COMPUTATIONALLY EFFICIENT METHOD OF SPEECH ENHANCEMENT BASED ON BLOCK REPRESENTATION OF SIGNAL IN STATE SPACE AND VECTOR QUANTIZATION

V. Semenov, A. Kovtonyuk, A. Kalyuzhny

"DELTA" S.P.E.
27 Academian Krymsky St., Kiev, 03142 Ukraine
e-mail: aleko@deltacorp.net

ABSTRACT

A computationally efficient speech enhancement method is proposed. Reduction of computations is achieved due to derived properties of block model of autoregressive (AR) signal. Decreasing of filtering error in comparison with traditional Kalman filter is shown. The problem of estimation of speech AR parameters is also considered. A two-phase computationally efficient estimation procedure, based on vector quantization of AR parameters, is proposed. On the first stage the initial approximation for optimal AR parameters is determined on a small number of AR quantums. Then the value of estimate is improved by efficient iterative procedure. The performance of resulting method is evaluated on real speech signals.

1. INTRODUCTION

Speech enhancement is a field of great practical importance. Speech quality and intelligibility might essentially deteriorate in the presence of background noise, essentially when the signal is subjected to subsequent processing. Speech coders, speech recognition and speaker identification systems that were designed to work on clean speech signals might be rendered useless in the presence of background noise. Majority of modern speech processing methods are based on AR model of speech:

\[ s(n) = - \sum_{k=1}^{p} a_k s(n-k) + g w(n), \]

where \( s(n) \) is a speech signal; \( w(n) \) is an excitation signal (white noise in the case of unvoiced speech and pulse train in the case of voiced speech); \( g \) is a gain factor; \( p \) is an order of AR filter; \( a_k, k = 1, \ldots, p \) - AR coefficients.

Suppose, that only observations:

\[ z(n) = s(n) + v(n), \]

are available, where \( v(n) \) is a white noise with variance \( \sigma_v^2 \).

The most popular technique for filtering of AR sequences is Kalman filter (KF). It assumes representation of models (1)-(2) in the state space (SS), which traditionally has the following form:

\[
\begin{align*}
\mathbf{x}(n) &= \mathbf{F}(n) \mathbf{x}(n-1) + \mathbf{G}(n) w(n), \\
\mathbf{z}(n) &= \mathbf{C} \mathbf{x}(n) + v(n),
\end{align*}
\]

Here matrices \( \mathbf{F} \) and \( \mathbf{G} \) can be expressed via the parameters of model (1), and the state vector is \( \mathbf{x}(n) = \begin{pmatrix} s(n-p+1) & s(n-p+2) & \ldots & s(n-1) \end{pmatrix}^T \). Note that two consequent state vectors of model (3) \( \{ \mathbf{x}(n) \} \) and \( \{ \mathbf{x}(n+1) \} \) differ only by 2 samples.

Application of KF to system (3) for the calculation of estimate \( \hat{s}(n) = E\{\mathbf{x}(n) / z(n), \ldots, z(1)\} \) is connected with high computational expenses, particularly because of renovation of all internal variables at each time step, which coincides with discretization unit. In spite of sparse structure of matrices in (3), this leads to high computational burden (expressions for the number of operations are given, for example, in [2]).

It is also necessary to note that renovation of internal variables for each observation sample has some contradiction with the procedure of parameters estimation, which is performed on the blocks of measurements.

Taking it into account, in this paper we propose approach to filtering of noisy AR signals, based on their block model. This model means the processing of signal by non-overlapping blocks of arbitrary length. It is shown that KF, based on such representation, provides significant reduction of computational expenses in comparison with traditional KF, grounded on (3). The reduction of computations is provided due to derived properties of block representation matrices.

Another problem is connected with the absence of effective methods of AR parameters estimation at noise background. The most popular approach to the solution of this problem lies in maximization of likelihood or a posteriori probability functions with the help of estimate-maximize (EM) algorithm [3, 4]. However, known methods do not provide global maximization and strongly depend on the choice of initial approximation. Taking it into account, in work [5] authors proposed noise-robust method of AR parameters estimation, based on vector quantization. The value of estimate was obtained by maximization of likelihood function (depending on noise characteristics) on a limited number of AR quantums. The main obstacle for the implementation of this methodology lies in its high computational expenses, because the necessary size of codebook may be very high. In this paper it is shown, how to decrease computational burden of this approach with the help of derived properties of block model.

Another way of decreasing of computational expenses - application of combined two-phase algorithm. On the first stage direct likelihood function (LF) maximization on a small number...
of quantums is used just as initial approximation. Then the exact value of estimate is determined by well-known iterative procedure [3], which provides global maximization under the assumption of qualitative initialization. Computational optimization of this procedure is considered. Such combined approach effectively solves important problem of iterative methods, connected with the choice of initial approximation.

2. BLOCK REPRESENTATION OF AR SIGNAL

In this section block representation of models (1)-(2) is described and properties, providing computational savings, are derived.

Let’s introduce following signal vector:

\[
S(k) = [s_t(k-l)/1, j = 1, ..., l + p + 1]T.
\]

In oppose to standard representation in SS (3), vector (4) includes array of \(l\) samples (\(l \geq p\)) and, besides, adjacent vectors do not overlap. Similarly to (4), let’s introduce parameters do not change in the limits of block (i.e., \(u_{ij} \leq i \leq j \leq l\), \(i < j \leq l + p\)). Formulas (1) and (2) can be represented in following block form (it is assumed that signal parameters do not change in the limits of block):

\[
\begin{align*}
\mathbf{A} \mathbf{S}(k) &= \mathbf{gW}(k), \\
\mathbf{Z}(k) &= \mathbf{S}(k) + \mathbf{V}(k),
\end{align*}
\]

where \(\mathbf{A}\) is a whitening matrix with elements

\[
\begin{align*}
A_{ij} &= 1, & i = j, \\
&= a_{ij}, & j < i \leq j + p, \\
&= 0, & i < j > j + p.
\end{align*}
\]

It can be seen from (5) that observation vector is equal to

\[
\mathbf{Z}(k) = \mathbf{gA}^{-1}\mathbf{W}(k) + \mathbf{V}(k).
\]

In the case when we are interested in the dependence on the previous time block, expression (7) must be replaced by system in the SS:

\[
\begin{align*}
\mathbf{S}(k) &= \mathbf{FS}(k-1) + \mathbf{gA}^{-1}\mathbf{W}(k), \\
\mathbf{Z}(k) &= \mathbf{S}(k) + \mathbf{V}(k).
\end{align*}
\]

Recursive formulae, connecting elements of matrix \(\mathbf{F}\) with parameters of AR model (1) have a form:

\[
\begin{align*}
F_{ij} &= 0, 1 \leq i \leq l, j < l - p, \\
F_{ij} &= - \sum_{k=1}^{\min(i-1, p)} a_k F_{i-k, j} + u_{ij}, 1 \leq i \leq l, l - p + 1 \leq j \leq l.
\end{align*}
\]

where

\[
\begin{align*}
u_{ij} &= 0, 1 \leq i \leq l, j < l - p, \\
u_{ij} &= a_{l-j+i}, 1 \leq i \leq p, l - p + 1 \leq j \leq l.
\end{align*}
\]

The basic property, connecting matrices \(\mathbf{A}\) and \(\mathbf{F}\), is a special structure of matrix \(\mathbf{T} = \mathbf{AF}\):

\[
\begin{align*}
\mathbf{T}_{ij} &= -a_{l-i+j}, i = 1, ..., p, j = l - p + i, ..., l, \\
\mathbf{T}_{ij} &= 0, \text{ in other cases}.
\end{align*}
\]

So, only \(1 + 2 + ... + p = 0.5p(p + 1)\) elements of matrix \(\mathbf{AF}\) are non-zero, while the size of \(\mathbf{AF}\) can be as high as 200 \(\times\) 200, corresponding to 25 ms (maximum period of speech quasistationarity) at sampling frequency of 8000 Hz. Moreover, this matrix does not need any calculations, because it consists of AR coefficients, taken with inverse sign. In future we will say that some matrix \(\mathbf{T}\) has AF-structure if \(\mathbf{T}_{ij} = 0, i > j - l + p\). It can be shown, that such matrices have following useful properties:

**Property 1.** Let \(\mathbf{C}\) be upper triangular matrix. Then matrices \(\mathbf{CT}\) and \(\mathbf{TC}\) have AF-structure.

**Property 2.** Let \(\mathbf{C}\) be upper triangular matrix. Then matrix \(\mathbf{D} = \mathbf{TC}^{-1}\) has AF-structure.

These facts play decisive role for the reduction of computations at filtering of AR processes.

3. PROCEDURE OF BLOCK FILTERING

Let’s apply KF to system (8):

\[
\begin{align*}
\mathbf{P}(k/k - 1) &= \mathbf{FP}(k - 1/k - 1)\mathbf{F}^T + \mathbf{g}^2\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{F}^T, \\
\mathbf{P}(k/k) &= (\mathbf{P}^{-1}(k/k - 1) + \mathbf{E}_{ij}^2)^{-1}, \\
\mathbf{S}(k) &= \mathbf{FS}(k - 1) + \mathbf{g}^2\mathbf{E}_{ij}(\mathbf{Z}(k) - \mathbf{FS}(k - 1)),
\end{align*}
\]

where \(\mathbf{S}(k)\) - vector of signal estimates on the \(k\)-th block, based on observation blocks \(\mathbf{Z}(1), ..., \mathbf{Z}(l)\); \(\mathbf{P}(k/k)\) and \(\mathbf{P}(k/k - 1)\) - filtering and prediction error matrices.

Computation of covariance matrices \(\mathbf{P}(k/k), \mathbf{P}(k/k - 1)\) and renovation of block estimates is now performed only when a new block of measurements is received.

Let’s derive key property of block KF: inverse error matrix \(\mathbf{P}^{-1}(k/k)\) has (\(2p + 1\))-diagonal structure, i.e. all its non-zero elements are situated on the central (\(2p + 1\)) diagonals.

From (11) one can show that:

\[
\begin{align*}
\mathbf{P}^{-1}(k/k) &= \mathbf{P}^{-1}(k/k - 1) + \mathbf{E}_{ij}^2, \\
&= (\mathbf{FP}(k - 1/k - 1)\mathbf{F}^T + \mathbf{g}^2\mathbf{A}^{-1}\mathbf{A}^{-1}\mathbf{F}^T)^{-1} + \mathbf{E}_{ij}^2.
\end{align*}
\]

Let’s rewrite (12) with the help of well-known matrix property:

\[
\begin{align*}
\mathbf{P}^{-1}(k/k) &= [\mathbf{FP}(k - 1/k - 1)\mathbf{F}^T + \mathbf{g}^2(\mathbf{A}^T\mathbf{A})^{-1}]^{-1} + \mathbf{E}_{ij}^2, \\
&= -\mathbf{g}^{-2}\mathbf{A}^T\mathbf{FP}^{-1}(k - 1/k - 1)\mathbf{F}^T \mathbf{A}^T\mathbf{E}_{ij}^2 + \mathbf{E}_{ij}^2, \\
&+ \mathbf{g}^{-2}\mathbf{A}^T\mathbf{FP}^{-1}(k - 1/k - 1) + \mathbf{g}^{-2}\mathbf{F}^T \mathbf{A}^T\mathbf{E}_{ij}^2 \mathbf{F}^T \mathbf{A}^T + \mathbf{E}_{ij}^2
\end{align*}
\]

Taking into account Cholesky decomposition

\[
\mathbf{P}^{-1}(k - 1/k - 1) + \mathbf{g}^{-2}\mathbf{F}^T \mathbf{A}^T\mathbf{E}_{ij}^2 \mathbf{F}^T \mathbf{A}^T = \mathbf{C}^T \mathbf{C},
\]

\((\mathbf{C}\) is upper triangular matrix), formula (13) will have a form:
P^{-1}(k/k) = g^{-2}A^T A - g^{-4}A^T A K (C^T C)^{-1}k g^{-2} A^T A + \sigma_n^{-2} E_i =
= g^{-2}A^T A - g^{-4}(A^T A F C^{-1}) (A^T A F C^{-1})^T + \sigma_n^{-2} E_i
\tag{15}

Consider component \( T = A^T A F C^{-1} \) in expression (15). Matrix \( D = A^T (AF) \) in accordance with lemma 1 has AF-structure, because \( A^T \) is upper triangular matrix. Then, due to lemma 2, \( T = DC^{-1} = A^T A F C^{-1} \) is also matrix with AF-structure (\( C^{-1} \) - upper triangular). At last, it is easy to see that matrix \( TT^T = (A^T A F C^{-1})(A^T A F C^{-1})^T \) is \((2p+1)\)-diagonal.

Matrix \( A \) has \( p+1 \) non-zero central diagonals; therefore matrix \( A^T A \) is \((2p+1)\)-diagonal. So, all components in right part of (15) are \((2p+1)\)-diagonal matrices. This completes the proof of the fact that \( P^{-1}(k/k) \) - \((2p+1)\)-diagonal matrix. Note, that covariance matrix of standard KF, based on (3), does not possess such property.

Further, third formula of algorithm (11) can be written as
\[
P^{-1}(k/k) [\sigma_n^2 (\hat{S}(k) - FS(k-1))] = Z(k) - FS(k-1).
\tag{16}
\]

Due to derived property of matrix \( P^{-1}(k/k) \), vector \( \hat{S}(k) \) can be efficiently determined from (16) via Cholesky decomposition of \( P^{-1}(k/k) \).

So, developed efficient method of block filtering consists of following stages (on \( k \)-th time block):

- renovation of \( P^{-1}(k/k) \) by formulae (14), (15);
- determination of desired estimate \( \hat{S}(k) \) by solution of (16).

Consider computational expenses, connected with application of proposed method for processing of \( N \) noisy observations. In particular case, when the size of block coincides with \( N \), computational expenses, required number of elementary operations is about \( p^2 N \). At the same time, traditional KF, based on (3), needs, according to [2], \( 4p^2 N \) operations. So, the computational expenses are reduced in 4 times.

Practical implementation of described procedure needs the values of speech AR parameters. Estimation of AR parameters at noise background is considered in the next section.

4. ESTIMATION OF AR PARAMETERS AT NOISE BACKGROUND

Among methods of AR parameters estimation in the presence of noise one can emphasize those based on the maximization of LF. Existing methods of solution of this problem do not provide global maximization. As a rule, they have iterative structure and strongly depend on the choice of initial approximation. That is why in work [5] authors proposed new noise-robust method of AR parameters estimation. The essence of this approach lies in maximization of LF (depending on noise characteristics) on a limited set of AR quantum \( \{a^{(m)}, g^{(m)}\} \), \( m = 1, \ldots, M \) instead of maximization on continuous space \( \{a, g\} \), \( a = [a_0, \ldots, a_p] \). The choice of quantums depends on the specific character of the problem and can be realized by \( K \)-means algorithm [6].

In oppose to traditional approaches, the proposed method has systematic error of estimation (bias). However, the value of bias is determined by such factors as the amount of quantums and can always be reduced to desired level. At the same time this bias is compensated by decreasing of statistical error of estimation (variance of estimate).

The main obstacle for the implementation of this methodology is that LF calculation for every quantum leads to significant computational expenses, when the number of quantums is relatively high. That is why a following method of economic LF computation, based on block model (7), is proposed.

Assume, that we have block \( Z \) of \( L \) values of noisy AR signal. From (7) one can obtain that the covariance matrix of vector \( Z \) is \((g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L \). So, under assumption that both the speech and noise sequences are Gaussian, LF can be written as
\[
p(Z|a, g) = (2\pi)^{-L/2} |det((g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L)|^{-1/2} \times
\exp[-0.5Z^T (g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L]^{-1/2} Z].
\tag{17}
\]

The expression in exponent function is equal to
\[
Z^T (g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L]^{-1/2} Z = Z^T A^T (g^2 E_L + \sigma_n^2 A A^T)^{-1} A Z.
\tag{18}
\]

Taking into account Cholesky decomposition
\[
g^2 E_L + \sigma_n^2 A A^T = C^T C,
\tag{19}
\]
formula (18) can be written as
\[
Z^T (g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L]^{-1} Z = Z^T A^T (C^T C)^{-1} A Z =
\tag{20}
= (C^{-1} A Z)^T C^{-1} A Z.
\]

Matrix \( g^2 E_L + \sigma_n^2 A A^T \) has \((2p+1)\)-diagonal structure (only \((2p+1)\) of its diagnals are non-zero). Then it can be shown that matrix \( C^{-1} A \) in decomposition (19) is lower \((p+1)\)-diagonal. This greatly simplifies the computation of vector \( X = C^{-1} A Z \), equivalent to the solution of equation \( C^T X = AZ \).

At last, from (19), one can obtain that
\[
det((g^2 A^{-1} A^T)^{-1} + \sigma_n^2 E_L] = det(C^T C) = \prod_{k=1}^{L} C_k^2.
\tag{21}
\]

This completes the calculation of function (17).

Another way of decreasing of computational expenses - application of combined two-phase algorithm. On the first stage direct LF maximization on a small number of quantums is used just as initial approximation. The exact value of estimate is determined by well-known iterative procedure [3], which provides convergence to global maximum of LF under the assumption of good initialization. Each iteration consists of solution of modified Yule-Walker equations with consequent optimal linear estimation based on the obtained parameters.
The main disadvantage of this iterative method, outlined in [3], lies in its high computational burden, connected with computation of optimal linear estimate at each iteration. This estimate is traditionally determined by double usage of KF (in forward and backward directions) [3] and is connected with high computational burden. However this difficulty can be completely avoided with the help of filtering method, described in section 3.

The described combined method allows to reduce estimation error in comparison with direct LF maximization due to exclusion of systematic component of error.

Proposed combined algorithm solves important problem of iterative methods, connected with the choice of initial approximation. In “Experiments” section is shown that the number of quantums, necessary for good initialization, can be relatively low.

5. EXPERIMENTS

As was pointed out in section 3, proposed procedure of block filtering provides reduction of computations in 4 times, as compared with traditional KF. It is also necessary to evaluate the performance of resulting enhancement method on real speech signals.

Russian sentences uttered by 4 male and 2 female speakers were used in experiments. They were mixed with white noise for different signal-to-noise ratios (SNR). The estimation of AR parameters was performed on non-overlapping frames of 20 ms (i.e. \( L = 160 \) samples at sampling frequency 8 kHz). The order of AR model was equal to 10.

At first, let’s consider the case, when the parameters of signal and noise are assumed to be known (AR parameters were determined on pure speech signal with the help of Levinson-Durbin algorithm). In table 1 one can see total output SNRs for the processing of noisy signal by following methods:

- traditional Kalman filter (TKF), based on (3);
- block KF (BKF) with block sizes \( l = 10(= p), \ l = 40, \ l = 160(= l) \).

Table 1: Output SNRs for evaluated filtering methods

<table>
<thead>
<tr>
<th>Input SNR</th>
<th>TKF</th>
<th>BKF (10)</th>
<th>BKF (40)</th>
<th>BKF (160)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 dB</td>
<td>7.78</td>
<td>9.46</td>
<td>9.97</td>
<td>10.12</td>
</tr>
<tr>
<td>5 dB</td>
<td>10.96</td>
<td>12.80</td>
<td>13.26</td>
<td>13.40</td>
</tr>
<tr>
<td>10 dB</td>
<td>14.42</td>
<td>16.06</td>
<td>16.47</td>
<td>16.63</td>
</tr>
<tr>
<td>15 dB</td>
<td>18.22</td>
<td>19.59</td>
<td>19.93</td>
<td>19.98</td>
</tr>
<tr>
<td>20 dB</td>
<td>22.23</td>
<td>23.27</td>
<td>23.44</td>
<td>23.52</td>
</tr>
</tbody>
</table>

From table 1 it is seen, that modifications of block KF have essential advantage over TKF (from 1.0 – 1.3 dB at SNR=20 dB to 1.7 – 2.3 dB at SNR=0 dB). Increasing of block length leads to decreasing of filtering error, because more observations are involved in filtering procedure.

Let’s consider the case, when parameters of signal and noise are unknown. For the estimation of speech AR parameters two-stage procedure, described in 4th section, was applied. On the first stage codebooks of different sizes (4, 16, 64, 128) were used. These codebooks were generated with the help of speakers, which did not take part in the creation of test signals. On the second stage for the iterative improvement of obtained estimates 5 iterations were used. Estimation of noise variance was performed on the beginning 100 msec, which did not contain speech signal.

Table 2 shows total output SNRs, obtained due to processing of noisy signal by block KF with block length \( l = L = 160 \) for different input SNRs. For every number of quantums SNRs corresponding to first and second (in brackets) stages of estimation are provided.

Table 2: Output SNRs for different modifications of BKF

<table>
<thead>
<tr>
<th>SNR</th>
<th>4 q.</th>
<th>16 q.</th>
<th>64 q.</th>
<th>128 q.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0 dB</td>
<td>8.52(9.22)</td>
<td>9.09(9.38)</td>
<td>9.17(9.31)</td>
<td>9.19(9.32)</td>
</tr>
<tr>
<td>5 dB</td>
<td>11.55(12.44)</td>
<td>12.23(12.62)</td>
<td>12.47(12.67)</td>
<td>12.49(12.64)</td>
</tr>
<tr>
<td>10 dB</td>
<td>14.86(15.77)</td>
<td>15.52(16.02)</td>
<td>15.73(16.09)</td>
<td>15.80(16.09)</td>
</tr>
<tr>
<td>15 dB</td>
<td>18.41(19.48)</td>
<td>18.95(19.64)</td>
<td>19.21(19.70)</td>
<td>19.27(19.70)</td>
</tr>
</tbody>
</table>

As can be seen from table 2, for the initialization of iterative algorithm it is enough to use 16 quantums. Consequently increasing of the number of quantums does not influence the value of filtering error. It also worth to note that all values of output SNR are within 1 dB lower than values obtained in “ideal” case, when parameters were known. This result illustrates the effectiveness of proposed estimation procedure.

6. CONCLUSION

In this paper a computationally efficient speech enhancement method was proposed. It is based on block model of AR signal in state space and vector quantization. Reduction of computations in 4 times in comparison with standard Kalman filter was obtained. Decreasing of filtering error in comparison with traditional Kalman filter was shown. A two-phase efficient procedure of speech AR parameters estimation was proposed. The effectiveness of resulting method was tested on real speech.

7. REFERENCES