Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 969–988

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Joint training of non-negative Tucker decomposition and discrete density hidden Markov models

Meng Sun*, Hugo Van hamme

Department of Electrical Engineering-ESAT, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, Bus 2441, B-3001 Leuven, Belgium

Received 7 February 2012; received in revised form 26 September 2012; accepted 28 September 2012
Available online 18 October 2012

Abstract

Non-negative Tucker decomposition (NTD) is applied to unsupervised training of discrete density HMMs for the discovery of sequential patterns in data, for segmenting sequential data into patterns and for recognition of the discovered patterns in unseen data. Structure constraints are imposed on the NTD such that it shares its parameters with the HMM. Two training schemes are proposed: one uses NTD as a regularizer for the Baum–Welch (BW) training of the HMM, the other alternates between initializing the NTD with the BW output and vice versa. On the task of unsupervised spoken pattern discovery from the TIDIGITS database, both training schemes are observed to improve over BW training in terms of pattern purity, accuracy of the segmentation boundaries and accuracy for speech recognition. Furthermore, we experimentally observe that the alternative training of NTD and BW outperforms the NTD regularized BW, BW training and BW training with simulated annealing.

Keywords: Non-negative Tucker decomposition; Hidden Markov models; Unsupervised training; Regularization; Speech recognition; Sequential pattern discovery; Vocabulary acquisition

1. Introduction

Hidden Markov models (HMM) are good at modeling sequential data and have thus been applied to a lot of tasks of sequential data processing, such as automatic speech recognition (ASR) (Rabiner, 1989), topic detection and segmentation (Blei and Moreno, 2001), handwriting recognition (Hu et al., 1996) and gene analysis (Pedersen and Hein, 2003). To estimate the unknown parameters of an HMM, usually the likelihood of the data is maximized with the expectation maximization (EM) method (Baum et al., 1970; Juang and Rabiner, 1990; Wu, 1983). Often, sequential data comes with sequential labels, resulting in a *supervised* training problem. For instance, in state-of-the-art ASR systems, the training data is labeled in terms of the words or phones and an HMM representation is learned for each label.

In some tasks the labels are not provided and the training is therefore *unsupervised*. For example, unsupervised vocabulary discovery aims to learn word-like sequential patterns from input speech data in a completely data-driven way. The motivation for studying unsupervised vocabulary discovery is fourfold. Firstly, it can serve to model early language acquisition of infants where they acquire spoken units from exposure to the spoken language. These units are

* Corresponding author. Tel.: +32 016 32 18 30; fax: +32 016 32 17 23.
E-mail addresses: mengsun@esat.kuleuven.be, sunmengccjs@gmail.com (M. Sun), hugo.vanhamme@esat.kuleuven.be (H. Van hamme).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.09.006

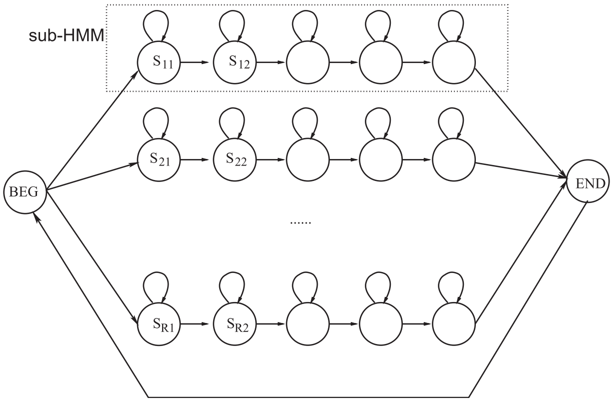

Fig. 1. The HMM topology used for sequential pattern discovery. Each parallel branch is called a sub-HMM.

subsequently related to multi-modal observations (e.g. images, actions) which is a form of weak supervision (Boves et al., 2007; Roy, 2003; Heckmann et al., 2009; Clemente et al., 2010). Secondly, a similar learning problem is created when one wants to communicate with robots in natural language to assign them a task. Such instructions are bound to contain environment-specific vocabularies that the robot needs to learn from interaction with humans. Thirdly, unsupervised learning of HMMs does not require transcriptions of speech, so the huge human effort to generate hand labels is avoided (Park and Glass, 2008; Zhang and Glass, 2010). Finally, unsupervised vocabulary acquisition can be applied in modeling out-of-vocabulary (OOV) words (ten Bosch et al., 2008; Jansen et al., 2010) and in the analysis of speech in underresourced languages or dialects (Zhuang et al., 2009).

In sequential pattern discovery, the only information that is available to us is that the data contains recurring *sequential patterns* showing some variation and possibly embedded in data that does not show much structure. In this article, it is assumed that each sequential pattern is adequately modeled by a sub-HMM, i.e. sequential pattern discovery is cast as the unsupervised estimation of the parameters of an HMM with the topology of Fig. 1. The objective function of unsupervised learning is usually the data likelihood which is maximized by using numerical optimization methods such as gradient descent algorithm (Levinson et al., 1982) or EM methods via optimizing an auxiliary function such as in Baum–Welch training (Baum et al., 1970; Khreich et al., 2012). Since EM is easier to apply than the gradient-based numerical optimization and is suitable for solving problems with a large number of parameters, it has been vastly adopted in ASR. Though the EM algorithm guarantees non-decrease of the data likelihood at each iteration, it can converge to local extrema due to the non-convexity of the optimization problem. Even in *supervised* learning careful initialization procedures with gradual increase of the model and data complexity are required to yield HMM parameter estimates that result in accurate recognition scores (Huang et al., 2001). In *unsupervised* learning, such procedures are more difficult to implement by the very nature of the problem: the data is unlabeled and its content is unknown. When applied blindly, the EM algorithm will produce poor local optima, as was observed in Smith (2006) and Johnson (2007) in discovering linguistic structure or as will be shown in Section 5.2.3 on word discovery where patterns corresponding to multiple words are found as illustrated in Fig. 5. Thus the unsupervised learning problem requires other solutions than supervised learning.

Our ultimate goal is pattern discovery in unlabeled data. In doing so, the avoidance of local optima in EM training emerges as an important issue which can be addressed with simulated annealing, i.e. perturbing the parameter estimates randomly with a magnitude that decreases with the iteration number (Paul, 1985). Simulated annealing is claimed to be able to yield the global optimum with probability one given *sufficiently slow* annealing and *enough* iterations. In practice, these conditions are hard to satisfy. In the example given in Paul (1985), a small HMM system with 8 states and 9 observation symbols required 60,000–400,000 iterations to attain high data likelihoods. On the other hand, we should notice that the target of sequential pattern discovery is not only to find a group of solutions with high data likelihood, but also to discover meaningful sequential patterns in the data. Hence, a model to jointly learn ground truth labels and HMM parameters is proposed in Siu et al. (2011). The iterative training procedures guarantee the non-decrease of the joint objective function. However, experiments reported in this article will show that the method is affected by the choice of initial ground truth labels which are treated as “correct” supervision in the following EM