Real-time Speech Enhancement System for Surgical Systems

Maria Avitha Francis[1], Marco Gimm[1], Thomas Becker[2], Gerhard Schmidt[1]

[1] *Digital Signal Processing and System Theory, Christian-Albrechts-Universität zu Kiel, Email: {af, mgi, gus}@tf.uni-kiel.de*

[2] *Klinik für Allgemeine, Viszeral-, Thorax-, Transplantations- und Kinderchirurgie, UKSH, Email: Thomas.Becker@uksh.de*

Abstract

The communication in surgery rooms might be difficult, especially if the surgeon is located in some distance from the patient due to the usage of minimally invasive operation devices. In order to improve the communication between the medical staff around the patient table and the surgeon controlling the operation device, dedicated signal processing can be applied. In particular, such speech enhancement systems focus on suppression of the stationary background noise caused by medical equipment, suppression of the feedback occurring due to a closed electro-acoustic loop, and lastly suppression of instationary noise, e.g. sharp clanging noises. All this should be done without compromising the quality and intelligibility of the speech signals.

In this work, we present an overview of a generic speech enhancement system comprising mainly of a traditional Wiener filter for background noise suppression, a model-based feedback suppression along with additional stages for multidirectional communication and linear and nonlinear time-domain processing for the attenuation of instationary noise. All this was implemented and tested on an embedded hardware platform with respect to the special requirements of a surgery room.

Introduction

Speech enhancement, to be specific, noise suppression and/or echo cancellation is required in several applications and its main objective is to improve the intelligibility and overall perceptual quality of the degraded speech signal [1]. The input to the speech enhancement system is the microphone signal that comprises of the desired speech as well as the undesired noise and feedback. In the case of a surgery room, noise can be broadly classified into two main types:

1. **Stationary Background Noise:** Major sources include medical equipment such as vacuum suction pumps, ECG monitors and electric or air-powered surgical instruments, being at least short time stationary. Also air conditioning system are often used in surgery rooms.

2. **Instationary Clanging Noise:** The dropping of scalpels, forceps and other metallic surgical equipment causes a sudden sharp clanging noise which is extremely loud and may startle the medical team resulting in discomfort and in the worst case scenario, surgical errors.

The speech enhancement system should be designed in order to suppress the aforementioned undesired noise and feedback signals. Also, in this application, since the speaker and the listener are present in the same closed environment, a large processing delay results in the loss of synchronicity between the direct speech signal and the signal played via the loudspeaker resulting in the listener perceiving the loudspeaker signal as an echo [2]. Hence, the parameters of all the implemented modules as well as the hardware should correspond to the least possible processing delay.

The block diagram of the speech enhancement system for a single input and output channel is depicted in Fig. 1.

Subband Processing

In many applications, such as in speech enhancement, it is advantageous to first split the input signal into M different frequency ranges known as subbands and to then process these subband signals in a parallel fashion. The splitting of the signal is achieved by using an analysis filterbank and depending on the properties of the lowpass, bandpass and highpass filters of the analysis filterbank, the sampling rates of these subband signals can be reduced [2]. So, instead of processing the fullband microphone input signal at a high sampling rate, the M subband signals are processed in a parallel fashion at a much lower sampling rate thereby the computation complexity is considerably reduced. Finally, after the subband processing, a synthesis filterbank recombines the M subband signals into a single fullband signal [3]. Fig. 2 illustrates the subband domain processing structure of the speech enhancement system depicted in Fig. 1.

The input to the analysis filterbank is the noisy microphone signal $y(n)$ that comprises of the speech of the local speaker $s(n)$, the background noise $b(n)$ and the feedback $r(n)$

$$y(n) \;=\; s(n) + b(n) + r(n). \tag{1}$$

Noise Estimation

The spectral behaviour of speech signals makes it possible to estimate and track noise even in speech frames. For example, with fricatives like /s/ and /f/, most of the energy is dominant in the higher frequencies making it possible to track noise in the lower frequencies. Similarly voiced sounds are dominant at lower frequencies making it possible to estimate the noise in the higher frequencies [1].

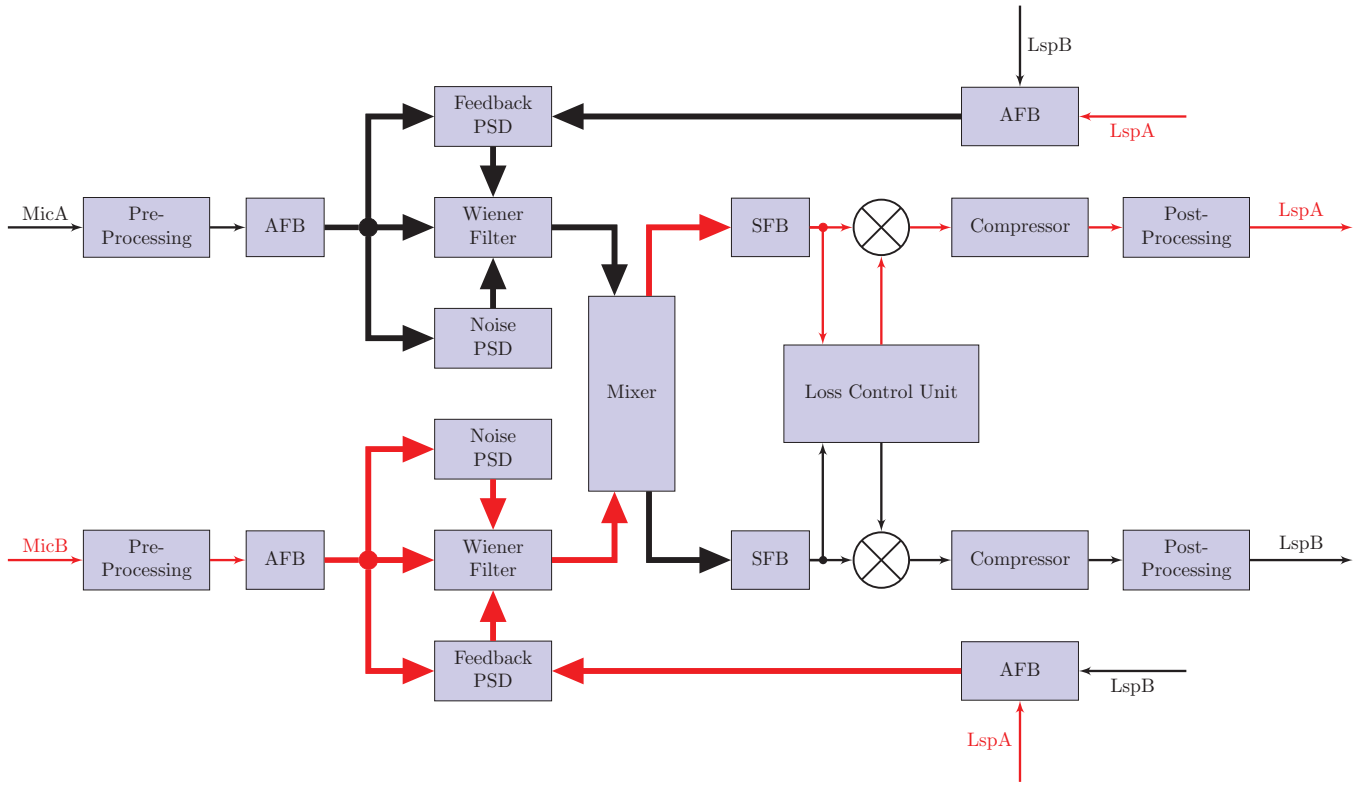

Figure 1: Block diagram of the speech enhancement system

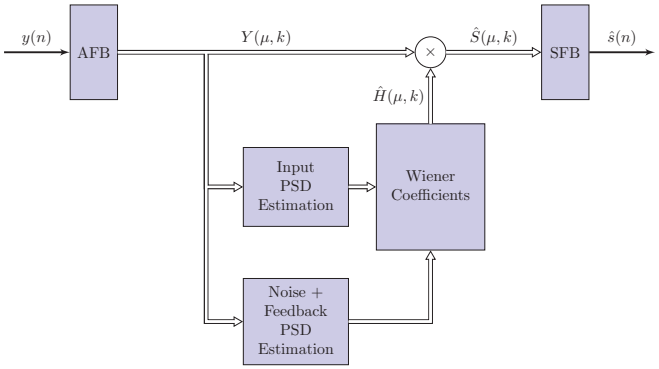

Figure 2: Subband domain processing structure

In order to reduce the variance, the magnitude square of the input spectrum $Y(\mu,k)$, is smoothed along the time axis using a first-order IIR filter according to

$$\overline{|Y(\mu,k)|}^2 \;=\; (1-\alpha_e)\overline{|Y(\mu,k-1)|}^2 \;+\; \alpha_e|Y(\mu,k)|^2, \tag{2}$$

with

$$|Y(\mu,k)|^2 \;=\; Y^2_{\text{real}}(\mu,k) \;+\; Y^2_{\text{img}}(\mu,k), \tag{3}$$

where $\overline{|Y(\mu,k)|}^2$ is the smoothed input spectra magnitude squared and α_e is the smoothing constant. Additionally, smoothing along the frequency axis in the positive and negative directions may also be applied.

Now, depending on the value of $\overline{|Y(\mu,k)|}^2$, the multiplicative constant Δ_{basic} is selected according to [1]

$$\Delta_{\text{basic}}(\mu,k) \;=\; \begin{cases} \Delta_{\text{inc}}, & \text{if } \overline{|Y(\mu,k)|}^2 \;>\; \hat{S}_{bb}(\mu,k-1), \\ \Delta_{\text{dec}}, & \text{if } \overline{|Y(\mu,k)|}^2 \;<\; \hat{S}_{bb}(\mu,k-1), \\ 1, & \text{else,} \end{cases} \tag{4}$$

where Δ_{inc} and Δ_{dec} are the incremental and decremental multiplicative constants respectively. The power spectral density of the background noise is now estimated as described in Eq. (5) [1]

$$\hat{S}_{bb}(\mu,k) \;=\; \Delta_{\text{basic}}(\mu,k)\cdot\hat{S}_{bb}(\mu,k-1). \tag{5}$$

The parameters Δ_{inc} and Δ_{dec} are chosen in order to let the estimation decreasing faster than it increases. This way, the estimation always follows the minimum of $|Y(\mu,k)|^2$, which is expected to be the background noise.

Feedback Estimation

After the signal processing, the output of the final module, in this application, the hard-soft limiter, is given as the input to the loudspeaker. Now, as soon as the loudspeaker plays this signal, feedback is generated. Fig. 3 illustrates the speech enhancement system depicted in Fig. 1 operating in a closed electro-accoustic loop.

Assuming the room impulse response $h_{\text{RIR}}(i)$ to be constant with respect to time, the feedback signal is obtained by convolving the loudspeaker signal $x(n)$ with the room

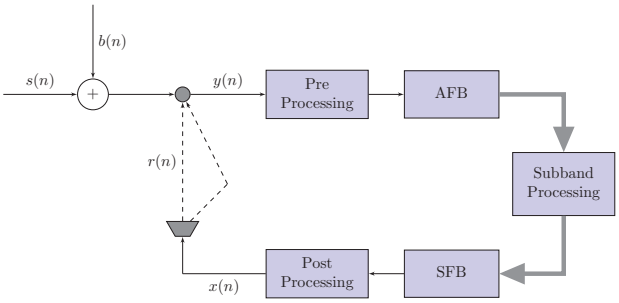

Figure 3: Illustration of the speech enhancement system operating in a closed electro-accoustic loop

impulse response according to according to Eq. (6) [4]

$$r(n) \;=\; \sum_{i=-\infty}^{\infty} x(n-i)h_{\mathrm{RIR}}(i). \tag{6}$$

In the model-based feedback suppression, the room impulse response is replaced with the Polacks model which is described in Eq. (7) [4]

$$h_{\mathrm{RIR,mod}}(i) \;=\; \begin{cases} 0\,, & \text{for } \; i < p, \\ w(i)\; e^{-\alpha(i-p)}\,, & \text{for } \; i \geq p, \end{cases} \tag{7}$$

where $w(i)$ is a Gaussian distributed random process and p is the latency between the loudspeaker and the microphone. The decay behavior α in dependency of the reverberation time T_{60} and the sampling rate f_s is given by [4]

$$\alpha \;=\; \frac{3\ln 10}{T_{60} f_s}. \tag{8}$$

Based on the modeled impulse response, the expected value of the energy envelop in the subband domain can be derived as [4]

$$E\left\{|H_{\mathrm{RIR,mod,A}}(\mu,k)|^2\right\} \;=\; \begin{cases} 0\,, \\ \text{for } k < P(\mu), \\ A(\mu)e^{-\gamma(\mu)(k-P(\mu))}\,, \\ \text{for } k \geq P(\mu). \end{cases} \tag{9}$$

The parameters $P(\mu)$, $A(\mu)$ and $\gamma(\mu)$ describe the latency, the coupling factor and the decay behavior respectively at subband index μ. The PSD of the feedback signal $\hat{S}_{rr}(\mu,k)$ is now estimated by the convolution of this model with the PSD of the loudspeaker signal $\hat{S}_{xx}(\mu,k)$ according to Eq. (10). The main advantage of the model based feedback suppression is that it is not subjected to any length limitation since the feedback PSD is estimated in a recursive manner [4]

$$\begin{aligned} \hat{S}_{rr,\mathrm{A}}(\mu,k) \;&=\; \sum_{i=P(\mu)}^{\infty} \hat{S}_{xx}(\mu,k-1)A(\mu)e^{-\gamma(\mu)(i\,-\,P(\mu))} \\ &=\; A(\mu)\hat{S}_{xx}\left(\mu,k-P(\mu)\right) \;+ \\ &\quad\; e^{-\gamma(\mu)}\hat{S}_{rr}(\mu,k-1). \end{aligned} \tag{10}$$

For the sake of simplification, the following is assumed:

$$P(\mu) \;\approx\; P \;=\; \left[T_D\;\frac{f_s}{L}\right] \quad \forall\;\mu \tag{11}$$

$$\gamma(\mu) \;\approx\; \gamma \;=\; \frac{2\;3\ln(10)\;L}{T_{60} f_s} \quad \forall\;\mu \tag{12}$$

The subband coupling factor is computed according to Eq. (13), where $|\bar{H}_{\mathrm{RIR}}(\mu,k\;=\;P)|$ is the subband impulse response at frame index P smoothed in the positive and negative frequency directions [4]. The simulated room impulse response is depicted in Fig. 4.

$$A(\mu) \;=\; |\bar{H}_{\mathrm{RIR}}(\mu,k\;=\;P)|^2 \tag{13}$$

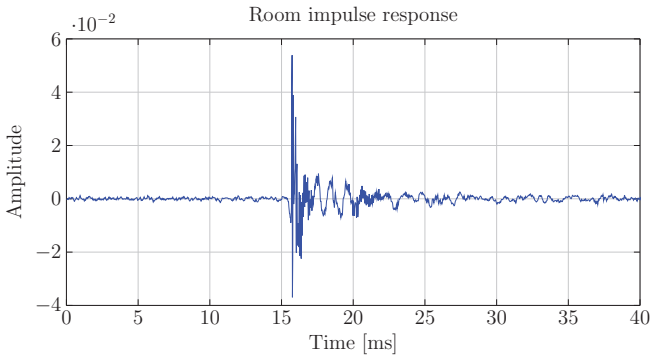

Figure 4: Simulated room impulse response

Wiener Filter

The Wiener solution in the subband domain $H_{\mathrm{w}}(\mu,k)$ is given by Eq. (14). In order to prevent musical noise, the estimated noise and feedback power spectral densities are multiplied with the overestimation factors K_b and K_r respectively [1]

$$H_{\mathrm{w}}(\mu,k) \;=\; 1 \;-\; \frac{K_b\cdot\hat{S}_{bb}(\mu,k) \;+\; K_r\cdot\hat{S}_{rr}(\mu,k)}{\hat{S}_{yy}(\mu,k)}. \tag{14}$$

The PSD of the noisy microphone input in the subband domain $\hat{S}_{yy}(\mu,k)$ is estimated according to [1]

$$\hat{S}_{yy}(\mu,k) \;=\; |Y(\mu,k)|^2. \tag{15}$$

The noisy microphone input spectrum is depicted in Fig. 5 and Fig. 6 shows the corresponding Wiener coefficients that are computed. From Fig. 5 and 6, it can be observed that the Wiener filter "opens" only in those subbands that correspond to speech. After filtering the noisy microphone signal with the Wiener filter, the estimated speech spectra $\hat{S}(\mu,k)$ in the subband domain is obtained according to Eq. (16)

$$Y(\mu,k)\cdot H_{\mathrm{w}}(\mu,k) \;=\; \hat{S}(\mu,k). \tag{16}$$

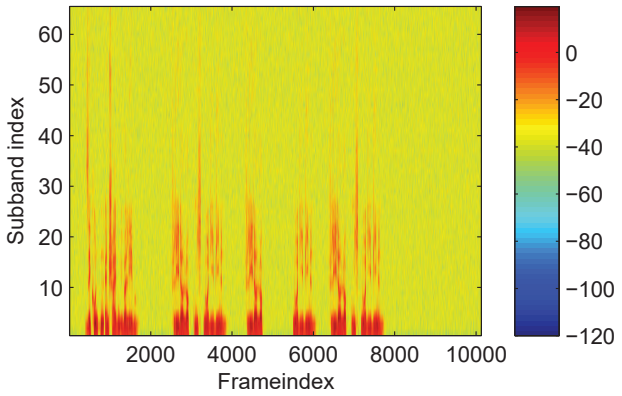

Figure 5: Spectrum of the noisy microphone input in dB

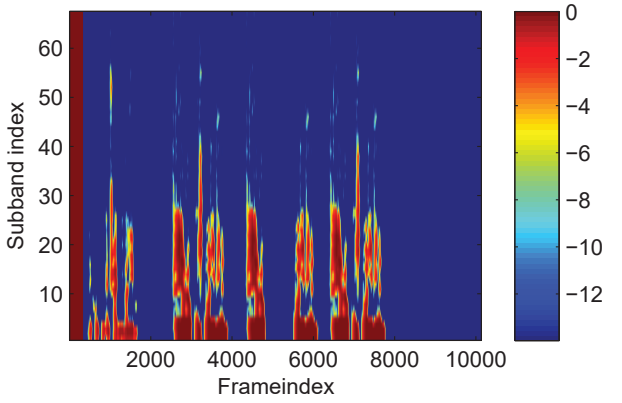

Figure 6: Wiener filter coefficients in dB

Loss Control Unit

Along with the model-based feedback suppression, the loss control unit is additionally implemented for the purpose of feedback suppression. The basic principle of the loss control unit is as follows: If voice activity is detected at one end, then an attenuation of 0 dB is inserted at the loudspeaker input of the inactive end while simultaneously inserting an attenuation of e.g. 45 dB at the loudspeaker input of the active end. During the "double talk" situation where both the sides are active simultaneously, an attenuation of e.g. 15 dB is inserted at the loudspeaker inputs of both the sides. The voice activity detection counter values and the corresponding computed attenuation factors is illustrated in Fig. 7 and Fig. 8 respectively.

Suppression of the Instationary Clanging Noise

Speech enhancement in a surgery room requires special focus on the suppression of sharp clanging noise which is caused when metallic surgical instruments are dropped. These clanging noises can be described as very short bursts of loud noise lasting for less than a second. Since the properties of these clanging noises are different from that of the stationary background noise, the Wiener filter is insufficient and a compressor module is required in

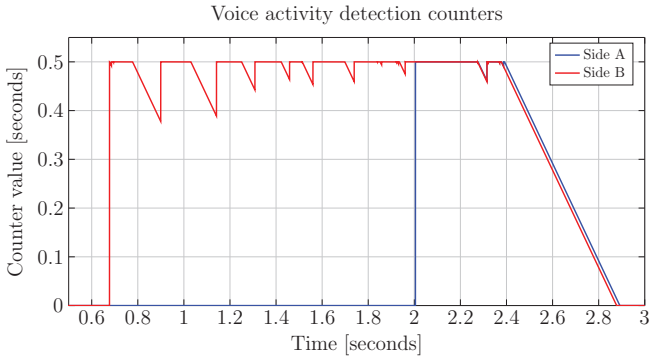

Figure 7: Voice activity detection time counters

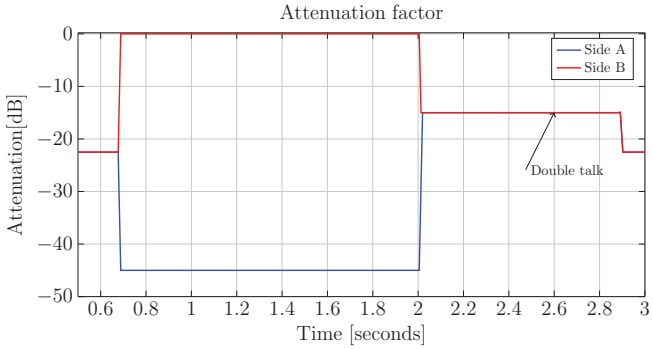

Figure 8: Attenuation inserted at the loudspeaker inputs

order to attenuate/suppress the peaks.

Conclusion and Outlook

In this paper, the design and implementation of a speech enhancement system specific to the requirements of a surgery room is discussed. Furthermore, with the real surgery room recordings obtained from Universitätsklinikum Schleswig-Holstein (UKSH), it was possible to simulate the surgery room environment and test the modules of speech enhancement system with real surgery room signals.

References

[1] Vasudev Kandade Rajan: Speech Enhancement in Hands-free Systems for Automobile Environments, Shaker Verlag, 2017.

[2] Eberhard Hänsler, Gerhard Schmidt: Acoustic Echo and Noise Control: A Practical Approach, Wiley-Interscience, 2004.

[3] P. P. Vaidyanathan: Multirate Systems and Filterbanks, Prentice Hall P R T, 1993.

[4] M. Gimm, P. Bulling, G. Schmidt: Energy-Decay Based Postfilter for ICC Systems with Feedback Cancellation.