EVALUATIONS OF AN ECHO SUPPRESSOR BASED ON A FREQUENCY-DOMAIN MODEL OF HIGHLY NONLINEAR RESIDUAL ECHO

Osamu Hoshuyama and Akihiko Sugiyama

Media and Information Research Laboratories

NEC Corporation, JAPAN

ABSTRACT

This paper presents evaluations of a nonlinear-echo suppressor based on a frequency-domain model of highly nonlinear residual echo. The residual-echo model, which is based on correlation between spectral amplitudes of residual echo and echo replica, is evaluated by a regression analysis. The results justify the nonlinear-echo suppressor structure. The output signals are subjectively evaluated by mean opinion score (MOS). The score of the nonlinear-echo suppressor is superior to an echo canceller with a linear adaptive filter by 0.8 points on a 5-point scale. In implementation on a DSP system, it is shown that the nonlinear-echo suppressor requires 6.1 to 16.8 MIPS depending on the memory allocation.

1. INTRODUCTION

Acoustic echo cancellation or suppression for hands-free communications with acoustically poor devices such as cellphones and laptop personal computers is a challenging problem. When a speech signal with large power is injected into a small loudspeaker mounted on a small shell, many mechanical contacts in the shell and the loudspeaker itself generate a nonlinear distorted echo [1].

An ordinary echo canceller with a linear adaptive filter can not suppress nonlinear-echo components that may be mixed with the linear echo. A linear adaptive filter models only the linear echo. The remaining nonlinear echo is often one tenth of its linear counterpart or even larger in amplitude. It is audible and degrades the quality of communication.

A practical approach to suppressing uncancelled nonlinear echo is nonlinear post filters [3][4]. Based on psychoacoustical effects, it suppresses uncancelled nonlinear echo as well as ambient noise, when the uncancelled nonlinear echo is sufficiently small compared to the near-end speech. However, the highly nonlinear echo with hands-free cellphone often violates the condition that the uncancelled nonlinear echo is relatively small. When they are designed to suppress highly nonlinear echo, quality of the near-end speech at the output becomes poor.

To suppress the highly nonlinear echo, a novel nonlinear-echo suppressor has been proposed [5]. It is based on a nonlinear-echo model supported by an observation that spectral amplitudes of the residual echo and the echo replica are significantly correlated. This nonlinear-echo suppressor can suppress the highly nonlinear residual echo to an inaudible level even when the distortion is one tenth of the echo. However, the evaluations in [5] are incomplete. The nonlinear-echo model is supported by 1 experimental result. Their output signals are evaluated only objectively in waveforms and spectrograms. Resource requirements such as total computations and memories are not available either. When applied to cellphones, the echo suppressor has to share the limited resources with other applications.

This paper presents detailed evaluations of the nonlinear-echo suppressor based on the spectral correlation between the residual echo and the echo replica. The nonlinear-echo suppressor is reviewed in the next section. Section 3 evaluates the nonlinear-echo model employed in the nonlinear-echo suppressor in various conditions. Section 4 shows subjective evaluation results of the nonlinear-echo suppressor using recorded data. Finally, in Section 5, the nonlinear-echo suppressor is implementated on a DSP (Digital Signal Processor) to assess resource requirements.

2. ECHO SUPPRESSOR BASED ON A FREQUENCY-DOMAIN MODEL OF HIGHLY NONLINUAL RESIDUAL ECHO

Figure 1 illustrates a hands-free system with the nonlinear-echo suppressor based on correlation between nonlinear residual echo and echo replica [5]. The signal at the microphone, $p(k)$, consists of the near-end signal $s(k)$, and the echo $e(k)$, where k is the time index. $e(k)$ contains both the linear and nonlinear components of the echo. In the echo canceller with a linear adaptive filter (EC-LAF), the residual signal $d(k)$ is calculated by subtracting the echo replica $y(k)$, which is generated by LAF from the far-end signal $x(k)$, from $p(k)$. The nonlinear-echo suppressor is based on a frequency-domain model of the nonlinear residual echo.

2.1. Frequency-Domain Model of Highly Nonlinear Residual Echo

The residual signal $d(k)$ after the EC-LAF is expressed as a sum of the near-end signal $s(k)$ and the residual echo $q(k)$ as

$$d(k) \quad = \quad s(k) + q(k). \tag{1}$$

When the EC-LAF cancels the linear echo almost completely, $q(k)$ mainly consists of the nonlinear component of the echo. A frequency-domain representation of $d(k)$ is obtained as

$$\mathbf{D}(m) \quad = \quad \mathbf{S}(m) + \mathbf{Q}(m), \tag{2}$$

where m is the frame index. The vectors $\mathbf{D}(m)$, $\mathbf{S}(m)$, and $\mathbf{Q}(m)$ are frequency domain representations of the signals $d(k)$, $s(k)$, and $q(k)$, respectively. For the i-th frequency bin, (2) becomes

$$D_i(m) \quad = \quad S_i(m) + Q_i(m). \tag{3}$$

The paper [5] models the nonlinear echo $|Q_i(m)|$ as the product of $\widehat{a_i}$ and echo replica $|Y_i(m)|$.

$$|Q_i(m)| \quad \simeq \quad |\widehat{Q_i}(m)| \quad \triangleq \quad \widehat{a_i} \cdot |Y_i(m)|, \tag{4}$$

where $\widehat{a_i}$ is a regression coefficient of $|Q_i(m)|$ and $|Y_i(m)|$. This model is based on an experimental result that $|Q_i(m)|$ and $|Y_i(m)|$ are significantly correlated.
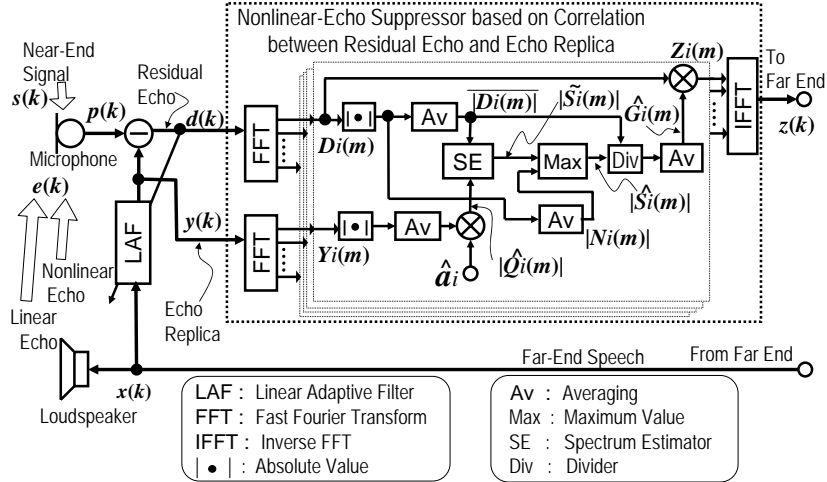
Fig. 1. Hands-Free System with Nonlinear-Echo Suppressor Based on Correlation between Nonlinear Residual Echo and Echo Replica.

2.2. Structure of Nonlinear-Echo Suppressor

Equation (3) can be viewed as an additive model of the nonlinear residual signal, which is widely known in noise suppression. Spectral subtraction [6] can be directly applied to the nonlinear-echo model in (4). However, the nonlinear-echo suppressor applies a "*spectral multiplication*" technique [7] to reduce subjectively annoying musical noise. In the framework of spectral multiplication, the output signal, $|Z_i(m)|$, is obtained as a product of a spectral gain $\widehat{G_i}(m)$ and the residual signal $|D_i(m)|$ as

$$|Z_i(m)| \quad = \quad \widehat{G_i}(m) \cdot |D_i(m)|. \quad (5)$$

$|Z_i(m)|$ is combined with the corresponding phase to reconstruct $Z_i(m)$, which is inversely transformed to the time-domain output signal $z(k)$ [5].

To obtain the spectral gain $\widehat{G_i}(m)$, spectral amplitude of the near-end signal $|S_i(m)|$ is estimated. $|D_i(m)|$ and $|Y_i(m)|$ have almost no cross correlation because they are decorrelated by the EC-LAF. Therefore, squaring and averaging both sides of (3) gives $\overline{|S_i(m)|}^2$ as

$$\overline{|S_i(m)|}^2 \simeq \overline{|D_i(m)|}^2 - \overline{|Q_i(m)|}^2. \quad (6)$$

By taking the square root of (6), and substituting $\overline{|Q_i(m)|}^2$ with $\widehat{a_i}^2 \cdot \overline{|Y_i(m)|}^2$ based on the model in (4), $|\widetilde{S_i}(m)|$, an approximation to $|S_i(m)|$, is obtained as follows.

$$|S_i(m)| \simeq \overline{|S_i(m)|} \quad (7)$$

$$\simeq |\widetilde{S_i}(m)| \triangleq \sqrt{\overline{|D_i(m)|}^2 - \widehat{a_i}^2 \cdot \overline{|Y_i(m)|}^2}. \quad (8)$$

The estimated spectral amplitude $|\widetilde{S_i}(m)|$ usually has a nonnegligible error, because the residual-echo model is an approximation. When the error is large, oversubtraction may occur resulting in attenuation of high frequency components or modulated near-end signal with the far-end signal. Especially, when the near-end signal is stationary like airconditioner noise, the modulation is annoying. To make the modulation artifact less audible, a spectral flooring is applied. The floor value is proportional to the stationary component of the near-end signal. The stationary component $|N_i(m)|$ in Fig. 1 is calculated by an averaging
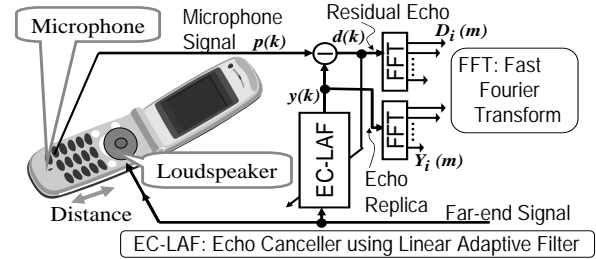

Fig. 2. Experimental Setup.

operation. Finally, the spectral gain $\widehat{G_i}(m)$ is calculated by averaging a ratio of $|\widehat{S_i}(m)|$ to $\overline{|D_i(m)|}$ [5].

3. EVALUATIONS OF NONLINEAR-ECHO MODEL

In order to evaluate the nonlinear-echo model for the nonlinear-echo suppressor, distribution of the residual echo $|D_i(m)|$ and echo replica $|Y_i(m)|$ at various frequencies, and property of the regression coefficients $\widehat{a_i}$ were investigated by a regression analysis. Figure 2 illustrates the experimental setup for the evaluations. In either of the evaluations, a common microphone inside a cellphone shell at the lower end with a different loudspeaker were used. The far-end signals were bandlimited to 0.3–3.4 kHz, and preprocessed by the AMR codec [8] at 12.2 kbps. The length of the signal were from 20 to 40 seconds. As shown in the right half of Fig. 2, a residual echo is obtained as the difference between the microphone signal and the echo replica generated by an adaptive filter. The residual echo and the echo replica were transformed into frequency-domain signals by FFTs (Fast Fourier Transforms) using Hanning window with a frame size M of 160 and a window size L of 256.

3.1. Spectral Correlation at Various Frequencies

For evaluating the distribution of the residual echo $|D_i(m)|$ and echo replica $|Y_i(m)|$, a loudspeaker mounted in the cellphone shell on the lower backside was used. Diaphragm diameter of the loudspeaker was 2.5 cm, and the distance between the loudspeaker and the microphone was 6 cm. The far-end signal was a
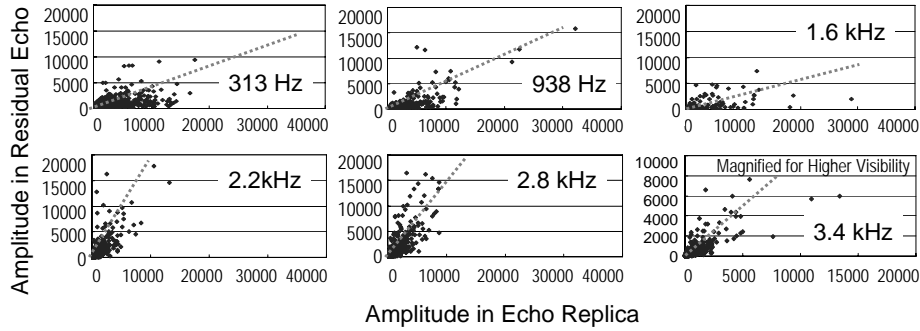
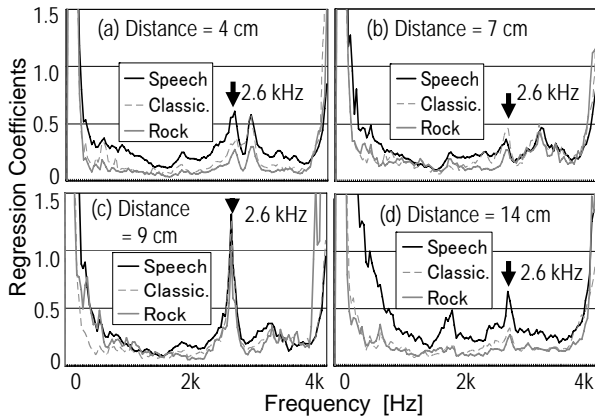**Fig. 3**. Spectral Correlation between Residual Echo and Echo Replica at Various Frequencies.

Fig. 4. Regression Coefficients as a Function of Frequency for Different Distances between Loudspeaker and Microphone, and Various Far-end Signal Categories.

dialogue between a man and a woman.

Figure 3 plots the spectral amplitudes $|Y_i(m)|$ and $|D_i(m)|$ of the echo replica and the residual signal for the same frame-index at various frequencies after convergence of the EC-LAF. Dots in the figure exhibit linear regression that is a sign of significant correlation between the residual echo and the echo replica at all the frequencies. Figure 3 indicates that the nonlinear-echo model represented by (4), where harmonics generated by the nonlinearity are not taken into account, is a good 1st order approximation of the nonlinear residual echo in the frequency domain.

3.2. Regression Analysis

To investigate the variation in the regression coefficients $\hat{a}_i$, a regression analysis was performed for different loudspeaker positions and far-end signals representing different characteristics. An independent sealed enclosure with a loudspeaker was attached to the surface of the cellphone shell by scotch tape to change the distance between the loudspeaker and the microphone. Diaphragm diameter of the loudspeaker was 1.2 cm, and the distances were set to 4, 7, 9, and 14 cm. To reduce the enclosure nonlinearity, which should not be the dominant nonlinearity in the experiment, a big enclosure with a volume of $110 cm^3$ was used. The far-end signals were a dialogue between a man and a woman, classical music (brass), and rock music. The lengths

of the signals were from 20 to 40 seconds. After convergence of the EC-LAF, $\hat{a}_i$, the regression coefficients between the spectral amplitudes of the residual echo and the echo replica, were calculated.

Regression coefficients as a function of frequency are shown in Fig. 4 (a), (b), (c), and (d) for different loudspeaker-microphone distances. A larger regression coefficient means higher nonlinearity. The regression coefficients at low-end and high-end frequencies are large for all the curves in Fig. 4. It is because the loudspeaker has poor response at those frequencies and causes high distortion.

By comparing all the curves in Fig. 4, it can be seen that the influence of the loudspeaker-microphone distance is larger than that of the far-end signal. All the curves have a peak at 2.6 kHz due to resonance of the cellphone shell. However, the height of the resonance peak basically depends more on the distance than on the far-end-signal characteristics. The speech and music signals have significantly different spectra. However, the shapes of the corresponding curves in Fig. 4 are similar in each graph. The variation, which is caused by a model error, in the regression coefficients is smaller than 0.3 in the frequency range from 0.5 to 3.5 kHz. Artifact by the model error is already compensated by the spectral multiplication and the spectral flooring. The variation of 0.3 is sufficiently small for the nonlinear-echo suppressor to use a common set of regression coefficients for suppressing nonlinear echo generated from various far-end signals.

4. SUBJECTIVE EVALUATION OF THE OUTPUT SIGNAL

A subjective evaluation was performed with 5 sets of recorded data obtained in quiet and noisy environments using a cellphone handset with a folding shell. A 1-inch loudspeaker was mounted in the cellphone on the lower backside. The distance between the loudspeaker and the microphone was approximately 6 cm. Loudness of the loudspeaker was set so that the echo level at the microphone is comparable to the near-end speech. In order to demonstrate the robustness of $\hat{a}_i$, a set of $\hat{a}_i$ for a female speaker was used, though the far-end talkers in the evaluation included a male speaker. All other parameters are the same as those in [5]

The output signals of the nonlinear-echo suppressor were evaluated by 5-grade mean opinion score (MOS) with headphone listening by ten nonprofessional subjects. As anchors, the near-end signal with nonlinear echo was included for grade 1, and the original near-end signal without echo for grade 5. Subjects were instructed with examples that there might be attenuation of
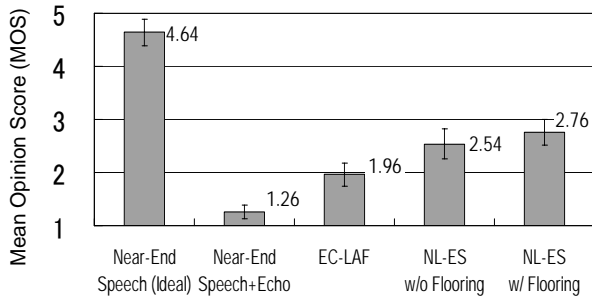
Fig. 5. Subjective Test Results.

Fig. 6. Computational Loads for Different Memory Allocations.

high frequency components or near-end signal modulation during double-talk periods.

Evaluation results are shown in Fig. 5. The number beside each bar represents the score obtained by the corresponding method. The vertical line on each bar indicates the 95% confidence interval. The EC-LAF obtained 1.96 points because the residual echo is still audible. The nonlinear-echo suppressor without flooring (NL-ES w/o flooring) suppresses the nonlinear residual echo almost completely. However, the near-end signal modulation is serious, thus, it scored 2.54 points. Its confidence interval has no overlap with that of the EC-LAF, which means that the nonlinear-echo suppressor has significant improvement in the output signal quality. The nonlinear-echo suppressor with flooring (NL-ES w/ flooring) obtained 2.76 points, which is the highest of all the methods and 0.8 point higher than that of the EC-LAF. Its confidence interval overlaps with that of the nonlinear-echo suppressor without flooring. However, it is likely that another subjective evaluation with more subjects will make the confidence intervals shorter, resulting in separated ones between the nonlinear echo suppressors with and without flooring.

5. DSP IMPLEMENTATION

In order to evaluate the resource requirements, the nonlinear-echo suppressor with an EC-LAF was implemented on a DSP starter kit (DSK) of TMS320C6416T running at 1 GHz [9]. The programming was carried out in C language with a compiler provided by Texas Instruments. In the implementation, three typical memory allocations were compared to assess the trade-off between the computational load and the internal memory size. Figure 6 shows computational loads for different memory allocations.

When both program codes and data are allocated to internal memory (SRAM: Synchronized RAM), the computations including EC-LAF are minimized to 6.1 MIPS (Million Instructions Per Second), although total usage of internal memory is 88 kBytes. When all the memory are allocated to external memory (SDRAM: Synchronized Dynamic RAM), which is the worst case, the total computations are 16.8 MIPS. In case 32 kBytes of fast cash memory (L2 cash) is available on the internal memory, even when only the external memory is used, the total computations are as small as 9.2 MIPS.

Hands-free communication test were performed using the DSK with the set of loudspeaker and microphone on the real cellphone mockup in Section 3. The users' comments were positive and agreed with the subjective evaluation results shown in Section 4. Echo was sufficiently small for conversation, and the degradation of the near-end signal was acceptable even in double-talk periods.
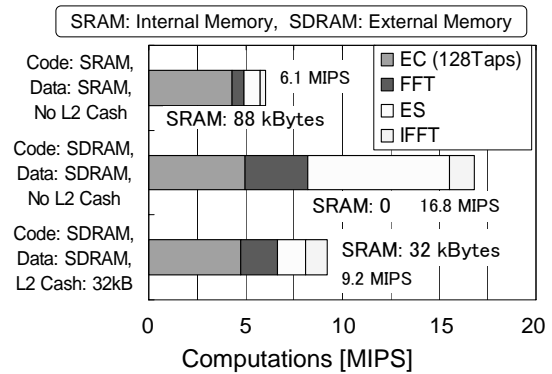
6. CONCLUSIONS

Evaluations of a nonlinear-echo suppressor based on a frequency-domain model of highly nonlinear residual echo has been presented. The residual-echo model, which is based on correlation between spectral amplitudes of residual echo and echo replica, was evaluated by a regression analysis. The results justify the nonlinear-echo suppressor structure. The subjective evaluation has shown that the MOS score of the nonlinear-echo suppressor with flooring is superior to an echo canceller with linear adaptive filter by 0.8 points on a 5-point scale. In implementation on a DSP system, it has been shown that the nonlinear-echo suppressor requires 6.1 to 16.8 MIPS depending on the memory allocation.

7. REFERENCES

[1] A. N. Birkett and R. A. Goubran, "Limitations of Hands-free Acoustic Echo Cancellers Due to Nonlinear Loudspeaker Distortion and Enclosure Vibration Effects," IEEE Proc. WASPAA'95, pp. 13–16, 1995.

[2] F. Kuech, A. Mitnacht, and W. Kellerman, "Nonlinear Acoustic Echo Cancellation Using Adaptive Orthogonalized Power Filters," IEEE Proc. ICASSP2005, Vol. III-105–108, 2005.

[3] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction," IEEE Trans. SAP, pp. 245–256, 2002.

[4] A. S. Chhetri, A. C. Surendran, J. W. Stokes, and J. C. Platt, "Regression-Based Residual Acoustic Echo Suppression," Proc. IWAENC2005, Sep. 2005.

[5] O. Hoshuyama, and A. Sugiyama, "An Acoustic Echo Suppressor Based on a Frequency-Domain Model of Highly Nonlinear Residual Echo," IEEE Proc. ICASSP2006, pp. V-269–272, 2006.

[6] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, pp. 113–120, 1979.

[7] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. ASSP, pp. 1109–1121, 1984.

[8] 3GPP TS26.90, "Adaptive Multi-rate Speech Codec; Transcode Functions," Mar. 2001.

[9] Spectrum Digital Inc., "TMS320C6416T DSK Technical Reference", Nov. 2004.