Selective-Tap Blind Dereverberation for Two-Microphone Enhancement of Reverberant Speech

Kostas Kokkinakis, *Member, IEEE*, and Philipos C. Loizou, *Senior Member, IEEE*

***Abstract*—In this letter, we propose a novel approach for two-microphone enhancement of speech corrupted by reverberation. Our approach steers computational resources to filter coefficients having the largest impact on the error surface and therefore only updates a subset of coefficients in every iteration. Experimental results carried out in a realistic reverberant setup indicate that the performance of the proposed algorithm is comparable to the performance of its full-update counterpart.**

***Index Terms*—Blind dereverberation, least-mean-square (LMS), perceptual evaluation of speech quality (PESQ), selective-tap filter updating, SIMO-FIR model.**

I. Introduction

ACOUSTIC reverberation is caused by multiple reflections and diffractions of sounds on the walls and objects in enclosed spaces. Reverberation is harmful to speech intelligibility since it blurs temporal and spectral cues, flattens formant transitions, reduces amplitude modulations associated with the fundamental frequency of speech and increases low-frequency energy, which in turn results in masking of higher speech frequencies [1]. Blind reverberation cancelation or *dereverberation* is a well-known technique, based on which we can reconstruct an estimate of a speech signal distorted by reverberation with *no prior* knowledge of the signal itself or the acoustical properties of the room [2]. To isolate the original or "true" source signal in a multipath propagation scenario, one needs to rely solely on information that can be collected from the microphones.

Gannot and Moonen [3] were the first to propose a multimicrophone dereverberation technique using a generalized singular-value decomposition (GSVD) approach. Nakatani and Miyoshi [4] achieved speech dereverberation by extraction of the harmonic components of clean speech after filtering the reverberant signal through a pre-trained harmonic filter. Wu and Wang [5] proposed to maximize the kurtosis of the linear prediction (LP) residual of the original clean speech and then to use a spectral subtraction algorithm to decrease late reverberation. Lee *et al.* [6] resorted to a binaural (two-channel) model to reformulate the problem of blind dereverberation as a single-input multiple-output (SIMO) inverse filtering problem. Jointly reducing spectral coloration due to late reverberant energy as well as background noise for speech enhancement in practical applications has also been the focus of other recent single- and multichannel speech dereverberation strategies (e.g., see [7]–[9]). The aforementioned dereverberation strategies perform well, as long as large amounts of processing power and training data can be made available. However, since most algorithms require a large number of taps to capture room impulse responses and since their computational complexity and processing delays are proportional to the tap length used, they are prohibitively expensive for use in practical applications (e.g., hearing aids).

In this letter, we derive a novel blind single-input two-output reverberant speech enhancement strategy, which stems from the multichannel least-mean-square (MCLMS) algorithm [10]. The proposed method uses second-order statistics to identify the acoustic paths in the time-domain and relies on a novel *selective-tap* criterion to update only a subset of the total number of filter coefficients in every iteration. Therefore, it substantially reduces computational requirements with only minimal degradation in dereverberation performance. The potential of the proposed low-complexity algorithm is verified and assessed through numerical simulations in realistic acoustical scenarios.

II. Problem Formulation and Algorithm

Consider the paradigm shown in Fig. 1 where speech is picked up by the two microphones of a hearing aid device. Let $s(k)$ represent the sound source, $h_1(k)$ and $h_2(k)$ denote the impulse responses of the two acoustic paths modeled using finite impulse response (FIR) filters and $x_1(k)$ and $x_2(k)$ be the reverberant signals captured by the two microphones of the device. In the noiseless two-microphone scenario, we exploit the correlation between the output signals of each microphone

$$x_i(k) = h_i(k) * s(k) \text{ and } x_j(k) = h_j(k) * s(k) \quad (1)$$

where $*$ denotes linear convolution and $i, j = 1, 2$. From (1) and for all $i \neq j$, it follows that

$$\begin{aligned} h_j(k) * x_i(k) &= h_j(k) * [h_i(k) * s(k)] \\ &= h_i(k) * x_j(k). \end{aligned} \quad (2)$$

Manuscript received May 14, 2009; revised June 30, 2009. First published July 14, 2009; current version published August 26, 2009. This work was supported by Grants R03DC008882 and R01DC007527 awarded from the National Institute on Deafness and Other Communication Disorders (NIDCD) of the National Institutes of Health (NIH). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jen-Tzung Chien.

The authors are with the Center for Robust Speech Systems, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX 75080 USA (e-mail: kokkinak@utdallas.edu, loizou@utdallas.edu).

Digital Object Identifier 10.1109/LSP.2009.2027658

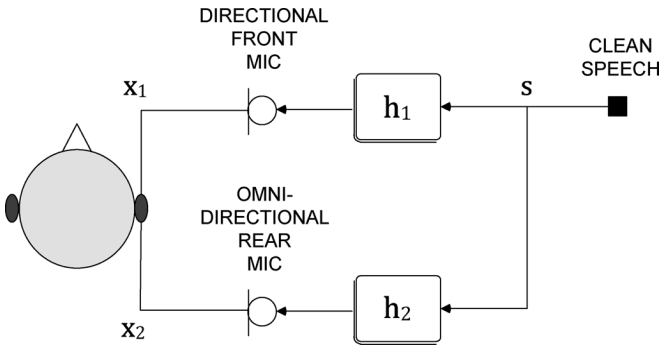

Fig. 1. Typical SIMO hearing aid setup, in which the two signals picked up by the directional (front) and omni-directional (rear) microphones are used to adaptively cancel the reverberation present in the background.

A. Multichannel LMS Algorithm

As shown in [10], an intuitive way to 'blindly' calculate the unknown acoustic paths is to minimize a cost function that penalizes correlated output signals between the ith and jth sensors[1], such that

$$J(k+1) = \arg\min_{\boldsymbol{h}} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} e_{ij}^2(k+1). \tag{3}$$

After rewriting (2) in vector notation, the error function $e_{ij}(k+1)$ in (3) can be further expanded to

$$e_{ij}(k+1) = \boldsymbol{x}_i^T(k+1)\,\widetilde{\boldsymbol{h}}_j(k) - \boldsymbol{x}_j^T(k+1)\,\widetilde{\boldsymbol{h}}_i(k) \tag{4}$$

defined for all $i, j = 1, 2$ and $i \neq j$, where $(\cdot)^T$ denotes vector transpose, vector $\widetilde{\mathbf{h}}_i(k) = [\widetilde{h}_i(0), \widetilde{h}_i(1), \ldots, \widetilde{h}_i(L-1)]^T$ represents the estimate of the time-invariant impulse response of the ith microphone at time instant k of order L and

$$\boldsymbol{x}_i(k) = [x_i(k), x_i(k-1), \ldots, x_i(k-L+1)]^T \tag{5}$$

is the corrupted (reverberant) speech picked up by the ith microphone. Accordingly, the update equation of the time-domain multichannel least-mean-square (MCLMS) algorithm is given by [10]

$$\widetilde{\boldsymbol{h}}(k+1) = \widetilde{\boldsymbol{h}}(k) - \mu \nabla J(k+1) \tag{6}$$

$$\nabla J(k+1) = \frac{\partial J(k+1)}{\partial \widetilde{\boldsymbol{h}}(k)} \tag{7}$$

$$= \frac{2\big[\widetilde{\mathbf{R}}_x(k+1)\widetilde{\boldsymbol{h}}(k) - J(k+1)\widetilde{\boldsymbol{h}}(k)\big]}{\big\|\widetilde{\boldsymbol{h}}(k)\big\|^2} \tag{8}$$

where $0 < \mu < 1$ is the learning parameter controlling the rate of convergence and speed of adaptation, $\widetilde{\boldsymbol{h}}(k)$ is the $(2L \times 1)$ composite channel response vector formed by the two separate channel coefficient vectors, such that

$$\widetilde{\boldsymbol{h}}(k) = \big[\widetilde{\boldsymbol{h}}_1^T(k), \widetilde{\boldsymbol{h}}_2^T(k)\big]^T \tag{9}$$

[1]To guarantee identifiability for the SIMO-FIR system described in (2), the two impulse responses $h_1(k)$ and $h_2(k)$ must be co-prime, namely they must share no common zeros and moreover the autocorrelation matrix of the source signal $\widetilde{\mathbf{R}}_{ss} = E\,[s(k)\,s^T(k)]$ needs to be of full rank.

and $\widetilde{\mathbf{R}}_x(k)$ is the $(2L \times 2L)$ autocorrelation matrix of the microphone signals, which is equal to

$$\widetilde{\mathbf{R}}_x(k) = \begin{pmatrix} \widetilde{\mathbf{R}}_{x_2x_2}(k) & -\widetilde{\mathbf{R}}_{x_2x_1}(k) \\ -\widetilde{\mathbf{R}}_{x_1x_2}(k) & \widetilde{\mathbf{R}}_{x_1x_1}(k) \end{pmatrix} \tag{10}$$

with $\widetilde{\mathbf{R}}_{x_ix_j}(k) = E\,[\boldsymbol{x}_i(k)\,\boldsymbol{x}_j^T(k)]$ valid for all $i, j = 1, 2$.

B. Selective-Tap Two-Channel LMS Algorithm

We formulate the new selective-tap approach based on the two-microphone configuration depicted in Fig. 1. By inspecting (6)–(8) we can see that as the adaptive algorithm approaches convergence, the cost function $J(k+1)$ diminishes and its gradient with respect to $\widetilde{\boldsymbol{h}}(k)$ becomes

$$\Delta\widetilde{\boldsymbol{h}}(k) \simeq -2\mu\,\widetilde{\mathbf{R}}_x(k+1)\,\widetilde{\boldsymbol{h}}(k) \tag{11}$$

after removing the unit-norm constraint from (8). From the above equation, it readily becomes evident that the convergence behavior of the MCLMS algorithm depends solely on the magnitude (element-wise) of the autocorrelation matrix $\widetilde{\mathbf{R}}_x$ estimated at each iteration k. The computational complexity and slow convergence of the MCLMS algorithm can be therefore reduced substantially by employing a simple tap-selection criterion to update only M out of L coefficients containing the largest values of the autocorrelation matrix [11]. The subset of the filter coefficients updated at iteration k can be determined from the $M \times M$ matrix $\mathbf{Q}(k)$, which is coined the tap-selection matrix

$$\mathbf{Q}(k) \triangleq \begin{pmatrix} \boldsymbol{q}_{ij}(0) & 0 & \cdots & 0 \\ 0 & \boldsymbol{q}_{ij}(1) & \ddots & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & 0 & 0 & \boldsymbol{q}_{ij}(M-1) \end{pmatrix} \tag{12}$$

where each element is given by

$$\boldsymbol{q}_{ij}(k) \triangleq [q_{ij}(k), q_{ij}(k-1), \ldots, q_{ij}(k-M+1)]^T \tag{13}$$

such that

$$q_{ij}(k-\ell) = \begin{cases} 1, & |\,\widetilde{\mathbf{R}}_x(k-\ell)\,| \in M \text{ maxima } |\,\widetilde{\mathbf{R}}_x\,| \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

where in a two-channel setup (12)–(14) are defined for $i, j = 1, 2$ and for all lags $\ell = 0, 1, \ldots, M-1$ and the operator $|\cdot|$ denotes absolute value. In order to calculate the different filter coefficients that are to be updated at different time instants, a fast sorting routine (e.g., see SORTLINE [12]) is executed at every iteration. After sorting, each block of the tap-selection matrix $\mathbf{Q}(k)$ contains M coefficients equal to one in the positions (or indices) calculated from (14) and zeros elsewhere, such that $M < L$ with $M = \mathrm{tr}\,[\mathbf{Q}(k)]$, where $\mathrm{tr}\,[\cdot]$ denotes the sum of the diagonal elements of matrix $\mathbf{Q}(k)$. To update only M taps of the equalizer $\widetilde{\boldsymbol{h}}(k)$, we write the *selective-tap two-channel least-mean-square* (SETA-TCLMS) algorithm as follows:

$$\widetilde{\boldsymbol{h}}_\ell(k+1) = \widetilde{\boldsymbol{h}}_\ell(k) - 2\lambda\,\widetilde{\mathbf{R}}_x(k-\ell+1)\,\widetilde{\boldsymbol{h}}_\ell(k) \tag{15}$$

where the update is carried out with learning rate λ only if ℓ corresponds to one of the first M maxima of $\mid \widetilde{\mathbf{R}}_x \mid$, whereas when $q_{ij}(k-\ell)=0$ then (15) becomes

$$\widetilde{\boldsymbol{h}}_\ell(k+1)=\widetilde{\boldsymbol{h}}_\ell(k) \tag{16}$$

where $\widetilde{\boldsymbol{h}}_\ell = [\widetilde{\boldsymbol{h}}_\ell(0),\widetilde{\boldsymbol{h}}_\ell(1),\ldots,\widetilde{\boldsymbol{h}}_\ell(M-1)]^T$. Note that for $M=L$, the SETA-TCLMS reduces to the *full-update* algorithm described in (6)–(8).

III. Experimental Results

The performance of the SETA-TCLMS algorithm is evaluated using sentences produced by five different speakers (three male and two female). The speech sources are approximately 10 s in duration and are recorded at a sampling rate of 8 kHz. All signals are taken from the IEEE database, which consists of phonetically balanced sentences, with each sentence being composed of approximately seven to 12 words [13]. Reverberant speech is generated by convolving "clean" speech with room impulse responses measured inside a $5\times 9\times 3.5$ m office using an experimental two-microphone hearing aid device mounted behind the ear of a KEMAR positioned at 1.5 m above the floor and at ear level [14].

The length of the acoustic impulse responses is approximately 1920 sample points and the reverberation time[2] is equal to $T_{60}=200$ ms in the 20–4000 Hz frequency band, which is a typical value encountered in most daily reverberant environments. All numerical simulations are carried out for a single-source and a two-microphone configuration. The SETA-TCLMS algorithm is executed with $L=2,048$, whereas the tap-selection length M is set to 2,048, 1,024, and 512 taps. The learning rates are explicitly tuned to yield the maximum possible steady-state performance. The enhanced speech is obtained by convolving the reverberant speech with the inverse of each estimated impulse response upon convergence.

A. Performance Evaluation

1) NPM: Since, in our experimental setup the acoustic channel impulse responses are known *a priori* the channel identification accuracy is calculated using the normalized projection misalignment (NPM) metric [10]

$$\mathrm{NPM}(k) \triangleq 20\log_{10}\frac{\|\,\varepsilon(k)\,\|}{\|\,\boldsymbol{h}\,\|} \tag{17}$$

with $\boldsymbol{h}=[\boldsymbol{h}_1^T,\boldsymbol{h}_2^T]^T$ and the projection misalignment $\varepsilon(k)$

$$\varepsilon(k)=\boldsymbol{h}-\frac{\boldsymbol{h}^T\widetilde{\boldsymbol{h}}(k)}{\widetilde{\boldsymbol{h}}^T(k)\widetilde{\boldsymbol{h}}(k)}\,\widetilde{\boldsymbol{h}}(k) \tag{18}$$

where for perfectly identified acoustic paths $\varepsilon(k)\rightarrow 0$.

2) Perceptual evaluation of speech quality (PESQ): Although the NPM metric can measure channel identification accuracy reasonably well, it might not always reflect the output speech quality. For that reason, we also assess the performance of the proposed algorithm using the PESQ [15]. The PESQ employs a sensory model to compare the original (unprocessed) with the enhanced (processed) signal, which is the output from the dereverberation algorithm, by relying on a perceptual model of the human auditory system. In the context of additive noise suppression, PESQ scores have been shown to exhibit a high Pearson's correlation coefficient of $\rho=0.92$ with subjective listening quality tests [16]. The PESQ measures the subjective assessment quality of the dereverberated speech rated as a value between 1 and 5 according to the five grade *mean opinion score* (MOS) scale. Here we use a modified PESQ measure [16], referred to as mPESQ, with parameters optimized towards assessing speech signal distortion, calculated as a linear combination of the average disturbance value D_{ind} and the average asymmetrical disturbance values A_{ind}[15], [16]

$$m\mathrm{PESQ}=a_0+a_1\,D_{\mathrm{ind}}+a_2\,A_{\mathrm{ind}} \tag{19}$$

such that

$$a_0=4.959,\ a_1=-0.191\ \text{ and }\ a_2=-0.006. \tag{20}$$

By definition, a high value of mPESQ indicates low speech signal distortion, whereas a low value suggests high distortion with considerable degradation present. In effect, the mPESQ score is inversely proportional to reverberation time and is expected to increase as reverberant energy decreases.

TABLE I
Dereverberation Performance For Five Speakers Averaged Across Both Microphones. All Values Obtained After Convergence

$L \mid M$	SPEAKER	NPM (dB)	OUTPUT PESQ
2,048 | 2,048	A	−10.97	4.12
	B	−10.83	4.09
	C	−12.62	4.27
	D	−11.01	4.18
	E	−11.24	4.21
2,048 | 1,024	A	−10.24	3.52
	B	−10.01	3.28
	C	−11.79	3.89
	D	−10.57	3.65
	E	−10.92	3.77
2,048 | 512	A	−7.18	3.27
	B	−9.67	3.12
	C	−8.52	3.50
	D	−7.31	3.52
	E	−7.98	3.63

B. Discussion

Table I contrasts the performance of the SETA-TCLMS algorithm relative to the performance of its full-update counterpart (see Section II-B). As it can be seen in Table I, the full-update TCLMS yields the best NPM performance and the highest mPESQ scores. Still, the degree of dereverberation remains largely unchanged when updating with the SETA-TCLMS using only $M=1,024$ filter coefficients. In fact, even when employing just $M=512$ taps, which accounts for a 75% reduction

[2] T_{60} defines the interval in which the reverberating sound energy, due to decaying reflections, reaches one millionth of its initial value. In other words, it is the time it takes for reverberation to drop by 60 dB below the original sound energy present in the room at any given instant.

in the total equalizer length (with a processing delay of just 64 ms at 8 kHz) the algorithm can estimate the room impulse responses with reasonable accuracy.

In terms of overall speech quality and speech distortion, the mPESQ score for the reverberant (unprocessed) speech signals, averaged across all five speakers and in both microphones, is equal to 2.72, which suggests that a relatively high amount of degradation is present in the microphone inputs. In contrast, after processing the two-microphone reverberant input signals with the SETA-TCLMS algorithm, the average mPESQ scores increase to 4.17, 3.62 and 3.40 when using $M = 2,048$, 1,024 and 512 taps, respectively. The estimated mPESQ values suggest that the proposed SETA-TCLMS algorithm can improve the speech quality of the microphone signals considerably, while keeping signal distortion to a minimum.

IV. Conclusions

We have developed a selective-tap blind identification scheme for reverberant speech enhancement using a two-microphone configuration. Numerical experiments carried out with speech signals in a moderately reverberant setup, indicate that the proposed two-channel dereverberation technique is capable of equalizing fairly long acoustic echo paths with sufficient accuracy and nearly no degradation. The proposed adaptive algorithm exhibits a low computational overhead and therefore is amenable to real-time implementation in portable devices (e.g., hearing aids).

References

[1] A. K. Nabelek and J. M. Picket, "Monaural and binaural speech perception through hearing aids under noise and reverberation with normal and hearing-impaired listeners," *J. Speech Hear. Res.*, vol. 17, pp. 724–739, 1974.

[2] S. Haykin, Ed., Unsupervised Adaptive Filtering New York, Wiley, 2000, vol. II, Blind Deconvolution.

[3] S. Gannot and M. Moonen, "Subspace methods for multi-microphone speech dereverberation," *EURASIP J. Appl. Signal Process.*, vol. 11, pp. 1074–1090, 2003.

[4] T. Nakatani and M. Miyoshi, "Blind dereverberation of single channel speech signal based on harmonic structure," in *Proc. ICASSP*, 2003, vol. 1, pp. 92–95.

[5] M. Wu and D. L. Wang, "A two-stage algorithm for one-microphone reverberant speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 3, pp. 774–784, May 2006.

[6] J.-H. Lee, S.-H. Oh, and S.-Y. Lee, "Binaural semi-blind dereverberation of noisy convoluted speech signals," *Neurocomputing*, vol. 72, pp. 636–642, 2008.

[7] E. A. P. Habets, "Multi-channel speech dereverberation based on a statistical model of late reverberation," in *Proc. ICASSP*, 2005, vol. 4, pp. 173–176.

[8] H. W. Löllmann and P. Vary, "A blind speech enhancement algorithm for the suppression of late reverberation and noise," in *Proc. ICASSP*, 2009, pp. 3989–3992.

[9] T. Yoshioka, T. Nakatani, T. Hikichi, and M. Miyoshi, "Maximum likelihood approach to speech enhancement for noisy reverberant signals," in *Proc. ICASSP*, 2008, pp. 4585–4588.

[10] Y. Huang and J. Benesty, "Adaptive multi-channel least mean square and Newton algorithms for blind channel identification," *Signal Process.*, vol. 82, pp. 1127–1138, 2002.

[11] T. Aboulnasr and K. Mayyas, "Complexity reduction of the NLMS algorithm via selective coefficient update," *IEEE Trans. Signal Process.*, vol. 47, pp. 1421–1424, 1999.

[12] I. Pitas, "Fast algorithms for running ordering and max/min calculation," *IEEE Trans. Circuits Syst.*, vol. 36, no. 6, pp. 795–804, Jun. 1989.

[13] IEEE Subcommittee, "IEEE recommended practice speech quality measurements"," *IEEE Trans. Audio Electroacoust.*, vol. 17, no. 3, pp. 225–246, Sep. 1969.

[14] B. G. Shinn-Cunningham, N. Kopco, and T. J. Martin, "Localizing nearby sound sources in a classroom: Binaural room impulse responses," *J. Acoust. Soc. Amer.*, vol. 117, pp. 3100–3115, 2005.

[15] Perceptual Evaluation of Speech Quality (PESQ), an Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Coders ITU-T Recommendation, 2001, ITU-T Recommendation P.862.

[16] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 1, pp. 229–238, Jan. 2008.