TOWARDS ROBUST ACOUSTIC ECHO CANCELLATION DURING DOUBLE-TALK AND NEAR-END BACKGROUND NOISE VIA ENHANCEMENT OF RESIDUAL ECHO

Ted S. Wada, Biing-Hwang (Fred) Juang

Center for Signal and Image Processing
Georgia Institute of Technology, 75 Fifth Street NW, Atlanta, GA 30308, USA
{twada,juang}@ece.gatech.edu

ABSTRACT

This paper examines the technique of using a noise suppressing nonlinearity in the adaptive filter error feedback loop of the acoustic echo canceler (AEC) based on the least mean square (LMS) algorithm when there are both double-talk and white background noise at the near-end. By combining the previously introduced noise suppressing technique with a compressive nonlinearity derived from the theory of robust statistics, consistently better results are obtained during double-talk as well as during single-talk when compared to the traditional approach of using only the compressive nonlinearity. It is shown that a compressive form of noise reducing nonlinearity can be derived also from the signal enhancement point of view when the noise probability density (pdf) is tailed more heavily and has a higher kurtosis than the Gaussian pdf. A combination of such a noise compressing nonlinearity and a noise suppressing nonlinearity is capable of producing results that are similar to that of the robust statistics approach during double-talk along with an added benefit of increased robustness during single-talk when there is only the background noise.

Index Terms—
acoustic echo cancellation, signal enhancement, robust statistics, robust adaptive filtering, nonlinear processing

1. INTRODUCTION

It was shown in [1, 2] that both the misalignment and the echo return loss enhancement (ERLE) from using the time-domain or the frequency-domain acoustic echo canceler (AEC) based on the least mean square (LMS) adaptive algorithm can be improved through the filter error enhancement procedure when there is either a *linear* distortion in the form of additive noise or a *nonlinear* distortion in the form of speech coding at the near-end. The enhancement is performed through the application of a noise suppressing nonlinearity to the estimation error before the error is reused for weights adaptation, as represented by the error suppression nonlinearity (ESN) in Figure 1. It was also shown in [1] that such a procedure is optimal in terms of the steady-state mean-square error (MSE) [3] or the mean-square deviation (MSD) [4] when right conditions are met. The results are consistent with the notion that reducing the distortion that may be present in the filter error enables an adaptive filter to better estimate the linear part of the system response.

The additive near-end noise during the AEC can be grouped into two major types. One is the background noise, such as air conditioner or car engine noise, that can often be ubiquitous and continuously present. In such a case, an adaptive step-size procedure can be used to scale down the step-size when the near-end signal-to-noise ratio (SNR) is low in order to avoid the filter divergence. The other type of noise that can be more troublesome than the background noise is the near-end speech, which is referred as the *double-talk* when the far-end speaker is concurrently talking. The double-talk can greatly disrupt the filter adaptation since it is usually larger in volume than the acoustic echo and is highly colored and non-stationary. The traditional approach is to stop the adaptation entirely during double-talk by using a double-talk detector (DTD). A more advanced approach is to use a *compressive* nonlinearity to limit the sudden outliers in the filter error that may leak through when a DTD fails to detect the event [5]. This technique is akin to removing an impulsive noise from the corrupted signal to get back the signal of interest and is very similar in goal to the ESN approach.

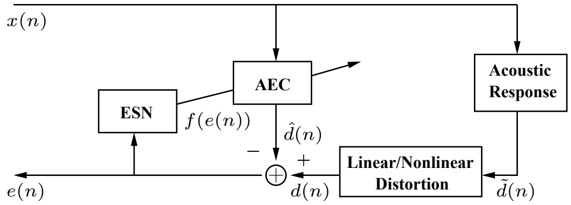

Fig. 1. AEC scheme with error suppression nonlinearity (ESN) for robustness against linear or nonlinear distortion to acoustic echo.

In this paper, the performances of the ESN in [1] and the compressive nonlinearity in [5] are evaluated in a simulated acoustic environment with the double-talk and the near-end white background present in the echo path. It is shown that a compressive form of the ESN can be derived from the signal enhancement point of view and that it is related in functionality to the compressive nonlinearity in [5]. It is also shown that each type of nonlinearity has its own merit and that the overall AEC performance can be improved by combining a noise *suppressing* nonlinearity and a noise *compressing* nonlinearity together to make the filter adaptation process less susceptible to additive distortions. The new results further support the fact that the filter error enhancement strategy can make an adaptive filter more robust to many types of disruption to the echo path.

The rest of the paper is organized as follows. First, noise reducing nonlinearities are derived through the minimum mean-square error (MMSE) approach. Second, a compressive nonlinearity derived from the robust statistics theory is presented and related to the MMSE noise compressing nonlinearity. Third, testing methods for verifying the performance of the nonlinearities are described, followed by simulation results. Finally, conclusion is given at the end.

This work is supported in part by the National Science Foundation Award IIS-0534221.

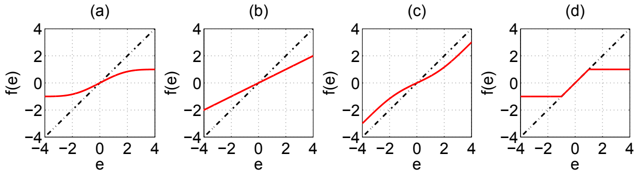

Fig. 2. Noise reducing nonlinearities: (a) for Gaussian $\tilde{e}$ and Laplacian v, (b) for Gaussian $\tilde{e}$ and v, (c) for Laplacian $\tilde{e}$ and Gaussian v, (d) for impulsive noise derived from the robust statistics theory.

2. NOISE REDUCING NONLINEARITIES

Let the noisy filter error e be modeled additively as $e = \tilde{e} + v$, where $\tilde{e}$ is the original filter error and v is the noise. Then there are three distinctive cases obtained from the MMSE estimation of $\tilde{e}$ by taking the conditional expectation $E[\tilde{e}|e]$ as follows.

2.1. Gaussian filter error and any noise distributions

If $\tilde{e}$ is zero-mean Gaussian distributed with the variance $\sigma_{\tilde{e}}^2$, then the MMSE estimate of e for any noise v is given by [1]

$$f_{MMSE}^{GA}(e) = \int_{-\infty}^{\infty} \tilde{e}\, p_{\tilde{e}|e}(\tilde{e}|e) d\tilde{e} = \frac{\int_{-\infty}^{\infty} \tilde{e}\, p_{e|\tilde{e}}(e|\tilde{e}) p_{\tilde{e}}(\tilde{e}) d\tilde{e}}{\int_{-\infty}^{\infty} p_{e|\tilde{e}}(e|\tilde{e}) p_{\tilde{e}}(\tilde{e}) d\tilde{e}}$$
$$= \frac{\int_{-\infty}^{\infty} \tilde{e}\, p_v(e-\tilde{e}) p_{\tilde{e}}(\tilde{e}) d\tilde{e}}{\int_{-\infty}^{\infty} p_v(e-\tilde{e}) p_{\tilde{e}}(\tilde{e}) d\tilde{e}} = -\sigma_{\tilde{e}}^2 \frac{p_e'(e)}{p_e(e)}. \quad (1)$$

Specifically, if v is zero-mean Laplacian distributed with the scaling parameter α_v, then (1) gives

$$f_{MMSE}^{GL}(e) = \frac{\sigma_{\tilde{e}}^2}{\alpha_v} \left[\frac{e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right) - e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right)}{e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right) + e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right)} \right], \quad (2)$$

where $\xi = e/\alpha_v$, $\psi = \sigma_{\tilde{e}}^2/\alpha_v^2$, and $\mathrm{erfc}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-r^2} dr$ is the complimentary error function. (2) is plotted for $\sigma_{\tilde{e}}^2 = \alpha_v = 1$ in Figure 2(a).

2.2. Any filter error and Gaussian noise distributions

If v is zero-mean Gaussian distributed with the variance σ_v^2, then the MMSE estimate of e for any filter error $\tilde{e}$ is given by

$$f_{MMSE}^{AG}(e) = \frac{\sigma_v^2 p_e'(e) + e p_e(e)}{p_e(e)} = e + \sigma_v^2 \frac{p_e'(e)}{p_e(e)}. \quad (3)$$

Specifically, if $\tilde{e}$ is zero-mean Laplacian distributed with the scaling parameter $\alpha_{\tilde{e}}$, then (3) gives [1]

$$f_{MMSE}^{LG}(e) =$$
$$\alpha_{\tilde{e}} \left[\frac{(\psi+\xi)e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) - (\psi-\xi)e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)}{e^{\xi}\mathrm{erfc}\left(\frac{\psi+\xi}{\sqrt{2\psi}}\right) + e^{-\xi}\mathrm{erfc}\left(\frac{\psi-\xi}{\sqrt{2\psi}}\right)} \right], \quad (4)$$

where $\xi = e/\alpha_{\tilde{e}}$ and $\psi = \sigma_v^2/\alpha_{\tilde{e}}^2$. (4) is plotted for $\alpha_{\tilde{e}} = \sigma_v^2 = 1$ in Figure 2(c).

2.3. Gaussian filter error and Gaussian noise distributions

If $\tilde{e}$ and v are zero-mean Gaussian distributed with the variances $\sigma_{\tilde{e}}^2$ and σ_v^2, respectively, then either (1) or (3) can be used to obtain

$$f_{MMSE}^{GG}(e) = \frac{\sigma_{\tilde{e}}^2}{\sigma_{\tilde{e}}^2 + \sigma_v^2} e, \quad (5)$$

which is the well-known Wiener filter used in the frequency-domain signal enhancement techniques. (5) is plotted for $\sigma_{\tilde{e}}^2 = \sigma_v^2 = 1$ in Figure 2(b).

2.4. Comparison of noise reducing nonlinearities

(5) can be thought of as the mid-point of the three cases, where (5) transforms into (2) or (4) when the noise probability density function (pdf) or the filter error pdf in the respective cases are tailed more heavily and is more peaky, i.e. has higher kurtosis, than the Gaussian pdf. That is, (2) *compresses* the signal amplitude at the high end while (4) *suppresses* it at the low end, each targeting the observed signal magnitude at which the noise probability is higher compared to the target signal. It can be shown that as $|e| \to \infty$, (2) levels off to plateaus at $\pm\sigma_{\tilde{e}}^2/\alpha_v$ and that (4) asymptotically reaches linearly sloped boundaries with the offset amount of $\pm\sigma_v^2/\alpha_{\tilde{e}}$. It can also be shown numerically (since closed-form solutions are not always attainable) by using the Gauss-Hermite quadrature to estimate the integrals that using the Laplacian pdf instead of the Gaussian pdf for $\tilde{e}$ in (2) gives the same compressive form of nonlinearity but with more suppression near the origin, whereas using a pdf with higher kurtosis than the Laplacian pdf (e.g. double-sided Gamma) for $\tilde{e}$ in (4) gives a "coring" form of nonlinearity that suppresses small amplitudes while preserving larger ones. Hereafter, (2) and (4) are referred as f_{comp} and f_{supp}, respectively.

3. COMPRESSIVE NONLINEARITY FOR DOUBLE-TALK

It was shown in [5] and by several others that a compressive form of nonlinearity can be derived for an impulsive noise through the robust statistics theory [6]. According to [5], the nonlinearity is defined as

$$f_{robust}(e) = \psi\left[\frac{|e|}{s}\right] \mathrm{sign}\,[e]\, s, \quad (6)$$

$$\psi\left[\frac{|e|}{s}\right] = \min\left[\frac{|e|}{s}, k_0\right], \quad (7)$$

$$s(n+1) = \lambda_s s(n) + \frac{1-\lambda_s}{\beta}\, \psi\left[\frac{|e(n)|}{s(n)}\right] s(n), \quad (8)$$

where (7) is the compressive function and (8) is the scaling function for some control parameters k_0, λ_s, and β. (6) is plotted in Figure 2(d) for $k_0 = s = 1$.

One can immediately see a similarity in the compressive form between (2) and (6), as (2) also limits large values in observed signal caused by the addition of noise with heavy-tailed distribution. The main difference is that (2) holds the output to within the $\pm\sigma_{\tilde{e}}^2/\alpha_v$ range, which allows for adaptive adjustment of the threshold as long as the statistics from both the signal of interest and the noise can be well estimated, while (6) limits the output range to $\pm s$, where it can be seen from (8) that the scaling factor is a smoothed estimate of the average magnitude of the observed filter error, i.e. $E[|e|]$. Therefore, (2) should be able to track the changes in the environment better than (6), while (6) is more effective than (2) in the ability to limit large fluctuations in signal value but may be too restrictive if it is used for scaling down the adaptive filter error, which ultimately leads to slower convergence.

4. TESTING METHODS

4.1. Double-talk detector

The Geigel DTD [7] is used here, whose decision rule is defined by

$$\begin{aligned} \max\{|x(n)| \dots |x - L + 1|\} &< T|y(n)| \;\rightarrow\; \text{DT}, \\ \max\{|x(n)| \dots |x - L + 1|\} &\geq T|y(n)| \;\rightarrow\; \text{no DT}, \end{aligned} \quad (9)$$

where x is the far-end signal, L is the tap length, y is the near-end signal, and T is the threshold value. Setting T too high increases the false detection rate, which is undesirable if further deceleration in the convergence speed from the freezing of adaptation is to be avoided. There are other more reliable DTD's based on the generalized cross-correlation (GCC) method, but (9) is used in this case for its simplicity and also to place more responsibility on a filter error nonlinearity to limit the effect of double-talk leakage.

4.2. Regularization procedure

Besides the normalized LMS (NLMS) algorithm, the affine projection algorithm (APA) [8] is used here to obtain a faster convergence for the case of a long impulse response. Just as with NLMS, APA is highly sensitive to the near-end noise that causes adaptation instability when the far-end signal is weakly excited. A simple yet effective solution for NLMS is using

$$\frac{||\boldsymbol{x}||^2}{||\boldsymbol{x}||^4 + \gamma\sigma_v^4} \quad (10)$$

to scale the step-size rather than with $1/||\boldsymbol{x}||^2$ for some constant γ [9]. (10) is nearly zero for very small far-end signal power, whereas it becomes the usual normalization factor when the far-end power is large than the near-end noise power. The same idea can be extended to APA by replacing the normalization matrix $[\,\boldsymbol{X}^T\boldsymbol{X} + \delta\boldsymbol{I}\,]^{-1}$ by

$$[\,\boldsymbol{X}^T\boldsymbol{X} + \gamma\sigma_v^2\boldsymbol{I}\,]^{-1}\boldsymbol{X}^T\boldsymbol{X}\,[\,\boldsymbol{X}^T\boldsymbol{X} + \gamma\sigma_v^2\boldsymbol{I}\,]^{-1}. \quad (11)$$

The application of (11) to APA can decrease the misalignment drastically when there is the near-end noise. However, there are still some instances of divergence even after the regularization procedure when the SNR is extremely low. Thus to reduce the chance of such an occurrence, the filter adaptation is allowed only during voice activity by using a voice activity detector (VAD). Hereafter, NLMS and APA combined with the regularization procedure are referred respectively as RNLMS and RAPA.

4.3. Simulation and system parameter estimation procedures

16-bit female and male speeches sampled at 8 kHz are used for the far-end and the near-end signals, respectively. The echo return loss (ERL) is set at 20 dB, the average far-end to double-talk ratio is set at 6 dB, and 10 dB SNR white Gaussian noise is added to the acoustic echo. The Geigel detector threshold T is set at 2 or 4, and the DTD hold time is set at 30 ms.

The step-size μ is set at 0.5 or 1 for RNLMS and RAPA, for last of which the fourth-order APA is used. The regularization parameter γ is set at 10^6 for RNLMS and 1 for RAPA. The misalignment, which measures the MSD performance of the AEC, is defined as

$$\text{Misalignment} \equiv 10\log_{10}\frac{\|\boldsymbol{h}(n) - \boldsymbol{w}(n)\|^2}{\|\boldsymbol{h}(n)\|^2} \quad \text{(dB)}, \quad (12)$$

where $\boldsymbol{h}$ and $\boldsymbol{w}$ are the true and the estimated impulse responses, respectively.

The parameters for f_{robust} are set at λ_s = 0.9969, k_0 = 1.1, and β = 0.6067, where the forgetting factor λ_s is chosen according to the formula $\tau_{\text{samples}} \approx 1/(1-\lambda_s)$, which gives 40 ms as the exponential decay time. The noise and the filter error statistics are estimated by

$$\hat{\sigma}_v^2(n) = \lambda_v\hat{\sigma}_v^2(n-1) + (1-\lambda_v)e^2(n), \quad (13)$$

$$\hat{\sigma}_{\tilde{e}}^2(n) = \lambda_{\tilde{e}}\hat{\sigma}_{\tilde{e}}^2(n-1) + (1-\lambda_{\tilde{e}})\max\{e^2(n) - \hat{\sigma}_v^2(n), 0\}, \quad (14)$$

where $\hat{\sigma}_v^2$ is calculated during silence and held constant otherwise, whereas $\hat{\sigma}_{\tilde{e}}^2$ is set equal to $\hat{\sigma}_v^2$ during silence and updated only during voice activity. λ_v and $\lambda_{\tilde{e}}$ are chosen with respective decay times of 100 ms and 40 ms. $\hat{\sigma}_{\tilde{e}}$ and s are forced to decay down to $\hat{\sigma}_v$ during double-talk. Using (13) and (14) to estimate the scaling parameter α for the Laplacian pdf will overestimate it by roughly a factor of 1.25, which translates into more restrictive thresholding for (2) but less for (4).

There are two nonlinearity combinations of interest: f_{supp} + f_{comp} and f_{supp} + f_{robust}. When two nonlinearities are used together, each nonlinearity is applied separately to e, and the output with smaller magnitude is taken as the effective output of the ESN.

5. SIMULATION RESULTS

Figure 3 shows the misalignment plots from using RNLMS with different combinations of the parameters μ and T and the nonlinearities f_{supp}, f_{comp}, and f_{robust}. A short impulse response with 100 coefficients (12.5 ms) is used. The double-talk occurs during some time between 2 and 4 seconds. It can be observed from all of the cases that in general f_{supp} + f_{comp} performs best during single-talk, i.e. when there is only the background noise, whereas f_{supp} + f_{robust} performs best during double-talk. As discussed in [1], the effectiveness of the ESN disappears when μ is decreased beyond a certain point due to excessive slowing of the convergence speed by the ESN.

Adding f_{comp} to f_{supp} further improves the misalignment during both single-talk and double-talk, and the combination does better than others especially when μ is large. This means that the compressive nature of f_{comp} has an additional stabilizing effect, as one form of disruptions to the echo path is the occurrence of sudden changes between the voiced speech, i.e. high SNR, and the unvoiced speech or the silence, i.e. low SNR. f_{supp} + f_{comp} also enables the ESN performance to reach close to that of f_{robust} during double-talk, where it was observed in other cases with different set of system parameters that almost the same performance can be obtained.

f_{robust} also benefits from the addition of f_{supp} during both single-talk and double-talk, most likely since f_{robust} by itself fails to suppress the lower magnitude noise that f_{supp} is better able to adaptively suppress. As expected, f_{supp} + f_{comp} does not do as well as f_{supp} + f_{robust} during double-talk since f_{comp} has a wider output range than f_{robust} and thus is unable to limit many large deviations as well as f_{robust} can. At this time, there is no practical way to enforce f_{comp} by considering the near-end speech as the noise since the near-end speech energy during double-talk cannot be reliably estimated, the problem of which is compounded by the fact that a DTD can sporadically give false decisions.

Finally, Figure 4 shows the misalignment plots from using RAPA with μ = 0.5, T = 4, and a longer impulse response of 800 coefficients (100 ms). f_{supp} + f_{comp} performs best for most of the time with this particular example. Although the misalignment for RAPA by itself is not shown in Figure 4, it was observed that the misalignment is decreased by over 20 dB after the inclusion of the nonlinearities, which further signifies the effectiveness of the ESN.

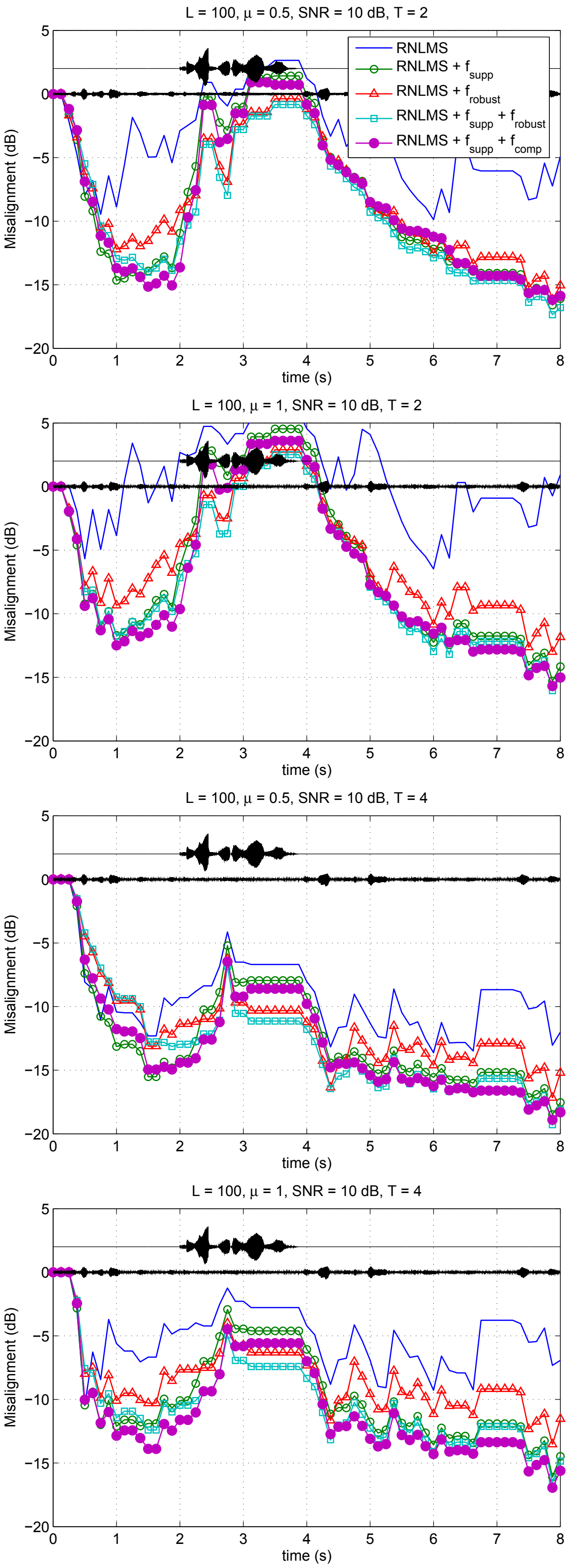

Fig. 3. Misalignment from using RNLMS and a short impulse response (12.5 ms) with different combinations of μ, T, f_{supp}, f_{comp}, and f_{robust}.

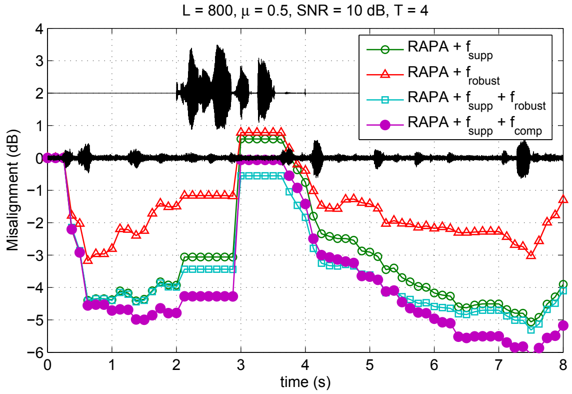

Fig. 4. Misalignment from using RAPA and a long impulse response (100 ms) with different combinations of f_{supp}, f_{comp}, and f_{robust}.

6. CONCLUSION

The technique of using a noise suppressing nonlinearity in the adaptive filter error feedback loop of the acoustic echo canceler (AEC) based on the least mean square (LMS) algorithm is extended to include a noise compressing nonlinearity to limit large deviations in the observed filter error due to the double-talk. By combining the previously developed noise suppressing nonlinearity with a compressive nonlinearity derived from the theory of robust statistics, consistently better results are obtained during both single-talk and double-talk when compared to using only the compressive nonlinearity. A compressive form of noise reducing nonlinearity can also be derived from the signal enhancement point of view, and a combination of noise suppressing and compressing nonlinearities is capable of providing performance similar to that of the compressive nonlinearity of the robust statistics approach during double-talk but also with an added benefit of increased robustness during single-talk when there is only the background noise. The new results further support the fact that the filter error enhancement strategy can make an adaptive filter more robust to many types of disruption to the echo path.

7. REFERENCES

[1] T.S. Wada and B.H. Juang, "Enhancement of residual echo for improved acoustic echo cancellation," in *Proc. EUSIPCO*, Sep. 2007, pp. 1620–1624.

[2] T.S. Wada and B.H. Juang, "Enhancement of residual echo for improved frequency-domain acoustic echo cancellation," in *Proc. WASPAA*, Oct. 2007, pp. 175–178.

[3] T.Y. Al-Naffouri and A.H. Sayed, "Adaptive filters with error nonlinearities: Mean-square analysis and optimum design," *Applied Signal Process.*, vol. 2001, no. 4, pp. 192–205, Oct. 2001.

[4] A. Mader, H. Puder, and G.U. Schmidt, "Step-size control for acoustic echo cancellation filters - and overview," *Signal Processing*, vol. 80, no. 9, pp. 1697–1719, Sep. 2000.

[5] T. Gänsler, S.L. Gay, M.M. Sondhi, and J. Benesty, "Double-talk robust fact converging algorithms for network echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 6, pp. 656–663, Nov. 2000.

[6] P.J. Huber, *Robust Statistics*, Wiley, 1981.

[7] D.L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Commun.*, vol. 8, no. 5, pp. 508–518, Sep. 2000.

[8] S. Haykin, *Adaptive Filter Theory*, Prentice Hall, 2002.

[9] A. Hirano and A. Sugiyama, "A noise-robust stochastic gradient algorithm with an adaptive step-size for mobile hands-free telephones," in *Proc. ICASSP*, May 1995, vol. 2, pp. 1392–1395.