boldsymbol{\eta}|\mathbf{x}) = \exp\{\boldsymbol{\xi}^\top T_\eta(\boldsymbol{\eta}) - A_\xi(\boldsymbol{\xi})\} \tag{B.12}$$

$$\boldsymbol{\xi} = \boldsymbol{\xi}_0 + \begin{bmatrix} T_x(\mathbf{x}) \\ 1 \end{bmatrix} \tag{B.13}$$

proves that (B.7) is the conjugate prior of $p(\mathbf{x}|\boldsymbol{\eta})$.

B.2 Distributions

We now describe the members of the exponential family of distributions used in this thesis. For each distribution we provide:

- the standard parametric form
- the vector of natural parameters $\boldsymbol{\eta}$
- the vector of sufficient statistics $T(\mathbf{x})$
- the log-normalizer $A(\boldsymbol{\eta})$
- the gradient of the log-normalizer which is also the expectation of the sufficient statistics.

B.2.1 Categorical

The categorical distribution is a probability distribution over a random variable which has K possible outcomes: $x \in \{1, \ldots, K\}$. It is parameterized by a vector of probabilities $\boldsymbol{\mu}$ where $0 < \mu_k < 1$ and $\sum_{k=1}^{K} \mu_k = 1$. Note that the natural parameters and the sufficient statistics of the categorical distribution are $K-1$ dimensional vectors.

$$p(x|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{\mathbb{1}[x=k]} \tag{B.14}$$

$$\eta_k = \ln\Big(\frac{\mu_k}{1 - \sum_{j=1}^{K-1} \mu_j}\Big) \tag{B.15}$$

$$T(\mathbf{x}) = \begin{bmatrix} \mathbb{1}[x=1] \\ \mathbb{1}[x=2] \\ \vdots \\ \mathbb{1}[x=K-1] \end{bmatrix} \tag{B.16}$$

$$A(\boldsymbol{\eta}) = \ln\Big(1 + \sum_{k=1}^{K-1} \exp\{\eta_k\}\Big) \tag{B.17}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_k} = \mu_k \tag{B.18}$$

B.2.2 Dirichlet

The Dirichlet distribution is a continuous probability distribution over K random variables $\mu_1, \mu_2, \ldots, \mu_K$ such that $0 < \mu_k < 1$ and $\sum_{k=1}^{K} \mu_k = 1$. The distribution is parameterized by a vector of K concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$. It is the conjugate prior of the categorical distribution. When $K = 2$, the Dirichlet distribution reduces to a Beta distribution.

$$p(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \mu_k^{(\alpha_k - 1)} \tag{B.19}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_K \end{bmatrix} = \begin{bmatrix} \alpha_1 - 1 \\ \vdots \\ \alpha_K - 1 \end{bmatrix} \tag{B.20}$$

$$T(\boldsymbol{\mu}) = \begin{bmatrix} \ln \mu_1 \\ \ln \mu_2 \\ \vdots \\ \ln \mu_K \end{bmatrix} \tag{B.21}$$

$$A(\boldsymbol{\eta}) = \Big(\sum_{k=1}^{K} \ln \Gamma(\eta_k + 1)\Big) - \ln \Gamma(\sum_{k=1}^{K} \eta_k + 1) \tag{B.22}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_k} = \ln \psi(\alpha_k) - \ln \psi(\sum_{k=1}^{K} \alpha_k) \tag{B.23}$$

where Γ and ψ are the gamma and digamma function respectively.

B.2.3 Gamma

The Gamma distribution is a continuous probability distribution over positive random variable $x > 0$. The distribution is governed by a shape parameter $a > 0$ and a rate parameter $b > 0$.

$$p(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp\{-bx\} \tag{B.24}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -b \\ a-1 \end{bmatrix} \tag{B.25}$$

$$T(x) = \begin{bmatrix} x \\ \ln x \end{bmatrix} \tag{B.26}$$

$$A(\boldsymbol{\eta}) = \ln \Gamma(\eta_2 + 1) + (\eta_2 + 1) \ln(-\eta_1) \tag{B.27}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{a}{b} \\ \psi(a) - \ln b \end{bmatrix}, \tag{B.28}$$

where Γ and ψ are the gamma and digamma function respectively.

B.2.4 Normal

The Normal distribution is a widely used continuous probability distribution over real vectors $\mathbf{x} \in \mathbb{R}^D$. It has two parameters: mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and a covariance matrix $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ is constrained to be a positive definite matrix. The distribution can also be expressed with the precision matrix $\boldsymbol{\Lambda} = \boldsymbol{\Sigma}^{-1}$.

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\Sigma|^{\frac{1}{2}}} \exp\Big\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \Big\} \tag{B.29}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu} \\ -\frac{1}{2} \operatorname{vec}(\boldsymbol{\Sigma}^{-1}) \end{bmatrix} \tag{B.30}$$

$$T(\mathbf{x}) = \begin{bmatrix} \mathbf{x} \\ \operatorname{vec}(\mathbf{x}\mathbf{x}^\top) \end{bmatrix} \tag{B.31}$$

$$A(\boldsymbol{\eta}) = -\frac{1}{4} \boldsymbol{\eta}_1^\top \operatorname{mat}(\boldsymbol{\eta}_2)^{-1} \boldsymbol{\eta}_1 - \frac{1}{2} \ln |-2 \operatorname{mat}(\boldsymbol{\eta}_2)| + \frac{D}{2} \ln 2\pi \tag{B.32}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_1} = \boldsymbol{\mu} \tag{B.33}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_2} = \operatorname{vec}\Big(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}^\top \Big) \tag{B.34}$$

where “vec‘ is the vectorization operation and “mat” is its inverse.

B.2.5 Normal-Wishart

The Normal-Whisart is a continuous probability distribution over a pair of real vector and positive definite matrix $\boldsymbol{\mu} \in \mathbb{R}^D, \boldsymbol{\Lambda} \in \mathbb{R}^{D \times D}$. It is the conjugate prior of the normal distribution with unknown mean and precision matrix. It is paremeterized by a mean $\mathbf{m}$,

a scaling factor β, a positive definite matrix $\mathbf{W}$ and a degree of freedom ν:

$$p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}, \beta, \mathbf{W}, \nu) = B \exp\{\frac{\nu - D}{2} \ln|\boldsymbol{\Lambda}| - \frac{\beta}{2} \operatorname{tr}(\boldsymbol{\Lambda}\mathbf{M})\} \tag{B.35}$$

$$\mathbf{M} = \beta(\boldsymbol{\mu} - \mathbf{m})(\boldsymbol{\mu} - \mathbf{m})^\top + \mathbf{W}^{-1} \tag{B.36}$$

$$B = \beta^{\frac{D}{2}} |\mathbf{W}|^{-\frac{\nu}{2}} \Big(2^{\frac{(\nu+1)D}{2}} \pi^{\frac{D(D+1)}{4}} \prod_{d=1}^{D} \Gamma(\frac{\nu + 1 - d}{2})\Big)^{-1} \tag{B.37}$$

$$\boldsymbol{\eta} = \begin{bmatrix} \boldsymbol{\eta}_1 \\ \eta_2 \\ \boldsymbol{\eta}_3 \\ \eta_4 \end{bmatrix} = \begin{bmatrix} \beta\mathbf{m} \\ -\frac{\beta}{2} \\ -\frac{1}{2}\operatorname{vec}(\beta\mathbf{m}\mathbf{m}^\top + \mathbf{W}^{-1}) \\ \frac{\nu - D}{2} \end{bmatrix} \tag{B.38}$$

$$T(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \begin{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\mu} \\ \boldsymbol{\mu}^\top\boldsymbol{\Lambda}\boldsymbol{\mu} \\ \operatorname{vec}(\boldsymbol{\Lambda}) \\ \ln|\boldsymbol{\Lambda}| \end{bmatrix} \tag{B.39}$$

$$A(\boldsymbol{\eta}) = -\ln B \tag{B.40}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_1} = \nu\mathbf{W}\mathbf{m} \tag{B.41}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_2} = \operatorname{tr}\Big(\frac{1}{\beta}\mathbf{I} + \nu\mathbf{W}\mathbf{m}\mathbf{m}^\top\Big) \tag{B.42}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}_3} = \nu \operatorname{vec}(\mathbf{W}) \tag{B.43}$$

$$\frac{\partial A(\boldsymbol{\eta})}{\partial \eta_4} = \sum_{d=1} \psi(\frac{\nu + 1 - d}{2}) + D\ln 2 + \ln|lnln\mathbf{W}|. \tag{B.44}$$