AN ACOUSTIC 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 proposes a new acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo. The proposed echo suppressor controls the gain in each frequency bin based on the new model where the highly nonlinear residual echo is approximated as a product of a regression coefficient and the echo replica in the frequency domain. To reduce annoying modulation by the error of the model, a flooring operation of estimated near-end signal level is introduced to the gain control. Simulation results with speech data recorded by a hands-free cellphone show that the proposed echo suppressor reduces the highly nonlinear residual echo to an almost inaudible level.

1. INTRODUCTION

Acoustic echo cancellation or suppression for hands-free cellphones is a challenging topic. One of the biggest problems is nonlinearity of the echo-path such as loudspeaker distortion and vibrations of the cellphone shell [1]. When a speech signal with large power is injected into a small loudspeaker, the loudspeaker itself and many mechanical contacts in the shell generate distorted echo. Among the sources of nonlinearity, mechanical nonlinearity with shell vibrations is said to be the dominant factor [2].

An ordinary echo canceller with a linear adaptive filter is used to eliminate linear echo. However, it 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 one tenth of its linear counterpart or even larger in amplitude. It is audible and degrades the quality of communication. A common solution to such a significant nonlinear echo is to mute the whole residual signal obtained at the output of the echo canceller. However, it often causes discontinuous speech during double talk periods.

Use of a nonlinear adaptive filter is also a popular solution to the nonlinear echo problem. Volterra adaptive filters fit the nonlinear echo-path model and can theoretically cancel nonlinear echoes [3]. Their drawbacks are heavy computational load and slow convergence. In the case of a cellphone handset, it is not possible for Volterra filters to track fast and frequent changes of the echo path. Simplified nonlinear adaptive filters that have fast tracking capability with reasonable computations have also been proposed (e.g. [4]). Unfortunately, the improvement by these simplified filters is limited to as much as 5 dB.

Single-input post filter is another approach to suppressing uncancelled nonlinear echo as well as ambient noise. When the uncancelled nonlinear echo is sufficiently small, a single-input post filter is applicable to suppressing the uncancelled nonlinear echo [5, 6]. However, they are not useful for high nonlinearity often encountered in hands-free cellphone handsets. The uncancelled nonlinear echo violates the condition that it is relatively small compared to those of the near-end speech.

This paper proposes a new echo suppressor based on a new frequency-domain model of highly nonlinear residual echo. In the next section, a new residual-echo model based on the spectral correlation between the residual echo and the echo replica is developed. An echo suppressor based on the new model is proposed in Section 3. Evaluation results using a hands-free cellphone handset are presented in Section 4.

2. FREQUENCY-DOMAIN MODEL OF HIGHLY NONLINEAR RESIDUAL ECHO

The signal at the microphone, $p(k)$, consists of the near-end signal $s(k)$, and the echo signal $e(k)$.

$$p(k) \quad = \quad s(k) + e(k) \tag{1}$$

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 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

$$d(k) \quad = \quad s(k) + q(k). \tag{2}$$

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 (2) is obtained as

$$\mathbf{D}(m) = \mathbf{S}(m) + \mathbf{Q}(m), \tag{3}$$

$$\mathbf{D}(m) = [D_0(m)\, D_1(m)\, \ldots\, D_{L-1}(m)]^T \triangleq \mathrm{FFT}[\mathbf{d}(mM)] \tag{4}$$

$$\mathbf{S}(m) = [S_0(m)\, S_1(m)\, \ldots\, S_{L-1}(m)]^T \triangleq \mathrm{FFT}[\mathbf{s}(mM)] \tag{5}$$

$$\mathbf{Q}(m) = [Q_0(m)\, Q_1(m)\, \ldots\, Q_{L-1}(m)]^T \triangleq \mathrm{FFT}[\mathbf{q}(mM)] \tag{6}$$

where $\mathrm{FFT}[\,\cdot\,]$ is a windowed fast Fourier transform with an overlap. m, M, and L represents the frame index, the frame size, and the window size. The residual-signal vector $\mathbf{d}(k)$, the near-end speech vector $\mathbf{s}(k)$, and the residual-echo vector $\mathbf{q}(k)$ are defined by

$$\begin{aligned}\mathbf{d}(k) &\triangleq [d(k)\, d(k-1)\, \ldots\, d(k-L+1)]^T,\\ \mathbf{s}(k) &\triangleq [s(k)\, s(k-1)\, \ldots\, s(k-L+1)]^T,\\ \mathbf{q}(k) &\triangleq [q(k)\, q(k-1)\, \ldots\, q(k-L+1)]^T.\end{aligned}$$

For the i-th frequency bin, (3) becomes

$$D_i(m) \quad = \quad S_i(m) + Q_i(m). \tag{7}$$

2.1. Spectral Correlation Between Residual Echo and Echo Replica

The highly nonlinear residual echo and the echo replica of an EC-LAF under the single-talk condition were investigated with a signal bandwidth of 4 kHz. A frequency-domain representation $\mathbf{Y}(m)$ of the echo replica $y(k)$ is obtained by

$$\mathbf{Y}(m) = [Y_0(m)\, Y_1(m)\, \ldots\, Y_{L-1}(m)]^T \triangleq \mathrm{FFT}[\mathbf{y}(mM)] \tag{8}$$

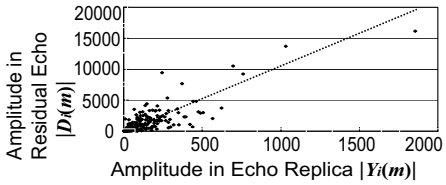

Fig. 1. Spectral Correlation between Residual Echo and Echo Replica at 2.2 kHz.

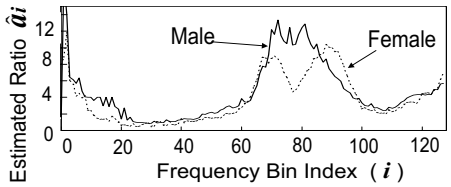

Fig. 2. Estimated Ratio $\widehat{a_i}$.

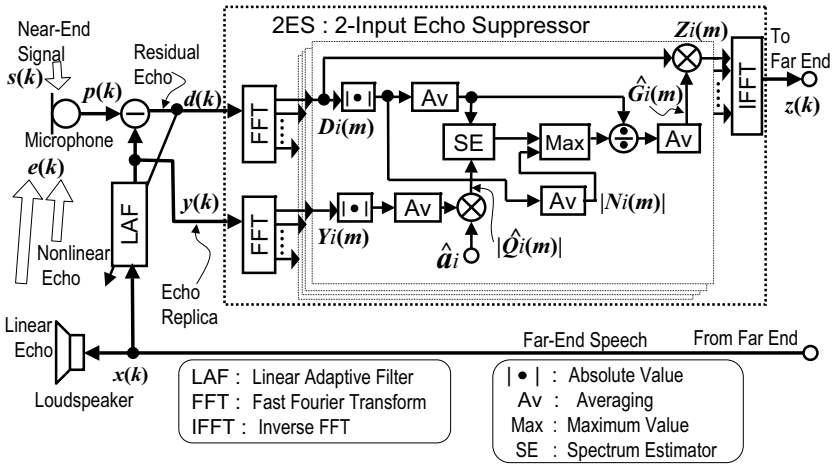

Fig. 3. Proposed Echo Suppressor.

where $\mathbf{y}(k) \triangleq [y(k)\, y(k-1)\, \ldots\, y(k-L+1)]^T$. The spectral amplitudes $|Y_i(m)|$ and $|D_i(m)|$ of the echo replica and the residual signal were obtained by the Fourier transform with a frame size M of 160 and a window size L of 256.

Figure 1 shows the relationship between the short-term average of $|Y_i(m)|$ and that of $|D_i(m)|$ at 2.2 kHz for the same frame-index. Dots in the figure exhibit linear regression that is a sign of significant correlation between the residual echo and the echo replica. Of course, distribution of the dots varies with the spectrum and level of the far-end signal. However, it stays in a limited range for various types of speech and music [7].

2.2. Model of Residual Echo Based on Spectral Correlation

Based on Fig. 1, a ratio of $|Q_i(m)|$ to $|Y_i(m)|$ is defined as $a_i(m)$. Using $a_i(m)$, a residual echo component $|Q_i(m)|$ is given by

$$|Q_i(m)| \quad = \quad a_i(m) \cdot |Y_i(m)|. \tag{9}$$

Equation (9) suggests that $|Q_i(m)|$ can be obtained from the echo replica $|Y_i(m)|$ if $a_i(m)$ is available. However, in reality, it is not the case. The ratio $a_i(m)$ should somehow be estimated.

Let us consider approximating $a_i(m)$ by $\widehat{a_i}$ using averaged absolute values of the residual echo and the echo replica. Then,

$$\widehat{a_i} \quad = \quad \frac{\overline{|Q_i(m)|}}{\overline{|Y_i(m)|}} \tag{10}$$

where overline $\overline{\cdot}$ means an averaging operation. When there is no near-end speech, *i.e.* $S_i(m) = 0$, (7) reduces to

$$D_i(m)_{\textit{Single talk}} \quad = \quad Q_i(m). \tag{11}$$

From (10) and (11), $\widehat{a_i}$ is given by

$$\widehat{a_i} \quad = \quad \frac{\overline{|D_i(m)|_{\textit{Single talk}}}}{\overline{|Y_i(m)|}}. \tag{12}$$

The estimated ratio $\widehat{a_i}$ can be obtained by advance experimental measurements in quiet environments, which are needed once for each set up of the loudspeaker and the microphone. Figure 2 shows $\widehat{a_i}$ as a function of frequency-bin index i for a male and a female speech. The difference of the curves is small enough to utilize one curve for the other.

By approximating $a_i(m)$ in (9) with a regression coefficient $\widehat{a_i}$, the residual echo $|Q_i(m)|$ is modeled as the product of $\widehat{a_i}$ and $|Y_i(m)|$.

$$|Q_i(m)| \quad \simeq \quad \widehat{|Q_i(m)|} \quad \triangleq \quad \widehat{a_i} \cdot |Y_i(m)|. \tag{13}$$

3. PROPOSED ECHO SUPPRESSOR

Equation (7) can be viewed as an additive model of the residual signal, which is widely known in noise suppression. The near-end signal corresponds to the speech to be enhanced and the nonlinear echo, to the additive noise. Spectral subtraction [8] is a popular technique in noise suppressors and can be directly applied to the nonlinear-echo model in (13). However, minimum mean-square error short-time spectral amplitude estimation (MMSE-STSA) [9] is adopted to reduce subjectively annoying musical noise as in [10]. Figure 3 shows the structure of the proposed 2-input echo suppressor (2ES) combined with an EC-LAF.

In the framework of MMSE-STSA, 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

$$|Z_i(m)| \quad = \quad \widehat{G_i}(m) \cdot |D_i(m)|. \tag{14}$$

$|Z_i(m)|$ is combined with $\angle D_i(m)$ to reconstruct $Z_i(m)$ as

$$Z_i(m) = |Z_i(m)| \cdot \exp(j \angle D_i(m)) \quad (i = 0, 1, \ldots, L-1). \tag{15}$$

In (15), $\angle D_i(m)$ is not modified at all because the phase is not important for speech intelligibility [9, 10].

The output signals $z(k)$ in the time domain are obtained as elements of the segmented frame which is reconstructed by the inverse Fourier transform as follows.

$$\mathbf{z}(k) \triangleq [z(k)\, z(k-1)\, \ldots\, z(k-L+1)]^T = \text{IFFT}[\mathbf{Z}(m)] \tag{16}$$

$$\mathbf{Z}(m) \triangleq [Z_0(m)\, Z_1(m)\, \ldots\, Z_{L-1}(m)]^T \tag{17}$$

where IFFT$[\,\cdot\,]$ is a windowed inverse FFT with overlap-add operations.

3.1. Estimation of Near-End Speech

The spectral gain $\widehat{G_i}(m)$ in (14) is ideally equal to the ratio of $|S_i(m)|$ to $|D_i(m)|$. However, it is not available because $|S_i(m)|$ is the ideal output. To obtain $\widehat{G_i}(m)$, spectral amplitude of the near-end signal $|S_i(m)|$ is estimated. $S_i(m)$ has almost no correlation with $Q_i(m)$ because they are independent. Therefore, squaring and averaging both sides of (7) gives

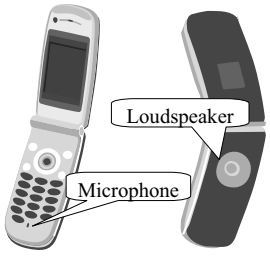

Fig. 4. Locations of Microphone and Loudspeaker on Hands-Free Cellphone.

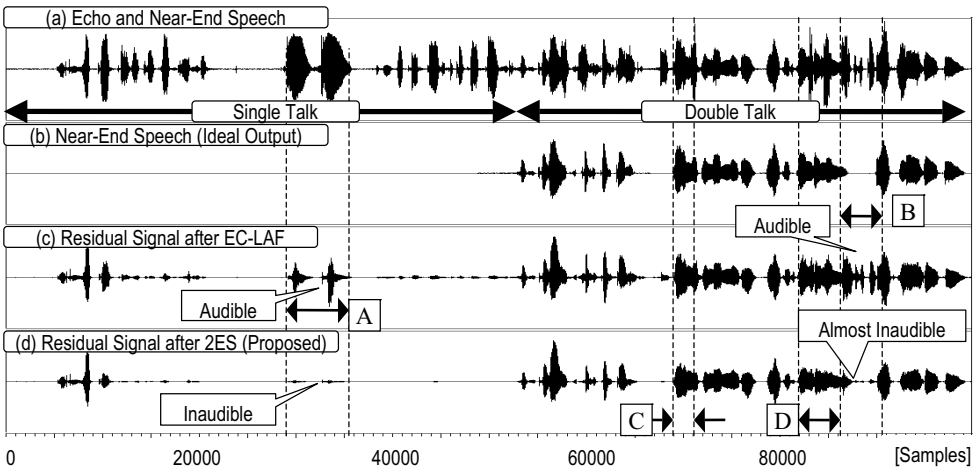

Fig. 5. Input and Output Signals in Quiet Environment.

$$\overline{|D_i(m)|}^2 \simeq \overline{|S_i(m)|}^2 + \overline{|Q_i(m)|}^2. \quad (18)$$

Hence, $\overline{|S_i(m)|}^2$ is obtained as

$$\overline{|S_i(m)|}^2 \simeq \overline{|D_i(m)|}^2 - \overline{|Q_i(m)|}^2. \quad (19)$$

By substituting $|Q_i(m)|$ with $\widehat{a_i}\cdot|Y_i(m)|$ based on (13), (19) becomes

$$\overline{|S_i(m)|}^2 \simeq \overline{|D_i(m)|}^2 - \widehat{a_i}^2 \cdot \overline{|Y_i(m)|}^2. \quad (20)$$

The square root of (20) gives $|\tilde{S}_i(m)|$, which is an approximation to $|S_i(m)|$, as follows.

$$|S_i(m)| \simeq \sqrt{\overline{|S_i(m)|}^2} \quad (21)$$

$$\simeq \sqrt{\overline{|D_i(m)|}^2 - \widehat{a_i}^2 \cdot \overline{|Y_i(m)|}^2} \triangleq |\tilde{S}_i(m)|. \quad (22)$$

$\overline{|D_i(m)|}$ and $\overline{|Y_i(m)|}$ are recursively calculated by two averaging operations as follows.

$$\overline{|D_i(m)|} = \begin{cases} \beta_{DA}\,|D_i(m)| + (1-\beta_{DA})\,\overline{|D_i(m-1)|}, \\ \quad |D_i(m)| > \overline{|D_i(m-1)|} \\ \beta_{DD}\,|D_i(m)| + (1-\beta_{DD})\,\overline{|D_i(m-1)|}, \\ \quad \text{otherwise} \end{cases} \quad (23)$$

$$\overline{|Y_i(m)|} = \begin{cases} \beta_{YA}\,|Y_i(m)| + (1-\beta_{YA})\,\overline{|Y_i(m-1)|}, \\ \quad |Y_i(m)| > \overline{|Y_i(m-1)|} \\ \beta_{YD}\,|Y_i(m)| + (1-\beta_{YD})\,\overline{|Y_i(m-1)|}, \\ \quad \text{otherwise} \end{cases} \quad (24)$$

where β_{DA}, β_{DD}, β_{YA}, and β_{YD} are constants to determine the time-constant of the averages ($0 < \beta_{DD} < \beta_{DA} \leq 1$ and $0 < \beta_{YD} < \beta_{YA} \leq 1$). Each averaging operation has two averaging constants for superior tracking capability typically represented by "fast attack and slow decay."

The estimated spectral amplitude $|\tilde{S}_i(m)|$ has usually an error, because the model of the residual echo is not precise. When the error is large, oversubtraction may occurs, which brings modulations by the far-end signal to the near-end signal. When the near-end signal is nonstationary like speech, most of the modulation effects are masked by the nonstationarity. However, when the near-end signal is stationary like airconditioner noise, the modulations are perceived as fluctuations.

To reduce the modulations, a spectral flooring [11] is introduced in the proposed echo suppressor. The floor value is proportional to the stationary component of the near-end signal. The stationary component $|N_i(m)|$ is calculated from $\overline{|D_i(m)|}$ by an averaging operation with a very-slow-attack-and-fast-decay response. An improved spectral amplitude $|\widehat{S_i(m)}|$ is obtained by the flooring operation as follows.

$$|\widehat{S_i(m)}| \triangleq \max\left(\gamma_D|N_i(m)|\,,\,|\tilde{S}_i(m)|\right) \quad (25)$$

where γ_D is a gain parameter for the flooring operation, and an operator $\max(\cdot,\cdot)$ stands for the maximum of the two arguments. $|N_i(m)|$ is calculated as follows.

$$|N_i(m)| = \begin{cases} \beta_{NA}\,\overline{|D_i(m)|} + (1-\beta_{NA})\,|N_i(m-1)|, \\ \quad \overline{|D_i(m)|} > |N_i(m-1)| \\ \beta_{ND}\,\overline{|D_i(m)|} + (1-\beta_{ND})\,|N_i(m-1)|, \\ \quad \text{otherwise} \end{cases} \quad (26)$$

where β_{NA} and β_{ND} are averaging constants satisfying $0 < \beta_{NA} < \beta_{ND} \leq 1$.

3.2. Spectral Gain Control

The spectral gain $\widehat{G_i}(m)$ is calculated by smoothing a ratio of $|\widehat{S_i(m)}|$ to $\overline{|D_i(m)|}$ as follows.

$$\widehat{G_i}(m) = \begin{cases} \beta_{GA}\,\tilde{G}_i(m) + (1-\beta_{GA})\,\widehat{G_i}(m-1), \\ \quad G_i(m) > \widehat{G_i}(m-1) \\ \beta_{GD}\,\tilde{G}_i(m) + (1-\beta_{GD})\,\widehat{G_i}(m-1), \\ \quad \text{otherwise} \end{cases} \quad (27)$$

$$\tilde{G}_i(m) = \frac{|\widehat{S_i(m)}|}{\overline{|D_i(m)|} + \delta_G} \quad (28)$$

where $\tilde{G}_i(m)$ is a temporary variable for calculating $\widehat{G_i}(m)$, and δ_G is a positive constant for avoiding zero division. The averaging constants β_{GA} and β_{GD} correspond to β_{DA} and β_{DD} in (23) satisfying $0 < \beta_{GD} < \beta_{GA} \leq 1$.

4. EVALUATIONS

Simulations of quiet and noisy environments were performed with recorded data using a folding hands-free cellphone with a 1-inch loudspeaker mounted at a lower backside as shown in Fig. 4. The

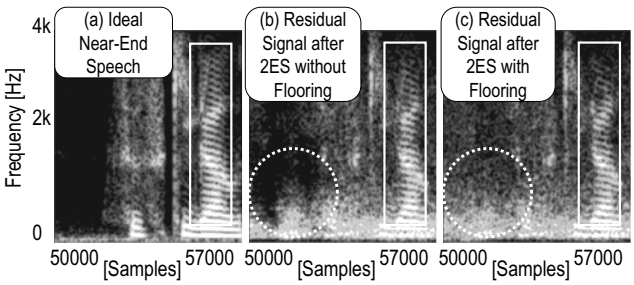

Fig. 6. Spectrograms in Noisy Environment.

distance between the loudspeaker and the microphone was approximately 6 cm. Loudness of the far-end speech was set so that the echo level at the microphone is comparable to the near-end speech.

For the EC-LAF, the number of filter coefficients was 128 and the coefficients were updated by the normalized least-mean-square algorithm with a double-talk control. The 2ES employed the window size L of 256, and the frame size M of 160 following the AMR codec standard. Hanning window was used as the windowing function. Constants for averaging operations were set to β_{DA}=β_{YA}=1.0, β_{DD}=β_{YD}=0.8, β_{NA}=0.001, β_{ND}=0.2, β_{GA}=0.8, and β_{GD}=0.3. The gain parameter γ_D for the flooring operation was 1.0. As a set of $\widehat{a_i}$, the dashed curve for a female in Fig. 2 was used, though the far-end talker in the evaluations was a male to demonstrate the robustness of $\widehat{a_i}$.

4.1. Quiet Environment

Figure 5 depicts simulation results under the quiet environment, where only near-end speech and far-end speech exist and there is no noise. The curve in (a) represents the echo and the near-end signal. (b) is the near-end signal, which is the ideal output signal. (c) and (d) are the output signals of the EC-LAF and the proposed echo suppressor, *i.e.* 2ES.

In the left half of Fig. 5, a single talk condition is implemented. The amplitude in (c) is one tenth or more of that in (a), which indicates that the EC-LAF canceled the echo only by 15 to 20 dB due to the distortion of the echo path. Referring to section A in (c), because distortion of the echo path is high, the residual echo of the EC-LAF is audible. On the other hand, the residual echo of the 2ES is inaudible as shown in (d).

In the right half of Fig. 5, near-end speech signals were added to the echo to implement a double-talk condition. Referring to section B in (b), the ideal output should be zero, that is achieved in (d) by the 2ES. The residual echo of the 2ES is almost inaudible as shown in the figure. However, the EC-LAF fails to cancel the distorted echo, whose residual echo is depicted in section B in (c). The residual echo after the EC-LAF is clearly audible.

It is also seen that the output signal of the 2ES, (d), exhibits some difference from the ideal near-end speech (b), in double-talk periods as sections C and D. Such a difference is mainly caused by attenuation at high frequencies. Despite the degradation, the quality of the output (d) in Fig. 5 was sufficiently good for speech communication on cellphones. In other environments with different loudness conditions and noise levels, the subjective qualities of the near-end speech signals at the output were also sufficiently good.

4.2. Noisy Environment

To demonstrate the effectiveness of the flooring operation, simulations of a noisy environment were also performed. A station noise was added to the signal in Fig. 5(a) as another near-end signal. Figure 6 shows spectrograms of the ideal near-end speech and the output signals of the 2ESs with and without the spectral flooring. Both the 2ESs caused attenuation in the middle of the figures at high frequencies. However, as shown in the rectangular boxes, they keep the harmonic structure of the near-end speech. The 2ES without the spectral flooring caused oversubtraction as depicted in Fig. 6(b) as a dark spot in the dotted circle. On the contrary, the 2ES with the spectral flooring preserves uniformity of the background noise as shown in (c), in which the dark spot in (b) disappeared. The difference is perceived as less modulated background noise.

5. CONCLUSIONS

This paper has proposed a new acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo. The proposed echo suppressor controls the gain in each frequency bin based on the new model where the highly nonlinear residual echo is approximated as a product of a regression coefficient and the echo replica in the frequency domain. To reduce annoying modulation by the error of the model, a flooring operation of estimated near-end signal level is introduced to the gain control. Simulation results with speech data recorded by a hands-free cellphone have shown that the proposed echo suppressor reduces the highly nonlinear residual echo to an almost inaudible level, and preserves uniformity of the background noise. The output signal quality is sufficiently good for speech communication on cellphones.

6. REFERENCES

[1] M. E. Knappe and R. A. Goubran, "Steady-State Performance Limitations of Full-Band Acoustic Echo Cancellers," IEEE Proc. ICASSP'94, pp. IV-81–84, 1994.

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

[3] A. J. M. Kaizer, "Modeling the Nonlinear Response of an Electrodynamic Loudspeaker by a Volterra Series Expansion," J. AES, Vol. 35, No. 6, 1997.

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

[5] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A Psychoacoustic Approach to Combined Acoustic Echo Cancellation and Noise Reduction," IEEE Trans. SAP, Vol. 10, No. 5, pp. 245 - 256, 2002.

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

[7] O. Hoshuyama and A. Sugiyama, "Regression Analysis of Residual Echo and Echo Replica for a Nonlinear-Echo Suppressor," to appear in Proc. IEICE General Conf. Mar. 2006.

[8] S. F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction," IEEE Trans. ASSP, Vol. ASSP-27, No. 2, pp. 113–120, Apr. 1979.

[9] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," IEEE Trans. ASSP, Vol. ASSP-32, No. 6, pp. 1109–1121, 1984.

[10] M. Kato, A. Sugiyama, and M. Serizawa, "A family of 3GPP-Standard Noise Suppressors for the AMR Codec and the Evaluation Results," IEEE Proc. ICASSP2003, pp. I-916–919, 2003.

[11] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of Speech Corrupted by Acoustic Noise," Proc. ICASSP'79, pp. 208-211, 1979.