Rao-Blackwellized particle filter를 이용한 순차적 음성 강조

박선호° 최승진
포항공과대학교대학원 컴퓨터 공학과
{titan°, seungjin}@postech.ac.kr

Rao-Blackwellized Particle Filtering for Sequential Speech Enhancement

Sunho Park°, Seunjin Choi
Postech, Computer Science and Engineering

요 약

we present a method of sequential speech enhancement, where we infer clean speech signal using a Rao-Blackwellized particle filter (RBPF), given a noise-contaminated observed signal. In contrast to Kalman filtering-based methods, we consider a non-Gaussian speech generative model that is based on the generalized auto-regressive (GAR) model. Model parameters are learned by a sequential Newton-Raphson expectation maximization (SNEM), incorporating the RBPF. Empirical comparison to Kalman filter, confirms the high performance of the proposed method.

1. 서 론

Speech enhancement is a fundamental problem, which aims at estimating clean speech, given noise-contaminated signals. Various speech enhancement methods have been developed. Specially in this paper, we consider the sequential speech enhancement algorithm.

In this paper, we consider the generalized auto-regressive (GAR) model for clean speech, in order to accommodate the non-Gaussian characteristics of speech. With the GAR model, we formulate the speech enhancement problem as a Rao-Blackwellized particle filtering. Associated model parameters are learned by a sequential Newton-Raphson expectation maximization (SNEM) method. Empirical comparison to the Kalman filter, confirms that the proposed method based on the Rao-Blackwellized particle fitler, is superior to alman filter, in the task of sequential speech enhancement.

2. Generalized Auto-Regressive Model

The auto-regressive (AR) model is a widely-used linear modelling method, where the current value of a time series, s_t, is expressed as a linear sum of its past values, $\{s_{t-d}\}$, and an innovation v_t:

$$s_t = \sum_{d=1}^{p} \alpha_d s_{t-d} + v_t. \quad (1)$$

The AR modelling involves determining coefficients $\{\alpha_d\}$ that provide a linear optimal fitting to given time series $\{s_t\}$, assuming that the innovation v_t is Gaussian.

The generalized auto-regressive (GAR) model is a non-Gaussian extension of the AR model, which adopts the same linear model (1) but assumes the innovation v_t is drawn from the generalized exponential (GE) distribution (a.k.a. generalized Gaussian) with mean zero [1] that is of the form

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

where $1/\beta$ determine width of the density and R is a parameter which determines a shape of distribution.

The GE 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 GAR model reflects the non-Gaussian characteristics of speech signals. However, in such a model, the probabilistic
inference is intractable, in contrast to Kalman filters. This leads us to consider the Rao-Blackwellised Particle Filter described in Sec. 4.

3. State-Space Models

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

$$y_t = s_t + n_t, \quad (1)$$

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

$$s_t = \sum_{d=1}^{p} \alpha_d s_{t-d} + v_t, \quad (2\text{-}3)$$
$$n_t = \sum_{d=1}^{q} \gamma_d n_{t-d} + u_t,$$

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

$$\boldsymbol{s_t} = [s_t, s_{t-1}, \cdots, s_{t-p+1}]^\top, \ \boldsymbol{n_t} = [n_t, n_{t-1}, \cdots, n_{t-q+1}]^\top.$$

Concatenating these two vectors, we define a state vector

$x_t = [s_t^\top, n_t^\top]^\top \in \mathbb{R}^{p+q}$. Accommodating generative models (2) and (3) for speech and noise, the state-space model that we consider, is of the form:

$$\mathbf{x}_t = \mathbf{A}\mathbf{x}_{t-1} + \mathbf{B}\mathbf{r}_t,$$
$$y_t = \boldsymbol{b}^\top \mathbf{x}_t,$$

where

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s & 0 \\ 0 & \mathbf{A}_n \end{bmatrix}, \mathbf{B} = \begin{bmatrix} \mathbf{b}_s & 0 \\ 0 & \mathbf{b}_n \end{bmatrix},$$
$$\mathbf{r}_t = [\mathrm{v}_t, \mathrm{u}_t]^\top, \ \mathbf{b}^\top = [\mathbf{b}_s^\top, \mathbf{b}_n^\top]$$

and

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

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

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

and $\boldsymbol{A}_n$ is constructed in a similar way.

4. Formulation with Rao-Blackwellised Particle Filter

The posterior density of the hidden state can be decomposed as [2]

$$p(\boldsymbol{x}_{0:t}|y_{1:t}) = p(\boldsymbol{n}_{0:t}|\boldsymbol{s}_{0:t}, y_{1:t}) p(\boldsymbol{s}_{0:t}|y_{1:t}), \quad (4)$$

which leads us to estimate the speech and noise separately.

The posterior density of the noise, $p(\boldsymbol{n}_{0:t}|\boldsymbol{s}_{0:t}, y_{1:t})$, can be analytically computed using the Kalman filter, if we know the marginal posterior density $p(\boldsymbol{s}_{0:t}|y_{1:t})$. Only the posterior density of speech, $p(\boldsymbol{s}_{0:t}|y_{1:t})$, is approximately calculated through a sampling method. This method, motivated by the decomposition (4), is known as RBPF.

Next section illustrates the details on the inference.

5. Inference

5.1 Inference for the state of noise

Let $\boldsymbol{s}_t^{(i)}$ $(i = 1,2,\cdots,N)$ be particles of clean speech and σ^2 be the variance of u_t in the noise AR model. We sample $\boldsymbol{s}_t^{(i)}$ by the method described in Sec. 5.2 and then propagate the mean $\boldsymbol{\mu}_t^{(i)}$ and covariance $\boldsymbol{\Sigma}_t^{(i)}$ of $\boldsymbol{n}_t$ with a Kalman filter as follows:

$$\begin{aligned} \boldsymbol{\mu}_{t|t-1}^{(i)} &= \boldsymbol{A}_n \boldsymbol{\mu}_{t-1|t-1}^{(i)}, \\ \boldsymbol{\Sigma}_{t|t-1}^{(i)} &= \boldsymbol{A}_n \boldsymbol{\Sigma}_{t-1|t-1}^{(i)} \boldsymbol{A}_n^\top + \sigma^2 \boldsymbol{b}_n \boldsymbol{b}_n^\top, \\ \boldsymbol{\Gamma}_t^{(i)} &= \boldsymbol{b}_n^\top \boldsymbol{\Sigma}_{t|t-1}^{(i)} \boldsymbol{b}_n, \end{aligned}$$

$$\begin{aligned} y_{t|t-1}^{(i)} &= \boldsymbol{b}_n^\top \boldsymbol{\mu}_{t|t-1}^{(i)} + \boldsymbol{b}_s^\top \boldsymbol{s}_t^{(i)}, \\ \boldsymbol{\mu}_{t|t}^{(i)} &= \boldsymbol{\mu}_{t|t-1}^{(i)} - \boldsymbol{\Sigma}_{t|t-1}^{(i)} \boldsymbol{b}_n \left[\boldsymbol{\Gamma}_t^{(i)}\right]^{-1} (y_t - y_{t|t-1}^{(i)}), \\ \boldsymbol{\Sigma}_{t|t}^{(i)} &= \boldsymbol{\Sigma}_{t|t-1}^{(i)} - \boldsymbol{\Sigma}_{t|t-1}^{(i)} \boldsymbol{b}_n \left[\boldsymbol{\Gamma}_t^{(i)}\right]^{-1} \boldsymbol{b}_n^\top \boldsymbol{\Sigma}_{t|t-1}^{(i)}, \end{aligned}$$

Finally the predictive density [2] is given by

$$p(y_t|y_{1:t-1}, \boldsymbol{s}_{0:t}) = N(y_t; y_{t|t-1}, \boldsymbol{\Gamma}_t).$$

5.2 Inference for the state of clean speech

For approximately estimating $p(\boldsymbol{s}_{0:t}|y_{1:t})$, we update the importance weights given by [2]

$$w_t^{(i)} \propto w_{t-1}^{(i)} \frac{p(s_t^{(i)}|\boldsymbol{s}_{t-1}^{(i)})}{\hat{p}(s_t^{(i)}|\boldsymbol{s}_{t-1}^{(i)})} \hat{p}(y_t|\boldsymbol{s}_{0:t-1}^{(i)}, y_{1:t-1}),$$

where $\hat{p}(s_t|\boldsymbol{s}_{t-1})$ is Gaussian-approximation allowing us to drawing new samples from Gaussian and $\hat{p}(y_t|\boldsymbol{s}_{0:t-1}, y_{1:t-1})$ is given by

$$\hat{p}(y_t|\boldsymbol{s}_{0:t-1}, y_{1:t-1}) = \int p(y_t|\boldsymbol{s}_{0:t}, y_{1:t-1}) \hat{p}(s_t|\boldsymbol{s}_{t-1}) ds_t.$$

From evaluation of the sequential importance weights, we can sequential estimate $p(\boldsymbol{s}_{0:t}|y_{1:t})$ within the RBPF.

6. Parameter Learning

Parameters to be learned, are $\theta = \{\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \sigma^2\}$, where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_p]^{\mathrm{top}}$ is the set of speech GAR model coefficients in (2), β is the parameter determining the width of the generalized exponential density in (2), $\boldsymbol{\gamma} = [\gamma_1, \cdots, \gamma_q]^\top$ is the set of noise AR model coefficients in (3), and σ_n^2 is the variance of e_t in the noise AR model (3).

In this section, we obtain prameter updating rule by a sequential Newton expectation maximization (SNEM) [3,4] allgorithm using the approximated posterior distribution of

$$\begin{aligned} \widehat{\boldsymbol{\theta}}_{t+1} &= \widehat{\boldsymbol{\theta}}_t + \boldsymbol{H}_{t+1}^{-1} \boldsymbol{\varphi}(\widehat{\boldsymbol{\theta}}_t), \\ \widehat{\boldsymbol{H}}_{t+1} &= \lambda \widehat{\boldsymbol{H}}_t + \boldsymbol{H}(\widehat{\boldsymbol{\theta}}_t), \end{aligned}$$

hidden variables determined by the RBPF.

where $\hat{\theta}_t$ is a solution of parameter at time t, ϕ and $\boldsymbol{H}$ are given by

$$\begin{aligned} \boldsymbol{\varphi}(\theta_t) &= \mathbb{E}\{\nabla_{\theta_t} \log p(\boldsymbol{x}_t; \theta_t) | y_{1:t}, \widehat{\theta}_{t-1}\}, \\ \boldsymbol{H}(\theta_t) &= -\mathbb{E}\{\nabla_{\theta_t}^2 \log p(\boldsymbol{x}_t; \theta_t) | y_{1:t}, \widehat{\theta}_{t-1}\}. \end{aligned}$$

The above equation need the expectation of given function with posterior density $p(\boldsymbol{x}_t|y_{1:t})$. This expectation is approximately estimated by using the RBPF. From updating rule for parameters, we recursively update parameters for both speech and noise model.

We denote that RBPF+SNEM is combining the RBPF and SNEM. In addition, RBPF+QNEM is combining the RBPF and Quasi-Newton EM which is reduced version of SNEM. Also we estimate posterior density $p(x_{1:t}|y_{1:t})$ by Kalman filter in the assumption that density of speech is Gaussian. In this case, we denote Kalman+SNEM algorithm. Although the Kalman+SNEM does not consider non-Gaussianity of speech, it show the better stability than Kalman-gradient-descent-sequential (KGDS) which updates parameters sequentially using gradient type method [4].

7. Experimental Result

For experiments, we use a speech signal from web site (http://www.ece.mcmaster.ca/~reilly/html/projects/dereverb/speechRHINTE.wav). It is re-sampled 8 Khz and first 5000 data oints are used for experiment. In the experiment, we compare our algorithm to the existing sequential speech algorithm which assume hat the density of speech is Gaussian.

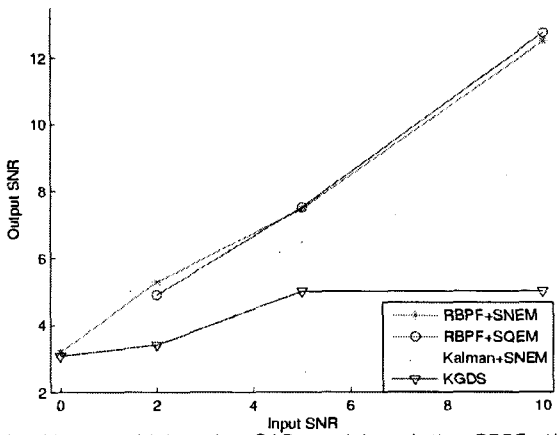
At first, we set the order of speech model to 12 which is generally used for speech modelling, i.e., p of the GAR model is 12. Next, to measure performance of given speech enhancement algorithm, we use an improvements of signal to noise ratio (SNR) between input and output signal. We obtain the output SNR from averaging independent 30 simulation for each experiment. we set the number of particles N to 200 and R of the GE density to 1.25 for the RBPF. Finally we set q of noise model to 5 and assume that this noise model is unknown and stationary.

Now, we compare our algorithm to the existing sequential speech enhancement methods. We calculate the output SNR for different condition where input SNR is changed from 0db to 10db.

From figure 1, we found that our methods outperform to the Kalman filter based method. Specially our proposed methods give the high performance when input SNR is above 5db. Since the observation is closer to speech signal in the high input SNR condition, the density of it is more different from Gaussian density. Hence, in this case, our method using the GE density is more appropriate than the Kaman filter based method. From above result, we conclude that that our methods using GE density outperform to the existing method based Gaussian density.

8. Conclusions

We proposed a new sequential speech enhancement



algorithm combining the GAR model and the RBPF. We used the GAR model as the peech model and the RBPF as an inference tool. Because the GAR model can deal with more general density than Gaussian, we can choose the density proper to real speech signal. Therefore, the estimated result is more close to the true speech signal.

In future, our algorithm can be extended to source separation. In this case, two GAR models correspond two speech signals. From this assumption, our sequential updating algorithm in the Sec. 6 can find the parameters of two speech signals sequentially.

Acknowledgments

This work was supported by ITEP Brain Neuroinformatics Program and Korea MIC under ITRC support program supervised by he IITA (IITA-2005-C1090-0501-0018).

참고 문헌.

[1] S. Choi, A. Cichocki and S. Amari , "Flexible Independent Component Analysis", Journal of VLSI Signal Processing, vol. 26, p.p. 25-38, 2000

[2] A Doucet, N. de Freitas, K. Murphy and S. Russell "Rao-Blackwellised particle filtering for dynamic Bayesian networks", In Proceedings of Uncertainty in Artificial Intelligence, p.p. 176-183, 2000

[3] D. M. Titterington, "Recursive Parameter Estimation Using Incomplete Data", Journal of the Royal Statistical Society, p.p. 257-267, vol. 46, 1984

[4] L. Frenkel and M. Feder, "Recursive Expectation-Maximization(EM) Algorithms for Time-Varying Parameters with Applications to Multiple Target Tracking", IEEE Trans. Signal Processing, vol. 47, p.p 306-320, 1999

[5] S. Gannot, D. Burshtein and E. Weinstein, "Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithms", vol. 6, pp. 373-385, 1998