Available online at www.sciencedirect.com
SCIENCE DIRECT®
Speech Communication 45 (2005) 435–454
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Generative factor analyzed HMM for automatic speech recognition

Kaisheng Yao [a,*], Kuldip K. Paliwal [b], Te-Won Lee [a]

[a] *Institute for Neural Computation, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0523, USA*
[b] *School of Microelectronic Engineering, Griffith University, Brisbane, Queensland 4111, Australia*

Received 10 November 2003; received in revised form 6 November 2004; accepted 4 January 2005

Abstract

We present a generative factor analyzed hidden Markov model (GFA-HMM) for automatic speech recognition. In a standard HMM, observation vectors are represented by mixture of Gaussians (MoG) that are dependent on discrete-valued hidden state sequence. The GFA-HMM introduces a hierarchy of continuous-valued latent representation of observation vectors, where latent vectors in one level are acoustic-unit dependent and latent vectors in a higher level are acoustic-unit independent. An expectation maximization (EM) algorithm is derived for maximum likelihood estimation of the model.

We show through a set of experiments to verify the potential of the GFA-HMM as an alternative acoustic modeling technique. In one experiment, by varying the latent dimension and the number of mixture components in the latent spaces, the GFA-HMM attained more compact representation than the standard HMM. In other experiments with varies noise types and speaking styles, the GFA-HMM was able to have (statistically significant) improvement with respect to the standard HMM.

Keywords: Hidden Markov models; Factor analysis; Mixture of Gaussian; Speech recognition; Expectation maximization algorithm

1. Introduction

In the automatic speech recognition (ASR) problem, one is presented with multi-dimensional data with D^y dimension where it is assumed that the data is generated from acoustic sources that are modeled as discrete state q in a hidden Markov model (HMM) (Rabiner and Juang, 1993). The transition of the states is assumed to encode the transition of the speech unit and the content of the uttered speech can be inferred by the well-known Viterbi algorithm (Viterbi, 1967). The task in speech modeling for ASR within the

* Corresponding author. Tel.: +1 858 822 2720; fax: +1 858 565 7440.
E-mail addresses: kyao@ti.com (K. Yao), k.paliwal@me.gu.edu.au (K.K. Paliwal), tewon@ucsd.edu (T.-W. Lee).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.01.002

HMM framework is to obtain a compact and accurate model of the observations. However, this is a hard problem, since the observation vector is high dimensional and the elements in the observation vector contain second as well as higher order statistical information. Traditional approaches in modeling speech observations in an HMM make use of mixture of Gaussians (MoG) with usually a diagonal covariance matrix in each state, which implicitly models the intra-frame correlations.

Despite its pattern recognition appearance, the speech model in an HMM can be viewed in statistics as a latent representation. In particular, the discrete state q is the discrete latent representation of the speech unit and the discrete Gaussian index m in the MoG is the discrete latent representation of the density in that state. In this context, it is therefore natural to describe the D^y dimensional observation vector $\boldsymbol{y}(t)$ at time t as correlated in terms of a smaller set of D^x dimensional continuous-valued latent vector $\boldsymbol{x}(t)$. In this case, the most straightforward description of the continuous-valued latent representation of $\boldsymbol{y}(t)$ is given by the following linear model

$$y_n(t) = \sum_{l=1}^{D^x} \Lambda_{nl} x_l(t) + v_n(t), \quad n = 1, \ldots, D^y, \qquad (1)$$

where $y_n(t)$ denotes the nth element in vector $y(t)$ at time t. The $y_n(t)$ depends on linear combination of elements in $\boldsymbol{x}(t)$ with matrix $\boldsymbol{\Lambda} = [\Lambda_{nl}]_{D^y \times D^x}$. The density of $\boldsymbol{y}(t)$ is also related to the D^y-dimensional noise $\boldsymbol{v}(t)$ with element $v_n(t)$. Note that the problem in Eq. (1) is general, since without certain constraints imposed on the model, the solution is non-trivial.

In the context of the continuous-valued latent representation, Eq. (1) presents solutions with different physical meanings depending on the different constraints on the model (Roweis and Ghahramani, 1999; Frey, 1999). In independent component analysis (ICA) (Comon, 1994; Bell and Sejnowski, 1995) the constraints are as follows: (1) $v_n(t) = 0$; i.e., no distortions in observation (no additive noise) $\boldsymbol{y}(t)$, (2) element $x_l(t)$ in $\boldsymbol{x}(t)$ is independent from each other, and (3) At most one element $x_l(t)$ is Gaussian distributed or $x_l(t)$ has usually a non-Gaussian density. Maximum likelihood estimation of $\boldsymbol{\Lambda}$ leads to the ICA solution (Pearlmutter and Parra, 1997). Another interesting related model is independent factor analysis (IFA) (Attias, 1998). IFA is obtained by the following constraints: (1) element of $\boldsymbol{x}(t)$, $x_l(t)$, is independent and distributed as non-Gaussian density, (2) $\boldsymbol{v}(t)$ is distributed as diagonal Gaussian density.

Though advanced algorithms (Bell and Sejnowski, 1995; Attias, 1998; Amari et al., 2000; Cardoso, 1997; Hyvarinen et al., 2001) have been derived for signal processing within the framework of continuous-valued latent representation, there are few works applied to ASR. One reason is that the MoG can approximate any observation vector distribution given a sufficient number of model parameters and enough training data. Thus, by increasing the amount of training data and/or increasing number of model parameters, speech models by MoGs in HMMs can reach high recognition accuracy for input speech. Due to this claim, one might expect that the above continuous-valued latent representation may not be useful in ASR. However, there are several important differences in speech recognition research. Firstly, the number of parameters in the model and the amount of training data increase monotonically in order to achieve an improved performance. Secondly, the larger the number of parameters in a model, the larger the amount of training data is needed in order to have accurate estimation of the parameters. Thirdly, given the amount of training data, even when the number of parameters is increased, the performance of the model can easily reach a saturation point. These observations could give rise to problems in spontaneous speech recognition since the amount of training data is not sufficient for a reliable estimation of all acoustic units. Some heuristically justified approaches have been applied to address the above problems, for example, the method of parameter-tying (Bellegarda and Nahamoo, 1990). But parameter-tying has its own drawbacks since it considerably requires some artistry to design the way to share parameters.

Continuous-valued latent representation can be useful for modeling speech in a compact way. Note