Part I

General discussion

1 Introduction

1.1 Machine learning and speech and language processing

Speech and language processing is one of the most successful examples of applying machine learning techniques to real problems. Current speech and language techniques embody our real-world information processing, automatically including information extraction, question answering, summarization, dialog, conversational agent, and machine translation (Jurafsky & Martin 2000). Among these, one of the most exciting applications of speech and language processing is speech recognition based voice search technologies (by Google, Nuance) and conversational agent technologies (by Apple) (Schalkwyk, Beeferman, Beaufays *et al*. 2010). These successful applications started to make people in general casually use speech interface rather than text interface in mobile devices, and the applications of speech and language processing are widely expanding.

One of the core technologies of speech and language processing is automatic speech recognition (ASR) and related techniques. Surprisingly, these techniques are fully based on statistical approaches by using large amounts of data. The machine learning techniques are applied to utilize these data. For example, the main components of ASR are acoustic and language models. The acoustic model (AM) provides a statistical model of each phoneme/word unit, and it is represented by a hidden Markov model (HMM). The HMM is one of the most typical examples of dealing with sequential data based on machine learning techniques (Bishop 2006), and machine learning techniques provide an efficient method of computing a maximum likelihood value for the HMM and an efficient training algorithm of the HMM parameters. The language model (LM) also provides an n-gram based statistical model for word sequences, which is also trained by using the large amount of data based on machine learning techniques. These statistical models and their variants are used for the other speech and language applications, including speaker verification and information retrieval, and thus, machine learning is a core component of speech and language processing.

Machine learning covers a wide range of applications in addition to speech and language processing, including bioinformatics, data mining, and computer vision. Machine learning also covers various theoretical fields including pattern recognition, information theory, statistics, control theory, and applied mathematics. Therefore, many people are studying and developing machine learning techniques, and the progress of machine learning is rather fast. By following the rapid progress of machine learning,

researchers in speech and language processing interact positively with the machine learning community or communities in the machine learning application field by importing (and sometimes exporting) advanced machine learning techniques. For example, the recent great improvement of ASR comes from this interaction for discriminative approaches (recent progress summaries for discriminative speech recognition techniques are found in Gales, Watanabe & Fossler-Lussier (2012), Heigold, Ney, Schluter *et al.* (2012), Saon & Chien (2012*b*), Hinton, Deng, Yu *et al.* (2012). The discriminative training of HMM parameters has been mainly studied in speech recognition research since the 1990s, and became a standard technique around the 2000s. In addition, the deep neural network replaces the emission probability of the HMM from the Gaussian mixture model (GMM) (or is used as feature extraction (Hermansky, Ellis & Sharma 2000, Grézl, Karafiát, Kontár *et al.* 2007) for the GMM) and achieves further improvement on the discriminative training based ASR performance. Actually, current successful applications of speech and language processing are highly supported by these breakthroughs based on the discriminative techniques developed through the interaction with the machine learning community. By following the successful experience, researchers in speech and language processing try to collaborate with the machine learning community further to find new technologies.

1.2 Bayesian approach

This book also follows the trend of tight interaction with the machine learning community, but focuses on another active research topic in machine learning, called the *Bayesian approach*. The Bayesian approach is a major probabilistic theory that represents a causal relationship of data. By dealing with variables introduced in a model as probabilistic variables, we can consider uncertainties included in these variables based on the probabilistic theory.

As a simple example of uncertainty, we think of statistically modeling several data (x_1, x_2, $\cdots$, x_N) by a Gaussian distribution $\mathcal{N}(x|\mu, \Sigma)$ with mean and variance parameters μ and Σ, as shown in Figure 1.1, i.e.,

$$p(x) \approx \mathcal{N}(x|\mu, \Sigma). \tag{1.1}$$

Now, we consider the Bayesian approach, where the mean parameter is uncertain, and is distributed by a probabilistic function $p(\mu)$. Since μ is uncertain, we can consider the several possible μs instead of one fixed μ, and the Bayesian approach considers representing a distribution of x by several Gaussians with possible mean parameters (μ_1, μ_2, and μ_3 in the example of Figure 1.1),

$$p(x) \approx \frac{1}{N} \sum_{\mu=\{\mu_1,\mu_2,\cdots,\mu_N\}} \mathcal{N}(x|\mu, \Sigma), \tag{1.2}$$

where μ_1, μ_2, $\cdots$ are generated from the distribution $p(\mu)$. The extreme case of this uncertainty consideration is to represent the distribution of x, which is represented by

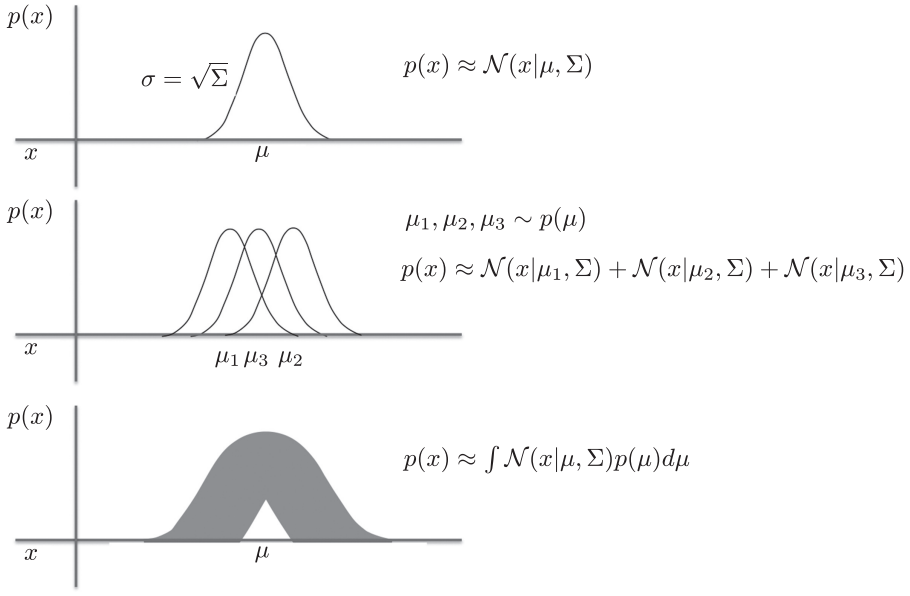

Figure 1.1 The uncertainty of the mean parameter of a Gaussian distribution.

all possible Gaussian distributions, with *all possible* mean parameters weighted by the probability $p(\mu)$, which is represented as the following integral equation:

$$p(x) \approx \int \mathcal{N}(x|\mu, \Sigma)p(\mu)d\mu. \tag{1.3}$$

This expectation over uncertain variables is called *marginalization*. The grayed band in Figure 1.1 provides an image of the expected distribution, where the mean parameter is marginalized over all possible infinite mean values. This is a unique aspect of the Bayesian approach that represents variables in a model by probabilistic distributions, and holds their *uncertainties*.

This uncertainty consideration often improves the generalization capability of a model that yields to mitigate the mismatch between training and unseen data and avoid over-fitting problems by the effect of regularization and marginalization. For example, Figure 1.2 points out the over-fitting problem. In real applications, we often face phenomena where observed data cannot cover all possible unobserved data (unseen data) and their distributions are mismatched. The Gaussian distribution with a fixed mean parameter that is estimated from observed data can well represent the observed data, but it cannot represent unseen data properly. This is a well-known over-fitting problem, that the model overly fits its parameters to represent observed data. However, since the Bayesian approach considers all possible mean parameters by the marginalization, some of these parameters would properly model unseen data more accurately than the fixed mean parameter case. The effect of this more powerful representation ability for unseen data leads to improved generalization capability, which is a famous advantage of the Bayesian approach. In addition, the example can also be viewed as

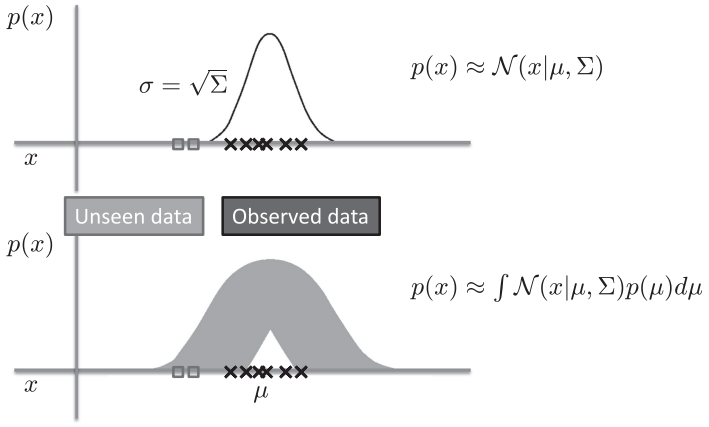

Figure 1.2 The Bayesian approach that holds uncertainty of the mean parameter has an ability to describe unseen data.

showing that the mean parameter is *regularized* not to overly fit the observed data to $p(\mu)$. The effect of setting the constraints for variables via their probabilistic distributions is a typical example of the *regularization*, and the regularization also leads to improved generalization capability.

Furthermore, the Bayesian theory provides a straightforward mathematical way to infer (predict) unobservable probabilistic variables by using the basic rules (including the Bayes theorem) within the basic probabilistic theory. The beauties of this mathematical treatment based on probabilistic theory and the expected robustness by considering uncertainties attract many machine learning researchers to study the Bayesian approach. Actually, Bayesian machine learning has also rapidly grown similarly to discriminative techniques, and various interesting approaches have been proposed (Bishop 2006, Barber 2012).

However, compared with the successful examples of the discriminative techniques, the applications of the Bayesian approach in speech and language processing are rather limited despite its many advantages. One of the most difficult problems is that the exact Bayesian approach cannot be applied in our speech and language processing without some approximations. For example, the Bayesian approach is based on the conditional distribution given datum x (that is called posterior distribution, $p(a|x)$). However, it is generally difficult to obtain the posterior distribution analytically, since we cannot solve the equation analytically to obtain the posterior distribution. The computation is often performed numerically, which limits the practical applications of the Bayesian approach, especially for speech and language processing that deals with large amounts of data. Thus, *how to make the Bayesian approach more practical for speech and language processing* is the most critical problem in Bayesian speech and language processing.

This book is aimed to guide readers in machine learning or speech and language processing to apply the Bayesian approach to speech and language processing in a systematic way. In other words, this book aims to bridge the gap between the machine learning community and the speech and language community by removing

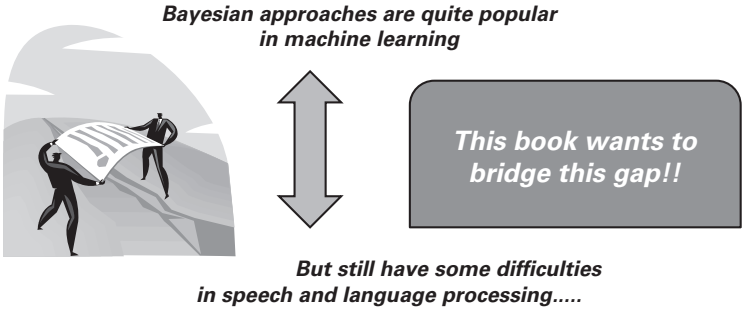

Figure 1.3 The aim of this book.

their preconceived ideas of the difficulty in these applications (Figure 1.3) (Watanabe & Chien 2012). The key idea for this guidance is *how to approximate the Bayesian approach* for specific speech and language applications (Ghahramani 2004). There are several approximations developed mainly in the machine learning and Bayesian statistics fields to deal with the problems. This book mainly deals with the following approximations:

- Chapter 4: Maximum a-posteriori approximation (MAP);
- Chapter 5: Evidence approximation;
- Chapter 6: Asymptotic approximation;
- Chapter 7: Variational Bayes (VB);
- Chapter 8: Markov chain Monte Carlo (MCMC).

Note that there are some other interesting approximations (e.g., loopy belief propagation (Murphy, Weiss & Jordan 1999, Yedidia, Freeman & Weiss 2003) and expectation propagation (Minka 2001)), but our book focuses on the above approximations. These approximations are described in the corresponding chapters in detail. We organize the chapters in Part II categorized by these approximation techniques, unlike application-oriented categorization (e.g., speech recognition and speech synthesis) as has appeared in many other speech and language books (Jurafsky & Martin 2000, Huang, Acero & Hon 2001), to emphasize how to approximate the Bayesian approach for practical applications. For example, the first sections in each chapter in Part II describe the introduction of the corresponding approximation, and provide the recipe for how to use the approximation in machine learning problems in general. The following sections in the chapter provide the approximated solutions of statistical models used in specific speech and language applications by following the recipe. This book mainly deals with popular statistical models including HMM, GMM, neural network, factor analysis, n-gram, and latent topic models. Table 1.1 summarizes approximated Bayesian inferences for statistical models discussed in this book. The applications covered by this book are typical topics in speech and language processing based on these statistical models with the approximated Bayesian treatment, and mainly related to automatic speech recognition.

Table 1.1 Approximated Bayesian inference for statistical models discussed in this book.

Approximation	HMM	GMM	Neural network	Factor analysis	*n*-gram	Latent topic model
MAP	4.3, 4.5	4.6			4.7	4.8
Evidence	5.2				5.3	
Asymptotic	6.3, 6.5	6.6	6.4			
VB	7.3			7.5		7.6
MCMC	8.3, 8.4				8.5	

1.3 History of Bayesian speech and language processing

There have been various studies to apply the Bayesian approach in speech and language processing. Although one of the aims of this book is to summarize these studies in a more systematic way in terms of an approximated Bayesian inference view, this section briefly reviews the history of these studies, according to time, by categorizing them within four trends.

The major earliest trend of using the Bayesian approach to speech and language processing started with *the statistical modeling* of automatic speech recognition in the 1980s. Furui (2010) reviews a historical perspective of automatic speech recognition and calls the technologies developed with statistical modeling around the 1980s *the third generation technology*. In a Bayesian perspective, this statistical modeling movement corresponded to the first introduction of *probabilistic variables* for speech recognition outputs (e.g., word sequences). As a result, the statistical modeling of automatic speech recognition formulates the speech recognition process as Bayes decision theory, and the noisy channel model is provided to solve the decision problem of determining most probable word sequences by considering the product of acoustic and language model distributions based on the Bayes theory (Jelinek 1976). The language model is used to provide a prior distribution of word sequences.

The second trend in the 1990s was to expand the Bayesian perspective from the speech recognition outputs to *model parameters* by regarding these as probabilistic variables. Maximum a-posteriori (MAP) estimation of HMM parameters is known as the most successful application of the Bayesian approach (Lee, Lin & Juang 1991, Gauvain & Lee 1994). This approach was used for speaker adaptation, where the prior distribution of HMM parameters is estimated from a large amount of speaker-independent data and the MAP estimation is used to estimate target speaker's HMM parameters with a small amount of target speaker data. The prior distribution regularizes the target speaker's HMM parameters to guarantee the performance of speaker independent HMMs instead of overly tuning to the model. Another example of the expansion was to treat *model structure* as a probabilistic variable in the late 1990s. This treatment enables *model selection* by selecting a most probable model structure from the posterior distribution of the model structure given training data (e.g., the numbers of HMM states and Gaussians). The Bayesian information criterion (BIC) and the minimum description

length (MDL) criterion are successful examples (Shinoda & Watanabe 1996, Chen & Gopinath 1999, Chou & Reichl 1999, Zhou & Hansen 2000) that cover wide applications of speech and language processing (e.g., acoustic model selection, speaker clustering, and speaker segmentation).

This second trend made the importance of the Bayesian approach come alive within the speech and language communities. From the late 1990s to 2000s, many other Bayesian studies have been applied to speech and language processing, which are classified as the third trend. The Bayesian techniques (MAP and BIC) used in the second trend miss the important Bayesian concept, *marginalization*, which makes the full treatment of the Bayesian approach difficult. By following the progress of VB, MCMC, Evidence approximation, and graphical model techniques in machine learning in the 1990s, people in speech and language processing started to apply more exact Bayesian approaches by *fully incorporating marginalization*. These studies covered almost all statistical models in speech and language processing (Bilmes & Zweig 2002, Watanabe, Minami, Nakamura & Ueda 2002, Blei, Ng & Jordan 2003, Saon & Chien 2011). The most successful approach in the third trend is latent Dirichlet allocation (LDA (Blei *et al*. 2003)), which provides a VB solution of a latent topic model from a probabilistic latent semantic analysis (Hofmann 1999*b*). LDA has been mainly developed in machine learning and natural language processing by incorporating a Gibbs sampling solution (Griffiths & Steyvers 2004), and structured topic models (Wallach 2006), and LDA was extended to incorporate *Bayesian nonparametrics* (Teh, Jordan, Beal & Blei 2006), which became the fourth trend in Bayesian speech and language processing.

The fourth trend is a still ongoing trend that tries to fully incorporate Bayesian nonparametrics with speech and language processing. Due to their computational costs and algorithmic difference from the standard expectation maximization (EM) type algorithms (Dempster, Laird & Rubin 1976), the applications were limited to latent topic models and related studies. However, recent computational progress (e.g., many core processing and GPU processing) has enabled broadening of the fourth trend in statistical models in speech and language processing other than extended latent topic models (Teh 2006, Goldwater 2007, Fox, Sudderth, Jordan *et al*. 2008, Ding & Ou 2010, Lee & Glass 2012).

This book covers all four trends and categorizes these popular techniques with the approximated Bayesian inference techniques.

1.4 Applications

This book aims to describe the following target applications:

- Automatic speech recognition (ASR, Figure 1.4)
  This is a main application in this book, that converts human speech to texts. The main techniques required in ASR are based on speech enhancement for noise reduction, speech feature extraction, acoustic modeling, language modeling, pronunciation

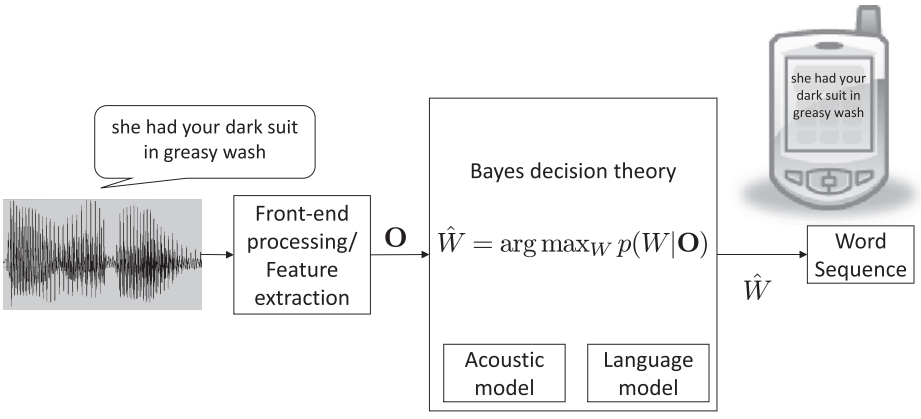

Figure 1.4 Automatic speech recognition.

lexicon modeling, and search. The book mainly deals with acoustic models represented by HMM and neural networks, and language models represented by *n*-gram and latent topic models.

- Speaker verification/segmentation/clustering (Figure 1.5)
  Speaker recognition techniques automatically provide a speaker identity given speech data. This often requires extracting speech segments uttered by target speakers from the recorded utterances (speaker segmentation). In addition, some of the real applications cannot have speaker labels in advance, and clustering speech segments to a specific speaker cluster (speaker clustering) is also another important direction. Gaussian or GMMs are used as statistical models. However, state-of-the-art speaker verification systems use GMMs as preprocessing, namely GMM parameters estimated from a speech segment are used as speaker features. The estimated features are further processed by a factor analysis to remove speaker-independent feature characteristics statistically.
- (Spoken) Information retrieval (IR, Figure 1.6)
  Information retrieval via document classification given a text (or spoken) query is another application of speech and language processing. The document can be represented by various units including newspaper articles, web pages, dialog conversations, sentences/utterances; which is used depends on applications. The most successful application for information retrieval is a search engine that uses a web page as a document unit. Voice search is an instance of a search engine based on spoken information retrieval where spoken terms are converted to text-form terms by using ASR and these terms are used as a query. The approach is based on the vector space model, which represents documents and queries as vectors in a vector space. The vector is simply represented by count or weighted count of unique words in a vocabulary. The approach often uses *n*-gram or latent topic models to provide more informative vector representation of documents.