DOUBLE-TALK ROBUST VSS-NLMS ALGORITHM FOR UNDER-MODELING ACOUSTIC ECHO CANCELLATION

Constantin Paleologu[1], *Silviu Ciochină*[1], *and Jacob Benesty*[2]

1: University Politehnica of Bucharest, Telecommunications Department
Iuliu Maniu Blvd., 1-3, Bucharest, Romania
e-mail: {pale, silviu}@comm.pub.ro
2: Universite du Quebec, INRS-EMT, Montreal, QC H5A 1K6, Canada
e-mail: benesty@emt.inrs.ca

ABSTRACT

Most of the adaptive algorithms used for acoustic echo cancellation (AEC) are designed assuming an exact modeling scenario (i.e., the acoustic echo path and the adaptive filter have the same length) and a single-talk context (i.e., the near-end speech is absent). In real-world AEC applications, the adaptive filter works most likely in an under-modeling situation, i.e., its length is smaller than the length of the acoustic impulse response, so that the under-modeling noise is present. Also, the double-talk case is almost inherent, so that a double-talk detector (DTD) is usually involved. Both aspects influence and limit the algorithm's performance. Taking into account these two practical issues, a double-talk robust variable step size normalized least-mean-square (VSS-NLMS) algorithm is proposed in this paper. This algorithm is nonparametric in the sense that it does not require any information about the acoustic environment, so that it is robust and easy to control in practice.

Index Terms— Acoustic echo cancellation, adaptive filters, variable step size normalized least-mean-square (VSS-NLMS) algorithm, double-talk robustness, under-modeling system identification.

1. INTRODUCTION

Adaptive algorithms designed for acoustic echo cancellation (AEC) have to consider several practical aspects such as the large length and time-varying nature of the echo path, and the presence of the near-end signal [1]. Mainly due to complexity reasons, the normalized least-mean-square (NLMS) algorithm and different versions of it [2] are frequently involved in AEC.

The performance of the NLMS based algorithms is governed by the step size parameter. Its value has to be large in order to achieve a high convergence rate or tracking (e.g., as in the case of an abrupt change of the echo path); on the other hand, low misadjustment (desired in the steady-state) is obtained using a small step-size. Consequently, a compromise choice should be made. This is the main motivation behind the development of the variable step size NLMS (VSS-NLMS) algorithms [3], [4] (and references therein), which control the value of the step size parameter according to these requirements. Nevertheless, most of these algorithms were derived assuming an exact modeling scenario, i.e., the length of the adaptive filter is equal to the length of the system that has to be modeled. In the context of AEC, due to the large length of the acoustic impulse response, the under-modeling situation (i.e., the length of the adaptive filter is smaller than the length of the echo path) is the rule. Hence, a residual echo, also known as the under-modeling noise, disturbs the algorithm performance.

A critical factor in echo cancellation is the presence of the near-end signal. It contains both the ambient noise and the near-end speech. When only the ambient noise is present, the adaptive algorithm can act on its own. The presence of the near-end speech is considered as a different case, also known as double-talk; it seriously affects the algorithm behaviour. Consequently, a double-talk detector (DTD) is usually involved, in order to slow down or completely halt the algorithm [1]. Since there is an inherent latency in the DTD decision, the adaptive algorithm should handle a small amount of double-talk without diverging. Nevertheless, most of the NLMS based algorithms face real difficulties in this context. As a consequence, several solutions for enhancing the double-talk robustness of these algorithms were developed [5], [6] (and references therein).

In this paper we propose a VSS-NLMS algorithm derived in a general framework that considers these two important aspects, i.e., the under-modeling case and the double-talk scenario. Due to its nature, this algorithm takes into account the under-modeling noise and it is also robust to double-talk. The proposed algorithm does not require any a priori information about the acoustic environment, so that it is easy to control in real-world AEC applications.

2. VSS-NLMS ALGORITHM FOR DOUBLE-TALK AND UNDER-MODELING SCENARIO

Let us consider the AEC configuration depicted in Fig. 1. The goal of this classical scheme is to identify an unknown system (i.e., acoustic echo path) using an adaptive filter. Both systems have finite impulse responses, defined by the real-valued vectors $\mathbf{h} = [h_0\ h_1\ \ldots\ h_{N-1}]^T$ and $\hat{\mathbf{h}}(n) = [\hat{h}_0(n)\ \hat{h}_1(n)\ \ldots\ \hat{h}_{L-1}(n)]^T$, where superscript ^{T} denotes transposition and n is the time index. Since the under-modeling scenario is more realistic in AEC, we impose that $L < N$.

This work was supported by the UEFISCSU Romania under Grants PN-II-"Idei" no. 65/01.10.2007 and no. 331/01.10.2007.

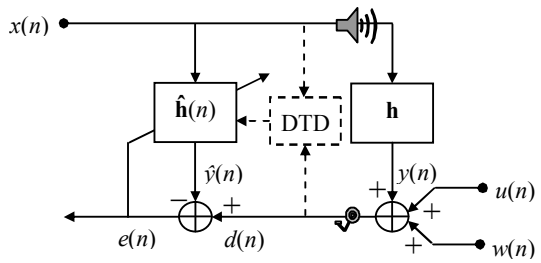

Fig. 1. AEC configuration.

The signal $x(n)$ is the far-end speech which goes through the acoustic impulse response $\mathbf{h}$, resulting the echo signal $y(n)$. This signal is picked up by the microphone together with the near-end speech $u(n)$ and the ambient noise $w(n)$, resulting the microphone signal $d(n)$. The output of the adaptive filter, $\hat{y}(n)$, provides a replica of the echo, which will be subtracted from the microphone signal. The DTD block controls the algorithm behaviour during double-talk; nevertheless, the proposed algorithm will be derived without involving the DTD decision.

In the most general situation, the desired signal can be written as

$$d(n)=y(n)+u(n)+w(n)=[\mathbf{x}^T_L(n)\ \mathbf{x}^T_{N-L}(n)][\mathbf{h}^T_L\ \mathbf{h}^T_{N-L}]^T+u(n)+w(n),$$

where the following notation has been introduced:

$$\mathbf{x}_L(n) = [x(n)\ x(n-1)\ \ldots\ x(n-L+1)]^T,$$

$$\mathbf{x}_{N-L}(n) = [x(n-L)\ x(n-L-1)\ \ldots\ x(n-N+1)]^T,$$

$$\mathbf{h}_L = [h_0\ h_1\ \ldots\ h_{L-1}]^T,\ \mathbf{h}_{N-L} = [h_L\ h_{L+1}\ \ldots\ h_{N-1}]^T.$$

Consequently, the echo signal is $y(n) = y_L(n) + y_{N-L}(n)$, where $y_L(n) = \mathbf{x}^T_L(n)\mathbf{h}_L$ and $y_{N-L}(n) = \mathbf{x}^T_{N-L}(n)\mathbf{h}_{N-L}$. The a priori error is defined using the adaptive filter coefficients at time $n-1$, as

$$e(n) = d(n) - \hat{y}(n) = d(n) - \mathbf{x}^T_L(n)\hat{\mathbf{h}}(n-1) = \\ = \mathbf{x}^T_L(n)[\mathbf{h}_L - \hat{\mathbf{h}}(n-1)] + y_{N-L}(n) + u(n) + w(n). \quad (1)$$

In a similar manner, the a posteriori error can be written based on the adaptive filter coefficients at time n, as

$$\varepsilon(n) = \mathbf{x}^T_L(n)[\mathbf{h}_L - \hat{\mathbf{h}}(n)] + y_{N-L}(n) + u(n) + w(n). \quad (2)$$

The update equation of the gradient based adaptive algorithm is

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \mu(n)\mathbf{x}_L(n)e(n), \quad (3)$$

where $\mu(n)$ is the step-size parameter (positive scalar). Thus, taking (1)–(3) into account, the a posteriori and a priori error signals are related by the following formula:

$$\varepsilon(n) = e(n)[1 - \mu(n)\mathbf{x}^T_L(n)\mathbf{x}_L(n)]. \quad (4)$$

At a first glance, the contributions of the under-modeling noise and the near-end signal do not appear explicitly in the above relation. So that, in order to derive an expression for the step size parameter, we may impose to cancel the a posteriori error signal, i.e., $\varepsilon(n) = 0$, assuming that $e(n) \neq 0$ [7]. As a result, $\mu_{\text{NLMS}}(n) = [\mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$, which is the step size of the classical NLMS algorithm. In practice, a positive constant (usually smaller than 1) multiplies this step size to achieve a proper compromise between the convergence rate and the misadjustment [2]. We should note that this straightforward approach holds in the noise-free sigle-talk scenario (i.e., $w(n) = 0$, $u(n) = 0$) and in the exact modeling situation (i.e., $N = L$). The differences from the ideal conditions can be explained as follows. If we impose to cancel the a posteriori error in the presence of the near-end signal and in the under-modeling case, it results from (2) that

$$\mathbf{x}^T_L(n)[\mathbf{h}_L - \hat{\mathbf{h}}(n)] = -y_{N-L}(n) - u(n) - w(n) \neq 0. \quad (5)$$

Hence, the adaptive filter estimate is biased. On the other hand, the proper condition $\mathbf{x}^T_L(n)[\mathbf{h}_L - \hat{\mathbf{h}}(n)] = 0$, leads to

$$\varepsilon(n) = e(n)[1 - \mu(n)\mathbf{x}^T_L(n)\mathbf{x}_L(n)] = y_{N-L}(n) + u(n) + w(n). \quad (6)$$

Consequently, in consistence with the approach proposed in [4] and [8], and assuming that the sequences $y_{N-L}(n)$, $u(n)$, and $w(n)$ are uncorrelated with each others, we can impose that $E\{\varepsilon^2(n)\} = E\{y^2_{N-L}(n)\} + E\{u^2(n)\} + E\{w^2(n)\}$, where $E\{\bullet\}$ denotes the mathematical expectation. Squaring (6), then taking the expectations, and assuming that $\mathbf{x}^T_L(n)\mathbf{x}_L(n) = LE\{x^2(n)\}$ for $L>>1$ (which is valid in AEC where the length of the adaptive filter is of the order of hundreds), it results that

$$E\{e^2(n)\}[1 - \mu(n)LE\{x^2(n)\}]^2 = \\ = E\{y^2_{N-L}(n)\} + E\{u^2(n)\} + E\{w^2(n)\}. \quad (7)$$

Regarding (7) as a quadratic equation, the solution for the step size parameter is

$$\mu(n) = \frac{1}{\mathbf{x}^T_L(n)\mathbf{x}_L(n)}\left[1 - \sqrt{\frac{E\{y^2_{N-L}(n)\} + E\{u^2(n)\} + E\{w^2(n)\}}{E\{e^2(n)\}}}\right] \quad (8)$$

Expression (8) is useless in a real-world AEC application since it depends on some parameters that are unavailable, i.e., $y_{N-L}(n)$, $u(n)$, and $w(n)$. In order to solve this issue, first we assume that $y_L(n)$ and $y_{N-L}(n)$ are uncorrelated. This holds especially when the correlation function of the input signal is a particular one, as in the case of white noise (frequently used in adaptive filters' analysis in the context of the independence assumptions). When the input signal is speech it is difficult to analytically state this assumption. Nevertheless, we can extend it based on the fact that $L >> 1$ in AEC scenario, and that for usual cases the correlation function has a decreasing trend with the time lag. Moreover, in general the first part of the acoustic impulse response $\mathbf{h}_L$ is more significant as compared to the tail $\mathbf{h}_{N-L}$. Thus, $E\{y^2_{N-L}(n)\} \approx E\{y^2(n)\} - E\{y^2_L(n)\}$. Secondly, we assume that the adaptive filter coefficients have converged to a certain degree, so that $E\{y^2_L(n)\} \approx E\{\hat{y}^2(n)\}$. We know that $d(n) = y_L(n) + y_{N-L}(n) + u(n) + w(n)$. Since all the sequences from the right term are uncorrelated with each others, it results that

$$E\{d^2(n)\} = E\{y^2_L(n)\} + E\{y^2_{N-L}(n)\} + E\{u^2(n)\} + E\{w^2(n)\}.$$

Hence, taking into account the previous assumptions, the most problematic term from (8) becomes

$$E\{y^2_{N-L}(n)\} + E\{u^2(n)\} + E\{w^2(n)\} = E\{d^2(n)\} - E\{\hat{y}^2(n)\},$$

and the expression of the step size parameter is

$$\mu(n)=\frac{1}{\mathbf{x}_L^T(n)\mathbf{x}_L(n)}\left[1-\sqrt{\frac{E\{d^2(n)\}-E\{\hat{y}^2(n)\}}{E\{e^2(n)\}}}\right] \quad (9)$$

In practice, (9) has to be evaluated in terms of power estimates, as

$$\mu(n)=\frac{1}{\mathbf{x}_L^T(n)\mathbf{x}_L(n)}\left[1-\frac{\sqrt{\hat{\sigma}_d^2(n)-\hat{\sigma}_{\hat{y}}^2(n)}}{\hat{\sigma}_e(n)}\right] \quad (10)$$

The notation $\hat{\sigma}_p^2(n)$ represents the power estimate of the sequence $p(n)$. These parameters can be computed recursively as $\hat{\sigma}_p^2(n)=\lambda\hat{\sigma}_p^2(n-1)+(1-\lambda)p^2(n)$, where $\lambda = 1 - 1/(KL)$ is a weighting factor, with $K > 1$; the initial value is $\hat{\sigma}_p^2(0)=0$.

Next, a few practical issues have to be considered. First, in order to avoid divisions by small numbers, a positive constant δ, known as the regularization factor, needs to be added to the first denominator in (10). Also, a small positive number ξ should be added to the second denominator of (10) to avoid division by zero. Secondly, under our assumptions, we have $E\{d^2(n)\} \geq E\{\hat{y}^2(n)\}$ and $E\{d^2(n)\} - E\{\hat{y}^2(n)\} \approx E\{e^2(n)\}$. Nevertheless, the estimates of these parameters could lead to some deviations from the previous theoretical conditions, so that we will take the absolute value of the step size parameter from (10). Consequently, the proposed algorithm uses the step size

$$\mu(n)=\frac{1}{\delta+\mathbf{x}_L^T(n)\mathbf{x}_L(n)}\left|1-\frac{\sqrt{\hat{\sigma}_d^2(n)-\hat{\sigma}_{\hat{y}}^2(n)}}{\xi+\hat{\sigma}_e(n)}\right| \quad (11)$$

In order to satisfy the assumption that the adaptive filter coefficients have converged to a certain degree, we could start the proposed algorithm using $\mu(n) = [\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$ in the first M iterations, with $M \geq L$. This option influences only the initial convergence rate.

It is interesting to notice that the step size of the proposed algorithm does not depend explicitly on the near-end signal or the under-modeling noise, even if it was developed under these conditions; consequently, a robust behaviour is expected. Moreover, since only the parameters available from the adaptive filter are required and there is no need for a priori information about the acoustic environment, it is easy to control in practice.

3. SIMULATION RESULTS

The simulations were performed in an AEC context, as shown in Fig. 1. The acoustic echo path is plotted in Fig. 2 (the sampling rate is 8 kHz). Its impulse response $\mathbf{h}$ has $N = 1000$ coefficients, while the adaptive filter length is $L = 500$. The input signal $x(n)$ is either a white Gaussian noise or a speech signal. An independent white Gaussian noise signal $w(n)$ is added to the echo signal $y(n)$, with 20 dB signal-to-noise ratio (SNR).

We compare the proposed algorithm with the NLMS algorithm with two different step sizes, (a) $0.5[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$ and (b) $0.05[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$, and with the nonparametric VSS-NLMS (NPVSS-NLMS) algorithm developed in [4]. This last algorithm was derived in a similar manner with the proposed one, but assuming an exact modeling ($N = L$) and a single-talk context.

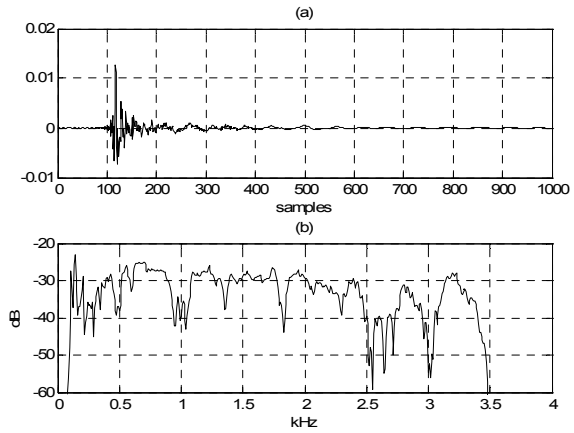

Fig. 2. (a) Acoustic impulse response; (b) Frequency response.

Its step size is $\mu(n)=[\delta+\mathbf{x}_L^T(n)\mathbf{x}_L(n)]^{-1}[1-\sigma_w/(\xi+\hat{\sigma}_e(n))]$, where σ_w^2 is the noise power (it has to be estimated during silences). A regularization factor $\delta = 30\sigma_x^2$ is used for all the algorithms, where σ_x^2 is the power of the input signal. The weighting factor λ (needed for power estimates) uses $K = 2$ for the white Gaussian input signal and $K = 6$ for the speech input signal. As it was specified before, the proposed algorithm uses a regular NLMS step size in the first $M = L$ iterations. This is a small value of M (in order to show one of the worst cases), so that a slower initial convergence rate is expected. In practice, according to the specific of application, a larger value could be used. The measure of performance is the normalized misalignment (in dB), defined as $20\log_{10}(\|\mathbf{h}-[\hat{\mathbf{h}}^T(n)\ \mathbf{0}^T_{N-L}]^T\|_2/\|\mathbf{h}\|_2)$, where $\|\bullet\|_2$ is the l_2 norm.

In the experiment presented in Fig. 3, using the white Gaussian input signal, the algorithms were stressed as follows. First, an abrupt change in the acoustic environment is introduced by shifting the acoustic impulse response to the right by 12 samples, after 5 seconds from the debut of the adaptive process. Secondly, a near-end sinusoidal burst, $u(n) = 0.5\sin(0.5\pi n)$ is introduced after 10 seconds, for a period of 2 seconds; the DTD is not involved. Finally, the SNR is changed from 20dB to 10dB after 15 seconds from the debut (assuming that the new value of σ_w^2 is not available yet for the NPVSS-NLMS algorithm). For comparison, the theoretical misalignment was measured as $20\log_{10}(\|\mathbf{h} - [\mathbf{0}^T_L\ \mathbf{h}^T_{N-L}(n)]^T\|_2/\|\mathbf{h}\|_2)$. In terms of convergence rate and tracking capabilities, the behaviour is almost the same for all the algorithms, except for the NLMS algorithm with the smaller step size, which is slower as expected. The proposed algorithm has a slightly lower final misalignment; its value is very close to the theoretical one and to the one achieved by the NLMS algorithm with the smaller step size. Most importantly, the proposed algorithm rules in the presence of the near-end signal variations.

In the second set of experiments we used speech as input, and also for the near-end component $u(n)$ (Fig. 4). Referring to (11), the quantity under the square-root provides an estimate of the near-end signal power plus the under-modeling noise power. Nevertheless, due to the specific nature of the speech signal (e.g., nonstationary character) the accuracy of the near-end speech power estimate is problematic, especially for long double-talk periods. Consequently, a DTD has to be involved in practice.

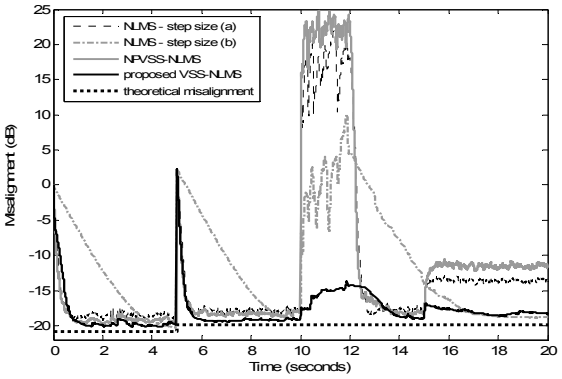

Fig. 3. Misalignment of the NLMS algorithm with two different step sizes (a) $0.5[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$ and (b) $0.05[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$, misalignment of the NPVSS-NLMS and proposed VSS-NLMS algorithms, and theoretical misalignment. The input signal is a white Gaussian noise, $N = 1000$, $L = 500$, $\lambda = 1 - 1/(2L)$, and SNR = 20dB. The impulse response changes after 5 seconds. A sinusoidal near-end burst (2 seconds) appears after 10 seconds. The SNR is decreasing from 20 dB to 10 dB after 15 seconds.

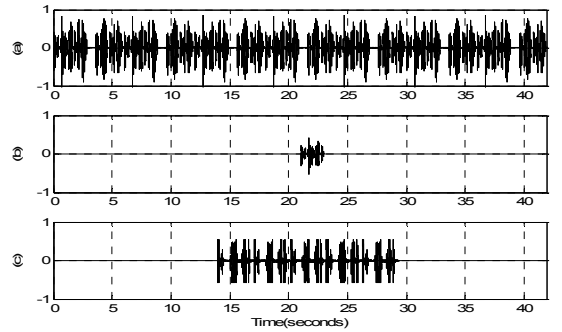

Fig. 4. (a) Far-end speech; (b) Near-end speech for simulation for Fig. 5a (without DTD); (c) Near-end speech for simulation for Fig. 5b (with Geigel DTD).

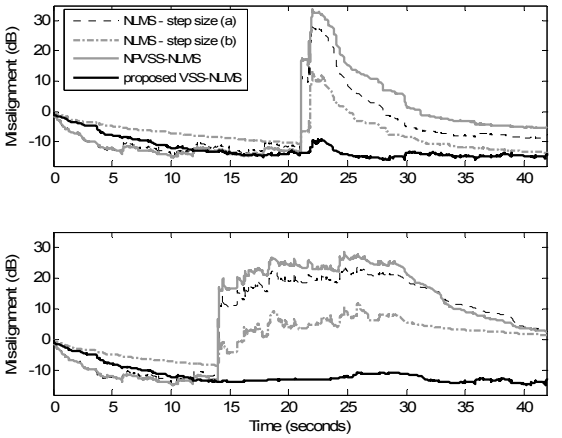

Fig. 5. Performance during double-talk: (Upper) without DTD (near-end speech from Fig. 4b); (Lower) with Geigel DTD (near-end speech from Fig. 4c). Misalignment of the NLMS algorithm with two different step sizes (a) $0.5[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$ and (b) $0.05[\delta + \mathbf{x}^T_L(n)\mathbf{x}_L(n)]^{-1}$, misalignment of the NPVSS-NLMS and proposed VSS-NLMS algorithms. The input signal is speech, $N = 1000$, $L = 500$, $\lambda = 1 - 1/(6L)$, and SNR = 20dB.

Two scenarios were considered. In the first one, the near-end speech is present for a short period of 2 seconds (Fig. 4b). The results of this simulation, performed without DTD, are presented in Fig. 5a. It can be noticed that the proposed algorithm is very robust in this case and outperforms by far the other algorithms. In the second scenario, the double-talk situation is stronger (Fig. 4c), so that a simple Geigel DTD [9] is involved (its settings are chosen assuming a 6dB attenuation, i.e., the threshold is equal to 0.5 and the hangover time is set to 240 samples). The results are shown in Fig. 5b. Also, it is clear that the proposed algorithm is very stable as compared to the others.

4. CONCLUSIONS AND PERSPECTIVES

The presence of the near-end signal and the under-modeling situation are two factors with a great impact on the performances of AEC applications. In this paper we have considered these aspects in the context of VSS-NLMS algorithms. The proposed algorithm is very simple and easy to control in practice, because it does not require any parameters from the acoustic environment. The simulation results indicate a good behaviour during double-talk situations. A comparison with other double-talk robust algorithms will be done in the near future.

5. REFERENCES

[1] J. Benesty, T. Gaensler, D. R. Morgan, M. M. Sondhi, and S. L. Gay, *Advances in Network and Acoustic Echo Cancellation*, Springer-Verlag, Berlin, Germany, 2001.

[2] S. Haykin, *Adaptive Filter Theory*, Fourth ed., Prentice-Hall, Upper Saddle River, NJ, 2002.

[3] H.-C. Shin, A. H. Sayed, and W.-J. Song, "Variable step-size NLMS and affine projection algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, Feb. 2004.

[4] J. Benesty, H. Rey, L. Rey Vega, and S. Tressens, "A nonparametric VSS NLMS algorithm," *IEEE Signal Processing Letters*, vol. 13, no. 10, pp. 581–584, Oct. 2006.

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

[6] T. van Waterschoot, G. Rombouts, P. Verhoeve, and M. Moonen, "Double-talk-robust prediction error identification algorithms for acoustic echo cancellation," *IEEE Trans. Signal Processing*, vol. 55, no. 3, pp. 846–858, Mar. 2007.

[7] D. R. Morgan and S. G. Kratzer, "On a class of computationally efficient, rapidly converging, generalized NLMS algorithms," *IEEE Signal Processing Letters*, vol. 3, no. 8, pp. 245–247, Aug. 1996.

[8] C. Paleologu, S. Ciochină, and J. Benesty, "Variable step-size NLMS algorithm for under-modeling acoustic echo cancellation," *IEEE Signal Processing Letters*, vol. 15, pp. 5–8, 2008.

[9] D. L. Duttweiler, "A twelve-channel digital echo canceler," *IEEE Trans. Communications*, vol. 26, pp. 647–653, May 1978.