ACOUSTIC NOISE REDUCTION USING A MULTIPLE-INPUT SINGLE-OUTPUT KALMAN FILTER

Alexander Kaps

signaltheorie

alexander.kaps@spg.tu-darmstadt.de
Darmstadt University of Technology,
Merckstrasse 25, 64283 Darmstadt, Germany

ABSTRACT

In this paper a model-based method for multi-microphone acoustic noise reduction is proposed. The model is able to handle speech, correlated and uncorrelated noise components. These are modelled using AR-processes. The components are parameterized by a combined procedure using Burg's algorithm and a modified autocorrelation method based on voice activity detection (VAD). Additionally, a spectral floor like scheme is applied to suppress musical tones. This model is the basis of a multiple-input single-output Kalman filter. It is implemented in a subband structure by application of polyphase filterbanks. This procedure reduces the model order of the AR-processes drastically without violating the GSM delay boundary of 39 ms. Results show that a significant level of noise reduction can be achieved while keeping user signal degradation low.

1. INTRODUCTION

Acoustic noise reduction methods play an important role in hands-free communication. In this paper we focus on the problem of acoustic noise reduction for hands-free telephone use inside a moving car. In this application, several approaches for a single microphone exist. Only a few of these use the Kalman filter algorithm. Newer methods utilizing more than one microphone generally perform the noise reduction in two steps. First, a beamformer is applied to the microphone outputs. Second, a postfilter, very similar to the single-channel methods, mentioned above is used to enhance the beamformer output. An overview of existing methods – single-channel and multi-channel – can be found in [1].

In this paper a signal model is developed for the multi-channel case which combines beamformer and postfilter functionality and can be used in a Kalman filter structure. This filter has been implemented in a subband scheme using a polyphase filter bank.

This paper is organized as follows: In section 2 the signal model is described which is the basis for the derivation of the Kalman filter in section 3. After that, parametrization of the model is addressed in section 4. Implementation issues and results are presented in sections 5 and 6 respectively. The publication ends with a conclusion and an outlook in section 7.

2. DESCRIPTION OF THE SIGNAL MODEL

In a multi-channel system, the microphone outputs can be divided into two main components: speech and noise. The noise can be further divided into noise that is correlated between the microphone signals and noise that is not. Therefore, an appropriate model should be able to represent all three components, i.e. *speech*, *correlated noise* and *uncorrelated noise*.

Although the proportion of the two noise components depends on the type and speed of the car as well as road conditions, it can be generally stated that the uncorrelated component is dominant over the correlated one. A diffuse noise field is modelled with uncorrelated noise only.

From now on, we restrict ourselves to $N = 2$ microphones. The signal model is shown in Fig. 1. It assumes a single speech source $s(k)$ (the speaker), which is modelled as AR-process of order p with excitation $v(k)$. It is linked to the speech components of the microphone signals $y_1(k)$ and $y_2(k)$ via two time-variant room impulse responses $h_{1,l}(k)$ and $h_{2,l}(k)$. Note that k denotes the discrete time and l the coefficient index.

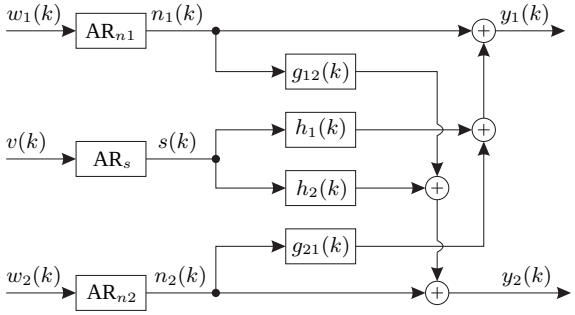

Figure 1: *Signal model.*

The signals $n_1(k)$ and $n_2(k)$ represent the uncorrelated noise components at each microphone. They are again modelled as AR-processes of order q excited by $w_1(k)$ and $w_2(k)$ respectively. These uncorrelated noise components of each microphone are linked to the other channel via the time-variant impulse responses $g_{21,l}(k)$ and $g_{12,l}(k)$ to model correlated noise components in the microphone signals.

The impulse responses $h_{1,l}(k)$, $h_{2,l}(k)$, $g_{21,l}(k)$ and $g_{12,l}(k)$

are time-variant. However, they have been denoted time-invariant in Fig. 1 for clarity.

The excitation signals $v(k)$, $w_1(k)$ and $w_2(k)$ are independently and identically distributed (iid) white noise processes that are uncorrelated to each other. An AR-process of order p with time-varying coefficients is defined as follows:

$$x(k) = \sum_{l=1}^{p} a_{s,l}(k-1)x(k-1) + u(k). \quad (1)$$

In order to obtain a state-space description necessary for the Kalman filter we apply Eq. 1 to our model and utilize matrix/vector notation:

$$\begin{aligned} \mathbf{s}(k) &= \mathbf{A}_s(k\text{-}1)\mathbf{s}(k\text{-}1) + [\,|\,]v(k) & (2)\\ \mathbf{n}_1(k) &= \mathbf{A}_{n1}(k\text{-}1)\mathbf{n}_1(k\text{-}1) + [\,|\,]w_1(k) & (3)\\ \mathbf{n}_2(k) &= \mathbf{A}_{n2}(k\text{-}1)\mathbf{n}_2(k\text{-}1) + [\,|\,]w_2(k) & (4) \end{aligned}$$

where $[\,|\,]$ represents a column vector of appropriate length (p or q) with all elements zero except the last one which equals one. Note that complex values are assumed due to the subband structure implementation. All vectors are defined as columns and are denoted in lower case bold face letters while matrices are in upper case bold face. The following examples show the element order of vectors representing a signal such as $s(k)$ and an impulse response such as $g_{21}(k)$ where the second index refers to the coefficient:

$$\begin{aligned} \mathbf{s}(k) &= [s(k\text{-}p\text{+}1), s(k\text{-}p\text{+}2), \ldots, s(k)]^{\mathrm{T}}\\ \mathbf{g}_{21}(k) &= [g_{21,q\text{-}1}(k), g_{21,q\text{-}2}(k), \ldots, g_{21,1}(k)]^{\mathrm{T}}. \end{aligned}$$

Each matrix is a combination of a shifting matrix and the AR-coefficients [1]:

$$\mathbf{A}_s(k) = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0\\ 0 & 0 & 1 & \cdots & 0\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & 0 & \cdots & 1\\ a_{s,p}(k) & a_{s,p\text{-}1}(k) & a_{s,p\text{-}2}(k) & \cdots & a_{s,1}(k) \end{bmatrix}.$$

In the same manner the output of the model can be written in matrix/vector notation:

$$\begin{aligned} y_1(k) &= \mathbf{h}_1^{\mathrm{H}}(k)\mathbf{s}(k) + [\,|\,]^{\mathrm{T}}\mathbf{n}_1(k) + \mathbf{g}_{21}^{\mathrm{H}}(k)\mathbf{n}_2(k) & (5)\\ y_2(k) &= \mathbf{h}_2^{\mathrm{H}}(k)\mathbf{s}(k) + \mathbf{g}_{12}^{\mathrm{H}}(k)\mathbf{n}_1(k) + [\,|\,]^{\mathrm{T}}\mathbf{n}_2(k). & (6) \end{aligned}$$

The complete model in state-space notation is shown in Fig. 2 where scalar quantities are indicated by thin, vector quantities by bold lines and where the time index k has been omitted for clarity.

3. KALMAN FILTER

In order to find the equations for the Kalman filter the model described in section 2 needs to be combined into one system and one measurement equation. Therefore, we rewrite Eqs. 2-4 as well as Eqs. 5-6 as one equation respectively:

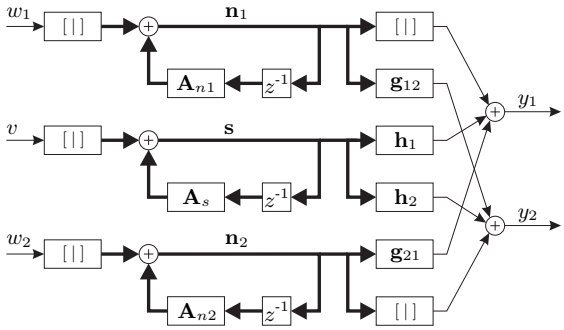

Figure 2: *Signal model in state-space notation.*

$$\begin{bmatrix} \mathbf{s}(k)\\ \mathbf{n}_1(k)\\ \mathbf{n}_2(k) \end{bmatrix} = \begin{bmatrix} \mathbf{A}_s(k) & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{A}_{n1}(k) & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{A}_{n2}(k) \end{bmatrix} \cdot \begin{bmatrix} \mathbf{s}(k\text{-}1)\\ \mathbf{n}_1(k\text{-}1)\\ \mathbf{n}_2(k\text{-}1) \end{bmatrix} \cdots + \begin{bmatrix} [\,|\,] & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & [\,|\,] & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & [\,|\,] \end{bmatrix} \cdot \begin{bmatrix} v(k)\\ w_1(k)\\ w_2(k) \end{bmatrix}$$

$$\begin{bmatrix} y_1(k)\\ y_2(k) \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^{\mathrm{H}}(k) & [\,|\,]^{\mathrm{T}} & \mathbf{g}_{21}^{\mathrm{H}}(k)\\ \mathbf{h}_2^{\mathrm{H}}(k) & \mathbf{g}_{12}^{\mathrm{H}}(k) & [\,|\,]^{\mathrm{T}} \end{bmatrix} \cdot \begin{bmatrix} \mathbf{s}(k)\\ \mathbf{n}_1(k)\\ \mathbf{n}_2(k) \end{bmatrix}.$$

These equations can now be abbreviated:

$$\begin{aligned} \mathbf{x}(k) &= \mathbf{A}(k-1)\mathbf{x}(k-1) + \mathbf{B}\mathbf{u}(k) & (7)\\ \mathbf{y}(k) &= \mathbf{C}^{\mathrm{H}}(k)\mathbf{x}(k) & (8) \end{aligned}$$

where $\mathbf{x}(k)$ denotes the state vector, $\mathbf{A}(k)$ the transition matrix, $\mathbf{B}\mathbf{u}(k)$ the system excitation, $\mathbf{y}(k)$ the measurement vector and $\mathbf{C}(k)$ the measurement matrix. Eq. 7 is referred to as *system equation*, Eq. 8 as *measurement equation*.

The following Kalman filter equations are not derived in detail. One can find detailed derivations in many textbooks, e.g. [1, 2]. Note that all estimates are linear, unbiased and optimal in terms of the mean square error, i.e. there is no better linear estimate.

3.1. Kalman Equations

In the following $\hat{\mathbf{x}}(k|k-1)$ is denotes the a-priori state estimate and $\hat{\mathbf{x}}(k|k)$ the a-posteriori state estimate. $\mathbf{P}_e(k|k-1)$ and $\mathbf{P}_e(k|k)$ denote the a-priori and a-posteriori covariance matrices of the estimation error.

3.1.1. Prediction

$$\begin{aligned} \hat{\mathbf{x}}(k|k\text{-}1) &= \mathbf{A}(k-1)\hat{\mathbf{x}}(k-1|k-1). & (9)\\ \mathbf{P}_e(k|k\text{-}1) &= \mathbf{A}(k-1)\mathbf{P}_e(k-1|k-1)\mathbf{A}^{\mathrm{H}}(k)\cdots\\ &\quad +\mathbf{B}\mathbf{P}_u(k)\mathbf{B}^{\mathrm{T}} & (10) \end{aligned}$$

3.1.2. Correction

$$\mathbf{K}(k) = \mathbf{P}_e(k|k-1)\mathbf{C}(k)\cdots \cdot\left[\mathbf{C}^{\mathrm{H}}(k)\mathbf{P}_e(k|k-1)\mathbf{C}(k)\right]^{-1} \quad (11)$$

$$\hat{\mathbf{x}}(k|k) = \hat{\mathbf{x}}(k|k-1)\cdots +\mathbf{K}(k)\left[\mathbf{y}(k)-\mathbf{C}^{\mathrm{H}}(k)\hat{\mathbf{x}}(k|k-1)\right] \quad (12)$$

$$\mathbf{P}_e(k|k) = \left[\mathbf{I}-\mathbf{K}(k)\mathbf{C}^{\mathrm{H}}(k)\right]\mathbf{P}_e(k|k-1) \quad (13)$$

$\mathbf{P}_u(k)$ denotes the matrix containing the variances of $v(k)$, $w_1(k)$ and $w_2(k)$:

$$\mathbf{P}_u(k) = \begin{bmatrix} \sigma_v^2(k) & 0 & 0 \\ 0 & \sigma_{w1}^2(k) & 0 \\ 0 & 0 & \sigma_{w2}^2(k) \end{bmatrix} \quad (14)$$

4. ESTIMATION OF THE PARAMETERS

In the following section, parametrization of the signal model is addressed and is divided into two subsections. The first deals with the estimation of the AR-parameters while finding the impulse responses is subject of the latter one.

4.1. AR-parameter estimation

Three sets of AR-parameters need to be estimated. One for the speech component and one for the noise component of each microphone. As well as the AR-coefficients, the excitation power or - more precisely - the covariance matrix of the excitation vector $\mathbf{u}(k)$ needs to be estimated.

4.1.1. Finding the AR-coefficients

The difficulty with estimating the AR-coefficients is due to the fact that the speech signal is always disturbed by noise. In speech pauses however, the noise itself can be measured directly.
Following the approach in [3] a combination of Burg's and a modified autocorrelation method with VAD is utilized to overcome the deficiencies of each method (see section 6). This structure is further improved by a spectral-floor like technique.
The input data at every microphone is processed in overlapping blocks. Each block is processed twice. First, they are used for the Burg estimator yielding a first set of speech AR-coefficients for each microphone. The sets from the two microphones are averaged to yield an improved estimate.
Second, they are utilized to calculate the periodogram $\hat{S}_{yy}(n)$ which is an estimate of the power spectral density (PSD) of the measurement. Depending on the VAD, the PSD is recursively smoothed and used as a noise PSD estimate $\hat{S}_{nn}(n)$ during speech pauses. During speech activity however, $\hat{S}_{nn}(n)$ is kept constant and the speech PSD is estimated by subtracting $\hat{S}_{nn}(n)$ from $\hat{S}_{yy}(n)$. In order to prevent unstable AR-models, negative values will be forced to zero. However, this causes audible musical tones and consequently a spectral floor like scheme is introduced to combat those effects:

$$\hat{S}_{ss}(n) = \max\{\hat{S}_{yy}(n) - \hat{S}_{nn}(n), 0\} \quad (15)$$

$$\hat{S}_{ss}^{\text{floor}}(n) = \max\{\hat{S}_{ss}(n), \gamma \cdot \max\{\hat{S}_{ss}(n)\}\} \quad (16)$$

The factor $\gamma \in [0;1]$ is used to adjust the spectral floor increasing the noise level at the output and masking the musical tones. Finally, the second set of speech AR-coefficients for each microphone is calculated by applying the inverse FFT to $\hat{S}_{ss}^{\text{floor}}(n)$. Again, the estimation is improved by averaging the two sets over the different microphones. The improved sets from Burg's and the autocorrelation method are then linearly combined.
The noise AR-coefficients are calculated by applying the inverse FFT to $\hat{S}_{nn}(n)$ at every microphone.

4.1.2. Finding the excitation power

By calculating the predictor error power, the excitation power of the AR-models can be estimated as shown in the following for the speech component:

$$\sigma_v^2(k) = r_{ss,0}(k) + \mathbf{a}_s^{\mathrm{H}}(k)\mathbf{R}_{ss}(k)\mathbf{a}_s(k)\cdots -2\cdot\mathrm{Re}\left\{\sum_{l=1}^{p} a_{s,l}^*(k) r_{ss,l}(k)\right\} \quad (17)$$

where $r_{ss,l}(k)$ is the autocorrelation sequence at time instance k, $\mathbf{R}_{ss}(k)$ the autocorrelation matrix and $\mathbf{a}_s(k)$ the vector of the AR-coefficients $a_{s,l}(k)$.

4.2. Estimation of the impulse responses

The impulse responses $h_{1,l}(k)$ and $h_{2,l}(k)$ can be used to compensate for a direction of arrival (DOA) other than broadside e.g. by usage of a fractional delay filter [4]. Additionally, it can be used to reverse the effects of reverberation, which is the subject of further research.
The impulse responses $g_{21,l}(k)$ and $g_{12,l}(k)$ model the correlated noise components. Setting $g_{21,l}(k) = g_{12,l}(k) = 0$ describes an uncorrelated noise field. However, having VAD available provides an easy way to enhance the model by applying two adaptive filters between the channels and in opposite directions, which adapt $g_{21,l}(k)$ and $g_{12,l}(k)$ during speech pauses. Another possibility is the utilization of a linear predictor in a similar way. Then, $g_{21,l}(k)$ and $g_{12,l}(k)$ represent the predictor coefficients instead of impulse responses.

5. IMPLEMENTATION

The system described in the previous sections has been implemented in a subband structure using polyphase filterbanks. A time-domain implementation is impractical due to the high model order (approximately 60) that would be required for the speech component [3]. Performing the Kalman filtering in subbands provides the advantage of using different model orders at different frequency bands to accommodate for the fact that speech and noise power decreases over frequency.
Specifically, a $M = 16$ bands polyphase filterbank with a length $L = 64$ prototype lowpass filter and subsampling rate of $r = 10$ was used. With data sampled at $f_s = 8$ kHz the lowpass causes a delay of 8 ms.
As the input signals are real valued, usage of the first nine bands is sufficient due to symmetry reasons. Initially, the model order for the speech component was set to $p = 6$ while using $q = 2$ coefficients for the noise components at the same time. The parameter estimation was performed for subband sample with a

block of length 32 and the sample to be calculated in its middle. By using those non-causal samples of half a block length an additional delay of 20 ms is introduced. Therefore, the system causes a total delay of 28 ms. Even with $r = 12$, which is the maximum subsampling rate one should use with a prototype lowpass of that length, the total delay becomes 32 ms. This is still less than the 39 ms required by ETSI for GSM [5].
As the DOA was broadside for the speech component, $h_1(k)$ and $h_2(k)$ were set $h_1(k) = h_2(k) = \delta_K(k)$.

6. RESULTS

For the simulations, plain speech signals were convolved with room impulse responses measured inside a car and then added to recorded car noise. As an example, the spectrogram of the noisy speech signal of microphone one is depicted in Fig. 3.

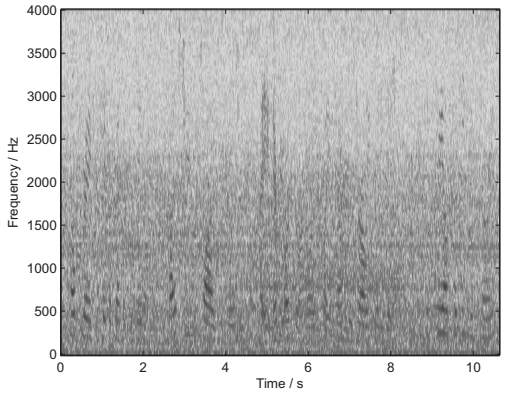

Figure 3: *Spectrogram of the noisy speech signal.*

If the linear combination of the two sets of AR-coefficients is performed in the way that only the autocorrelation method is used, the maximum noise reduction is achieved. However, the speech signal is degraded and the speech quality is poor.
Using Burg's method alone, i.e. utilizing the autocorrelation method only for noise AR-parameter estimation, results in good speech quality but less noise reduction. Therefore, various mixtures of the AR-coefficient sets estimated with Burg's method and the autocorrelation method were tested. Best results are obtained by using a mixture of approximately 70% autocorrelation and 30% Burg's method.
The remaining musical tones can be traded against noise reduction performance by application of a spectral floor setting of $\gamma = 0.05$. Then, the increased noise level masks most of the musical tones providing a good speech quality.
Finally, model estimation was improved by adaptation of $g_{21,l}(k)$ and $g_{12,l}(k)$ via the NLMS algorithm with small step sizes such as $\mu = 0.01$. First results are shown in Fig. 4 where an SNR gain of approximately 6 dB was achieved.
Using a subsampling rate of $r = 12$ and less AR-coefficients at higher frequency band as well as estimating the AR-coefficients not every subband sample but every five samples was already examined in the single-channel case. Those former results indicate that the computational complexity can be drastically reduced without significant loss in performance.

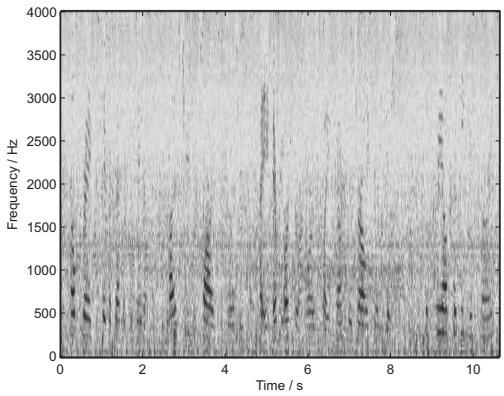

Figure 4: *Spectrogram after noise reduction using the combined method and spectral floor resulting in an SNR gain of 6 dB.*

7. CONCLUSION

A new AR-process based model, which represents speech, correlated and uncorrelated noise components and in which DOA compensation and de-reverberation techniques can be easily integrated, was introduced. It was shown, that the model can be described in the state-space domain making it suitable for a multiple-input single-output Kalman filter implementation. Two methods providing stable AR-models – Burg's method and the autocorrelation method – were presented for estimating the AR-coefficients of speech and noise from a noisy speech signal. Additionally, a method similar to the spectral floor in the single-channel case was applied to combat musical tones. The system was implemented in a subband structure in order to handle the necessary order of the AR-processes. The input signal was decomposed into 16 subbands using a polyphase analysis filterbank with a prototype lowpass filter of length 64 subsampling rate 10. After performing the Kalman filtering the output signal was formed by the corresponding polyphase synthesis filterbank. Results showed that the autocorrelation method yields better noise reduction compared to Burg's method but also a poor speech quality. Burg's method behaves almost contrarily, providing good speech quality with less noise reduction. Combining both methods together with the spectral floor scheme to suppress the remaining musical tones provided the best results.

8. REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*, John Wiley & Sons, Inc., 2004.

[2] M.H. Hayes, *Statistical Digital Signal Processing and Modeling*, John Wiley & Sons, Inc., 1996.

[3] H. Puder, "Kalman-filters in subbands for noise reduction with enhanced pitch-adaptive speech model estimation," *Euro. Trans. on Telecom.*, vol. 13, no. 2, pp. 139–148, 2002.

[4] V. Välimäki and T.I. Laakso, "Principles of fractional delay filters," in *Proc. ICASSP*, Istanbul, Turkey, 2000.

[5] ETS 300 903 (GSM 03.50), *Transmission Planing Aspects of the Speech Service in the GSM Public Land Mobile Network (PLMS) System*, ETSI, France, 1999.