1

Fundamental Frequency Estimation In Speech Signals With Variable Rate Particle Filters

Geliang Zhang, and Simon Godsill, *Member, IEEE*

***Abstract*—Fundamental frequency estimation, known as pitch estimation in speech signals is of interest both to the research community and to industry. Meanwhile, the particle filter is known to be a powerful Bayesian inference method to track dynamic parameters in non-linear state-space models. In this paper, we propose a speech model under a time-varying source-filter speech model, and use variable rate particle filters (VRPF) to develop methods for estimation of pitch periods in speech signals. A Rao-Blackwellised variable rate particle filter (RB-VRPF) is also implemented. The proposed VRPF and RBVRPF are compared with a state-of-the-art pitch estimation algorithm, the YIN algorithm. Simulation results show that more accurate estimation of pitch can be obtained by VRPF and RBVRPF even under strong background noise conditions.**

***Index Terms*—variable rate particle filters, pitch estimation, Rao-Blackwellisation, source-filter model**

I. Introduction

ROBUST pitch estimation algorithms for speech signals have wide application and thus have been proposed in many papers. Previous algorithms are mainly based on time domain and frequency domain techniques; for example, a robust algorithm for pitch tracking (RAPT) algorithm and YIN, a fundamental frequency estimator for speech and music [1] [2] [3]. Most time domain algorithms are based on autocorrelation methods and the average magnitude difference function (AMDF) method, which can be used to estimate the periods of speech signals [4]. Recently a robust frequency domain algorithm for pitch estimation has also been proposed [5]. Some researchers have proposed a statistical method which chooses peaks from short time spectrum of speech signals [6]. Other methods which have been proposed to estimate glottal waves can also be used to extract pitch periods, include [7] [8]. However, very few of them can give satisfactory pitch tracking results under strong noise conditions, for example, when the Signal-to-Noise Ratio (SNR) is as poor as -5 dB to -10 dB.

Particle filters have been used widely in tracking applications since their development in recent decades [9] [10] [11]. However, little work has been done to apply the particle filter to pitch tracking. Recently it has been shown that the particle filter approach can be used to track pitch period, using a quasi-periodic speech signal model [12]. In this paper, we propose another particle filter approach to address the pitch tracking problem using a source-filter speech signal model, which is capable of tracking pitch period under very noisy conditions.

A possible source-filter model that can be used to capture the pitch period of speech signals is the time-varying autoregressive (AR) model driven by some source signals. Because of the near-periodic properties of voiced speech signals, the driving sources should themselves be near-periodic signals. A promising driving source model is proposed here in this paper, with an accompanying speech waveform model. Experiments have been carried out to test the performance of the proposed algorithm in various SNR conditions, showing that the proposed method can track pitch periods more successfully than state-of-the-art algorithms under noisy conditions.

The paper is organised as follows. Section 2 describes the source-filter speech model. In Section 3, a detailed description of the application of variable rate particle filters of the problem is presented. An initialization step for the particle filter using a joint optimization approach is derived in Section 4. Section 5 describes the details of the Rao-Blackwellisation approach to the previous variable rate particle filter. Section 6 gives experiment results for the proposed methods and compares them with the YIN algorithm. Finally, conclusions are drawn in Section 7.

II. Speech Model

A. A time-varying AR source-filter model

It is known that in human speech the fundamental period has a lower bound T_{low} and an upper bound T_{upp}. The n-th period T_n is thus modeled as,

$$T_n = T_{n-1} + \tau_n, \;\; T_{low} < T_n < T_{upp}. \tag{1}$$

The speech signal at time t in the current period n_t are represented as a periodic source input to a M-th order time-varying AR model,

$$s_t = \sum_{p=1}^{M} a_{n_t}^{p} s_{(t-p)} + V_t, \tag{2}$$

where V_t denotes the input source to the AR model and can be modeled as either near-periodic signals or glottal pulse sequences, while the current period n_t is $n : t \in [P_n, P_{n+1}]$. P_n is the time when the n-th period starts, i.e. $P_n = \sum_{i=1}^{n-1} T_i$.

The AR coefficients $a_{n_t}^{p}$ are assumed to change randomly between periods, but remain fixed within each period,

$$a_{n_t}^{p} = a_{n_t-1}^{p} + \tau_{a,p}, \tag{3}$$

where $\tau_n \sim U(\tau_{min}, \tau_{max})$. U refers to the uniform distribution and $\mathcal{N}$ denotes the Gaussian distribution through out the paper. $\tau_{min} = max[-\tau_T, T_{low} - T_{n-1}]$, and $\tau_{max} =$

The authors are with the Signal Processing and Communication Laboratory, Engineering Department, University of Cambridge, CB2 1PZ, UK. e-mail: (gz246@cam.ac.uk; sjg30@cam.ac.uk.

$min[\tau_T, T_{upp} - T_{n-1}]$, where τ_T is a fixed hyperparameter. $\tau_{a,p}$ can be sampled from $\mathcal{N}(0, \sigma^2_{a,p})$.

Finally, the voiced speech signal s_t is observed in Gaussian noise:

$$y_t = s_t + G_t. \tag{4}$$

G_t is sampled from $\mathcal{N}(0, \sigma^2_G)$. Values of hyperparameters such as σ_G and τ_T in these distributions are related with the extent of variations of parameters in the speech model and are given in the experiment section. We use $\mathbf{a}_{n_t}$ to denote $a^{1:M}_{n_t}$.

The characteristics of the speech signal model are largely determined by the input source, V_t. V_t can potentially be modeled using different quasi-periodic models, resulting in different performances. A particular source model is proposed here, described in the following subsection.

B. Input sources modeled as almost periodic signals

Because the voiced speech signal is almost periodic, with time-varying period, it is suggested that the input source can be modeled as an almost periodic signal itself. Such an approach has previously been used to model spectroscopy signals and music signals [13] [14]. Here it is proposed that a similar method can be used to model the input source to the speech production model as well,

$$V_t = \sum_{k=0}^{K} A^k_{n_t} \cos(kw_0^{n_t} t) + B^k_{n_t} \sin(kw_0^{n_t} t) + W_t \tag{5}$$

$$A^k_{n_t} = A^k_{n_t-1} + \epsilon_{A,k} \tag{6}$$

$$B^k_{n_t} = B^k_{n_t-1} + \epsilon_{B,k} \tag{7}$$

$$W_t \sim N(0, \tau^2_w) \tag{8}$$

Here $\epsilon_{A,k}$ and $\epsilon_{B,k}$ can be sampled from Gaussian distributions $N(0, \sigma^2_{A,k})$ and $N(0, \sigma^2_{B,k})$.

In equation (5), $K + 1$ denotes the number of harmonic components, i.e. cosine and sine waves, used in the input source model. The variable $w_0^{n_t}$ refers to the fundamental frequency of the current speech signal, which is the inverse of current pitch period T_n. Here we assume that $A^k_{n_t}$ and $B^k_{n_t}$ change slowly that they can be assumed fixed within each pitch period. In order to simplify the notation, we use $\mathbf{A}_{n_t}$ and $\mathbf{B}_{n_t}$ to denote $A^{1:K}_{n_t}$ and $B^{1:K}_{n_t}$.

III. Implementation of variable rate particle filter

We can use Bayesian filtering to recursively estimate hidden states $x_{1:t}$ from observable states $y_{1:t}$ [10], [15], using the following prediction and updating equations,

$$p(x_{1:t}|y_{1:t-1}) = p(x_t|x_{1:t-1})p(x_{1:t-1}|y_{1:t-1}) \tag{9}$$

$$p(x_{1:t}|y_{1:t}) \propto p(y_t|x_{1:t}, y_{1:t-1})p(x_{1:t}|y_{1:t-1}) \tag{10}$$

Thus if we can set the initial prior $p(x_1)$, we can use (9) and (10) to calculate the posterior distribution of $p(x_{1:t}|y_{1:t})$ and its marginal distributions once a new observable state y_t is received [10].

The variable rate particle filter approach uses a set of random, weighted 'particles' $x^{(i)}_{1:t}$ to approximate the posterior distribution for the unknown state variable sequence $x_{1:t}$ from the noisy data $y_{1:t}$,

$$p(x_{1:t}|y_{1:t}) \approx \sum_{i=1}^{N} w^{(i)}_t \delta(x_{1:t} - x^{(i)}_{1:t}), \tag{11}$$

where N denote the number of total particles.

In order to deal with the analytic intractability of the speech signal model and considering the fact that the period T is asynchronous with the sample time t, we adopt the variable rate particle filter here. Compared with the standard particle filter, the variable rate particle filter (VRPF) applies to cases when the state variables arrive at unknown times relative to the observation process [16]. This makes VRPF suitable for this speech signal model in which the pitch period arrives at a random rate relative to the observed speech signal samples.

In the VRPF, at a time t, the unknown parameters in the problem are $s^{1:t}$, $T_{1:n_t}$, $\mathbf{A}_{1:n_t}$, $\mathbf{B}_{1:n_t}$, and $a^{1:M}_{1:n_t}$. Fixed hyperparameters $(\sigma_T, \sigma_G, \sigma_V, \sigma_{a,p}, \tau_w, \sigma_{A,k}, \sigma_{B,k})$ are assumed here. Thus the hidden state vector $x_{1:t}$ is defined as,

$$x_{1:t} = [s_{1:t}, a^{1:M}_{1:n_t}, \mathbf{A}_{1:n_t}, \mathbf{B}_{1:n_t}, T_{1:n_t}]. \tag{12}$$

The algorithm of variable rate particle filter used here can be summarized as follows.

Algorithm 1:

Goal: Tracking $T_{1:n_t}$ which are contained in $x_{1:t}$, given $y_{1:t}$ and T_1.

```
1) Initialize $\{x^{(i)}_1\}^N_{i=1}$. Set up all the fixed hyperparameters and $\{w^{(i)}_1\}^N_{i=1} = \frac{1}{N}$. To initialize $\{x^{(i)}_1\}^N_{i=1}$, use the joint estimation technique based on the first period of speech data $y^{1:T_1}$, according to (14), see later. Then sample $\{s^{(i)}_1\}^N_{i=1}$ based on $\mathbf{a}_1, \mathbf{A}_1, \mathbf{B}_1, T_1$ according to (2) and (5). Set $P^{(i)} = T_1$.
2) for $t$ = 1:$t_{end}$ do
   a) for i = 1:$N$ do
      i) Set $n^{(i)}_t = n^{(i)}_{(t-1)}$.
      ii) While $t > P^{(i)}$:
         A) Add a new pitch period, $n^{(i)}_t \leftarrow n^{(i)}_t + 1$.
         B) Sample a new pitch period and other coefficients:
            $T^{(i)}_{n_t} \sim U(\max[T^{(i)}_{n_{t-1}} - \sigma_T, T_{low}], \min[T^{(i)}_{n_{t-1}} + \sigma_T, T_{upp}]),$
            $a^{n^{(i)}_t}_{(i),p} \sim \mathcal{N}(a^{n^{(i)}_t - 1}_{(i),p}, \sigma^2_{a,p}),$
            $A^{n^{(i)}_t}_{(i),k} \sim \mathcal{N}(A^{n^{(i)}_t - 1}_{(i),k}, \sigma^2_{A,k}),$
            $B^{n^{(i)}_t}_{(i),k} \sim \mathcal{N}(B^{n^{(i)}_t - 1}_{(i),k}, \sigma^2_{B,k}),$
         C) Update $P^{(i)} \leftarrow P^{(i)} + T^{(i)}_{n_t}$.
```

```
        iii) Sample new signal value: $s_i^t$ based on $a_{i,p}^{n_t^{(i)}}, A_{i,k}^{n_t^{(i)}}, B_{i,k}^{n_t^{(i)}}, T_{n_t^{(i)}}^{(i)}$ as (2) and (5). Now $[x_{1:t}^{(i)}] = [x_{1:t-1}^{(i)}, x_t^{(i)}]$, as defined in (34).
        iv) Compute importance weight $w_t^{(i)}$ of each particle:
                $w_t^{(i)} \propto w_{t-1}^{(i)} p(y_t | s_t^{(i)}, \sigma_G).$
    b) end for
    c) Renormalize $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}}$, $i = 1, 2, ..., N$.
    d) If $t = k * BlockSize$, where $k$ is a positive integra,
        i) Resample $\{x_{1:t}^{(i)}\}_{i=1}^N$ when $N_{eff} < N/2$. $N_{eff}$ denotes the effective sample size and is calculated as $N_{eff} = 1/\sum_{i=1}^N \tilde{w}_t^{(i)}$.
        ii) $\hat{T}_{n_t} = \sum_{i=1}^N \tilde{w}_t^{(i)} T_{n_t}^{(i)}$.
3) end for
```

One thing we would like to mention here is that it is not suggested that we make the decision of whether or not to resample when every signal sample becomes available, as it will decrease the robustness of the algorithm. Rather, we make the decision only after a certain length of samples have been processed, which is equal to the length of a pre-determined window size. The time length of this window function is usually about 32, 64 or 128 ms depending on the context.

IV. Initialization of particle filters

A. Motivation

In order to apply the particle filter to estimate the pitch period T_0 of speech signal using the time-varying AR model, it is necessary to estimate all the parameters used in the model except for the fixed parameters. However, if too many parameters need to be tracked simultaneously without any prior knowledge about their initial value, it means we need to estimate the state vector within a high dimensional space, which needs exponentially growing computation and number of particles to track them [17]. For example, in the case of modeling input sources as almost periodic signals, if 20 harmonics are assumed to be existing in the input source ($K = 20$) and a 11-order AR model is used ($M = 11$), the number of parameters involved in the whole model is $2 * K + M + 1$, which is 52. In order to estimate a 52-dimensional vector using a moderate number of particles, for example, 1000 particles, it will be necessary to have a good initialization method at the beginning of the algorithm to make the particle filter work.

B. Joint source-filter estimation method

It has been proposed in [8] that a joint source-filter optimization approach can be used to estimate glottal flow using the LF model of the glottal flow derivative when the input source is modeled as glottal pulses. It is suggested in our paper that after some modification on the model used in the input source, this joint source-filter optimization approach can be also applied here when the input source is modeled as almost periodic signals as a joint source-filter estimation technique to initialize the parameters used in the whole model. Details of how this technique can be modified to apply when input sources are modeled as almost periodic signals here are described in A.

Here we just display the results. If we write the parameters of the proposed almost periodic source-filter model except for T_n, i.e., $\{a_p, A_k, B_k\}_{p=1:M,\ k=0:K}$ (upper index n_t omitted here, see A), into a vector $\mathbf{a}$, where

$$\mathbf{a} = \begin{pmatrix} a_1, \ldots, a_M, & A_0, \ldots, A_K, & B_0, \ldots, B_K \end{pmatrix}^T \tag{13}$$

Then it is possible to jointly estimate the parameters in $\mathbf{a}$ using the following equation:

$$\mathbf{a} = \mathbf{R}^{-1}\mathbf{p} \tag{14}$$

where

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & -\mathbf{R}_2 \\ -\mathbf{R}_2^T & \mathbf{R}_3 \end{pmatrix} \tag{15}$$

where

$$\mathbf{R}_1 = \begin{pmatrix} C_{xx}(1,1) & \ldots & C_{xx}(M,1) \\ \vdots & \ldots & \vdots \\ C_{xx}(1,M) & \ldots & C_{xx}(M,M) \end{pmatrix} \tag{16}$$

$$\mathbf{R}_2 = \begin{pmatrix} \mathbf{R}_{2A} & \mathbf{R}_{2B} \end{pmatrix} \tag{17}$$

where

$$\mathbf{R}_{2A} = \begin{pmatrix} C_{Ax}^0(0,1) & \ldots & C_{Ax}^K(0,1) \\ \vdots & \ldots & \vdots \\ C_{Ax}^0(0,M) & \ldots & C_{Ax}^K(0,M) \end{pmatrix}, \tag{18}$$

and

$$\mathbf{R}_{2B} = \begin{pmatrix} C_{Bx}^0(0,1) & \ldots & C_{Bx}^K(0,1) \\ \vdots & \ldots & \vdots \\ C_{Bx}^0(0,M) & \ldots & C_{Bx}^K(0,M) \end{pmatrix}, \tag{19}$$

$$\mathbf{R}_3 = \begin{pmatrix} C_{CC}^{0,0}(0,0) & \ldots & C_{CC}^{0,2K+2}(0,0) & & \\ C_{CC}^{1,0}(0,0) & \ldots & C_{CC}^{1,2K+2}(0,0) & & \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ C_{CC}^{2K,0}(0,0) & \ldots & C_{CC}^{2K,2K+1}(0,0) & & \\ C_{CC}^{2K+1,0}(0,0) & \ldots & C_{CC}^{2K+1,2K+1}(0,0) & & \end{pmatrix}, \tag{20}$$

and

$$\mathbf{p} = \begin{pmatrix} -C_{xx}(0,1) \\ \vdots \\ -C_{xx}(0,M) \\ C_{Ax}^0(0,0) \\ \vdots \\ C_{Ax}^K(0,0) \\ C_{Bx}^0(0,0) \\ \vdots \\ C_{Bx}^K(0,0) \end{pmatrix}. \tag{21}$$

And those parameters used in the paper have been defined as,

$$\begin{cases} C_{xx}(i,j) & = \sum_{t=1}^{N_T} x^{t-i} x^{t-j} \\ C_{Ax}^k(i,j) & = \sum_{t=1}^{N_T} x^{t-i} A_k \cos(kw_0(t-j)) \\ C_{Bx}^k(i,j) & = \sum_{t=1}^{N_T} x^{t-i} B_k \sin(kw_0(t-j)) \\ C_{CC}^{m,n} & = \sum_{t=1}^{N_T} \mathbf{c}^m(t) \mathbf{c}_t^n \end{cases} \tag{22}$$

And $\mathbf{c}^k(t)$ denotes the $k-th$ element of the vector $\mathbf{c}(t)$ which is defined as follows:

$$\mathbf{c}(t) = \begin{pmatrix} A_0\cos(0w_0t) \\ A_1\cos(1w_0t) \\ \vdots \\ A_K\cos(Kw_0t) \\ B_0\sin(0w_0t) \\ B_1\sin(1w_0t) \\ \vdots \\ B_K\sin(Kw_0t) \end{pmatrix} \tag{23}$$

V. Rao-Blackwellisation of variable rate particle filter

It is well known that the optimal solution for linear state space models and Gaussian noise is the Kalman filter [18]. The Rao-Blackwellised particle filter separates the linear part and the nonlinear part of the model. Then it utilizes the Kalman filter to solve the linear and Gaussian part, and uses the particle filter to solve the nonlinear/non-Gaussian part of the model. This strategy enables the Rao-Blackwellised particle filter to save a lot of computing effort and reduce variance. As a reference for Rao-Blackwellised particle filters, please see [9] [15]. Some people have already adapted the Rao-Blackwellisation strategy to the variable rate particle filter in some other applications, such as for tracking and financial applications [19] [20].

Here we adapt the Rao-Blackewellisation method to the variable rate particle filter used in the previous section. To illustrate the process, we mainly rely on the framework provided in the section II.G in [9].

Recall that the dynamic model of speech signals used in the time-varying AR model is:

$$\begin{aligned} y^t &= s^t + G^t, \\ s^t &= \sum_{p=1}^{M} a_p^{n(t)} s^{(t-p)} + V^t. \\ a_p^{n(t)} &= a_p^{n(t)-1} + \tau_{a,p}, \\ T_n &= T_{n-1} + \tau_n, \\ V^t &= g^t + W^t \\ &= \sum_{k=0}^{K} A_k^{n(t)} \cos(kw_0t) + B_k^{n(t)} \sin(kw_0t) + W^t \\ A_k^{n(t)} &= A_k^{n(t)-1} + \epsilon_{A,k} \\ B_k^{n(t)} &= B_k^{n(t)-1} + \epsilon_{B,k}. \end{aligned} \tag{24}$$

Once the nonlinear parameter T_n is determined, this model can be written into the standard linear Gaussian state-space model as follows,

$$\begin{aligned} \mathbf{x}_t^L &= \mathbf{A}_t \mathbf{x}_{(t-1)}^L + \mathbf{u}_t^L, \\ s_t &= \mathbf{B}_t \mathbf{x}_{(t)}^L + v_t^L, \\ y_t &= s_t + G_t, \end{aligned} \tag{25}$$

where $\mathbf{x}_L^t$ is defined as

$$\mathbf{x}_t^L = \begin{pmatrix} a_1^{n(t)} \\ \vdots \\ a_M^{n(t)} \\ A_0^{n(t)} \\ \vdots \\ A_K^{n(t)} \\ B_0^{n(t)} \\ \vdots \\ B_K^{n(t)} \end{pmatrix}. \tag{26}$$

while $\mathbf{A}_t$ and $\mathbf{B}_t$ are functions of nonlinear states $x_{1:t}^N$ and are defined later in (30) and (31).

Meanwhile, the nonlinear part of state, $x_N^{0:t}$ can be expressed as

$$\mathbf{x}_{1:t}^N = [T_{1:n_t}, s_{1:t}]. \tag{27}$$

The other terms in the state space model are as follows,

$$\mathbf{u}_t \sim \begin{cases} \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),\ t > P_i \geq t-1, \\ \mathbf{0}, t \leq P_i \end{cases} \tag{28}$$

which accounts for the fact that x^L is fixed during each pitch period.

$\mathbf{\Sigma}$ is a diagonal matrix with the leading diagonal as:

$$\begin{pmatrix} \sigma_{a,1}^2 \\ \vdots \\ \sigma_{a,M}^2 \\ \sigma_{A,0}^2 \\ \vdots\ , \\ \sigma_{A,K}^2 \\ \sigma_{B,0}^2 \\ \vdots \\ \sigma_{B,K}^2 \end{pmatrix}. \tag{29}$$

And $\mathbf{B}_t$ is defined as

$$\mathbf{B}_t = \begin{pmatrix} s^{t-1} \\ \vdots \\ s^{t-M} \\ \cos(0w_0\phi(t)) \\ \vdots \\ \cos(Kw_0\phi(t)) \\ \sin(0w_0\phi(t)) \\ \vdots \\ \sin(Kw_0\phi(t)) \end{pmatrix}^T \tag{30}$$

$\phi(t)$ refers to the time since the start of the current period, i.e. $\phi(t) = t - \sum_{i=1}^{n_t-1} T_i$.

And,

$$\mathbf{A}_t = \mathbf{I}; \tag{31}$$

Thus, it is possible to separate the linear/Gaussian and nonlinear/non-Gaussian parts of the model and apply the Rao-Blackwellisation to the variable rate particle filter described in the last section.

Here we give the details of the derivation to estimate $p(x^N_{1:t}|y_{1:t})$ recursively from $p(x^N_{1:t-1}|y_{1:t-1})$, using Bayes' Theorem and the prediction error decomposition (PED) as below, mainly either taken directly or modified from the section II.G of [9].

Generally, we want to choose two proposal functions, i.e. $q(s_t|x^N_{1:t-1})$ and $q(T_{n_t}|T_{n_t-1})$, for the following probability distribution:

$$\begin{aligned} p(\mathbf{x}^N_{1:t}|y_{1:t}) &= p(T_{1:n_t}, s_{1:t}|y_{1:t}) \\ &\propto p(y_t|s_t, T_{n_t})p(T_{n_t}, s_t|T_{1:n_{t-1}}, s_{1:t-1})p(x^N_{1:t-1}|y_{1:t-1}) \\ &= p(y_t|s_t)p(T_{n_t}, s_t|T_{1:n_{t-1}}, s_{1:t-1})p(x^N_{1:t-1}|y_{1:t-1}) \\ &= p(y_t|s_t)p(T_{n_t}|T_{1:n_{t-1}})p(s_t|s_{1:t-1}, T_{1:n_t})p(x^N_{1:t-1}|y_{1:t-1}), \end{aligned} \tag{32}$$

and the PED term can be obtained as follows,

$$p(s_t|s_{1:t-1}, T_{1:n_t}) = \mathcal{N}\left(s_t|\mu_{s_t}, C_{s_t}\right), \tag{33}$$

where

$$\begin{aligned} \mu_{s_t} &= \mathbf{B}_t\mu_{t|1:t-1}, \\ C_{s_t} &= \mathbf{B}_t C_{t|1:t-1}\mathbf{B}_t^T + C_v. \end{aligned} \tag{34}$$

.

Then the weight w_t can be obtained by:

$$w_t \propto \frac{p(y_t|s_t)p(T_{n_t}|T_{1:n_{t-1}})p(s_t|s_{1:t-1}, T_{1:n_t})p(x^N_{1:t-1}|y_{1:t-1})}{q(s_t|x^N_{1:t-1})q(T_{n_t}|T_{n_t-1})} \tag{35}$$

The second proposal function here is chosen as $q(T_{n_t}|T_{n_t-1}) = p(T_{n_t}|T_{1:n_{t-1}})$ so that these two terms cancel out. And the first proposal function can be chosen as $q(s_t|x^N_{1:t-1}) = p(s_t|s_{1:t-1}, T_{1:n_t}, y_t)$. This is calculated as follows,

$$\begin{aligned} q\left(s_t|x^N_{1:t-1}\right) &= p\left(s_t|s_{1:t-1}, y_t, T_{1:n_t}\right) \\ &= \mathcal{N}\left(\hat{s}_t, 1/\phi_t\right), \end{aligned} \tag{36}$$

where

$$\begin{aligned} \hat{s}_t &= \theta_t/\phi_t, \\ \theta_t &= y_t/\sigma_G^2 + \mu_{s_t}/C_{s_t}, \\ \phi_t &= 1/\sigma_G^2 + 1/C_{s_t}. \end{aligned} \tag{37}$$

Terms used here such as μ_{s_t} and C_{s_t} involves calculating $\mu_{t|1:t-1}$ and $C_{t|1:t-1}$, i.e. the first two moments of $p(x^L_t|y_{1:t-1}, x^N_{1:t-1})$ obtained through the standard Kalman filter described as below. As a starting point, suppose we know $p(\mathbf{x}^L_{t-1}|y_{1:t-1}, \mathbf{x}^N_{1:t-1})$, which a Gaussian, denoted by

$$p(\mathbf{x}^L_{t-1}|y_{1:t-1}, \mathbf{x}^N_{1:t-1}) = \mathcal{N}\left(\mathbf{x}^L_{t-1}|\mu_{t-1|1:t-1}, C_{t-1|1:t-1}\right). \tag{38}$$

Then, to predict the linear part of state at step t while new data point s_t is not yet available, use the following:

$$p(\mathbf{x}^L_t|y_{1:t-1}, \mathbf{x}^N_{1:t}) = \mathcal{N}\left(\mathbf{x}^L_t|\mu_{t|1:t-1}, C_{t|1:t-1}\right) \tag{39}$$

where

$$\begin{aligned} \mu_{t|1:t-1} &= \mathbf{A}_t\mu_{t-1|1:t-1}, \\ C_{t|1:t-1} &= \mathbf{A}_t C_{t-1|1:t-1}\mathbf{A}_t^T + C_u. \end{aligned} \tag{40}$$

When s_t is sampled, we have the measurement update step:

$$p(\mathbf{x}^L_t|s_{1:t}, \mathbf{x}^N_{1:t}) \propto \mathcal{N}\left(\mathbf{x}^L_t|\mu_{t|1:t}, C_{t|1:t}\right) \tag{41}$$

where

$$\begin{aligned} \mu_{t|1:t} &= \mu_{t|1:t-1} + K_t\left(s_t - \mathbf{B}_t\mu_{t|1:t-1}\right), \\ C_{t|1:t} &= \left(I - K_t\mathbf{B}_t\right)C_{t|1:t-1}, \\ K_t &= C_{t|1:t-1}\mathbf{B}_t^T\left(\mathbf{B}_t C_{t|1:t-1}\mathbf{B}_t^T + C_v\right)^{-1}. \end{aligned} \tag{42}$$

And the update of importance weights are thus modified accordingly:

$$w_t^i \propto \frac{(2\pi)^{1/2}}{(\phi_t)^{1/2}} N(y_t|\hat{s}_t, \sigma_G^2) N(\hat{s}_t|\mu_{s_t}, C_{s_t}) w_{t-1}^i. \tag{43}$$

Then the Rao-Blackwellised version of variable rate particle filter is very similar with the variable rate particle filter algorithm described in section 'Implementation of variable rate particle filter', except the propagation part and the weight calculation.

The propagation part of the Rao-Blackwellised variable rate particle filter pitch tracking algorithm (RBVRPF) can be summarized as follows. Express $x_{1:t} = [\mathbf{x}^N_{1:t}, \mathbf{x}^L_{1:t}]$.

Algorithm 2:

```
1) Initialize $\{x_1^{(i)}\}_{i=1}^N$. Set up all the fixed hyperparameters and $\{w_1^{(i)}\}_{i=1}^N = \frac{1}{N}$. To initialize $\{x_1^{(i)}\}_{i=1}^N$, use the joint estimation technique based on the first period of speech data $y_{1:T_1}$, following equation (14). Then sample $\{s_1^{(i)}\}_{i=1}^N$ based on $[\mathbf{a_1}, \mathbf{A_1}, \mathbf{B_1}, \mathbf{T_1}]$ according to equation (24). Set $P_i = T_1$.
2) for $t$ = 1:$t_{end}$ do
   a) for i = 1:N do
      i) Set $n_t^{(i)} = n_{t-1}^{(i)}$.
      ii) While $t > P_i$:
         A) Add a new pitch period, $n_t^{(i)} \leftarrow n_t^{(i)} + 1$.
         B) Sample a new pitch period:
            $T_{n_t}^{(i)} \sim U(\max[T_{n_t-1}^{(i)} - \sigma_T, T_{low}], \min[T_{n_t-1}^{(i)} + \sigma_T, T_{upp}]),$
         C) Update $P_i \leftarrow P_i + T_{n_t}^{(i)}$.
      iii) Update $p(\mathbf{x}_t^L|y_{1:t}, \mathbf{x}_{1:t}^N)$ from $p(\mathbf{x}_{t-1}^L|y_{1:t-1}, \mathbf{x}_{1:t-1}^N)$ using equations (38) - (42).
      iv) Sample new signal value $s_t^{(i)}$ according to (36) and (37) .
      v) Compute importance weight $w_t^{(i)}$ of each particle according to (43):
   b) end for
   c) If $t$ = $k * BlockSize$, where $k$ is a positive integra,
      i) Renormalize $\tilde{w}_t^{(i)} = \frac{w_t^{(i)}}{\sum_{i=1}^N w_t^{(i)}}$, $i$ = $1, 2, ..., N$.
      ii) Resample $\{x_{1:t}^{(i)}\}_{i=1}^N$ when $N_{eff} < N/2$. $N_{eff}$ denotes the effective sample size and is calculated as $N_{eff} = 1/\sum_{i=1}^N \tilde{w}_t^{(i)}$.
```

iii) $\hat{T}_{n_t} = \sum_{i=1}^{N} w_t^{(i)} T_{n_t}^{(i)}$.

3) **end for**

It is worth noting that using the estimation of mean value $\hat{T}_{n_t}$ during this algorithm is not necessarily a good point estimator in many situations, for example, due to possible pitch doubling, etc. Actually what we have obtained from particle filters is a distribution of $p(T_{n_t}) \approx \sum_{i=1}^{N} w_t^i \delta(T_{n_t} - T_{n_t}^{(i)})$, and thus it may be better to use some other estimate from distribution to obtain a better point estimate of the target variable. However here we only choose the naive estimator here and test its performance.

VI. Experimental results

Speech data in this experiment are taken from the PTDB-TUG database [21]. Speech signals in PTDB-TUG database are recorded with laryngograph signals, which can be used to extract true pitch period. During the performance test stage, we randomly chosen 10 sentences from the database. To diversify the speakers, 5 sentences were spoken by five different males and 5 sentences were spoken by five different females. These sentences were then added with white Gaussian noise with five different SNRs. Then we randomly cut one utterance out of each of the sentence. Each utterance has a length of 187.5 ms. These 10 utterances were then downsampled from 48kHz to 16kHz. The other details in the simulations are as follows:
(1) Block size for particle filter processing: 8ms. So we have approximately 235 measurements of all the testing speech signals in total.
(2) Signal-to-Noise Ratio (SNR) of the noisy speech signals tested in the experiment are 10dB, 5dB, 0dB, -5dB and -10dB.
(3) During the following experiments, other parameters in VRPF are set as $\sigma_{a,p}^2 = 0.001$, $\sigma_T = 10$, $\sigma_{A,k}^2 = 0.005$, $\sigma_{B,k}^2 = 0.005$, $\tau_w = 0$,$\sigma_G^2 = 0.1$, $T_{low} = 40(samples)$, $T_{upp} = 200(samples)$, number of particles $N_p = 1000$, $M = 11$, $K = 20$.
Parameters in RBVRPF are set as $N_p = 1000$, $M = 8$, $K = 10$, $\sigma_{A,k}^2 = \sigma_{B,k}^2 = 0.0004$, $\sigma_{a,p}^2 = 0.001$, $\sigma_T = 10$, $T_{low} = 40(samples)$, $T_{upp} = 200(samples)$, $\tau_w = 0.003$. σ_G^2 is set to be equal to the variance of the added noise level. If $\sigma_G^2 < 0.04$, then set $\sigma_G^2 = 0.04$ to increase the stability of the particle filters.
(4) The first pitch period T_0 is assumed to be known for VRPF and RBVRPF. However, to show RBVRPF's robustness of this prior knowledge, the first pitch periods of all particles in RBVRPF method are chosen to be uniformly distributed from $T_0 - 10(samples)$ to $T_0 + 10(samples)$ during its initialization process.

Regarding the choices of values for these hyperparameters, we set them according to their physical interpretations as well as taking prior knowledge into considerations. For example, T_{low} and T_{upp} are set as those values because it is generally known that the fundamental frequency of human voices are mostly between 80Hz and 400Hz, and with the sampling frequency to be 16kHz, we can obtain the lower and upper bound of the possible fundamental period T range from 40 to 200 samples. The order of AR model, M is chosen to be an integer between 8 to 16 as most AR speech models follow this rule, and the number of harmonic components, K is chosen to be between 10 to 20 so that the signal model can include most energy of the voiced speech. Practice shows that the proposed algorithms here are not sensitive to the values of M and K as long as they fall into the approximate ranges stated above. Other parameters, were informally tuned using a training set of another set of 10 utterances with the same length/format as the test data, spoken by another different 10 speakers chosen randomly and gender balanced. The tuning process was performed as finding out suitable values for those hyper-parameters such as τ_w, σ_T, $\sigma_{a,p}^2$, $\sigma_{A,k}^2$, $\sigma_{B,k}^2$ and σ_G^2 so that the algorithm can produce satisfactory results for most of the utterances. They are not necessarily optimal but the overall performance of the proposed algorithms are not very sensitive to these hyper-parameters as long as these values are set within a value range which is appropriate to capture the characteristic of the slow time-varying process by which the voice signal is produced. A more formal inference procedure for these parameters is left for further study.

A. Preliminary tests

We have tested the proposed VRPF and RBVRPF methods in noisy environments from 0 dB to -10 dB input SNR levels for a vowel sound. In all the figures afterwards, 'SD' denote for 'standard deviation' of a single run of particle filters. Fig.1 shows the waveform for this vowel sound. Fig.2 shows the tracking result of the clean speech example of the VRPF method, along with the results of YIN algorithm and RAPT algorithm. The true pitch period, T0 value is extracted by the RAPT algorithm using the corresponding laryngograph recording data. From this figure we can see the proposed particle filter approach can estimate the pitch period quite well from clean speech.

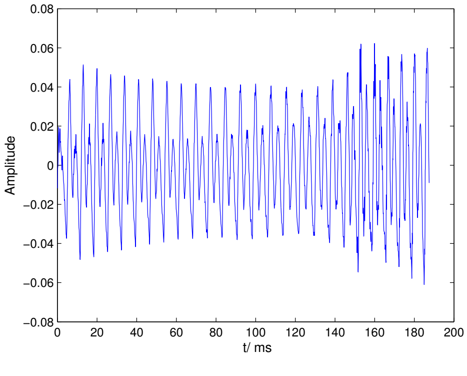

Fig. 1. Waveform of speech sample1.

Fig.3, Fig.4 and Fig.5 show the tracking results of these three methods in 0dB, -5dB and -10dB SNR input scenarios. The RAPT algorithm can not give meaningful result in 0dB and -5dB SNR input speech, under which conditions results are not displayed. The YIN algorithm is still robust in -5 and -10 dB SNR condition, but the performance is not that good compared with that in 0dB SNR condition. In contrast, the proposed method tracks the pitch period relatively accurate under both strong background noise conditions.

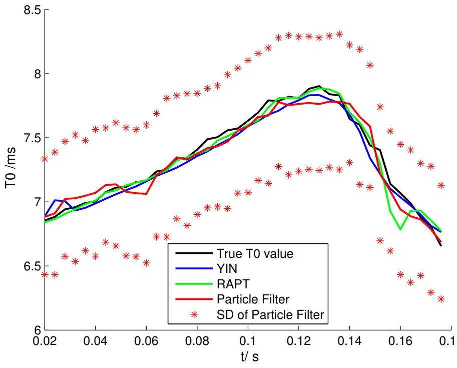

Fig. 2. Comparison of T0 estimated from three methods with the true T0 value, clean speech.

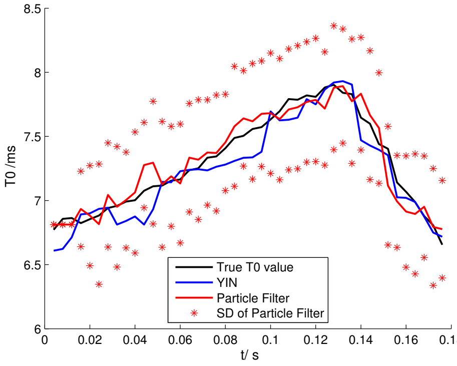

Fig. 3. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = 0dB.

Fig.6∼9 show the results of pitch tracker obtained from the RBVRPF method and the YIN algorithm in clean speech, 0dB, -5dB and -10dB SNR input scenarios, respectively. We can notice that the RBVRPF method is capable of extracting the pitch period in all these different noisy conditions, especially for the -10 dB SNR input scenario where the YIN algorithm provides a poor estimate of the correct pitch period.

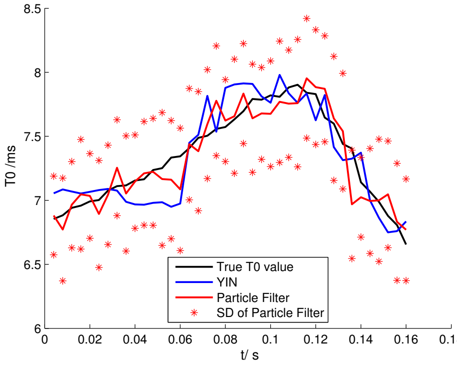

Fig. 4. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = -5dB.

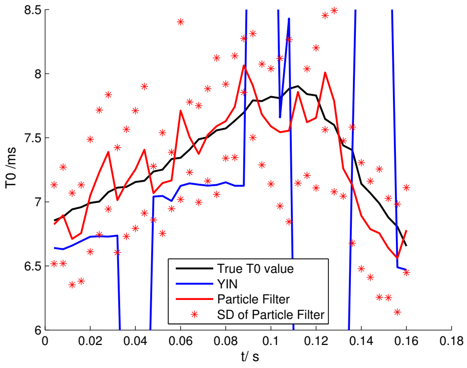

Fig. 5. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = -10dB.

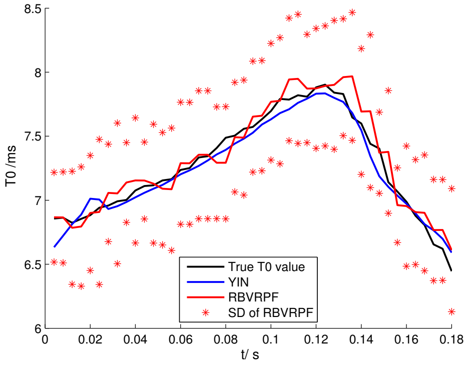

Fig. 6. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = 30dB.

B. Performance Comparison

To investigate the performance of the proposed particle filter approaches, the VRPF and RBVRPF using time-varying AR model, we compared them with one established state-of-the-art algorithm, the YIN algorithm [3]. The criterion used here include gross pitch error (GPE) rate and the mean and standard deviation of fine pitch error (FPE), which were defined in [22]. According to [22], GPE counts for an estimation error larger than 1 ms in fundamental period (T0). Gross pitch error (GPE) rate used here is calculated through gross pitch

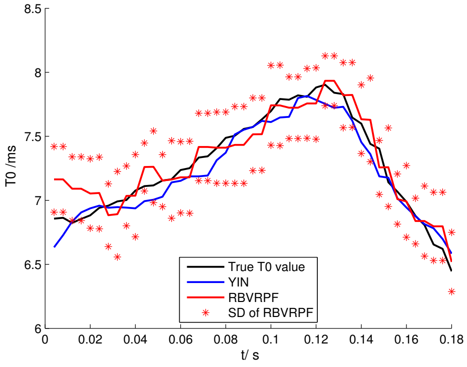

Fig. 7. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = 0dB.

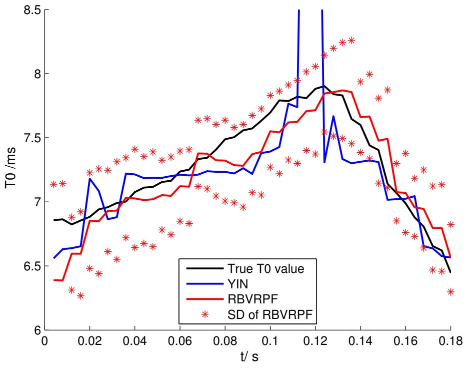

Fig. 8. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = -5dB.

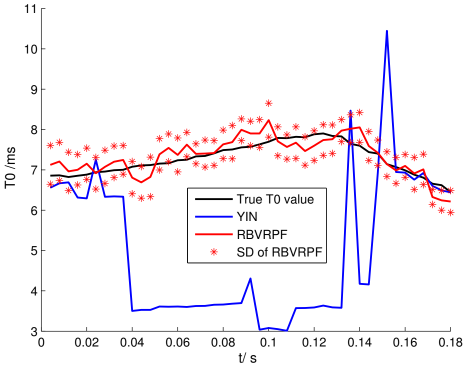

Fig. 9. Comparison of T0 estimated from three methods with the true T0 value. Input SNR = -10dB.

error count divided by the total number of pitch estimation values. The calculation of FPE excludes those errors which are included within the GPE rate. To estimate the GPE rate with consistency for these three algorithms, for each SNR input we simulate them with Gaussian noise for 10 times. For each simulation, we use the same 10 utterances with different randomly generated Gaussian noise. Thus for the 10 times of simulations, the noises added to the speech signals are different from time to time while all the noise levels are adjusted in accordance with the input SNR.

Table I shows that the proposed VRPF and RBVRPF based on time-varying AR model driven by almost periodic signals have less gross pitch error rate than the YIN algorithm when SNR = -5 and -10 dB. RBVRPF approach shows a significant better tracking result than the VRPF approach when SNR is -10 dB, indicating that Rao-Blackwellisation improves the performance of particle filter on that SNR scenario. In other higher SNR input conditions, GPE rate of all methods are quite low, which indicate that all these algorithms can estimate the pitch period with a proper accuracy. We also notice that the proposed RBVRPF is not better than the VRPF when the input SNR is higher than 0 dB. Possible reasons are that when the system noise is small, the particle filter suffers from the degenerate problem [10]. The Rao-Blackwellisation marginalises the parameter space and the calculation of linear part is also vulnerable to the degeneracy problem when the system noise is small [23]. Table II shows the standard deviations of mean GPE rate for these three algorithms, calculated from 10 times of simulations, which again suggests that the VRPF and RBVRPF produce at least as stable average GPE rates as the YIN algorithm when SNR = -5 and -10 dB.

Table III and Table IV shows the mean and standard deviation of the FPE of the proposed algorithms and the YIN algorithm. We can find out that the mean fine pitch error of the three algorithms are of no significant difference. When comparing the standard deviation of the fine pitch error, the VRPF algorithm is slightly larger than RBVRPF, and the RBVRPF is slightly larger than the YIN algorithm.

As a discussion of the performances of the three algorithms, we may conclude that the RBVRPF give the best GPE result when SNR is -10 dB. In SNR = -5 dB, the performances of the VRPF and the RBVRPF are of no significant difference, while both are better than the YIN algorithm. In other scenarios, the YIN algorithm is preferred because it is accurate and fast. The FPE results of the three algorithms are similar and can be ignored in most applications if there is no particular interest in it.

TABLE I
MEAN OF GROSS PITCH ERROR (GPE) RATE OF THE YIN ALGORITHM AND THE PROPOSED VRPF METHOD

Input SNR /dB	10	5	0	-5	-10
YIN	5.06%	4.97%	4.24%	11.72%	42.90%
VRPF	4.24%	4.45%	4.61%	6.76%	19.36%
RBVRPF	5.61%	6.33%	5.70%	6.27%	14.48%

TABLE II
STANDARD DEVIATION OF GROSS PITCH ERROR (GPE) RATE OF THE YIN ALGORITHM AND THE PROPOSED VRPF METHOD FROM TEN RUNS

Input SNR /dB	10	5	0	-5	-10
YIN	0.24%	0.33%	0.62%	4.13%	7.02%
VRPF	0.25%	0.29%	0.28%	2.04%	7.11%
RBVRPF	1.56%	1.73%	1.90%	2.03%	7.11%

TABLE III
MEAN FINE PITCH ERROR (FPE) RATE OF THE YIN ALGORITHM AND THE PROPOSED VRPF METHOD

Input SNR /dB	10	5	0	-5	-10
YIN	-0.067	-0.066	-0.064	-0.070	-0.024
VRPF	-0.047	-0.036	-0.012	-0.008	0.030
RBVRPF	-0.014	-0.036	-0.039	-0.045	-0.039

TABLE IV
STANDARD DEVIATION OF FINE PITCH ERROR (FPE) RATE OF THE YIN ALGORITHM AND THE PROPOSED VRPF METHOD

Input SNR /dB	10	5	0	-5	-10
YIN	0.047	0.047	0.046	0.039	0.043
VRPF	0.087	0.097	0.110	0.153	0.182
RBVRPF	0.055	0.058	0.075	0.094	0.092

VII. Conclusion

In this paper, we proposed a speech model based on time-varying AR model driven by almost periodic signals. Using this model, we used variable rate particle filters (VRPF) to track pitch period of voiced speech signals. The detailed implementation of Rao-Blackwellised variable rate particle filter (RBVRPF) is also shown in this paper. To test the proposed methods, we compare them with the well-established pitch estimation algorithm, the YIN algorithm. Experiments results show that the RBVRPF approach tracks pitch period with least Gross Pitch Error (GPE) among the three methods when SNR is less than or equal to -5 dB. In practice, it might be better to use the YIN algorithm when SNR is higher than -5 dB.

To the best of the authors' knowledge, very few pitch tracking methods today can give good pitch tracking results under strong noise condition. Here the proposed RBVRPF approach can give satisfactory pitch tracking results even if input SNR < -5 dB, thus it may be used on other speech signal processing techniques to improve their performance when strong noise presents. It is worth noting that the proposed methods work on voiced speech. Thus in practical applications, it will be helpful to combine them with voice activity detection (VAD), a topic left for future investigation.

Also the proposed speech model, i.e. the time-varying AR model driven by almost periodic signals, can serve as a basis to build other speech processing techniques, such as speech denoising or source estimation. For multi-speaker scenarios, it may even serve as a speech model for source separation methods.

Appendix A
Derivation of joint source-filter optimization for almost-periodic source

As a reminder that the observed samples s_t with a length of N produced by the time-varying AR model are given by,

$$s_t = \sum_{p=1}^{M} a_p^{n_t} s_{(t-p)} + V_t, \tag{44}$$

Where V_t would be given by

$$\begin{aligned} V_t =& g_t + W_t \\ =& \sum_{k=0}^{K} A_{n_t}^k \cos(kw_0 t) + B_{n_t}^k \sin(kw_0 t) + W_t \end{aligned} \tag{45}$$

$$\text{where} \quad g_t = \sum_{k=0}^{K} [A_{n_t}^k \cos(kw_0 t) + B_{n_t}^k \sin(kw_0 t)]$$

A_t^k and B_t^k are amplitude of the sine and cosine harmonic waveforms in the input sources in the t-th time sample. $W_t \sim \mathcal{N}(0, \tau_w^2)$ is assumed in [8]. Then given the observed speech signal samples, the goal of the joint source-filter optimization is to find a set of parameters $A_{n_t}^k$, $B_{n_t}^k$, $a_{n_t}^p$ (k=1 K, p=1 M) such that the cost function J defined as follows,

$$\begin{aligned} J =& E[(W_t)^2] \\ &\approx \frac{1}{N-M+1} \sum_{t=M}^{N-1} (w_t)^2 \\ &\propto \sum_{t=M}^{N-1} (s_t + \sum_{p=1}^{M} a_{n_t}^p s_{(t-p)} - v_t)^2 \qquad \text{[using Eq.45]} \\ &= \sum_{t=M}^{N-1} (s_t + \sum_{p=1}^{M} a_{n_t}^p s_{(t-p)} \\ &- \sum_{k=0}^{K} [A_{n_t}^k \cos(kw_0 t) + B_{n_t}^k \sin(kw_0 t)])^2 \end{aligned} \tag{46}$$

Since it is only expected the parameters of the first period of the speech are needed to be initialized, the cost function J in Eq.(46) can be calculated from time sample 1 to N_T, where N_T denotes the first pitch period time of input speech data. Besides, w_0 is a constant within each period, and therefore its upper index n_t is omitted during the whole paper.

$$\begin{aligned} J =& \sum_{t=1}^{N_T} (s_t + \sum_{p=1}^{M} a_{n_t}^p s_{(t-p)} \\ &- \sum_{k=0}^{K} [A_{n_t}^k \cos(kw_0 t) + B_{n_t}^k \sin(kw_0 t)])^2 \end{aligned} \tag{47}$$

It is worth noting that given a fixed input speech s_t, prior estimation of fundamental frequency w_0, and fixed hyper-parameters K, M, the cost function J is a quadratic function of the parameter set $a_{n_t}^p$,$A_{n_t}^k$, $B_{n_t}^k$. And we need only to estimate these parameters for the first period of voiced speech, then this function J depends on a^p,A^k, B^k (the upper time index n_t are omitted since $n_t = 1$). Thus, in order to minimize J, it can be obtained by setting $\frac{\partial J}{\partial a^p} = 0$, $p = 1, 2, ..., M$, and $\frac{\partial J}{\partial A^k} = 0$, $\frac{\partial J}{\partial B^k} = 0$, $k = 0, 1, 2, ..., K$ and solving this set of $2 * (K + 1) + M$ linear equations.

Expanding $\frac{\partial J}{\partial a_p} = 0$, $p = 1, 2, ..., M$ will lead to the following equations:

for r = 1,2,...,M :

$$\begin{aligned} &\frac{\partial J}{\partial a^r} = \\ &\sum_{t=1}^{N_T} x_{t-r}[x_t + \sum_{p=1}^{M} a^p x_{t-p} - \sum_{k=0}^{K} (A^k \cos(kw_0 t) + B^k \sin(kw_0 t))] \\ &= 0 \end{aligned} \tag{48}$$

And this equation can be written as:

for r = 1,2,...,M :

$$\begin{aligned} &\sum_{t=1}^{N_T} x_{t-r} \sum_{p=1}^{M} a^p x_{t-p} - \sum_{t=1}^{N_T} x_{t-r} \sum_{k=0}^{K} (A^k \cos(kw_0 t) + B^k \sin(kw_0 t)) \\ &= - \sum_{t=1}^{N_T} x_{t-r} x_t \end{aligned} \tag{49}$$

Expanding $\frac{\partial J}{\partial A^k} = 0$, $k = 1, 2, ..., K$ will lead to a equation similar with Eq. (49):

$$\begin{aligned} &\text{for k = 0,1,2,...,}K: \\ &-\sum_{t=1}^{N_T} \cos(kw_0t) \sum_{k=0}^{K} (A^k cos(kw_0t) + B^k sin(kw_0t)) \\ &-\sum_{t=1}^{N_T} \sum_{p=1}^{M} a^p x_{t-p} \cos(kw_0t) = -\sum_{t=1}^{N_T} cos(kw_0t) x^t \end{aligned} \tag{50}$$

And similarly, from $\frac{\partial J}{\partial B^k} = 0$, $k = 1, 2, ..., K$ we can get:

$$\begin{aligned} &\text{for k = 0,1,2,...,}K: \\ &-\sum_{t=1}^{N_T} \sin(kw_0t) \sum_{k=0}^{K} (A^k \cos(kw_0t) + B_k \sin(kw_0t)) \\ &-\sum_{t=1}^{N_T} \sum_{p=1}^{M} a^p x_{t-p} \sin(kw_0t) = -\sum_{t=1}^{N_T} \sin(kw_0t) x_t \end{aligned} \tag{51}$$

If we combine the set of equations from Eq.(49) to Eq.(51), we can write them in matrix form as follow following the same pattern used in [8]:

$$\mathbf{Ra} = \mathbf{p} \tag{52}$$

where

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 & -\mathbf{R}_2 \\ -\mathbf{R}_2^T & \mathbf{R}_3 \end{pmatrix} \tag{53}$$

where

$$\mathbf{R}_1 = \begin{pmatrix} C_{xx}(1,1) & \ldots & C_{xx}(M,1) \\ \vdots & \ldots & \vdots \\ C_{xx}(1,M) & \ldots & C_{xx}(M,M) \end{pmatrix} \tag{54}$$

$$\mathbf{R}_2 = \begin{pmatrix} \mathbf{R}_{2A} & \mathbf{R}_{2B} \end{pmatrix} \tag{55}$$

where

$$\mathbf{R}_{2A} = \begin{pmatrix} C_{Ax}^0(0,1) & \ldots & C_{Ax}^K(0,1) \\ \vdots & \ldots & \vdots \\ C_{Ax}^0(0,M) & \ldots & C_{Ax}^K(0,M) \end{pmatrix} \tag{56}$$

and

$$\mathbf{R}_{2B} = \begin{pmatrix} C_{Bx}^0(0,1) & \ldots & C_{Bx}^K(0,1) \\ \vdots & \ldots & \vdots \\ C_{Bx}^0(0,M) & \ldots & C_{Bx}^K(0,M) \end{pmatrix} \tag{57}$$

$$\mathbf{R}_3 = \begin{pmatrix} C_{CC}^{0,0}(0,0) & \ldots & C_{CC}^{0,2K+2}(0,0) & & \\ C_{CC}^{1,0}(0,0) & \ldots & C_{CC}^{1,2K+2}(0,0) & & \\ \vdots & \vdots & \ldots & \vdots & \vdots \\ C_{CC}^{2K,0}(0,0) & \ldots & C_{CC}^{2K,2K+1}(0,0) & & \\ C_{CC}^{2K+1,0}(0,0) & \ldots & C_{CC}^{2K+1,2K+1}(0,0) & & \end{pmatrix} \tag{58}$$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_M \\ A_0 \\ \vdots \\ A_K \\ B_0 \\ \vdots \\ B_K \end{pmatrix} \tag{59}$$

and

$$\mathbf{p} = \begin{pmatrix} -C_{xx}(0,1) \\ \vdots \\ -C_{xx}(0,M) \\ C_{Ax}^0(0,0) \\ \vdots \\ C_{Ax}^K(0,0) \\ C_{Bx}^0(0,0) \\ \vdots \\ C_{Bx}^K(0,0) \end{pmatrix} \tag{60}$$

And those parameters used in the paper have been defined as,

$$\begin{cases} C_{xx}(i,j) & = \sum_{t=1}^{N_T} x_{t-i} x_{t-j} \\ C_{Ax}^k(i,j) & = \sum_{t=1}^{N_T} x_{t-i} A^k \cos(kw_0(t-j)) \\ C_{Bx}^k(i,j) & = \sum_{t=1}^{N_T} x_{t-i} B^k \sin(kw_0(t-j)) \\ C_{CC}^{m,n} & = \sum_{t=1}^{N_T} \mathbf{c}^m(t) \mathbf{c}^{n_t} \end{cases} \tag{61}$$

And $\mathbf{c}^k(t)$ denotes the $k-th$ element of the vector $\mathbf{c}(t)$ which is defined as follows:

$$\mathbf{c}(t) = \begin{pmatrix} A^0 \cos(0w_0t) \\ A^1 \cos(1w_0t) \\ \vdots \\ A^K \cos(Kw_0t) \\ B^0 \sin(0w_0t) \\ B^1 \sin(1w_0t) \\ \vdots \\ B^K \sin(Kw_0t) \end{pmatrix} \tag{62}$$

Acknowledgment

The authors would like to thank CSC Cambridge International Scholarship and Natural Science Foundation of China(No.61463035) for providing financial support.

References

[1] D. Talkin, *A Robust Algorithm for Pitch Tracking (RAPT)*. Elsevier, 1995.

[2] A. De Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *Journal of the Acoustical Society of America*, vol. 111, no. 4, p. 1917, 2002.

[3] D. Sharma and P. A. Naylor, *Evaluation of pitch estimation in noisy speech for application in non-intrusive speech quality assessment*. 17th European Signal Processing Conference (EUSIPCO 2009), 2009.

[4] R. Chakraborty, D. Sengupta, and S. Sinha, "Pitch tracking of acoustic signals based on average squared mean difference function," *Signal, image and video processing*, vol. 3, no. 4, pp. 319–327, 2009.

[5] S. Gonzalez and M. Brookes, “Pefac-a pitch estimation algorithm robust to high levels of noise,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 2, pp. 518–530, 2014.

[6] W. Chu and A. Alwan, “Safe: a statistical approach to f0 estimation under clean and noisy conditions,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 933–944, 2012.

[7] P. A. Naylor, A. Kounoudes, J. Gudnason, and M. Brookes, “Estimation of glottal closure instants in voiced speech using the DYPSA algorithm,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 34–43, 2007.

[8] P. K. Ghosh and S. S. Narayanan, “Joint source-filter optimization for robust glottal source estimation in the presence of shimmer and jitter,” *Speech Communication*, vol. 53, no. 1, pp. 98–109, 2011.

[9] O. Cappé, S. J. Godsill, and E. Moulines, “An overview of existing methods and recent advances in sequential monte carlo,” *Proceedings of the IEEE*, vol. 95, no. 5, pp. 899–924, 2007.

[10] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, “A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking,” *Signal Processing, IEEE Transactions on*, vol. 50, no. 2, pp. 174–188, 2002.

[11] B. Ristic, S. Arulampalm, and N. Gordon, *Beyond the Kalman filter: Particle filters for tracking applications*. Artech House Publishers, 2004.

[12] G. Zhang and S. Godsill, “Tracking pitch period using particle filters,” in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2013, pp. 1–4.

[13] I. Trajkovic, C. Reller, M. Wolf, and H.-A. Loeliger, “Modelling and filtering almost periodic signals by time-varying fourier series with application to near-infrared spectroscopy,” *Proc. 17th European Signal Proc. Conf. (EUSIPCO)*, pp. 632–636, 2009.

[14] M. Davy, S. Godsill, and J. Idier, “Bayesian analysis of polyphonic western tonal music,” *The Journal of the Acoustical Society of America*, vol. 119, p. 2498, 2006.

[15] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and computing*, vol. 10, no. 3, pp. 197–208, 2000.

[16] S. Godsill, J. Vermaak, W. Ng, and J. F. Li, “Models and algorithms for tracking of maneuvering objects using variable rate particle filters,” *Proceedings of the IEEE*, vol. 95, no. 5, pp. 925–952, 2007.

[17] C. Snyder, T. Bengtsson, P. Bickel, and J. Anderson, “Obstacles to high-dimensional particle filtering,” *Monthly Weather Review*, vol. 136, no. 12, pp. 4629–4640, 2008.

[18] R. E. Kalman, “A new approach to linear filtering and prediction problems,” *Journal of Basic Engineering*, vol. 82, pp. 35–45, 1960.

[19] S. Godsill, “Particle filters for continuous-time jump models in tracking applications,” vol. 19, pp. 39–52, 2007.

[20] H. L. Christensen, J. Murphy, and S. J. Godsill, “Forecasting high-frequency futures returns using online langevin dynamics,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 4, pp. 366–380, 2012.

[21] G. Pirker, M. Wohlmayr, S. Petrik, and F. Pernkopf, *A Pitch Tracking Corpus with Evaluation on Multipitch Tracking Scenario*, 2011.

[22] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, “A comparative performance study of several pitch detection algorithms,” *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[23] F. Lindsten, T. B. Schön, and L. Svensson, “A non-degenerate rao-blackwellised particle filter for estimating static parameters in dynamical models,” vol. 16, no. 1, pp. 1149–1154, 2012.