DEEP ADAPTATION CONTROL FOR STEREOPHONIC ACOUSTIC ECHO CANCELLATION

Amir Ivry *Israel Cohen* *Baruch Berdugo*

Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering
Technion – Israel Institute of Technology, Technion City, Haifa 3200003, Israel

ABSTRACT

We introduce a general and data-driven adaptation-control framework for stereophonic acoustic-echo cancellation. The adaptation update rule for the filters that estimate the actual echo paths is compactly expressed with the widely-linear model in the complex time domain. A single step-size parameter that governs the behavior of the adaptation process is optimized by minimizing the misalignment between the actual echo paths and their filtered estimate. The relation between acoustic signals and the optimal step-size is learned via a deep neural network. In test mode, the optimal step-size prediction is inferred by the network and fed to the sign-error normalized least mean-squares (SNLMS) adaptive filter for echo-paths tracking. Real and simulated data show advantageous performance in single and double-talk scenarios across various acoustic setups.

Index Terms— Stereophonic acoustic echo cancellation, adaptation control, variable step-size, sign-error NLMS, deep learning.

1. INTRODUCTION

In stereophonic hands-free speech communication, the near-end microphones may capture three types of acoustic signals; the desired speech, additional noises, and reverberant echoes. The echoes are nonlinearly distorted versions of the far-end signal played by loudspeakers and reverberate to the microphones via echo paths [1]. These echoes may impede conversation intelligibility as perceived by the far-end participant. The stereophonic acoustic echo cancellation (SAEC) task is two-fold; tracking the near-end echo-paths and subtracting them from the microphones signals, and communicating the undistorted desired-speech signal to the far-end [2].

The popular normalized least mean-squares (NLMS) adaptive filter is numerically stable and efficient [3, 4]. Its sign-error NLMS (SNLMS) variation employs the polarity of the adaptation error [5] and is favorable over the NLMS due to its protection against abrupt noises [6–8]. The adaptation of the SNLMS filter is governed by the step-size parameter, which balances the convergence pace and the adaptation accuracy of the filter. Controlling the step-size is desirable in scenarios of frequent acoustic changes, e.g., echo-path variations and single-to-double-talk transitions. The variable step-size (VSS) problem has motivated Haubner et al. to employ deep learning for near-end speech [9] and noise [10] evaluation, and to reduce the error of the adaptive process [11]. Meta-learning-based solutions have also recently emerged in [12]. The a priori adaptation error and the far-end signal undergo feature extraction for VSS estimate in [13] and a non-parametric VSS (NPVSS) minimized the adaptation error in [14]. The mean-error sigmoid VSS (SVSS) combines adaptation-error history with current adaptation-error estimate [15].

The methods in [9–11] model the far-end signal as linear with its respective echo-signal, and the studies in [13–15] consider the echo path as time-invariant. Unfortunately, both assumptions restrict performance in realistic setups and may cause low adaptation accuracy with slow convergence-pace [16]. On top of that, parameter-tuning, as in the NPVSS [14], involves heuristics that are inaccurate in practice. Thus, SAEC in real-life scenarios remains a relevant challenge and an active research area.

Inspired by [17], we mitigate these disparities by introducing a data-driven framework for deep learning-based VSS (DVSS) that avoids heuristics and does not require acoustic setup hypotheses. First, the update rule of the adaptation process, governed by the step size, integrates the widely-linear model in the complex time domain. The mismatch between the actual echo paths and their filtered estimate is quantified by the normalized misalignment, which is then minimized with respect to the step size. A neural network (NN) relates acoustic signals to the optimal step-size in training, and the predicted step-size feeds the SNLMS filter in real time for tracking the echo paths. The described framework is novel for SAEC.

We compare our approach with the competition by considering a pair of near-end loudspeakers and microphones, although this framework generalizes to any number of channels. Experimenting with 100 h from the AEC-challenge corpus [18] reveals the consistent advantage of the DVSS in single and double-talk periods across various acoustic setups. The DVSS-SNLMS system also re-converges more rapidly and accurately after abrupt echo-path changes and is more robust to single-to-double-talk transitions.

2. PROBLEM FORMULATION

Our DVSS-SNLMS setup is in Figure 1. The left and right near-end microphones $m_{\text{L}}(n)$ and $m_{\text{R}}(n)$ at time index n are, respectively,

$$m_{\text{L}}(n) = y_{\text{L}}(n) + s_{\text{L}}(n) + w_{\text{L}}(n), \tag{1}$$

$$m_{\text{R}}(n) = y_{\text{R}}(n) + s_{\text{R}}(n) + w_{\text{R}}(n), \tag{2}$$

where $s_{\text{L}}(n)$ and $s_{\text{R}}(n)$ are the near-end speech signals, $w_{\text{L}}(n)$ and $w_{\text{R}}(n)$ represent environmental and system noises, and $y_{\text{L}}(n)$ and $y_{\text{R}}(n)$ are the nonlinear reverberant echo signals, as correspondingly captured by the left and right microphones:

$$y_{\text{L}}(n) = \mathbf{h}_{\text{LL}}^{T}(n)\,\mathbf{x}_{\text{NL,L}}(n) + \mathbf{h}_{\text{RL}}^{T}(n)\,\mathbf{x}_{\text{NL,R}}(n), \tag{3}$$

$$y_{\text{R}}(n) = \mathbf{h}_{\text{LR}}^{T}(n)\,\mathbf{x}_{\text{NL,L}}(n) + \mathbf{h}_{\text{RR}}^{T}(n)\,\mathbf{x}_{\text{NL,R}}(n). \tag{4}$$

Here, $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ respectively denote the L-recent samples from the left and right far-end signals, i.e., $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$, subsequent to nonlinear distortions by nonideal hardware [16]:

$$\mathbf{x}_{\text{NL,L}}(n) = [x_{\text{NL,L}}(n), \ldots, x_{\text{NL,L}}(n-L+1)]^{T}, \tag{5}$$

$$\mathbf{x}_{\text{NL,R}}(n) = [x_{\text{NL,R}}(n), \ldots, x_{\text{NL,R}}(n-L+1)]^{T}, \tag{6}$$

and each of the column vectors $\mathbf{h}_{\text{LL}}(n)$, $\mathbf{h}_{\text{RL}}(n)$, $\mathbf{h}_{\text{LR}}(n)$, $\mathbf{h}_{\text{RR}}(n)$ has L samples and represents an echo path from the loudspeakers

979-8-3503-2372-6/23/$31.00 ©2023 IEEE

to the microphones, also known as a room impulse response (RIR). Instead of tracking $4L$ real-valued coefficients, we turn to the more compact widely-linear model [19] by defining the complex signals:

$$\mathbf{h}\left(n\right) = \mathbf{h}_1\left(n\right) + j\mathbf{h}_2\left(n\right), \tag{7}$$

$$\mathbf{h}'\left(n\right) = \mathbf{h}'_1\left(n\right) + j\mathbf{h}'_2\left(n\right), \tag{8}$$

where $j = \sqrt{-1}$, and

$$\mathbf{h}_1\left(n\right) = 0.5\left[\mathbf{h}_{\text{LL}}\left(n\right) + \mathbf{h}_{\text{RR}}\left(n\right)\right], \tag{9}$$

$$\mathbf{h}_2\left(n\right) = 0.5\left[\mathbf{h}_{\text{RL}}\left(n\right) - \mathbf{h}_{\text{LR}}\left(n\right)\right], \tag{10}$$

$$\mathbf{h}'_1\left(n\right) = 0.5\left[\mathbf{h}_{\text{LL}}\left(n\right) - \mathbf{h}_{\text{RR}}\left(n\right)\right], \tag{11}$$

$$\mathbf{h}'_2\left(n\right) = -0.5\left[\mathbf{h}_{\text{RL}}\left(n\right) + \mathbf{h}_{\text{LR}}\left(n\right)\right]. \tag{12}$$

The complex echo signal $y\left(n\right) = y_{\text{L}}\left(n\right) + jy_{\text{R}}\left(n\right)$ can now be expressed in a widely-linear manner by $y\left(n\right) = \tilde{\mathbf{h}}^H\left(n\right)\tilde{\mathbf{x}}_{\text{NL}}\left(n\right)$:

$$\tilde{\mathbf{h}}\left(n\right) = \begin{bmatrix}\mathbf{h}\left(n\right)\\ \mathbf{h}'\left(n\right)\end{bmatrix}, \tag{13}$$

$$\tilde{\mathbf{x}}_{\text{NL}}\left(n\right) = \begin{bmatrix}\mathbf{x}_{\text{NL}}\left(n\right)\\ \mathbf{x}^*_{\text{NL}}\left(n\right)\end{bmatrix}, \tag{14}$$

where $\mathbf{x}_{\text{NL}}\left(n\right) = \mathbf{x}_{\text{NL,L}}\left(n\right) + j\mathbf{x}_{\text{NL,R}}\left(n\right)$. The superscripts ^{H} and * correspondingly notate the transpose-conjugate and conjugate operations. As a result, the complex microphone signal $m\left(n\right) = m_{\text{L}}\left(n\right) + jm_{\text{R}}\left(n\right)$ can be formulated by

$$m\left(n\right) = \tilde{\mathbf{h}}^H\left(n\right)\tilde{\mathbf{x}}_{\text{NL}}\left(n\right) + s\left(n\right) + w\left(n\right), \tag{15}$$

where $s\left(n\right) = s_{\text{L}}\left(n\right) + js_{\text{R}}\left(n\right)$ and $w\left(n\right) = w_{\text{L}}\left(n\right) + jw_{\text{R}}\left(n\right)$.

The echo estimation $\hat{y}\left(n\right) = \hat{\tilde{\mathbf{h}}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right)$, where $\tilde{\mathbf{x}}\left(n\right)$ and $\tilde{\mathbf{x}}_{\text{NL}}\left(n\right)$ follow the same notation, is evaluated by tracking the $2L$ complex-coefficients of $\hat{\tilde{\mathbf{h}}}\left(n\right)$ with the SNLMS adaptive filter. Subsequently, the complex near-end speech estimate can be drawn by

$$\begin{aligned} e\left(n\right) &= m\left(n\right) - \hat{y}\left(n\right) \\ &= \left(y\left(n\right) - \hat{y}\left(n\right)\right) + s\left(n\right) + w\left(n\right), \end{aligned} \tag{16}$$

where $e\left(n\right) = e_{\text{L}}\left(n\right) + je_{\text{R}}\left(n\right)$. Our focus is two-fold; tracking and cancelling the echo signal, i.e. nullifying $y\left(n\right) - \hat{y}\left(n\right)$, and avoiding distortion of the near-end speech, i.e. preserving $s\left(n\right)$.

3. DVSS-SNLMS FILTER FOR SAEC

3.1. Modeling the SNLMS Filter and Step-size in Double-talk

By placing (15) and the definition of $\hat{y}\left(n\right)$ into (16), we respectively derive the a priori and a posteriori errors of the SNLMS filter [4]:

$$\epsilon\left(n\right) = \tilde{\mathbf{h}}^H\left(n\right)\tilde{\mathbf{x}}_{\text{NL}}\left(n\right) - \hat{\tilde{\mathbf{h}}}^H\left(n-1\right)\tilde{\mathbf{x}}\left(n\right) + s\left(n\right) + w\left(n\right), \tag{17}$$

$$e\left(n\right) = \tilde{\mathbf{h}}^H\left(n\right)\tilde{\mathbf{x}}_{\text{NL}}\left(n\right) - \hat{\tilde{\mathbf{h}}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right) + s\left(n\right) + w\left(n\right). \tag{18}$$

The update rule of the $2L$ complex-valued filter coefficients is [5]:

$$\hat{\tilde{\mathbf{h}}}\left(n\right) = \hat{\tilde{\mathbf{h}}}\left(n-1\right) + \mu\left(n\right)\tilde{\mathbf{x}}\left(n\right)\text{sign}\left(\epsilon^*\left(n\right)\right), \tag{19}$$

where $\hat{\tilde{\mathbf{h}}}\left(0\right)$ is a column vector of $2L$ zeros, the step-size is given by $\mu\left(n\right) \in \mathbb{R}$, and $\text{sign}\left(z\right) = z/|z|$ for every $z \in \mathbb{C}$, where $|\cdot|$ is the absolute value. From (17)–(19):

$$e\left(n\right) = \epsilon\left(n\right) - \mu\left(n\right)\text{sign}\left(\epsilon\left(n\right)\right)\tilde{\mathbf{x}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right). \tag{20}$$

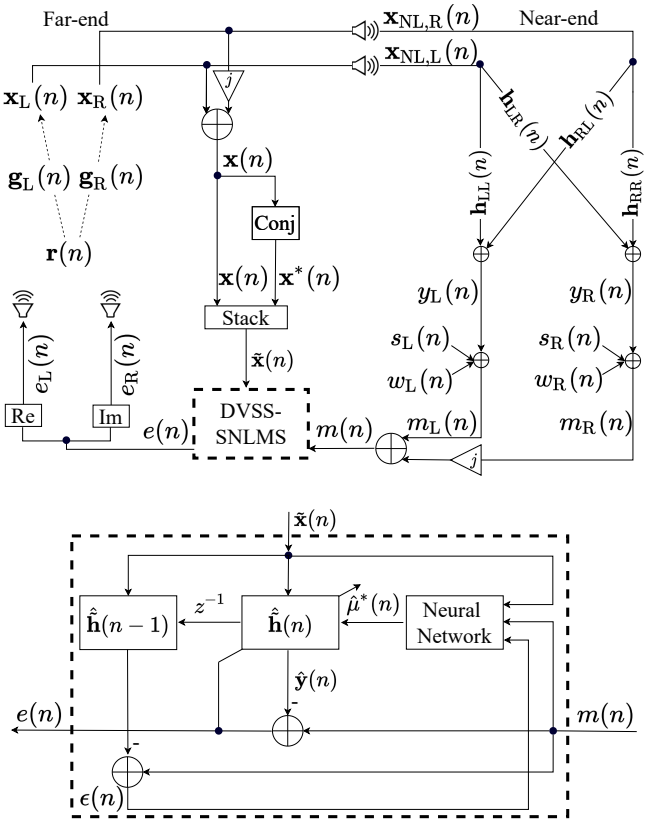

Figure 1: Top - SAEC scenario under the widely-linear model. Bottom - the DVSS-SNLMS block, where a NN estimates the step-size $\hat{\mu}^*\left(n\right)$ and the SNLMS filter estimates the acoustic paths via $\hat{\tilde{\mathbf{h}}}\left(n\right)$.

We now force the a posteriori error to a complete echo-cancellation and extract the corresponding expression of the step-size $\mu\left(n\right)$ [19]. Assuming $s\left(n\right)$ and $w\left(n\right)$ are zero-mean and uncorrelated [4]:

$$\sigma^2_e\left(n\right) = \sigma^2_s\left(n\right) + \sigma^2_w\left(n\right), \tag{21}$$

where $\sigma^2_e\left(n\right) = E\left[|e\left(n\right)|^2\right]$ and $\sigma^2_s\left(n\right)$, $\sigma^2_w\left(n\right)$ follow the same definition. Now, the $E\left[|\cdot|^2\right]$ operator is applied on both sides of (20), and then (21) is substituted into (20). This process yields

$$\mu\left(n\right) = c + \sqrt{\frac{\sigma^2_s\left(n\right) + \sigma^2_w\left(n\right) - \sigma^2_\epsilon\left(n\right)}{E\left[\left(\tilde{\mathbf{x}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right)\right)^2\right]} - c^2}, \tag{22}$$

where $c = E\left[|\epsilon|\tilde{\mathbf{x}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right)\right]/E\left[\left(\tilde{\mathbf{x}}^H\left(n\right)\tilde{\mathbf{x}}\left(n\right)\right)^2\right]$.

3.2. Step-size Optimization with a Data-driven Approach

The mismatch between the adaptive and true filter coefficients is often assessed using the normalized misalignment measure [19]:

$$\begin{aligned} \mathcal{D}\left(n\right) &= \frac{\left\|\tilde{\mathbf{h}}\left(n\right) - \hat{\tilde{\mathbf{h}}}\left(n\right)\right\|_2}{\left\|\tilde{\mathbf{h}}\left(n\right)\right\|_2} \\ &= \frac{\left\|\tilde{\mathbf{h}}\left(n\right) - \hat{\tilde{\mathbf{h}}}\left(n-1\right) - \mu\left(n\right)\tilde{\mathbf{x}}\left(n\right)\text{sign}\left(\epsilon^*\left(n\right)\right)\right\|_2}{\left\|\tilde{\mathbf{h}}\left(n\right)\right\|_2}, \end{aligned} \tag{23}$$

where (19) was employed to the second transition and $\|\cdot\|_2$ is the ℓ_2 norm. We now solve a constrained nonlinear optimization problem [20] to yield the optimal step-size. Formally, the normalized misalignment is minimized with respect to the step-size, in dB:

$$\mu^*(n) = \underset{0<\mu(n)<2}{\arg\min}\ 20\log_{10}\mathcal{D}(n), \tag{24}$$

where the condition $0 < \mu(n) < 2$ is dictated by the stability requirements of NLMS-based adaptive filters [4]. The active-set optimization algorithm [21] is utilized to perform optimization. According to (23), the only values involved in deriving $\mathcal{D}(n)$ are the far-end and the a priori error signals. This data-driven approach does not require heuristic parameter tuning to estimate $\mu^*(n)$.

3.3. Deep Adaptation to the Optimal Step-size

A deep NN is integrated into our system to model the relation derived in (22) between acoustic signals and the optimal step-size $\mu^*(n)$. Despite the near-end speech and noise signals being inaccessible in reality, the microphone signal can serve as an approximation. Thus, the microphone signal, along with the far-end and a priori error signals, are mapped to their respective step-size. The NN architecture is of a convolutional form [22] with six input channels, one for the real and one for the imaginary part of each of the three input signals. These six waveforms undergo short-time Fourier transform (STFT) [23] separately before being fed to the NN. The specific architecture is standard and follows the one in [17]. During training, optimization is carried to minimize the ℓ_2 norm between the optimal step-size $\mu^*(n)$ and the output of the network. During inference, the step-size estimate $\hat{\mu}^*(n)$ is evaluated by the network and injected to an SNLMS filter that tracks the echo paths. Addressing complexity analysis, the NN and the SNLMS filter consume 4.2 Million floating-point operations per second (Mflops) and 4.8 Megabytes (MB) of memory, by employing 1.05 Million parameters. Embedding this system into real-life edge devices for hands-free speech communication is thus considered feasible in terms of resources [24]. One example of dedicated hardware for this task is the NDP120 neural processor by Syntiant™ [25].

4. EXPERIMENTAL SETUP

4.1. Database Acquisition

The database corpus utilized in this study includes 100 h of noisy and clean segments taken from the AEC-challenge [18], where 25 h are simulated recordings, and 75 h are real recordings. The AEC-challenge data involves scenarios with no echo-paths change, where the near-end speaker and devices do not move, and scenarios with echo-paths change, where either the near-end speaker or devices are moving. We consider both double-talk periods and single-talk periods with far-end speakers only. Practically, audio clips are assigned to the original far-end source signal $\mathbf{r}(n)$ and to the near-end speech and noise signals, where $s_{\text{L}}(n) = s_{\text{R}}(n)$ and $w_{\text{L}}(n) = w_{\text{R}}(n)$ in this study. To produce the far-end signals $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$, $\mathbf{r}(n)$ is randomly propagated via one of 4500 pairs of RIRs that generate $\mathbf{g}_{\text{L}}(n)$ and $\mathbf{g}_{\text{R}}(n)$, i.e., the acoustic paths between $\mathbf{r}(n)$ and the left and right far-end microphones, respectively. To account for realistic acoustic environments, one of 4500 simulated nonlinear functions is applied to every $\mathbf{x}_{\text{L}}(n)$ and $\mathbf{x}_{\text{R}}(n)$ pair in a random fashion. These nonlinearities are modeled after realistic power amplifiers and loudspeakers in current hands-free hardware [16]. Each pair of nonlinearly-distorted far-end signals $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ is randomly propagated via one of 4500 foursomes of near-end RIRs. All RIRs are generated using the Image Method [26] with L coefficients and reverberation times RT_{60}, where $\text{RT}_{60} \sim U[0.2, 0.5]$ seconds. The near-end stereo-echo-to-speech ratio (SESR) and stereo-echo-to-noise ratio (SENR) levels were drawn from $[-10, 10]$ dB and $[0, 40]$ dB, respectively, where SESR=$10\log_{10}\left[|y(n)|^2/|s(n)|^2\right]$ and SENR=$10\log_{10}\left[|y(n)|^2/|w(n)|^2\right]$ in dB [19]. These ratios are derived by running 20 ms frames that overlap by 50%.

4.2. Preprocessing, Training, and Inference

We recognize the well-known non-uniqueness problem in setups of SAEC, where strong coherence between $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ may degrade the adaptation process [27]. To mitigate that, we apply the following channel-wise transformation introduced in the context of the widely-linear model [19]. First, we define the positive and negative half-wave rectifiers [28]:

$$\mathbf{x}'_{\text{NL,L}}(n) = \mathbf{x}_{\text{NL,L}}(n) + 0.5\left[\mathbf{x}_{\text{NL,L}}(n) + \|\mathbf{x}_{\text{NL,L}}(n)\|\right], \tag{25}$$

$$\mathbf{x}'_{\text{NL,R}}(n) = \mathbf{x}_{\text{NL,R}}(n) + 0.5\left[\mathbf{x}_{\text{NL,R}}(n) - \|\mathbf{x}_{\text{NL,R}}(n)\|\right]. \tag{26}$$

With the element-wise operation $\tan\boldsymbol{\theta}(n) = \mathbf{x}'_{\text{NL,R}}(n)/\mathbf{x}'_{\text{NL,L}}(n)$:

$$\mathbf{x}''_{\text{NL,L}}(n) = \cos\boldsymbol{\theta}(n)\,\|\mathbf{x}_{\text{NL}}(n)\|, \tag{27}$$

$$\mathbf{x}''_{\text{NL,R}}(n) = \sin\boldsymbol{\theta}(n)\,\|\mathbf{x}_{\text{NL}}(n)\|, \tag{28}$$

where eqs. (27) and (28) use element-wise arithmetic. This transformation modifies only phase information, so employing $\mathbf{x}''_{\text{NL,L}}(n)$ and $\mathbf{x}''_{\text{NL,R}}(n)$ instead of $\mathbf{x}_{\text{NL,L}}(n)$ and $\mathbf{x}_{\text{NL,R}}(n)$ allows a desired reduction in coherence with the advantage of little distortion.

The entire 100 h batch of data is split to yield training, validation, and test sets of sizes 80 h, 10 h, and 10 h, respectively. The split is random, but constrained to preserve balance and avoid bias by following the principles in [29]. Using the training and validation parts, the step-size is evaluated once every 8 ms according to (24) with parameter values of $\mu(0) = 3 \times 10^{-5}$ and $L = 2400$. Echo-paths are abruptly changed once every t seconds, where $t \sim U[4, 10]$, which characterizes in-the-wild conversations. Waveforms undergo STFT with running time frames that are 16 ms long and have 50% overlap. Before being inserted into the network, every STFT representation of every channel is attached to its 96 ms past context. Training the network using back-propagation involves learning rate of 10^{-4} that decays by 10^{-6} every 5 epochs, mini-batch size of 32 ms, and 40 epochs, using Adam optimizer [30]. The real-time inference is done on the test set. After the artificial gain of the network is calibrated according to [31], the step-size estimate is injected from the network output into the SNLMS, which constantly evaluates the echo paths. Training the network took 32 minutes for every 1 h of input data from all channels. The inference time for the end-to-end system, from the network entry to the echo-paths estimate, is 26 ms on average using an Intel Core i7-8700K CPU @ 3.7 GHz with two GPUs of Nvidia GeForce RTX 2080 Ti.

4.3. Performance Measures

In single-talk periods with only far-end signals and noise presence, we estimate the echo suppression level between the microphone and enhanced signals using the echo return loss enhancement (ERLE) [32], defined as $10\log_{10}\left[|m(n)|^2/|e(n)|^2\right]$. In double-talk, we consider both the signal-to-distortion ratio (SDR) [33] and the perceptual evaluation of speech quality (PESQ) [34]. The SDR

Table 1: Performance with no echo-paths change.

	SDR [dB]	PESQ	ERLE [dB]	Norm. Mis. [dB]
DVSS	**3.31±0.6**	**2.45±0.3**	**16.1±4.8**	**-17.3±4.4**
NPVSS	2.47±1.0	2.01±0.4	12.9±5.9	-14.8±4.6
NNVSS	2.41±1.0	1.86±0.5	12.5±6.0	-14.2±4.8
SVSS	2.12±0.9	1.73±0.5	10.7±6.5	-13.1±4.8
SNLMS	1.93±1.1	1.60±0.3	9.7±6.8	-11.6±5.2

Table 2: Performance with echo-paths change.

	SDR [dB]	PESQ	ERLE [dB]	Norm. Mis. [dB]
DVSS	**3.05±0.8**	**2.27±0.4**	**10.9±6.3**	**-12.5±5.7**
NPVSS	2.20±1.2	1.79±0.5	7.9±6.5	-9.9±5.9
NNVSS	1.98±1.1	1.71±0.5	7.4±6.8	-9.4±6.1
SVSS	1.91±1.3	1.63±0.5	7.0±6.8	-9.2±5.9
SNLMS	1.74±1.5	1.51±0.3	6.6±6.3	-8.2±6.0

Table 3: Convergence times [sec] and success rates [%].

DVSS	NPVSS	NNVSS	SVSS	SNLMS
4.4s, 79%	7.1s, 63%	8.5s, 55%	8.6s, 51%	9.1s, 48%

is defined by $10\log_{10}\left[|s\left(n\right)|^2/|e\left(n\right)-s\left(n\right)|^2\right]$ and is affected by both echo levels and speech distortion levels. The PESQ we report is the average of the PESQ score between $s_{\mathrm{L}}\left(n\right)$ and $e_{\mathrm{L}}\left(n\right)$, and the PESQ score between $s_{\mathrm{R}}\left(n\right)$ and $e_{\mathrm{R}}\left(n\right)$. These measures are derived with running time frames of 20 ms with an overlap of 50%. For a complete view of performance, we report the adaptation convergence times and convergence success rates. Convergence is considered achieved when $\mathcal{D}\left(n\right)$ falls below -10 dB and is considered successful if $\mathcal{D}\left(n\right)$ remains below -10 dB until echo-paths change [4].

5. EXPERIMENTAL RESULTS

Our DVSS-SNLMS approach is matched against the VSS approaches in [13–15], correspondingly abbreviated "NNVSS", "NPVSS", and "SVSS". These competing algorithms were integrated with the widely-linear model and the SNLMS filter for an unbiased comparison. The SNLMS filter with step-size of $\mu = 3\times10^{-5}$, briefly "SNLMS", is the classic approach baseline. Performance in Tables 1 and 2 is outlined with mean and standard deviation (std) values, and Table 3 shows average test set values.

We distinguish between the performance when no echo-paths changes occur, i.e., in Table 1, from segments where echo-paths change, as in Table 2. In both cases, we only consider the post-convergence of the adaptive filter. In Table 1 and Table 2, the mean value of the results reflects the advantage of the DVSS method over the competition. The ERLE stresses the leading echo suppression of the DVSS method, and the SDR and PESQ measures reveal its ability to maintain low speech distortion and high speech quality. It is also noted that the low std values of the DVSS indicate the stability of its performance across various acoustic setups. Table 3 affirms that our method achieves the shortest re-convergence times and the most successful convergence rates in scenarios with no echo-path changes. Unlike the competition, our method has shown a prominent ability to track echo paths by adapting the step-size accurately and rapidly while maintaining high robustness and generalization

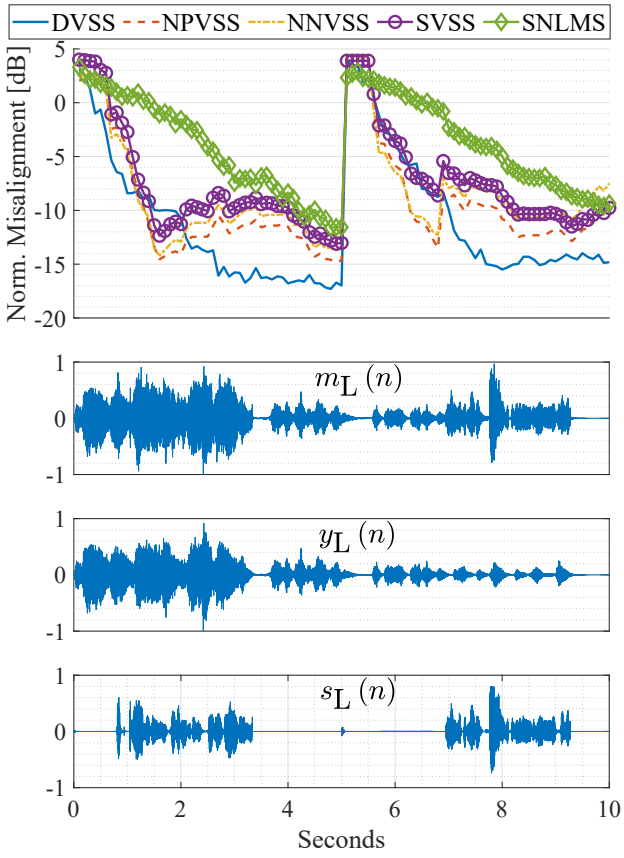

Figure 2: Convergence comparison. Near-end echo paths abruptly change at 5 s. SESR and SENR values regularly vary.

capabilities. This can be associated with our approach avoidance of heuristic parameter tuning and of making acoustic assumptions that often mismatch realistic scenarios.

Fig. 2 depicts the desired convergence behavior of the DVSS-SNLMS filter in a two-fold manner; it shows the most rapid convergence and re-convergence after abrupt echo-paths change, and it is also the least disturbed by the occurrence of double-talk. On the contrary, competing VSS-based methods slightly diverge due to double-talk, which impedes their convergence success afterward.

6. CONCLUSIONS

Controlling the step-size in adaptive filtering can allow for optimally operate between convergence rate and adaptation accuracy. This study attempts to bring this ability a step closer to practice by introducing a general adaptation-control framework that is both non-parametric and does not require acoustic assumptions and apply it to SAEC. Using the widely-linear model, we first derive the optimal step-size by minimizing the filter misalignment in the complex time domain. Then, we train a neural network to predict this optimal step-size from acoustic data in real time. Based on this step-size estimate, the SNLMS filter tracks the echo paths and performs well over competition across various acoustic setups. Future work may focus on generalization to scenarios where near-end microphones capture different versions of the speech and noise signals.

7. REFERENCES

[1] M. M. Sondhi, D. R. Morgan, and J. L. Hall, "Stereophonic acoustic echo cancellation - an overview of the fundamental problem," *IEEE Signal Process. Lett.*, vol. 2, no. 8, pp. 148–151, 1995.

[2] J. Benesty, D. R. Morgan, and M. M. Sondhi, "A better understanding and an improved solution to the specific problems of stereophonic acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 6, no. 2, pp. 156–165, 1998.

[3] S. G. Sankaran and A. Beex, "Stereophonic acoustic echo cancellation using NLMS with orthogonal correction factors," in *Proc. IWAENC*. Citeseer, 1999, pp. 40–43.

[4] C. Paleologu, S. Ciochină, J. Benesty, and S. L. Grant, "An overview on optimized NLMS algorithms for acoustic echo cancellation," *EURASIP J. Adv. Signal Process.*, vol. 2015, no. 1, pp. 1–19, 2015.

[5] M. M. U. Faiz and A. Zerguine, "A steady-state analysis of the ε-normalized sign-error least mean square (NSLMS) adaptive algorithm," in *Proc. Conf. Signals, Sys., Comput. (ASILOMAR)*. IEEE, 2011, pp. 538–541.

[6] N. Freire and S. C. Douglas, "Adaptive cancellation of geomagnetic background noise using a sign-error normalized LMS algorithm," in *Proc. ICASSP*, vol. 3. IEEE, 1993, pp. 523–526.

[7] J. Ni, J. Chen, and X. Chen, "Diffusion sign-error LMS algorithm: Formulation and stochastic behavior analysis," *Signal Process.*, vol. 128, pp. 142–149, 2016.

[8] M. Liu, M. J. Wang, and B. Y. Song, "An efficient architecture of the sign-error LMS adaptive filter," in *Proc. Solid-State and Integrated Circ. Tech. (ICSICT)*. IEEE, 2016, pp. 753–755.

[9] T. Haubner, M. M. Halimeh, A. Brendel, and W. Kellermann, "A synergistic kalman-and deep postfiltering approach to acoustic echo cancellation," in *Proc. EUSIPCO*. IEEE, 2021, pp. 990–994.

[10] T. Haubner, A. Brendel, M. Elminshawi, and W. Kellermann, "Noise-robust adaptation control for supervised acoustic system identification exploiting a noise dictionary," in *Proc. ICASSP*. IEEE, 2021, pp. 945–949.

[11] T. Haubner, A. Brendel, and W. Kellermann, "End-to-end deep learning-based adaptation control for frequency-domain adaptive system identification," in *Proc. ICASSP*. IEEE, 2022, pp. 766–770.

[12] J. Casebeer, N. J. Bryan, and P. Smaragdis, "Meta-AF: Meta-learning for adaptive filters," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 31, pp. 355–370, 2022.

[13] S. Meier and W. Kellermann, "Relative impulse response estimation during double-talk with an artificial neural network-based step size control," in *Proc. IWAENC*. IEEE, 2016, pp. 1–5.

[14] J. Benesty, H. Rey, L. R. Vega, and S. Tressens, "A nonparametric VSS NLMS algorithm," *IEEE Signal Process. Lett.*, vol. 13, no. 10, pp. 581–584, 2006.

[15] M. Hamidia and A. Amrouche, "Improved variable step-size NLMS adaptive filtering algorithm for acoustic echo cancellation," *Digital Signal Process.*, vol. 49, pp. 44–55, 2016.

[16] A. Ivry, I. Cohen, and B. Berdugo, "Nonlinear acoustic echo cancellation with deep learning," in *Proc. Interspeech*, 2021, pp. 4773–4777.

[17] ——, "Deep adaptation control for acoustic echo cancellation," in *Proc. ICASSP*. IEEE, 2022, pp. 741–745.

[18] R. Cutler, A. Saabas, T. Parnamaa, M. Loida, S. Sootla, *et al.*, "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*, 2021, pp. 4748–4752.

[19] J. Benesty, C. Paleologu, T. Gänsler, and S. Ciochină, *A Perspective on Stereophonic Acoustic Echo Cancellation*. Springer Science & Business Media, 2011, vol. 4.

[20] A. Ruszczynski, *Nonlinear optimization*. Princeton university press, 2011.

[21] W. W. Hager and H. Zhang, "A new active set algorithm for box constrained optimization," *SIAM J. on Opt.*, vol. 17, no. 2, pp. 526–557, 2006.

[22] S. Albawi, T. A. Mohammed, and S. Al-Zawi, "Understanding of a convolutional neural network," in *Proc. Int. Conf. Eng. Tech.* IEEE, 2017, pp. 1–6.

[23] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, 1984.

[24] *ETSI ES 202 740: Speech and multimedia Transmission Quality (STQ); Transmission requirements for wideband VoIP loudspeaking and handsfree terminals from a QoS perspective as perceived by the user*, ETSI Std., 2016.

[25] "NDP120 Syntiant™ Neural Processor," https://www.syntiant.com/ndp120, 2021.

[26] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. America*, vol. 65, no. 4, pp. 943–950, 1979.

[27] T. Gansler and J. Benesty, "New insights into the stereophonic acoustic echo cancellation problem and an adaptive nonlinearity solution," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 257–267, 2002.

[28] D. R. Morgan, J. L. Hall, and J. Benesty, "Investigation of several types of nonlinearities for use in stereo acoustic echo cancellation," *IEEE Trans. Speech Audio Process.*, vol. 9, no. 6, pp. 686–696, 2001.

[29] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *Proc. ICASSP*. IEEE, 2021, pp. 126–130.

[30] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.

[31] A. Ivry, I. Cohen, and B. Berdugo, "Objective metrics to evaluate residual-echo suppression during double-talk," in *Proc. WASPAA*, 2021.

[32] *ITU-T Rec. G.168: Digital network echo cancellers*, ITU-T Std., Feb. 2012.

[33] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, 2006.

[34] *ITU-T Rec. P.862.2: Wideband extension to recommendation P. 862 for the assessment of wideband telephone networks and speech codecs*, ITU-T Std., Oct. 2017.