<table>
<thead>
<tr>
<th>Title</th>
<th>Recovery of speech signal from contaminated input (Accepted version)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Lee, Huan Hong; Foo, Say Wei</td>
</tr>
<tr>
<td>Date</td>
<td>2001</td>
</tr>
<tr>
<td>URL</td>
<td><a href="http://hdl.handle.net/10220/4609">http://hdl.handle.net/10220/4609</a></td>
</tr>
</tbody>
</table>

Rights: © IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE. This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.
Recovery of Speech Signal from Contaminated Input

Huan Hong Lee              Say Wei Foo
Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117576
huanhong@pacific.net.sg
elefoosw@nus.edu.sg

Abstract
In this paper, investigation is made on the application of an adaptive FIR filter to estimate the desired audio signal, in this case speech signal, from an input of the signal contaminated with background music. Results show that the signal to noise ratio is significantly increased. Informal listening tests also show that very significant improvement to the quality of the recovered signal can be achieved. The condition for the success of the method hinges on the availability of information of the interfering source such as the background music.

1. Introduction
Very often, recording of speech signal is made in the presence of unwanted signal such as background music. In certain cases, the background music may be so loud and render the speech inaudible or unintelligible. In situations where the background music or noise can be identified, it may be possible to 'subtract' the background noise from the 'contaminated' input and enhance the desired signal.

In this paper, the application of adaptive FIR filter for background music cancellation is investigated. In Section 2, the system for adaptive noise cancellation is presented. The adaptive FIR filter and least mean square method of gradient descent are described in Sections 3 and 4 respectively. Details of the experiments and performance of the system are covered in Section 5 and the concluding remarks given in Section 6.

2. Adaptive Noise Cancellation System
The block diagram of the system is given in Fig.1. The input signal, \( p(t) \), consists of the desired signal \( s(t) \) and a noise component \( n_0(t) \). \( n_0(t) \) in turn is the sum of the reference signal \( n_1(t) \) and random noise. The reference signal \( n_1(t) \) may be background music or other known interference signal.

We shall assume that \( s(t) \), \( n_0(t) \), \( n_1(t) \) and \( y(t) \) are statistically stationary and have zero means, and that \( s(t) \) is uncorrelated with \( n_0(t) \) and \( n_1(t) \) but \( n_1(t) \) is correlated with \( n_0(t) \).

The output \( z \) is
\[
z = s + n_0 - y
\]

Squaring (1), we have
\[
z^2 = s^2 + 2s(n_0 - y) + (n_0 - y)^2
\]

Taking the expected value of (2), we obtain
\[
E[z^2] = E[s^2] + E[2s(n_0 - y)] + E[(n_0 - y)^2]
\]

Since \( s \) is not correlated with \( n_0 \) and \( y \), the term vanishes and therefore
\[
E[z^2] = E[s^2] + E[(n_0 - y)^2]
\]
The signal power is a fixed quantity for the frame of interest. It may be estimated by minimizing the term \( E[(n_0 - y)^2] \).

\[
E_{\text{min}}[z^2] = E[s^2] + E_{\text{min}}[(n_0 - y)^2] \tag{5}
\]

When the adaptive filter is adjusted so that \( y \) is the best estimate of \( n_0 \), \( z \) is then the best estimate of the desired signal \( s \).

3. Adaptive FIR Filter

The adaptive FIR filter is shown in Fig.2. The impulse response of the filter is determined by the weight settings. The adaptation process adopts an algorithm that automatically seeks an optimal filter impulse response by adjusting the weights.

![Fig.2 Adaptive FIR filter](image)

For a FIR filter with \( M \) taps, \( M \) samples of the input signal appear simultaneously on all input lines at discrete time indexed by \( t \).

Denote the input signal vector of the adaptive linear combiner as \( U(k) \). Mathematically,

\[
U^T(k) = [u(t), u(t-1), \ldots, u(t-M+2), u(t-M+1)]^T \tag{6}
\]

The weight vector is defined as

\[
W^T(k) = [w_1(t), w_2(t), \ldots, w_{(M-2)}(t), w_{(M-1)}(t)]^T \tag{7}
\]

The output \( y(t) \) of the adaptive filter is given by

\[
y(t) = \sum w_k(t)u(t-k) \tag{8}
\]

where \( w_k(t) \) denotes the \( k^{th} \) weight of the filter and \( U(t-k) \) denotes the corresponding input sample at time \( t \). In matrix form,

\[
y(t) = U^T(t)W(t) \tag{9}
\]

Denote the contaminated input signal as \( p(t) \), where \( p(t) = s(t) + n_0(t) \). The output \( z(t) \) is then given by

\[
z(t) = p(t) - y(t) = p(t) - U^T(t)W(t) \tag{10}
\]

Squaring \( z(t) \) and taking the expectation of the resulting terms, we have

\[
E[z^2(t)] = E[p^2(t)] - 2E[p(t)U^T(t)W(t)] + E[W^T(t)U(t)U^T(t)W(t)] \tag{11}
\]

Define \( \phi(p, r) \) as the vector of cross-correlation between the reference signal and the input signal and \( \phi(r, r) \) as the auto-correlation matrix of the reference signal, then (11) may be written as

\[
E[z^2(t)] = E[p^2(t)] - 2\phi(p, r) + W^T(t)\phi(r, r)W(t) \tag{12}
\]

Equation (12) shows that the estimate of the signal is a quadratic function of the weights. One method to adjust the weights is the method of steepest descent with the gradient given by

\[
\nabla E[z^2(t)] = -2\phi(p, r) + 2\phi(r, r)W(t) \tag{13}
\]

The optimum weight vector can be expressed as

\[
W^* = \phi(p, r)\phi^{-1}(r, r) \tag{14}
\]

This method involves finding the inverse of the correlation matrix and the computational load is high. An alternative approach is to use the least mean square method of gradient descent.

4. Least Mean Square Method of Gradient Descent

The basic equation is given by

\[
W_{j+1} = W_j - \mu \nabla [z^2(t)]
\]

\[
= W_j - \mu [2z(t)\nabla z(t)]
\]

\[
= W_j - \mu [2z(t)\nabla [p(t) - U(t)W^T(t)]]
\]

\[
= W_j + \mu z(t)U(t) \tag{15}
\]
By expanding the matrix equation above, we obtain
\[
\begin{align*}
    w_0(t+1) &= w_0(t) + 2\mu z(t)u(t) \\
    w_1(t+1) &= w_1(t) + 2\mu z(t)u(t-1) \\
    \vdots & \\
    w_{M-1}(t+1) &= w_{M-1}(t) + 2\mu z(t)u(t-M+1)
\end{align*}
\]

This relationship, known as the Widrow-Hoff Least Mean Square algorithm, is used to update the weights of the adaptive FIR filter. Although the method makes use of gradients of mean square error function, it does not require differentiation in the final form. In order for the adaptation process to converge, the parameter \( \mu \) must meet the following requirement.

\[
    | -2\mu \lambda_q | < 1 \quad \text{for } q=1,2,3,\ldots M
\]

where \( \lambda_q \) is the \( q \)th eigenvalue of the autocorrelation matrix \( \phi(r,r) \) and \( \lambda_{\max} \) is the largest eigenvalue of \( \phi(r,r) \). This convergence condition can be related to the total input signal power as follows.

\[
    \lambda_{\max} < E[U(t)U^T(t)] \\
    < \sum_{i=1}^{N} E[u_i^2] \\
    < \text{Total tap input power}
\]

It follows that satisfactory convergence can be obtained with

\[
    0 < \mu < 1/\text{Total tap input power}
\]

The flowchart for the computation of the weights of the adaptive FIR filter is given in Fig.3.

5. Experiments and System Performance

Several experiments using speech signals contaminated by background music and noise of different levels were conducted.

One example using simulated signals is illustrated in Fig.4. A speech (8.53 seconds long) corresponding to the sentence, "We pride ourselves on the high moral standards of our program, and I mean that sincerely, we don't resort to sex or crime or violence or drinking." was corrupted by a music signal and random white noise to simulate a speech given in a party, being 'drowned' by the background music in a noisy environment. The signals are sampled at 11025 samples per second and quantized with 8 bits per sample.

![Flowchart of weight adjustment](image)

The waveforms of the signals are presented in Fig.4. The contaminated speech signal is presented in Fig.4(a); the original clean music signal is given in Fig.4(b); the original clean speech signal is given in Fig.4(c) and the recovered signals using 2048-tap filter and 256-tap filter are shown in Fig.4(d) and Fig.4(e) respectively.

Filters with different number of taps are used. The signal to noise ratios of the input signal and the output signal are computed for the different number of taps used.

The input and output signal to noise ratios are calculated as follows.

\[
    \text{SNR (input)} = 10\log_{10} \left( \frac{E[z^2(t)]}{E[n_0^2(t)]} \right) \quad (19)
\]

\[
    \text{SNR (output)} = 10\log_{10} \left( \frac{E[z^2(t)]}{E[z^2(t) - s^2(t)]} \right) \quad (20)
\]
Fig. 4(a) Speech contaminated with music

Fig. 4(b) Original clean music signal

Fig. 4(c) Original clean speech signal

Fig. 4(d) Recovered speech signal using 2048-tap FIR filter

Fig. 4(e) Recovered speech signal using 256-tap FIR filter
The output SNRs for the different number of taps used are given in Table I. It can be seen that the SNR increases with the number of taps used. The reason of increasing SNR with the number of taps is possibly due to the fact that the initial latency of the filter must be long enough to cover the time delay when noise reference signals travel to the primary sensor.

<table>
<thead>
<tr>
<th>No. of Taps</th>
<th>SNR (input)</th>
<th>SNR (output)</th>
</tr>
</thead>
<tbody>
<tr>
<td>256</td>
<td>1.9086</td>
<td>2.6838</td>
</tr>
<tr>
<td>512</td>
<td>1.9086</td>
<td>4.3039</td>
</tr>
<tr>
<td>1024</td>
<td>1.9086</td>
<td>4.3351</td>
</tr>
<tr>
<td>2048</td>
<td>1.9086</td>
<td>5.8557</td>
</tr>
</tbody>
</table>

The mean square errors of residual noise for different numbers of taps as the adaptation progresses are shown in Fig.5. From the figure, it can be seen that the errors are lower for larger number of taps used and that thousands of iterations are required to reach the steady state.

6. Conclusion

The application of an adaptive FIR filter for the recovery of speech signal from input signal drowned in background music is investigated. Least Mean Square Method of Gradient Descent is used for weight adjustment of the adaptive filter.

It is established that a minimum number of taps is required of the adaptive filter to achieve good results. The minimum number of taps is required to account for the time difference for signals to travel to the primary sensor and the reference sensor.

Experiments and informal listening tests have shown that significant improvement of the speech signal can be achieved using the method.

The method can readily be extended to cases where the desired signal is corrupted by more than one source of 'noise' as long as some information of the 'noise' such as background music is available and that the desired signal is not correlated to the 'noise'.

Acknowledgment

The authors would like to thank DSTA for their support.

References