Spectral Modulation Sensitivity Based Perceptual Acoustic Echo Cancellation

Wei-Lun Chuang, Kah-Meng Cheong, Chung-Chien Hsu and Tai-Shih Chi

Department of Electrical and Computer Engineering
National Chiao Tung University, Hsinchu, Taiwan 300, R.O.C.
s900925@gmail.com, jiamingnctu@gmail.com, hsu.chung.chien@gmail.com, tschi@mail.nctu.edu.tw

Abstract
Perceptual acoustic echo cancellers were developed by mainly considering human hearing thresholds of different acoustic frequencies in the past. In addition to the different frequency sensitivities, the human brain further analyzes sounds in terms of their spectral and temporal modulations. In this paper, we extend the perceptual normalized least mean square (P-NLMS) algorithm by adding a second pre-emphasis filter which accounts for brain’s spectral modulation sensitivity. Simulation results and listening tests demonstrate our system effectively reduces the perceived residual echo during convergence due to faster convergence rates in hearing-sensitive spectral modulation bands and acoustic frequency bands.

Index Terms: spectral modulation, hearing threshold, perceptual acoustic echo cancellation.

1. Introduction
In hands-free telecommunication systems, acoustic echoes inevitably arise when speaker signals feed back into microphones and thus deteriorate speech quality. Adaptive algorithms are often used in acoustic echo cancellers (AECs) to reduce acoustic echoes [1][2]. The research field of AEC is well-established and many algorithms have been proposed over the past few decades [3]. Most of these algorithms were developed to increase the convergence rate of the adaptive filter using variable step sizes or to reduce the computational complexity. However, in early days, these algorithms were seldom evaluated from perceptual viewpoints by assessing the perceived residual echo. Many researchers have successfully applied hearing perceptual properties to the development of speech signal processing, for instance, the masking property of the human auditory system adopted in transparent coding [4]. Recently, in the AEC field, Gordy and Goubran showed the steady-state echo return loss enhancement (ERLE) and the mean square error (MSE) cannot really reveal whether residual echo is perceivable. In addition, they showed the perceived residual echo can be suppressed by increasing the rate of convergence on those acoustic frequency bands to which human hearing is most sensitive [5].

In addition to the fact that human hearing possesses different sensitivity to different acoustic frequencies, our brain analyzes sounds in terms of their joint spectro-temporal modulations [6]. The concept of joint spectro-temporal modulation filtering has been successfully used in many speech-related applications, such as noise reduction [7], voice activity detection [8] and speech intelligibility assessment [9]. The detection thresholds of spectro-temporal modulations have been measured in psycho-acoustic experiments and demonstrated that it is separable in terms of detection thresholds of spectral and temporal modulations [10].

The concept of using different spectral weights pertaining to hearing thresholds of acoustic frequencies in AEC adaptive filter design is extended by further including detection thresholds of spectral modulations. Since hearing thresholds only address the minimum magnitudes of the sound that human can hear at different acoustic frequencies, we further emphasize on spectral modulations, where the brain is more sensitive, in building a perceptual AEC. In other words, the detection thresholds of spectral modulations are considered when analyzing the input far-end speech signal in the AEC problem.

The rest of the paper is organized as follows. In section 2, the perceptual effect due to different spectral modulation sensitivities is demonstrated. In section 3, the proposed perceptual AEC, an extension of the adaptive filter structure of [5], is introduced with objective and subjective test results. Finally, conclusions and future work are given in section 4.

2. Spectral Modulation Sensitivity
In this section, a brief introduction of an auditory model is given to demonstrate the perceptual effect due to different detection thresholds of spectral modulations. This auditory model consists of two parts: (1) an early cochlear module, shown in Fig. 1, which transforms the input acoustic signal into an auditory spectrogram, and (2) a central cortical module which further analyzes the auditory spectrogram based on joint spectro-temporal modulations. Detailed descriptions of this auditory model can be accessed in [6]. The speech signal is then analyzed using this auditory model and weighted by spectral modulation sensitivities. Based on these findings, we can build a perceptual AEC, which puts more emphasis on more critical spectral modulations while cancelling the echo.

Fig. 1. Stages of the early cochlear module of the auditory model.

2.1. Cochlear module
The early cochlear module transforms the input acoustic signal from a waveform into an auditory spectrogram by considering critical neural activities along the peripheral auditory pathway. As shown in Fig. 1, this cochlear module first filters the input sound s(t) using 128 constant-Q band-pass filters which simulate the quasi-logarithmic frequency selectivity of the cochlea. The output of each filter is then passed through a non-linear compression stage (hair cell model) and a lateral inhibitory network (LIN). The non-linear compression models the saturation of the inner hair cells, and the LIN sharpens the frequency resolution of the cochlea and models the frequency masking effect. Outputs of the LIN are further processed by an
envelope extractor (a half-wave rectifier followed by a low-pass filter) to produce an auditory spectrogram which represents a more robust energy pattern compared with the Fourier spectrogram. More details of this cochlear module can be found in [6].

Fig. 2. (a) Spectral profile of a moving ripple stimulus; (b) Detection thresholds of spectral modulations [11].

2.2. Spectral modulation sensitivity

Joint spectro-temporal modulations of speech have been shown to provide important cues to speech intelligibility [9][10]. However, to build an AEC with frame based pre-emphasis filters, only the spectral modulation sensitivity of human hearing is considered in this paper. The human sensitivity to spectral modulations has been studied in [11]. The role of spectral modulations in human perception of speech is characterized by the sensitivity measure using ripple stimuli. Each ripple stimulus is constructed by equally spaced frequency components along the logarithmic frequency axis. As shown in Fig. 2(a), the spectral profile of each ripple stimulus is:

\[
S(x) = A \sin(2 \pi \Omega \cdot x + \Phi)
\]  

where \(A\) is defined as the modulation depth (ranging from 0 to 1); \(x\) represents the logarithmic frequency: \(\Omega\) (scale) is defined as the ripple density in units of cycles/octave (cyc/oct); and the phase \(\Phi\) determines the starting position of the sine function in radians or degrees. The measured detection thresholds of spectral modulations with \(\Omega = 0.5-8\) cyc/oct are shown in Fig. 2(b). This curve depicts the spectral modulation sensitivity and is called the spectral modulation transfer function (MTF). This MTF resembles a low-pass character, which indicates tested subjects are mostly sensitive to \(\Omega = 1-4\) cyc/oct spectral modulations and quickly lose their ability in discriminating faster changing spectral profiles (\(\Omega > 5\)).

Based on the spectral modulation sensitivity, we extend the AEC in [5], which puts more emphasis on perceptually important acoustic frequencies, to further consider perceptually important spectral modulation bands. The spectral modulation gain function (SMGF) is obtained by inverting the detection thresholds of spectral modulations, i.e., a higher/lower gain is assigned to the modulation band with a lower/higher threshold. To examine the effect of the SMGF on a spectral profile, we built a bank of spectral modulation bandpass filters [6]. The impulse response \(h_{\text{scale}}\) of the spectral modulation filter tuned to \(\Omega\) can be written as:

\[
h_{\text{scale}}(x; \Omega) = \Omega_c (1 - \Omega_c^2 x^2) e^{-\Omega_c^2 x^2/2}
\]  

where the filter bandwidth is increased with the center frequency \(\Omega\). To encode the different spectral modulation sensitivity, a spectral profile is first passed through the spectral modulation filter bank, and the output of each filter is multiplied by the corresponding gain. Since the spectral modulation filtering is simply a linear operation, a perfect reconstruction is attainable.

Fig. 3. (a) An arbitrary Fourier spectrum of an vowel; (b) the corresponding enhanced spectrum by incorporating SMGF; (c) auditory spectrum of the enhanced speech signal.

Fig. 3(a) shows an arbitrary Fourier spectral profile (magnitude spectrum) of a sample speech signal at a given time instance. The harmonic structure can be clearly seen at all frequencies. Fig. 3(b) shows the enhanced spectral profile by incorporating the SMGF. The harmonics in high frequency region (approximately >2 kHz) are suppressed because they are resolved as higher scale (more cycles per octave) modulations, which are less critical to human hearing. In other words, the high frequency region would be suppressed and become smoother due to the lower gains of the higher-scale spectral modulation filters. Fig. 3(c) illustrates the auditory spectrum, produced by the cochlear module, of the enhanced speech signal in Fig. 3(b). Clearly, the harmonic structure in the high frequency region is diminished while the harmonic structure in the low frequency region is further enhanced due to the quasi-logarithmic frequency selectivity of the cochlea. This spectral profile produced by the cochlear module of the auditory model mimics the spectral profile preserved along the auditory pathway. This figure clearly shows that humans are more sensitive to the harmonic structure of speech in lower frequency regions, and to the formant structure of speech in higher frequency regions.

3. Proposed System

The goal of an AEC is to remove annoying echoes in reverberant environments as fast as possible. Thus, the performance of the AEC should be addressed by assessing the residual echo \(\hat{h}(n)\) perceived by the far-end user [5]. In this paper, we extend the P-NLMS algorithm [5] by adding a second pre-emphasis filter which reflects the SMGF of the brain. The basic block diagram of the P-NLMS algorithm is shown in Fig. 4. Its function can be expressed by:

\[
\begin{align*}
y(n) &= h^T(n) x(n) \\
d(n) &= y(n) + v(n) \\
\delta(n) &= [h(n) - \hat{h}(n)]^T x(n) = \Delta h^T(n) x(n) \\
e(n) &= \delta(n) + v(n)
\end{align*}
\]  

where \(h(n)\) is the echo path that can be modeled by a linear finite impulse response (FIR) of length L. The error signal \(e(n)\) is composed of the residual echo \(\delta(n)\), which results from the misadjustment \(\Delta h^T(n)\) between the actual echo path and the adaptive filter, and the near-end background noise \(v(n)\).
3.1. Acoustic frequency weighted NLMS algorithm

The effect of the SMGF on a spectral profile is demonstrated in the previous section. Consequently, it is intuitive to put more emphasis on those perceptually critical aspects when reducing residual echoes. We adopt the echo canceller structure in [5], which proposed a frequency-weighted version of normalized least mean square (NLMS) method by adding a pre-emphasis filter. It showed that it is feasible to increase the rate of convergence in certain frequency bands. Let $x_f(n)$ and $e_f(n)$ be the input and the error signal filtered by the added pre-emphasis filter $f(n)$:

\[
\begin{align*}
    x_f(n) &= f^T(n)x(n) \\
    e_f(n) &= f^T(n)e(n)
\end{align*}
\]

where the NLMS filter coefficient update function becomes

\[
\hat{h}(n+1) = \hat{h}(n) + \mu \frac{e_f(n)x_f(n)}{x_f^T(n)x_f(n)}
\]

The choice of $f(n)$ significantly affects the overall performance. The design criterion of the pre-emphasis filter is to increase the eigenvalues of those hearing-sensitive bands which put a limitation to the AEC’s convergence rate [12].

3.2. SMGF weighted pre-emphasis filter design

In [5], a fixed filter $f_1(n)$ was roughly generated by inverting the curve of the absolute hearing threshold to mimic the acoustic frequency gain function (AFGF) of human hearing. To fully incorporate SMGF effects, a time-varying filter, which emphasizes the harmonic structure in lower frequency regions and the formant structure in higher frequency regions, is needed and can be generated by applying the SMGF to the input spectra of far-end speech. However, since the echo path is also time-varying in real environments plus the slow convergence rate of the adaptive filter, the effect of adding a time-varying pre-emphasis filter to NLMS algorithm cannot be accurately applied to the corresponding output frame in real-time implementations using current technologies. In other words, once the adaptive filter tunes to reducing more echoes at desired frequency bands, the spectral profile of input speech has already changed due to fast varying formant structures, which is known directly related to the articulation of speech.

Thus, in this paper, we focus on the harmonic structure which varies more slowly than the formant structure. Combining the fact that humans are sensitive to harmonic structures in low frequency regions, a fixed filter $f_2(n)$ was designed to roughly contain the low-frequency harmonic structure. In this study, for a given utterance, the magnitude response of $f_2(n)$ was derived by extracting frequencies of the first eight peaks below 2 kHz from the first several speech frames in a Fourier spectrogram. An example is shown in Fig. 5. In real systems, this filter can be set/reset by a table look-up method after performing the pitch detection every couple hundred milliseconds. Accordingly, the filtered version of input and error signals become

\[
\begin{align*}
    x_f(n) &= f_1^T(n) f_2(n)x(n) \\
    e_f(n) &= f_1^T(n) f_2(n)e(n)
\end{align*}
\]

4. Experiment Evaluations

In this section, the steady-state performance of the proposed echo canceller is assessed by simulations. We used white Gaussian noise as our input signal $x(n)$. A 16 kHz 2500-sample meeting room impulse response was extracted from the Aachen Impulse Response (AIR) database [13] as our $h(n)$. A white noise $v(n)$ extracted from NOISEX-92 was added as the background noise. Three adaptive filters were investigated in our simulations: (1) standard NLMS; (2) P-NLMS which uses $f_1(n)$ as a pre-emphasis filter [5]; (3) our proposed system with $f_1(n)$ and $f_2(n)$ as pre-emphasis filters. The step size $\mu = 0.1$ and filter length $M = 2500$ samples were fixed for all three adaptive filters.

In addition to the objective MSE measure, subjective listening tests similar to the ones in [5] were also conducted for evaluations. The adaptive filter coefficients $\hat{h}(n)$ were extracted for each algorithm at five time instances during convergence, and then the mis-adjustment $\Delta \hat{h}(n)$ was calculated. Error signal of each algorithm was finally constructed by adding the residual echo signal (\(\pm \Delta \hat{h}(n)x(n)\)) and the background noise $v(n)$ with the 20dB residual-echo-to-background-noise ratio as in equation (7).

\[
\begin{align*}
    e_{NLMS}(n) &= \Delta \hat{h}^T_{NLMS}(n)x(n) + v(n) \\
    e_{P-NLMS}(n) &= \Delta \hat{h}^T_{P-NLMS}(n)x(n) + v(n) \\
    e_{PROPOSED}(n) &= \Delta \hat{h}^T_{PROPOSED}(n)x(n) + v(n)
\end{align*}
\]

Ten sets of utterances were randomly selected from TIMIT corpus as our input signal $x(n)$. Five subjects aged between 22 and 25 were recruited for the two-alternative preference test [14]. Subjects were asked to tell which residual echo signal was more perceivable. To reduce test bias, pairs of test signals were presented in a random order. Two listening tests were conducted: (I) NLMS versus the proposed algorithm; and (II) P-NLMS versus the proposed algorithm.
Fig. 6. Magnitude response of mis-adjustment vectors after n=21000 samples. (a) NLMS and proposed method. (b) P-NLMS and proposed method.

Fig. 7. MSE of NLMS, proposed and P-NLMS during convergence period.

The objective MSE measures of these three algorithms are plotted in Fig. 7. It shows the P-NLMS and our proposed algorithm both converged at a slower rate than the standard NLMS. This is because both algorithms reached lower MSE in desired perceptual sensitive bands at the expense of higher MSE in other bands as shown in Fig. 6. However, all three algorithms reached about the same level of MSE after they converged.

For listening tests, we extracted mis-adjustment vectors at time instances n_i = {7000, 14000, 21000, 28000, 35000} within the converging period. The preference test results of test (I) and test (II) are shown in Table I and II, which display the number of times (out of ten) each subject reported the residual echo of NLMS or P-NLMS was more perceivable than the echo of the proposed algorithm. From Table I, residual echo of NLMS during n_3~n_4 was considered to be more perceivable than the echo of the proposed algorithm, especially at n_3. These results are very similar to results shown in [5] which compared NLMS with P-NLMS. Indicated by Table II, our proposed algorithm performs better than P-NLMS in reducing echo during n_4~n_5. Besides, these results suggest that residual echoes of all three algorithms have no significant difference after algorithms converge.

### Table I. Preference test results of the test (I).

<table>
<thead>
<tr>
<th>Subject</th>
<th>n_1</th>
<th>n_2</th>
<th>n_3</th>
<th>n_4</th>
<th>n_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>8</td>
<td>9</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>B</td>
<td>7</td>
<td>7</td>
<td>8</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>D</td>
<td>5</td>
<td>8</td>
<td>8</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>E</td>
<td>6</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

### Table II. Preference test results of the test (II).

<table>
<thead>
<tr>
<th>Subject</th>
<th>n_1</th>
<th>n_2</th>
<th>n_3</th>
<th>n_4</th>
<th>n_5</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6</td>
<td>5</td>
<td>7</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>C</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>D</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>E</td>
<td>5</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

### 5. Conclusions and Discussions

In this paper, we extend the P-NLMS AEC [5] by adding a second pre-emphasis filter, which emphasizes the harmonic structure at low frequencies due to the spectral modulation sensitivity of human hearing. Subjective listening test results show the proposed AEC produces less residual echo than P-NLMS AEC within the converging period. Due to the real-time processing limitation, only the spectral modulation sensitivity of human hearing is considered in this work. In the future, we will build an AEC by considering the joint spectro-temporal modulation sensitivity of human hearing [10] just like the way human behaves.

### 6. Acknowledgements

This research is supported by National Science Council, R.O.C. under Grant NSC 101-2220-E-009-065 and the Biomedical Electronics Translational Research Center, National Chiao Tung University.
7. References


