CODEBOOK CONSTRAINED ITERATIVE NOISE CANCELLATION WITH APPLICATIONS TO SPEECH ENHANCEMENT*

Yan Gao, Jing Lu, Kai Yu, and Bo-Ling Xu

The Institute of Acoustics, National Key Laboratory of Modern Acoustics,

Department of Electrical Science and Engineering , Nanjing University, 210093 Nanjing, P.R. China

Fax: ++86-25-3315557 E-mail: assp@nju.edu.cn

ABSTRACT

The performance of widely-used adaptive noise canceling(ANC) deteriorates much when the desired signal is leaked into the reference channel or when there are uncorrelated noises present in the reference channel. This paper proposes a dual-microphone scheme, named Iterative Noise Canceling (INC), to overcome the drawbacks mentioned above. The proposed INC system, in which a codebook-based speech quality measure is employed to control a modified iterative Wiener filter (MIWF), can automatically reduce noises in the primary input until convergence occurs. In comparison with traditional ANC algorithm, the evaluation using real noises and voices recorded in a car shows the noise reduction performance is dramatically improved, even in cases that the reference SNR is close to 0 dB.

1. INTRODUCTION

In real environments, the presence of interfering noises always greatly degrades the performance of speech communication systems. Some techniques have been developed to solve the problem over the past decades, including spectral subtraction, all-pole modeling /noncausal wiener filtering[1], MMSE estimation etc.. Most of them are mainly under the assumption that the interfering signal is stationary, additive and nonspeechlike. Since the needed statistics of noises only can be estimated during speech pauses, those single-channel approaches present a poor performance when interference is time-varying. Whereas, in a conventional structure of ANC [2]-[4], there are two microphones: the primary microphone to obtain the noise-corrupted speech and the reference one to obtain a correlated component of the noise present in the primary input. The reference input is processed by an adaptive Wiener filter to generate a replica of the noise component in the primary input. Hence, it can automatically grasp the property of time-varying noises and lead to significant improvement to noise suppression, without making any strict hypotheses on the character of noise in advance.

In real acoustic environments, it is always inevitable that the target speech is also sampled by the reference microphone. Such a "leakage" makes the adaptive filter partially suppress the speech in the primary channel, resulting in notable distortion. Therefore, the signal-to-noise ratio (SNR) of reference signal is required to be very low. Particularly, according to the principle of noise minimizing method, any level of reference SNR above 0 dB will lead to a great distortion.

On the other hand, a diffuse background noise field lies in many environments, such as crowds, automobiles and flight jets. The perturbations can be characterized by a correlated component and an uncorrelated component. The uncorrelated noise in the primary input cannot be canceled by ANC. Moreover, the uncorrelated noise in reference channel not only is introduced into the output terminal by the adaptive filter, but also serves as an interference for adaptation of ANC.

An environment with a diffuse noise field like a car is considered in this paper. In order to obtain a reference signal which is correlated with the noise component in primary channel as much as possible, the two microphones should be placed closely. Unfortunately, this placement encounters a relatively high SNR of reference signal. As a result, thus performance of the ANC is limited [4].

We proposed a new structure, iterative noise canceling (INC). The initial relevant research was reported in [5] by Y.Cao etc., in which a modified iterative wiener filtering technique was briefly described for speech separation. This paper gives a further theoretical analysis of that approach in order to draw some conclusions about its behavior in real environments and to give a more regular MIWF. It is shown that the MIWF technique can provide a better performance than conventional ANC techniques when the reference noise is not very "clean".

The interference component in primary channel is attenuated consistently across iterations until optimal quality of output speech occurs. Since any further iteration will lead to a great impairment, there is clearly a need for a criterion of convergence based on some objective quality measures for occasions requiring automatic enhancement. Similar problems had also been encountered in [1] [6]. An obvious problem with those widely used measures is that outside of simulation the clean speech is unavailable, and hence, comparative evaluation is impossible[1][6]. A speaker classifier was introduced to determine the convergence of MIWF in [5]. However, the method cannot work well when the desired speaker is changed. In this paper, we proposed a new objective measure, which derived from some comparison between processed speech and a codebook, so the problem has been well solved in a simple way.

2. BACKGROUND OF INC

In a real dual-channel speech acquisition system, each microphone acquires not only the target speech, but also the interfering signals from the other sources. For simplicity, our theoretic analysis is limited to the two-source case. But later we will present experimental results demonstrating that our proposed system can actually work well in a diffuse noisy background in a car. Let s_0 and n_0 be the target speech and interfering noise obtained by the main microphone, mic1. Using the linear filters *A* and *B* to model the difference caused by spacial transfer channels between the signals received by main microphone and reference microphone, mic2. Fig.1 illustrated the dual-microphone system, which can be described in the time domain as

$$y_1(t) = s_0(t) + n_0(t)$$

$$y_2(t) = s_0(t) * a(t) + n_0(t) * b(t) \qquad (1)$$

where *a(t)* and *b(t)* represent the impulse responses of filters *A* and *B* respectively.

* This paper is supported by National Science Foundation of China (69872014)

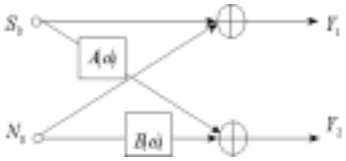

Fig.1. Block diagram of the dual-microphone acquisition system

Assume the signal and the noise are independent, and the noise is additive. Our problem of noise reduction could be solved if we could build a noncausal filter as follows,

$$H_{opt}(\omega) = \frac{P_{s_0}(\omega)}{P_{s_0}(\omega) + P_{n_0}(\omega)} \tag{2}$$

where $P_{s_0}(\omega)$ and $P_{n_0}(\omega)$ denote the power spectral densities (PSD) of the desired signal and the interfering noise, respectively. According to the theory of Wiener filtering, Eq(2) provides the optimum estimator of target speech $s_0(t)$ in a sense of minimum mean-squared error. But obviously, the Wiener filter of Eq.(2) can not be applied directly since the spectrums $P_{s_0}(\omega)$ and $P_{n_0}(\omega)$ cannot be known. Consider the following filter which is derived from Eq.(2),

$$H(\omega) = 1 - k \cdot \frac{P_{n_0}(\omega)}{P_{s_0}(\omega) + P_{n_0}(\omega)} \qquad k \le 1 \tag{3}$$

Eq.(3) provides an explanation to noncausal Wiener filter in a different way when k equals to 1. The optimal estimate of target speech can also be obtained by subtracting the optimal estimate of noise component from the noisy signal. Based on this view, we can construct a sub-optimal solution to iteratively approach the ideal filter described as Eq.(2). Consider if k<1 (for example k=0.1), then Eq.(3) means to partially subtract the noise component. Thus a scheme of iterative noise cancellation (INC) can be formed by running the filter again and again, as a result, it reduces the noise component little by little while maintaining the target speech s_0 at a certain level of magnitude until a critical situation occurs. Here we are facing two problems which should be resolved:

1. How to construct such a iterative wiener filter based on the signals acquired by our current two-microphone system?
2. Since any further iterations will lead to a great distortion of target speech after an optimal output has been obtained, how to detect the presence of such a "critical situation" across iterations?

3. MODEFIED ITERATIVE WIENER FILTER

The modified iterative Wiener filter (MIWF) technology was formally proposed to suppress the competing speech in [5]. Here we will further study it and present a more regular formulation of MIWF:

$$H_i(\omega) = \frac{(1+\alpha_i) P_{y_1^{(i)}}(\omega)}{P_{y_1^{(i)}}(\omega) + \beta_i P_{y_2}(\omega)} \qquad i = 1, 2, \cdots \tag{4}$$

where

$$\alpha_i = k \frac{\left|\hat{A}(\omega)\right|^2}{\left|\hat{B}(\omega)\right|^2} \prod_{k=0}^{i-1} \left|H_k(\omega)\right|^2 \tag{5}$$

$$\beta_i = k \frac{1}{\left|\hat{B}(\omega)\right|^2} \prod_{k=0}^{i-1} \left|H_k(\omega)\right|^2 \tag{6}$$

$$Y_1^{(i+1)}(\omega) = Y_1^{(i)}(\omega) \cdot H_i(\omega) \tag{7}$$

where $Y_1^{(i)}(\omega)$ is the output signal of Wiener filter at the (*i-1*)th iteration. At the beginning, $Y_1^{(1)}(\omega)$ is replaced by the primary input $Y_1(t)$ and $|H_0(\omega)|$ equals to 1. k denotes the step factor which can be varying across iterations. And it is required that $k << 1$ (for example $k = 0.1$).

A. How does it work

Now let's put an insight into the process of the iterative filtering – how does the proposed system work? Considering the first step of it, the modified wiener filter at first iteration can be transformed into

$$H_1(\omega) = 1 - \frac{k\left(1 - \frac{|A(\omega)|^2}{|B(\omega)|^2}\right) P_{n_0}(\omega)}{\left(1 + k\frac{|A(\omega)|^2}{|B(\omega)|^2}\right) P_{s_0}(\omega) + (1+k) P_{n_0}(\omega)} \tag{8}$$

$$\text{if} \qquad \hat{A}(\omega) = A(\omega) \quad \text{and} \quad \hat{B}(\omega) = B(\omega) \tag{9}$$

If the following conditions are satisfied,

$$\text{(I) } k << 1, \quad \text{(II) } k\frac{|A(\omega)|^2}{|B(\omega)|^2} << 1, \quad \text{(III) } \frac{|A(\omega)|^2}{|B(\omega)|^2} << 1 \tag{10}$$

then Eq.(8) can be approximately transformed into

$$H_1(\omega) \approx 1 - k_1 \cdot \frac{P_{n_0}(\omega)}{P_{s_0}(\omega) + P_{n_0}(\omega)} \tag{11}$$

which is quite similar to Eq.(2). As indicated before, Eq.(11) means to subtract some noise component from the input noisy signal. In fact, it can be proved, if the first step of approximation is acceptable, similar transformation can be performed again and again as iterations are going on. So the following wiener filters can all be the similar forms

$$H_i(\omega) \approx 1 - k \cdot \frac{P_{n_{i-1}}(\omega)}{P_{s_0}(\omega) + P_{n_{i-1}}(\omega)} \tag{12}$$

Note that the noise components vary in every step while the PSD of signal remains the same. Thus the process can be described as the continuous reduction of the noise component presented in the primary channel. In fact, the prerequisites described in Eq.(10) indicates that the performance of MIWF doesn't depend on the SNR of reference input but the difference of SNR between the two input noisy signals. This is a rather big merit of MIWF for it is comparatively easier to obtain a large difference between the two input noisy signals in practical use while in a traditional ANC, the prerequisite of a very low SNR in reference channel is hard to satisfied.

B. Impact of SNR Difference

As indicated above, the performance of noise cancellation depends on the SNR difference between the two input channels in a large sense. This difference can be described as following:

$$M_0(\omega) = \frac{P_{SNR2}(\omega)}{P_{SNR1}(\omega)} = \frac{|A(\omega)|^2}{|B(\omega)|^2} \tag{13}$$

Since $k << 1$, Eq.(8) can be approximately changed into

$$H_1(\omega) = 1 - k\frac{P_{n_0}(\omega)}{P_{s_0}(\omega) + P_{n_0}(\omega)} + k \cdot M_0 \frac{P_{n_0}(\omega)}{P_{s_0}(\omega) + P_{n_0}(\omega)} \tag{14}$$

The third term on the right side of Eq.(14) indicates that a distortion is introduced into the noise component, besides some part of the noise has been suppressed. The more significant the SNR difference is, the lower such distortion is. It is also shown that the level of SNR difference does not influence the distortion of the target speech, while the same case in ANC inevitably impacts the distortion level of the target speech. A further quantitative analysis suggests that if

the primary SNR is 6dB greater than the reference SNR, ANC using LMS algorithm presents the output with 25% distortion of target speech, but the MIWF produces output with about 3.3% distortion of noise at the first step when $k = 0.1$.

C. Impact of Channel-Estimate Error

The analysis above is under an assumption that the channels *A* and *B* could be accurately estimated. Since error cannot be avoided, it is necessary to study the influence of channel-estimate error on the behavior of MIWF. In such a case, the Wiener filter at the first step can be written as

$$H_1(\omega) \approx 1 - \frac{k\frac{A^2(\omega) - \hat{A}^2(\omega)}{\hat{B}^2(\omega)} P_s(\omega) + k\left(\frac{B^2(\omega)}{\hat{B}^2(\omega)} - \frac{\hat{A}^2(\omega)}{\hat{B}^2(\omega)}\right) P_n(\omega)}{P_s(\omega) + P_n(\omega)} \quad (15)$$

From eq.(15), we can find the Wiener filter leads to not only the distortion of noise component but also that of target component. The following conclusions are drawn from our study:

1. MIWF is not sensitive to the error of $\hat{A}(\omega)$, when the SNR difference is notable, i.e. not less than 6dB. But MIWF is somewhat sensitive to the error of $\hat{B}(\omega)$.
2. Error of $\hat{A}(\omega)$ mainly causes the distortion of target signal, while error of $\hat{B}(\omega)$ influences the noise reduction.
3. The impairment is comparatively small if $\hat{A}(\omega)$ tends to be lower than $A(\omega)$, or if $\hat{B}(\omega)$ tends to be higher than $B(\omega)$.

Details about our analysis will be presented in another paper. One interesting point regarding the above conclusions is that, although the performance is more sensitive to the error of $\hat{B}(\omega)$, it is easier to estimate $B(\omega)$ in practice—We can get it during speech pauses. Later, our experiments will show that a simple setting of $\hat{A}$ and $\hat{B}$ can get a good performance.

4. CODEBOOK-BASED CRITERION FOR CONVERGENCE DETECTION

With INC going on, it has been observed that the quality of filtering output speech is getting better and better. After an optimal step happens, however, the speech quality is significantly decreased. The phenomenon indicates that an objective speech quality measure can be employed to serve as the criterion for convergence detection. As far as we know, nearly all the widely used objective quality measures are some comparison between processed noisy speech and original clean speech [7]. But in real speech enhancement applications, the clean speech is inevitably unknown. As the behavior of mankind, machine's ability of evaluating how greatly a signal sounds like a speech should be built on a broad knowledge of speech features. So we use unsupervised learning techniques to derive the requisite knowledge of voice. A simple approach to this is through pattern clustering of clean speech spectra. In the present work, a codebook-based objective speech quality measure has been firstly proposed and applied to form the criterion for convergence detection.

Let {a} be a set of LPC vectors derived from clean speech data of a set of selected speakers. The perceptual difference between two LPC vectors is well correlated to the IS (Itadura-Satio) distortion measure. A close approximation to the IS measure has be widely used in vector quantization [8],

$$d(a_x, a_y) = (a_x - a_y)^{\mathrm{T}} R_x (a_x - a_y) \quad (16)$$

where a_x and a_y are any two LPC vectors of the same order, and R_x is the normalized autocorrelation matrix corresponding to the LPC vector a_x. Using the IS distortion measure, the problem of pattern clustering can be solved applying the iterative K-means algorithm [8], where K is the order of 1024 (since such a codebook size has been found sufficient in speech coding applications).

Based on the formed codebook, a new speech quality measure, codebook-based distortion, has been introduced as following,

$$d_{cbk}(x) = \frac{1}{1024} \sum_{i=1}^{1024} d(x, y_i) \quad (17)$$

Codebook-based distortion is the mean distance between feature vector x and all the vectors in the codebook. It is quite different from VQ algorithm, in which the nearest distance between a test vector and each cluster is always desired. Since the codebook represents a special class of data (speech feature vectors) in the sound feature space, due to interclass fuzzy border and intraclass dispersed distribution, the measure that how exactly a LPC vector is speechlike should be derived through the mean distance to the whole speech class.

5. EXPERIMENTS

Experiments were conducted under a simulated environment based on the signals sampled in a real car acoustic conditions. The data acquisition system consists of two microphones set vertically in a car with a distance of 30 cm. The noise signals were accumulated when the car was running in a highway while the speech signals were sampled in a quiet circumstance when the car was held still. Thus by adjusting the magnitude of speeches in two channels before the mixture is made, various SNR can be obtained. As far as the codebook is concerned, LPC parameters extracted from recorded speech sentences of four male and four female speakers were used as the training vectors. Some of the parameters concerning the preprocessing are sampling frequency 8kHz, LPC model order 12, frame window width 32ms and successive frame overlap 8ms. Since the primary microphone should be placed more closely to speaker's mouth than the reference one in practice, we simulate the situation where the car noises obtained by two microphones have the same SPL level while the SPL levels of speeches are different due to the different distances from speaker's mouth to microphones. In such a case, the parameters *k*, $\hat{A}$ and $\hat{B}$ are fixed to 0.05, 0.1 and 1, respectively. As indicated before, the performance of MIWF is not sensitive to the error of $\hat{A}$. Our experimental results also support this point.

To demonstrate the efficiency of the codebook-based criterion for convergence detection, two objective speech quality measures were introduced for comparison, one is the LPCC linear distance between the processed speech and the original clean speech and the other is IS distortion measure. Figure 2 shows the variation of three measures across iterations. (1), (2) and (3) comes from the test results of three inner speakers whose data are within the training set which forms the codebook while (4) is the test results of an outer speaker whose data is out of the training set. It is obvious that the proposed distortion holds well consistency with two objective measures in inner-speaker tests. So we expect that if a general codebook is derived from a large set of speakers' data, it could be suitable for any speaker that convergence of MIWF can be easily detected by checking whether the proposed distortion is increased. In the present experiments, totally 20 inner-speaker tests have been done and the rate of correct detection is 100%.

A comparison of performance between the proposed INC and the traditional ANC using LMS algorithm (order is 200) under various SNR differences is given in Fig.3 and Fig.4, with SNR 0dB and 5dB in the main channel respectively. It has been shown that INC's performance is dramatically better than ANC's. From the evaluation of Itakura Distortion, we can learned that the proposed system is less sensitive to the reference SNR than ANC, especially when SNR difference is between 8dB to 20dB

Figure 5 illustrates the difference between the processed waveforms of INC and ANC algorithm. Compared with the original noisy signal (the first waveform), although the result of ANC (the second waveform) partially suppresses the noise component, the desired speech is also attenuated. However, INC holds the desired

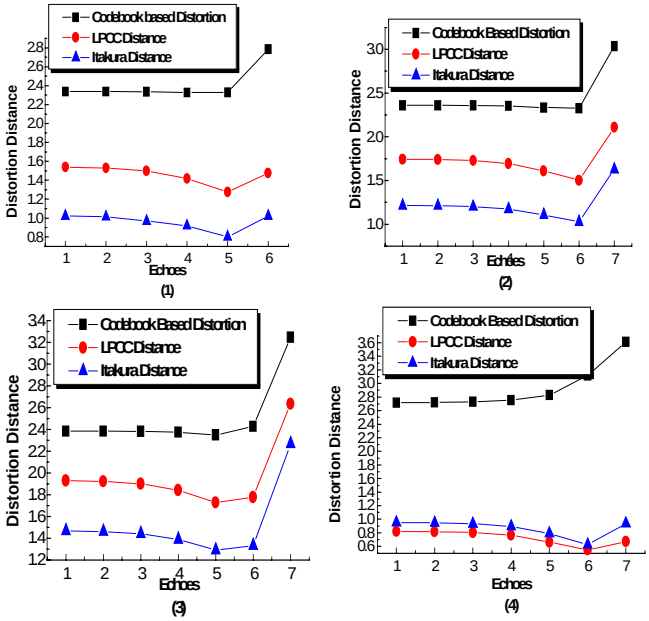

Figure 2. Variations of codebook-based quality across iterations

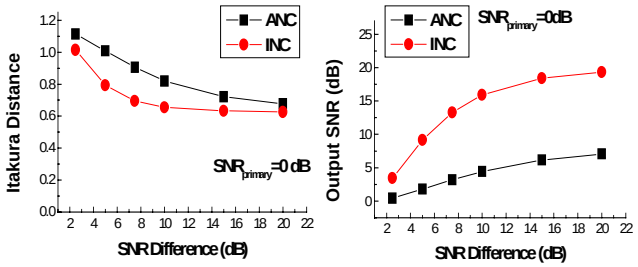

Figure 3. Comparison of performance between INC and ANC (Primary SNR is 0dB)

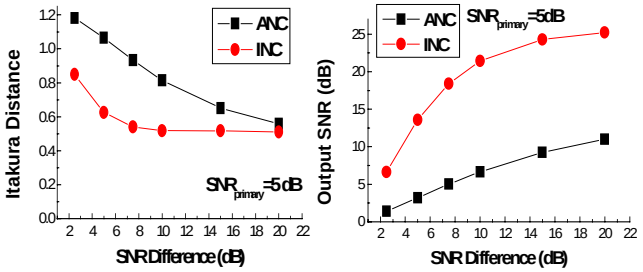

Figure 4. Comparison of performance between INC and ANC (Primary SNR is 5dB)

speech well while greatly suppresses the noise component. As shown in Fig.5, due to the diffuse noise field in a car, ANC cannot reduce the uncorrelated noise. Whereas, since INC is a sub-optimal solution to the noncausal Wiener filter, it can significantly suppress the incoherent noise in the domain of PSD.

6. CONCLUSION

In this paper, we propose an iterative noise canceling scheme: a modified iterative Wiener filter controlled by a codebook- based speech quality measure. It is shown that the proposed INC structure results in dramatically improved speech enhancement, even when the reference signal is not "clean". Experimental results demonstrate that in a real diffuse noise field such as a running car, INC gives much better speech enhancement result than the traditional noise cancellation methods such as ANC. Furthermore, because of the comparatively loose conditions required by it, INC system can be expected to have a wide use in automatic noise reduction.

REFERENCES

1. F.S.Lim and A.V.Oppenheim, "All-pole modeling of dgraded speech", *IEEE Trans. on ASSP*, Vol.26, pp.197-210(1978).
2. B.Widrow et.al., "Adaptive noise cancelling: princeples and applications," *Proc. IEEE*, vol.63, pp.1692-1716, Dec. 1975.
3. W.A.Harrison, J.S.Lim, and E.Singer, " A new application of adaptive Noise cancellation," *IEEE Trans. Acoust., Speech, Signal Processing*, vol.34, pp.21-27, Jan. 1986.
4. V.Parsa, P.A.Parker, and R.N.Scott, "Performance analysis of a crosstalk resistant adaptive noise canceller", *IEEE Trans. Circuits Syst.*, vol. 43, pp, 473-482, 1996.
5. Y.Cao, Sridha Sridharan, and Miles Moody, "Co-talker separation using the "Cocktail Party Effect" ," *J. Audio Eng. Soc.*, vol. 44. No. 12, 1996.
6. J.H.L.Hansen and M.A.Clements, "Iterative speech enhancement with spectral constraints", *ICASSP'87*, vol.I, pp.189-192.
7. J.H.L.Hansen and B.L.Pellom, " An effective quality evaluation protocol for speech enhancement algorithms", b*ICSLP-98*, Sidney, Australia, 1998
8. Y.Linde, A.Buzo, and R.Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, no.1, pp.84-94, Jan. 1980

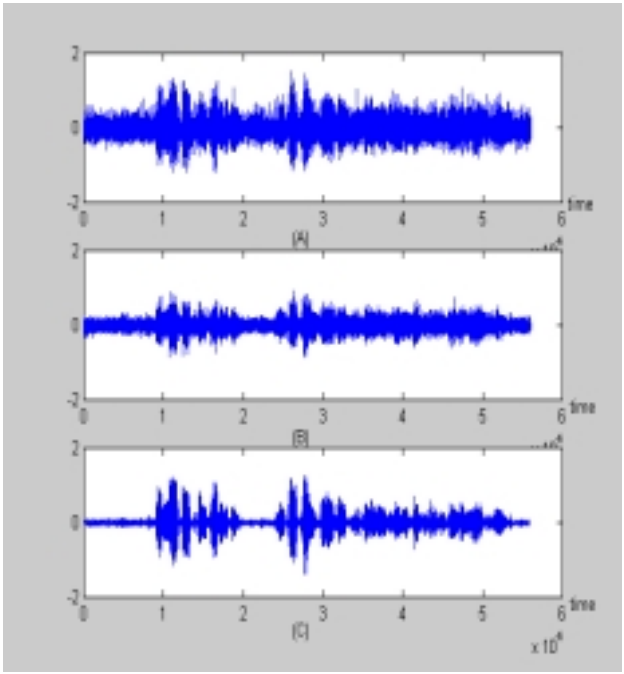

Figure 5. Comparison of result waveforms between INC and ANC (SNR difference between two inputs is 10 dB)