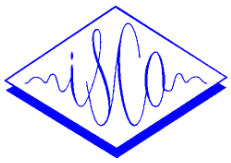

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

Design and Optimization of a Two Microphone Speech Enhancement System

Rainer Martin

Institute of Communication Systems and Data Processing (IND), Aachen University of Technology, Templergraben 55, 52056 Aachen, Germany, Phone: +49 241 806984, Fax: +49 241 8888186, e-mail: martin@ind.rwth-aachen.de

ABSTRACT

This contribution presents a novel two microphone speech enhancement system and its optimization using objective measures of speech quality. The speech enhancement system was designed for the voice communication system of a Computed Tomography system (CT scanner) which is used for the observation of the patient during CT examinations. The purpose of the system is to reduce noise and thus to enhance the speech signal of the patient which is transmitted to the control desk. The enhancement is based on the spatial coherence of noise and speech signals and is accomplished in two separate frequency bands by means of adaptive filters, scalar adaptive weighting and a highpass filter. For the optimization of the algorithms we used objective criteria which include the distortion of the clean speech signal and the reduction of noise during speech activity and during speech pause. The real time version of the speech enhancement system is implemented on a single Motorola DSP 56001.

1. INTRODUCTION

The voice communication between the patient in a Computed Tomography scanner (CT scanner) and the operator at the control desk is disturbed by acoustic noise which originates from the CT scanner. The noise is due to the rotating x-ray imaging system as well as to numerous cooling fans. To reduce the fatigue of the operator it is very desirable to reduce the level of noise transmitted from the gantry of the scanner to the control desk. It is also crucial that the speech enhancement algorithm maintains a high level of speech intelligibility and that the residual noise sounds very natural.

Because of these requirements and the fact that the noise is non-stationary a multi-microphone enhancement technique is most appropriate. Since it was required that the real time implementation of the system should not need more than one DSP a minimization of hardware and computational demands was necessary leading to the two microphone enhancement system described below. The system was designed and optimized for the CT application but may be used in other environments as well.

The disturbed speech signal is split into two frequency bands. In the band above 800 Hz we use an adaptive filter which attenuates incoherent signal components, i.e. noise. This filter is adapted using the Normalized Least Mean Square algorithm (NLMS) and a spectral smoothing technique. In the frequency band between 240 Hz and 800 Hz an adaptively controlled scalar factor attenuates noise only when the speaker is not active. The frequency band below 240 Hz is attenuated by 20 dB by means of a highpass filter. The enhancement algorithm will be discussed in detail in the next section.

The optimization of a speech enhancement system requires time consuming listening experiments. While the objective evaluation of speech quality has received considerable attention in the context of speech coding (e.g. [1]) the objective assessment of speech quality of a speech enhancement system is not yet well developed. We use an approach similar to the one proposed in [2] which evaluates the distortion of the clean speech signal and the attenuation of the noise signal during speech activity and during speech pause. We show that this method is very helpful to find the optimal parameter settings and to shortcut some of the listening experiments. Section 3 outlines the optimization of this system and some experimental results.

2. TWO MICROPHONE SPEECH ENHANCEMENT SYSTEM

The design of the speech enhancement system is based on measurements of the power spectral density and the coherence function of noise and speech. Figure 1.A shows the spectral density of the acoustic noise. The gantry of the CT scanner produces broadband noise peaking at frequencies below 1200 Hz. The (magnitude squared) coherence function of the noise signals is shown in Figure 1.B. There is only little correlation of noise signals above 500 Hz. Based on coherence measurements and listening tests the microphone distance was optimized. The microphones are mounted inside the gantry tunnel such that the patients head is close to the microphones.

The spectral density and the coherence suggest a speech enhancement scheme which processes the speech signal in two frequency bands: Above 800 Hz a linear phase adaptive filter is used which suppresses incoherent signal components (noise) and passes highly coherent speech signals. Processing the frequency band below 800 Hz with this adaptive filter would result in annoying fluctuating residual noise ("musical tones") caused by residual correlation of noise signals at these frequencies. The

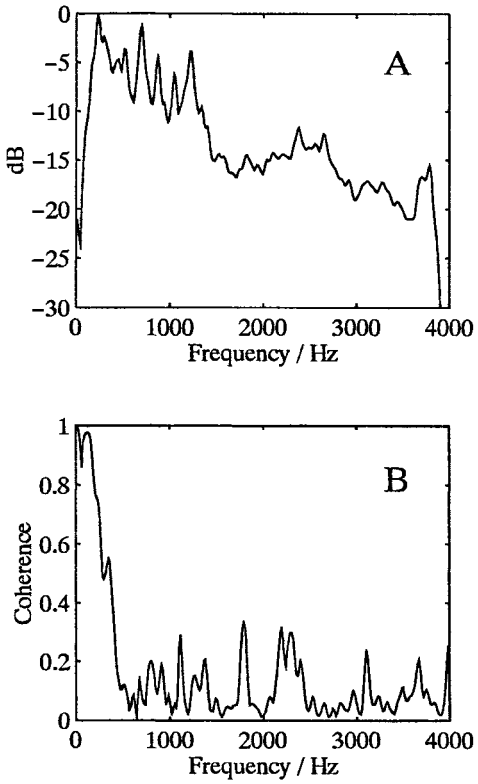

Figure 1: Long term power spectral density (A) and (magnitude squared) coherence (B) of acoustic noise of the CT scanner. The microphone distance is $d_{mic} = 0.4\ m$.

noise in the band between 240 and 800 Hz is therefore suppressed by an adaptive scalar factor. This factor is controlled by the speech activity of the person inside the gantry tunnel. The speech activity is determined by an SNR estimator which is also used to increase the robustness of the time delay estimation algorithm [3]. The frequency band below 240 Hz is attenuated by 20 dB by means of a second order recursive highpass filter.

Figure 2 shows a block diagram of the resulting algorithm. The main features of the speech enhancement system are

- 3–tap FIR preemphasis filters to whiten the speech and noise signals;
- Robust time delay estimation algorithm using cross-correlation and an SNR estimator as described in [3];
- Processing in two frequency bands:
  - 800–3600 Hz: adaptive FIR filtering using the NLMS algorithm and spectral smoothing;
  - 240–800 Hz: adaptive scalar weighting $b(k)$ based on the estimated SNR [3].
- Deemphasis filter to restore the spectral characteristics of the speech signal.

2.1. Processing in the 800–3600 Hz Band

The adaptive filters with the coefficient vectors $\mathbf{H_1}(k)$ and $\mathbf{H_2}(k)$ of order N_H (N_H even) are updated using a linear phase version of the NLMS algorithm (see Figure 2, $T_H = N_H/2$)

$$\mathbf{H_1}(k+1) = \mathbf{H_1}(k) + \alpha e_1(k) \frac{(\mathbf{I}+\mathbf{I}^R)\mathbf{Y}_{\mathrm{hppre1}}(k)}{\mathbf{Y}^T_{\mathrm{hppre1}}(k)\mathbf{Y}_{\mathrm{hppre1}}(k)}$$

$$\mathbf{H_2}(k+1) = \mathbf{H_2}(k) + \alpha e_2(k) \frac{(\mathbf{I}+\mathbf{I}^R)\mathbf{Y}_{\mathrm{hppre2}}(k)}{\mathbf{Y}^T_{\mathrm{hppre2}}(k)\mathbf{Y}_{\mathrm{hppre2}}(k)} \quad (1)$$

where $\mathbf{I}$ and $\mathbf{I}^R = \begin{pmatrix} 0 & 0 & 0 & .. & & 1 \\ 0 & 0 & .. & & .. & \\ 0 & .. & & 1 & & \\ .. & & 0 & & & .. \\ & 1 & & & .. & 0 \\ .. & & & .. & 0 & 0 \\ 1 & & .. & 0 & 0 & 0 \end{pmatrix}$ denote the identity matrix and a modified reflection matrix, respectively. The error signals $e_1(k)$ and $e_2(k)$ are given

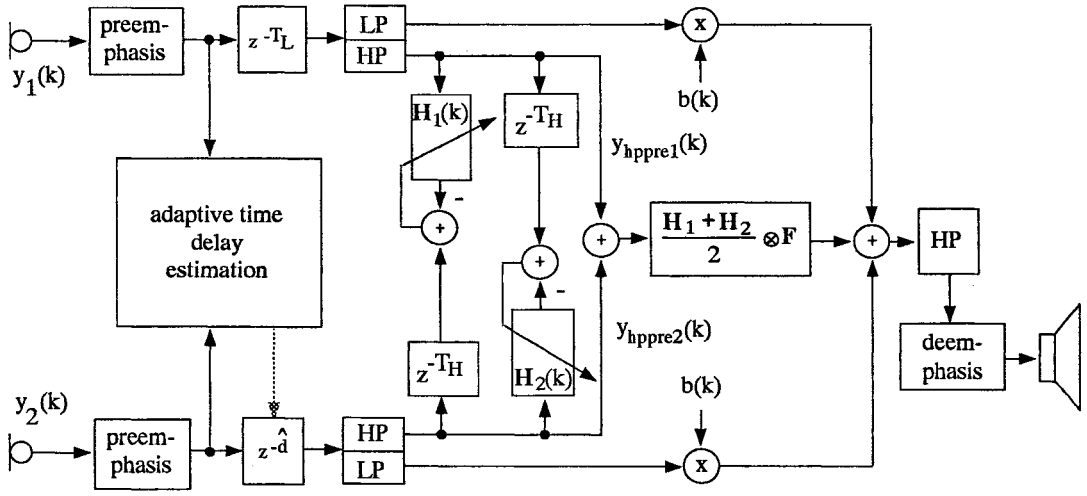

Figure 2: Block diagram of the two microphone speech enhancement system. LP: lowpass filter, HP: highpass filter.

by

$$e_1(k) = y_{hppre2}(k - T_H) - \mathbf{Y}^T_{\text{hppre1}}(k)\mathbf{H}_1(k)$$
$$e_2(k) = y_{hppre1}(k - T_H) - \mathbf{Y}^T_{\text{hppre2}}(k)\mathbf{H}_2(k) \quad (2)$$

and

$$\mathbf{Y}_{\text{hppre1}}(k) = (y_{hppre1}(k),\ \ldots, y_{hppre1}(k - N_H))^T$$
$$\mathbf{Y}_{\text{hppre2}}(k) = (y_{hppre2}(k),\ \ldots, y_{hppre2}(k - N_H))^T \quad (3)$$

denote the vectors of the preemphasis and highpass filtered input signals. To filter the sum of the highpass filtered input signals $y_{hppre1}(k)$ and $y_{hppre2}(k)$ we use the mean of the two coefficient vectors $\mathbf{H}_1(k)$ and $\mathbf{H}_2(k)$ and introduce an additional smoothing window $\mathbf{F}^T = (f_0, f_1, \ldots, f_{N_H})$

$$\mathbf{H}(k) = \frac{\mathbf{H}_1(k) + \mathbf{H}_2(k)}{2} \otimes \mathbf{F} \quad (4)$$

The symbol $\otimes$ denotes the multiplication of two vectors by elements. The window function is used to smooth the frequency response of the adaptive filter. A Kaiser window [4] with a shape parameter of $\beta_{Kaiser} \in [3,5]$ results in good speech quality and increased noise reduction. The linear phase of the adaptive filters with the coefficient vectors $\mathbf{H}_1(k)$ and $\mathbf{H}_2(k)$ has the advantage that when the vectors $\mathbf{H}_1(k)$ and $\mathbf{H}_2(k)$ are averaged (equ. 4) the amplitude spectra of these filters also average without additional errors due to different phase spectra.

2.2. Processing in the 240–800 Hz Band

The attenuation factor $b(k)$ ($b_{min} \leq b(k) \leq b_{max}$) is adjusted according to the estimated SNR. Whenever the estimated SNR is below a preselected threshold the attenuation is slowly and successively increased until a maximum attenuation of 40 dB (corresponding to $b_{min} = 0.01$) is reached. Whenever the estimated SNR is larger than the threshold the attenuation is rapidly decreased to a minimum value of 3 dB ($b_{max} = 0.5$). The SNR threshold is set to 3 dB.

The attenuation factor $b(k)$ is computed using the recursive system

$$b(k+1) = b(k)\beta_1 + b_{max}(1 - \beta_1),\ SNR > threshold$$
$$b(k+1) = b(k)\beta_2 + b_{min}(1 - \beta_2),\ SNR \leq threshold \quad (5)$$

The smoothing constants β_1 and β_2 are set to $\beta_1 = 0.9996$ and $\beta_2 = 0.99999$.

3. OPTIMIZATION USING OBJECTIVE CRITERIA

We assess the speech quality delivered by the speech enhancement system in terms of

- distortion of the speech signal;
- noise reduction during speech activity;
- noise reduction during speech pause.

These criteria can be measured during simulation of the speech enhancement system according to the scheme outlined in Figure 3. It is required however, that the speech signals are recorded seperately from the noise signals and that the processing is done by means of linear filters. For the purpose of measuring the above quantities the adaptive filter with coefficient vector $\mathbf{H}(k) = \frac{\mathbf{H}_1(k)+\mathbf{H}_2(k)}{2} \otimes \mathbf{F}$ is duplicated such that the undisturbed speech signal and the noise signal can be processed independently. The speech signal distortion can then be measured as the segmental SNR of the processed signal $\tilde{s}(k)$ with respect to the unprocessed delayed speech signal $s(k - T_H)$

$$SNR^s_{\tilde{s}-s}(m) = 10 \cdot \log_{10}\left(\frac{\sum_{i=mN}^{mN+N-1} s^2(i - T_H)}{\sum_{i=mN}^{mN+N-1} (\tilde{s}(i) - s(i - T_H))^2}\right) \quad (6)$$

$$SEGSNR^s_{\tilde{s}-s} = \frac{1}{K}\sum_{m=0}^{K-1} \max\left(SNR^s_{\tilde{s}-s}(m), 0\right) \quad (7)$$

To reduce the influence of speech pause we average only those speech signal frames which exhibit an SNR larger than 0 dB. Since we use a linear phase filter in our speech enhancement algorithm the segmental SNR is a measure for the amplitude distortion of the speech signal.

The attenuation of the noise signals during speech activity NR_{active} and during speech pause NR_{pause} is measured as the power ratio of the noise signal before and after the adaptive filter

$$NR_{active} = 10\log_{10}\left(\overline{P_n(k - T_H)}/\overline{P_{\tilde{n}}(k)}\right),\ \overline{P_s}(k) \neq 0 \quad (8)$$

$$NR_{pause} = 10\log_{10}\left(\overline{P_n(k - T_H)}/\overline{P_{\tilde{n}}(k)}\right),\ \overline{P_s}(k) = 0 \quad (9)$$

where $\overline{P_n(k)}$ and $\overline{P_{\tilde{n}}(k)}$ denote the average power of the unprocessed and the processed noise signal, respectively. $\overline{P_s}(k)$ denotes the short time power of the speech signal.

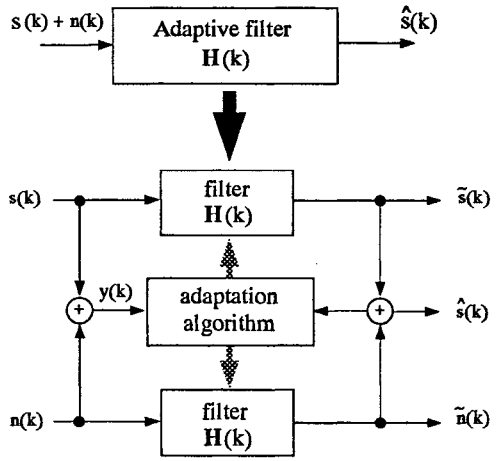

Figure 3: Signal flow for the evaluation of objective measures.

The plots in Figure 4 show the result of the objective evaluation as outlined above applied to the adaptive filter $\mathbf{H}(k)$. Figure 4.A plots the objective measures as a function of the input SNR. For this experiments adaptive filters of order 64 and a rectangular window $\mathbf{F}$ were used (equivalent to $\beta_{Kaiser} = 0$). It can be seen that the speech signal distortion increases as the input SNR decreases. The system should not be used at SNR below 0 dB. The noise reduction during speech pause is about 5–6 dB. As an example for the objective optimization procedure we plot the objective criteria versus the order (Figure 4.B) and the step size (Figure 4.C) of the adaptive filters $\mathbf{H}_1(k)$, $\mathbf{H}_2(k)$, and $\mathbf{H}(k)$. In accordance with extensive listening test it was found that a filter order of 64 and a step size of $\alpha = 0.1$ yields the best speech quality. A smaller step size reduces fluctuations of the residual noise but increases the perceived reverberation of the speech signal. It is also obvious from Figure 4 that a smoothing window significantly increases the noise reduction during speech pause without detrimental effects to the quality of the speech signal.

The system was implemented on a single 32 MHz DSP56001. It runs in real time at a sampling rate of 8 kHz with about 60% of the DSP's computing power. The remaining 40% of computing power are occupied by a real time operating system and other vital functions. The DSP implementation of the system, which includes the adaptive scalar weighting and highpass filters yields about 14 dB noise reduction during speech pause. This noise reduction system is currently used in the Siemens SOMATOM PLUS 4 CT scanner with excellent results.

4. CONCLUSIONS

The speech enhancement system presented in this contribution significantly reduces noise in the speech transmission system of a CT scanner. It was shown that speech enhancement systems can be optimized using appropriate objective criteria. If the speech enhancement system is realized by means of linear filters the optimization scheme as outlined in this paper is very practical. The amount of listening tests can be significantly reduced.

The system achieves a high intellegibility. The residual noise sounds very natural. All test persons agreed that the enhanced speech signal is preferable to the unprocessed noisy speech signal.

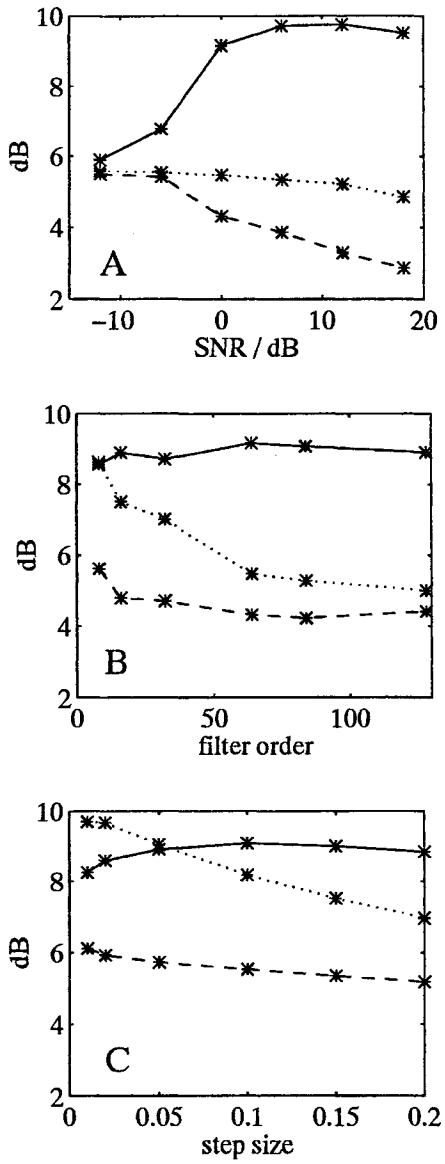

Figure 4: Distortion of speech signal and noise reduction during speech activity and during speech pause vs. input SNR (A), order (B), and step size (C) of adaptive filters.
A: $\beta_{Kaiser} = 0$, step size: $\alpha = 0.1$, filter order $N_H = 64$
B: $SEGSNR^s_n \approx 0$ dB, $\beta_{Kaiser} = 0$, step size: $\alpha = 0.1$
C: $SEGSNR^s_n \approx 0$ dB, $\beta_{Kaiser} = 3$, filter order $N_H = 64$
(——): Segmental SNR of speech;
(· · · · ·): noise attenuation during speech pause;
(– – –): noise reduction during speech activity.

ACKNOWLEDGMENTS

This research was supported by Siemens Medical Systems, Erlangen Germany. The author thanks Dr. G. Dehner and Prof. Dr. P. Vary for many stimulating discussions and Mr. M. Oberhoff for some of the DSP assembly language coding.

REFERENCES

[1] S. Quackenbush, T. Barnwell III, and M. Clements: "Objective Measures of Speech Quality", Prentice Hall, 1988.

[2] R. Le Bouquin, G. Faucon, and A. Akbari Azirani: "Proposal of a Composite Measure for the Evaluation of Noise Cancelling Methods in Speech Processing", Proc. EUROSPEECH '93, pp. 227-230, Berlin, September 1993.

[3] R. Martin: "An Efficient Algorithm to Estimate the Instantaneous SNR of Speech Signals", Proc. EUROSPEECH '93, pp. 1093-1096, Berlin, September 21-23, 1993.

[4] J. Kaiser: "Nonrecursive Digital Filter Design Using the I_o-sinh Window Function", Proc. IEEE Int. Symp. on Circuits and Syst., pp. 20-23, Apr. 22-25, 1974.