NN3A: NEURAL NETWORK SUPPORTED ACOUSTIC ECHO CANCELLATION, NOISE SUPPRESSION AND AUTOMATIC GAIN CONTROL FOR REAL-TIME COMMUNICATIONS

Ziteng Wang, Yueyue Na, Biao Tian, Qiang Fu

Alibaba Group, China

ABSTRACT

Acoustic echo cancellation (AEC), noise suppression (NS) and automatic gain control (AGC) are three often required modules for real-time communications (RTC). This paper proposes a neural network supported algorithm for RTC, namely NN3A, which incorporates an adaptive filter and a multi-task model for residual echo suppression, noise reduction and near-end speech activity detection. The proposed algorithm is shown to outperform both a method using separate models and an end-to-end alternative. It is further shown that there exists a trade-off in the model between residual suppression and near-end speech distortion, which could be balanced by a novel loss weighting function. Several practical aspects of training the joint model are also investigated to push its performance to limit.

Index Terms— echo cancellation, noise suppression, automatic gain control

1. INTRODUCTION

The demand for real-time communications (RTC) has grown rapidly in recent years. The three often required modules for RTC are acoustic echo cancellation (AEC), noise suppression (NS) and automatic gain control (AGC), as shown in Fig.1. AEC is designed to remove echo from the near-end microphone signal. It usually consists of a linear echo cancellation filter and a residual echo suppressor (RES). NS is subsequently to reduce the background noise, and AGC is to adjust the processed signal to a proper sound level. Many neural network based algorithms have been developed to tackle each problem separately, and casual and real-time ones are preferred as seen in the recent DNS-Challenges [1, 2] and AEC-Challenges [3, 4].

The deep learning based algorithms first made breakthrough in the NS task. Xu et al. [5] proposed to find a spectral mapping function between noisy and clean speech signals. Wang et al. [6] proposed to learn a mapping from noisy features to time-frequency masks of the target, such as the ideal binary mask, the phase-sensitive mask (PSM) [7], or the later introduced complex ratio mask (CRM) [8]. Advanced neural networks have been developed, such as the convolutional time-domain audio separation network (Conv-TasNet) [9] and the deep complex convolution recurrent network (DCCRN) [10] , and remarkable results are achieved.

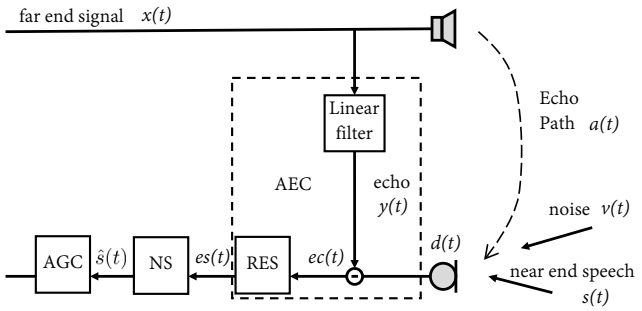

Fig. 1. Typical AEC/NS/AGC modules for real-time communications.

The neural network based AEC algorithms follow a similar paradigm by considering the echo signal as *undesired noise*. [11, 12] investigated end-to-end AEC relying on only neural networks. Nevertheless, a joint approach of linear filtering and neural network based RES has shown to be more effective [13, 14]. In the latter case, the linear filtered output, the microphone signal, the far-end signal, as well as the estimated echo, could all be taken as model inputs [15, 16, 17].

The need for an AGC is to amplify speech segments to an intelligible sound level, while not amplifying noise only segments [18, 19]. This can be properly handled once the near-end speech activity results are obtained, usually from a model based voice activity detector (VAD) [20, 21].

Though the three modules each involves a neural network model, they are separately designed and a joint model remains not investigated. The motivations are: a neural network model trained to suppress echo is likely to reduce noise at the same time, and multi-task training could benefit the overall performance. A joint model is also more compact from the engineering perspective. In this paper, we consider the three modules systematically and propose a NN3A algorithm, building upon our previous work [13] for echo cancellation. The NN3A algorithm retains a linear adaptive filter and adopts a multi-task model for joint residual echo suppression, noise reduction and near-end speech activity detection. It is shown that there exists a trade-off in the model between

residual suppression and near-end speech distortions. And an intuitive loss weighting function is introduced to balance the trade-off to meet the *zero* echo leakage requirement in typical RTC applications. The proposed algorithm outperforms both a method using separate models and an end-to-end neural network based alternative.

2. THE NN3A ALGORITHM

Consider a typical single microphone single loudspeaker setup, the microphone signal at time t is expressed as:

$$d(t) = x(t) * a(t) + s(t) + v(t) \tag{1}$$

where $x(t), s(t)$ and $v(t)$ are respectively the far-end signal, the near-end speech and the ambient noise. $a(t)$ is the echo path and $*$ denotes convolution. The task of AEC, NS and AGC is to recover a properly scaled version of the near-end speech $s(t)$, given the microphone signal and the far-end signal.

Without loss of generality, the algorithm is developed in the frequency domain. The frequency representations of the variables are denoted by their capitals, such as D, X, S.

2.1. Linear filter

A linear adaptive filter is first adopted to remove an echo estimation Y from the microphone signal:

$$E_{t,f} = D_{t,f} - \underbrace{\mathbf{w}_{L,f}^H \mathbf{x}_{L,f}}_{Y_{t,f}} \tag{2}$$

where $\mathbf{x}_{L,f} = [X_{t,f}, X_{t-1,f}, ..., X_{t-L+1,f}]^T$, f is the frequency index, $(\cdot)^T$ denotes transpose and $(\cdot)^H$ denotes Hermitian transpose. L is the filter tap and the filter coefficients are derived using the weighted recursive least square (wRLS) algorithm [13] as:

$$\begin{aligned}
\gamma_{t,f} &\leftarrow |E_{t,f}|^{\beta-2}, \\
\mathbf{R}_{L,f} &\leftarrow \sum_t \gamma_{t,f} \mathbf{x}_{L,f} \mathbf{x}_{L,f}^H, \\
\mathbf{r}_{L,f} &\leftarrow \sum_t \gamma_{t,f} \mathbf{x}_{L,f} D_{t,f}^*, \\
\mathbf{w}_{L,f} &\leftarrow \mathbf{R}_{L,f}^{-1} \mathbf{r}_{L,f}
\end{aligned} \tag{3}$$

where $\beta \in [0, 2]$ is a shape parameter related to the speech source prior, and $(\cdot)^*$ denotes complex conjugate.

2.2. Multi-task model

After linear filtering, a neural network based multi-task model is designed to further suppress the residual echo and remove the ambient noise:

$$\hat{S}_{t,f} = M_{t,f} E_{t,f} \tag{4}$$

where $M_{t,f}$ is the time-frequency mask inferred from the available signal set $\mathbf{f}_t = \{E_{t,f}, Y_{t,f}, D_{t,f}, X_{t,f}\}$. The inference process is expressed as:

$$\begin{aligned}
\mathbf{h}_t^0 &= \text{ReLU}(\text{Linear}(\mathbf{f}_t)) \\
\mathbf{h}_t^{j+1} &= \text{DFSMN}(\mathbf{h}_t^j), \quad j \in [0, 1, ..., J-1] \\
M_{t,-} &= \text{Sigmoid}(\text{Linear}(\mathbf{h}_t^{j+1})) \\
P_t &= \text{Sigmoid}(\text{Linear}(\mathbf{h}_t^{j+1}))
\end{aligned} \tag{5}$$

where $\mathbf{h}_t^j$ is output of the jth layer, P_t is the probability of near-end speech activity, and Linear$(\cdot)$ denotes an affine transformation layer with proper size of parameters. The deep feed-forward sequential memory network (DFSMN) [22] is chosen to model temporal dependency in time series, and one DFSMN layer is expressed as:

$$\begin{aligned}
\tilde{\mathbf{h}}_t^j &= \text{Linear}(\text{ReLU}(\text{Linear}(\mathbf{h}_t^{j-1}))), \\
\mathbf{h}_t^j &= \mathbf{h}_t^{j-1} + \tilde{\mathbf{h}}_t^j + \sum_{\tau=0}^{\mathcal{T}} \mathbf{m}_\tau^j \odot \tilde{\mathbf{h}}_{t-\tau}^j
\end{aligned} \tag{6}$$

where $\mathbf{m}_\tau$ is the time-invariant memory parameter weighting the history output $\tilde{\mathbf{h}}_{t-\tau}$ and $\odot$ denotes element-wise multiplication.

Given a predefined training target, such as the PSM

$$\bar{M}_{t,f} = \text{Real}(\frac{S_{t,f}}{E_{t,f}}), \tag{7}$$

the network model is trained to minimize both a mean squared error loss $\mathcal{L}_{mask}$ and a binary cross-entropy loss $\mathcal{L}_{vad}$, where

$$\begin{aligned}
\mathcal{L}_{mask} &= \sum_{t,f} \alpha_{t,f} |M_{t,f} - \bar{M}_{t,f}|^2, \\
\mathcal{L}_{vad} &= \sum_t -\bar{P}_t \log(P_t) - (1 - \bar{P}_t) \log(1 - P_t)
\end{aligned} \tag{8}$$

where $\bar{P}_t \in \{0, 1\}$ is the oracle near-end speech activity. It is empirically found that a model trained under the vanilla MSE loss cannot completely remove the residual echo, which fails to meet the usual human auditory requirement of zero echo leakage. Hence a weighting function is introduced as:

$$\alpha_{t,f} = \alpha - \bar{M}_{t,f}, \quad \alpha > 1 \tag{9}$$

The weighting function puts more weight on echo dominant time-frequency bins, where $\bar{M}_{t,f}$ is small, and less weight otherwise.

2.3. Post-processing

The sound level of the processed signal is subject to change in different applications. Therefore, the AGC algorithm is put in a post-processing stage as

$$\hat{s}(t) = g(P_t) \text{IFFT}(\hat{S}_{t,f}) \tag{10}$$

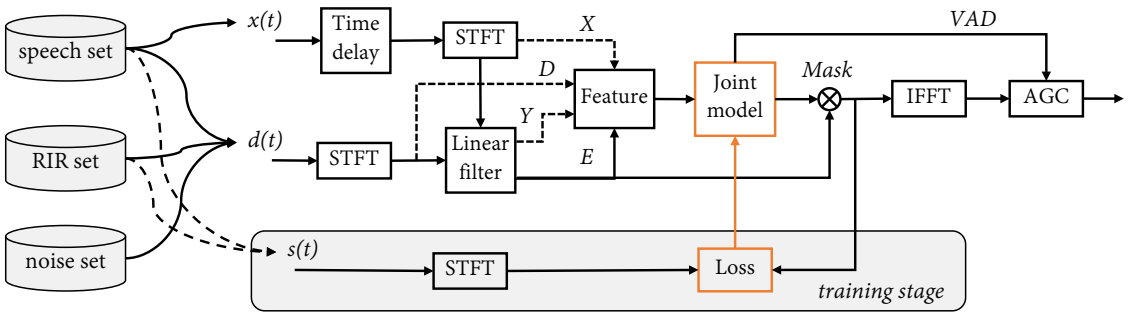

Fig. 2. Flowchart of the proposed NN3A algorithm. The gray blocks are only for the training stage.

and $g(\cdot)$ is a customized function, which consists of a peak level detector for computing gain and a gain controller for adjusting gain[19].

The flowchart of the proposed NN3A algorithm is presented in Fig 2. To handle possible delays between the microphone signal and the far-end signal, an additional time delay compensation module could be added before processing [13].

2.4. Related work

Combining speech activity detection and echo suppression has been discussed in [11]. Nevertheless, the study considered a pure model based approach and the experiments were limited to small-scale simulated data. An intuitive loss weighting method was proposed in [23] for spectral mapping based echo suppression, while a risk of sub-band nullification was introduced at the same time.

3. EXPERIMENTS

3.1. Setup

Experiments are conducted following the setup in the AEC-Challenge [3]. The training set covers near-end (NE) single talk (ST), far-end (FE) single talk, and double talk (DT) cases. Besides the data provided in the Challenge, a simulation pipeline is built to generate more training samples as in Fig 2. The speech set, the room impulse response (RIR) set and the noise set are all drawn from the DNS-Challenge [1]. The signal-to-echo ratio (SER) are uniformly sampled from [-10, 20] dB, and the signal-to-noise (SNR) ratio are sampled from [0, 40] dB. The simulated data are mixed as in (1), with 30% $x(t) = 0$, 20% $v(t) = 0$, and 10% $a(t) = 0$ implying a muted loudspeaker. Finally, approximate 1k hours of training data are used during training.

For STFT, the frame size is 20 ms and the hop size is 10ms. For the wRLS filter, the filter tap $L = 5$, and the source prior shape parameter $\beta = 0.2$. The multi-task model consists of $J = 12$ DFSMN layers, each layer with 512 nodes, and the temporal order $\mathcal{T} = 20$, leading to 6.7 M parameters. The number of operations per second is 0.335 GMACs.

Table 1. Step-wise evaluation of the proposed algorithm. + denotes cascade, and & denotes joint processing.

	NE ST PESQ	FE ST ERLE	DT PESQ
Orig	1.65	0	1.86
Linear	1.65	5.49	2.23
+RES	1.84	34.39	2.70
+RES+NS	2.49	38.28	2.72
+RES&NS	2.42	35.11	2.75
NN3A, $\alpha_{t,f} = 1$	2.47	37.25	**2.79**
NN3A, $\alpha = 1.1$	**2.57**	**45.13**	2.69

Table 2. Comparison of different multi-task model inputs $\mathbf{f}_t = \{E, Y, D, X\}$, as well as two end-to-end setups.

	NE ST PESQ	FE ST ERLE	DT PESQ
EX	2.57	45.13	2.69
EY	2.58	52.79	2.71
EYD	**2.61**	53.37	**2.75**
EYDX	2.59	**53.99**	2.74
DX	2.54	43.59	2.52
DTLN [12]	2.25	34.01	2.65

The evaluation set consists of three parts:

NE ST: 480 near-end speech utterances, contaminated by noise types of {car, subway, howling, keyboard, office, white} with SNR of {-5, 0, 5, 10} dB.

FE ST: 300 far-end single talk real recordings from the Challenge blind testset.

DT: the top 500 double talk utterances in the Challenge synthetic set.

The ITU-T recommendation P.862 perceptual evaluation of speech quality (PESQ) and echo return loss enhancement

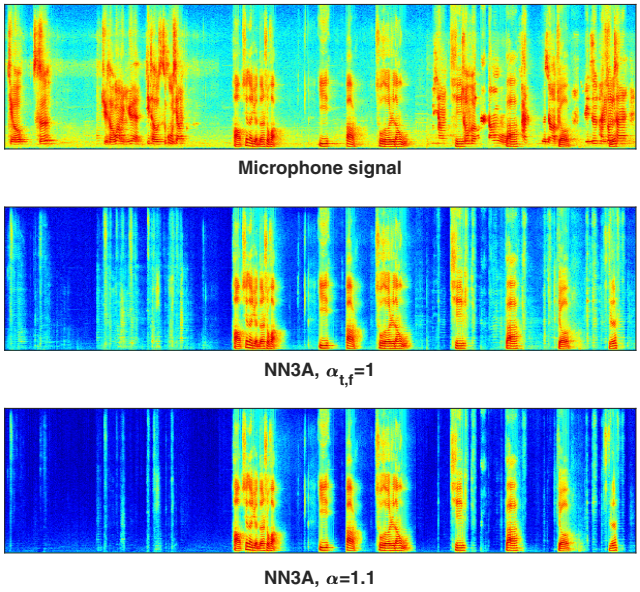

Fig. 3. Illustration of the effect of loss weighting. The sample audio covers FE ST, NE ST and DT consecutively.

(ERLE) scores are reported.

3.2. Results and Analysis

We first show step-wise evaluations of the algorithm in Table 1. A model that only performs residual echo suppression is denoted as RES, which also shows the ability to suppress noise, increasing the PESQ score by 0.19 for noisy NE ST. Stacking RES with a separately trained NS model, the RES+NS setup outperforms a joint RES&NS model in suppressing residuals. But cascading two models is likely to over-suppress the near-end speech, as shown by the lower DT PESQ score (2.72 compared with 2.75). The proposed vanilla NN3A algorithm (with $\alpha_{t,f} = 1$), which combines residual suppression with near-end speech activity detection, improves the results overall.

To achieve minimum echo leakage, $\alpha = 1.1$ is set in (9). As expected, the residual suppression performance is largely improved with a ERLE score of 45.13 dB compared with the 37.25 dB reference, at the cost of more speech distortion. An audio sample is given in Fig. 3 to illustrate the effect of loss weighting. The residual echo in the last signal is finally reduced to below audible level.

In Table 2, possible algorithm configurations are investigated to improve the performance. It is observed that a combination of the linear filtered output E, the estimated echo Y and the microphone signal D, leads to much higher scores than our previous EX setup [13]. And using the estimated echo as input is a better choice than using the far-end signal. The findings are consistent with a recent preprint [17].

The NN3A algorithm is also compared with a setup that

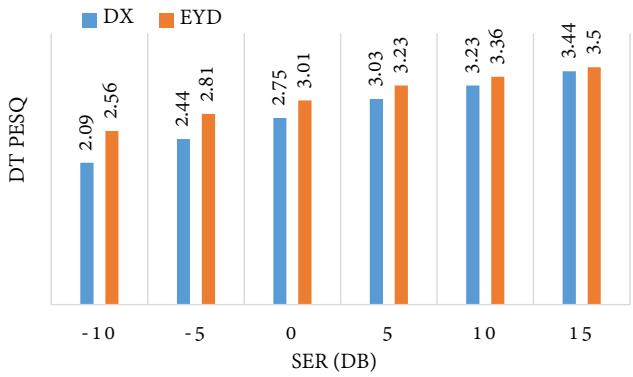

Fig. 4. Comparison of an end-to-end setup (DX) and the NN3A algorithm (EYD) in different SERs.

discards the linear filter, denoted as DX, and a publicly available DTLN (10.4 M) model [12]. There is a clear advantage of the joint signal processing and deep learning approach over the pure model based ones. In Fig. 3, the DX setup is compared with the NN3A-EYD setup in different SERs. The results verifies the benefit of linear filtering, and the gap is larger especially in low SER scenarios.

4. CONCLUSION

This paper presents a NN3A algorithm for real-time communications. The algorithm retains a linear filter and introduces a multi-task model for joint residual echo suppression, noise reduction and near-end speech activity detection. The multi-task model is shown to outperform both cascaded models and end-to-end alternatives. A novel loss weighting function is introduced to balance tradeoff between residual suppression and speech distortion, and minimum echo leakage could be achieved by tuning a weighting factor.

5. REFERENCES

[1] Chandan KA Reddy, Harishchandra Dubey, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 6623–6627.

[2] Chandan KA Reddy, Harishchandra Dubey, Kazuhito Koishida, Arun Nair, Vishak Gopal, Ross Cutler, Sebastian Braun, Hannes Gamper, Robert Aichner, and Sriram Srinivasan, "Interspeech 2021 deep noise suppression challenge," *Proc. Interspeech*, 2021.

[3] Kusha Sridhar, Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Hannes Gamper, Sebastian Braun, Robert Aichner, and Sriram Srinivasan, "ICASSP 2021

acoustic echo cancellation challenge: Datasets, testing framework, and results," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 151–155.

[4] Ross Cutler, Ando Saabas, Tanel Parnamaa, Markus Loide, Sten Sootla, Marju Purin, Hannes Gamper, Sebastian Braun, Karsten Sorensen, Robert Aichner, et al., "Interspeech 2021 acoustic echo cancellation challenge," in *Proc. Interspeech*, 2021.

[5] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal processing letters*, vol. 21, no. 1, pp. 65–68, 2013.

[6] Yuxuan Wang, Arun Narayanan, and DeLiang 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.

[7] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 708–712.

[8] Donald S Williamson, Yuxuan Wang, and DeLiang Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 24, no. 3, pp. 483–492, 2015.

[9] Yi Luo and Nima Mesgarani, "Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[10] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," *Proc. Interspeech 2020*, pp. 2472–2476, 2020.

[11] Hao Zhang, Ke Tan, and DeLiang Wang, "Deep learning for joint acoustic echo and noise cancellation with nonlinear distortions.," in *INTERSPEECH*, 2019, pp. 4255–4259.

[12] Nils L Westhausen and Bernd T Meyer, "Acoustic echo cancellation with the dual-signal transformation lstm network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7138–7142.

[13] Ziteng Wang, Yueyue Na, Zhang Liu, Biao Tian, and Qiang Fu, "Weighted recursive least square filter and neural network based residual echo suppression for the aec-challenge," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 141–145.

[14] Jean-Marc Valin, Srikanth Tenneti, Karim Helwani, Umut Isik, and Arvindh Krishnaswamy, "Low-complexity, real-time joint neural echo control and speech enhancement based on percepnet," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 7133–7137.

[15] Guillaume Carbajal, Romain Serizel, Emmanuel Vincent, and Eric Humbert, "Multiple-input neural network-based residual echo suppression," in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 231–235.

[16] Amin Fazel, Mostafa El-Khamy, and Jungwon Lee, "Cad-aec: Context-aware deep acoustic echo cancellation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6919–6923.

[17] Jan Franzen and Tim Fingscheidt, "Deep residual echo suppression and noise reduction: A multi-input fcrn approach in a hybrid speech enhancement system," *arXiv preprint arXiv:2108.03051*, 2021.

[18] Peter L Chu, "Voice-activated agc for teleconferencing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*. IEEE, 1996, vol. 2, pp. 929–932.

[19] Fitzgerald J Archibald, "Software implementation of automatic gain controller for speech signal," *Texas Instruments SPRAAL1 White Paper*, 2008.

[20] Thad Hughes and Keir Mierle, "Recurrent neural networks for voice activity detection," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7378–7382.

[21] Sebastian Braun and Ivan Tashev, "On training targets for noise-robust voice activity detection," *arXiv preprint arXiv:2102.07445*, 2021.

[22] Shiliang Zhang, Ming Lei, Zhijie Yan, and Lirong Dai, "Deep-fsmn for large vocabulary continuous speech recognition," in *ICASSP*. IEEE, 2018, pp. 5869–5873.

[23] Amir Ivry, Israel Cohen, and Baruch Berdugo, "Deep residual echo suppression with a tunable tradeoff between signal distortion and echo suppression," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 126–130.