ADAPTIVE KALMAN FILTERING OF SPEECH SIGNALS, BASED ON A BLOCK MODEL IN THE STATE SPACE AND VECTOR QUANTIZATION OF AUTOREGRESSIVE FEATURES

A.A. Kovtonyuk, A.Ya. Kalyuzhny, V.Yu. Semenov
Scientific-Production Enterprise "DELTA"
27 Acad. Krymsky St., Kiev, 03142 Ukraine
e-mail: aleko@delta-net.kiev.ua

ABSTRACT: A novel method of adaptive Kalman filtering (KF) of noisy speech is proposed. The method is based on block model of autoregressive (AR) signal in the state space (SS). It is shown that such representation allows to reduce computational expenses and to decrease filtering error as compared with known methods. Also the essentially new method of estimation of AR parameters in the presence of noise is developed. This method is based on the usage of optimal Bayesian estimation and vector quantization.

INTRODUCTION

The problem of speech enhancement now has great importance due to the development of automatic speech recognition systems, intended for usage in adverse noisy conditions. This problem is also urgent for systems of digital telephony, because their efficiency degrades quickly in the presence of background noise.

Earlier this problem was typically solved with the help of filtering methods, which did not use specific features of speech (Boll (1979), reviews of Lim & Oppenheim (1979), Kybic (1998)). At the present time a majority of speech enhancement methods are based on the AR model of speech generation:

$$s(n)=-\sum_{k=1}^{p} a_k s(n-k)+gw(n), \qquad (1)$$

where $s(n)$ and $w(n)$ are the discrete values of speech and excitation (non-correlated noise in the case of unvoiced speech or pulse train in the case of voiced speech) respectively; p is an order of AR filter; g and a_k $(k=1,...,p)$ are the model gain and filter coefficients respectively.

We assume that only observations $z(n)$, which are formed by the sum of pure speech $s(n)$ and white noise $v(n)$ with zero mean and known variance σ_v^2, are available:

$$z(n)=s(n)+v(n) \qquad (2)$$

Thus, speech enhancement procedure includes 2 stages: estimation of AR parameters of noisy speech, and filtering with the help of measured parameters.

Along with the Wiener filtering methods, which use the AR model of speech (Lim & Oppenheim (1979), review of Kybic (1998)), increasing interest is attracted to the usage of Kalman filter. The results of Gannot (1998), Goh (1999), and Kybic (1998) show the increasing of enhancement efficiency due to the adaptation by AR parameters as compared with the filtering approaches, which do not use specific features of speech.

However, there are several difficulties with the application of this methodology. The first is connected with the absence of convenient and at the same time effective methods of AR parameters estimation in the presence of noise. Whereas the methods of optimal filtering are relatively well studied, this part of the problem still remains open. The standard methods of AR parameters calculation, which are widely used in speech coding (for instance, autocorrelation method (ACM) in Markel & Gray (1976), Rabiner & Schafer (1978)), lose their efficiency in the presence of background noise. So, there is a need for noiseproof methods of estimation.

Among such methods we must emphasize those based on the principle of maximum likelihood. One of the first papers in this direction belongs to Lim & Oppenheim (1978). They proposed an iterative

method (a kind of EM algorithm), every iteration of which consisted of the application of ACM with consequent Wiener filtering based on the obtained parameters. This method, however, does not lead even to the locally optimal estimate, and this influences the quality of enhanced speech. Several similar estimation methods, based on the EM algorithm, were mentioned by Gannot (1998). All of them give, in the best case, convergence to the local maximum of likelihood function. But the problem of global maximization still remains open.

Another problem under discussion is connected with the high computational expenses of modern speech enhancement procedures. Even without adaptation, Kalman filtering needs essentially higher computational efforts as compared with Wiener filtering (see, for example, Goh (1999)).

Taking these difficulties into account, our goal is the development of: 1) more convenient and effective noiseproof methods of AR model estimation; 2) more efficient methods of Kalman filtering.

KALMAN FILTERING BASED ON THE BLOCK MODEL IN THE STATE SPACE

Kalman filter, as opposed to Wiener filter, is better accommodated to the processing of nonstationary signals and data of finite length. Its usage implies representation of AR model (1) in the SS. In papers by Gannot (1998), Goh (1999), Kybic (1998) the iterative kind of representation was used:

$$\begin{cases} \mathbf{x}(n) = \mathbf{F}(n)x(n-1) + \mathbf{G}(n)w(n) \\ z(n) = \mathbf{C}\mathbf{x}(n) + v(n) \end{cases} \tag{3}$$

where $\mathbf{x}(n) = (s(n-p+1)\ s(n-p+2)\ \ldots\ s(n-1)\ s(n))^T$ is the state vector, matrices $\mathbf{F}(n)$ and $\mathbf{G}(n)$ are defined by AR parameters.

KF based on (3) needs computation of all necessary matrices after every shift on discretization interval T, and thus leads to significant computational expenses. Also, such filtering has some contradiction with the estimation of parameters, performed on the blocks of data. That's why we propose a new block model of AR signal in the SS, which allows us to reduce computational expenses.

Let's introduce a new state vector:

$$\mathbf{s}^{(j)} = (s((j-1)l+1)\ s((j-1)l+2) \ldots s(jl))^T \tag{4}$$

As opposed to the previous case, vector (4) includes the array of l samples ($l \geq p$) and, also, adjacent vectors do not intersect and are separated by the time interval lT. Then, using AR model (1) and measurement equation (2), we obtain the following representation in the SS (it is assumed that AR parameters do not change in the limits of one time block):

$$\begin{cases} \mathbf{s}^{(j)} = \mathbf{F}^{(j)}\mathbf{s}^{(j-1)} + \mathbf{G}^{(j)}\mathbf{w}^{(j)} \\ \mathbf{z}^{(j)} = \mathbf{s}^{(j)} + \mathbf{v}^{(j)} \end{cases} \tag{5}$$

Here $\mathbf{w}^{(j)}$, $\mathbf{v}^{(j)}$, $\mathbf{z}^{(j)}$ are respectively excitation, noise and measurement vectors, which are formed similarly to (4).

We obtained a relationship between $\mathbf{F}$ and $\mathbf{G}$ matrices and the parameters of AR model (1) (we drop time superscripts for the simplicity). The elements of matrix $\mathbf{F}$ can be calculated in the recursive way:

$$\begin{cases} \mathbf{F}_{ij} = 0, & 1 \leq i \leq l,\ j \leq l-p \\ \mathbf{F}_{ij} = -\sum_{k=1}^{\min(i-1,p)} a_k \mathbf{F}_{i-k,j} + u_{ij}, & 1 \leq i \leq l,\ l-p+1 \leq j \leq l \end{cases} \tag{6}$$

where

$$\begin{cases} u_{ij} = 0, & 1 \le i \le l,\ j \le l - p \\ u_{ij} = a_{l-j+i}, & 1 \le i \le p,\ l - p + 1 \le j \le l \end{cases} \tag{7}$$

So, the first $l - p$ columns of matrix $\mathbf{F}$ are equal to zero. The remaining columns are formed by applying filter (1) to the sequence $\mathbf{u}_j = \{a_{l-j+1}\ a_{l-j+2} \ldots\ a_p\ 0\ 0 \ldots 0\}$ of length l (where j is a number of column).

Further, $\mathbf{G}$ is a lower triangular matrix, the elements of which can be obtained in a similar recursive way:

$$\begin{cases} \mathbf{G}_{ii} = g, & 1 \le i \le l \\ \mathbf{G}_{ij} = -\sum_{k=1}^{\min(i-1,p)} a_k \mathbf{G}_{i-k,j}, & 2 \le i \le l, 1 \le j \le i \end{cases} \tag{8}$$

We have to notice that $\mathbf{F}$ and $\mathbf{G}$ are sparse matrices and their structures are essential for the reduction of computational expenses of KF.

After the calculation of $\mathbf{F}$ and $\mathbf{G}$ matrices, we can apply KF to obtain an optimal linear mean-square estimate (see, for example, Sage & Melse (1972)):

$$\begin{cases} \mathbf{V}(k/k-1) = \mathbf{F}^{(k)}\mathbf{V}(k-1/k-1)\mathbf{F}^{(k),T} + \mathbf{G}^{(k)}\mathbf{G}^{(k),T} \\ \mathbf{V}(k/k) = \mathbf{V}(k/k-1) - \\ \quad - \mathbf{V}(k/k-1)(\mathbf{V}(k/k-1) + \sigma_v^2 \mathbf{E}_l)^{-1}\mathbf{V}(k/k-1) \\ \hat{\mathbf{s}}^{(k)} = \mathbf{F}^{(k)}\hat{\mathbf{s}}^{(k-1)} + \dfrac{1}{\sigma_v^2}\mathbf{V}(k/k)(\mathbf{z}^{(k)} - \mathbf{F}^{(k)}\hat{\mathbf{s}}^{(k-1)}) \end{cases} \tag{9}$$

Here $\hat{\mathbf{s}}^{(k)}$ is the estimate of k-th block values based on observations $\mathbf{z}^{(1)}$, ..., $\mathbf{z}^{(k)}$; $\mathbf{V}(k/k)$ and $\mathbf{V}(k/k-1)$ are covariance matrices of filtering and prediction errors respectively.

The number of calculations of KF internal matrices is now reduced in l times (as compared with KF based on (3)). On the other hand, one can notice that the sizes of these matrices increase in l/p times. As the number of operations, necessary for inversion of a matrix, is proportional to the third degree of its size, direct realization of algorithm (9) would increase the total number of operations in $O(l^2)$ times as compared with the standard KF based on (3). That is why we developed economical equivalent of algorithm (9) based on the properties of matrices in representation (5). It will be shown (see "Experiments" section) that this method not only allows to reduce computational expenses in comparison with the standard KF, but also provides their decreasing when the block length is increased (the increasing of a block length also diminishes filtering error, because more future observations are involved). KF taking into account future observations is already discussed in the literature (two-pass and three-pass smoothing – see review in Kybic (1998)), but such algorithms have quite a difficult realization (for instance, they use iterative KF as one of the stages).

ESTIMATION OF AR PARAMETERS AT NOISE BACKGROUND

To apply proposed KF we have to know the values of matrices $\mathbf{F}$ and $\mathbf{G}$, which, in turn, are defined by AR parameters of (1). As was previously mentioned, standard methods of AR parameters calculation lose their efficiency in the presence of background noise. Among noiseproof methods we must emphasize those based on the principle of maximum likelihood. The search for local maxima of a likelihood function leads to a complicated set of non-linear equations, which cannot be immediately solved. That is why different iterative methods are used for its solution. While these procedures are quite awkward and laborious, they do not guarantee convergence to the optimal estimates of AR parameters.

That is why we propose a novel method of estimation, based on vector quantization (VQ). We refrain from AR parameters estimation on the continuous $(p+1)$-dimension space. As was shown in our investigations, to realize adaptive KF it is enough to determine what vector from a limited set of typical AR parameters $\{\mathbf{a},g\}=\{(a_1,...,a_p)^T,g\}^T$ corresponds to the noisy speech frame being processed. The number of such vectors (quantums) can be relatively low (several hundred). In this case there is no need in the exact maximization of the likelihood function $p(\mathbf{Z}/\mathbf{a},g)$ (or a posteriori probability function in the case of Bayesian estimation), because the estimation procedure is reduced to the verification of a limited set of hypotheses. This means a comparison of $p(\mathbf{Z}/\mathbf{a}^{(k)},g^{(k)})$, $k=1,...,N$, where N is a number of quantums; $\{\mathbf{a}^{(k)},g^{(k)}\}$ is a k-th quantum of AR parameters; $\mathbf{Z}$ is a vector of observations. Such a verification can be realized by direct maximization of the likelihood function (or a posteriori probability function) on the set of AR quantums. The optimal estimate is represented by quantum maximizing the chosen function.

The calculation of typical quantums can be performed using well-known ideas of VQ (Buzo (1980), Makhoul (1985)), which were introduced in connection with the problem of low-rate speech coding (in the noise-free case). The main idea of VQ lies in the approximation of possible AR parameters of speech $\{\mathbf{a},g\}=\{(a_1,...,a_p)^T,g\}^T$ by a relatively low quantity of AR quantums. These quantums can be determined by applying an iterative clustering K-means algorithm to the learning speech sequence. The accuracy of representation of possible AR parameters depends on the application and can be improved by increasing the number of quantums, extending the size of the learning sequence and using as many speakers with different voice characteristics as possible. Especially high effectiveness can be reached in systems intended for a limited number of speakers.

Now we will describe in more detail realization of the proposed estimation procedure. Consider frame $\mathbf{Z}$ of noisy speech. We assume that its length L is big enough (probably of several hundred samples), so the dependence on the previous frame can be neglected and one can write (5) in the following form (we drop superscripts for the number of frame):

$$\mathbf{Z}=\mathbf{G}\mathbf{W}+\mathbf{V} \tag{10}$$

where $\mathbf{W}$ and $\mathbf{V}$ are the corresponding blocks of excitation and noise. For observations (10) one can write the likelihood function:

$$p(\mathbf{Z}/\mathbf{a},g\}=\frac{1}{(2\pi)^{\frac{L}{2}}\sqrt{\det(\mathbf{G}\mathbf{G}^T+\sigma_v^2\mathbf{E}_L)}}\exp[-\frac{1}{2}\mathbf{Z}^T(\mathbf{G}\mathbf{G}^T+\sigma_v^2\mathbf{E}_L)^{-1}\mathbf{Z}] \tag{11}$$

As the maximum likelihood estimate we choose a quantum of AR parameters maximizing (11) (for every quantum $(\mathbf{a}^{(k)},g^{(k)})$, $k=1,...,N$ it is necessary to calculate corresponding matrix $\mathbf{G}$ by formula (8)). In the case of Bayesian estimate (maximum a posteriori probability estimate) we have to maximize sequence $\{p(\mathbf{Z}/\mathbf{a}^{(k)},g^{(k)})*P(\mathbf{a}^{(k)},g^{(k)})\}$ containing N elements, where $p(\mathbf{Z}/\mathbf{a}^{(k)},g^{(k)})$ is defined by (11) and $P(\mathbf{a}^{(k)},g^{(k)})$ is the probability of k-th quantum appearance (it is determined beforehand using the learning sequence). We developed effective procedure of calculation of (11), so that the described verification of hypotheses can effectively be realized even for a high quantity of quantums. Also suboptimal approaches for reduction of calculations, as in Buzo (1980), can be used.

Our approach can independently be applied to the important problem of noiseproof speech coding. In previous works coding of noisy speech consisted of 2 stages – estimation of AR parameters was made and only then their quantization was performed (Lim & Oppenheim (1979)). Here these two stages are replaced by one.

EXPERIMENTS

In this section we will describe modeling of estimation and filtering methods considered in previous sections. We have to note that all described experiments were held on the artificial AR sequences with parameters typical for human speech. It allowed us to control the conditions of the experiments.

Figure 1 shows the Itakura-Saito errors of AR parameters estimation (as functions of the signal-to-noise ratio - SNR) resulted from the standard autocorrelation method (ACM) of linear prediction, from widely used Wiener-EM algorithm of Lim & Oppenheim (1978) and from our maximum likelihood (ML) method, using 32 quantums of AR parameters. The order of the AR model was equal to 10. AR parameters of signal were changed every 160 samples.

From results presented in Figure 1 one can see that our method is advantageous as compared with the Wiener-EM algorithm. In particular, for SNR=0 dB, Itakura-Saito distortion is decreased on more than 20 percent.

Now let's consider modeling of filtering methods. Special attention is paid here to investigation of KF

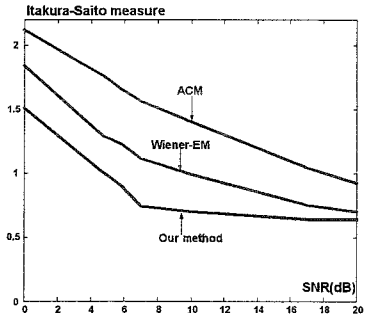

Figure 1. Comparison of different estimation methods.

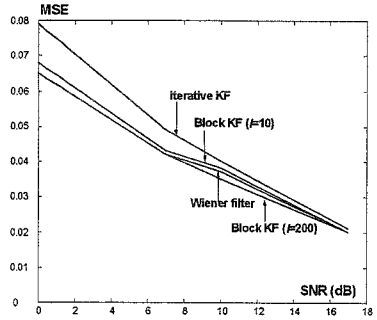

Figure 2. Mean-square error for non-adaptive methods.

methods based on proposed block SS representation. At first we will describe results for the situation when AR parameters are exactly known, i.e. there is no need for adaptation. Figure 2 illustrates filtering mean-square errors (MSE) for the following methods: standard iterative KF (based on representation (3)), block KF (l=200), block KF (l=10) and Wiener filtering (l=200). The order of AR model was equal to 10. AR parameters of signal were changed every 200 samples.

It can be seen that all methods provide lower error values as compared with iterative KF. In particular, for SNR of 0 dB, the advantage of block KF (l=200) and Wiener filter is approximately 25%.

We also studied the question of computational expenses of block KF. All data are presented in table 1.

Filtering method	Flops	Relative flops
Iterative KF	246929	1
Block KF, l=10	111200	0.45
Block KF, l=20	103040	0.42
Block KF, l=50	88692	0.36
Block KF, l=100	83846	0.34
Block KF, l=200	81423	0.33
Wiener filtering	70487	0.29
Optimal linear estimate	45778	0.19

Table 1. Computational expenses of filtering methods

As can be seen, all variants of block KF are advantageous as compared with iterative KF based on (3). Besides, the number of necessary computations decreases when the block length l is increased. Another significant result is the alignment of computational expenses of block KF (which takes into account all previous history of signal) and Wiener filter (which operates only with current frame of noisy signal). We also considered optimal linear estimate on the blocks of $l = 200$ samples (it corresponds to the case $\mathbf{F} = \mathbf{0}$ in (5)). Earlier this estimate was obtained by the double usage of standard KF based on (3) (see review in Kybic (1998)). That is why the computational savings of our

method are more than tenfold. Also such optimal linear estimation provides essential economy (in 1.5 times) in comparison with Wiener filter.

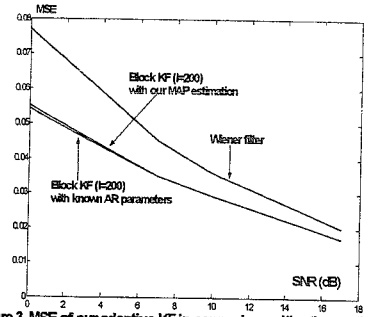

Figure 3. MSE of our adaptive KF in comparison with other methods.

Now let's consider adaptive filtering (fig. 3). The order of AR model was equal to 14. AR parameters of signal were changed every 200 samples. We can see that block KF $(l = 200)$ with adaptation by our method of MAP estimation works almost like block KF $(l = 200)$ with known parameters. The proposed method has a 45% advantage in filtering error at SNR=0 dB when compared with standard Wiener filter (with transfer function defined by spectrum of observations and the spectrum of noise).

CONCLUSION

In this paper we have considered a novel method of adaptive Kalman filtering of noisy speech. The method is based on the block model of autoregressive signal in the state space. It was shown that such a representation allows us to reduce computational expenses and to decrease filtering error as compared with earlier methods. We also developed an essentially new method of estimation of autoregressive parameters in the presence of noise. It is based on the usage of optimal Bayesian estimation and vector quantization. This method can independently be applied to the important problem of noiseproof speech coding.

REFERENCES

Boll, S. (1979) "Suppression of acoustic noise in speech using spectral subtraction", IEEE Transactions on ASSP-27, 113-120.

Buzo, A. & Gray, A.H. & Gray, R.M. & Markel, J.D. (1980) "Speech coding based on vector quantization", IEEE Transactions on ASSP-28, 562-574.

Gannot, S. & Burnstein, D. & Weinstein, E. (1998) "Iterative and sequential Kalman filter-based speech enhancement algorithms", IEEE Transactions on speech and audio processing-6, 373-385.

Goh, Z. & Tan, K.-C. & Tan, B. (1999) "Kalman filtering speech enhancement method based on voiced/unvoiced speech model", IEEE Transactions on speech and audio processing-7, 510-525.

Kybic, J. (1998) "Kalman filtering and speech enhancement. Diploma work", http://cmp.felk.cvut.cz/~kybic/dipl.

Lim, J. & Oppenheim, A. (1978) "All-pole modeling of degraded speech", IEEE Transactions on ASSP-26, 197-210.

Lim, J. & Oppenheim, A. (1979) "Enhancement and bandwidth compression of noisy speech", Proc. of IEEE, vol.67, 1586-1604.

Makhoul, J. & Roucos, S. & Gish, H. (1985) "Vector quantization in speech coding", Proc. of IEEE, vol.73, #11.

Markel, J. & Gray, A. (1976) "Linear prediction of speech", Springer-Verlag, 26-48.

Rabiner, L. & Schafer, R. (1978) "Digital processing of speech signals", Englewood Cliffs, NJ: Prentice Hall, 370-372.

Sage, A.P. & Melse, J.L. (1972) "Estimation theory with application to communication and control", N.-Y. McGraw-Hill, 251-317.