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

1 Introduction **2**
- 1.1 Motivations . . . 2
- 1.2 Related works . . . 3
- 1.3 Thesis Contributions . . . 5

2 Non-Parametric Bayesian Phone-Loop Model **7**
- 2.1 Bayesian formulation of the AUD problem . . . 7
  - 2.1.1 Non-parametric Bayesian AUD . . . 8
- 2.2 Model . . . 11
  - 2.2.1 Acoustic Model . . . 11
  - 2.2.2 Base measure . . . 13
  - 2.2.3 Generative Process . . . 14
- 2.3 Conclusion . . . 16

3 Generalized Subspace Model for Sound Representation **18**
- 3.1 Generalized Subspace Model . . . 18
  - 3.1.1 Definition . . . 19
- 3.2 Dirichlet Process Subspace Hidden Markov Model . . . 20
  - 3.2.1 Revisiting the base measure . . . 20
  - 3.2.2 Approximating the phonetic subspace of the target language . . . 22
- 3.3 Conclusion . . . 23

4 Phonotactic Language Model **24**
- 4.1 Non-Parametric Bigram Phone-Loop Model . . . 24
  - 4.1.1 Hierarchical Dirichlet Process . . . 25
- 4.2 Results . . . 26
- 4.3 Conclusion . . . 26

5 Conclusion **28**
- 5.1 Future work . . . 28
  - 5.1.1 Acoustic Modeling . . . 28
  - 5.1.2 Language Modeling . . . 29
- 5.2 Summary of contributions . . . 30

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.

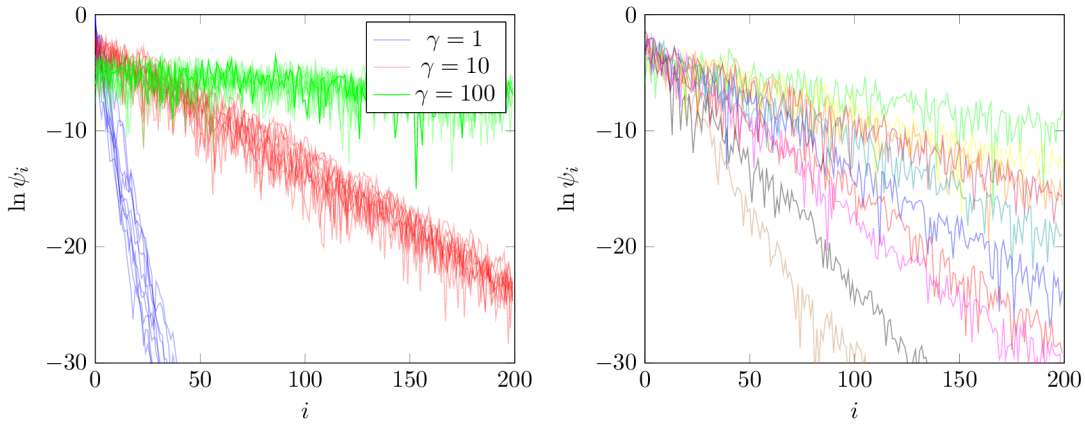
Many Dirichlet process-based models use the Chinese restaurant process as inference scheme (Lee and Glass, 2012; Beal et al., 2002). The Chinese restaurant process is a sampling scheme that draws, in the limit, samples from the posterior distribution over the model's parameters marginalized over all 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 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, 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:

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

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

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 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 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. 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:

- $\mathbf{s}^{u_l} = s^{u_l}_1, \dots, s^{u_l}_{N_l}$ is the sequence of indices of the HMM states for acoustic unit u_l

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

- $\mathbf{c}^{u_l} = c_1^{u_l}, \dots, c_{N_l}^{u_l}$ 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. 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:

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

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

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 $\pi_1, ..., \pi_C$ are constrained such that $0 < \pi_i < 1$ and $\sum_{i=1}^{C}\pi_i = 1$. Finally, the log-normalizers

are defined as:

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

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

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

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

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 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 and the prior over mean vector $\boldsymbol{\mu}$ and the (inverse) covariance matrix $\boldsymbol{\Sigma}^{-1}$ is a Normal-Wishart distribution. (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. 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.

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:

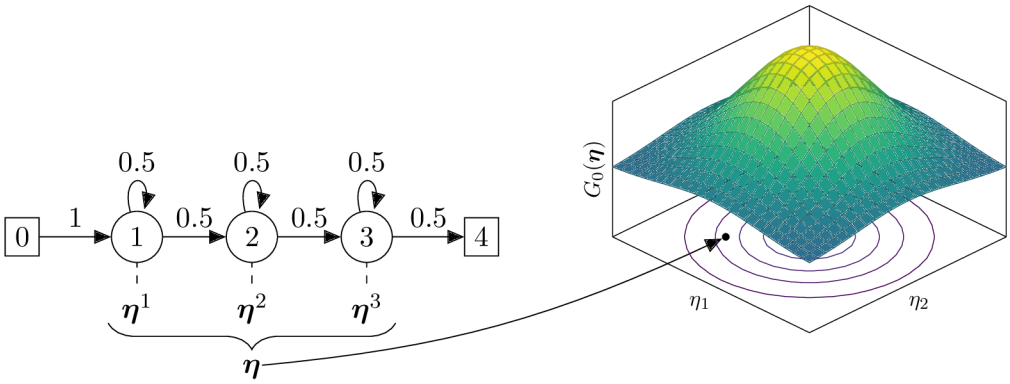


Figure 2.2: Model of an acoustic unit and its relation with the base measure. Each acoustic unit is parameterized by a vector of natural parameters $\boldsymbol{\eta}$ corresponding to the concatenation of all the HMM states' parameters. The base measure, G_0, is a density over the acoustic (natural) 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).

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 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.

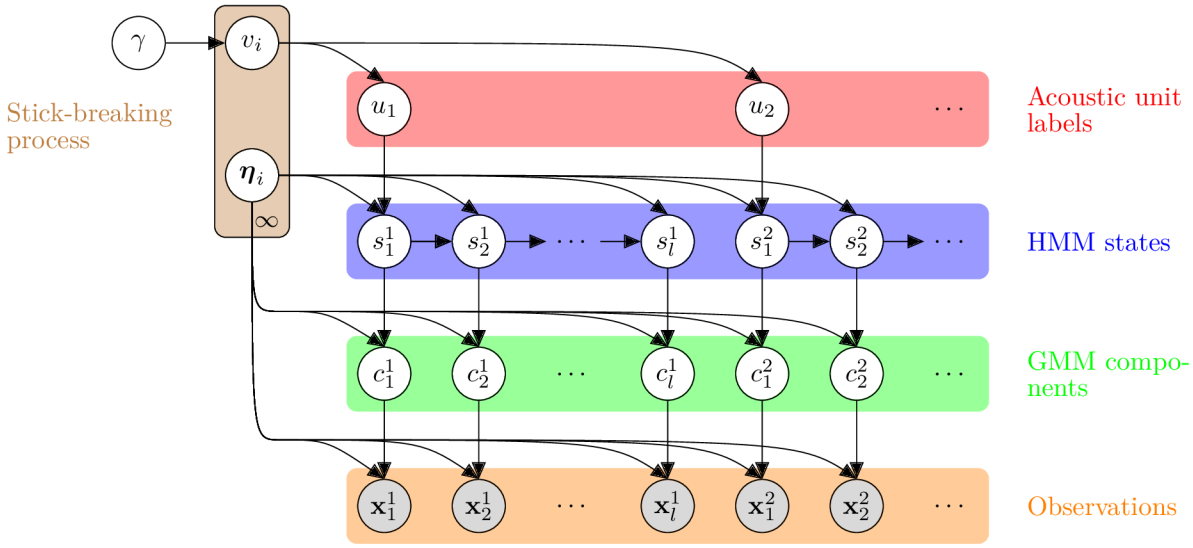


Figure 2.3: Bayesian network of the non-parametric acoustic unit clustering model for a given segmentation. a_i^j refers to the variable a associated to the ith segment of the jth unit. l is the duration of the first unit u_1. Note that in practice the segmentation is unknown and the inference needs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

- $\boldsymbol{\xi}_0$ (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.

2.3 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:

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 the *Subspace Hidden Markov Model* (SHMM) to represent phones in a low-dimensional space. Finally, in section 3.2, 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 $\mathbf{X}_1, \dots, \mathbf{X}_K$ where the ith set has N_i observations: $\mathbf{X}_i = \mathbf{x}_{i1}, \dots, \mathbf{x}_{iN_i}$. 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, 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. $\mathbf{h}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}) \quad \forall i \in \{1, 2, \dots, K\}$
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.

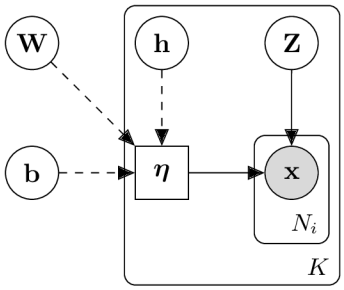


Figure 3.1: Graphical model of the Generalized Subspace Model. Dashed edges pointing to a square node represent a deterministic relation.

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.

3.2 Dirichlet Process Subspace Hidden Markov Model

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.

3.2.1 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 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. 3.2.

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 $\mathbf{h}_i \sim G_0 \quad i \in \{1, 2, \dots\}$
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 $\mathbf{s} = s_1, \dots, s_l$ 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 $\mathbf{x}_k \sim \mathcal{N}(\boldsymbol{\mu}_{u_j}^{s_k, c_k}, \boldsymbol{\Sigma}_{u_j}^{s_k, c_k})$

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. We call this new model the *Dirichlet Process Hidden Markov Model* (DP-SHMM) and its graphical representation is shown in Fig. 3.3. Interestingly, the base measure is not a proper density function in

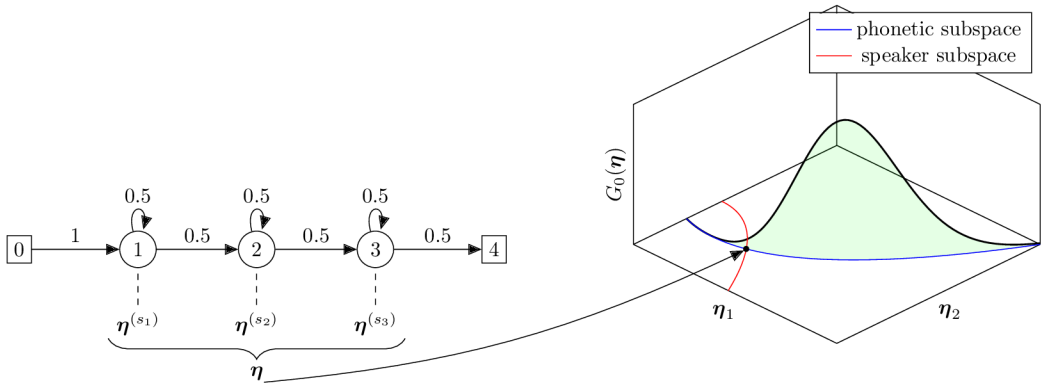


Figure 3.2: Base measure of the SHMM Dirichlet Process Mixture model.

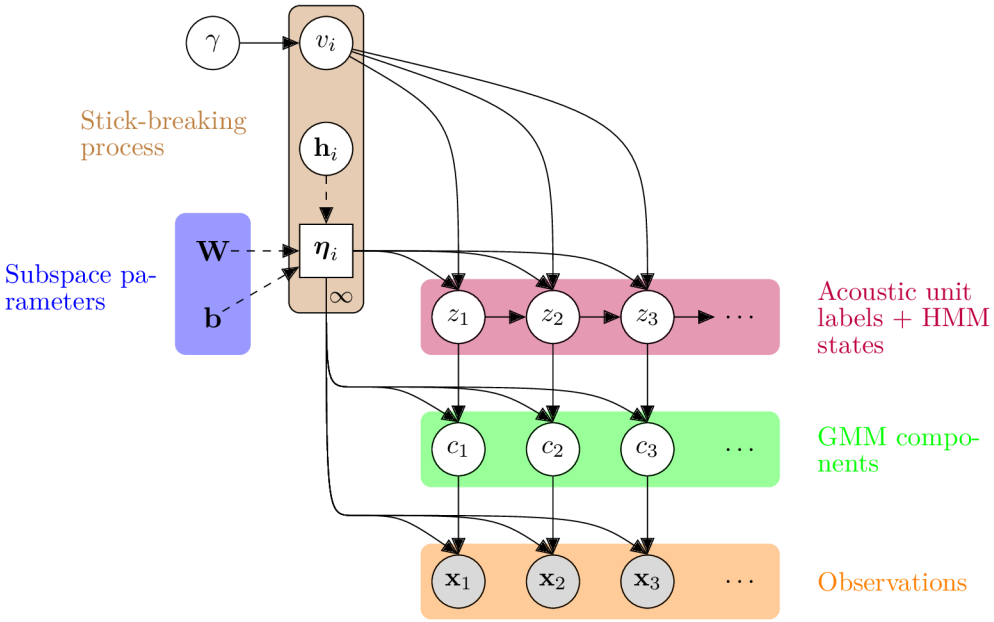


Figure 3.3: Bayesian network of the Dirichlet Process Subspace Hidden Markov Model (DP-SHMM). The atoms of the Dirichlet process are constrained to live in a low-dimensional subspace parameterized by $\mathbf{W}$ and $\mathbf{b}$.

the $\boldsymbol{\eta}$ space, however, a sample from the Dirichlet Process, $G \sim G_0$, is indeed a discrete probability distribution over the atoms $\mathbf{h}_1, \mathbf{h}_2, \ldots$:

$$G(\mathbf{h}) = \sum_{i=1}^{\infty} \psi_i \delta_{\mathbf{h}_i}(\mathbf{h}). \tag{3.2}$$

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 $q(\mathbf{h}_1), q(\mathbf{h}_2), \ldots$

3.2.2 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.

3.3 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 capturing 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 $\mathbf{h}_1, \mathbf{h}_2, \dots$ 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. We propose a „corrected“ version of the bigram AUD model to control how the acoustic and language model affects the learning.

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 of the HDP and then apply the mean-field approximation.

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.

4.2 Results

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

4.3 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 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

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.

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.

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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 M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

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.

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.

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.

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.

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. 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.

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.

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.