Improved Performance of Multimicrophone Speech Enhancement Systems

Zhen Yang, Klaus Uwe Simmer, Alexander Wasiljeff

University of Bremen, Department of Physics and Electrical Engineering,
P.O. Box 330 440, D-28334 Bremen, Germany

RÉSUMÉ

Nous traitons ici des systèmes adaptifs pour améliorer la qualité des signaux de paroles en utilisant un réseau de N microphones. Le système réalisé est une synthèse de deux méthodes proposées par Zelinski et Simmer. Nous présentons les resultats d' une étude menée dans notre laboratoire et discutons la qualité d'amélioration du rapport signal sur bruit.

ABSTRACT

In this paper, we analyse the performance of multimicrophone speech enhancement systems. We propose some modifications to Zelinski's method as well as an algorithm which synthesizes these modifications and Simmer's method by setting two thresholds on the basis of the input SNR in the frequency domain. Finally we present the results of our study of this new noise suppression system.

1. INTRODUCTION

In our paper we discuss the noise reduction in reverberant environments such as offices, conference rooms or classrooms. Microphone arrays have been proposed and constructed to enhance speech signals under noisy conditions [1] [4] [5].

To compensate for the movement of the signal source and the effect of multipath, time delay compensation methods are used [10] which synchronize the input signals to the microphones of the array. Thus the array is steered into the source direction. By coherent addition of the synchronized speech signals, the noise may be suppressed [2] [5]. Remaining noise can be reduced by a post-processing Wiener filter. A typical processing system is shown in Fig.1. Conventional adaptive filters (e.g. those presented by Frost, Duvall, Griffiths-Jim) may achieve 6-8 dB SNR improvement (Fig. 2). With Zelinski´s method [1] we can expect over 10 dB noise reduction when the input SNR is lower than 10 dB (Fig. 2) because of better estimation of the speech spectrum. Recently one of our authors, K.U. Simmer, has further reduced the noise incurred by higher frequency distortion [2], but this method unfortunately degrades the performance if the input SNR is lower than 0 dB. On the basis of the algorithms mentioned, we have analysed the Wiener-Hopf equation and have tried several possible new modifications.

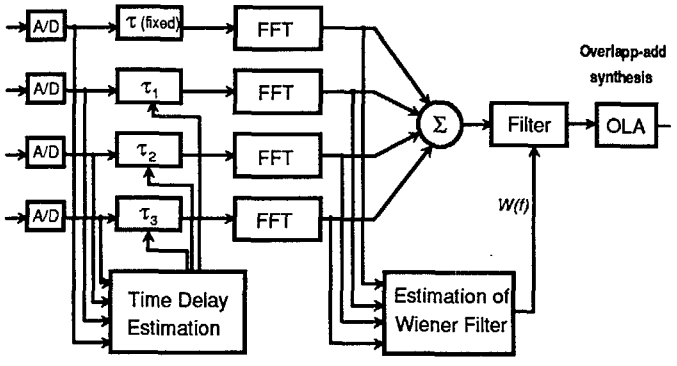

Fig. 1 Noise Suppression System.

2. ANALYSIS OF THE WIENER FILTER

The output of this noise suppression system in Fig. 1 is:

$$Y(k) = \overline{X}(k) W_{opt}(k) \qquad (1)$$

$\overline{X}(k)$ is the average input signal and $W_{opt}(k)$ is the transfer function of the Wiener filter. The difference between Zelinski´s and Simmer´s methods lies in the estimation of the autocorrelation spectrum of the input signal [1] [2]. We rewrite their transfer functions of the Wiener filter as follows:

$$W_{opt}(k) = \frac{P_s(k)}{\frac{1}{N}\sum_i |X_i(k)|^2} = \frac{C(k)}{A(k)} \quad \textbf{(Zelinski)} \qquad (2)$$

$$W_{opt}(k) = \frac{P_s(k)}{\left|\frac{1}{N}\sum_i X_i(k)\right|^2} = \frac{C(k)}{A'(k)} \quad \textbf{(Simmer)} \qquad (3)$$

$$C(k) = \alpha(k)\,\mathrm{Re}\sum_i \sum_{j=i+1} X_i(k) X_j^*(k) = \alpha(k) C'(k) \qquad (4)$$

Equation (2) considers the input of the beamformer whereas in (3) the output of the beamformer is used for the Wiener filter. $C(k)$ is the estimated autospectral density function of the input speech signal. The signal at the receiver of this Wiener filter is $\overline{X} = \sum X_i$. From the derivation of the optimal transfer function of this filter [2], the best Wiener filter is (3) rather than (2). This is true, as we can see from Fig. 2, when the input SNR is greater than 10 dB. For lower SNR, however, (2) is better than (3). This is related to the Wiener-Hopf equation:

$$\sum_{j \in I} w(j) R_{xx}(l-j) = R_{ss}(l), \; l \in I \qquad (5)$$

I is the number of coefficients $w(j)$.

Equation (5) is based on the assumption of uncorrelated speech and noise:

$$P_{sn}(k) = E\left[S(k)N^*(k)\right] = 0 \qquad (6)$$

Eq. (6) holds only in a statistical sense. As we use only four input signals to estimate this function, it is not allowed to assume that their estimated cross-spectrum is still zero if the SNR is low because the cross-spectrum is sufficiently large compared to the rather weak auto-spectrum of the speech. Under this condition the Wiener-Hopf equation should be modified as:

$$\sum_{j\in I} w(j)R_{xx}(l-j) = R_{ss(l)} + R_{sn}(l), l \in I \qquad (7)$$

R_{sn} is the cross-correlation of speech and noise. In other words the numerator *C(k)* of the optimal Wiener filter (3) should also be modified. Note that:

$$X_i(k)X_i^*(k) = (P_s(k) + P_{sn}(k)) + (P_{ns}(k) + P_n(k)) \qquad (8)$$

In low SNR situations, the sum of the third and fourth terms is generally greater than zero, therefore the first two terms (the right hand side of (7)) can not be greater than the product of $X_i(k)X_i^*(k)$. This is the motivation for changing the numerator of the transfer function. The denominator in (2) is greater than that in (3), it reduces the transfer function in low SNR case, thus it has better performance.

Another problem we are concerned with is how much SNR improvement we can expect with this kind of filter. The estimation error of the linear Wiener filter may be written as:

$$E_{error} = E\left[\left[s(m) - \sum_j w(j)\bar{x}(m-j)\right]^2\right] = E\left[e^2(m)\right] \qquad (9)$$

where s(m) is the input speech sample at the time point m, $\bar{x}(m)$ is the sum of speech and noise, *e(m)* is the error signal at the output of this system. The power density spectrum of *e(m)* is:

$$P_e(f) = P_s(f)(1 - W(f) - W^*(f) + |W(f)|^2) + P_{\bar{n}}(f)|W(f)|^2 \qquad (10)$$

$P_{\bar{n}}(f)$ is the average power density spectrum of input noise. Applying the Wiener-Hopf equation we get the minimum:

$$P_e(f) = \frac{P_s(f)P_{\bar{n}}(f)}{P_{\bar{x}}(f)} \qquad (11)$$

The minimum power of the error signal e(m) is

$$E_{\min} = \int_0^{f_o} \frac{P_s(f)P_{\bar{n}}(f)}{P_{\bar{x}}(f)}\,df = \int_0^{f_o} \frac{P_s(f)P_{\bar{n}}(f)}{P_s(f) + P_{\bar{n}}(f)}\,df \qquad (12)$$

f_0 is the stop-frequency of this system, it equals 8KHz in our experiments. Now we discuss the minimum error power in three different situations and assume that the input noise is wide band noise.

A. the input SNR is very low

$$E_{\min} = \int_0^{f_o} \frac{P_s(f)}{1 + P_s(f)/P_n(f)}\,df \qquad (13)$$

because $P_s(f)/P_n(f) << 1$. (Fig. 3 shows the spectrum of -20 dB input SNR, obviously this hypothesis holds.)

$$E_{\min} \approx \int_0^{f_o} P_s(f)df = p_s \qquad (14)$$

p_s is the power of the speech signal. This means the minimum estimation error tends to the input speech signal under the low SNR condition. The noise reduction in this system is

$$NR(dB) = 10LOG(N_i) - 10LOG(N_o) = -10LOG(\frac{p_s}{N_i}) \qquad (15)$$

N_i is the input noise power and N_o is the output noise power.

B. the input SNR is high

In this situation, we cannot assume $P_s(f)/P_n(f) << 1$ because the energy of speech is concentrated mainly in the low frequency regions. We take input SNR=20 dB as example (Fig. 4). The power spectral-density is composed of two different frequency ranges:

$f < 1KHz$ $\quad P_x(f) \approx P_s(f) \rightarrow P_n(f)/P_s(f) << 1$

$1KHz < f < 8KHz$ $\quad P_x(f) = P_s(f) + P_n(f)$

(notice that the unit of the y-axis in this figure is dB). We thus have:

$$E_{\min} = \int_0^{1KHz} P_{\bar{n}}(f)df + \int_{1KHz}^{8KHz} \frac{P_s(f)P_{\bar{n}}(f)}{P_s(f) + P_{\bar{n}}(f)}\,df \qquad (16)$$

but we have:

$$N_i = \int_0^{8KHz} P_{\bar{n}}(f)df = \int_0^{1KHz} P_{\bar{n}}(f)df + \int_{1KHz}^{8KHz} P_{\bar{n}}(f)df \qquad (17)$$

It can be seen that the noise which lies in the frequency range 0 to 1KHz can hardly be suppressed. This is the inherent disadvantage of this linear system. The noise reduction occurs mainly in the frequency range *f*>1KHz, where the speech component is small. From (16):

$$E_{\min} \geq \int_0^{1KHz} P_{\bar{n}}(f)\,df \qquad (18)$$

and we assume the power spectrum of noise is a constant C/KHz, then $N_i = 10\log(C \cdot 8KHz)dB$ and $N_o \geq 10\log(C \cdot 1KHz)dB$. The difference between $N_i$and N_o is the noise reduction, so we can expect NR $\leq$10log8-10log1 = 9.03 dB noise reduction.

C. in general

$$NR = 10LOG(N_i) - 10LOG(E_{\min}) = 10LOG\frac{N_i}{\int_0^{f_0} \frac{P_{\bar{n}}(f)}{1 + P_{\bar{n}}(f)/P_s(f)}\,df} \qquad (19)$$

With the increase of input SNR, $P_s(f)$ will increase or $P_n(f)$ decrease. Both will make the NR decrease.

We concluded that: 1. The maximum noise reduction equals the negative input SNR, if the SNR is very low. 2. If the SNR is very high, there is almost no noise reduction in low frequency region. The improvement depends on the frequency spectrum of the speech. With SNR=20 dB, maximum NR=9.03 dB.

3. Generally NR is between the case A and B and the greater the input SNR, the smaller the NR.

3. PERFORMANCE IMPROVEMENT OF THE WIENER FILTER

Zelinski uses (2) to estimate the Wiener filter, where C(k) consists of the average of the cross-spectra C'(k) (4) and a weight factor $\alpha(k)$ which depends on the $P_s(k)$ and $P_n(k)$ (the estimation of noise which is composed of the negative items of C'(k)) [1]. He has in fact only set one limit to reduce the effect of noise in the procedure to estimate *C(k)*, i.e. if C'(k)<0, set it to zero. The $P_n(k)$ in $\alpha(k)$ also consists of the previous negative items of C'(k). In experiments, we have found that C'(k) is often greater than $X_i(k)X_i^*(k)$. If this occurs, we reduce C'(k) to approximate (7). Our method is as follows:

1. calculate C'(k) (4) 2. calculate $Y_i(k) = X_i(k)X_i^*(k)$ i=1,2...N and compare $Y_i(k)$ with C'(k). If the latter is bigger, change it. 3. add $Y_i(k)$ to find A(k), 4. calculate $W_{opt}(k)$. We do not need $\alpha(k)$ as in Zelinski's method and have tried three modifications to *C(k)*

$$a.)\ C_a(k) = X_i(k)X_i^*(k) \qquad (20)$$

$$b.)\ C_b(k) = C'(k) - \frac{1}{N}\left[C'(k) - X_i(k)X_i^*(k)\right] \qquad (21)$$

$$c.)\ C_c(k) = C'(k) - \frac{1}{2}\left[C'(k) - X_i(k)X_i^*(k)\right] \qquad (22)$$

The modification b. is explained as follows: Since the noise at the ith microphone has the largest projection on the speech signal, i.e. $N_i(k)$ is comparable to $S(k)$, we subtract its effect on C'(k). In the modification equation (22) we consider that every $X_i(k)$ occurs just half the total number of times in the calculation of C'(k) and the other $X_j(k)$ also contain the same $S(k)$ component. The noise reduction achieved can be seen in Fig. 5. All modifications improve the performance in low SNR cases, but they degrade the result in high SNR cases. As the third term in (8) plays an important role for high SNR the term $\mathrm{Re}\left[P_{sn}(k)\right] + P_n(k)$ may be smaller than zero.

We have seen that with our modifications, there is an improvement in noise reduction for low SNR cases and we recall that we can get good results by using Simmer's method in high SNR cases. Therefore we have developed a synthesis method to utilize their different characteristics. We think that the energy of speech concentrates in the low frequency region, mainly [0, 2KHz]. Our new method is as follows:

1. calculate C(k) (just as defined by Zelinski, but without $\alpha(k)$)
2. let $C = \sum_{k=0}^{N1} C(k)$ N1 corresponds to 2KHz
3. let $A = \sum_{K=0}^{N} \left|\overline{X}(k)\right|^2$ N corresponds to 8KHz
4. let B=*C/A*
5. if B < η_1
   use the modified algorithm a to change C(k)
   if $\eta_1 < B < \eta_2$
   use the modified algorithm b to change C(k)
   else
   keep C(k)
6. if $B < \eta_2$

   calculate $A(k) = \frac{1}{N}\sum_i \left|X_i(k)\right|^2$

   else

   calculate $A(k) = \left|\frac{1}{N}\sum_i X_i(k)\right|^2$
7. find $W_{opt}(k)$ and $\hat{S}(k)$

Fig. 5 shows its noise reduction performance. It is better than the performance of Zelinski's algorithm regardless whether the input SNR is high or low, and has approached the performance limit of linear estimation (just as discussed in section 3.). This noise reduction curve can be adjusted if we change the thresholds η_1 and η_2. In Fig. 5 η_1=0.05,η_2=0.5. If we increase η_1 then the left side of this curve moves up a little, but the right side moves down a little. We have also observed in experiments that η_1 plays a more important role than η_2.

Fig. 6 shows the result of a speaker independent DTW-based isolated word recognizer before and after the proposed speech enhancement system. The vocabulary of the recognizer consists of 10 digits and 30 computer command terms. Speakers were at a fixed position.

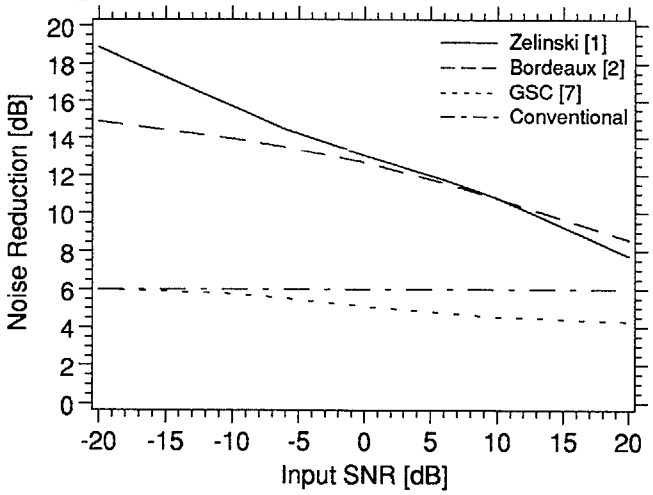

Fig. 2 Noise Reduction in Office Room.

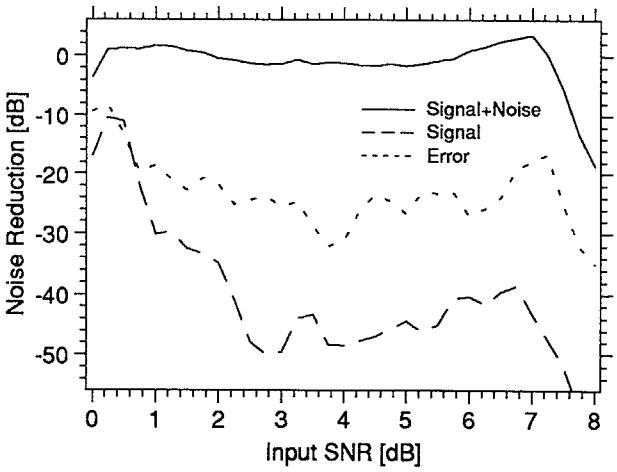

Fig. 3 Power Spectra of Speech, Noise and Error of Wiener Filter.

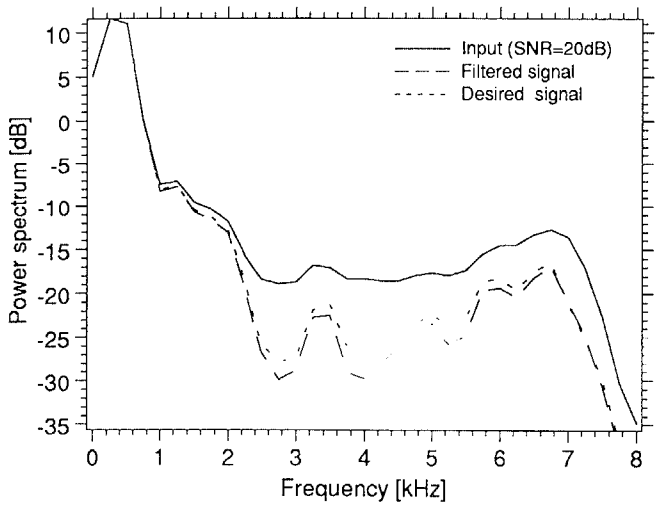

Fig. 4 Power Spectra of Input, Filtered and Desired Signal.

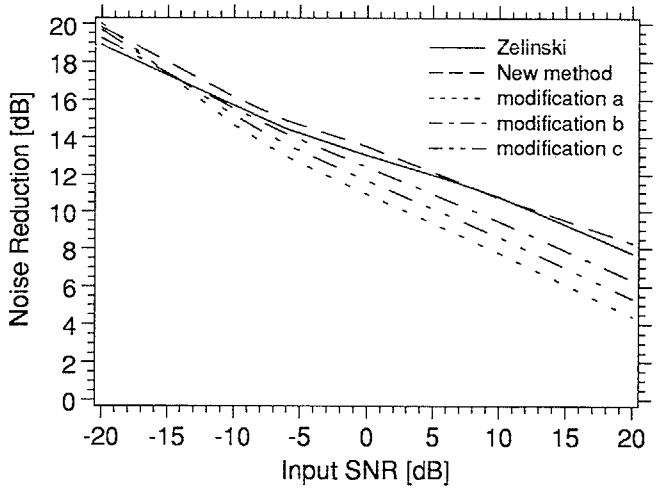

Fig. 5 Noise Reduction of Proposed Modifications.

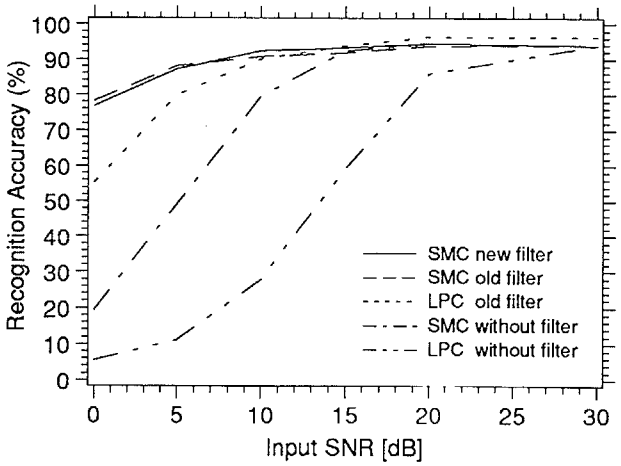

Fig. 6 Speech Recognition Accuracy.

White noise has been recorded in an office room with four microphones. This recorded noise and undisturbed speech have been added in the computer to simulate various SNRs. 8th order SMC and LPC-cepstrum coefficients have been computed without preemphasis using the old [10] and the new speech enhancement systems. Fig. 6 shows the good performance of the short-time modified coherence representation (SMC) [11] combined with the proposed speech enhancement system.

4. CONCLUSIONS

There is a performance limit to suppressing noise in microphone array systems by using Wiener linear filter. This means that we can only reduce the noise level by no more than the absolute value of the input SNR when this SNR is very low. Inversely, when it is high, our achievement depends on the distribution of the frequency spectrum of the speech. In our example it is no more than 9.03dB. Generally the noise reduction is between these two extreme values, the higher the input SNR, the smaller the noise reduction we can expect. We can improve the performance of this estimation system in low SNR cases by modifying the numerator of the Wiener filter. This means we have deleted the effect of the bigger noise components in the input signal. A new synthesis method seems to give the best estimate of the input speech in linear estimation situations.

REFERENCES

[1] R. Zelinski "A Microphone Array with Adaptive Post-Filtering for Noise Reduction in Reverberant Rooms", ICASSP-88, New York, pp. 2578-2581, 1988.

[2] K. U. Simmer, A. Wasiljeff "Adaptive Microphone Arrays for Noise Suppression in the Frequency Domain", Second Cost 229 Workshop on Adaptive Algorithms in Communications, Bordeaux, France, 1992.

[3] J. B. Allen et al. "Multimicrophone signal-processing technique to remove room reverberation from speech signals" J.Acoust. Soc. America, Vol.62, No.4, pp. 912-915, 1977.

[4] Y. Kaneda, M. Tohyama "Noise Suppression Signal Processing Using 2-Point Received Signals " Electronics and Communications in Japan, Vol.67-A, No.12, pp. 19-28, Apr. 1984.

[5] J. L. Flanagan et al. "Computer-steered Microphone Arrays for Sound Transduction in Large Rooms", J. Acoust. Soc. America, Vol. 78, No. 5, pp.1508-1518, 1985.

[6] K. M. Duvall "Signal Cancellation in Adaptive Antennas: The Phenomenon and Remedy" Ph. D. Thesis, Stanford University, California, U.S.A., Aug. 1983.

[7] L. J.Griffiths and C.W.Jim "An alternative approach to linearly constrained adaptive beamforming "IEEE Trans.on Antennas, Vol.30, No.1, pp. 27-34, 1982.

[8] B. Widrow et al. "Signal Cancellation Phenomena in Adaptive Antennas: Causes and Cures" IEEE Trans. on Antennas, Vol.30 p. 469, 1982.

[9] K. U. Simmer, A.Wasiljeff "Analysis and comparison of systems for adaptive array processing of speech signals in a noisy environment" Treizième Colloque GRETSI, Juan-Les-Pins, pp.529-532, 1991.

[10] K. U. Simmer, P. Kuczynski, A.Wasiljeff "Time delay compensation for adaptive multichannel speech enhancement systems" ISSSE-92, Paris, pp. 660-663, Sept. 1992.

[11] D. Mansour, B. H. Juang, "The short-time modified coherence representation and noisy speech recognition", IEEE, ASSP - 37, no. 6, pp 795 - 804, 1989.

[12] J. Hinrichs, "Adaptive Kurzzeitkohärenzanalyse und konfidenzabhängige Wortrückweisung zur automatischen Einzelworterkennung bei gestörten Sprachsignalen.", Ph. D. Thesis, University of Bremen, 1992.

[13] Y. Grenier, "A Microphone array for Car Environments", Speech Communications, Vol. 12, No. 1, pp. 25-39, 1993.