Digital Signal Processing 73 (2018) 106–116

Contents lists available at ScienceDirect

Digital Signal Processing

www.elsevier.com/locate/dsp

Variational Bayesian learning for removal of sparse impulsive noise from speech signals

Hongjie Wan [a,*], Xin Ma [b], Xuebin Li [a]

[a] Department of Information Engineering, Beijing University of Chemical Technology, China
[b] School of Information Science and Engineering, Shandong University, China

ARTICLE INFO

Article history:
Available online 21 November 2017

Keywords:
Speech enhancement
Sparse impulsive noise
Variational Bayesian
Kalman smoother
Additive outlier

ABSTRACT

In this paper, a new variational Bayesian (VB) learning algorithm is proposed to remove sparse impulsive noise from speech signals. The clean signal is modeled using an autoregressive (AR) model on frame basis. The contaminated signal is modeled as the sum of the AR model of the clean speech signal, a sparse noise term and a dense Gaussian noise term. The sparse noise and the dense Gaussian noise terms model the large additive values caused by the impulsive noise and the small additive values or Gaussian noise, respectively. A hierarchical Bayesian model is constructed for the contaminated signal and a VB framework is used to estimate the parameters of the model. The AR model parameter estimation, the speech signal recovery and the sparse impulsive noise removal are carried out simultaneously. The proposed algorithm starts from random initial values and it does not require training and a threshold as compared to other methods. Experiments are performed using a standard speech database and impulsive noise generated from a probabilistic impulsive noise model and real impulsive noise. The comparison of obtained results with other methods demonstrates the performance of the proposed method.

1. Introduction

Various types of background noises can degrade the quality of speech signals, which may affect speech intelligibility and automatic speech recognition [1]. Among these background noises, one specific noise is the impulsive noise. The impulsive noise can be found in adverse communication channel environment, signal dropouts [2,3], archived gramophone recordings [4], and in daily life such as the clicks of keyboard and the hitting of rain drops on the windshield of a moving car [5]. Broadband noise also exists with the impulsive noise [6–9], and elimination of both these noises is difficult [7].

Various methods have been proposed to minimize the effects of impulsive noise. These methods can mainly be classified into two categories – AR based [2,6–8,10–14] and not AR based [5,15–17]. In [2], the residuals of the speech signal obtained by inverse filtering are further processed by matched filtering to detect the impulsive noise and then the corrupted signal is interpolated by the neighbors. Warped linear prediction based detection method has been proposed to improve the detection performance [10]. Bayesian model is used to detect and estimate the signal in [11] but the AR coefficients are not estimated simultaneously. The simultaneous estimation of the AR coefficients, the detection of the impulsive noise and the reconstruction of the signal have been carried out in [8,12,13]. In addition, these methods can also eliminate one type of the broadband noise – the white Gaussian noise in the same time. In [13], the extended Kalman filter (EKF) based method is proposed and it has been proved that the combination of the forward and backward time (reversing the time axes) prediction improves the accuracy of signal estimation. Bidirectional detection and reconstruction have also been proposed in [7], where the detection results using normal time signal and the time reversed signal have been combined using some fusion rules designed by the authors, and the normal time and time reversed reconstruction results have been linearly combined with the weights obtained based on minimum variance unbiased estimation. The features of the above mentioned methods mainly include one or more of the three aspects described as follows. The first is that, the estimation of the AR coefficients, the detection of the impulsive noise and the reconstruction of the original signal are not carried out simultaneously. The second is that the combination of the forward and backward estimation is not integrated. The third is that thresholds should be used. The methods that are not based on AR model including the wavelet based method [5,15], the SD-ROM (Signal Dependent Rank Order Mean) based method [16] and STFT (short time Fourier transform) method [17], wherein threshold is also used for detec-

* Corresponding author.
E-mail addresses: wanhj@mail.buct.edu.cn (H. Wan), max@sdu.edu.cn (X. Ma), lixb@mail.buct.edu.cn (X. Li).

https://doi.org/10.1016/j.dsp.2017.11.007
1051-2004/

tion. Although many algorithms have been proposed to remove the impulsive noise but most of them are focused on music signals.

The samples corrupted by large impulses can also be regarded as additive outliers that are different from innovation outliers. Several methods have been proposed to deal with these outliers including the detection of the outliers [18,19] and the use of distributions robust to the presence of outliers [20]. If some successive data points corrupted by the audio impulsive noise are large then they can be regarded as outliers. Multiple successive outliers are also called patches of outliers, but are different with audio impulsive noise as audio impulsive noise also contains small values. Therefore, processing of the large values in audio impulsive noise processing method is related with methods dealing with patches of outliers [11], but these methods cannot be used to deal with impulsive noise directly. These methods include the diagnostic method for ARIMA (autoregressive integrated moving average model) [21], the Gibbs sampling based method for AR model [22], and the diagnostic method based on local influence for ARIMA model [23]. However, simultaneous estimation of the signal, detection of the outliers and reconstruction of the signal are not integrated in these methods.

In this work, a new method is proposed to recover a speech signal that is contaminated by sparse impulsive noise. The word sparse is used here because the number of samples corrupted by the impulsive noise in speech signal may be fewer than the total number of speech samples [3,4]. In the proposed method, the estimation of the AR coefficients, estimation of the sparse impulsive noise and recovery of the speech signal are carried out in an integrated manner. The clean speech signal is modeled by an autoregressive (AR) model on frame basis. The corrupted signal is modeled as the sum of three terms, the AR model of the clean signal, a Gaussian noise term and a sparse noise term. The modeling of the corrupted signal is similar to that of robust principal component analysis [24–26], but is different from the traditional modeling methods of additive noise in AR models using Kalman filtering for speech enhancement [27–29]. The large values in the impulsive noise are modeled by the sparse noise term while the small values are modeled by the dense Gaussian distribution. The white Gaussian noise is also modeled by the dense Gaussian distribution if white Gaussian noise exists. The model is explained in detail in Section 2 and Section 3. The relation between the clean speech signal generation process of the AR model and the contaminated observation can be transformed into a state space model. A hierarchical Bayesian model is constructed to obtain the parameters of the state space model. The sparsity of the impulse noise is automatically achieved by placing the Automatic Relevance Determination (ARD) prior over the sparse noise term [30]. The ARD prior has been used in sparse signal recovery, including relevance vector machine [31], compressive sensing [32] and robust principal component analysis [26]. In order to estimate the parameters of the hierarchical Bayesian model, a maximum likelihood based method [33] or a Markov Chain Monte Carlo (MCMC) method [34] can be used. Variational Bayesian (VB) framework [35] can be adopted to approximate the posterior of the parameters because it can avoid overfitting and its computational cost is far lower than the MCMC. In this paper, the expected statistics of the hidden states are estimated by Kalman smoother [36] iterations and the posteriors of other parameters are estimated through M steps of a VB framework. In the proposed method, the forward and backward estimation is combined by the Kalman smoother iteration, which is different to the Kalman filtering as it only proceeds in a single direction. The proposed method is an unsupervised method that starts with random initial values and has no requirements of threshold and tuning of the parameters.

The organization of the rest of the paper is as follows. In Section 2, the formulation of speech signal corrupted by the additive impulsive noise is transformed into a state space model, and then a hierarchical Bayesian model is constructed accordingly. The clean signal, the parameters of the model and the sparse impulsive noise are estimated using a VB framework in Section 3. Experiments and analysis using sentences selected from standard speech database and probabilistic impulsive noise generation model are presented in Section 4 followed by Section 5 concluding this paper.

2. Bayesian modeling for speech corrupted by sparse impulsive noise

2.1. Modeling of speech signal corrupted by sparse impulsive noise

Let $\mathbf{Y} = \{y_1, \cdots, y_n, \cdots, y_N\}$ be one frame of the clean speech signal, and then it can be modeled by an AR model as follows

$$y_n = \sum_{m=1}^{p} a_m y_{n-m} + \varepsilon_n \tag{1}$$

where $\mathbf{a} = [a_1, a_2, \cdots, a_p]$ is the vector containing the AR coefficients, p is the model order, and ε_n is the innovation at time n. The frame length N is 320 which corresponds to 20 ms of speech signal under the 16 kHz sampling frequency.

Assuming the clean signal y_n in Eq. (1) is contaminated by the sparse noise e_n corresponding to the large values in the impulsive noise, and then the corrupted signal can be expressed as:

$$x_n = y_n + v_n + e_n \tag{2}$$

where v_n is the dense Gaussian noise which models the small impulse values or the white Gaussian noise.

Let one frame of the corrupted signal be denoted as $\mathbf{x} = \{x_1, x_2, \cdots, x_N\}$, then the relationship of the corrupted signal and the underlying AR model must be transformed into state space form:

$$\mathbf{y}_n = \mathbf{A}\mathbf{y}_{n-1} + \mathbf{c}\varepsilon_n \tag{3}$$

$$x_n = \mathbf{c}^T \mathbf{y}_n + v_n + e_n \tag{4}$$

where $\mathbf{y}_n = \left[y_n, y_{n-1}, \cdots, y_{n-p+1}\right]^T$ is the state vector at time n, and $\mathbf{c} = [1, 0, \cdots, 0]^T$ is a constant vector. An auxiliary state $\mathbf{y}_0$ is introduced here as the initial state, then the collections of the state vectors can be written as $\bar{\mathbf{Y}} = \{\mathbf{y}_0, \mathbf{y}_1, \cdots, \mathbf{y}_N\}$. The state transition matrix $\mathbf{A}$ of (3) is:

$$\mathbf{A} = \begin{bmatrix} & \mathbf{a}^T \\ \mathbf{I}_{p-1} & 0_{(p-1)\times 1} \end{bmatrix}_{p\times p} \tag{5}$$

In above equation, $\mathbf{I}_{p-1}$ represents an $(p-1)\times(p-1)$ identity matrix, while $0_{(p-1)\times 1}$ denotes an $(p-1)\times 1$ vector with zero elements.

2.2. Probabilistic formulations

The innovation term ε_n is assumed to be a normally distributed variable with zero mean and covariance precision β:

$$p(\varepsilon_n) = N(\varepsilon_n | 0, \beta^{-1}) \tag{6}$$

The sparse noise term e_n is normally distributed with zero mean and precision λ_n at each time n as in Eq. (7), which is an ARD prior [30]. Using this ARD prior, most of the λ_n will tends to infinity value or a relatively large value, and thus the sparse noise term e_n will concentrate at zero values. This is the process of automatic relevance determination that provides the sparsity of the impulsive noise. Further details about ARD can be found in [30].

$$p(e_n) = N(e_n | 0, \lambda_n^{-1}) \tag{7}$$