Performance Limitations of Acoustic Echo Cancellers for Handsfree Telephony

A. N. Birkett , R. A. Goubran* and M. E. Knappe***

* Department of Systems and Computer Engineering,
Carleton University, 1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6, Canada
Tel: (613) 520-5740, Fax: (613) 520-5727
e-mail: birkett@sce.carleton.ca, goubran@sce.carleton.ca

** Cisco Systems, Inc.
170 West Tasman Drive, San Jose, CA, 95134, USA
Tel: (408) 527-3849, Fax: (408) 526-4287
email mknappe@cisco.com

EDICS Classification SA 2.3.1

Indexing Terms: Acoustics, Echo Cancellers, Adaptive Filters, DSP, Room Acoustics

ABSTRACT

This paper examines the limitations of acoustic echo cancellers which are commonly used in hands-free full-duplex telephones. The various factors affecting the performance of echo cancellers in terms of achievable steady-state echo cancellation and convergence characteristics are discussed. New techniques which can overcome these limitations are proposed, and evaluated using field test data acquired from commercial hands-free telephones in different rooms.

1.0 INTRODUCTION

Full-duplex hands-free telephony is achieved by using an adaptive acoustic echo canceller (AEC) to model the transfer function between the loudspeaker and microphone. This way it is possible to estimate the echo at the microphone and subtract it from the transmitted signal. In an ideal environment with a perfectly linear loudspeaker, a vibration-free telephone, no background or circuit noise, and a static anechoic chamber, a simple transversal Finite Impulse Response (FIR) adaptive filter using the Normalized Least Mean

Squares (NLMS) algorithm [1] could achieve perfect cancellation. However, achieving a perfect model in a realistic environment is a difficult problem. The limitations addressed in this paper include: undermodelling of the acoustic impulse response (AIR) of the room, room noise (fans, air conditioning), near end speech disturbance (double talk), and the ability of a particular algorithm to quickly converge and dynamically track a changing AIR while objects move inside the room. Handsfree telephones (HFTs) however include many other components which are not usually accounted for in conventional AEC designs but need to be considered in order to achieve optimal results during the identification process. These include electronic circuit noise, finite precision and truncation effects that occur when the analog signal is processed in digital form, vibration and resonances in the plastic enclosure as the loudspeaker emits a signal, microphone mechanical vibration sensitivities (as opposed to acoustical sensitivity) and nonlinearities which can occur in the loudspeaker and signal amplifiers.

A typical HFT is illustrated in Figure 1(a) and normally consists of two Adaptive Filters (AF). The first AF is used to remove acoustic echoes and the second AF is used for cancelling echoes from an imperfect hybrid as well as reflections from the line. Conventional AECs utilize a linear adaptive transversal filter to model the room impulse response and cancel the echo signal. The NLMS algorithm is the baseline by which performance of alternative models is measured but the linear transversal filter architecture is incapable of reducing (for example) nonlinear distortion and will almost always be in the undermodelled state for a typical room acoustic impulse response. As a result, a revised echo path model is required which includes all of the above limitations (See Figure 1(b)).

In this paper we address the relative seriousness of these limitations to the achievable echo cancellation. In Section 2 we outline the relevant performance requirements according to the currently available standards, review the characteristics of reverberant rooms and including reverberation time and how this affects performance, and present some measurement procedures. In Section 3, we present a summary of the performance limitations, with particular attention to the vibration and nonlinear distortion problem which have

received little attention in the literature. We provide both simulation and experimental results which illustrate the relative magnitudes of these limitations. In Section 4, we present some new methods that can be used to combat the limitations due to vibration and loudspeaker nonlinearity and provide experimental results to verify the efficacy of the proposed methods. Finally in Section 5, we discuss the results and present conclusions.

2.0 AEC PERFORMANCE EVALUATION AND MEASUREMENT

2.1 Performance Recommendations

The basic objective of a handsfree AEC is to provide ease of communication for conversational purposes. Very little work has been done to correlate objective criteria and subjective test results with regard to acoustic echo control [2]. Quantities such as naturalness of transmitted speech and quality of conversation with regard to easiness of speaking and interruption are not well defined in the literature, although these are most important to the user. The quantity most recognized as the measure of the AEC performance is the steady state Echo Return Loss Enhancement (ERLE) during single talk mode which is defined as [3];

$$ERLE(dB) = \lim_{N \to \infty} \left[10\log \frac{E[p^2(n)]}{E[e^2(n)]} \right] \cong 10\log \left[\frac{\sigma^2_p}{\sigma^2_e} \right] \quad (1)$$

where σ^2_p and σ^2_e refer to the variances of the primary and error signals respectively and E is the statistical expectation operator. However, other objective performance specifications may be found in [4] - [9] as follows;

TABLE 1. AEC Performance Requirements for Handsfree Telephones

Quantity	Description	Value
TIC	Initial convergence time	1 sec, 20 dB
TRDT	Recovery time after double talk	1 sec, 20 dB
TCLWPV	Echo loss during echo path variation	>10 dB
TRPV	Recovery time after echo path variation	1 sec, 20 dB
TCLWST	Echo loss in single talk	>45 dB
TCLWDT	Echo loss in double talk mode	>25 dB
ARDT	Received speech attenuation in double talk mode	>6 dB

TABLE 1. AEC Performance Requirements for Handsfree Telephones

Quantity	Description	Value
ARST	Transmitted speech attenuation in double talk mode	>6 dB
DRST	Received speech distortion in double talk mode	currently under study
DRDT	Transmitted speech distortion in double talk mode	currently under study
TONST	Break-in time in single talk mode	20 ms, 3 dB
TONDT	Break-in time in double talk mode	20 ms, 6 dB

Another objective measure of performance not listed in Table 1 is the attainable *early to late ratio* which is defined as the ratio of energy received before 40 ms to that received after.

Experimental results in [2] show that the annoyance due to acoustic echo level is strongly dependent of the background noise level, and that the annoyance of the background noise subjectively masks the echo. Mean opinion scores on speech signals also degrade with increased echo delay. High figures of TCLWST up to 45 dB are often proposed in the case of large transmission delays and since current technology/algorithms are generally unable to provide such high attenuation, additional variable losses in the receive and/or transmit path are frequently used. An important aspect is the ERLE subjectively required for HFTs which depends highly on the environment. For example, [10] reports an assessment performed in an audio teleconferencing environments, where for an overall round trip delay time of 100 ms and a reverberation time of 400 ms a 40 dB echo return loss is considered necessary. Other performance metrics have been adopted by the Freetel consortium to determine the performance of AECs in single talk mode only. These are listed in Table 2 [11].

TABLE 2. Freetel Evaluation criteria

Metric	Description
ERLE max	The maximum value of segmental ERLE in dB attained during the test signal (max. 2 second convergence time).
ERLE mean	The average value of segmental ERLE in dB calculated over the whole test signal.
ERLE Std	The standard deviation of ERLE mean.
TIC	The time in ms to achieve ERLE mean.
TIC 10 dB	The time in ms to attain 10 dB of segmental ERLE.

In the context of improving the subjective quality of handsfree terminals, both speech enhancement (i.e. noise reduction) and echo reduction should therefore be taken into account for obtaining an overall quality enhancement.

2.2 Reverberant Room Characteristics

The characteristics of the acoustic impulse response (AIR) of the room have a direct bearing on the type of adaptive filter structure should be used to obtain a reasonable model of the venue. For example, venues with short echo decay times will allow for more complex algorithmic processes whereas venues with long echo decay times will require the use of simpler algorithms, often resulting in an approximation to the actual room characteristics. The question remains as to what is a good approximation. If we consider a rectangular room, the number of vibrational modes N in the frequency range from 0 to f is given by [12];

$$N = \frac{4\pi}{3}V\left(\frac{f}{c}\right)^3 + \frac{\pi}{4}S\left(\frac{f}{c}\right)^2 + \frac{L}{8}\left(\frac{f}{c}\right) \qquad (2)$$

where V is the volume of the room, S is the area of all walls, L the sum of all edge lengths of the walls of the rectangular room and c is the speed of sound. For example: for a room with dimensions L_x=3m, L_y=4m and L_z=5m, the number of eigenfrequencies in the range [0, 3.4 kHz] is 247787. Since this number represents half the number of poles necessary to completely model the physical phenomenon, it is obvious that exact cancellation would require an extremely complicated architecture. Acoustic reverberation is so complicated that it can only be investigated under statistical considerations. The eigenfrequencies are highly overlapped and therefore they can be reduced to averages to provide a much more parsimonious number of modes [13]. It has been observed in the frequency response of typical rooms are composed of a sequence of maxima and minima spaced by about 5 Hz apart. If we model each maximum/minimum pair by a 2nd order IIR filter section, the total number of parameters is far less than described by (2) however, it is still quite large.

Implications for Model Selection. Results presented in [13] use a Hankel Norm approximation to show that a normalized error bound of -30 dB can be obtained when an all zero transfer function with 512 coefficients is modelled by an IIR structure with 128 parameters. However, this number is dependent on the decay characteristics of the room impulse response. A similar error can be also obtained using an all-zero filter with 128 coefficients. Results presented in [14] also show that IIR structures contribute very little improvement in ERLE, however at the expense of added complexity. Due to current processing limitations, time and frequency domain FIR structures would appear to be the obvious choice. Considering the fact that a typical AIR impulse response may be several thousand milliseconds duration, we must be satisfied to model the physical phenomenon with an undermodelled system and try to obtain the best fit according to the reverberation characteristics.

Reverberation Time. The reverberation time T_R which is defined as the length of time necessary for all reflections in a room to decay by 60 dB and is defined by [12];

$$T_R = \frac{6.91}{\delta} = \frac{-13.8}{c\left(\frac{1}{L_x} + \frac{1}{L_y} + \frac{1}{L_z}\right)\ln\beta} \quad (3)$$

where δ is the average damping constant of all the surfaces in the room, and β is the reflection coefficient varying between 0 and 1. β has a frequency dependence and generally low frequencies have a higher reflection coefficient than higher frequencies for most reflecting surfaces.Typical values of T_R range from 0.3s (living rooms) up to 10s (large churches), with values of δ ranging from 1 to 20 s^{-1}. Recommendation G.167 [4] defines the reverberation time averaged over the transmission bandwidth in a typical test room of volume 50 m^3. Equation (3) is intended for regularly shaped rooms free of furniture and people. For irregularly shaped venues, or typical furnished rooms, experimental results are required to determine accurately the echo decay characteristics. Experimental results presented in [15] for automobiles suggests a factor of 1 dB reduction in echo for every 1 ms of delay. If we model the echo path with a time-domain FIR filter structure and let the number of taps in the delay line span T_R, then we should be able to cancel to -60 dB.

The remaining uncancelled tail portion of the AIR manifests itself as a finite error at the output of the AEC. Increasing the number of taps to cover the AIR beyond T_R results in added complexity, greater algorithmic noise and slower convergence.

2.3 Measurement Procedures

The measurements presented in the following section are performed in either a low-noise, furnished conference room (approximately 11.7 m x 5.8 m x 3 m), or inside an anechoic chamber. Two commercially available HFTs are used in the experiments. The second HFT has improved vibration characteristics. Each HFT has been modified to allow access to the primary and reference electrical signals and are placed either on top of a conference table (conference room recording) or on a 1m square board on the floor of the anechoic chamber. The reference source signal consists of white noise which is bandlimited from 300 Hz to 3400 Hz. The filtered reference signal is then amplified such that the loudspeaker produces a sound pressure level (SPL) anywhere from 60dB to 100dB as measured 0.5m directly above the loudspeaker, depending on the HFT used. The primary and reference signals are then recorded onto a TEAC Digital Audio Recorder (DAT). The DAT signals are downloaded to a computer via an ARIEL DSP96 board sampling at 16 kHz for processing at a later time. Depending on the test being performed, between 32,000 and 80,000 samples are recorded, which generally gives enough time for the algorithms to converge to a steady state.

3.0 LIMITATIONS OF AECs

3.1 Primary Signal Noise Contributions

Noise components in the primary signal include room noise, microphone circuit noise and quantization noise. These can be modelled as white noise sources with variances σ_R^2,σ_M^2 and σ_Q^2 as indicated in Figure 1(b). The effect of these noise components is to reduce the achieved ERLE according to the following formula;

$$ERLE(dB) \approx 10\log\left[\frac{\sigma^2_p}{\sigma^2_T}\right] \quad (4)$$

$$\text{where } \sigma^2_T = \sigma^2_R + \sigma^2_M + \sigma^2_Q \quad (5)$$

Figure 2 (a) shows the effect on the ERLE as the noise σ^2_T component is increased.

Room Noise . Uncorrelated room noise is usually the largest contributor to the overall noise introduced into the primary path. Assuming a linear transfer function between the reference and primary signals, room noise contribution becomes the asymptote for the achievable converged ERLE [16] in the absence of other effects.

Microphone/Circuit Noise. Circuit noise is separate from the external room noise, and is modelled as uncorrelated noise which is generated mainly in the sensing electronics for the microphone. A typical electret microphone will be biased through a dropping resistor of a few kΩ to provide a bias voltage for the microphone element. The output voltage change ΔV from such a microphone is defined by;

$$\Delta V = \alpha V_b \Delta C \quad (6)$$

where α is a constant, V_b is the bias voltage across the electret capacitive element and ΔC is the change in capacitance due to the impinging sound wave. Thermal noise due to the bias resistor will be added to the microphone which itself has a noise level in the vicinity of -100 dBV. Amplification of the microphone output signal to line levels of 100 mVrms will amplify this noise, however, except for anechoic conditions, this noise is generally far below typical acoustic room noise and is not considered further.

A/D Quantization Effects. Quantization noise is introduced when the primary data is sampled using an analog to digital (A/D) converter. This noise is assumed to have a uniform distribution with a variance equal to:

$$\sigma^2_q = \frac{2^{-2B_d}}{12} \quad (7)$$

where B_d is the number of binary quantization levels (i.e. bits). In the model of Figure 1(b), the quantization noise introduced in the primary path is uncorrelated with the primary signal so appears at the output essentially unchanged (reference signal quantization noise on the other hand is modified by the adaptive filter transfer function converges). The effect of quantization noise on performance is illustrated in Figure 2 (b) and (c). When there is no quantization, the converged ERLE is independent of the signal level of the primary or reference signal levels and the ERLE will converge to the noise floor essentially limited by the IEEE floating point representation. This is because the ERLE is determined by the *ratio* of primary to error power as governed by (1) and not the individual primary or error signal power. However, when the primary signal is quantized, the maximum level of converged ERLE will be determined by the ratio of the primary signal power to the quantization noise at the location of ADC. As the number of bits in the quantized signal increases, the achievable ERLE will also increase. Figure 2 (c) illustrates this effect. As a result it is often prudent to scale the input signals to a normalized range of +/- 1.0 before quantization.

3.2 Fixed Point Internal Arithmetic in DSPs

In a fixed point digital implementation of a particular algorithm, internal word lengths will introduce truncation and quantization in addition to the quantization noise introduced during the A/D conversion. A full analysis for the LMS algorithm can be found in [17] which states that the total output mean square error for the LMS algorithm can be expressed as:

$$J = J_{min} + \frac{1}{2}\mu J_{min} tr(\mathbf{R}) + \frac{N\sigma_c^{\,2}}{2a^2\mu} + \frac{1}{a^2}(|\mathbf{w}_o| + c)\sigma_d^{\,2} \qquad (8)$$

where

J is the mean square error at the output
J_{min} is the minimum mean squared error
$\sigma_c 2$ is the variance introduced by the coefficient quantization
$\sigma_d 2$ is the variance introduced by the data (sampling) quantization
a is a scaling factor used to bring the maximum levels to +/- 1.0
N is the number of taps in the FIR structure
μ is the step size parameter

$\mathbf{w}_o$ is the optimum weiner filter coefficient vector
c is a constant

The implications of this formula for fixed point processors are important. Although one may be tempted to reduce the step size μ to reduce the excess mean square error (i.e. the second term), it may result in a large quantization error generated in the third term. There exists an optimum value of μ which minimizes the total output MSE, however, it is too small to allow the algorithm to converge completely. Figure 2(d) illustrates the MSE as a function of the adaptation step size μ where the number of bits B_d in the data representation and B_c in the coefficient representation are the same.

Floating point representation generally consists of an integer followed by a mantissa of an arbitrary number of bits. For example, the IEEE format has an implied one followed by a 23 bit mantissa. The coefficient quantization noise σ_c^2 may be represented in a similar fashion to (7) where B=23 for the IEEE floating point convention. However, depending on the number of bits in the ADC, σ_d^2, may be different. These values may then be applied in (7) to obtain the degradation in the MSE. For the floating point simulations shown, single point precision was used, with a resulting noise floor of approximately -128 dB.

In practise an ERLE of 25 to 35 dB seems to be the physical limit to the achievable ERLE in real systems. The results presented here have shown that the limitations due to noise and truncation effects are far below this limit, and therefore should not have a significant impact on the final ERLE value.

3.3 Vibration and Resonances in the Enclosure

A major part of the AIR is due to “direct coupling” between the loudspeaker, enclosure and microphone. This coupling is usually much larger in amplitude than the received echoes in the case of HFTs. Some of this coupling comes from the air path between the loudspeaker and microphone, however, a substantial part is due to vibrational coupling in the handset itself. In typical handsfree telephones, the amount of acoustic coupling between the loudspeaker and microphone may exceed 12 dB, and therefore in order to ensure stability, the total attenuation required will be greater than 12 dB(i.e. the sum of ARDT and ARST-

-See Table 1). These additional losses will have some detrimental impact on overall speech quality so reduction of this direct coupling is desirable. Vibrational coupling can be modelled with fixed parameters (if the parameters are known) or using the adaptive filter with a small step size for the early part of the reflection. This technique is called *Beta-grading* and is described in [18] and [15]. However, slowly converging filter taps will add misadjustment and gradient estimate noise to the error output, unless its tap updates are “frozen” after an initial start-up period.

Rattling of the handset and keys is also encountered in the HFT domain. Rattling is nonlinear and chaotic and can be modelled as uncorrelated noise. Recent measurements have shown that in HFTs with plastic enclosures, rattling and vibration cause an increase in the uncorrelated noise signal introduced into the primary path. Figure 3 (a) shows the effects that rattling and vibration have on the achievable ERLE of HFT #1 as measured in an anechoic chamber. The basic loudspeaker and microphone configuration will have the best achievable ERLE. The performance drops when the components are added into the enclosure. When the keys are allowed to rattle, the ERLE drops even further and finally, when the handset is placed on the set, a 10 dB reduction in ERLE is observed at 90 dB SPL as compared to the case with microphone and loudspeaker only. Figure 3(b) illustrates the effect that rattling, vibration and nonlinearity have on the primary power spectral density (PSD) of HFT #1. It is clear that when the components are mounted inside the enclosure, the out-of-band signals (distortion) increase substantially with an increase in the reference signal level. Figure 3(c) shows the PSD of the loudspeaker and microphone components only removed from HFT #1. Notice that the distortion in the frequency range 4-8 kHz is significantly reduced.

Microphone vibrational sensitivity. A microphone element with low mechanical vibration sensitivity will reduce the vibration effect mentioned previously and minimize the magnitude of the first part of the AIR. The mechanical sensitivity of a microphones will depend on the orientation of the microphone element with respect to the axis of vibration. Vibrational sensitivity displays a frequency dependence with the lower frequencies being more sensitive than higher frequencies. A typical electret microphone will exhibit

a peak vibration response in the low frequency ranges in the vicinity of 300 Hz. Table 3 lists some typical measured audio sensitivities of electret microphones and Table 4 lists the corresponding mechanical vibrational sensitivities in dbV/G.

TABLE 3. Acoustic Sensitivity

Orientation	Microphone Sensitivity	Nominal Acoustic Level	Output Voltage
Parallel	-30 (to -40) dBV/Pa	-30 dBPa	-64 dBV

TABLE 4. Mechanical Vibrational Sensitivity

Orientation	Frequency	Sensitivity	Acceleration	Output Voltage
Parallel	300 Hz	-32 dBV/G	1 G	-32dBV
Parallel	1KHz	-36 dBV/G	1 G	-36dBV
Perpendicular	1KHz	-65 dBV/G	1 G	-65 dBV

If we model the vibrational acceleration using a one dimensional harmonic motion $x = r\cos\omega t$, the acceleration *a* acting on the microphone element is $a = -\omega^2 r\cos\omega t$. A microphone element must travel a radius of 2.8 μm to generate 1G acceleration (9.8m/s^2) at 300 Hz. This small distance requires laser measurement techniques to determine accurately. However, it is reasonable to assume that given such small distances, microphone output due to vibrational coupling is not negligible when compared to the acoustical coupling. Our measurements (See Figure 3) seem to confirm this hypothesis. Methods for minimizing vibration effects are presented in Section 4.1 .

3.4 Nonlinearities in the Transfer Function

Generated mainly in the loudspeaker, nonlinear distortion effectively puts a limit on the achievable ERLE when using algorithms based on linear mechanics. Several algorithms have been proposed in the literature to deal with loudspeaker nonlinearity. [19] presents a nonlinear state space model for compensation of loudspeaker nonlinearity. A 3rd order Volterra filter is presented in [20], however this does not compensate for loudspeaker hysteresis effects. In [21] an inverse loudspeaker model is developed using a Time Delay Neural Network to provide single point room equalization.

A loudspeaker has several sources of nonlinearity including non-uniform magnetic field and nonlinear suspension system [22][19]. A loudspeaker consists of an electrical part and a mechanical part. The electrical part is the voice coil and the mechanical part consists of the cone, the suspension system and the air load. The two parts interact through the magnetic field resulting in a nonlinear force deflection characteristic f_M *o*f the loudspeaker cone suspension system, usually approximated [20] by;

$$f_M = \alpha x + \beta x^2 + \delta x^3 \quad (9)$$

where α, β and δ are modelling constants and x is the displacement of the voice coil. Suspension system nonlinearity manifests itself as soft clipping at the loudspeaker output and results in odd-order harmonics under large signal conditions. The nonlinear distortion consists mainly of cubic terms and can easily be 5 to 10 percent of the total output, especially when dealing with small loudspeakers that operate at high volumes, which is generally the case for speakerphones. However, there is also significant nonlinear distortion at extremely *low* levels of reference signal amplitude, and this distortion is mainly caused by unbalanced two-point suspension - the surround and spider [23]. It is reasonable to expect that a nonlinear model can also improve performance at these low levels as well. Nonlinear loudspeaker distortion effects can be observed in Figure 3 (c) which shows the PSD of the primary signal with the loudspeaker and microphone components removed and the loudspeaker placed in a standard baffle inside an anechoic chamber (this removes the effect of vibration, noise and echo). Notice that there is an increase in the out-of-band signal level which is essentially nonlinear components of the original bandlimited (reference) signal. However, the level of distortion is much less than that due to vibration (shown in Figure 3 (b)). A method to combat the effect of nonlinear distortion is presented in Section 4.2 .

3.5 Undermodelling of the AIR

In this section we investigate the effect of using an FIR structure to model a transfer function where the number of parameters in the candidate system will be less that required to exactly identify the system. This gives the undermodelled system: $deg(\hat{H}) < deg(H)$. Poltmann [24] showed that the achievable ERLE

is a function of both the step size and magnitude of the modelled and undermodelled AIR coefficients. For a system modelled by an FIR transfer function the achievable steady state ERLE can be calculated from;

$$ERLE = 10\log\left[\frac{2-\mu}{2}\left(\frac{\|h\|^2}{\|\Delta h\|^2}+1\right)\right] \quad (10)$$

where

$$\|h\|^2 = \sum_{i=0}^{M-1} h^2(i) \quad (11)$$

represent the modelled coefficients up to order M and,

$$\|\Delta h\|^2 = \sum_{i=M}^{\infty} h^2(i) \quad (12)$$

represents the tail portion of the AIR from M to infinity. For small values of μ this value is approximated by the Total Impulse response Power (TIP) to the uncancelled Tail Power (TP) of the AIR originally proposed by Knappe and Goubran [16], who show that the TIP/TP ratio defines the achievable ERLE up to approximately 20 dB. Beyond this point, other effects dominate. Actual ERLE measurements in [16] show that even at ratios of (S+N)/N of greater that 40 dB, the ERLE did not go beyond 25 dB. It was proposed that the most likely causes of this ERLE limitation is loudspeaker nonlinearities. The TIP/TP ratio is invaluable for determining the optimum number of AEC filter taps to use given a certain loudspeaker, microphone and enclosure. The impulse response of HFT #2 inside a conference room is shown in Figure 4 (a) and Figure 4 (b) shows the calculated TIP/TP vs. ERLE ratio compared to the measured ERLE. The ERLE will follow the TIP/TP ratio very closely up to a certain number of taps according to (10), however, in experimental recordings, nonlinearities and other effects limit the achievable ERLE.

3.6 Algorithmic Limitations

Dynamic Tracking in Nonstationary Conditions. The initial convergence of a particular algorithm identifies the room configuration, however as objects move and the input signal characteristics become nonstationary, the tracking ability of the algorithm becomes important. For example, although RLS based algorithms have fast convergence and have been shown theoretically to have tracking capability equivalent or better than the LMS algorithm in low noise [25], it has been found in [26] that algorithms based on instantaneous gradient estimates like the LMS family outperform RLS algorithms in conference room conditions using real speech where the SNR of the primary signal is often quite low.

Speech and Quasi-periodic Training signals. LMS based algorithms suffer from poor convergence when trained by quasi-periodic signals with highly coloured spectra, like speech. Often, a combination of architectures and algorithms is necessary to obtain satisfactory performance. A brief summary is presented in [15]. A comparative analysis of eight different algorithms is presented in [11] showing measured performance metrics (See Table 2) for the single talk mode only using both USASI noise signals and speech signals. Of eight algorithms tested, the generalized multi-delay filter (GMDF) [27], which is based on [28] obtains the best performance metrics. The unconstrained fast LMS [29] and wavelet decomposition technique [30] also produce good results. An algorithm presented in [31] uses a fast Newton training scheme to obtain performance enhancement with speech signals. However, measurements were obtained using a short impulse response (for use in automobile environments). No results were presented for an HFT in a highly reverberant venue.

Effect of Step Size on MMSE. The NLMS algorithm will produce a mean squared error J_{tot} that is in excess of the minimum mean-squared error J_{min} depending on the step size parameter α. The expression for J_{tot} for the Normalized LMS algorithm with stationary input signals can be approximated by;

$$J_{tot} = J_{min}\left(\frac{2}{2-\alpha}\right) \quad (13)$$

For a white noise input, the misadjustment is a factor of α only and that for α=1, a 3 dB increase in J_{min} can be expected. The exact theoretical form of the misadjustment depends on the specific algorithm employed and the input signal characteristics, but it is usually proportional to both α and the filter order M. For examples and theoretical forms on specific algorithms, see Widrow and Stearns [32]. Since J_{min} is generally a few orders of magnitude less than external environmental noise etc., this subject is not considered further.

3.7 Double Talk

Double talk (DT) occurs during periods when the far end speaker and near end speaker are simultaneously talking. The effect of DT is to increase the noise in the primary signal (similar to additive room noise described in Section 3.1) causing a temporary decrease in the ERLE and a slowing of the convergence and tracking ability. In a full duplex system, it is often necessary to freeze the adaptive filter coefficients such that divergence of the tap weights does not occur. The most drastic form of DT control is push-to-talk (half-duplex or single-talk mode) which was the defacto standard until the advent of adaptive filters for removing echo. The literature is full of techniques for performing DT, for example [27] describes a method of detecting local speaker activity by comparing the spectral shapes of the primary and reference signals, using an appropriate distance. A large distance is and indicator of the presence of a local talker. The method described in [33] proposes a short term cross correlation between the error output $e(n)$ and the canceller output $\widehat{y}(n)$ for controlling the step size. The correlation is used to obtain fast convergence during single-talk periods and low sensitivity during double-talk periods. Other methods are outlined in [34] and [24].

3.8 Summary

The major limitations to ERLE are caused by physical vibration, environmental noise, and algorithmic limitations like tracking ability in nonstationary conditions and convergence in highly coloured environments. Limitations like loudspeaker nonlinearity do have an effect on the achievable ERLE however the magni-

tude of distortion is highly dependent on the frequency and volume of the signal. The combination of these factors is illustrated in Figure 4(c) which shows relative levels of the achievable ERLE as a function of the physical limitations of undermodelling, room noise, vibration and nonlinear distortion.

4.0 METHODS TO IMPROVE PERFORMANCE

4.1 Vibration Isolation

A reduction in the vibration distortion can be obtained by (i) placing the microphone element inside a vibration isolator (ii) placement of the microphone in a vibrationally "silent" area in the enclosure (iii) selecting a microphone with low vibrational sensitivity. If all three precautions are taken, direct coupling between the loudspeaker and microphone is minimized. Figure 3(d) illustrates the PDF of the primary signals obtained with HFT #2 that has been designed with these purposes in mind. Note that the level of out-of-band distortion is reduced compared to the HFT measurements shown in Figure 3(b).

4.2 Nonlinear Loudspeaker Distortion Compensation

A method for successfully combating nonlinear loudspeaker distortion is presented here and is based on the work contained in [35]. The proposed structure is shown in Figure 5. The proposed structure consists of both nonlinear and linear sections. The nonlinear section consist of a two layer neural network that cancels the first part of the AIR where most of the energy is contained. The weight update equations for the nonlinear portion are based on the gradient backpropagation algorithm [36] with a normalized adaptive step size. The nonlinear node consists of a linearized hyperbolic tangent function which is linear for inputs below a user definable amplitude a, where $0 \le a \le 1$. Based on measurements reported in [37], the parameter a was set to 0.2 since it was found that this produced an ERLE approximately 1.5 dB higher than with a conventional (i.e. a=0) sigmoid. The node activation function $f(\,)$ is defined by;

$$f(s) = \begin{cases} s & ;|s| \le a \\ \text{sign}(s)\left[(1-a)\cdot \tanh\left(\dfrac{|s|-a}{1-a}\right)+a\right] & ;|s| > a \end{cases} \quad (14)$$

where s is the input. In Figure 5, the output $y(k)$ of the neural network portion at time k is defined by;

$$y(k) = w^{(2)}(k)x^{(2)}(k) + w_b{}^{(2)}(k) \quad (15)$$

$$x^{(2)}(k) = f(s(k)) \quad (16)$$

$$s(k) = \mathbf{w}^{(1)}(k)^T\mathbf{x}^{(1)}(k) + w_b{}^{(1)}(k) \quad (17)$$

where $\mathbf{x}^{(l)}(k)$ represents the input vector to layer l, $\mathbf{w}^{(l)}(k)$ represents the weight vector in layer l, $w^{(l)}{}_b(k)$ represents the single bias weight for layer l, $s(k)$ represents the input to the nonlinear node and T is the transpose operator. The weight update equations are described by;

$$\mathbf{w}^{(l)}(k+1) = \mathbf{w}^{(l)}(k) - \mu_{TDNN}(k)\delta^{(l+1)}(\dot{k}) \cdot \mathbf{x}^{(l)}(k) \quad (18)$$

$$w_b^{(l)}(k+1) = w_b{}^{(l)}(k) - \mu_{TDNN}(k)\delta^{(l+1)}(\dot{k}) \quad (19)$$

$$\delta^{(l+1)}(k) = \begin{cases} -2e_1(k) & ;l=2, \text{output layer} \\ f'(s(k))\delta^{(l+2)}(k)w^{(l+1)}(k) & ;l=1, \text{hidden layer} \end{cases} \quad (20)$$

where $e_1(k) = p(k) - y(k)$, $f'(\,)$ represents the derivative of the activation function at the input value $s(k)$, $\delta^{(l+1)}(k)$ represents the local gradient "delta" term in layer $l+1$, and $\mu_{TDNN}(k)$ is the normalized step size parameter defined by;

$$\mu_{TDNN}(k) = \frac{\alpha}{2 + \mathbf{x}^{(1)}(k)^T\mathbf{x}^{(1)}(k) + [x^{(2)}]^2} \quad (21)$$

The parameter α is a number between 0 and 2, and is set to 0.5. The weights in the linear portion of the proposed structure are updated using the NLMS algorithm;

$$\mathbf{w}_{FIR}(k+1) = \mathbf{w}_{FIR}(k) - \left[\frac{\alpha}{1 + \mathbf{x}_{FIR}(k)^T\mathbf{x}_{FIR}(k)}\right]e_2(\dot{k}) \cdot \mathbf{x}_{FIR}(k) \quad (22)$$

$$w_b(k+1) = w_b(k) - \left[\frac{\alpha}{1 + \mathbf{x}_{FIR}(k)^T \mathbf{x}_{FIR}(k)}\right] e_2(k) \quad (23)$$

Measurements: It was found that the vibration characteristics of HFT #1 limited the performance enhancement of the proposed algorithm. As a result, HFT #2 (which has improved vibration characteristics) was selected for the measurements presented here. Primary and reference signal are obtained using the method of Section 2.3 .These samples are then applied to both the proposed structure and a 600 tap linear adaptive filter which has DC bias compensation and weights updated in the same fashion as (22) and (23). In the proposed structure, the number of taps in the nonlinear section delay line equals 200 to cover the bulk of the loudspeaker impulse response. The number of taps in the linear section is 400 for a total impulse length of 600 taps. For each SPL, both algorithms are tested with the same input data of length 80,000 to allow convergence to a steady state at which point the average ERLE is measured and plotted.

The experimental results shown in Figure 5(c) show that at low volumes in the vicinity of 60 dB SPL, the proposed structure improves the ERLE by 3 dB as compared to the linear adaptive filter. As mentioned previously, at low level the two point suspension can cause nonlinear distortion. It would appear that in this range even though there is little nonlinear distortion in this range. In the low volume ranges, room noise becomes a dominant limitation and the proposed structure offers some improvement. In the medium volume range from 70-75 dB SPL, the proposed structure performs about 1 dB poorer than the linear filter due to an extra bias weight variance not included in the linear filter, and also because $f(s)$ will generate some small amount of distortion for any $|s| > a$ even when the inputs are linear. However, in the vicinity of 80 to 95 dB SPL where nonlinear effects dominate, the proposed structure clearly outperforms the linear filter in terms of converged ERLE and demonstrates over 8 dB improvement at 90 dB SPL.

Figure 6 shows the error PSDs as obtained by using (a) and FIR structure trained using the NLMS algorithm and (b) the proposed nonlinear structure. The error signal for the proposed structure is smaller than in the FIR case, however, the nonlinear structure regenerates some error near the sampling frequency. In gen-

eral, a nonlinear filter can create energy at frequencies not present in the input signal [38], and this effect is evidenced here.

5.0 CONCLUSIONS

We have investigated the physical performance limitations in handsfree telephones and have determined that primary signal noise, enclosure vibration, loudspeaker nonlinearity, choice of algorithm, and measurement noise all contribute to limit performance. Experimental results showing the relative magnitude of the vibration and loudspeaker nonlinearity have been presented since literature about these limitations is sparse. Methods to reduce the vibrational coupling have been presented along with a new nonlinear algorithm based on the application of a time domain neural network. Neural networks are relatively new to the realm of digital signal processing, however, the clear improvements provided by the nonlinear filter suggest that further work with structures of this nature are warranted.

Physical measurements have determined that resonances and vibrations within the enclosure can be a more serious limitation than nonlinear distortions generated within the loudspeaker. In fact, the results presented in Section 4.2 were obtained on HFT #2 since results (not presented here) obtained using HFT #1 with components mounted inside the enclosure failed to improve the ERLE in the nonlinear range. It seems clear in this case that vibration distortion masks loudspeaker nonlinearity and that addressing this effect is necessary before the application of an algorithm to identify signal characteristics that would be masked by such distortion.

6.0 ACKNOWLEDGEMENTS

The authors wish to thank NSERC, Carleton University, Nortel and the Telecommunications Research Institute of Ontario (TRIO) for their financial support.

7.0 REFERENCES

[1] S. Haykin, *Adaptive Filter Theory*, 3rd ed., Prentice-Hall Information Systems Series, 1996.

[2] A. Gilloire, “Performance Evaluation of Acoustic Echo Control: required Values and Measurement Procedures”, *Annales des Telecommunications*, Vol. 49, No. 7-8, Jul.-Aug. 1994, pp. 368-372.

[3] R. Wehrmann, J.V.D. List, P. Meissner, “A Noise Insensitive Compromise Gradient Method for the Adjustment of Adaptive Echo Cancellers”, *IEEE Trans. Comm*. COM-28, No. 5, 1980, pp. 753-759.

[4] *** "General Characteristics of International Telephone Connections and International Telephone Circuits: Acoustic Echo Controllers" Recommendation G.167, *ITU-TSS* (03/1993)

[5] *** "General Characteristics of International Telephone Connections and International Telephone Circuits: Echo Ccancellers" Recommendation G.165, *ITU-TSS* (03/1993)

[6] *** "Transmission Performance of Group Audio Terminals (GATS)", Recommendation P.30 *CCITT Blue Book* (1988) Vol. 5.

[7] ***, "Transmission Characteristics of Hands-free Telephones", Recommendation P.34, *CCITT Blue Book* (1988), Vol. 5.

[8] *CCITT Blue Book* (1988) P.34, P.50, P.51,P.76,P.79,and supplement No. 2 of the P recommendations.

[9] *** "Technical Characteristics of Telephony Terminals- Part 3: PCM A-law, loudspeaking and handsfree function", *ETSI draft I-ETS* 300 245-3: ISDN (April, 1993).

[10] H. Yasukawa, M. Ogawa, M. Nishino, “Echo Return Loss Required for Acoustic Echo Controller based on subjective assessment”, *IEICE Trans. E* ., Vol. 74, 1991, pp 629-705.

[11] P. Naylor, J. Alcazar, J. Boudy, Y. Grenier, “Enhancement of Hands-free Telecommunications”, in *Annales des Telecommunications*, Vol. 49, No. 7-8, Jul.-Aug. 1994, pp.373-379.

[12] H. Kuttruff, *Room Acoustics*, London, U.K.: Elsevier, 1991, 3rd. Ed.

[13] M. Mboup, M. Bonnet, “IIR Filtering for Acoustic Echo Cancellation”, *Asilomar Conference on Signals, Systems and Computers*, Vol. 1, pp. 203-206, November 1991.

[14] S. Gudvangen, S.J. Flockton, “Modelling of Acoustic Transfer-Functions for Echo Cancellers”, in *IEE Proceedings on Vision , Image and Signal Processing*, 1995, Feb. Vol. 142, No.1, PP. 47-51.

[15] S. McCaslin, N. Van Bavel, “Effects of Quasi-periodic Training Signals on Acoustic Echo Canceller Performance”, in *Annales des Telecommunications*, Vol. 49, No. 7-8, Jul.-Aug. 1994, pp. 380-385.

[16] M.E. Knappe, R.A. Goubran,”Steady State Performance Limitations of Full-Band Acoustic Echo Cancellers”, *ICASSP* 1994, Adelaide, South Australia, Vol. 2, pp. 73-76.

[17] C. Caraiscos, B.Liu, “A Roundoff Error Analysis of the LMS Adaptive Algorithm”, *I.E.E.E. Trans. on Acoustics, Speech and Sig. Proc*. Vol. ASSP-32, No. 1, Feb. 1984, pp. 34-41.

[18] S. Makino, Y. Keneda, N. Koizumi, “Exponentially Weighted Stepsize NLMS Adaptive Filter Based on the Statistics of Room Impulse Response”,*IEEE Trans. on Speech and Audio Proc.*, Vol. 1, No. 1, Jan 1993, pp. 101-108.

[19] X. Y. Gao, W. M. Snelgrove, “Adaptive Linearization of a Loudspeaker”, *ICASSP* 1991 Vol. 3, pp 3589-3592.

[20] A.J.M. Kaizer, *On the Design of Broadband Electrodynamic Loudspeakers and Multiway Loudspeaker Systems*, Ph.D. Thesis, Eindhoven University of Technology, The Netherlands, 1986, Chapter 6.

[21] P. Chang, C. Lin, B. Yeh, “Inverse Filtering of a Loudspeaker and Room Acoustics Using Time-delay Neural Networks”, *Journal of the Acoustic Society of America*, Vol 95, No. 6, June 1994, pp. 3400-3408.

[22] H.F. Olsen, *Acoustical Engineering*, Toronto, D. Van Nostrand Company,Inc., 1964.

[23] Dr. A. Van Shyndel, Nortel, Personal Communication.

[24] R. D. Poltmann, “Stochastic Gradient Algorithm for System Identification Using Adaptive FIR-Filters with too Low Number of Coefficients”, *IEEE Trans. on Circ. and Syst.*, Vol. 35, No. 2, Feb. 1988, pp. 247-250.

[25] E. Eleftheriou, D.D. Falconer, “Tracking Properties and Steady State Performance of RLS Adaptive Filter Algorithms”, *IEEE Trans. on Acoust., Speech and Sig. Proc*., Vol ASSP-34, No. 3, June, 1986, pp. 499-510.

[26] H. Yuan, *Dynamic Behavior of Acoustic Echo Cancellation*, M. Eng. Thesis, Carleton University, Ottawa, Canada, 1994.

[27] J. Prado, E. Moulines, “Frequency Domain Adaptive Filtering with Applications to Acoustic Echo Cancellation”, in *Annales des Telecommunications*, Vol. 49, No. 7-8, Jul.-Aug. 1994, pp. 414-428.

[28] J. Soo, K. Pang, “Multidelay Block Frequency Domain adaptive Filter”, *IEEE Trans. on Acoustics, Speech and Sig. Proc*., Vol. 38, No. 2, Feb. 1990, pp. 373-376.

[29] D. Mansour, A. Gray, “Unconstained frequency domain adaptive filter”, *IEEE Trans. ASSP*, Vol. 30, No. 5, pp. 726-734, 1982.

[30] O. Rioul, M. Vetterli, “Wavelets and Signal Processing”, *IEEE Signal Processing Magazine*, Oct. 1991, pp.14-38.

[31] T. Petillon, A. Gilloire, S. Theodoridis, “A Fast Newton Transversal Filter: An Efficient Scheme for Acoustic Echo Cancellation in Mobile Radio”, *IEEE Trans. on Sig. Proc.* Vol. 42, No. 3, March 1994, pp. 509-518.

[32] B. Widrow, S. D. Stearns, *Adaptive Signal Processing*, Englewood Cliffs, N.J.: Prentice-Hall, 1985.

[33] H. P. Meana, et. al., “A Time Varying Step Size Normalized LMS Echo Canceler Algorithm”, *Proceedings ICASSP 1994*, Vol 2, pp. 249-252.

[34] S. M. Kuo, Z. Pan, “Distributed Acoustic Echo Cancellation System with Double-talk Detector”, *J. Acoust. Soc. Am*. No. 6, Dec.1993, pp. 3057-3060.

[35] A.N. Birkett, R. A. Goubran, “Limitations of Handsfree Acoustic Echo Cancellers due to Nonlinear Loudspeaker Distortion and Enclosure Vibration Effects”, in *1995 IEEE ASSP Workshop on Appl. of Sig. Proc. to Aud. and Acoustics*, New Paltz, New York, Oct. 1995.

[36] Y. Pao, *Adaptive Pattern Recognition and Neural Networks*, Addison-Wesley Publishing Company, Inc, 1989.

[37] A.N. Birkett, R. A. Goubran, “Acoustic Echo Cancellation Using a NLMS-Neural Network Structures”, *ICASSP* 1995, Detroit, MI., Vol. 5, pp. 3035-3038.

[38] W. G. Knecht, “Nonlinear Noise Filtering and Beamforming Using the Perceptron and Its Volterra Aproximation”, *IEEE Trans. Audio and Acoustics*, Vol. 2, No. 1, Jan. 1994, pp. 55-62.

[39] Y. Haneda, S. Makino, Y. Kaneda, “Common Acoustical Pole and Zero Modelling of Room Transfer Functions”, *I.E.E.E. Transactions on Speech and Audio Proc.* Vol. 2, No. 2, April 1994, pp. 320-328.

[40] H. Schutze, “Convergence of Acoustic Echo Cancelers for Hands-Free Telephones Operating Under Feedback Conditions”, *IEEE Trans. on Speech and Audio Proc.*, Vol. 1, No. 2, April 1994, pp. 241-249.

[41] P.L. De Leon, D. M. Etter, “Experimental Results with Increased Bandwidth Analysis Filters in Oversampled Subband Acoustic Echo Cancelers”, IEEE Signal Processing Letters, Vol. 2, No. 1, Jan. 1995, pp. 1-3.

[42] D. R. Morgan, “Slow Assymptotic Convergence of LMS Acoustic Echo Cancellers”, *IEEE Transactions on Speech and Audio Processing*, Vol. 3., No. 2, March 1995, pp. 126-136.

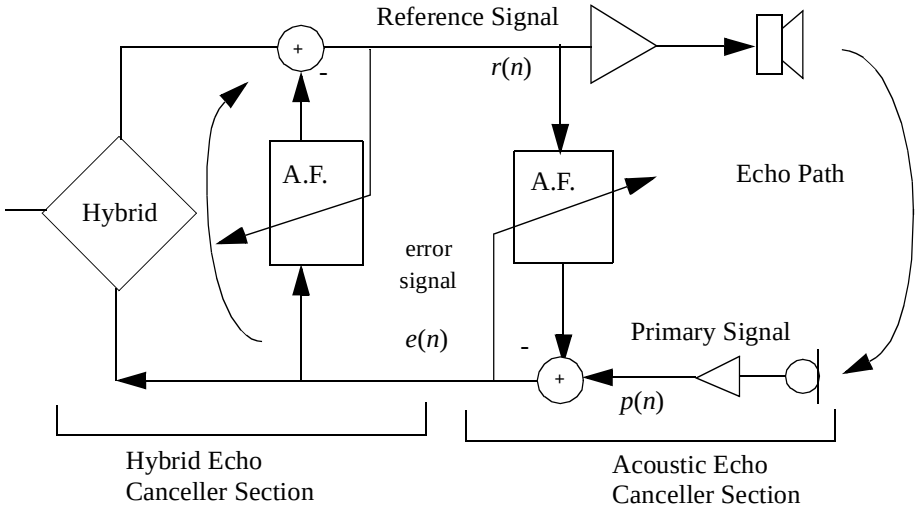
[43] J. G. Ryan, M. E. Knappe, H. Yuan, R.A. Goubran, T. Aboulnasr, “Audio Quality Enhancement in Noisy Environments”, *Final Report submitted to Data Network Planning Section*, Bell Canada, May 1992.

[44] E. Hansler, “The Hands-Free Telephone Proplem: An Annotated Bibliogray Update”, *Ann. Telecommun.* Vol. 49, No. 7-8, 1994, pp. 360-367.

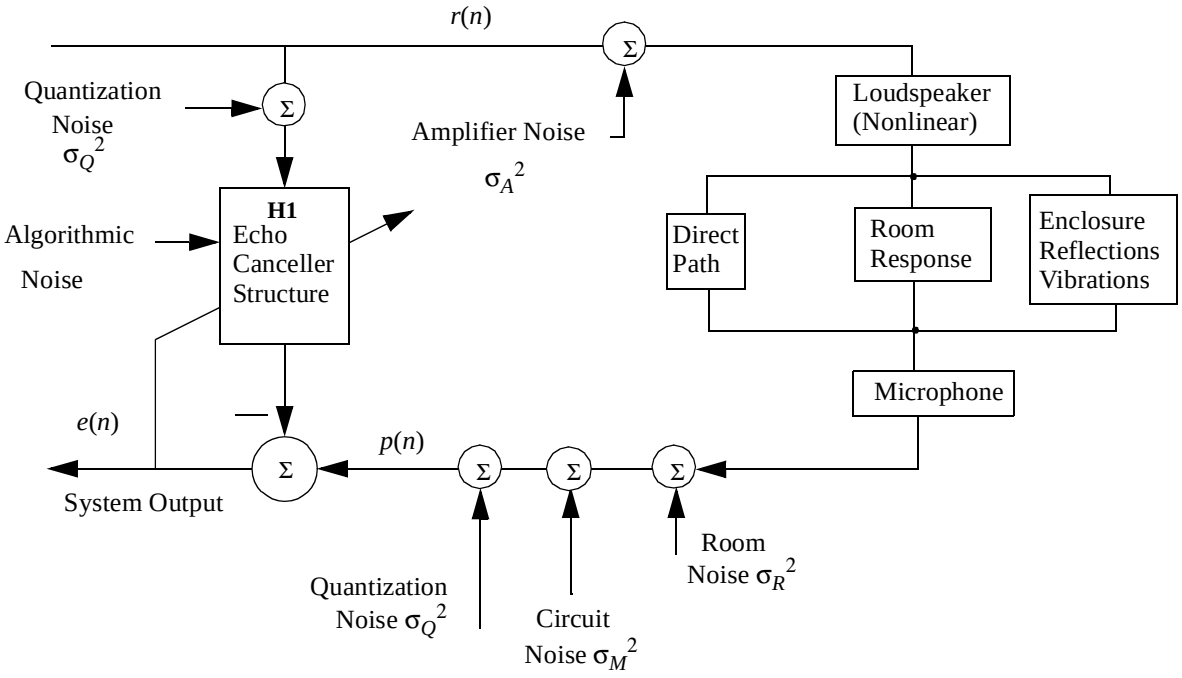
[45] E. Martine, A. Gilloire, P. Le Scan, “CAD of Signal processing Architectures: An Application to Acoustic Echo Cancellation”, in *Annales des Telecommunications*, Vol. 49, No. 7-8, Jul.-Aug. 1994, pp. 447-459

8.0 ILLUSTRATIONS

.



(a)



(b)

FIGURE 1. (a) Adaptive Acoustic Echo Canceller Structure. The hybrid echo canceller is also shown for reference. Variables *p*(*n*) and *e*(*n*) are the primary and error signals. (b) Complete AIR model includes enclosure reflections and vibration as well as transducer nonlinear responses. The model should also include room, quantization and circuit noise.

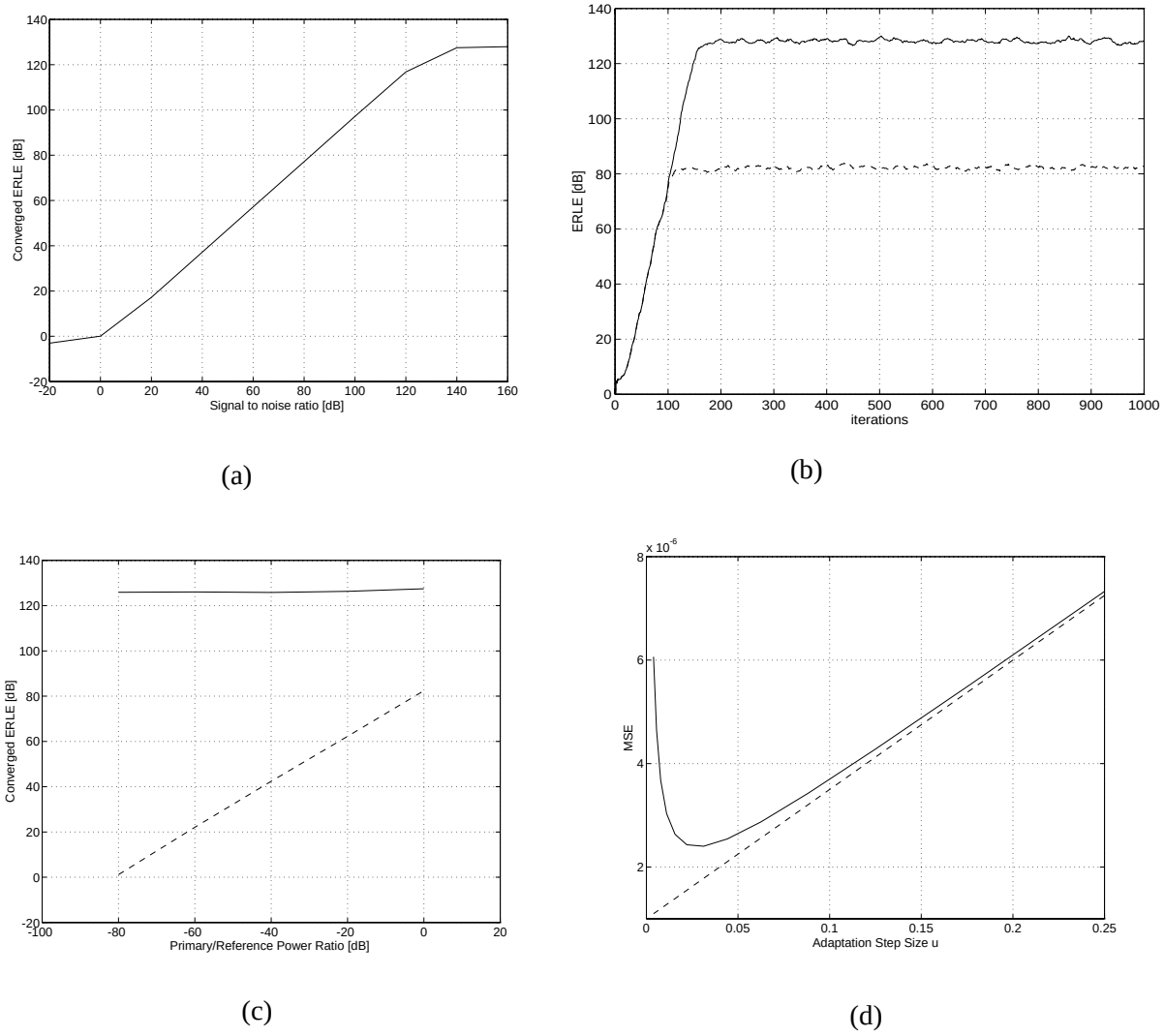


FIGURE 2. (a) Converged ERLE for additive noise in the primary path. At low noise levels, the ERLE is limited by the floating point noise floor of 127 dB. At high noise levels, the algorithm is unable to converge properly. (b) Comparison of converged ERLE curves using the NLMS algorithm with a white noise input. Quantizing the primary signal to 15 bits plus sign results in a loss of accuracy and limits the achievable ERLE. (c) Converged ERLE for a variable primary signal level where the primary signal is quantized to 15 bits plus sign (dotted line) and for an unquantized primary signal (solid line). (d) Total MSE as a function of μ where $B_d=B_c=16$ bits, MMSE=1e-6, M=500 taps, reference input variance $\sigma_r^2=0.1$. The dashed line shows the equivalent infinite precision case.

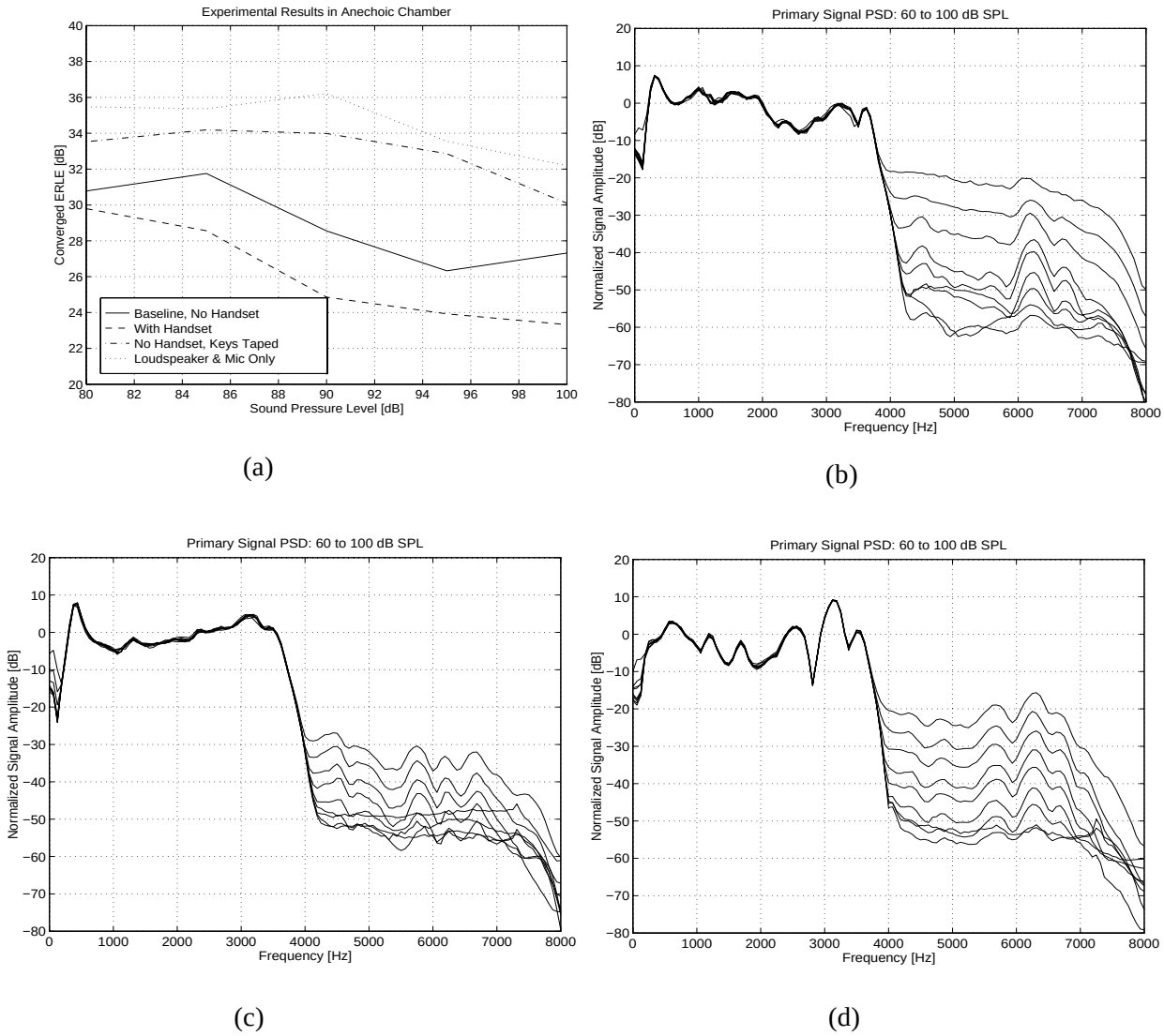
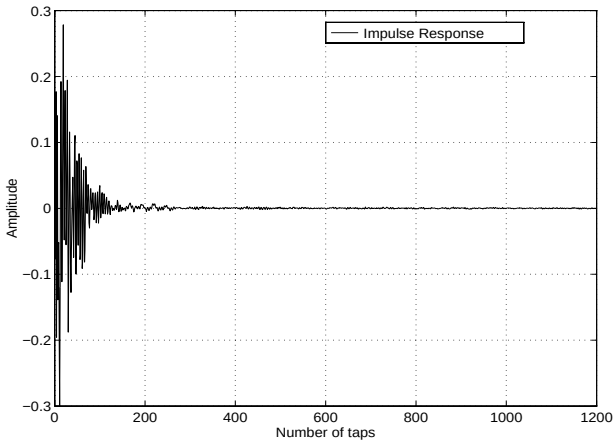
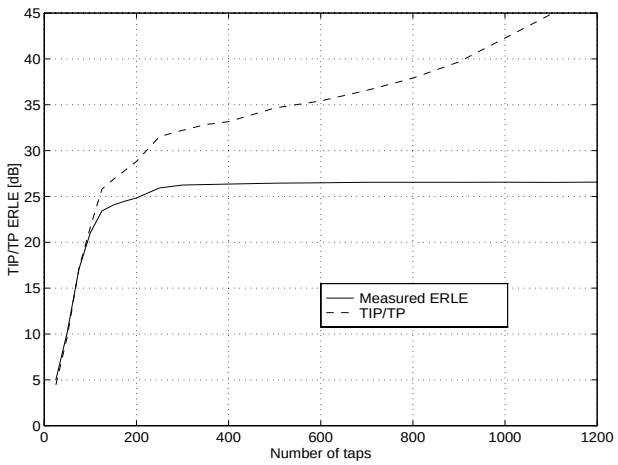


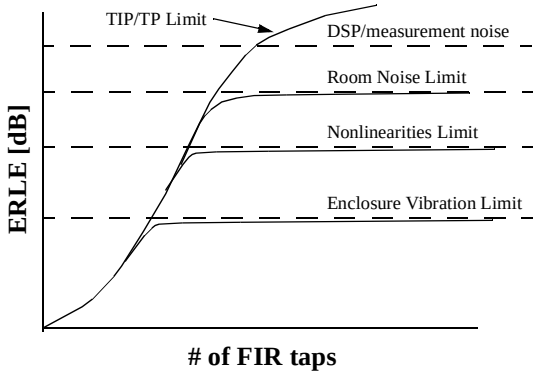
FIGURE 3. Effects on achievable ERLE due to nonlinear distortion, vibration and rattling. (a) HFT #1. Converged ERLE as volume in increased from 60 dB SPL to 100 dB SPL (b) HFT #1. Primary signal PSD with loudspeaker and microphone inside the HFT enclosure. Out-of-band components increase in level as the volume is increased from 60 dB SPL to 100 dB SPL. (c) same as (b) but with components removed from enclosure and mounted inside a standard baffle. (d) HFT #2 PSD with improved vibration characteristics.



(a)

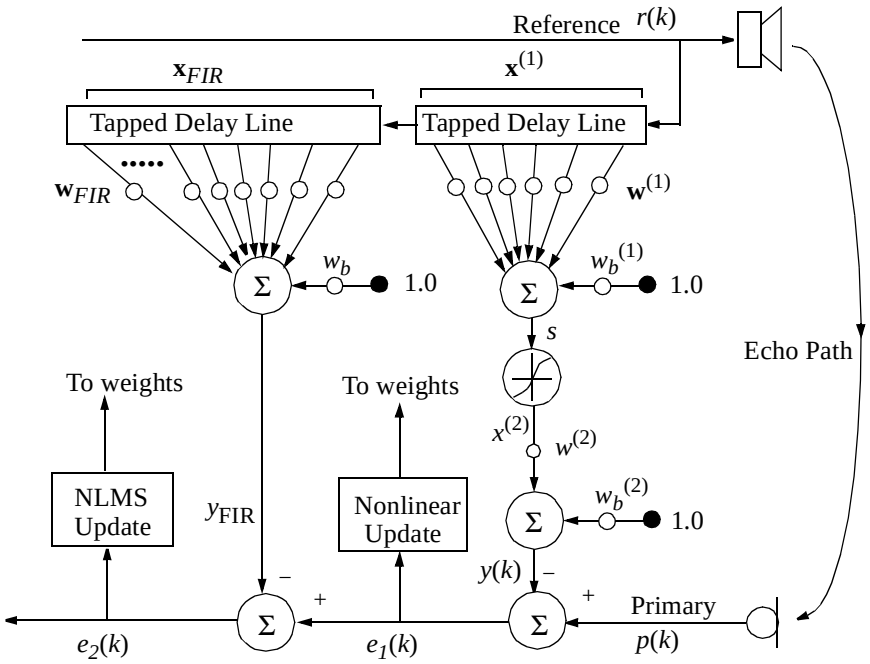


(b)

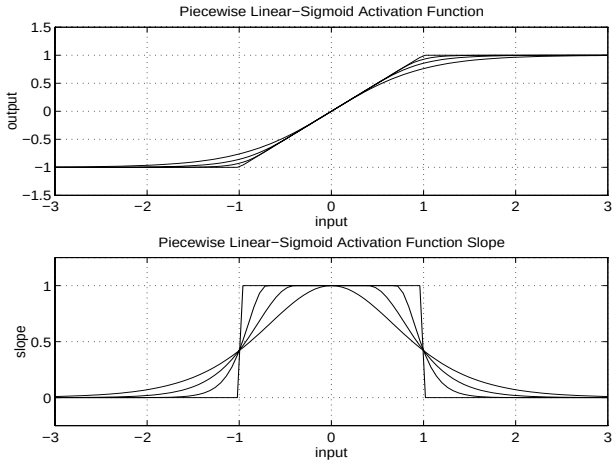


(c)

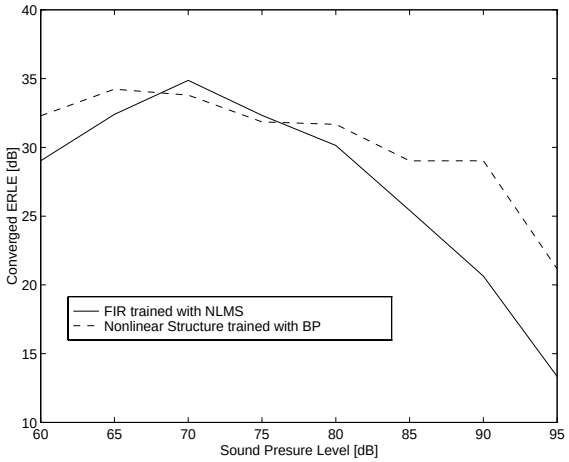
FIGURE 4. HFT #2. (a) Impulse response measured in a furnished conference room. (b) calculated TIP/TP and measured ERLE using NLMS algorithm. (c) Achievable ERLE as a function of physical limitations.



(a)



(b)



(c)

FIGURE 5. Nonlinear structure used to combat nonlinear distortion. (a) structure consist of a cascade of a conventional FIr adaptive filter and a 2 layer neural network. (b) piecewise linear activation function used as the nonlinear element. (c) Experimental results for the proposed structure.

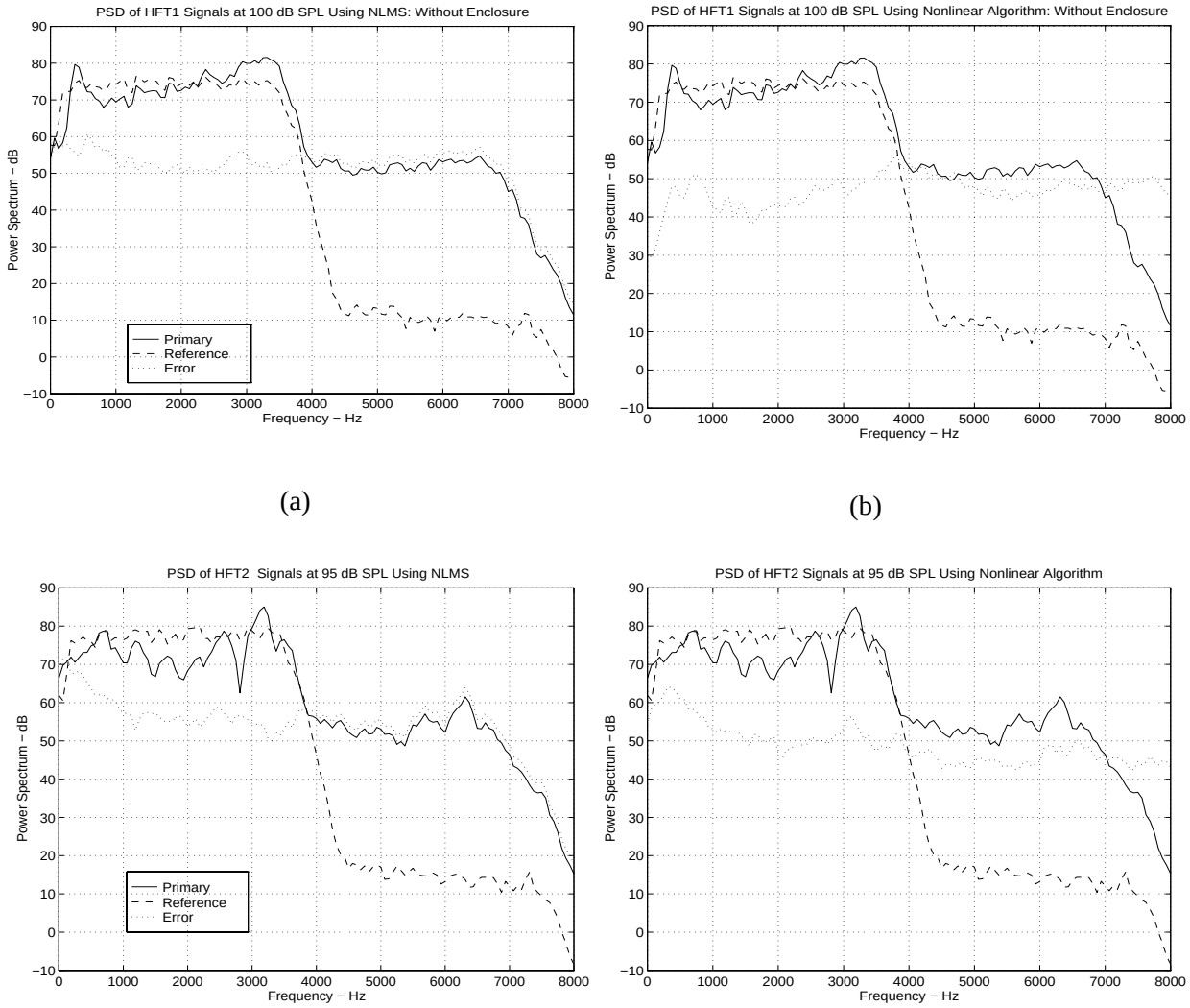


FIGURE 6. Comparison of primary, reference and error signal PSDs at 100 dB SPL. (a) HFT #1, components removed. FIR trained with the NLMS algorithm (b) HFT #1, components removed. Nonlinear structure trained with backpropagation. Error is lower than (a) but the preprocessor regenerates distortion products near the Nyquist frequency. (c) HFT #2, unmodified. NLMS algorithm. (d) HFT #2, unmodified. nonlinear algorithm.