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

DTD-free AEC via a Sliding DFT Window for ICA-based single Parameter Estimation

E. S. Gower, T. Tsalaile, M. Kgwadi, S. Masupe

Abstract*—In this paper, an independent component analysis (ICA) acoustic echo cancellation (AEC) algorithm is introduced where a sliding discrete Fourier transform window is adopted such that there is only one AEC parameter to estimate (reduced computational load), as opposed to thousands of coefficients modeling the room response. Conventional adaptive filtering techniques such as the least mean square (LMS) algorithm often fail under double-talk condition (and excessive noise) due to a corrupted measure of the objective function (i.e. minimization of the error output). Recent study has shown that ICA allows continual adaptation of the AEC parameters, hence it is adopted here as the optimization method of our AEC parameter. Simulation results are used to illustrate the superiority of the proposed algorithm over the LMS methods.

***Index Terms*—acoustic echo cancellation, blind deconvolution, double-talk detection, mutual information.**

I. INTRODUCTION

The current acoustic echo cancellation (AEC) algorithms are based mostly on adaptive filtering techniques [1-3]. The loudspeaker signal $l(n)$ is filtered by the loudspeaker-environment-microphone (LEM) impulse response $h_l(n)$ to give the far-end signal $d(n) = h_l(n) * l(n)$, where $*$ is the convolution operator. The talker signal $s(n)$ is filtered by the talker-environment-loudspeaker (TEM) impulse response $h_s(n)$ resulting in the near-end signal $m(n) = h_s(n) * s(n)$. These two signals along with the observed room noise $v(n)$ are captured by the microphone to give $x(n) = d(n) + m(n) + v(n)$. In the absence of the near-end signal, an adaptive filter $\widehat{h(n)}$ is used to model the LEM impulse response to give the far-end estimated signal $\widehat{d(n)}$. The estimated signal is then subtracted from the microphone signal resulting in the error signal $e(n) = r(n) + v(n)$, where $r(n) = d(n) - \widehat{d(n)}$ is the residue signal. Assuming minimal room noise, the estimated response $\widehat{h(n)}$ can converge to the desired LEM response $h(n)$ and the echo can be effectively cancelled. The problem arises on observation of the near-end signal as now the filter $\widehat{h(n)}$ will diverge resulting in poor echo cancellation.

To circumvent the problem of filter divergence due to the near-end signal, double talk detection (DTD) algorithms such as cross-correlation methods [4, 5], the Geigel algorithm [6], and other variants are employed to detect the presence of the near end signal, in which case the adaptive process of modeling the LEM filter is frozen. Assuming minimal changes in the LEM enclosure, the far-end signal can still be effectively suppressed. Unfortunately, there are cases where the near-end signal is observed for long durations and since in most cases it is the source of this signal that controls the location of the microphone, changes in the LEM enclosure and hence the filter are far too frequent. This implies that the current estimate of the LEM filter in the adaptive process is inadequate to effectively reduce the residual signal and echo cancellation fails.

In this paper, we propose the use of blind source separation (BSS) based on a blind deconvolution algorithm using a single frequency bin to address the problem a changing LEM filter in the presence of the near-end signal. By employing source separation as opposed to suppression of the far-end signal, there is no need for a DTD algorithm as the signals $l(n)$ and $s(n)$ can simply be separated, after which a mutual information check can be performed between each of the separated signals and loudspeaker signal to identify which is the echo (far-end signal) and the desired signal for transmission (near-end signal). The key point is that the filter modeling should not be frozen because of the near-end signal, and this is possible if BSS is employed. This paper is structured as follows: Section II introduces the proposed echo cancellation method. Simulation results for echo cancellation in the presence of the near-end signal for the proposed algorithm are in Section III. Discussions and summary remarks follow in Section IV.

II. THE PROPOSED AEC ALGORITHM

The proposed AEC algorithm is illustrated in Fig. 2. The loudspeaker signal is filtered by the LEM impulse response $h_l(n)$ such that the far-end signal is $d(n) = h_l(n) * l(n)$, whereas the near-end signal is given by $m(n) = h_s(n) * s(n)$.

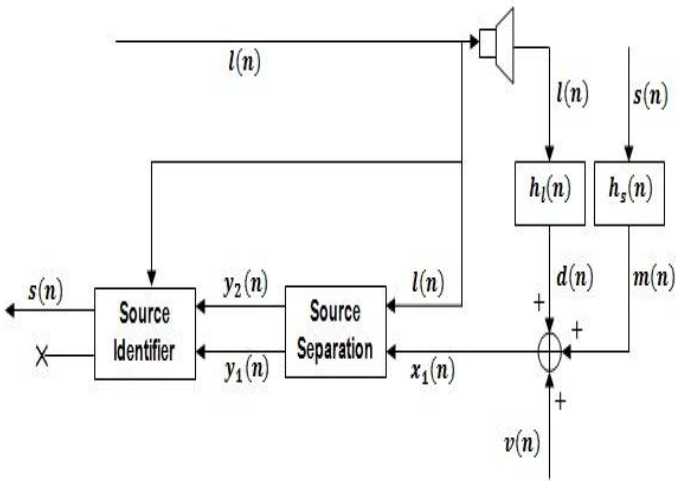

Fig. 1. The loudspeaker and microphone signals are used as inputs the source separation algorithm of convolutive mixtures. The sources are identified using the correlation coefficient.

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

There are two major signal processing stages;

A. PHASE 1: The Deconvolution Process

The inputs to the blind deconvolution algorithm are the loudspeaker signal $x_1(n) = l(n)$ and the microphone captured signal $x_2(n) = d(n) + m(n) + v(n)$, for $n\epsilon[1, N]$. The output signals are $y_1(n)$ and $y_2(n)$ which are approximations to the loudspeaker signal $l(n)$ and the talker signal $s(n)$, but due to the inherent output permutation ambiguity of instantaneous BSS algorithms to be employed for deconvolution, it is not clear at this point where each signal is output (this is considered in the second phase of the AEC algorithm). If we let $\boldsymbol{X}(n) = [x_1(n), x_2(n)]^T$, $\boldsymbol{S}(n) = [l(n), s(n)]^T$ and $\boldsymbol{A}(n)$ be the observation vector, source signal vector and mixing filter matrix respectively then we have the convolutive model

$$\boldsymbol{X}(n) = \boldsymbol{A}(n) * \boldsymbol{S}(n) + \boldsymbol{V}(n) = \sum_{k=0}^{K-1} \boldsymbol{A}(k)\boldsymbol{S}(n-k) + \boldsymbol{V}(n), \quad (1)$$

where $\boldsymbol{V}(n) = [\boldsymbol{0}, v(n)]^T$ is the observed noise vector (the noise captured by the microphone), and $*$ denotes the convolution operator with K as the maximum filter length. Based on the properties of the discrete Fourier transform (DFT), time domain convolution can be implemented as frequency domain point-to-point multiplication. Taking the L-point short-time Fourier transform (STFT) of (1) using a window size of length P, for $L > P \gg K$ to avoid circular convolution and allow sufficient spectral resolution of the filtering effects,

$$\boldsymbol{X}(\omega_k, n) = \boldsymbol{A}(\omega_k, n)\boldsymbol{X}(\omega_k, n)\boldsymbol{V}(\omega_k, n), \quad (2)$$

for $k\epsilon[0, L-1]$ and $\omega_k = \frac{2\pi k}{L}$. Given the de-mixing matrix $\boldsymbol{W}(\omega_k, n)$ the process can be formulated as follows;

$$\boldsymbol{Y}(\omega_k, n) = \boldsymbol{W}(\omega_k, n)\boldsymbol{X}(\omega_k, n), \quad (3)$$

with $\boldsymbol{Y}(\omega_k, n)$ as the source estimates vector in time-frequency domain for the frequency bin ω_k, $k\epsilon[0, L-1]$. In [7], it is shown that the sources can be extracted in the time-domain via the sliding-DFT method as

$$\boldsymbol{Y}'(n) = \boldsymbol{W}(\omega_k, t)\boldsymbol{X}'(\omega_k, n), \quad (4)$$

where the virtual source and virtual observation vectors $\boldsymbol{Y}'(n)$ and $\boldsymbol{X}'(\omega_k, n)$ respectively are given by

$$\boldsymbol{Y}'(n) = \boldsymbol{Y}(n) - \boldsymbol{Y}(n+P)e^{-j\omega_k P},$$

$$\boldsymbol{X}'(\omega_k, n) = \boldsymbol{X}(n) - \boldsymbol{X}(n+P)e^{-j\omega_k P}.$$

Based on the de-mixing model given by (4), only one frequency bin of choice ω_k need be considered, and this eliminates the permutation and amplitude ambiguities in the output vector for signal reconstruction, hence the choice of this algorithm for reduced computational complexity. The virtual sources are obtained by running any instantaneous BSS algorithm [8, 9], and to obtain source estimates from the virtual sources using the frequency bin of choice $\omega_k = 0$ we use

$$\boldsymbol{Y}(n+P) = [\boldsymbol{Y}(n) - \boldsymbol{Y}'(n)], \quad \text{for } n\epsilon[1, N-P], \quad (5)$$

where N is the total length of the virtual observations. Therefore, $\boldsymbol{Y}(t+P)$ can be calculated iteratively given the values $\boldsymbol{Y}(1), \boldsymbol{Y}(2), \cdots, \boldsymbol{Y}(P)$ are known. These values can be simply assumed to be zero which means that $\boldsymbol{X}(1), \boldsymbol{X}(2), \cdots, \boldsymbol{X}(P)$ are also set to zero. Consequently, for a filter length of K this introduces errors to $\boldsymbol{X}(P+1), \boldsymbol{X}(P+2), \cdots, \boldsymbol{X}(P+K)$. However, these errors are negligible as they are only observed on the initial adaptation of the algorithm (the same with adaptive filtering prior to convergence).

B. Phase 2: Identifying the Near-End Signal

There are two output signals from the blind deconvolution algorithm, the far-end signal and the near-end signal. For AEC, it is the near-end signal that is transmitted for communication which means that the far-end signal must be discarded. Therefore it is necessary to check where the near-end and far-end signals are output. It should be noted that this output ambiguity is due to the applied instantaneous BSS algorithm to the single frequency bin of choice. This is different from using multiple frequency bin deconvolution algorithms where the permutations can occur at every bin, leading to multiple checks for signal reconstruction. This check can be performed using the simple correlation coefficient measurement between the loudspeaker signal $l(n)$ and each of the output signals $y_1(n)$ and $y_2(n)$. The highest value would mean that the far-end signal is detected whereas the lowest correlation coefficient corresponds to the desired near-end signal. The correlation coefficient is given by

$$C_i(l(n), y_i(n)) = \frac{E[(l(n) - \mu_l)^T(y_i(n) - \mu_i)]}{\sigma_l \sigma_i}, \quad (6)$$

where μ_l and σ_l are the mean and standard deviation of the loudspeaker signal $l(n)$, with μ_i and σ_i as the mean and standard deviation of the i^{th} output signal, for $i\epsilon[1,2]$, and $E[.]$ is mathematical expectation. There are other possibly more accurate measures for identifying the near-end signal such as measuring the mutual information via spacing estimates of entropy [9], but it is the computational simplicity of (6) that makes the correlation coefficient preferable especially given the experimental success rate of this measure for DTD applications. That is, if (6) is a successful measure in identifying the presence of the near-end signal in DTD, then it is also a good measure for distinguishing between the near-end and far-end signal. In fact, any DTD method can be employed for this task.

III. SIMULATION RESULTS

The LEM and TEM impulse responses were generated using MATLAB for the length $K = 1024$, and these are plotted in Fig. 2.

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

The two speech signals of length $N = 20000$ 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 30dB are illustrated in fig. 3.

The DFT length and window size of the blind deconvolution algorithm were set such that $(L = 8192) > (P = 4096) \gg (K = 1024)$. After source separation, (6) was used to identify the desired talker signal, as well as a measure of the residual echo signal for which the correlation coefficient was $C\left(l(n), \widehat{s(n)}\right) = 0.09$, for the estimated talker signal $\widehat{s(n)}$. The results are in fig. 4.

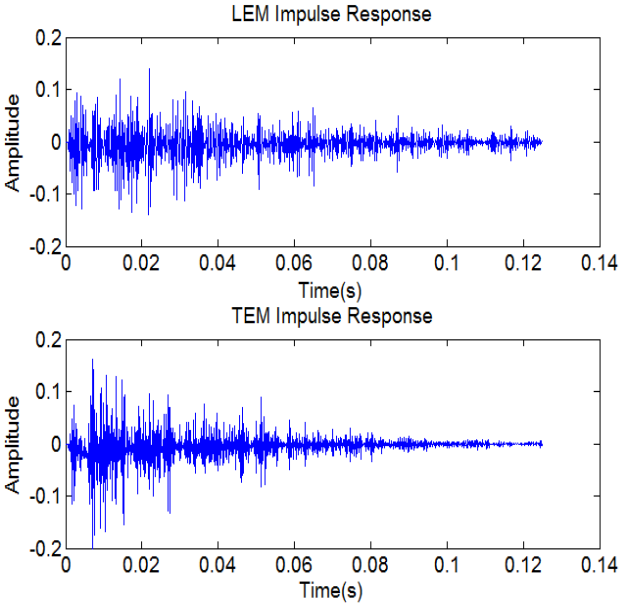

Fig. 2. Plots of the LEM and TEM impulse responses for filter length $K = 1024$.

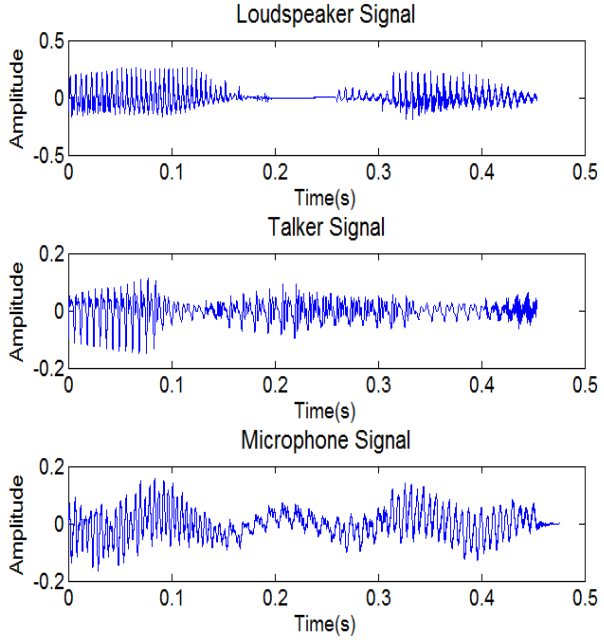

Fig. 3. The loudspeaker (top) and talker (middle) signals along with Gaussian noise are used to give the microphone (bottom) signal with a SNR=30dB.

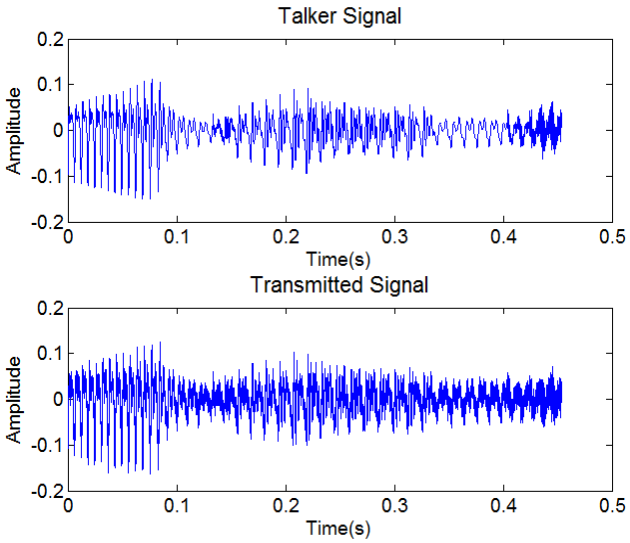

Fig. 4. The transmitted signal resembles the talker signal where the correlation between the loudspeaker and transmitted signal is 0.09.

The proposed method was further simulated in comparison with the conventional normalized least-mean-square (NLMS) algorithm for AEC under noisy conditions, but no near-end signal, where the echo residue is reflected by the correlation coefficient between the loudspeaker signal and the transmitted signal. The results are illustrated in Fig. 5.

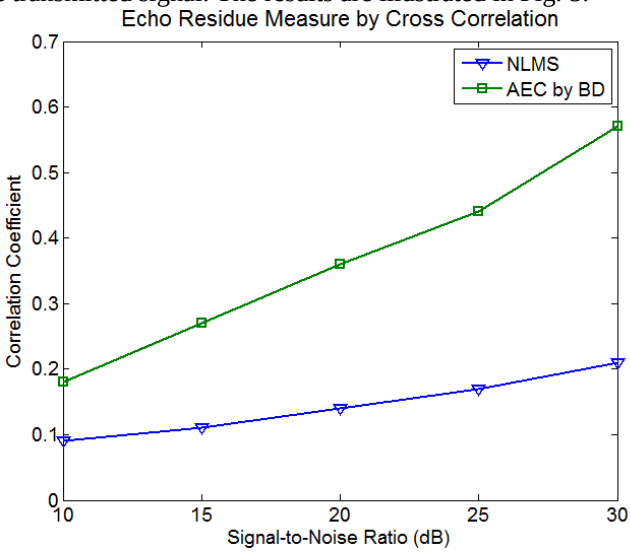

Fig. 5. In general, the correlation between the loudspeaker and transmitted signals is less for the blind deconvolution approach compared with the NLMS method. This is because deconvolution exploits statistical relations whereas NLMS focuses on reducing the noisy error signal for echo cancellation.

The blind deconvolution method achieves better echo suppression than the NLMS algorithm. This is because the conventional approach adapts the echo cancellation filter via minimization of the error signal, which under noise conditions is $e(n) = d(n) - \widehat{d(n)} + v(n)$. While the objective is to match $\widehat{d(n)}$ to $d(n)$, it is not possible due to additive noise, much in the same way that the near-end signal causes the filter to diverge. The deconvolution algorithm works by minimizing the mutual information between the

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

loudspeaker signal $l(n)$ and the microphone output $x(n) = h_l(n) * l(n) + v(n)$. This is still achievable even in the presence of noise, and the transmitted signal is usually free of the loudspeaker signal.

IV. CONCLUSION

A new method for AEC via a blind deconvolution algorithm using a single frequency bin has been introduced. The use of an arbitrary frequency bin helps to eliminate the amplitude and permutation ambiguities of multiple frequency bin deconvolution algorithms for signal reconstruction. However, there is a need to check where the desired signal is output, and this is done using any DTD algorithm. In this paper the correlation coefficient was employed, where a relatively low value indicates the estimated talker signal. The major benefit of this method is that the echo signal is still suppressed even when the near-end signal and room noise are observed. With conventional adaptive filtering techniques, it is possible to accurately model the LEM impulse response filter for echo cancellation in the presence of the near-end signal are severe room noise.

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, no. 3, 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] J. Liu, "A Novel Adaptation Scheme in the NLMS Algorithm for Echo Cancellation, " IEEE Signal Processing Letters, vol. 8, no. 1, pp. 20-22, Jan. 2001.

[7] G. Zhang, J. Li, and C. Li, "A Novel Blind Deconvolution Algorithm using Single Frequency Bin, " journal of Zhejiang University SCIENCE A, vol. 8, no. 8, pp. 1271-1276, Jul. 2007.

[8] E. S. Gower and M. O. J. Hawksford, "Learning Over complete Dictionaries using a Cauchy Mixture Model for Sparse Decay, "WASET International Journal of Electrical and Electronics Engineering, vol. 5, no. 2, Mar. 2011.

[9] 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.