SPEECH ENHANCEMENT WITH KALMAN-FILTERS IN SUBBANDS

Henning Puder

Signal Theory, Darmstadt University of Technology
Merckstr. 25, 64283 Darmstadt, Germany
henning.puder@nt.tu-darmstadt.de

ABSTRACT

This paper presents an approach for a Kalman filter in subbands to enhance noise-corrupted speech signals. We focus especially on the estimation of the necessary speech models. The resulting algorithm has less than 39 ms delay and thus fulfills the ETSI requirements for mobile telephony.

1. INTRODUCTION

In recent years, speech processing systems for cars, such as hands-free telephones or voice controlled operations, have become increasingly popular. They all require noise reduction, increasing both the communication comfort and the recognition rate of voice controlled systems. Especially for hands-free car phones, customers' quality demands are steadily increasing, while developers still are required to limit the delay of the algorithms to 39 ms. In this contribution, a noise reduction algorithm based on Kalman filtering is proposed that requires parametric spectral estimation of both speech and noise signals. These parametric models are able to resolve precisely the pitch components of speech signals. Since pitch components are necessary to preserve the natural sound of speech [1], the Kalman filter offers a high potential for good noise reduction. Additionally, we will show that, with this algorithm, it is possible to fulfill the ETSI delay requirements for mobile telephony.

In order to avoid high model orders – necessary to model the pitch components – the Kalman filtering is performed in subbands.

In the following, we propose a Kalman filter for noise reduction in subbands. We focus especially on the necessary estimation of the speech and noise models. Additional enhancement can be obtained by adjusting and enhancing the AR models of the lower frequency bands, which usually exhibits the lowest SNR (signal to noise ratio), to the estimated pitch frequency.

2. KALMAN FILTERS FOR COLOURED NOISE

The disturbed speech signal $x(k)$ can be considered as the sum of the pure speech and the car noise: $x(k) = s(k) + n(k)$. These superimposing signals can be modelled as AR processes, with white noise processes $w(k)$ and $\eta(k)$, respectively

$$s(k) = \sum_{i=1}^{p} a_i(k)\, s(k-i) + w(k) \tag{1}$$

$$n(k) = \sum_{i=1}^{q} b_i(k)\, n(k-i) + \eta(k) \tag{2}$$

where p and q, indicate the model orders [2].

With $\mathbf{s}(k) = [s(k-p+1), \ldots, s(k)]^T$ and $\mathbf{n}(k) = [n(k-q+1), \ldots, n(k)]^T$ these equations can also be written in the state-space domain,

$$\begin{aligned} \mathbf{s}(k) &= \mathbf{A}_s(k-1)\,\mathbf{s}(k-1) + \mathbf{g}_s\, w(k) && (3)\\ s(k) &= \mathbf{h}_s^T\,\mathbf{s}(k) && (4)\\ \mathbf{n}(k) &= \mathbf{A}_n(k-1)\,\mathbf{n}(k-1) + \mathbf{g}_n\, \eta(k) && (5)\\ n(k) &= \mathbf{h}_n^T\,\mathbf{n}(k) && (6) \end{aligned}$$

with

$$\mathbf{A}_s(k) = \begin{bmatrix} 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \\ a_p(k) & a_{p-1}(k) & \cdots & a_1(k) \end{bmatrix}_{p\times p}, \quad \mathbf{g}_s = \mathbf{h}_s = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}_{p\times 1} \tag{7}$$

For $\mathbf{A}_n(k)$, $\mathbf{g}_n$ and $\mathbf{h}_n$ the variables a and p have to be replaced by b and q.

The common notation in the state-space domain is given by:

$$\begin{aligned} \mathbf{x}(k) &= \mathbf{A}_x(k-1)\,\mathbf{x}(k-1) + \mathbf{G}\,\mathbf{v}(k) && (8)\\ x(k) &= \mathbf{h}_x^T\,\mathbf{x}(k) && (9) \end{aligned}$$

with:

$$\mathbf{x}(k) = \begin{bmatrix} \mathbf{s}(k) \\ \mathbf{n}(k) \end{bmatrix} \quad \mathbf{v}(k) = \begin{bmatrix} w(k) \\ \eta(k) \end{bmatrix} \quad \mathbf{G} = \begin{bmatrix} \mathbf{g}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{g}_n \end{bmatrix}$$

$$\mathbf{A}_x(k) = \begin{bmatrix} \mathbf{A}_s(k) & \mathbf{0} \\ \mathbf{0} & \mathbf{A}_n(k) \end{bmatrix} \quad \mathbf{h}_x = \begin{bmatrix} \mathbf{h}_s \\ \mathbf{h}_n \end{bmatrix} \tag{10}$$

The Kalman equations for the given model can be denoted as:

$$\begin{aligned} \hat{\mathbf{x}}(k) &= \mathbf{A}_x(k-1)\,\hat{\mathbf{x}}(k-1) + \\ &\quad \mathbf{K}(k)[x(k) - \mathbf{h}_x^T\mathbf{A}_x(k-1)\,\hat{\mathbf{x}}(k-1)] && (11)\\ \mathbf{K}(k) &= \mathbf{P}(k|k-1)\mathbf{h}_x\,[\mathbf{h}_x^T\mathbf{P}(k|k-1)\mathbf{h}_x]^{-1} && (12)\\ \mathbf{P}(k|k-1) &= \mathbf{A}_x(k-1)\mathbf{P}(k-1)\mathbf{A}_x^T(k-1) + \\ &\quad \mathbf{G}\mathbf{V}(k)\mathbf{G}^T && (13)\\ \mathbf{P}(k) &= [\mathbf{I} - \mathbf{K}(k)\,\mathbf{h}_x^T]\,\mathbf{P}(k|k-1) && (14) \end{aligned}$$

where $\hat{\mathbf{x}}(k)$ is the estimate of $\mathbf{x}(k)$, $\mathbf{K}(k)$ the Kalman gain, $\mathbf{P}(k|k-1) = \mathrm{E}\{[\mathbf{x}(k) - \mathbf{A}_x(k-1)\,\hat{\mathbf{x}}(k-1)]\,[\mathbf{x}(k) - \mathbf{A}_x(k-1)\,\hat{\mathbf{x}}(k-1)]^T\}$ the prediction-error covariance matrix, and $\mathbf{P}(k) = \mathrm{E}\{[\mathbf{x}(k) - \hat{\mathbf{x}}(k)]\,[\mathbf{x}(k) - \hat{\mathbf{x}}(k)]^T\}$ the filtering-error covariance matrix. The covariance matrix of $\mathbf{v}(k)$ is defined as $\mathbf{V}(k) = \mathrm{diag}(\sigma_w^2(k), \sigma_\eta^2(k))$.

Thus, the estimate of the undisturbed speech signal can be obtained by $\hat{s}(k) = [\mathbf{h}_s^T \mathbf{0}]\,\hat{\mathbf{x}}(k)$.

When determining the AR model of a speech signal, one can observe that model orders $ord \geq f_{sample}/f_{pitch}$ are necessary to resolve the pitch components. At a sampling frequency of $f_{sample} = 8$ kHz, this results in model orders larger than 80 for pitch frequencies $f_{pitch} < 100$ Hz.

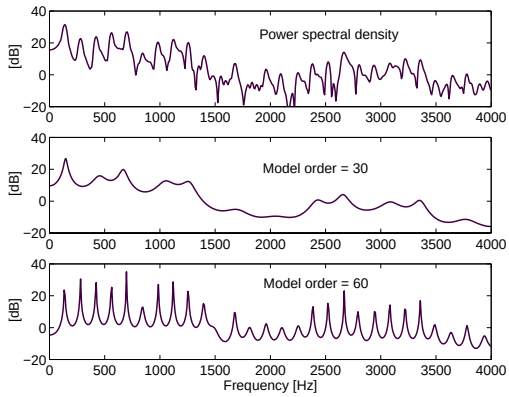

Figure 1: Comparison of the reference PSD (upper) with AR-models of different order (middle and lower) for a voiced signal block with $f_{pitch} = 140$ Hz.

Fig. 1 compares the power spectral density (PSD) of a voiced signal block (upper graph) with AR-speech models of order $ord = 0.5{*}f_{sample}/f_{pitch}$ (middle graph) and $ord = f_{sample}/f_{pitch}$ (lower graph). It demonstrates the necessity of a high model order.

On the one hand, these high orders cause problems for the estimation of the speech models. On the other hand, the Kalman equations become very large, resulting in a high computional demand. Therefore, we first decompose the speech signal in subbands. These subband signals are then filtered separately with Kalman filters of lower order. For the frequency decomposition, we chose a 16 channel filterbank.

3. MODEL ESTIMATION

The methods for the estimation of the speech and noise models have to cope with the fact that speech and noise cannot be measured separately. Noise is superimposed on speech and the noise model can only be estimated during speech pauses.

To estimate the models, the input signal is decomposed in blocks of $N = 32$ samples, which is equivalent to 48 ms when using a subsampling rate of 12 for the filterbank. After every $K = 5$ samples, the models are updated. The best results are obtained when the current filtered input sample lies in the middle of the signal block used for the model estimation. This results in a signal delay of $N/2$ samples ($\hat{=}$24 ms). Additional delay is provided by the filterbank. Our 16-channel filterbank utilises a prototype lowpass filter with length 64, corresponding to a signal delay of 8 ms. In total, the algorithm produces a delay of 32 ms.

3.1. Speech models

For the speech signal, model orders of about 5-7 should be chosen for the lower frequency bands, in order to obtain sufficient resolution of the pitch components. For the higher frequency bands, the model order can be reduced to two. We compared different methods for AR modeling [2, 3]. In particular, we investigated the EM algorithm of [4]. This algorithm describes a recursive iteration for every signal block. The iteration utilizes the Kalman estimates for an enhanced model estimation and vice versa. At the beginning, a good initial model estimate is necessary to allow a first application of the Kalman equations. At the end of every iteration, the Kalman filter yields estimates of the error covariance matrix $\mathbf{P}(k)$ and the state $\hat{\mathbf{x}}(k)$. These are used to build the correlation matrix

$$\mathbf{Q}(k) = \sum_{k_0=k-N/2+1}^{k+N/2} \mathbf{P}(k_0) + \hat{\mathbf{x}}(k_0)\hat{\mathbf{x}}^T(k_0). \qquad (15)$$

This matrix contains estimates of the autocorrelation matrices of the signals $s(k)$ and $n(k)$ which can be utilized for enhanced estimates of the signal models. Performing 10-15 iterations, the algorithm usually converges towards better model estimates. In particular, the speech models exhibit higher maxima at multiples of the pitch frequency. Nevertheless, the algorithm also tends to amplify small model maxima of voiceless sections and speech pauses. This yields stronger musical tones. Therefore, we decided to apply a non-recursive estimation method and enhance the speech models explicitly during voiced sections based on the estimated pitch frequency (see Section 4).

Burg's method [5], which minimizes the sum of the powers of the forward ($e_i^+(k)$) and backward ($e_i^-(k)$) prediction error, proved to be the most powerful of these direct estimation methods. We used noisy speech as input signal. As long as the pitch components exhibit larger power than the car noise, the results are satisfactory (see also Section 4). The Burg algorithm estimates the Parcor coefficients $\Gamma_i(k); i = 1 \ldots p$

$$i = 0 : e_0^+(k_0) = e_0^-(k_0) = x(k_0);$$
$$k_0 \in [k - N/2 + 1 \ldots k + N/2] \qquad (16)$$

$$i = 1 \ldots p : \Gamma_i(k) = \frac{2 \sum_{k_0=k-\frac{N}{2}+i}^{k+\frac{N}{2}} e_{i-1}^+(k_0)[e_{i-1}^-(k_0-1)]^*}{\sum_{k_0=k-\frac{N}{2}+i}^{k+\frac{N}{2}} \left\{|e_{i-1}^+(k_0)|^2 + |e_{i-1}^-(k_0-1)|^2\right\}} \qquad (17)$$

$$e_i^+(k_0) = e_{i-1}^+(k_0) - \Gamma_i(k)\, e_{i-1}^-(k_0-1) \qquad (18)$$

$$e_i^-(k_0) = e_{i-1}^-(k_0-1) - \Gamma_i^*(k)\, e_{i-1}^+(k_0) \qquad (19)$$

which can be converted into the AR parameters ($a_i(k)$) with the Levinson-Durbin recursion. The time index

k lies in the middle of the signal block for which the models are estimated.

Besides the AR parameters, the variance $\sigma_w^2(k)$ has to be estimated. The estimate determined by the Burg algorithm described above is distorted by the noise and cannot be utilized. Our approach is the following:

$$\begin{aligned}\sigma_w^2(k) &= \mathrm{E}\{|s(k) - \sum_{i=1}^{p} a_i^*(k)\, s(k-i)|^2\} \\ &= s_{ss,k}(0) - 2\,\mathrm{Re}\{\sum_{i=1}^{p} a_i^*(k) s_{ss,k}(i)\} \\ &\quad + \sum_{i=1}^{p}\sum_{j=1}^{p} a_i^*(k) a_j(k)\; s_{ss,k}(i-j) \quad (20)\end{aligned}$$

utilizing the estimated AR parameters $a_i(k)$ and the autocorrelation values $s_{ss,k}(i)$; $i = 0, \ldots, p$ which are calculated as follows:

$$s_{ss,k}(i) = IDFT\{S_{PSD}(j,k)\} \quad (21)$$

$$S_{PSD}(j,k) = \max\{X_{PSD}(j,k) - N_{PSD}(j,k), 0\} \quad (22)$$

$$X_{PSD}(j,k) = |X(j,k)|^2 \quad (23)$$

$$N_{PSD}(j,k) = \begin{cases} \beta\, N_{PSD}(j,k-1) & \text{speech pause} \\ +(1-\beta)|X(j,k)|^2 & \\ N_{PSD}(j,k-1) & \text{else} \end{cases} \quad (24)$$

The autocorrelation values $s_{ss,k}(i)$ are based on the difference of the speech and noise PSDs (smoothed with $\beta = 0.93$). The spectral components $X(j,k)$ are the values of the Short Time Fourier transform of the noisy subband signal $x(k)$. The index j indicates the frequency. To ensure that the PSD of the resulting ACF is greater than zero for any frequency bin, we do not determine the difference of the ACF in the time domain. Instead, we calculate the difference in the frequency domain using the PSDs and limit the result to zero before transforming it back into the time domain. Simulations proved the remarkable advantage of this method. In the above equations, we did not mark the different subband signals explicitly. Of course, the calculations are performed for every subband.

We also tried to calculate the AR parameters $a_i(k)$ with the Yule-Walker equations based on $s_{ss,k}(i)$, instead of using the Burg algorithm. This method gave worse results because the models varied more intensively in time, which provoked musical tones.

3.2. Noise models

Since car noise spectra do not exhibit large maxima, model orders of 2 are sufficient for every subband. For the model estimation we used the smoothed $N_{PSD}(j,k)$ of Eqn. 24. Three autocorrelation coefficients, $s_{nn,k}(i)$, $i = 0, \ldots, 2$ determined by inverse Fourier-Transform are sufficient to calculate the two coefficients $b_1(k)$ and $b_2(k)$ of the noise model. Stability problems seldom occur and can be checked with the Levinson-Durbin recursion. The variance $\sigma_\eta^2(k)$ can then be calculated with Eqn. 20 where s, a and p should be replaced by n, b and q.

3.3. Overestimation and Residual Noise

Due to the variance of the noise power, which cannot be modelled with the estimate according to Eqn. 24, the noise reduction algorithm based on Kalman filters also generates slight musical tones. They are less powerful and annoying compared to the Wiener solution and can be eliminated by a slight overestimation of $N_{PSD}(j,k)$ in Eqn.22 and $\sigma_\eta^2(k)$.

It is also appropriate to let some residual noise pass to preserve the natural sound of the speech. This can be obtained by adding a fraction (3-8 %) of the disturbed input signal to the output of every subband Kalman filter.

4. ENHANCEMENT OF SPEECH MODEL ESTIMATION

We already mentioned in Sec. 3.1 that it is possible to enhance the quality of the noise reduction algorithm specifically by amplifying or generating the maxima of the speech models at multiples of the pitch frequency. In the following, we propose a method that realizes this idea.

For the lowest subband, it is difficult to resolve the pitch components. Instead, model maxima (poles of the AR polynomial) occur below the pitch frequency (see Fig. 2 left, grey line), due to the large noise components. For the higher frequency bands, this problem is less important. Usually, the models represent the pitch structure quite well, however, with smaller maximum values compared to the models of the clear speech signal.

First, we propose a method to enhance the speech model for the lowest subband with the following three-step procedure:

1. Detect the pitch frequency based on the second subband (250 - 750 Hz), (s. Fig. 2, right).
2. Suppress maxima of the speech model far below the pitch frequency.
3. Place poles of the AR model at multiples of the pitch frequency.

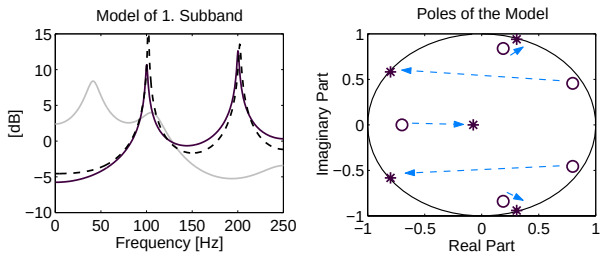

Figure 2: Speech signal models. The estimated model (grey) of the first subband is modified (black).

Fig. 2, right side, depicts the location of the poles of the initial estimation (o) and after the modification with estimated pitch frequency (*). According to step 2, the amplitude of the poles at frequencies below the pitch frequency is reduced, in order to attenuate the low-frequency noise components. The other poles are moved towards multiples of the pitch frequency.

The resulting model is depicted in Fig. 2, black graph on the left. The estimated model is clearly a close approximate to the true model (dashed graph).

For comparison, the model of the initial estimation is given by the grey graph.

While step 2 is performed permanently, based on the last available pitch estimate, the third step is only executed when the pitch frequency can be estimated.

The greatest quality improvement is obtained by the modifications of the model described for the lowest subband. Further enhancements are possible by amplifying the model maxima at multiples for the pitch frequency for the subbands 2-4. This can be performed by moving the poles, corresponding to the estimated pitch frequency, closer to the unit circle. Fig. 3 depicts the result for an example of the second subband. However, the model maxima should only be amplified

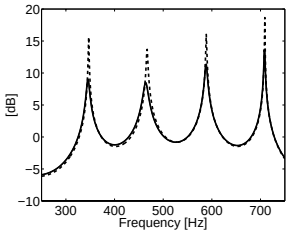

Figure 3: Speech signal models. The estimated model (solid) of the second subband is modified (dashed).

but not changed in frequency because the pitch components are no longer located exactly at multiples of the pitch frequency at higher frequencies.

The pitch frequency, necessary for the improved model estimation, is estimated from the second subband (250-750 Hz). This is a good choice because on the one hand, the pitch components are quite strong in this frequency band compared to the higher bands, and on the other hand, in the case of car noise, the SNR is much higher in this frequency band, compared to the lowest one. A frequency band of 500 Hz (even 666 Hz if one considers the overlapping regions to the neighbouring bands at a subsampling rate of 12) usually contains two to four pitch components, and this is sufficient to estimate the pitch frequency. The algorithm first determines the poles which are the closest to the unit circle. A pitch component is then detected if other poles are located close to the unit circle at double, half, etc., of the frequency of the closest pole. The pitch frequency is then estimated by averaging over the relevant poles. Based on these steps, we have designed a reliable algorithm for pitch estimation that considers past decisions and eliminates unlikely decisions.

5. COMPUTATIONAL LOAD

Due to the matrix multiplications, the Kalman filter seems to involve an enormous computational load. However, as the matrices and vectors have few non-zero elements, only $4(p+q)+8$ multiplications per subband are necessary. Additional effort is required for the estimation of the signal models. Here, the Burg algorithm and the FFTs and IFFTs that are necessary to determine $\sigma_w^2(k)$ make the largest contribution. As the models only have to be estimated for signal blocks every K samples, the order of the computational load remains comparable to the classical Wiener solution. The methods described to enhance the estimates of the speech model require additional computational power, especially due to the calculation of the poles of the models and the pitch estimation. The effort strongly depends on the platform of implementation. Nevertheless, it is also limited to reasonable orders, because the model enhancement is performed at a subsampled rate for only a few subbands.

6. RESULTS AND CONCLUSIONS

In this paper, we presented a noise reduction algorithm based on Kalman filtering. To limit the model orders to reasonable values, the input signal is first decomposed in 16 subbands. We focused especially on the estimation of the speech model parameters. The best results were obtained for the speech model estimated with the Burg algorithm. To enhance the estimates, we moved the poles of the the lowest subband towards multiples of the pitch frequency. The determinations of the necessary pitch estimates are based on the second subband. Additionally, the pitch maxima of the higher subbands may also be amplified. With its 32 ms delay, the algorithm fulfills the ETSI requirements for mobile telephones. One result obtained for a highly

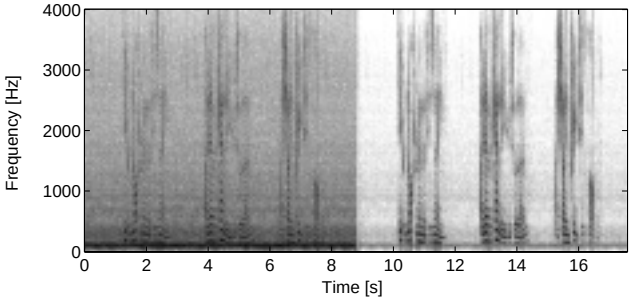

Figure 4: Disturbed speech signal (left) and enhanced signal (right)

corrupted speech signal is depicted in Fig. 4. Two properties of the output signal should be emphasized: The uniform residual noise without musical tones and the strong pitch components for the low frequencies generate a speech signal which sounds natural.

7. REFERENCES

[1] J. Tilp: *Single-Channel Noise Reduction with Pitch-Adaptive Post-Filtering,* Proc. EUSIPCO-2000, vol. 3, pp. 1851-1854, Tampere, Finland, September 2000

[2] W.R. Wu, P.C. Chen: *Subband Kalman Filtering for Speech Enhancement,* IEEE Trans. on Circuits and Systems-II, vol. 45, no. 8, pp. 1072-1083, August 1998

[3] G. Doblinger: *An Adaptive Kalman Filter for the Enhancement of Noisy AR Signals,* Proc. of the IEEE ISCAS-98, vol. 5, pp. 305-308, 1998

[4] S. Gannot et. al.: *Iterative and Sequential Kalman Filter-Based Speech Enhancement Algorithm,* IEEE Trans. on Speech and Audio Processing, vol. 6, no. 4, July 1998

[5] M.H. Hayes: *Statistical Digital Signal Processing and Modelling,* John Wiley & Sons, Inc., New York, 1996