Nonlinear Acoustic Echo Cancellation with Deep Learning

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
sivry@campus.technion.ac.il, icohen@ee.technion.ac.il, bbaruch@technion.ac.il

Abstract

We propose a nonlinear acoustic echo cancellation system, which aims to model the echo path from the far-end signal to the near-end microphone in two parts. Inspired by the physical behavior of modern hands-free devices, we first introduce a novel neural network architecture that is specifically designed to model the nonlinear distortions these devices induce between receiving and playing the far-end signal. To account for variations between devices, we construct this network with trainable memory length and nonlinear activation functions that are not parameterized in advance, but are rather optimized during the training stage using the training data. Second, the network is succeeded by a standard adaptive linear filter that constantly tracks the echo path between the loudspeaker output and the microphone. During training, the network and filter are jointly optimized to learn the network parameters. This system requires 17 thousand parameters that consume 500 million floating-point operations per second and 40 kilo-bytes of memory. It also satisfies hands-free communication timing requirements on a standard neural processor, which renders it adequate for embedding on hands-free communication devices. Using 280 hours of real and synthetic data, experiments show advantageous performance compared to competing methods.

Index Terms: Nonlinear acoustic echo cancellation, deep learning, hands-free communication, on-device implementation

1. Introduction

Hands-free communication often involves a conversation between two speakers located at near-end and far-end points. The near-end microphone captures the desired-speech signal and two interfering signals: echo produced by a loudspeaker playing the far-end signal, and background noises. The acoustic coupling between the loudspeaker output and the microphone may lead to degraded speech intelligibility in the far-end due to echo presence [1]. This problem prompted numerous studies regarding acoustic echo cancellation (AEC) systems that aim to remove echo and preserve the near-end speech [2]. In recent years, however, miniaturization of electronic components in hands-free devices, e.g., smart phones, smart speakers, and wearable devices, caused non-negligible nonlinear (NL) distortions in the echo path between the far-end signal and the loudspeaker output [3].Consequently, AEC systems that assume an echo path that is linear often fail in practice [4].

To mitigate this mismatch, various nonlinear acoustic echo cancellation (NLAEC) approaches were proposed to identify the NL echo path. The Volterra series showed success in modeling systems with weak nonlinearities and memory using NL basis functions, while often requiring high computational complexity [5]. A simplified version is given by the block-oriented Hammerstein and Wiener models, which describe NL systems without memory and linear systems with memory [6]. Also, adaptive functional link filters [7], Bayesian state-space modeling [8], and kernel-based methods [9] are commonly used for NLAEC. Avargel and Cohen considered this problem from a time-frequency point-of-view and applied multiplicative function approximation [10], sub-band adaptive filtering [11], and an efficient Volttera series modeling using cross-band terms [12], [13]. Neural networks (NNs) provide an alternative framework for a more accurate NL modeling compared to classic approaches [14], [15], [16], [17]. For instance, Malek and Koldovsky [18] estimated the NL echo path with a fully-connected NN (FCNN) that assumes the Hammerstein model, followed by an adaptive linear filter to track the acoustic path. Recently, Halimeh et al. [19] constructed an FCNN that assumes the Wiener-Hammerstein model and captures both the NL and linear echo paths.

Despite showing promising results, the performance of these methods is still challenging in real-life scenarios, which may be associated with two of their attributes. First, these models are not accurately designed according to the physical behavior of distortions that modern hands-free devices apply to the far-end signal. Second, they are mostly parametric, i.e., they require that memory lengths and NL basis functions are predetermined. E.g., in [5], [6], the presented models assume a given number of memory taps, and in [18], [19], fixed NL activation functions are employed inside the NN. These drawbacks may produce sub-optimal solutions in real setups.

To address these two gaps, we make two contributions that are inspired by the physical behavior of modern hands-free devices. We first introduce a novel NN architecture that is specifically designed to model the distortions these devices induce between receiving and playing the far-end signal. Second, we construct this NN with trainable memory length and NL activation functions that are not parameterized in advance, but are rather optimized during the training stage based on the training data. The NN output is inserted into a standard adaptive linear filter that constantly tracks the acoustic path from the loudspeaker output to the microphone. The end-to-end system, from the input of the NN to the output of the linear filter, forms the proposed NLAEC system. During training, the NN and the linear filter are jointly optimized to learn the NN parameters. In testing, the NN is used for inference and is not updated, while the linear filter is adapted to the time-varying acoustic paths.

This system requires 17 thousand parameters that consume 500 million floating-point operations per second (Mflops) and 40 kilo-bytes (KB) of memory, which renders it applicable for embedding on hands-free communication devices. It also meets the timing requirements of the AEC challenge [20], and more generally the constraints of hands-free communication standards [21] on a standard neural processor.

This research was supported by the Pazy Research Foundation and the ISF-NSFC joint research program (grant No. 2514/17). The authors thank Stem Audio for providing equipment and technical guidance.

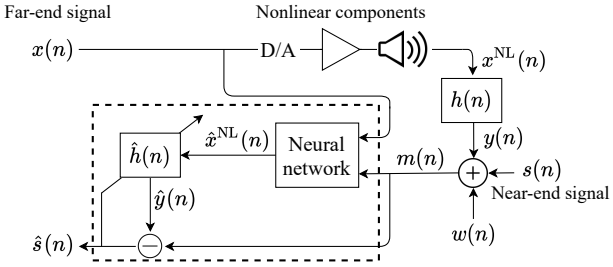

Figure 1: *NLAEC scenario and proposed system (bordered). The nonlinear components are modeled with a neural network and the acoustic path with a standard adaptive linear filter.*

Performance is evaluated against two recent NN-based NLAEC methods in [18] and [19], and to a linear AEC method. Experiments are conducted with 280 h of both synthetic and real data, which include half-duplex and full-duplex periods affiliated with various acoustic environments, devices, speakers, and noise and echo levels. Results show leading performance of the proposed NLAEC system in terms of echo cancellation and speech distortion levels, generalization and stability to various setups, robustness to high levels of noise and echo, and convergence and re-convergence rates.

The reminder of this paper is organized as follows. In Section 2, we formulate the problem. In Section 3, we introduce the proposed NLAEC system. In Section 4, we describe the experimental setup. In Section 5, we demonstrate the performance of the proposed system. Finally, we conclude in Section 6.

2. Problem Formulation

Figure 1 depicts the scenario and proposed system for NLAEC. Let $s\left(n\right)$ be the near-end speech signal and let $x\left(n\right)$ be the far-end speech signal. The microphone signal $m\left(n\right)$ is given by

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

where $w\left(n\right)$ represents additive environmental and system noises and $y\left(n\right)$ is a nonlinear reverberant echo that is generated from $x\left(n\right)$. The far-end signal, $x\left(n\right)$, is first distorted by electrical components that produce $x^{\mathrm{NL}}\left(n\right)$, and then $x^{\mathrm{NL}}\left(n\right)$ propagates via a linear acoustic path $h\left(n\right)$, namely $y\left(n\right)=x^{\mathrm{NL}}\left(n\right)*h\left(n\right)$. The proposed NLAEC system attempts to estimate $y\left(n\right)$ by using an NN to find $\hat{x}^{\mathrm{NL}}\left(n\right)$, which is an estimate for $x^{\mathrm{NL}}\left(n\right)$, and filtering the result with an adaptive linear filter that tracks the acoustic path, denoted by $\hat{h}\left(n\right)$:

$$\hat{y}\left(n\right)=\hat{x}^{\mathrm{NL}}\left(n\right)*\hat{h}\left(n\right). \tag{2}$$

The signal transmitted to the far-end is given by

$$\hat{s}\left(n\right)=m\left(n\right)-\hat{y}\left(n\right)=s\left(n\right)+\left(y\left(n\right)-\hat{y}\left(n\right)\right)+w\left(n\right). \tag{3}$$

Our goal is to cancel the echo $y\left(n\right)$ by eliminating the term $y\left(n\right)-\hat{y}\left(n\right)$, without distorting the speech signal $s\left(n\right)$.

3. Nonlinear Acoustic Echo Cancellation

The proposed NLAEC system is comprised of two parts. First, an NN models the physical behavior of distortions applied between the far-end signal and the loudspeaker output, caused by non-ideal electrical components in practical hands-free devices.

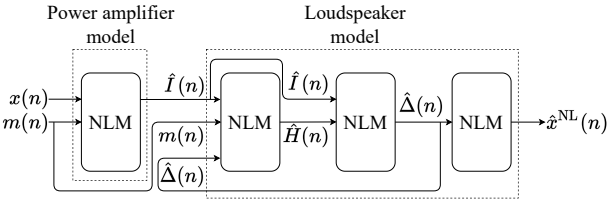

Figure 2: *Proposed neural network architecture.*

Second, a standard adaptive linear filter tracks the acoustic echo path from the loudspeaker output to the microphone.

In order to understand our system, it is helpful to understand how the above-mentioned electrical components behave. Modern hands-free devices often apply distortions between receiving the far-end signal and playing it in the near-end. These distortions are created by three different electrical components; a digital-to-analog converter (D/A), a power amplifier, and a loudspeaker [22], [23], [24], [25]. This study uses a 16-bit data precision, so the signal-to-quantization-noise ratio is sufficiently high and the D/A distortions are numerically negligible [22]. Thus, the D/A is not modeled. Ideally, the power amplifier should increase the energy of its input signal without distortions by using the power supply from the device battery. However, low-powered hands-free devices drive the amplifier to operate close to saturation, which yields distortions. The specific NL behavior of each amplifier depends on its saturation curve, ranging from a soft-clipped sigmoid, to a hard-clipped rectified function, and in extreme cases, it may exhibit a square waveform behavior [22].

The loudspeaker component is responsible for the majority of distortions. In this study, the widely-used electro-dynamic loudspeaker model is considered, which exhibits four major types of nonlinearities; electrical, magnetic, mechanical, and acoustical [25]. The electrical signal, $I\left(n\right)$, is received from the amplifier output and creates a magnetic field signal of strength $H\left(n\right)$ around the voice coil, which renders it an electromagnet. The relation between $I\left(n\right)$ and $H\left(n\right)$ is NL and depends on the coil displacement signal, $\Delta\left(n\right)$. Both $I\left(n\right)$ and $H\left(n\right)$ lead to polarity changes in the electromagnet that moves the coil back and forth with force that also has NL relations with $\Delta\left(n\right)$. This movement creates air pressure that is translated into acoustic sound waves that depend on $\Delta\left(n\right)$ and its temporal derivatives. This relation is NL as well due to wave propagation and mechanical nonlinearities, caused by stiffness of the loudspeaker spider. Both the power amplifier and loudspeaker components may depend on previous observations.

The above NL behavior is modeled using an NN that is comprised of two cascaded parts: a power amplifier model, and a loudspeaker model, depicted in Figure 2. First, the amplifier is modeled with 3 identical gated recurrent units (GRUs) that contain 16 cells each [26] and dropout [27] in the recurrent layers, an FCNN with a one-neuron output, and a piecewise linear unit activation function layer (PLU) with trainable parameters [28]. This entire NL model (NLM) is fed with the far-end and microphone waveform signals, since the latter contains information about the distortions of the former. Second, the loudspeaker is modeled by a sequence of 3 consecutive NLMs. It receives the output of the amplifier, i.e., the estimated excitation current $\hat{I}\left(n\right)$ that drives the loudspeaker. Similarly to the amplifier model, $\hat{I}\left(n\right)$ is concatenated to the microphone signal, and the first NLM learns the electrical-to-

magnetic NL model from $\hat{I}(n)$ to $\hat{H}(n)$. Then, the predicted $\hat{H}(n)$ is concatenated to $\hat{I}(n)$ and inserted to the second NLM, which learns the magnetic-to-mechanical NL model and predicts $\hat{\Delta}(n)$. Then, $\hat{\Delta}(n)$ is inserted to the third NLM, which learns the mechanical-to-acoustic NL model and estimates the distorted far-end signal at the output of the loudspeaker, i.e., $\hat{x}^{\text{NL}}(n)$. Since $\hat{\Delta}(n)$ also affects $\hat{H}(n)$, the first NLM is fed with the output of the second NLM using a skip-connection. The NLM unit is adjusted to receive between 1 to 3-dimensional input signals across the NN model. Following this NN, a linear adaptive filter models the acoustic path between the loudspeaker output and the microphone. This filter contains 150 samples and was developed by Phoenix Audio Technologies™ using a filter bank approach. The NN and the linear filter construct the proposed end-to-end NLAEC system.

To the best of our knowledge, the proposed NN architecture is used in this study for the first time. The NN is based on the GRU, whose internal gate-based mechanism is optimized for NL sequence-to-sequence mapping in the waveform domain. Also, the GRU keeps relevant past information without discarding it through time, while neglecting irrelevant data. Thus, the optimal memory length is implicitly learned by the NN during training and should not be set in advance. The trainable PLU parameters are also adjusted during training to optimally describe various saturation curves of the power amplifier and other NL behaviors exhibited by the loudspeaker. Thus, the NL behavior of the NN is not restricted to a predetermined set of NL basis functions. In addition, the GRU consumes low computational resources and requires short inference time.

The NLAEC system contains 17 thousand parameters that consume 500 Mflops and 40 KB of memory. Thus, its integration on hands-free devices is enabled, e.g., using the NDP120 neural processor by Syntiant™ [29]. Timing constraints of hands-free communication on that processor are also met [21].

4. Experimental Setup

4.1. Database Acquisition

Two data corpora are employed in this study; the AEC challenge database [20], and a database recorded in our lab, both sampled at 16 kHz. These corpora include single-talk and double-talk periods both with and without echo-path change. In the case of no echo-path change, there is no movement in the room during the recording. In the other case, either the near-end speaker or the device are constantly moving during the recording. In [20], two open sources of synthetic and real recordings are introduced. The synthetic data includes 100 h, and the real data contains 140 h of audio clips, generated from $5,000$ hands-free devices that are used in various acoustic environments. In both real and synthetic cases, signal-to-echo ratio (SER) and signal-to-noise ratio (SNR) levels were distributed on $[-10, 10]$ dB and $[0, 40]$ dB, respectively. Additional real recordings were conducted in our lab to test the generalization of the system to unseen setups and its robustness to extremely low levels of SERs. This database is fully described in [30]. For completion, it contains 40 h of recordings from the TIMIT [31] and LibriSpeech [32] corpora with SNR levels of 32 ± 5 dB and SER levels distributed on $[-20, -10]$ dB.

Formally, the SER and SNR captured by the microphone are defined as SER=$10\log_{10}\left[\|s(n)\|_2^2/\|y(n)\|_2^2\right]$ and SNR=$10\log_{10}\left[\|s(n)\|_2^2/\|w(n)\|_2^2\right]$ in dB, and are calculated with 50% overlapping time frames of 20 ms.

Table 1: *Performance metrics for NLAEC.*

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

4.2. Data Processing, Training, and Testing

The real and synthetic data from [20] is randomly split to create 185 h of training set and 45 h of validation set. The test set contains only real data that is comprised of the remaining 10 h from [20] and all 40 h from [30]. Each set is divided into 10 s segments that contain recordings in different setups. This leads to frequent re-convergence during transitions between segments, both without and with echo-path change. These sets are balanced to prevent bias in results, as detailed in [30].

During training, the NN and the succeeding linear filter are jointly optimized to learn the NN parameters. Optimization is done by minimizing the ℓ_2 distance between the output of the NLAEC, $\hat{s}(n)$, and the desired-near-end speech $s(n)$. To train the NN, back-propagation through time is used with a learning rate of 0.0005, mini-batch size of 32 ms, and 20 epochs, using Adam optimizer [33]. Also, automatic differentiation [34] is applied, since the loudspeaker modeling involves temporal derivatives of its input signals. Training duration was typically 15 minutes per 10 h of data on an Intel Core i7-8700K CPU @ 3.7 GHz with two GPUs of type Nvidia GeForce RTX 2080 Ti.

During testing, the NN is used for inference only and is not updated. The linear filter receives the outputs of the NN and is continuously adapted to account for time variations of the acoustic path. An artificial gain may be introduced by the NN, which is compensated as shown in [35].

4.3. Performance Measures

To evaluate performance, the echo return loss enhancement (ERLE) [36] is used. It measures echo reduction between the degraded and enhanced signals when only a far-end signal and noise are present. For double-talk periods, we use the signal-to-distortion ratio (SDR) [35] that takes echo suppression and speech distortion into account, and the perceptual evaluation of speech quality (PESQ) [37], [38]. The PESQ is calculated over an entire 10 s segment. The ERLE and SDR are calculated with 50% overlapping frames of 20 ms, and are defined in Table 1.

5. Experimental Results

The performance of the proposed NLAEC system is compared against two competing NN-based methods in [18] and [19], notated "Malek" and "Halimeh", respectively. To approximate the linear echo path, the proposed system and "Malek" are implemented here with an identical adaptive linear filter mentioned in Section 3, while "Halimeh" employs a linear echo approximation via an NN. As benchmark, the linear filter is also applied alone, and this method is denoted by "Linear". Measures are reported by their mean and standard deviation (std) values, with respect to the test set specified in each experiment. Unless stated otherwise, the format of the results is presented as mean±std. In this study, convergence was reached if the normalized misalignment between consecutive linear echo approximations was lower than -30 dB [39].

Table 2: *Performance with no echo-path change.*

	Proposed	Halimeh	Malek	Linear
ERLE	**26.4±5.1**	23.1±5.9	22.6±6.7	21.3±7.2
PESQ	**3.17±0.4**	2.88±0.5	2.64±0.5	2.02±0.7
SDR	**5.37±0.4**	4.83±0.6	4.37±0.8	3.01±0.9

Table 3: *Performance with echo-path change.*

	Proposed	Halimeh	Malek	Linear
ERLE	**23.2±6.0**	19.2±7.7	18.0±8.3	16.9±8.9
PESQ	**2.92±0.5**	2.54±0.7	2.31±0.6	1.91±0.6
SDR	**5.08±0.6**	4.25±0.9	3.82±0.9	2.52±1.0

Table 4: *Performance before convergence.*

	Proposed	Halimeh	Malek	Linear
ERLE	**19.7±7.5**	14.9±8.1	13.8±8.8	11.0±9.6
PESQ	**2.56±0.6**	1.98±0.7	1.91±0.7	1.75±0.6
SDR	**4.71±0.9**	3.58±1.2	3.04±1.3	1.54±1.3

Table 5: *Convergence time in seconds.*

Proposed	Halimeh	Malek	Linear
4.6±0.7	6.6±1.1	7.3±1.4	7.9±1.8

Results for segments with no echo-path change are given in Table 2 and for segments with echo-path change are given in Table 3, both after convergence. Compared to competition, the proposed method achieves enhanced echo cancellation in single-talk periods according to the ERLE measure. In double-talk periods, less speech distortion and better speech quality are obtained, as suggested by the SDR and PESQ scores, respectively. Also, a lower std measure is achieved, which projects better stability of our method across various setups. Scenarios of echo-path change lead to overall decline in performance relative to no echo-path change, as expected. However, our method still prevails competition across all measures in terms of both higher mean and lower std. Based on the above, our method allows enhanced modeling of the NL echo path, which improves both the estimation of acoustic paths with no echo-path change, and the tracking of acoustic paths with echo-path change.

In addition, we investigate the performance before convergence and during re-convergence for segments with no echo-path change. Due to the test set segmentation described in Section 4.2, re-convergence frequently occurs during transitions between segments. As shown in Table 4, performance is collectively impeded relative to the converged case in Table 2. However, our method still prevails across all measures in terms of both mean and std values. This indicates the high sensitivity of competing methods to converged echo approximation, while our model captures the behavior of the echo even from degraded measurements. We also examine the convergence time of each method. According to Table 5, our method achieves the shortest convergence time compared to competition. Again, it can be suggested that enhanced modeling of the NL echo path is obtained by the proposed NN, which allows the succeeding linear filter to be adjusted more accurately and rapidly.

Next, performance with no echo-path change is examined in various SNR and SER levels, after convergence. As shown in

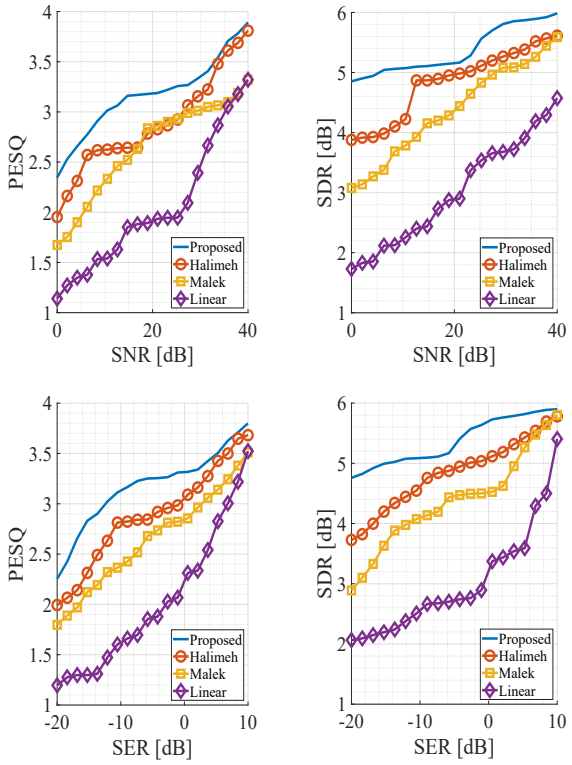

Figure 3: *Comparison of average PESQ and SDR measures in various SNR and SER levels.*

Figure 3, all methods suffer from decline in performance when acoustic conditions deteriorate. However, our method outperforms competition in both PESQ and SDR measures across all SNR and SER levels, which projects high generalization ability to various levels of noise and echo. The relatively stable behavior of the proposed method, especially in low levels of SNRs and SERs, indicates high robustness to high levels of noise and echo that often occur in practice. Interestingly, in severely degraded conditions of 0 dB SNR and of -20 dB SER, the proposed method achieves roughly 1 dB higher SDR and 0.5 higher PESQ score on average than the competition in second place.

6. Conclusion

We have presented an NLAEC system that comprises a novel NN architecture and a succeeding standard adaptive linear filter. To describe the distortions modern hands-free devices induce between receiving and playing the far-end signal, we constructed the NN of a power amplifier model followed by a loudspeaker model. The adaptive filter is fed by the NN and tracks the acoustic path from the loudspeaker output to the microphone. The NN parameters are updated during training using joint optimization of the NN and the filter. The NLAEC implementation is adequate for integration on hands-free devices, and can meet timing requirements of hands-free communication standards on a standard neural processor. Experiments with 280 h of real and synthetic recordings demonstrate the improved performance of our method compared to competition in terms of echo suppression and desired-signal distortion, generalization and stability in various setups, robustness to high levels of noise and echo, and convergence and re-convergence times.

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 Processing Letters*, 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] A. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. Workshop on Applications of Signal Processing to Audio and Accoustics (WASPAA)*. IEEE, 1995, pp. 103–106.

[4] M. I. Mossi, N. W. Evans, and C. Beaugeant, "An assessment of linear adaptive filter performance with nonlinear distortions," in *Proc. ICASSP*. IEEE, 2010, pp. 313–316.

[5] A. Guérin, G. Faucon, and R. Le Bouquin-Jeannès, "Nonlinear acoustic echo cancellation based on Volterra filters," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 672–683, 2003.

[6] M. Scarpiniti, D. Comminiello, R. Parisi, and A. Uncini, "Comparison of Hammerstein and Wiener systems for nonlinear acoustic echo cancelers in reverberant environments," in *Proc. International Conference on Digital Signal Processing*. IEEE, 2011, pp. 1–6.

[7] D. Comminiello, M. Scarpiniti, L. A. Azpicueta-Ruiz, J. Arenas-García, and A. Uncini, "Functional link adaptive filters for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1502–1512, 2013.

[8] S. Malik and G. Enzner, "State-space frequency-domain adaptive filtering for nonlinear acoustic echo cancellation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 2065–2079, 2012.

[9] S. Van Vaerenbergh, L. A. Azpicueta-Ruiz, and D. Comminiello, "A split kernel adaptive filtering architecture for nonlinear acoustic echo cancellation," in *Proc. EUSIPCO*. IEEE, 2016, pp. 1768–1772.

[10] Y. Avargel and I. Cohen, "Nonlinear acoustic echo cancellation based on a multiplicative transfer function approximation," in *Proc. IWAENC*. Citeseer, 2008, pp. 1–4.

[11] ——, "Adaptive nonlinear system identification in the short-time Fourier transform domain," *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3891–3904, 2009.

[12] ——, "Modeling and identification of nonlinear systems in the short-time Fourier transform domain," *IEEE Transactions on Signal Processing*, vol. 58, no. 1, pp. 291–304, 2009.

[13] ——, "Representation and identification of nonlinear system in the short-time Fourier transform domain," in *Speech Processing in Modern Communication: Challenges and Perspectives*, I. Cohen, J. Benesty, and S. Gannot, Eds. Springer, 2010, ch. 3, pp. 49–88.

[14] A. N. Birkett and R. A. Goubran, "Acoustic echo cancellation using NLMS-neural network structures," in *Proc. ICASSP*, vol. 5. IEEE, 1995, pp. 3035–3038.

[15] A. B. Rabaa and R. Tourki, "Acoustic echo cancellation based on a recurrent neural network and a fast affine projection algorithm," in *Proc. Annual Conference Industrial Electronics Society (IECON)*, vol. 3. IEEE, 1998, pp. 1754–1757.

[16] A. Janczak, *Identification of nonlinear systems using neural networks and polynomial models: a block-oriented approach*. Springer Science & Business Media, 2004, vol. 310.

[17] S. Zhang and W. X. Zheng, "Recursive adaptive sparse exponential functional link neural network for nonlinear AEC in impulsive noise environment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 9, pp. 4314–4323, 2017.

[18] J. Malek and Z. Koldovskỳ, "Hammerstein model-based nonlinear echo cancelation using a cascade of neural network and adaptive linear filter," in *Proc. IWAENC*. IEEE, 2016, pp. 1–5.

[19] M. M. Halimeh, C. Huemmer, and W. Kellermann, "A neural network-based nonlinear acoustic echo canceller," *IEEE Signal Processing Letters*, vol. 26, no. 12, pp. 1827–1831, 2019.

[20] R. Cutler, A. Saabas, T. Parnamaa, M. Loida, S. Sootla, H. Gamper *et al.*, "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*, Sep. 2021, submitted.

[21] *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.

[22] A. Dobrucki, "Nonlinear distortions in electroacoustic devices," *Archives of Acoustics*, vol. 36, no. 2, pp. 437–460, 2011.

[23] W. Klippel, "Loudspeaker nonlinearities–causes, parameters, symptoms," in *Audio Engineering Society Convention 119*. Audio Engineering Society, 2005.

[24] R. Ravaud, G. Lemarquand, T. Roussel, and V. Lemarquand, "Ranking of the nonlinearities of electrodynamic loudspeakers," *Archives of Acoustics*, vol. 35, no. 1, pp. 49–66, 2010.

[25] M. Soria-Rodríguez, M. Gabbouj, N. Zacharov, M. S. Hamalainen, and K. Koivuniemi, "Modeling and real-time auralization of electrodynamic loudspeaker non-linearities," in *Proc. ICASSP*, vol. 4. IEEE, 2004, pp. 81–84.

[26] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," *preprint arXiv:1412.3555*, 2014.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] A. Nicolae, "PLU: The piecewise linear unit activation function," *preprint arXiv:1809.09534*, 2018.

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

[30] A. Ivry, I. Cohen, and B. Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *Proc. ICASSP*, Jun. 2021.

[31] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. LDC93S1, 1993.

[32] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *Proc. ICASSP*, 2015, pp. 5206–5210.

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

[34] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito *et al.*, "Automatic differentiation in pytorch," in *Proc. Neural Information Processing Systems (NIPS)*, 2017.

[35] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

[37] *ITU-T Rec. P.862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, ITU-T Std., Feb. 2001.

[38] *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.

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