Neural Networks 21 (2008) 1401–1409

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A constrained sequential EM algorithm for speech enhancement

Sunho Park, Seungjin Choi *

Department of Computer Science, Pohang University of Science and Technology, San 31 Hyoja-dong, Nam-gu, Pohang 790-784, Republic of Korea

ARTICLE INFO

Article history:
Received 7 September 2007
Received in revised form
24 February 2008
Accepted 11 March 2008

Keywords:
Expectation maximization (EM)
Generalized auto-regressive model
Generalized exponential density
Kalman filter
Rao-Blackwellized particle filter
Speech enhancement

ABSTRACT

Speech enhancement is a fundamental problem, the goal of which is to estimate clean speech s_t, given a noise-contaminated signal $s_t + n_t$, where n_t is white or colored noise. This task can be viewed as a probabilistic inference problem which involves estimating the posterior distribution of hidden clean speech, given a noisy observation. Kalman filter is a representative method but is restricted to Gaussian distributions only. We consider the generalized auto-regressive (GAR) model in order to capture the non-Gaussian characteristics of speech. Then we present a constrained sequential EM algorithm where Rao-Blackwellized particle filters (RBPFs) are used in the E-step and model parameters are updated in a sequential manner in the M-step under positivity constraints for noise variance parameters. Numerical experiments confirm the high performance of our proposed method, compared to Kalman filter-based methods, in the task of sequential speech enhancement.

1. Introduction

Signals measured through microphones in real-world environments are always noisy data, thus, the enhancement of speech or the elimination of noise, plays a critical role for successful subsequent speech processing. Speech enhancement aims at estimating clean speech s_t, given a noise-contaminated signal $y_t = s_t + n_t$ where n_t is white (temporally independent) or colored (temporally dependent) noise. Various methods have been developed for speech enhancement. These include Wiener filter method (Lim & Oppenheim, 1978), spectral subtraction method (Boll, 1979), hidden Markov model (HMM)-based method (Sameti, Sheikhzadeh, Deng, & Brennan, 1998), signal subspace method (Ephraim & Trees, 1993), Kalman filter method (Paliwal & Basu, 1987), and H_∞ filter-based method (Shen & Deng, 1999). Most of these methods are iterative algorithms in nature.

Speech enhancement becomes more important than ever, as speech quality plays a critical role in automatic speech recognition systems. In embedded or mobile environments, sequential speech enhancement is more desirable than batch methods, since it requires less memory and lower computational complexity. Kalman filter has been a useful tool in speech enhancement (Paliwal & Basu, 1987) and was further extended, incorporating the expectation maximization (EM) optimization (Gannot, Burshtein, & Weinstein, 1998; Weinstein, Oppenheim, Feder, & Buck, 1994). Kalman gradient descent sequential (KGDS) algorithm is an exemplary sequential speech enhancement method (Gannot et al., 1998). However, Kalman filter is restricted to the case where speech modeling is based on Gaussian distribution. In contrast, particle filter can handle non-Gaussian distributions, because it is a sequential importance sampling method, computing statistical expectation with respect to rather complex distributions. Recently, particle filter was also used for speech enhancement (Vermaak, Andrieu, Doucet, & Godsill, 2002), where speech signal is modeled by time-varying auto-regressive (AR) model, assuming that the innovation sequence follows Gaussian distribution.

As in Gannot et al. (1998) and Vermaak et al. (2002), we formulate the speech enhancement as a probabilistic inference problem, employing a state space model. In this framework, the task of sequential speech enhancement involves integrations with respect to the posterior distribution over hidden variables (states vector), given an incoming observation signal (see Eq. (12)). In contrast to most existing methods, we use the generalized auto-regressive (GAR) model for a speech generating process, in order to capture the non-Gaussian characteristics of speech. Noise is assumed to be a colored Gaussian random process, where the conventional AR model is used with a Gaussian innovation sequence. Exploiting the analytical structure of the state space model, we employ the Rao-Blackwellization (Casella & Robert, 1996; Doucet, Godsill, & Andrieu, 2000b) where the posterior distribution over hidden variables is decomposed as two parts, one of which is analytically calculated by Kalman filter and the other of which is estimated by particle filter. We use Rao-Blackwellized particle filter (RBPF) for probabilistic inference which involves the sequential calculation of the integration with respect to the posterior distribution. Our earlier work is found in Park and Choi (2006).

In this paper we present a constrained sequential EM algorithm for speech enhancement, where the inference is carried out by

* Corresponding author. Tel.: +82 54 279 2259; fax: +82 54 279 2299.
E-mail address: seungjin@postech.ac.kr (S. Choi).

0893-6080/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2008.03.001

RBPFs in the E-step and model parameters are updated in a sequential manner under positivity constraints for noise variance parameters in the M-step. This RBPF-based constrained sequential EM is referred to as RBPF + csEM throughout this paper. The contributions of the paper are summarized:

- We employ the GAR model where the innovation sequence follows the generalized exponential distribution, which reflects the non-Gaussian characteristics. In contrast to the AR model, in such a case, the probabilistic inference becomes intractable. The posterior distribution over hidden variables (both speech and noise state variables) is computed by combining the particle filter with the Kalman filter. That is, the probabilistic inference is carried out by the RBPF.
- We propose a constrained sequential EM which estimates model parameters recursively with positivity constraints for some of parameters. It turns out that the constrained sequential EM together with Kalman filter (which is referred to as KF + csEM) also outperforms KGDS.

The rest of this paper is organized as follows. The next section describes a formulation of the probabilistic sequential speech enhancement. A state space model is introduced, where clean speech follows the generalized auto-regressive model and the noise is based on the standard auto-regressive model. Section 3 explains the inference part in our probabilistic sequential speech enhancement, while parameter estimation based on the proposed constrained sequential EM is described in Section 4. Experimental results and comparisons to Kalman filter-based methods, are given in Section 5, emphasizing the high performance of the proposed method. Finally, conclusions are drawn in Section 6.

2. Problem formulation

We first formulate the task of speech enhancement as a probabilistic inference problem, presenting a state space model by considering a GAR model for the speech generating process. Then we outline the proposed constrained sequential EM algorithm for speech enhancement, the details of which are illustrated in subsequent sections.

2.1. GAR model

AR model is a widely-used linear modeling method, where the current value of a time series, s_t, is expressed as a linear sum of its past values, $\{s_{t-\tau}\}$, and an innovation v_t:

$$s_t = \sum_{\tau=1}^{p} \alpha_\tau s_{t-\tau} + v_t. \tag{1}$$

AR modeling involves determining coefficients $\{\alpha_\tau\}$ that provide a linear optimal fitting (in mean squared error sense) to a given time series $\{s_t\}$, assuming that the innovation v_t is Gaussian. AR model captures the dependence of the current value of a time series on its past values, through a linear model. The innovation contains a truly new information that is not found in past values of time series. AR modeling has been widely used in speech processing (Lim & Oppenheim, 1978).

GAR model is a non-Gaussian extension of the AR model, where the same linear model (1) is used but the innovation v_t is assumed to be drawn from the generalized exponential distribution (a.k.a. generalized Gaussian) with mean zero (Box & Tiao, 1992; Choi, Cichocki, & Amari, 2000) that is of the form

$$p(v; R, \beta) = \frac{R\beta^{1/R}}{2\Gamma(1/R)} \exp\left\{-\beta |v|^R\right\}, \tag{2}$$

where $\Gamma(\cdot)$ is the gamma function, $1/\beta$ determines the width of the density, and R is a parameter involving a shape of distribution.

Generalized exponential distribution accommodates a wide class of unimodal probability distribution. For example, $p(v; R, \beta)$ becomes Gaussian distribution for $R = 2$ and Laplacian distribution for $R = 1$. The value of R close to 1, well approximate the distribution of the innovation sequence for real speech signal (see Fig. 2(a)). In contrast to the AR model where the probabilistic inference is carried out by the Kalman filter, the probabilistic inference in the GAR model is intractable. This leads us to consider the Rao-Blackwellized particle filter that is described in Section 3.

2.2. State space model

The noise-contaminated observed signal y_t is modeled as a linear sum of clean speech s_t and noise n_t:

$$y_t = s_t + n_t, \tag{3}$$

where the clean speech and noise follow GAR and AR models, respectively, i.e.,

$$s_t = \sum_{\tau=1}^{p} \alpha_\tau s_{t-\tau} + v_t, \tag{4}$$

$$n_t = \sum_{\tau=1}^{q} \gamma_\tau n_{t-\tau} + u_t, \tag{5}$$

where v_t obeys the generalized exponential distribution and u_t is drawn from Gaussian distribution, i.e.,

$$v_t \sim p(v_t; R, \beta) = \frac{R\beta^{1/R}}{2\Gamma(1/R)} \exp\left\{-\beta |v|^R\right\}, \tag{6}$$

$$u_t \sim \mathcal{N}(u_t; 0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2} u_t^2\right\}. \tag{7}$$

We assume that s_t and n_t are statistically independent.

We define $\boldsymbol{s}_t \in \mathbb{R}^p$ and $\boldsymbol{n}_t \in \mathbb{R}^q$ as

$$\boldsymbol{s}_t = [s_t, s_{t-1}, \ldots, s_{t-p+1}]^\top, \tag{8}$$

$$\boldsymbol{n}_t = [n_t, n_{t-1}, \ldots, n_{t-q+1}]^\top. \tag{9}$$

Concatenating these two vectors, we define a state vector $\boldsymbol{x}_t = [\boldsymbol{s}_t^\top, \boldsymbol{n}_t^\top]^\top \in \mathbb{R}^{p+q}$. Accommodating generative models (4) and (5) for speech and noise, the state space model that we consider, is of the form:

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{B}\boldsymbol{r}_t, \tag{10}$$

$$y_t = \boldsymbol{b}^\top \boldsymbol{x}_t, \tag{11}$$

where

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_s & 0 \\ 0 & \boldsymbol{A}_n \end{bmatrix} \in \mathbb{R}^{(p+q)\times(p+q)},$$

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{b}_s & 0 \\ 0 & \boldsymbol{b}_n \end{bmatrix} \in \mathbb{R}^{(p+q)\times 2},$$

$$\boldsymbol{r}_t = [v_t, u_t]^\top \in \mathbb{R}^2,$$

$$\boldsymbol{b} = \begin{bmatrix} \boldsymbol{b}_s \\ \boldsymbol{b}_n \end{bmatrix} \in \mathbb{R}^{p+q},$$

and

$$\boldsymbol{b}_s = [1, 0, \ldots, 0]^\top \in \mathbb{R}^p,$$

$$\boldsymbol{b}_n = [1, 0, \ldots, 0]^\top \in \mathbb{R}^q.$$

The state transition matrix $\boldsymbol{A} \in \mathbb{R}^{(p+q)\times(p+q)}$ is a block diagonal matrix where $\boldsymbol{A}_s \in \mathbb{R}^{p\times p}$ and $\boldsymbol{A}_n \in \mathbb{R}^{q\times q}$ are given by

$$\boldsymbol{A}_s = \begin{bmatrix} \alpha_1 & \alpha_2 & \cdots & \cdots & \alpha_p \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{bmatrix},$$