ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

Acoustic Echo Cancellation by Informax Based Source Deflation of the Loudspeaker Signal

Thato Tsalaile, Ephraim Gower, Monageng Kgwadi, Shedden Masupe

Abstract— ***In this paper, an informax based source deflation algorithm of the loudspeaker (far-end) signal for acoustic echo cancellation is introduced. The aim is to continually model the loudspeaker-environment-enclosure filter even under double-talk and noisy conditions, something the current methods fail to do. The deflation filter is learned using the informax principle where a prior knowledge about the near-end signal's approximate probability density function is required for optimal filter convergence. Simulation results are used to illustrate the performance of the algorithm under double-talk conditions, as well as simulation comparisons to the normalized least-mean-square algorithm for echo cancellation under varying noise conditions with no double-talk***

Index Terms—**acoustic echo cancellation, double talk detection, informax principle.**

I. INTRODUCTION

The current acoustic echo cancellation (AEC) algorithms mostly utilize adaptive filtering techniques optimized using the least mean square (LMS) algorithm and its variants [1-3].The loudspeaker signal $l($ is filtered by the loudspeaker-environment-enclosure (LEM) filter $h_l($ resulting in the far-end signal $d(n) = h_l(n) * l($ for $n\epsilon[0, N -$ where is the convolution operator. It is this signal that must be cancelled from the transmitted signal, so that it is only the near-end signal $m(n) = h_s(n) * s($ that is played through the loudspeaker, be it for teleconferencing or public announcement systems, where $h_s($ is the talker-environment-microphone (TEM) filter and $s($is the talker signal, for $n\epsilon[0, N -$. These two signals, along with room noise $v($ are captured by the microphone such that the output is $x(n) = d(n) + m(n) + v($, for $n\epsilon[0, N -$. With no observed near-end signal, an adaptive filter $\widehat{h}($ is used to model the LEM filter, and then using the available loudspeaker signal, the far-end signal estimate $\widehat{d}(n) = \widehat{h}(n) * l($ is obtained, for $n\epsilon[0, N -$. The estimated signal is subtracted from the microphone signal resulting in the error signal $e(n) = r(n) + v($, for the residue signal $r(n) = d(n) - \widehat{d}($for $n\epsilon[0, N - 1$Given minimal room noise, the adaptive filter can converge to the LEM filter resulting in effective echo cancellation. However, this echo cancellation process fails when then near-end signal is observed (double-talk) as the adaptive filter diverges from the desired LEM filter, and the same thing occurs under excessive noise conditions.

To avoid the problem of filter divergence due to the near-end signal or excessive noise, double-talk-detection (DTD) algorithms such as those based on cross-correlation methods [4,5] and other variants, are used to detect the presence of the near end signal as well as excessive noise, to freeze the LEM filter modeling process.

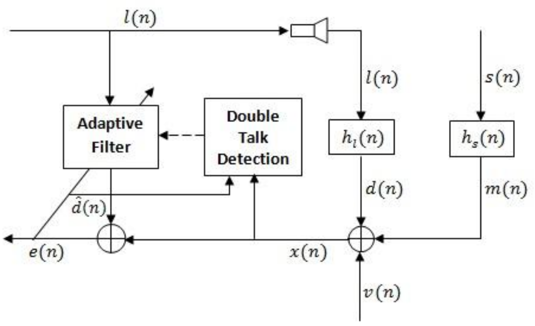

Fig. 1. An adaptive filter is used to model the LEM enclosure filter, then using the available loudspeaker signal an estimate for the far-end signal is subtracted from the microphone signal given no double-talk-detection

The frozen coefficients of the adaptive filter are continually used to estimate the far-end signal, and given the LEM path does not change much during these periods, the echo signal can still be cancelled. The system is illustrated in Fig. 1. However, the reality is that the near-end signal is frequently observed, especially in public announcement systems, and in a continually changing LEM enclosure. It is for this reason that adaptive filtering techniques often fail in echo cancellation due to the excess mean square error of the adaptation algorithm. In this paper, we propose the use of a source deflation or cancellation algorithm where the deflation filter is directly learned using the informax principle to cancel the echo signal $d($ By employing this approach, the far-end (or echo) signal can still be cancelled even under double-talk conditions. The main idea is to address the limitations of conventional echo cancellers when there is double-talk and excessive noise. This paper is structured as follows: In Section II, we introduce the proposed source deflation algorithm. The simulation results are presented in Section III. Discussion and summary remarks are in Section IV.

II. THE PROPOSED AEC ALGORITHM

The proposed AEC algorithm is illustrated in Fig. 2, where the inputs to the source deflation algorithm are the microphone signal $A. \; x_1$ and the loudspeaker signal $A. \; x_2(n) = l$ for $A. \; n\epsilon[0, N$ -. Besides echo suppression, it is further desired that the near-end signal remain unfiltered so that the natural effects of the room are preserved in the communication signal.

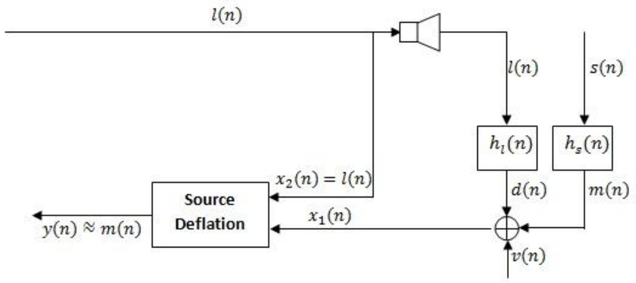

Fig. 2. Using the available loudspeaker signal, a deflation filter is used to suppress the far-end signal from the microphone signal even under doubletalk and noisy conditions. The processed signal is an estimate of the near-end signal.

A. Problem Formulation

Given the microphone signal $x_1(n) = h_1(n) * l(n) + h_s(n) * s($ and the loudspeaker signal $x_2(n) = l($, we have the mixing model.

$$\begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix} = \begin{pmatrix} h_l(n) & h_s(n) \\ \delta(n) & 0 \end{pmatrix} * \begin{pmatrix} l(n) \\ s(n) \end{pmatrix}$$

for all $n \epsilon [0, N -$, where $\delta($ is the n coefficient of the Dirac impulse response, assuming causal room impulse response (RIR) filters of length . The task is to find the loudspeaker signal deflation matrix W) such that

$$\begin{pmatrix} y(n) \\ x_2(n) \end{pmatrix} = \begin{pmatrix} w_{11}(n) & w_{12}(n) \\ w_{21}(n) & w_{22}(n) \end{pmatrix} * \begin{pmatrix} x_1(n) \\ x_2(n) \end{pmatrix}$$

for all $n \epsilon [0, N -$. From the model we have $x_2(n) = [w_{21}(n) * h_l(n) + w_{22}(n)] * l(n) + w_{21}(n) * h_s(n) * s(n)$, where it is known that $x_2(n) = l($ Therefore it is necessary that $w_{21}(n) =$ and $w_{22}(n) = \delta($, for all $n \epsilon [0, N -$. The resulting signal after deflation is given by $y(n) = [w_{11}(n) * h_l(n) + w_{12}(n)] * l(n) + w_{11}(n) * h_s(n) * s(n)$, where the goal is to end up with $y(n) = h_s(n) * s(n) = m($, the unfiltered near-end signal. This implies that $w_{11}(= \delta($, and therefore we have

$$y(n) = [h_l(n) + w_{12}(n)] * l(n) + m(\qquad (1)$$

for all $n \epsilon [0, N -$. Based on (1) the task of finding a $2 \, b$ deflation matrix has been reduced to learning an optimal vector $\widehat{w_{12}(n)} = -h_l($ for all $n \epsilon [0, N -$. It is interesting to note that (1) requires the estimation of the LEM enclosure filter just like the conventional LMS methods do, leading to subtraction of the estimated far-end signal from the observed microphone signal. In other words, $y(n) = d(n) - \bar{d}(\bar{n}) + m(n) = r(n) + m($ The advantage here is that it would be possible to suppress the echo even under double-talk conditions, which is the main aim of this paper. Clearly, it is not possible to use LMS variants as the presence of the near-end signal would lead to filter divergence. We propose the use of the informax principle to learn the optimal echo deflation vector $\bar{w}_{12}(\bar{}$ for all $n \in [0, N -$.

B. Learning the Echo Deflation Filter

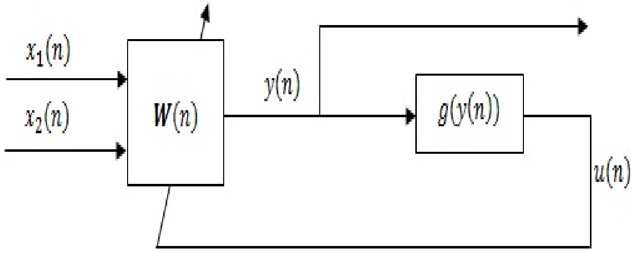

Fig. 3. The processed signal $y(n)$ is passed through a non-linear squashing function to give the output $u(n)$, which is used to optimize the deflation filter $w_{12}(n)$. Some knowledge of the density function of $m(n)$ is useful to set $g(\cdot)$ as close as possible to its cumulative distribution.

Fig. 3 illustrates the use of the information-maximization (informax) process to learn the optimal filter $A. \; \widehat{w_{12}(n)} \approx -h_l$ for echo suppression, for all $A. \; n \epsilon [0, N$ -. A non-linear point-to-point squashing function $A. \; g$ is used to give the signal $A. \; u$ for all $A. \; n \epsilon [0, N$ -. From this, the probability distribution of $A.$ (the variable realized by samples of the signal $A. \; u($ can be written as

$$A. \; p(u) = \frac{}{\left|\frac{\partial}{\partial}\right|}$$

Which means that if the non-linear function $u = g(y)$ is chosen as the cumulative distribution of the variable Y (realized by samples of the signal $y(n)$), then $p(u)$ reduces to a uniform density function. It is well known that the uniform density function has the most differential entropy (uncertainty associated with a variable) as all values are equally likely. Therefore, maximum information is transferred from the input to the output when $p(u)$ matches a uniform distribution, and it is this informax principle that we shall use to learn the optimal deflation filter $\widehat{w_{12}(n)}$., for all $n \epsilon [0, N - 1]$. For further reading on the informax principle, the reader is referred to [6].

It is desired that the output signal $y(n) \approx m(n)$, thus some a prior knowledge about the density function of the near-end signal would serve well in determining the function $g(.)$. If $u \approx g(m)$, then the output signal $y(n) \approx m(n)$, the desired near-end signal. In [6], it was observed that speech signals tend to have a leptokurtic density function which resembles the Laplacian or double exponential distribution. This lead to the use of $u(n) = tanh \, y(n)$ as the non-linear squashing function, a function we shall adopt in the derivations.

Remark: In public announcement systems, the echo signal is the delayed near-end signal. This means that it is possible to estimate the type of density function from the available loudspeaker signal $A.\ l$ which can lead to more effective echo suppression.

Using the deflation matrix $A.\ \mathbf{W}$ where $A.\ w_{11}\ (n) = w_{22}\ (n)) = \delta(n), and\ w_{21}\ (n)$, the processed signal can be written as

$$A.\ y(n) = x_1\ (n) + \sum_{k=0}^{L-1} w_{12}(k)\, x_2(n - \qquad (2)$$

for all $A.\ n\epsilon[0, N$ -.The differential entropy of the output $A.\ u$ with respect to the loudspeaker signal $A.\ x_2$ for $A.\ n \in [0, N$ - is given by

$$A.\ -E[ln\, p(u)] = -E\left[ln \frac{p(}{\frac{\partial u}{\partial x_2}}\right.$$

which can be expanded to

$$-E[ln\, p(u)] = -E[ln\, p(x_2(\mathrm{n}))] + E\left[ln \left|\frac{\partial u(n)}{\partial x_2(n)}\right|\right]$$

where the loudspeaker signal $A.\ x_2$ is chosen as it is the one directly processed by the filter of interest $A.\ w_{12}$ for all $A.\ n\epsilon[0, N$ -, and $A.$ denotes mathematical expectation whereas the ratio $A.\ \frac{\partial}{\partial x}$ defines how the loudspeaker signal affects the output signal $A.\ u$ for $A.\ n\epsilon[0, N$ -. Since the task is to find the optimal filter $A.\ \widetilde{w_{12}}$, for all $A.\ n\epsilon[0, N$ -, such that the amount of information transferred from the input to the output is maximal, the contrast function can be written as

$$A.\ \widehat{w_{12}(n)} = arg\, max_{w_{12}(n)}\, E\left[ln \left|\frac{\partial u(}{\partial x_2(}\right.\right. \qquad (3)$$

for all $n\epsilon[0, N-1]$, with the term $-E[ln\ |p(x_2)|]$ left out because the deflation filter coefficients are not dependent on the observed signals. Using the chain rule

$$\frac{\partial u(n)}{\partial x_2(n)} = \frac{\partial u(n)}{\partial y(n)} \cdot \frac{\partial y(n)}{\partial x_2(n)}$$

From (2) then $\frac{\partial u(n)}{\partial x_2(n)} = w_{12}(0)$ and $u\,'(n) = \frac{\partial u(n)}{\partial y(n)} = 1 - u^{\,2}(n)$, for the non-linear squashing function $u(n) = tanh\ y(n).$ It follows that $ln\ \left|\frac{\partial u(n)}{\partial x_2(n)}\right| = ln\ |u'(n)\,\Delta w_{12}(0)|$. For optimization of the contrast function given by (3) using the gradient ascent rule, the adaptation step for the first coefficient is such that

$$\Delta w_{12}(0) \propto \frac{\delta\ ln\,|u'(n)\Delta w_{12}(0)|}{\delta\ w_{12}(0)},$$

which evaluates to

$$\Delta w_{12}(0) \propto \frac{1}{w_{12}(0)} - 2u(n)x_2(n) \qquad (4)$$

For $k\epsilon[1, L-1]$, then following the same derivations the learning rule is

$$\Delta w_{12}(k) \propto -2u(n)x_2(n-k) \qquad (5)$$

III. RESULTS AND ANALYSIS

The LEM and TEM impulse responses were generated using MATLAB for the length $K = 10$, and these are plotted in Fig.4. The two speech signals of length $N = 200$ samples used as the loudspeaker and talker signals as well as the resulting microphone output with addition of Gaussian noise for a signal to noise ratio (SNR) of $30c$ are illustrated in Fig.5.

The learning rules (4) and (5) were used to learn the deflation filter coefficients for echo cancellation, after which the cross-correlation measure between the loudspeaker signal $l($ and the output $y($ for all $n\epsilon[0, N-$, is used to illustrate the effectiveness of the proposed AEC algorithm. Low values of the cross-correlation indicate good separability as opposed to high values. The cross-correlation coefficient is given by

$$C(l(n), y(n)) = \frac{E[(l(n)-\mu_l)^T(y(n)-\mu}{\sigma_l\sigma_y}, \qquad (6)$$

where and are the mean and standard deviation of the loudspeaker signal $l($ with and as the mean and standard deviation of the processed signal $y($ for all $n\epsilon[0, N-$, After deflation, (6) was used to measure the residual echo, for which the correlation coefficient was $C(l(n), y(n)) = 0.$. The results are illustrated in Fig. 6. There are other possibly more accurate measures for determining the deflation of the loudspeaker signal in the signal $y($ such as measuring the mutual information via spacing estimates of entropy [7], but it is the computational simplicity of (6) that makes it attractive as well as its experimental success as a measure for DTD applications, which is essentially what we are doing. For example, no double-talk implies successful source deflation and this is reflected by a low correlation value.

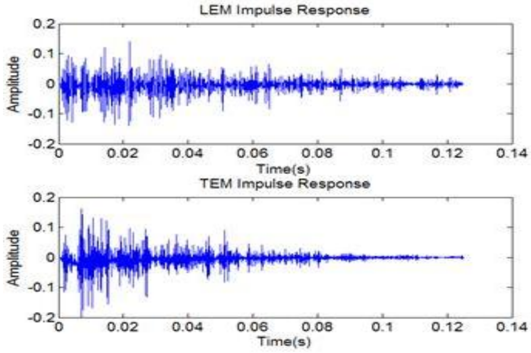

Fig. 4. Plots for the LEM and TEM impulse responses for filter length K = 1024.

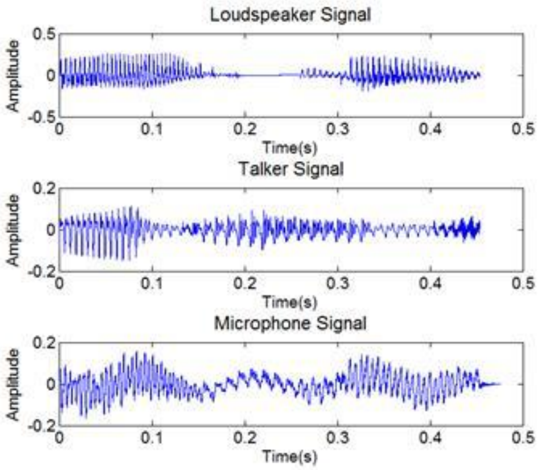

Fig.5. The loudspeaker (top) and talker (middle) signals along with Gaussian noise are used to give the microphone (bottom) signal with a SNR=30dB.

The proposed algorithm is further simulated in comparison with the normalized least-mean-square (NLMS) algorithm under varying noisy conditions, where the echo residue is again reflected by the values of (6). The results are illustrated in Fig. 7, where the informax based deflation algorithm achieves better echo cancellation results than the NLMS method. This is because the conventional approach adapts the echo cancellation filter via the minimization of the error signal, which under noisy conditions is $e(n) = d(n) - \widehat{d(n)} + v($ for all $n \in [0, N -$ The objective of this approach is to match $\hat{d}($ to $d($ and this is not possible due to additive noise, much in the same way that the near-end signal causes the adaptive filter to diverge. The deflation algorithm works by suppressing the loudspeaker signal from the microphone signal, which is still possible in the presence of noise and therefore the processed signal is usually echo free.

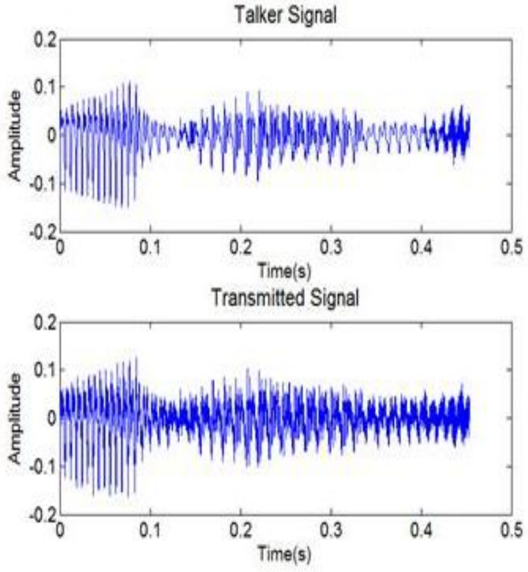

Fig. 6. The transmitted signal resembles the talker signal where the correlation between the loudspeaker and transmitted signal is 0.08.

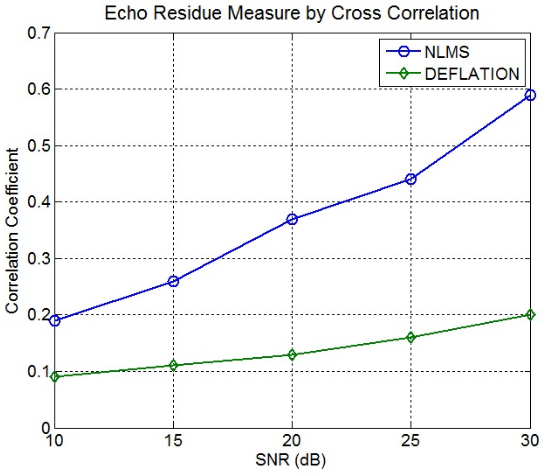

Fig. 7. Comparison between the source deflation algorithm and NLMS approaches: In general, the correlation between the loudspeaker and transmitted signals is less for the proposed method compared with the NLMS method. This is because noise causes filter divergence with the NMLS approach.

IV. CONCLUSION

A new algorithm for AEC via source deflation of the loudspeaker signal from the microphone signal has been introduced. The optimal deflation filter, which is the negated LEM enclosure filter, is learned via maximization of the information transferred from the observed loudspeaker signal to the processed signal for communication. The amount of information transferred is measured by means of differential entropy, where it is known that the uniform density function has the most entropy as all values are equally likely. Thus, with *a prior* knowledge about the near-end's probability density function, it is possible to cancel the loudspeaker signal where optimality coincides with maximal differential entropy of the non-linearly squashed output signal.

REFERENCES

[1] W. Sheng, Q. Xiaojung, and M. Ming, “Stereo Acoustic Echo Cancellation Employing Frequency-Domain Preprocessing and Adaptive Filter,” IEEE Transactions on Audio, Speech, and Language Processing. vol. 19, pp. 614-623, Mar. 2011.

[2] P. Ahgren, “Acoustic Echo Cancellation and Double Talk Detection using Estimated Loudspeaker Impulse Responses,” IEEE Transactions on Speech and Audio Processing. vol. 13, no. 6, pp. 1231-1237, Nov. 2005.

[3] C. Paleologu, S. Ciochina, and J. Benesty, “Variable Step-Size NLMS Algorithm for Under-Modeling Acoustic Echo Cancellation,” IEEE Signal Processing Letters. vol. 15, pp. 5-8, 2008.

[4] J.C. Jenq, and S. F. Hsieh, “Acoustic Echo Cancellation using Iterative Maximal Length Correlation and Double Talk Detection,” IEEE Transactions on Speech and Audio Processing. vol. 9, no. 8, pp. 932-942, Nov. 2001.

[5] C. Schuldt, F. Lindstrom, and I. Claesson, “A Delay-Based Double-Talk Detector,” IEEE Transactions on Audio, Speech, and Language Processing. vol. 20, no. 6, pp. 1725-1733, Aug. 2012.

[6] A.J. Bell, and T.J. Sejnowski, ”An Information-Maximization Approach to Blind Separation and Blind Deconvolution, ” Neural Computation, vol.7, no. 6, pp. 1004-1034, 1995.

[7] E.G. Learned-Miller and J.W. Fisher, “ICA using Spacing Estimates of Entropy,” Journal of Machine Learning Research. vol. 4, pp. 1271-1295, Dec. 2003.