SPEECH ENHANCEMENT BY MULTI-CHANNEL CROSSTALK RESISTANT ADAPTIVE NOISE CANCELLATION

Qingning Zeng *Waleed H. Abdulla*

Department of Electrical and Computer Engineering, Faculty of Engineering, University of Auckland, Auckland, New Zealand

ABSTRACT

A novel Multi-channel Crosstalk Resistant Adaptive Noise Cancellation (MCRANC) algorithm is presented in this paper to enhance noise carrying speech signals. The algorithm would permit locating the microphones in close proximity as it cancels out the crosstalk effect. Results have indicated that this method outperforms the commonly used techniques in the sense of SNR improvement and speech intelligibility. A SNR improvement of 17.8dB using MCRANC keeping highly intelligible speech was achieved in our experiments versus 9.1dB using Multi-channel ANC (MANC) with far less speech quality.

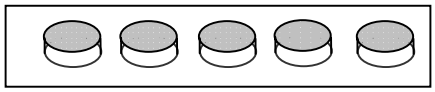
1. INTRODUCTION

In many applications such as in mobile and hands-free phones, speech enhancement systems are expected to be small in size [1][2]. This implies that the distances between the employed microphones should be very small. However, microphones located in close proximity undergo serious crosstalk effect. This effect would violate the operating conditions of Multi-channel ANC (MANC) method [3][4], which is one of the prominent methods for noise cancellation. Though two-channel Crosstalk Resistant ANC (CRANC) methods have been investigated in [5][6][7][8], they appear to be unstable and computationally expensive. Thus, a new Multi-channel Crosstalk Resistant ANC (MCRANC) is proposed in this paper. It does not only extend the two-channel CRANC method to multi-channel signal processing, but also has very good stability and employs only two adaptive FIR filters while the two-channel CRANC method in [8] needs three filters. Experiments show that MCRANC outperforms the CRANC technique since it can make the speech enhancement system more efficient in noise reduction and smaller in size.

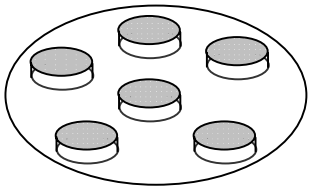
2. SYSTEM DESCRIPTION

2.1 Microphone array

We assume that N+1 microphones are closely placed. These microphones form an array. The array might be any formation such as Uniform Linear Array (ULA), planar array or solid array. We have no strict limitations for the position of microphones. Fig. 1 shows some of the commonly used arrays.



(a)



(b)

Fig.1 Two possible layouts of microphone arrays: (a) ULA (b) planar array

2.2 Crosstalk formulation

Suppose a speech signal $s(k)$ and noise $n(k)$ are generated by independent sources. As shown in figure 2 they arrive at microphone M_i through multi-paths and acquired as $s_i(k)$ and $n_i(k)$. The impulse responses of the intermediate media between the speech and noise sources and the acquiring microphone M_i are $h_{si}(k)$ and $h_{ni}(k)$ respectively. The audio signal acquired by microphone M_i can be represented by $x_i(k) = s_i(k) + n_i(k)$, where $i = 0,1,\cdots,N$, $k = 0,1,2,\cdots$, N+1 is the number of microphones employed in the speech enhancement system, and k is the discrete time index. Since the actual signal from every microphone contains noise signal and speech signal, it is called the crosstalk of the noise and speech.

Let us consider $x_0(k)$ as the main channel signal acquired by microphone M_0 and $x_i(k)$ ($i = 1,\cdots,N$) as the

referential signals acquired by the other N microphones. Assume that the main-channel signal is correlated with the referential-channel signals which is a valid assumption as the microphones are located in close proximity. Since the referential signals contain both speech and noise, common Adaptive Noise Cancellation (ANC) [9] and Multi-channel ANC (MANC) methods will not be proper methods for speech enhancement.

From figure 2 we have

$$x_i(k) = s_i(k) + n_i(k) \tag{1}$$

$$s_i(k) = h_{si}(k) * s(k) \tag{2}$$

$$n_i(k) = h_{ni}(k) * n(k) \qquad i = 0,1,\cdots,N \tag{3}$$

where $*$ is the convolution sign, $h_{si}(k)$ and $h_{ni}(k)$ is the time domain impulse response correspondence of the z-domain response $H_{si}(z)$ and $H_{ni}(z)$.

Let the impulse response of the intermediate environment between the input signal s_i and the output signal s_j be $h_{s_j s_i}(k)$, and the impulse response of the environment carrying the source n_i transferring to n_j be $h_{n_j n_i}(k)$, then

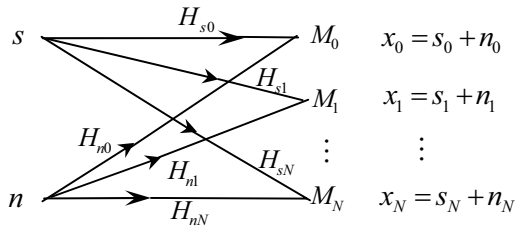


Fig.2 Speech and noise propagation between the emitting sources and the acquiring microphones.

$$s_j(k) = h_{s_j s_i}(k) * s_i(k) \tag{4}$$

$$n_j(k) = h_{n_j n_i}(k) * n_i(k) \qquad i, j = 0,1,\cdots,N \tag{5}$$

Through (2)-(5)

$$H_{s_j s_i}(z) = \frac{H_{sj}(z)}{H_{si}(z)} \tag{6}$$

$$H_{n_j n_i}(z) = \frac{H_{nj}(z)}{H_{ni}(z)} \qquad i, j = 0,1,\cdots,N \tag{7}$$

2.3 Environment scenario

In the practical environment shown in Fig. 3, noise emitted from a certain source may propagate to microphone M_i through multiple paths including direct, reflection, and refraction paths. The noise may also be emitted from multiple sources. But we may consider those noises are from a combined source and all propagation paths are included in the combined transfer function $H_{ni}(z)$ or the impulse response $h_{ni}(k)$.

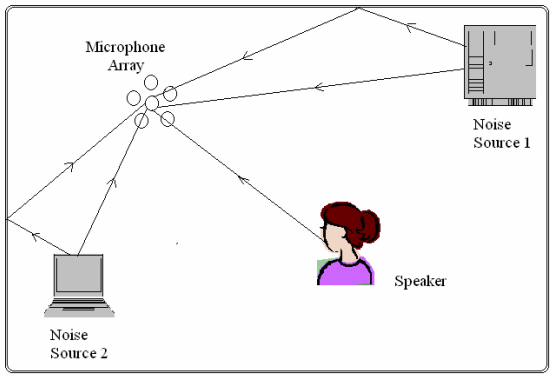


Fig.3 Noisy speech environment

3. MCRANC algorithm

Figure 4 shows the proposed scheme of the speech enhancement system. It employs a Voice Activity Detector (VAD) and two digital filter A and B.

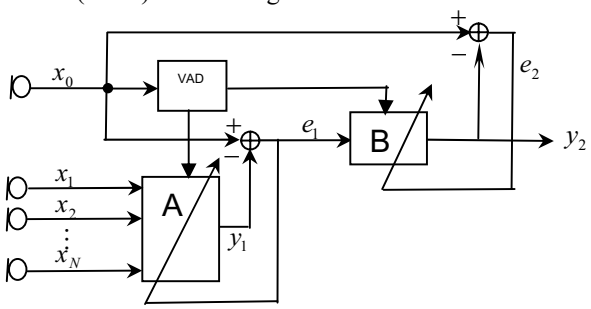


Fig. 4 Speech enhancement system

During the Non-Voice Periods (NVP) of the utterances, we may assume $s_0(k) = 0$ and $s_i(k) = 0$. Thus, the referential signals are used to cancel the main signal. In this case we have

$$x_0 = y_1 + e_1 \tag{8}$$

That is

$$n_0(k) = w\bar{n}(k) + err(k) \tag{9}$$

where $x_0 = n_0(k)$, $e_1 = err(k)$ is the prediction error, $y_1 = w\bar{n}(k)$ is the output of filter A. w is the weight vector of the FIR filter A, i.e.

$$w = (w_1, w_2, \cdots, w_N) \tag{10}$$

where $w_i = (w_{i0}, w_{i1}, \cdots, w_{iL})$, $\bar{n}(k)$ is the vector of noise signal

$$\bar{n}(k) = [\bar{n}_1(k), \bar{n}_2(k), \cdots, \bar{n}_N(k)]^T \tag{11}$$

where $\bar{n}_i(k) = [n_i(k), n_i(k-1), \cdots, n_i(k-L)]^T$.

Let the minimal error power be denoted by $P[err^0(k)]$ and the corresponding optimal weight vector by

$$w^0 = (w_1^0, w_2^0, \cdots, w_N^0)$$

$$= (w_{10}^0, w_{11}^0, \cdots, w_{1L}^0, \cdots\cdots, w_{N0}^0, w_{N1}^0, \cdots, w_{NL}^0) \quad (12)$$

We only need to adjust the weights of filter A to minimize the square sum of e_1 in Fig. 4 to obtain w^0 and $P[err^0(k)]$. Theoretically $P[err^0(k)]$ is inversely proportional to the number of the referential channels used.

During the Voice Period (VP) which just follows the previous NVP, we may assume the environment remains unchanged and accordingly we may keep the optimal weights of filter A unchanged, thus

$$\begin{aligned} y_1(k) &= w^0 \bar{x} = w^0 (\bar{s} + \bar{n}) \\ &= w^0 \bar{s} + [n_0(k) - err^0(k)] \end{aligned} \quad (13)$$

where $\bar{x}$ and $\bar{s}$ represent the acquired speech plus noise and the pure speech respectively, and may be expressed in a similar way to $\bar{n}$ in equation (11). Then from equations (8) and (13)

$$\begin{aligned} e_1(k) &= x_0(k) - y_1(k) = [s_0(k) + n_0(k)] - y_1(k) \\ &= [s_0(k) + n_0(k)] - [w^0 \bar{s} + n_0(k) - err^0(k)] \\ &= s_0(k) - w^0 \bar{s} + err^0(k) \\ &= p(k) + err^0(k) \end{aligned} \quad (14)$$

where

$$p(k) = s_0(k) - w^0 \bar{s} \quad (15)$$

Take the z-transform of (14) and (15) to get

$$E_1(z) = P(z) + Err^0(z) \quad (16)$$

$$\begin{aligned} P(z) &= S_0(z) - Z[\sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 s_i(k-j)] \\ &= S_0(z) - Z[\sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 h_{s_i s_0}(k-j) * s_0(k-j)] \\ &= [1 - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 z^{-2j} H_{s_i s_0}(z)] S_0(z) \\ &= \tilde{H}(z) S_0(z) \end{aligned} \quad (17)$$

where

$$\tilde{H}(z) = 1 - \sum_{i=1}^{N}\sum_{j=0}^{L} w_{ij}^0 z^{-2j} H_{s_i s_0}(z) \quad (18)$$

If the system function of filter B is $\tilde{H}^{-1}(z) = [\tilde{H}(z)]^{-1}$, then by using (17) and (18) we get

$$\begin{aligned} Y_2(z) &= \tilde{H}^{-1}(z) E_1(z) = \tilde{H}^{-1}(z)[\tilde{H}(z) S_0(z) + Err^0(z)] \\ &= S_0(z) + \tilde{H}^{-1}(z) Err^0(z) \end{aligned} \quad (19)$$

Thus

$$y_2(k) = s_0(k) + \tilde{h}^{-1}(k) * err^0(k) \quad (20)$$

where $\tilde{h}^{-1}(k)$ is the inverse z-transform of $\tilde{H}^{-1}(z)$ and $*$ is the convolution symbol.

As commonly assumed in ANC the noise n_0 is uncorrelated with the speech signal s_0 and the mean value of n_0 is zero [9]. Then in order that the system transfer function of filter B approximates $\tilde{H}^{-1}(z)$, we need only to adjust the coefficients of filter B to minimize the square sum of e_2 . This is because

$$\begin{aligned} \|e_2(k)\|^2 &= \|x_0(k) - y_2(k)\|^2 = \|s_0(k) + n_0(k) - y_2(k)\|^2 \\ &= \|n_0(k)\|^2 + \|s_0(k) - y_2(k)\|^2 \\ &\quad + 2n_0(k)[s_0(k) - y_2(k)] \end{aligned} \quad (21)$$

and

$$E(e_2^2) = E(n_0^2) + E[(s_0 - y_2)^2] \quad (22)$$

From (22), we may conclude that to minimize $E(e_2^2)$ we need to minimize $E[(s_0 - y_2)^2]$ which implies minimizing the error between y_2 and s_0. The input e_1 of filter B is mainly related to s_0 , as indicated in equations (14)~(15), and by assuming that the main signal and the referential signals are correlated, it is feasible to minimize the error between y_2 and s_0.

To adapt the system to the dynamical changes in the environment, the weights of filter A during the NVP periods and the weights of filter B during the VP periods should be retrained. If the environment including temperature, space, noise sources and speech source does not change, we do not need to retrain the optimal weights of filter A and B. This also implies that we may use a common voice activity detector for our MCRANC system.

4. EXPERIMENT

Our experiments were carried out in a common research room. Four microphones $M_0, M_1, \cdots, M_3$ were closely placed. The maximum distance between any two microphones was 2cm. The noise was generated from an improperly tuned radio situated at one meter from the microphones. The speech was from a male at 0.5 meter from the microphones. The sampling rate was 8K Hz. Normalized Least Mean Square (NLMS) algorithm was employed to adjust the weights of FIR filters A and B.

Fig.5 (a) shows the noisy speech signal x_0 acquired by the main microphone. The signals acquired by the referential microphones are almost the same as x_0 . Fig.5 (b) depicts the enhanced speech signal by ordinary MANC method. The SNR improvement is 9.1 dB but the speech is seriously damaged. In Fig.5 (c) the enhanced speech by two-channel CRANC method is shown with SNR improvement of 8.6 dB [8]. While Fig.5(d) illustrates the enhanced speech by the proposed MCRANC, which has achieved a SNR improvement of 17.8 dB. Both enhanced speech signal in (c) and (d) have high speech quality.

Fig.6 shows zoomed sections of non-speech parts from Fig.5. From this figure we may see that MANC and MCRANC have high noise cancellation abilities.

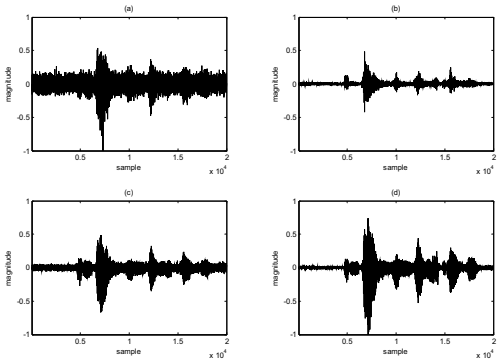


Fig.5 Experiment results
(a) Noisy speech signal
(b) Enhance speech by common MANC
(c) Enhanced speech by 2-channel CRANC
(d) Enhanced speech by proposed MCRAN

Fig.7 also illustrates zoomed sections but of speech parts from Fig.5. From that figure we may notice that MANC has greatly damaged the speech signal while both CRANC and MCRANC have almost preserved the original clean speech signal. However, MCRANC outperforms CRANC for MCRANC contains less noise than CRANC.

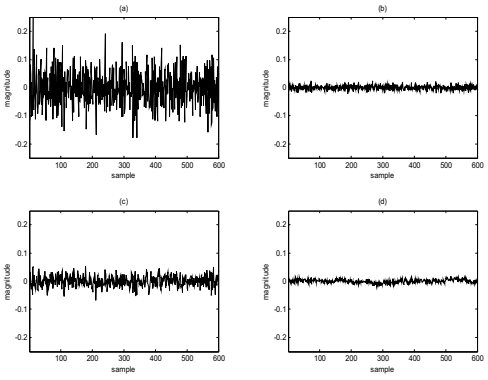


Fig.6 A section of Fig.5 (non-speech section)

5. CONCLUSION

A novel algorithm named MCRANC to enhance noise carrying speech signals is mathematically formulated and evaluated in this paper. It can make the speech enhancement system more efficient in noise reduction and smaller in size. A better speech quality with better SNR improvement compared to its rival techniques has been achieved with the proposed technique.

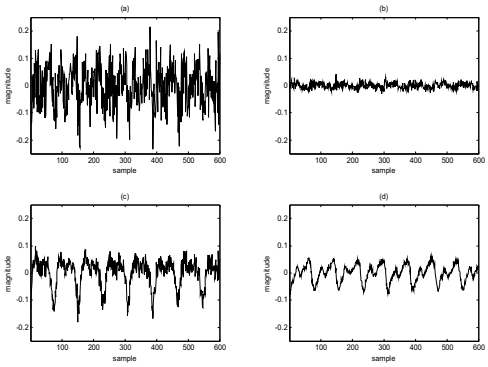


Fig.7 A section of Fig.5 (speech section)

6. REFERENCES

[1] R. L. Bouquin, "Enhancement of Noise Speech Signals: Application to Mobile Radio Communications", *Speech Communication*, (18), pp. 3-19, 1996

[2] R. Martin, "Small Microphone Arrays with Postfilters for Noise and Acoustic Echo Reduction", *Microphone Arrays*, M. Brandstein and D. Ward (eds.), Springer-Verlag, pp.255-276, 2001

[3] A. Hussain, "Multi-sensor Adaptive Speech Enhancement Using Diverse Sub-band Processing", *International Journal of Robotics & Automation*, 15, (2), pp.78-84, 2000

[4] M. Dahl, I. Claesson, S. Nordebo, "Simultaneous Echo Cancellation and Car Noise Suppression Employing a Microphone Array", *ICASSP-97*, Munich, Germany, pp. 239-242, 1997

[5] R. L. Zinser, G. Mirchandani, J. B. Evans, "Some Experimental and Theoretical Results Using a New Adaptive Filter Structure for Noise Cancellation in the Presence of Cross-talk", *Proc. ICASSP*, Tampa, 3, pp. 1253-1256, 1985

[6] G. Mirchandani, R. L. Zinser and J. B. Evans, "A new Adaptive Noise Cancellation Scheme in the Presence of Crosstalk", *IEEE Transaction on Circuits and System*, 39, (10), pp. 681-694, 1992

[7] S. M. Kuo, W. M. Peng, "Asymmetric Crosstalk-Resistant Adaptive Noise Canceler", *Proc. IEEE workshop on Signal Processing System*, October, pp. 605-614, 1999

[8] G. Madhavan, H. D. Bruin: "Crosstalk Resistant Adaptive Noise Cancellation", *Annals of Biomedical Engineering*, 18, pp. 57-67, 1990

[9] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ, USA, 1996