ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-6654-0540-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICASSP43922.2022.9746557

DEEP ADAPTATION CONTROL FOR ACOUSTIC ECHO CANCELLATION

Amir Ivry *Israel Cohen* *Baruch Berdugo*

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

ABSTRACT

We propose a general framework for adaptation control using deep neural networks (NNs) and apply it to acoustic echo cancellation (AEC). First, the optimal step-size that controls the adaptation is derived offline by solving a constrained nonlinear optimization problem that minimizes the adaptive filter misadjustment. Then, a deep NN is trained to learn the relation between the input data and the optimal step-size. In real-time, the NN infers the optimal step-size from streaming data and feeds it to an NLMS filter for AEC. This data-driven method makes no assumptions on the acoustic setup and is entirely non-parametric. Experiments with 100 h of real and synthetic data show that the proposed method outperforms the competition in echo cancellation, speech distortion, and convergence during both single-talk and double-talk.

***Index Terms*—** Acoustic echo cancellation, adaptation control, variable step-size, double-talk, deep learning.

1. INTRODUCTION

Hands-free speech communication often involves a conversation between two speakers located at near-end and far-end points. During double-talk, the near-end microphone captures the desired-speech signal in addition to an echo produced by a loudspeaker that nonlinearly distorts and plays the far-end signal. The acoustic coupling between the loudspeaker and the microphone may lead to degraded speech intelligibility in the far-end due to echo presence [1]. Acoustic echo cancellation (AEC) aims to identify the echo path with an adaptive filter and create a replica of the echo that is subtracted from the microphone signal [2].

The normalized least mean squares (NLMS) filter is a popular adaptive filter since it is numerically stable and computationally efficient [3]. The NLMS integrates the normalized step-size parameter that governs the often conflicting fast convergence requirements and low misadjustment. Therefore, it is highly desirable to control the step-size during adaptation in practical scenarios of time-varying echo paths and double-talk. This problem has motivated numerous variable step-size (VSS) related studies. For example, Haubner et al. employed neural networks (NNs) for near-end estimation [4], noise estimation [5], and minimizing the error using adaptation control in the frequency domain [6]. Meier and Kellermann [7] employed a deep NN that maps statistical features of the far-end and a priori error signals to an analytically derived VSS. A batch of classic approaches includes the non-parametric VSS (NPVSS) that adjusts the step-size by reducing the squared error at each instant [8], the mean error sigmoid VSS (SVSS) that applies decompoisition of the error into sub-blocks [9], and Huang's VSS (HVSS) that estimates the system noise power to control the step-size update [10].

However, existing approaches make restricting assumptions in real-life setups, e.g., assuming a linear relationship between the echo and the far-end signals [4]– [10], and adopting a time-invariant echo-path [8]. In practice, these assumptions result in filter misadjustment and slow convergence rates during echo-path changes [11]. Also, such methods require tuning parameters that are difficult to control in real-life scenarios. For example, the NPVSS [8] involves estimating the noise power, which is challenging during double-talk.

We address these gaps by presenting a deep VSS (DVSS) framework. First, we solve a constrained nonlinear optimization problem that minimizes the normalized misalignment between the actual and estimated echo path. Second, we present a deep NN that learns the relation between the far-end, microphone, and a priori error signals and the optimal step-size. Finally, the trained NN produces the VSS estimate in real-time, which is fed to the NLMS filter for echo cancellation. This data-driven method makes no acoustic assumptions and is completely non-parametric. The end-to-end system, from the NN input to the NLMS output, comprises the proposed DVSS-NLMS filter. Notably, the DVSS framework can be generalized and is not restricted to NLMS-type algorithms.

For evaluation, we use 100 h of recordings from the AEC-challenge database [12] and compare the DVSS to five competing methods. Experiments show that the DVSS is advantageous in echo cancellation and speech distortion in double-talk, is more robust to high levels of speech and noise, and has a better generalization to various nonlinearities. The DVSS also achieves the best re-convergence times and success rates following abrupt echo-path changes during single-talk and double-talk across different acoustic conditions.

This work was supported by the Pazy Research Foundation.

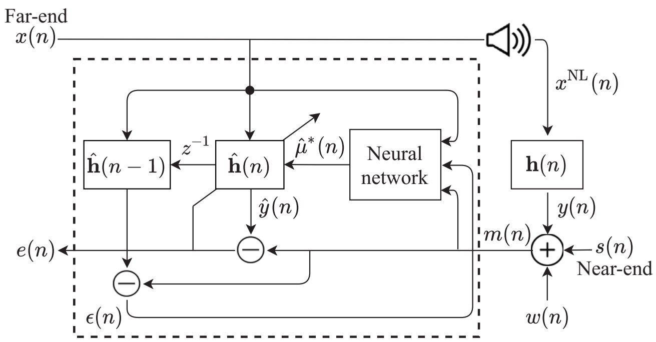

Fig. 1: AEC scenario and proposed system (bordered). The NN produces the DVSS estimate $\hat{\mu}^*(n)$, which is fed to an NLMS filter that generates the acoustic path estimation $\hat{\mathbf{h}}(n)$.

2. PROBLEM FORMULATION

Figure 1 illustrates the DVSS-NLMS configuration. The microphone signal $m(n)$ at time index n is given by

$$m(n) = y(n) + s(n) + w(n), \tag{1}$$

where $s(n)$ is the near-end speech signal, $w(n)$ represents environmental and system noises, and $y(n) = \mathbf{x}_{\mathrm{NL}}^T(n)\,\mathbf{h}(n)$ is a nonlinear and reverberant echo. $\mathbf{x}_{\mathrm{NL}}(n)$ denotes the L most recent samples of the far-end signal, $\mathbf{x}(n)$, after undergoing nonlinear distortions by nonideal components, and the echo path $\mathbf{h}(n)$ is modeled as a finite impulse response filter with L coefficients:

$$\mathbf{x}_{\mathrm{NL}}(n) = \left[x_{\mathrm{NL}}(n), \ldots, x_{\mathrm{NL}}(n-L+1)\right]^T, \tag{2}$$

$$\mathbf{h}(n) = \left[h_0(n), h_1(n), \ldots, h_{L-1}(n)\right]^T. \tag{3}$$

An NLMS adaptive filter with L coefficients tracks the echo path estimate $\hat{\mathbf{h}}(n)$ and echo estimate $\hat{y}(n) = \mathbf{x}^T(n)\,\hat{\mathbf{h}}(n)$:

$$\mathbf{x}(n) = \left[x(n), x(n-1), \ldots, x(n-L+1)\right]^T, \tag{4}$$

$$\hat{\mathbf{h}}(n) = \left[\hat{h}_0(n), \hat{h}_1(n), \ldots, \hat{h}_{L-1}(n)\right]^T. \tag{5}$$

Then, an estimate of the near-end speech signal is given by

$$\begin{aligned} e(n) &= m(n) - \hat{y}(n) \\ &= (y(n) - \hat{y}(n)) + s(n) + w(n). \end{aligned} \tag{6}$$

Our goal is to estimate $\hat{\mathbf{h}}(n)$ and to cancel the echo by eliminating $y(n) - \hat{y}(n)$, without distorting the speech $s(n)$.

3. DEEP VARIABLE STEP-SIZE ALGORITHM

3.1. General NLMS Filter Model in Double-talk

The a priori and a posteriori error signals of the NLMS adaptation process are, respectively, given by [3]:

$$\epsilon(n) = \mathbf{x}_{\mathrm{NL}}^T(n)\,\mathbf{h}(n) - \mathbf{x}^T(n)\,\hat{\mathbf{h}}(n-1) + s(n) + w(n), \tag{7}$$

$$e(n) = \mathbf{x}_{\mathrm{NL}}^T(n)\,\mathbf{h}(n) - \mathbf{x}^T(n)\,\hat{\mathbf{h}}(n) + s(n) + w(n). \tag{8}$$

Also, NLMS-type adaptive filters follow the update rule:

$$\hat{\mathbf{h}}(n) = \hat{\mathbf{h}}(n-1) + \mu(n)\,\mathbf{x}(n)\,\epsilon(n), \quad \hat{\mathbf{h}}(0) = \mathbf{0}^T, \tag{9}$$

where $\mu(n)$ is a positive step-size that controls the trade-off between convergence rate and adaptation misalignment and $\hat{\mathbf{h}}(0)$ has L zeros. From (7)–(9), we have

$$e(n) = \epsilon(n)\left[1 - \mu(n)\,\mathbf{x}^T(n)\,\mathbf{x}(n)\right]. \tag{10}$$

To derive the general expression for $\mu(n)$, we impose echo cancellation from the a posteriori error, namely:

$$e(n) = s(n) + w(n). \tag{11}$$

Assuming $s(n)$ and $w(n)$ are uncorrelated [3], substituting (11) into (10) yields

$$\mu(n) = \frac{1}{L\,\mathrm{E}\left[\mathbf{x}^2(n)\right] + \delta}\left[1 - \sqrt{\frac{s^2(n) + w^2(n)}{\epsilon^2(n)}}\right], \tag{12}$$

where $\mathrm{E}[\cdot]$ denotes empirical expectation and $\delta > 0$ is a regularization parameter added to avoid division by zero.

3.2. Data-driven Generation of the Optimal Step-Size

The normalized misalignment $\mathcal{D}(n)$ quantifies the mismatch between the actual and estimated echo paths in dB:

$$\begin{aligned} \mathcal{D}(n) &= 20\log_{10}\left[\frac{\|\mathbf{h}(n) - \hat{\mathbf{h}}(n)\|_2}{\|\mathbf{h}(n)\|_2}\right] \\ &= 20\log_{10}\left[\frac{\|\mathbf{h}(n) - \hat{\mathbf{h}}(n-1) - \mu(n)\,\mathbf{x}(n)\,e(n)\|_2}{\|\mathbf{h}(n)\|_2}\right]. \end{aligned} \tag{13}$$

The optimal step-size $\mu^*(n)$ is the solution of the constrained nonlinear optimization problem that minimizes $\mathcal{D}(n)$:

$$\mu^*(n) = \underset{0<\mu(n)<2}{\arg\min}\; \mathcal{D}(n), \tag{14}$$

where the constraint complies with the stability condition of NLMS-type algorithms [3]. This optimization process is carried out using the active-set optimization algorithm [13]. According to (13), merely the far-end and a priori error signals are required for $\mu^*(n)$. This allows a non-parametric and data-driven approach to estimate $\mu^*(n)$.

3.3. Optimal Step-Size Learning Using Neural Networks

Deriving $\mu^*(n)$ in practice is time-consuming and requires knowledge of the echo path. Thus, a deep NN is built to learn the relation between available data measurements and $\mu^*(n)$ during training, and to produce an estimate $\hat{\mu}^*(n)$ in real-time. According to (12), the step-size involves information of the far-end, a priori error, and near-end speech and noise signals. Even though the near-end signals are not available in practice, they comprise the available microphone signal.

Thus, we propose a deep NN that receives the far-end, a priori error, and microphone signals as inputs and maps them to the corresponding optimal step-size.

We employ a convolutional NN [14] with three input channels, one for each input signal, and a single-neuron output for the step-size. Each input channel is fed with its corresponding waveform signal's short-time Fourier transform (STFT) [15] amplitude. The first convolution layer employs a 3×3 kernel size, stride of 3, dilation of 5, and padding of 1, followed by 2-D batch normalization and a ReLU activation layer, and has 3 input and 16 output channels. A second convolution layer follows the same filtering specifications, but has 16 input and 16 output channels. A fully-connected NN unit receives the 16 filters and propagates their flatten version through a 1920×512 layer, followed by 1-D batch normalization, a ReLU activation function, and a dropout layer with a probability of 0.5. Finally, this outcome is concatenated to a second fully-connected layer with dimensions 512×1 that ends with a sigmoid activation function. The objective function is the ℓ_2 distance between the NN prediction and the optimal step-size $\mu^*(n)$.

In real-time, the NN produces $\hat{\mu}^*(n)$, which is fed to the succeeding NLMS. This end-to-end system contains 1 Million parameters that consume 4 Million floating-point operations per second (Mflops) and 4.6 Megabytes (MB) of memory. Thus, its integration on hands-free devices is enabled with hands-free communication timing constraints met [16], e.g., using the NDP120 neural processor by Syntiant™ [17].

4. EXPERIMENTAL SETUP

4.1. Database Acquisition

The AEC challenge database [12] is employed in this study. This corpus is sampled at 16 kHz and includes single-talk and double-talk periods both with and without echo-path change. No echo-path change means no movement in the room during the recording, and echo-path change means either the near-end speaker or the device are moving during the recording. The corpus includes 25 h of synthetic data and 75 h of real clean and noisy data. To account for realistic acoustic environments, every far-end signal randomly undergoes one of 4500 simulated nonlinear modifications, generated according to the physical behavior of power amplifiers and loudspeakers in modern hands-free devices [11]. Also, every nonlinearly-distorted signal is randomly propagated via one of 4500 real room impulse responses that are taken from the corpus in [18] with their first L coefficients. The echo-to-speech ratio (ESR) and echo-to-noise ratio (ENR) levels were distributed on $[-10, 10]$ dB and $[0, 40]$ dB, respectively, and are defined as ESR=$10\log_{10}\left[\|y(n)\|_2^2/\|s(n)\|_2^2\right]$ and ENR=$10\log_{10}\left[\|y(n)\|_2^2/\|w(n)\|_2^2\right]$ in dB, both calculated with 50% overlapping time frames of 20 ms.

Table 1: Performance measures for evaluation.

Measure	Definition
ERLE	$10\log_{10}\frac{\|m(n)\|_2^2}{\|e(n)\|_2^2}\Big\|_{\text{Far-end single-talk}}$
SDR	$10\log_{10}\frac{\|s(n)\|_2^2}{\|e(n)-s(n)\|_2^2}\Big\|_{\text{Double-talk}}$

4.2. Data Processing, Training, and Testing

Initially, the 100 h of real and synthetic data are randomly split to create 80 h of training, 10 h of validation, and 10 h of test sets. All sets are balanced to prevent biased results, as detailed in [19]. The training and validation sets are used for step-size generation via (14) with $\mu(0) = 3 \times 10^{-5}$, $L = 150$ ms, and $\hat{\mathbf{h}}(0) = \mathbf{0}^T$ being a vector of L zeros. The step-size is generated every 8 ms to avoid unnecessary heavy computations. An abrupt change in echo path reoccurs every t seconds, where $t \sim U[4.5, 5.5]$, resembling real-life scenarios. The signals are transformed by the STFT using 16 ms frames and 8 ms shifts. Past information of 96 ms is concatenated before entering the NN. Training the NN is done using back-propagation through time with a 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 [20]. In real-time, the NN infers the test set and is not updated. The NLMS receives the optimal step-size estimate from the NN and continuously tracks the echo path. The NN may introduce an artificial gain, which is compensated as in [21]. Training duration was 30 minutes per 1 h of data, and the batch inference time of the end-to-end system, i.e., the NN and adaptive filter, is 24 ms on an Intel Core i7-8700K CPU @ 3.7 GHz with two GPUs of Nvidia GeForce RTX 2080 Ti.

4.3. Performance Measures

To evaluate the performance, the echo return loss enhancement (ERLE) [22] is used. It measures echo reduction between the degraded and enhanced signals when only a far-end signal and noise are present. In double-talk, we use the signal-to-distortion ratio (SDR) [23] that takes echo suppression and speech distortion into account, and the perceptual evaluation of speech quality (PESQ) [24]. All measures are calculated with 50% overlapping frames of 20 ms, and the ERLE and SDR are defined in Table 1. Convergence times and success rates are also given. Convergence occurs when $\mathcal{D}(n)$ falls under -10 dB and is successful if that holds for the remaining echo path. We also report the value of $\mathcal{D}(n)$ as given in (13).

5. EXPERIMENTAL RESULTS

Using the entire test set, the DVSS method is compared against four competing VSS-based methods in [7]– [10], respectively notated "NNVSS", "NPVSS", "SVSS", and

Table 2: Performance with no echo-path change.

	SDR	PESQ	ERLE	Norm. Mis.
DVSS	**3.51±0.4**	**2.52±0.3**	**21.3±4.6**	**-22.8±4.2**
NNVSS	2.48±0.9	1.78±0.4	15.5±5.7	-16.8±4.9
NPVSS	2.81±0.8	2.06±0.5	16.8±6.7	-18.1±5.7
SVSS	2.21±0.9	2.03±0.6	15.0±5.5	-16.3±5.0
HVSS	2.86±0.6	2.12±0.4	18.1±6.5	-19.9±6.2
NLMS	2.09±1.1	1.62±0.3	14.2±5.8	-15.5±4.9

Table 3: Performance with echo-path change.

	SDR	PESQ	ERLE	Norm. Mis.
DVSS	**3.16±0.6**	**2.31±0.5**	**16.9±5.7**	**-18.3±5.2**
NNVSS	2.11±1.1	1.75±0.5	11.9±5.5	-11.9±4.9
NPVSS	2.57±1.0	1.99±0.6	15.9±7.7	-17.4±7.1
SVSS	2.03±1.2	1.80±0.7	15.0±6.1	-13.4±5.9
HVSS	2.62±0.9	2.03±0.5	12.7±5.7	-15.1±4.2
NLMS	1.95±1.4	1.56±0.3	10.2±4.1	-11.0±3.0

Table 4: Convergence times [seconds] and success rates [%].

DVSS	NNVSS	NPVSS	SVSS	HVSS	NLMS
3.4s, 95%	5.9s, 77%	6.6s, 75%	5.6s, 83%	7.0s, 71%	7.9s, 58%

"HVSS". All methods are implemented with the NLMS filter, which is also implemented with a constant step-size of $\mu = 3 \times 10^{-5}$ as the benchmark, notated "NLMS". In Tables 2 and 3, measures are reported by their mean and standard deviation (std) values in the format mean±std. In Table 4, the average convergence times and success rates are reported.

Results with no echo-path change are given in Table 2 and with echo-path change are shown in Table 3, both after convergence. According to the ERLE measure, the proposed method achieves leading echo cancellation in single-talk. The DVSS yields less speech distortion and better speech quality during double-talk, respectively deduced by the SDR and PESQ scores. A lower std value is also achieved, which implies better stability of the DVSS across various setups. Although scenarios of echo-path change lead to expected performance decline relative to no echo-path change, our method outperforms competing methods across all measures in terms of mean and std. Furthermore, by Table 4, our method achieves the fastest average re-convergence time and highest convergence success rate compared to the competition. Thus, the data-driven DVSS that requires no acoustic assumptions and is entirely non-parametric, can track the VSS in practical acoustic conditions with double-talk with high generalization and robustness, and adjust the VSS most accurately and rapidly.

Convergence comparison is illustrated in Fig. 2, where the ESR and ENR continuously vary, and after 5 s, an abrupt echo-path change occurs. On the other hand, the DVSS-NLMS filter continues to converge during double-talk and is only disturbed by the abrupt echo-path change. Also, the DVSS rapid convergence and re-convergence are demonstrated. However, all VSS-based competing methods experience divergence due to double-talk, which degrades their adaptation process. This supports previous conclusions regarding the DVSS superiority in real acoustic conditions, including double-talk and echo-path changes.

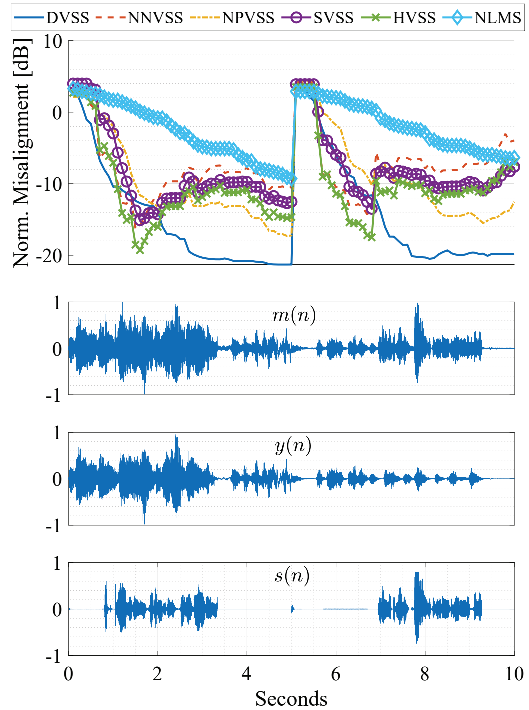

Fig. 2: Convergence comparison. Abrupt echo-path change occurs after 5 s, and ESR and ENR values regularly change.

6. CONCLUSIONS

We have introduced a general framework for real-time adaptation control using deep learning. We first performed optimal VSS generation that is entirely non-parametric and makes no acoustic assumptions via minimization of the filter misalignment. Second, the relation of the data and the optimal VSS was learned via a deep NN. Finally, in real-time, the NN yields a VSS estimate that is fed into the adaptive filter that continuously tracks the echo path. Experiments using 100 h of real and synthetic data showed superior performance of the DVSS over the competition in AEC using the NLMS filter. In particular, the DVSS is preferable during double-talk in terms of echo cancellation and speech distortion, and characterized by faster convergence following abrupt echo-path changes.

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, T. Gänsler, D. R. Morgan, M. M. Sondhi, S. L. Gay, et al., *Advances in Network and Acoustic Echo Cancellation*, New York: Springer, 2001.

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

[4] T. Haubner, M. M. Halimeh, A. Brendel, W. Kellermann, et al., "A synergistic Kalman-and deep postfiltering approach to acoustic echo cancellation," *preprint arXiv:2012.08867*, 2020.

[5] 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.

[6] T. Haubner, A. Brendel, and W. Kellermann, "End-to-end deep learning-based adaptation control for frequency-domain adaptive system identification," *preprint arXiv:2106.01262*, 2021.

[7] 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.

[8] 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.

[9] 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.

[10] H. C. Huang and J. Lee, "A new variable step-size NLMS algorithm and its performance analysis," *IEEE Trans. on Signal Process.*, vol. 60, no. 4, pp. 2055–2060, 2011.

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

[12] 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.

[13] 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.

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

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

[16] "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," 2016.

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

[18] I. Szöke, M. Skácel, L. Mošner, J. Paliesek, and J. Černockỳ, "Building and evaluation of a real room impulse response dataset," *IEEE J. of Selected Topics in Signal Process.*, vol. 13, no. 4, pp. 863–876, 2019.

[19] 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.

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

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

[22] "ITU-T Rec. G.168: Digital network echo cancellers," Feb. 2012.

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

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