Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 1029–1048

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Incremental word learning: Efficient HMM initialization and large margin discriminative adaptation

Irene Ayllón Clemente [a,b,*], Martin Heckmann [b], Britta Wrede [a]

[a] Bielefeld University, Research Institute for Cognition and Robotics – CoR-Lab, D-33615 Bielefeld, Germany
[b] Honda Research Institute Europe GmbH, D-63073 Offenbach am Main, Germany

Received 22 September 2011; received in revised form 19 April 2012; accepted 19 April 2012
Available online 28 April 2012

Abstract

In this paper we present an incremental word learning system that is able to cope with few training data samples to enable speech acquisition in on-line human robot interaction. As with most automatic speech recognition systems (ASR), our architecture relies on a Hidden Markov Model (HMM) framework where the different word models are sequentially trained and the system has little prior knowledge. To achieve good performance, HMMs depends on the amount of training data, the initialization procedure and the efficiency of the discriminative training algorithms. Thus, we propose different approaches to improve the system. One major problem of using a small amount of training data is over-fitting. Hence we present a novel estimation of the variance floor dependent on the number of available training samples. Next, we propose a bootstrapping approach in order to get a good initialization of the HMM parameters. This method is based on unsupervised training of the parameters and subsequent construction of a new HMM by aligning and merging Viterbi decoded sequences. Finally, we investigate large margin discriminative training techniques to enlarge the generalization performance of the models using several strategies suitable for limited training data. In the evaluation of the results, we examine the contribution of the different stages proposed to the overall system performance. This includes the comparison of different state-of-the-art methods with our presented techniques and the investigation of the possible reduction of the number of training data samples. We compare our algorithms on isolated and continuous digit recognition tasks. To sum up, we show that the proposed algorithms yield significant improvements and are a step towards efficient learning with few examples.

Keywords: Incremental word learning; Learning from few examples; Speech recognition; Multiple sequence alignment; Discriminative training; Bootstrapping

1. Introduction

In cognitive robotics researchers target for the understanding and modeling of cognitive processes in order to endow robotic systems with more complex behavior and the ability of autonomous learning. Robotic systems have to acquire knowledge during unconstrained interaction and transfer this knowledge to new situations in order to realize daily activities as well as to interact with humans, e.g. to assist people requiring care. This requires the development of methods to directly interact with human tutors rather than hard-coded programs or predefined behaviors to fulfill only specific predefined tasks.

A basic ability to achieve this goal is a communication between human and machine, i.e. a personal robot has to understand written or verbal communication (Charles and Bailey, 1992). It would enable a natural interaction with the user and the possibility to get a direct feedback during interaction as well as to transfer knowledge without the requirement of programming skills. Since we communicate using different dialects and vocabulary, it is crucial that robots not only know our native language, but also

* Corresponding author at: Bielefeld University, Research Institute for Cognition and Robotics – CoR-Lab, D-33615 Bielefeld, Germany.
E-mail addresses: iayllon@cor-lab.uni-bielefeld.de (I. Ayllón Clemente), martin.heckmann@honda-ri.de (M. Heckmann), bwrede@cor-lab.uni-bielefeld.de (B. Wrede).

0167-6393/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.specom.2012.04.005

are able to learn new words of our own lexicon, e.g. dialects.

Language acquisition is a complex mental ability of humans (Bosch et al., 2009). Nowadays automatic speech recognition (ASR) systems are considered as potential computational models of these corresponding cognitive skills. Despite the huge advances of ASR in the last decade, all conventional ASR systems still perform substantially worse than humans (Bosch et al., 2008). In addition, until now no computational model can precisely explain the acquisition of language and communication skills (Boves et al., 2007). Thus, our target is the efficient construction of speech models in order to enable the acquisition of knowledge in robots. In this way, an interactive learning system where a human tutor teaches a robot is a very attractive scenario for our long-term goal.

The framework we propose is inspired by the process of speech acquisition in children. Compared to current systems children learn through constant communication and interaction with their parents and other children. Hence for auditory learning in small children, we assume that a closed loop of speech perception and production plays an important role. While some authors concentrate on jointly solving both aspects (Moore, 2007; Minematsu, 2010), others (Van Segbroeck and Van hamme, 2009; Bosch et al., 2009) constrain their work on the robust perception, i.e. the recognition of words, as it is in itself still a widely unsolved problem in ASR systems. Moreover, our learning framework should constitute a user-friendly system adequate to acquire language in an unsupervised way and to require a low tutoring time, i.e. few training samples.

In ASR systems, the standard statistical model to represent the structure of speech is the continuous density Hidden Markov Model (CDHMM), where Gaussian Mixture Models (GMM) are usually employed to model the feature distribution in the hidden states. In HMMs, the amount of available training data to obtain the estimates of the model parameters, the initialization of the models and the discriminative training algorithms used are critical factors. These aspects are explained in the following paragraphs.

With conventional off-line training techniques, a large amount of labeled training data is required to estimate an optimal set of parameters. Unfortunately, it is very difficult to obtain this data in interactive learning; hence researchers aim to train the system in an unsupervised manner and/or to train it with a smaller number of training data samples (Iwahashi, 2006). One of the main drawbacks of using a small amount of training data is the over-fitting problem (Bishop, 2006). In this case the learning algorithms of the HMMs may fit the model parameters to some specific features of the dataset but do not generalize to unseen examples (Ghahramani, 2001). There are different methods to avoid over-fitting. A common method is to apply standard cross-validation learning (Devijver and Kittler, 1982). However, such an approach relies on off-line training and we cannot ensure enough training data on on-line systems to realize this approach. Alternative approaches to enable learning with few training samples apply regularization methods with a penalty term (Neukirchen and Rigoll, 1998) or in other cases, they reduce the number of free parameters (Siu et al., 1999). Following this idea, Neukirchen et al. (1998) and Kadous (2002) proposed to use 'parameter tying'. The idea is to set up an equivalence relation between HMM parameters in different states. In this way the number of independent parameters in the model is reduced. The problem here is to define these relations *a priori*, which is also difficult in an incremental learning scenario without prior knowledge. Nevertheless, the widely used technique to avoid over-fitting is the integration of further constraints like the so called variance floor. Here the training algorithms can be modified by including a lower threshold on the variance parameters, a variance floor (Melin, 1998; Sim and Gales, 2006; Melin and Lindberg, 1999). One simple way of computing the variance floor is to estimate the global variance of the speech segments and then scale it by a predefined factor (Stuttle, 2004). In contrast, other researchers as Dong et al., 2008 use a method to adapt the variance floor to each dimension of the features or to floor the variances using a percentile variable (Lee et al., 1992). Additionally to the variance floor, the minimum allowed variance (Young et al., 2006) is another threshold applied to the floor, which can also replace it in cases where no variance floor is computed. Other authors suggest that the variances of the GMMs should be fixed at the beginning and not updated in each iteration of the re-estimation of the GMMs (Liu, 1994; Rosenberg et al., 1998; Mokbel and Collin, 1999).

Since the training process in HMMs is based on Expectation-Maximization, the efficiency of the procedure strongly depends on the initialization of the parameters. The standard initialization approach was proposed by Rabiner (1989). It segments the training data frames into states by means of K-means clustering and Viterbi decoding. This initialization technique is executed in a supervised manner requiring the use of many labeled training data samples. However this bootstrapping method gives the possibility to iteratively estimate new models. Another state-of-the-art initialization method is 'flat start'. In this case all Gaussians Mixtures Models (GMMs) of the classes are initialized with identical parameters equal to the global speech mean and variance of the training data (Itaya et al., 2005; Young et al., 2006). The advantage of this method is that the models can be easily initialized without the need of labels, i.e. unsupervised. Nevertheless, this technique does not allow an efficient iterative construction of new models. Either the global mean and variance are exclusively computed from the training data of the new cluster, or all the computations have to be repeated each time a new class is introduced into the system. Some instances of similar initialization techniques can be found in Nathan et al. (1996) and Smith (2002). Moving towards an unsupervised initialization technique for incremental learning in interactive environments, some authors like Brandl et al. (2008) and Iwahashi (2006) initialize the system by means of an unsu-