MULTIPLE-INPUT NEURAL NETWORK-BASED RESIDUAL ECHO SUPPRESSION

Guillaume Carbajal[†⋆] *Romain Serizel*[⋆] *Emmanuel Vincent* [⋆] *Éric Humbert* [†]

[†]Invoxia SAS, 2 Rue Maurice Hartmann, 92130 Issy-les-Moulineaux, France
[⋆] Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

ABSTRACT

A residual echo suppressor (RES) aims to suppress the residual echo in the output of an acoustic echo canceler (AEC). Spectral-based RES approaches typically estimate the magnitude spectra of the near-end speech and the residual echo from a single input, that is either the far-end speech or the echo computed by the AEC, and derive the RES filter coefficients accordingly. These single inputs do not always suffice to discriminate the near-end speech from the remaining echo. In this paper, we propose a neural network-based approach that directly estimates the RES filter coefficients from multiple inputs, including the AEC output, the far-end speech, and/or the echo computed by the AEC. We evaluate our system on real recordings of acoustic echo and near-end speech acquired in various situations with a smart speaker. We compare it to two single-input spectral-based approaches in terms of echo reduction and near-end speech distortion.

Index Terms— Acoustic echo cancellation, residual echo suppression, neural network, deep learning.

1. INTRODUCTION

In telecommunications, acoustic echo is a well-known issue. A speaker from a near-end point interacts with another speaker at a far-end point. Due to the acoustic coupling between the loudspeaker and the microphone at the near-end, the far-end speaker receives a delayed version of his/her own voice, known as the acoustic echo. Acoustic echo cancellation aims to eliminate this echo while leaving the near-end speech undistorted. The acoustic echo canceler (AEC) is a classic solution to this problem: the echo path is modeled as a long, linear filter and its estimation is then subtracted from the microphone signal [1]. However, a linear relationship between the acoustic echo and the far-end speech is assumed, which is not the case in real conditions [2]. Nonlinear responses of the loudspeaker and the microphone result in far-end speech distortion. In addition, the echo is often loud, especially when the microphone is close to the loudspeaker. Therefore, in real applications, the AEC reduces only partly the echo. Nonlinear AEC [3, 4] further reduces the echo but a residual echo always remains.

To overcome this limitation, a residual echo suppressor (RES) is typically employed. A RES is a short, nonlinear filter applied to the AEC output to estimate the near-end speech. It is akin to a noise suppression filter [5]. It often involves first estimating the spectral amplitude of the residual echo based on a single input signal, that is either the far-end speech or the echo computed by the AEC, and then deriving the coefficients of the RES [6–10]. This model-based approach can also be applied directly to the microphone signal: the filter is then called an acoustic echo suppressor (AES) [11–13]. While it significantly reduces the residual echo, this type of filter can distort the near-end speech, particularly in difficult situations such as double-talk or when the AEC has not converged.

Neural networks (NN) have recently become the state of the art in noise suppression [14–16], but they have rarely been used in acoustic echo cancellation. Schwarz et al. [17] used an NN to estimate the spectral amplitude of the residual echo using the spectral amplitude of the far-end as the single input, and derived the coefficients of the RES with a classic model-based approach. Madrid Portillo [18] used an NN to estimate directly the coefficients of an AES, using the spectral amplitude of two signals: the microphone and the far-end signal.

In this paper, we propose to extend these approaches to estimate the coefficients of the RES. Firstly we use multiple signals rather than a single signal as the NN inputs. Secondly we compute directly the coefficients of the RES through a mask, that we estimate with an NN trained according to the phase-sensitive cost in [16]. We refer to our approach as multiple-input phase-sensitive NN-based RES. It benefits from both the information of the different signals and the particular cost function. The validity of the proposed approach is experimentally verified with real echo recordings. We compare our results to a RES based on Valin's linear residual echo model [19][1] and to the NN-based RES in [17], including in double-talk situations and when the AEC has not converged.

This paper is organized as follows. We describe the classic AEC setting in Section 2 and review existing RES methods in Section 3. We present our approach in Section 4 and evaluate it in Section 5. We conclude in Section 6.

[1] Valin used this model to derive an AEC instead of a RES, but we found that it also works well as a RES in practice.

2. ACOUSTIC ECHO CANCELLATION

2.1. Signal model

For an arbitrary time-domain signal, $a(t)$ denotes this signal at time t. The signal is transformed in the time-frequency domain by a short-time Fourier transform (STFT) characterized by the window shape, the frame length L, the Fourier transform size N and the frame overlap O. $\underline{A}(m,n)$ is the resulting complex-valued spectrum at frame index m and frequency bin n. The spectral amplitude is denoted by $A(m,n)$ and the phase by $\theta_A(m,n)$. For the sake of conciseness, we will omit indexes t, m and n in the remaining of the paper.

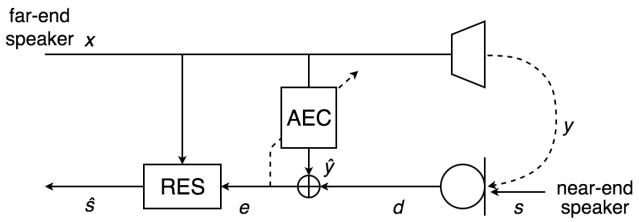


Fig. 1: General setting for an acoustic echo canceler (AEC) followed by a residual echo suppressor (RES).

The general problem setting is illustrated in Fig. 1. The microphone signal d is the sum of the near-end signal s and the acoustic echo y, which is a nonlinearly distorted version of the far-end signal x:

$$d = s + y. \tag{1}$$

2.2. AEC

The output signal of the AEC e consists of the near-end signal s and the residual echo z, that remains after subtracting the estimated echo $\widehat{y}$ from the microphone signal d:

$$e = s + z \tag{2}$$
$$= s + y - \widehat{y}. \tag{3}$$

Most AECs are based on time-domain adaptive filtering using various methods to track double-talk and echo path change. In particular, Valin [19] implemented SpeexDSP[2], an AEC using variable step-size for robust double-talk detection.

3. RESIDUAL ECHO SUPPRESSION

The RES applies a mask $\widehat{M}$ to $\underline{E}$ to produce $\underline{\widehat{S}} = \widehat{M}\underline{E}$. The coefficients of $\widehat{M}$, i.e., the RES filter coefficients, are real-valued and different in each time-frequency bin. The estimated near-end signal $\widehat{s}$ is then recovered by the inverse STFT. This signal contains the potentially distorted near-end signal s_{RES} and the post-residual echo z_{RES}:

$$\widehat{s} = s_{\text{RES}} + z_{\text{RES}}. \tag{4}$$

[2]https://github.com/xiph/speexdsp

3.1. Single-input vs. multiple-input methods

We categorize RES methods depending on the signal(s) used to estimate M. Single-input methods rely on a single signal, for example the magnitude spectrum of the far-end speech X [6–10] or the echo $\widehat{Y}$ estimated by the AEC [19, 20]. Schwarz et al. [17] proposed a single-input[3] RES using X and two linear combinations of X. However, they exploited neither the AEC output E nor $\widehat{Y}$. Yet, E contains the information of the residual echo and $\widehat{Y}$ approximates the long-term dependencies in the echo which are not included in X. Their results show that the residual echo is not always suppressed, especially when the AEC has not converged and in double-talk situations. Conversely, multiple-input methods rely on various signals to estimate M. Madrid Portillo [18] estimated an AES using the microphone signal D in addition to X, which improved performance compared to using D only but he did not benefit from using a long AEC filter as a prior step.

3.2. Spectral-based vs. mask-based methods

We can also categorize RES methods depending on the steps used to estimate M. On the one hand, spectral-based methods compute $\widehat{M}$ in two steps. The underlying idea is to subtract the residual echo estimate from $\underline{E}$. In the first step, they compute an estimate $\widehat{Z}$ of the magnitude spectrum of the residual echo Z [6–10, 19, 20]. In the second step, they derive $\widehat{M}$ from $\widehat{Z}$ according to a rule such as the Wiener filtering rule

$$\widehat{M} = \max\left(M_{\text{min}}, 1 - \mu\frac{\widehat{Z}^2}{E^2}\right) \tag{5}$$

with M_{min} the masking floor and μ the overestimation factor. Many methods estimate Z using the linear models $\widehat{Z} = \lambda X$ [6–10] or $\widehat{Z} = \lambda\widehat{Y}$ [19, 20], with λ a frequency- and time-dependent scalar, which do not account for the nonlinear distortions. Schwarz et al. [17] estimated Z using a (nonlinear) multilayer perceptron with two hidden layers instead. Yet, the rule (5) does not directly fit the target (ground truth) mask M. This may result in poor near-end speech transmission during double-talk or poor residual echo reduction. On the other hand, mask-based methods compute $\widehat{M}$ in one step by directly fitting the target mask M. Madrid Portillo [18] trained an AES using a multilayer perceptron with two hidden layers according to two alternative target masks: the ideal binary mask (IBM) or the ideal ratio mask (IRM) [14] (see Table 1).

4. MULTIPLE-INPUT NN-BASED RES

In this paper, we propose to estimate the RES filter coefficients M with a multiple-input NN. Specifically, we use E, X, and/or $\widehat{Y}$ as inputs and the phase-sensitive filter (PSF) [16] (see Table 1) as the target output. Contrary to single-input

[3]The authors use a different definition of single- vs. multiple-input.

and/or spectral-based methods, this enables us to benefit from the information of E, X, and/or $\widehat{Y}$ at the same time and to fit directly the target mask. To our knowledge, this is the first use of a multiple-input NN and the PSF in the context of RES. As we aim at comparing our RES to Schwarz' NN-based RES [17], we use a multilayer perceptron with two hidden layers. Figure 2 shows an example of the topology. We use the mean-square error (MSE) between the output mask $\widehat{M}$ and the target mask M as the training cost.

Ideal binary mask (IBM)	$M = \mathbb{1}_{S>Z}$
Ideal ratio mask (IRM)	$M = \frac{S}{\sqrt{S^2+Z^2}}$
Ideal amplitude mask (IAM)	$M = \frac{S}{E}$
Phase-sensitive filter (PSF)	$M = \frac{S}{E}\cos(\theta_S - \theta_E)$

Table 1: Example target masks.

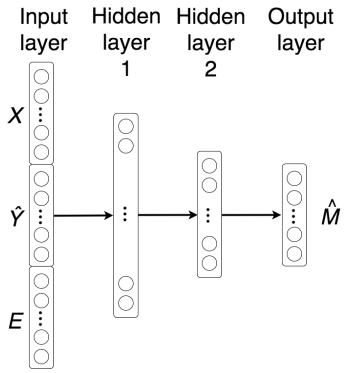


Fig. 2: Example multiple-input NN-based RES.

5. EXPERIMENTS

In the following, we assess the impact of E, X, and $\widehat{Y}$ on the performance of our method. We also evaluate the different target masks in Table 1, including the ideal amplitude mask (IAM), and neglect the IBM which performed more poorly. Finally we compare our method with two single-input spectral-based methods: a RES based on Valin's linear residual echo model [19] and Schwarz's NN-based RES [17]. We assess the performance in terms of echo reduction during single-talk and double-talk, and near-end distortion during double-talk. In addition, we analyze the performance before and after AEC convergence in single-talk and double-talk scenarios. We don't evaluate Madrid Portillo's AES as this paper focuses on RES only.

5.1. Dataset

We performed the experiments using the Librispeech clean development corpus [21], which consists of 5.4 h of audio books. We split this corpus into training, validation, and test sets with disjoint speakers. In each set, all speakers were used at least once as near-end and far-end speakers. At the near end, the distance between the speaker and the microphone was 1 m, whereas the distance between the loudspeaker (playing the far-end signal) and the microphone was 5 cm.

For the training and validation sets, the far-end speech was played and recorded at a rate of 16 kHz with Triby, a smart speaker device developed by Invoxia, and the near-end speech was played by a regular loudspeaker. The echo recordings were done without near-end speech. We measured the room impulse response (RIR) between the regular loudspeaker and Triby in the room, and convolved the utterances with this RIR to simulate the near-end speech. The recordings were done in a room of size $3 \times 3 \times 3$ m. The background noise level was about 50 dBA and the reverberation time about 0.2 s. The validation set was used to tune the NN hyperparameters.

For the test set, no simulation was conducted. The near-end and far-end speech were separately recorded in a different room of size $7 \times 7 \times 3$ m and 0.5 s reverberation time and summed together. The smart speaker was different than the one used in the train and validation sets. This close-to-real recording protocol makes it possible to measure performance accurately, which is not possible with real-world recordings for which the ground truth near-end speech is unknown. Our protocol is similar to [13], however more realistic as it includes the nonlinearities in the far-end speech.

We considered 9 scenarios defined by 3 near-end positions and 3 situations: near-end talk, far-end talk, and double-talk (simultaneous near-end and far-end speech). For the test set, only 2 situations were considered: far-end talk and double-talk. Given a scenario, each set consists of n 15 s utterances. For the train, validation and test sets, we took respectively $n = \{629, 205, 208\}$. During each utterance, the echo path was considered constant. However, it varied from one utterance to another. The near-end speech was set to a constant level and the far-end speech was played at three different levels to account for the loudspeaker nonlinearities, resulting in signal-to-echo ratios (SER) of -9, -12 and -15 dB.

5.2. Algorithm parameters

We processed the training, validation, and test data using the AEC implementation in SpeexDSP. We used a time-domain filter length of 2560 samples, implemented via a FFT size of 640 and a 50% overlap . This setting produced a good compromise between echo reduction and near-end distortions. From various observations, we set the AEC convergence time to 4 s after the begining of the utterance, which is consistent with Valin's observations. We implemented all RES methods using an STFT with a Hanning window, $L = 640$, $N = 1024$, and $O = 50\%$. Regarding the NN, we used 1024 neurons and $\tanh$ for the activation function at each hidden layer. With a 2.70 GHz CPU, computing e and $\widehat{s}$ for a 15 s utterance took respectively 2.3% and 1.7% of real time.

$\widehat{g} = \frac{\langle \mu, s \rangle}{\|s\|^2}\Big\|_{\text{double talk}}$	$\tilde{s} = \widehat{g} \cdot s \quad \tilde{y} = \widehat{g} \cdot y$
$\text{ERLE}_{\text{AEC or AEC+RES}}$	$10 \log_{10} \frac{\|\tilde{y}\|^2}{\|\nu\|^2}$
$\text{SDR}_{\text{AEC or AEC+RES}}$	$10 \log_{10} \frac{\|\tilde{s}\|^2}{\|\tilde{s}-\mu\|^2}$
$\text{SAR}_{\text{AEC+RES}}$	$10 \log_{10} \frac{\|\tilde{s}\|^2}{\|\tilde{s}-s_{\text{RES}}\|^2}$

Table 2: Evaluation metrics. For AEC only, $\nu = y - \widehat{y}$ and $\mu = e$. For AEC+RES, $\nu = z_{\text{RES}}$ and $\mu = \widehat{s}$.

		NN inputs			
	Double-talk	e	e, x	$e, \widehat{y}$	$e, x, \widehat{y}$
ERLE	Yes	10.8	19.3	16.5	20.3
ERLE	No	12.3	22.6	18.5	23.5
SDR	Yes	-2.7	3.6	1.0	4.1

Table 3: Average ERLE (dB) and SDR (dB) achieved by the proposed RES with various NN inputs.

5.3. Metrics

We evaluated all systems in terms of the echo return loss enhancement (ERLE), which measures the echo reduction, and the signal-to-distortion Ratio (SDR) [22], which measures the overall distortion (including both residual or post-residual echo and near-end speech distortion). During double-talk, these two metrics are essential. In addition, the signal-to-artifacts ratio (SAR) measures near-end speech distortion alone. The AEC itself induces little near-end speech distortion. Some RES may introduce an attenuation g on $\widehat{s}$ which results in both artificial increases of echo reduction and distortions with usual metrics. We needed metrics invariant to this attenuation. We assumed the attenuation g was constant over time. Based on the work of Vincent et al. [22], we estimated $\widehat{g}$ during double-talk for each utterance and applied it to s and y. The evaluation metrics are defined in Table 2.

5.4. Choice of NN inputs and target masks

We investigated the performance of E either as a single input or in combination with X and/or $\widehat{Y}$. Table 3 shows the results averaged over all choices of target masks. Using E and X together provides the best performance in terms of both ERLE and SDR. This performance is significantly larger than using E only, and comparable to using all three inputs. We also investigated the performance of the different target masks. Table 4 shows the results averaged along all choices of NN inputs. Using the PSF as a training target provides the best performance in terms of both ERLE and SDR.

5.5. Comparison to Valin's and Schwarz' RES

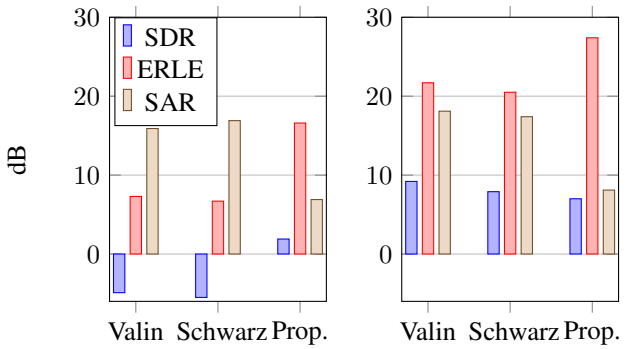
Eventually we obtained the best performance in ERLE using E, X and $\widehat{Y}$ as the inputs and the PSF as the target mask. Table 5 compares this setting to Valin's [19] and Schwarz' [17] RES. Our method significantly outperforms the others in terms of both ERLE and SDR. Figure 3 provides further analysis in a double-talk situation. After AEC convergence, all three methods achieve satisfactory results (ERLE above 20 dB and SAR above or close to 10 dB). Before AEC convergence, the performance of our method remains acceptable, while the ERLE achieved by the other methods drops below 10 dB and results in overwhelming post-residual echo compared to the near-end speech. This is confirmed by informal listening tests. Similar conclusions in terms of echo reduction can be drawn in a far-end talk situation. We conclude that our method is more robust to the lack of AEC convergence.

		Target mask		
	Double-talk	IRM	IAM	PSF
ERLE	Yes	14.8	16.7	17.8
ERLE	No	16.1	18.7	20.2
SDR	Yes	0.2	1.7	2.5

Table 4: Average ERLE (dB) and SDR (dB) achieved by the proposed RES with various target masks.

		AEC only	AEC+RES		
	Double-talk		Valin [19]	Schwarz [17]	Prop. RES
ERLE	Yes	10.6	12.5	11.8	21.2
ERLE	No	12.2	13.8	13.3	24.4
SDR	Yes	-1.1	0.4	-0.2	4.9

Table 5: Average ERLE (dB) and SDR (dB) achieved by the proposed RES compared to other RES and to AEC only.



(a) Before AEC convergence (b) After AEC convergence

Fig. 3: Detailed analysis during double-talk.

6. CONCLUSION AND FUTURE WORK

We proposed a multiple-input phase-sensitive NN-based RES, providing greater residual echo reduction than single-input spectral-based approaches. This reduction appears to be robust to different scenarios and to different rooms, as shown by training and testing in two different rooms. In the future, we will tune the tradeoff between echo reduction and near-end speech distortion using smoothing techniques [23, 24].

7. REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic Echo And Noise Control: A Practical Approach*, Adaptive and learning systems for signal processing, communications, and control. Wiley-Interscience, Hoboken, N.J, 2004.

[2] A. N. Birkett and R. A. Goubran, "Limitations of handsfree acoustic echo cancellers due to nonlinear loudspeaker distortion and enclosure vibration effects," in *Proc. WASPAA*, 1995, pp. 103–106.

[3] 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. DSP*, 2011, pp. 1–6.

[4] C. Hümmer, C. Hofmann, R. Maas, A. Schwarz, and W. Kellermann, "The elitist particle filter based on evolutionary strategies as novel approach for nonlinear acoustic echo cancellation," in *Proc. ICASSP*, 2014, pp. 1315–1319.

[5] S. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 27, no. 2, pp. 113–120, 1979.

[6] S. Gustafsson, R. Martin, and P. Vary, "Combined acoustic echo control and noise reduction for hands-free telephony — state of the art and perspectives," in *Proc. EUSIPCO*, 1996, pp. 1107–1110.

[7] C. Beaugeant, *Réduction de bruit et contrôle d'écho pour les applications radiomobiles*, Ph.D. thesis, Université de Rennes 1, 1999.

[8] S. Gustafsson, R. Martin, P. Jax, and P. Vary, "A psychoacoustic approach to combined acoustic echo cancellation and noise reduction," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 245–256, 2002.

[9] G. Enzner, R. Martin, and P. Vary, "Unbiased residual echo power estimation for hands-free telephony," in *Proc. ICASSP*, 2002, pp. 1893–1896.

[10] A. S. Chhetri, A. C. Surendran, J. W. Stokes, and J. C. Platt, "Regression-based residual acoustic echo suppression," in *Proc. IWAENC*, 2005.

[11] C. Avendano, "Acoustic echo suppression in the STFT domain," in *Proc. WASPAA*, 2001, pp. 175–178.

[12] C. Faller and J. Chen, "Suppressing acoustic echo in a spectral envelope space," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1048–1062, 2005.

[13] H. Huang, C. Hofmann, W. Kellermann, J. Chen, and J. Benesty, "A multiframe parametric Wiener filter for acoustic echo suppression," in *Proc. IWAENC*, 2016, pp. 1–5.

[14] Y. Wang, A. Narayanan, and D. L. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[15] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *Proc. GlobalSIP*, 2014, pp. 577–581.

[16] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015, pp. 708–712.

[17] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral feature-based nonlinear residual echo suppression," in *Proc. WASPAA*, 2013, pp. 1–4.

[18] J. Madrid Portillo, "Deep learning applied to acoustic echo cancellation," M.S. thesis, Aalborg University, 2017.

[19] J. M. Valin, "On adjusting the learning rate in frequency domain echo cancellation with double-talk," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1030–1034, 2007.

[20] O. Hoshuyama and A. Sugiyama, "An acoustic echo suppressor based on a frequency-domain model of highly nonlinear residual echo," in *Proc. ICASSP*, 2006, pp. 269–272.

[21] 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.

[22] 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.

[23] C. Breithaupt, T. Gerkmann, and R. Martin, "Cepstral smoothing of spectral filter gains for speech enhancement without musical noise," *IEEE Signal Processing Letters*, vol. 14, no. 12, pp. 1036–1039, 2007.

[24] E. Vincent, "An experimental evaluation of Wiener filter smoothing techniques applied to under-determined audio source separation," in *Proc. LVA/ICA*, 2010, pp. 157–164.