ESMERALDA: A Development Environment for HMM-Based Pattern Recognition Systems

— Extended Abstract —

Gernot A. Fink and Thomas Plötz
Dortmund University, Robotics Research Institute, Intelligent Systems Group
Otto-Hahn-Str. 8, 44227 Dortmund, Germany
{Gernot.Fink,Thomas.Ploetz}@udo.edu

Abstract

In this paper we describe ESMERALDA – an integrated Environment for Statistical Model Estimation and Recognition on Arbitrary Linear Data Arrays – which is a development toolkit for building statistical recognizers operating on sequential data as e.g. speech, handwriting, or biological sequences. ESMERALDA primarily supports continuous density Hidden Markov Models (HMMs) of different topologies, and with user-definable internal structure. Furthermore, the framework supports the incorporation of Markov chain models (realized as statistical n-gram models) for long-term sequential restrictions, and Gaussian mixture models (GMMs) for general classification tasks. In recent years various applications within different challenging research areas have been realized using ESMERALDA. **Availability:** The software is open source and can be retrieved under the terms of the GPL from `sourceforge.net/projects/esmeralda`.

1 Introduction

In recent years Hidden Markov Models (HMMs), as the most prominent variant of segmentation free classification approaches, have become a standard technology for the automatic analysis of sequential data. Originally applied to speech recognition tasks, HMMs nowadays also serve as classification framework for numerous alternative domains like handwriting recognition, gesture recognition, or biological sequence analysis.

Depending on the actual application domain, specialized model architectures have been developed and, especially for vector data, Gaussian mixture models describing the particular feature space were integrated. Furthermore, in order to represent long-term sequential restrictions, HMMs can be complemented with Markov chain models. Over the years efficient algorithms for both model training and evaluation have been developed.

The classification of sequential, probably noisy data using HMMs is based on its alignment to the particular models. Consequently, the effectiveness of HMM based classification is mainly caused by the fact that an explicit segmentation of data to be analyzed is not required.

For successfully applying HMM based recognizers to “real-world” scenarios the ESMERALDA toolkit has been developed. Therefore, also practical aspects have been considered, like memory as well as computational effiency for both robust estimation of statistical recognition systems and their evaluation on unknown data. In this paper an overview of the framework is given including a summary of applications realized so far using ESMERALDA.

In section 2 the concepts behind Markovian models are outlined. Afterwards, ESMERALDA’s architecture is described. Section 4 provides an overview of applications realized using ESMERALDA followed by a summary.

2 Markov-Model Concepts

HMMs describe a two-stage stochastic process with hidden states and observable outputs[1]. State transitions and generation of outputs are probabilistic. The output elements – or emissions – generated per state can be symbolic but mostly are continuous, i.e. vectors from some high-dimensional feature space, which is more suitable for pattern recognition applications. Continuous output distributions are modeled via mixtures of Gaussians (cf. e.g. [1]).

HMMs are attractive because there exist efficient algorithms for estimating the model parameters and for decoding the model on new data, which is equivalent to an integrated segmentation and classification of the associated sequence data. The efficiency arises from the fact that HMMs can only store one internal state as context for future actions, which is also called the Markov Property.

However, in many applications it is desirable to be able to describe long-term dependency also with a statistical model. In speech recognition, for example, where individual HMM states describe parts of elementary phonetic units, correlations between occurrances of subsequent words can not be captured using HMMs alone. This is where Markov-chain Models come into play.

Markov-chain models can be used to statistically describe the probability of the occurrance of entire symbol sequences. The sequence probability is defined based on conditional probabilities of some symbol – or word – occurring in the context of its $n-1$ predecessor words. The models are therefore referred to as n-gram or language models (cf. e.g. [2]).

In principle the conditional probabilities required can be derived from training data. However, even for moderate sizes of n (e.g. 2 for bigram models) most n-gram events necessary for deriving robust statistical estimates will not be observed in (small) training sets. Therefore, for robust estimation of n-gram models raw probabilities need to be smoothed appropriately in order to obtain useful probability estimates for events not observed in the training data.

As HMMs and n-gram models are quite similar to each other they can be rather easily combined into an integrated model. In order to balance between the different granularities of the models a weighted combination of the different scores is necessary. Furthermore, as n-gram models span considerably longer contexts than HMMs also the search procedures used for integrated model decoding are more complex.

3 System Architecture

The goal of ESMERALDA is to put together a tractable set of conceptually simple yet powerful techniques in an integrated development environment. The system consists of a modular architecture (cf. figure 1). Separate modules for estimating mixture density models in conjunction with HMMs and for building N-gram models are provided. Technically, every module contains a library with an API as well as stand-alone programs for manipulating the appropriate models and associated data.

The three fundamental modules, namely for mixture densities, HMMs and N-gram models provide the following methods:

Mixture densities: k-means and LBG-based unsupervised mixture estimation, respectively; EM based model training; maximum a-posteriori (MAP) adaptation; estimation of linear feature space transforms (PCA, LDA); efficient two-stage decoding

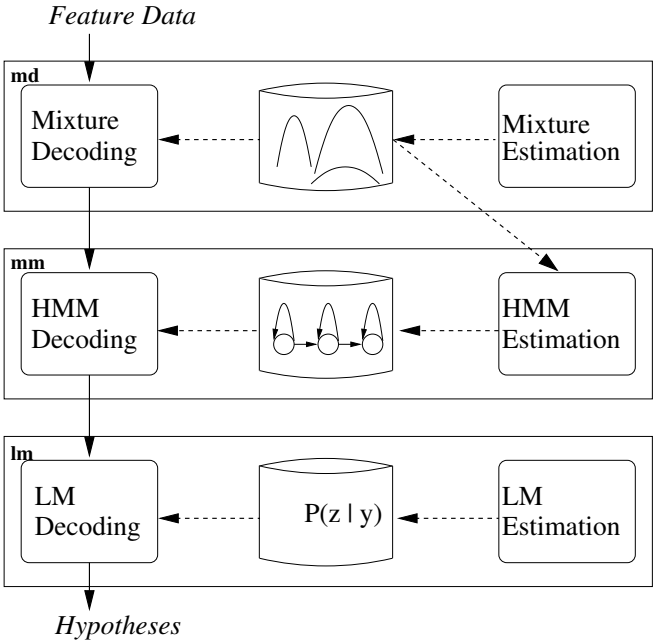
HMMs: declarative configuration language for building structured models from elementary units; initialization; training based on Baum-Welch reestimation; efficient Viterbi beam-search decoding; (semi-)supervised model adaptation (MLLR/MAP)

n-gram: memory efficient estimation of n-gram statistics; n-gram estimation based on different smoothing techniques (most notably absolute discounting and backing-off/interpolation); efficient decoding of long-span models

ESMERALDA was designed with special focus on the development of recognition systems which can be embedded into “real-world”

[1] Due to space limitations we did not include any formulas in this extended abstract. The mathematical details can be found in the referenced literature.

Figure 1: System architecture of ESMERALDA.



applications. Therefore, a command line interface has been developed allowing for pipelined operation by cascading the particular modules.

4 Applications

ESMERALDA was successfully applied to a number of challenging pattern recognition problems which are summarized in the following.

4.1 Speech Recognition

Originally designed for speech recognition purposes the use of ESMERALDA within this domain has a fairly long history. Allowing for batch as well as for interactive speech recognition, signal recording and feature extraction (based on MFCCs) modules are integrated.

Within an incremental speech recognizer all calculations from feature extraction to language model search are carried out strictly time-synchronously [3]. In order to be able to produce recognition results for an utterance while the user is still speaking, i.e. the end of the input signal is not yet reached, an incremental processing strategy was developed. Additionaly the recognizer is capable of applying the constraints of a context-free grammar in conjunction with a statistical language model [4]. In [5] acoustic and articulatory information have been combined using ESMERALDA for robust speech recognition.

As prominent examples, the toolkit has been used for the development of online speech recognizers embedded in intelligent human-robot interaction systems including automatic speaker identification [6]. Furthermore, a recognizer for accessing non-safety relevant functions of cars was realized including online adaptation to changing acoustic environments [7].

4.2 Handwriting Recognition

In order to recognize handwritten script in the last few years HMM-based techniques have been applied very successfully [8]. Depending on the recording process, handwritten script is either processed as online (i.e. trajectories of pen movements captured by pressure sensitive tablets) or offline data (i.e. digital document images acquired by e.g. scanners or video cameras). In the latter case lines of script are extracted from the images of the handwriting data and usually subject to several pre-processing and normalization operations. Subsequently, sequential data is extracted, by means of a sliding window approach resulting in a stream of features.

In recent years ESMERALDA has been successfully used for realizing offline handwriting recognition systems (cf. e.g. [9]). Feature streams are calculated from lines of handwritten script which are automatically extracted from the particular documents analyzed. By means of ESMERALDA, HMMs with Bakis topology are estimated for letters which are combined to word models using the framework's configuration language. The integration of Bi-gramm models restricts the decoding reasonably.

ESMERALDA has also been used for the realization of a whiteboard reading system which recognizes handwritten notes [9].

4.3 Biological Sequence Analysis

The functions of proteins, which are of major interest for life-science applications, are more or less directly connected to their primary structure, i.e. the underlying sequence of amino acids. Due to the linear structure of this biological data (complex) Profile HMMs with specific topologies have been applied very successfully to genomics and proteomics tasks [10].

By means of the ESMERALDA framework substantial enhancements of the basic approach to sequence alignment have been developed (cf. e.g. [11]) improving the detection of remotely related protein sequences which is especially relevant for pharmaceutical purposes. Therefore, a new feature representation for protein sequences was developed allowing for semi-continuous protein family HMMs with less complex model architectures.

5 Summary

In this paper we presented ESMERALDA, which evolved from a system for HMM-based speech recognition to a general development environment for statistical pattern recognition systems based on Markov models. The system showed its versatility in several application areas and is now available as open source software under the terms of the GNU General Public License (GPL).

Acknowledgements Most of the work described in this paper has been pursued at the authors' former affiliation, namely the University of Bielefeld, Germany, Applied Computer Science Group.

References

[1] L. R. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[2] X. Huang et al. *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Prentice Hall, 2001.

[3] G. A. Fink et al. Incremental speech recognition for multimodal interfaces. In *Proc. Annual Conference of the IEEE Industrial Electronics Society*, volume 4, pages 2012–2017, 1998.

[4] S. Wachsmuth et al. Integration of parsing and incremental speech recognition. In *Proc. European Signal Processing Conference*, volume 1, pages 371–375, 1998.

[5] K. Kirchhoff et al. Combining acoustic and articulatory information for robust speech recognition. *Speech Communication*, 37(3-4):303–319, 2002.

[6] G. A. Fink and T. Plötz. Integrating speaker identification and learning with adaptive speech recognition. In *2004: A Speaker Odyssey – The Speaker and Language Recognition Workshop*, pages 185–192, 2004.

[7] T. Plötz and G. A. Fink. Robust time-synchronous environmental adaptation for continuous speech recognition systems. In *International Conference on Spoken Language Processing*, 2002.

[8] T. Starner et al. On-line cursive handwriting recognition using speech recognition methods. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing*, volume 5, pages 125–128, 1994.

[9] M. Wienecke et al. Toward automatic video-based whiteboard reading. *Int. Journal on Document Analysis and Recognition*, 7(2–3):188–200, 2005.

[10] R. Durbin et al. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, 1998.

[11] T. Plötz and G. A. Fink. Pattern recognition methods for advanced stochastic protein sequence analysis using HMMs. *Pattern Recognition, Special Issue on Bioinformatics*, 39:2267–2280, 2006.