Y^2-Net FCRN for Acoustic Echo and Noise Suppression

Ernst Seidel, Jan Franzen, Maximilian Strake, Tim Fingscheidt

Institute for Communications Technology, Technische Universität Braunschweig,
Schleinitzstr. 22, 38106 Braunschweig, Germany

{e.seidel, j.franzen, m.strake, t.fingscheidt}@tu-bs.de

Abstract

In recent years, deep neural networks (DNNs) were studied as an alternative to traditional acoustic echo cancellation (AEC) algorithms. The proposed models achieved remarkable performance for the separate tasks of AEC and residual echo suppression (RES). A promising network topology is a fully convolutional recurrent network (FCRN) structure, which has already proven its performance on both noise suppression and AEC tasks, individually. However, the combination of AEC, postfiltering, and noise suppression to a single network typically leads to a noticeable decline in the quality of the near-end speech component due to the lack of a separate loss for echo estimation. In this paper, we propose a two-stage model (Y^2-Net) which consists of two FCRNs, each with two inputs and one output (Y-Net). The first stage (AEC) yields an echo estimate, which—as a novelty for a DNN AEC model—is further used by the second stage to perform RES and noise suppression. While the subjective listening test of the Interspeech 2021 AEC Challenge mostly yielded results close to the baseline, the proposed method scored an average improvement of 0.46 points over the baseline on the blind testset in double-talk on the instrumental metric DECMOS, provided by the challenge organizers.

Index Terms: acoustic echo cancellation, echo suppression, convolutional LSTM, convolutional neural network

1. Introduction

Acoustic echoes arise if a speech communication system's microphone picks up again the echo that was just played by the system's loudspeaker in so-called handsfree mode. If not suppressed, the far-end (FE) speaker is forced to listen to his or her own echo, which considerably lowers the quality of the conversation. Accordingly, acoustic echo cancellation (AEC) is a widely researched topic.

The traditional approach to AEC is based on the application of adaptive filters [1, 2, 3] to estimate the impulse response (IR) of the loudspeaker-enclosure-microphone (LEM) system. Based on the IR, an echo estimate can be computed and subtracted from the microphone signal. While the normalized least mean squared (NLMS) algorithm [4] and Kalman filters [5, 6, 7] are likely the most prominent solutions, various adaptive filter approaches were continuously improved. Typically a subsequent postfilter is used to suppress residual echo further [8, 9].

The use of deep neural networks (DNNs) for AEC was proposed just recently, first in the form of residual echo suppression (RES) postfilters (PFs) [10, 11]. Newer approaches combine AEC with RES and noise suppression in a single DNN [12, 13], treating the echo suppression task as a source separation problem. However, while the reported echo suppression was impressive, it was accompanied by a noticeable decline in the quality of the resulting near-end (NE) speech component. This observation was confirmed in our previous work [14], in which it was

shown that it is the better choice to have a separate AEC DNN to estimate the echo.

Based on these findings we propose a two-stage model in which the first Y-Net (AEC DNN) receives the loudspeaker and microphone signals, estimating an echo, while the second Y-Net (PF DNN) receives the microphone signal after echo subtraction and, as a further novelty, the echo estimate output of the first Y-Net, instead of the loudspeaker signal. As the fully convolutional recurrent network (FCRN) structure has proven its strong performance in coded speech enhancement [15], in noise suppression [16, 17], including a second-ranked Interspeech 2020 Deep Noise Suppression Challenge proposal [18], and in AEC [14], we adapt this topology for our Y-Net.

The remainder of this paper is structured as follows: In Section 2, an overview of the two-stage network structure and its individual stages is given. The training and experimental setup is described in Section 3. In Section 4, the experimental results of our model are discussed and compared to the AEC Challenge's baseline approach. Section 5 provides conclusions.

2. System Overview and Proposed Model

The speech enhancement in this paper is completely performed by DNNs. The proposed model is based on the findings of [17] and [14], in which it was already shown that FCRN models perform well separately on AEC and noise suppression. However, it was also found that training a single network for the combined task of AEC and noise suppression leads to significant NE speech distortions. As a result, we propose a two-stage model (Y^2-Net) in which the tasks of AEC and noise suppression are performed by consecutive network stages (Y-Nets).

The proposed Y^2-Net model structure for our training setup is depicted in Figure 1. The input signals are the loudspeaker reference signal $x(n)$ and the microphone signal $y(n) = s(n) + d(n) + n(n)$ with NE speech $s(n)$, noise $n(n)$, and echo $d(n)$, and n being the sample index. For synthetic datasets, $d(n)$ can be computed by applying an LEM model to the reference signal $x(n)$, simulating the effects of loudspeaker distortions and near-end room characteristics.

Our Y^2-Net operates entirely in the DFT domain at a sampling rate of 16 kHz. After applying a 1st order high-pass on both $x(n)$ and $y(n)$ to eliminate time-varying biases [19], the signals are divided into frames of $N_T = 424$ samples with a frame shift of 50% and subject to a square root Hann windowing. The frames are zero-padded so that a $K = 512$-point DFT can be applied. The resulting frequency domain representations $X_\ell(k)$ and $Y_\ell(k)$ with frame index ℓ and frequency bin index $k \in \mathcal{K} = \{0, 1, ..., K{-}1\}$ are fed into the first stage of the model (Y-Net AEC). This stage is trained to provide an echo estimate $\hat{D}_\ell(k)$. Unlike other DNN AEC systems (e.g., [12, 13]), but just as classical hands-free systems and also our recent AEC DNN work [14], $\hat{D}_\ell(k)$ is then subtracted from $Y_\ell(k)$ to obtain
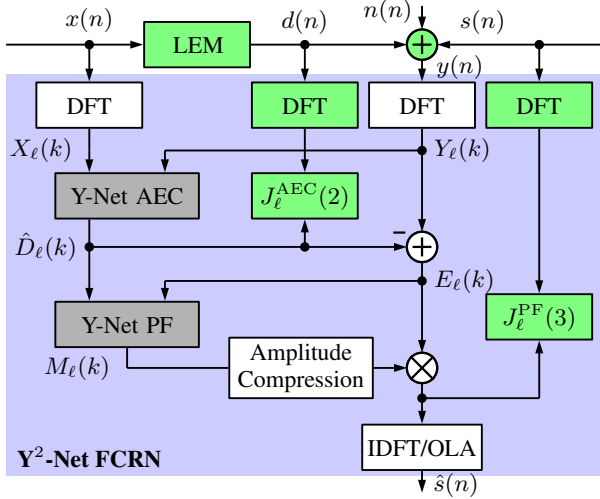
Figure 1: *Proposed Y^2-Net model (light blue), comprising AEC and PF, embedded into the synthetic training setup (green parts). Details of the Y-Net structure are displayed in Fig. 2.*
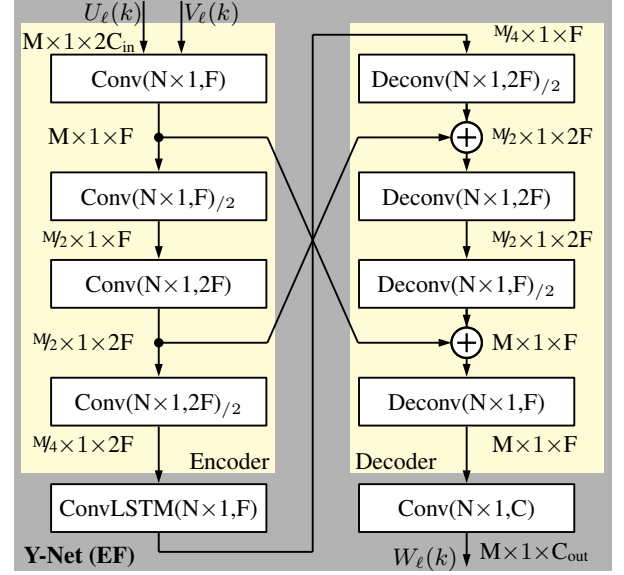

Figure 2: *Proposed Y-Net FCRN structure (here with EF) used for each stage of the proposed Y^2-Net model (see Figure 1) with DFT-domain inputs $U_\ell(k)$ and $V_\ell(k)$, and output $W_\ell(k)$.*

the intermediate enhanced signal $E_\ell(k)$. Subsequently, the second model stage (Y-Net PF) performs RES and noise suppression by computing a complex mask $M_\ell(k)$ based on the inputs $E_\ell(k)$ and $\hat{D}_\ell(k)$—a concept so far only known for RES after traditional AEC filters [11]. We follow the second-ranked non-real-time approach of the Interspeech 2020 Deep Noise Suppression Challenge (with a real-time capable model) [18] by employing mask amplitude compression and multiplying its result with $E_\ell(k)$ according to

$$\hat{S}_\ell(k) = E_\ell(k) \cdot \tanh(|M_\ell(k)|) \cdot \frac{M_\ell(k)}{|M_\ell(k)|}, \quad (1)$$

and transform $\hat{S}_\ell(k)$ back into the time-domain by applying a 512-point IDFT. The resulting frames are cut back to N_T samples before applying another square root Hann window and reconstructing the final estimated output signal $\hat{s}(n)$ by overlap-add (OLA). The challenge rules require an algorithmic latency defined as $T_1 = T + T_s \le 40\,\text{ms}$. Our model yields a frame length of $T = 26.5\,\text{ms}$ and a frame shift of $T_s = 13.25\,\text{ms}$, resulting in $T_1 = 39.75\,\text{ms} \le 40\,\text{ms}$.

The topology of the Y-Net FCRN is depicted in Figure 2. The feature dimensions of each layer are denoted as *feature axis × time axis × number of feature maps*. The kernel size is $N \times 1$ with $N = 24$ in feature direction and 1 in time direction. Determining the number of filter kernels throughout the network layers, we choose $F = 70$.

The Y-Net FCRN uses an encoder-decoder structure with two generic inputs $U_\ell(k)$ and $V_\ell(k)$ and one output $W_\ell(k)$ (hence the name Y-Net). The encoder features four convolutional layers. Every other layer is configured with a stride of 2 in the feature axis (marked as $_{/2}$), thereby halving the feature axis each time. The first two layers use F kernels, while the last two layers use $2F$ kernels. The decoder is designed inversely to the encoder, using strides on deconvolutional layers [20] to restore the original feature axis length. All convolutional and deconvolutional layers use a leaky ReLU activation function [21]. The bottleneck between encoder and decoder consists of a convolutional LSTM [22, 17] layer with F kernels, a tanh() activation function, and the "hard sigmoid" recurrent activation function as defined by `TensorFlow2` [23]. Two skip connections

are employed. The convolutional output layer after the decoder block yields $W_\ell(k)$ and uses C kernels and a linear activation.

Although the specific input and output features for each stage are selected individually, both stages' inputs and outputs consist of spectral frequency bins. These bins are divided into two channels each ($C = 2$), containing their real and imaginary components. By concatenation along the channel axis, $2C = 4$ real-valued input channels are obtained. To preserve the feature axis size M, it must be chosen to be dividable by 4. With a 512-point DFT yielding 257 non-redundant spectral frequency bins, the frequency dimension is set to $M = 260$ and channels are zero-padded accordingly.

Note that we investigate the Y-Net AEC either having identical early fusion (EF) as shown in Figure 2, dubbing the model as Y^2-Net EF/EF, or an AEC variant called late fusion (LF), where separate encoder paths are used for the inputs $U_\ell(k) = X_\ell(k)$ and $V_\ell(k) = Y_\ell(k)$. Each encoder path of LF models has the same layers and parameters as the single path of EF, and both paths are concatenated just before the ConvLSTM block. The skip connections then originate from the encoder of $Y_\ell(k)$. The corresponding entire model with EF-type PF (Fig. 2) is labeled Y^2-Net LF/EF. The effects of EF and LF are investigated in more detail in [14]. As a Y-Net with LF has considerably more parameters to train than with EF, we choose $F = 60$ for the Y-Net AEC of Y^2-Net LF/EF.

3. Experimental Setup

3.1. Dataset

The model is trained with synthetic files from the database provided by Microsoft for the Interspeech 2021 AEC Challenge [24], in the following labeled as $\mathcal{D}_{\text{syn}}$. The audio files were created with varying conditions, including single-talk, double-talk, both NE and FE noise, as well as simulated nonlinear loudspeaker distortions. The dataset consists of 10,000 audio samples of 10 s length each. All speaker audio was taken from the `LibriVox` project [25] by picking a 10 s audio sam-

ple of a randomly selected speaker for the FE signal. The NE signal is then chosen from a different speaker and contains 3 s-7 s of audio padded to the length of the FE signal. The IRs for the FE echoes are generated using Microsoft's `Project Triton` [26] technology with reverberation times T_{60} between 0.2 s and 1.2 s. For 80% of the files the FE signal is also subject to one of various nonlinear functions before applying the IR to simulate loudspeaker distortions. A more in-depth description can be found in [27]. A number of 500 files of $\mathcal{D}_{\text{syn}}$ were created with unseen speakers and IRs, and are used as preliminary synthetic testset $\mathcal{D}_{\text{syn}}^{\text{test,pre}}$ for instrumental measurements. The remaining files are split into 8,000 files ($\mathcal{D}_{\text{syn}}^{\text{train}}$) for training and 1,500 for validation between epochs ($\mathcal{D}_{\text{syn}}^{\text{val}}$).

For each audio sample, the dataset contains the noisy microphone signal $y(n)$ and the reference signal $x(n)$, as well as the speech components for NE speech $s(n)$ and echo $d(n)$ at the microphone. The NE noise $n(n)$ in the provided data is only present during the duration of NE speech, which we find as not representative for real scenarios. Therefore, we augmented the dataset with custom noise by adding an additional NE noise component randomly taken from the database provided by Microsoft for the Interspeech 2021 Deep Noise Suppression Challenge [28] with an SNR between 0 dB and 40 dB. As NE speech and FE echo were not altered, the SER of audio files in $\mathcal{D}_{\text{syn}}$ lies between -10 dB and 10 dB as originally designed by the dataset creators. The FE noise also remains unchanged.

Furthermore, the performance of the network is evaluated on two real test datasets, both being provided by the challenge organizers. The preliminary real testset $\mathcal{D}_{\text{real}}^{\text{test,pre}}$ and the final blind test set $\mathcal{D}_{\text{real}}^{\text{test}}$ each consist of 800 real-world recordings comprising double-talk, FE single-talk, or NE single-talk.

3.2. Training Targets and Loss Computation

The stages of the model are trained with different targets according to the assigned task. The Y-Net AEC computes an echo estimate $\hat{D}_\ell(k)$, which is then subtracted from the microphone signal to yield an enhanced intermediate signal $E_\ell(k)$. The Y-Net PF, in turn, provides a complex mask $M_\ell(k)$, which is applied to $E_\ell(k)$ according to (1), yielding the final estimated enhanced signal $\hat{S}_\ell(k)$. The training is conducted using a standard MSE loss, which is computed for the individual network stages as

$$J_\ell^{\text{AEC}} = \frac{1}{K} \sum_{k \in \mathcal{K}} \left| \hat{D}_\ell(k) - D_\ell(k) \right|^2 \tag{2}$$

$$J_\ell^{\text{PF}} = \frac{1}{K} \sum_{k \in \mathcal{K}} \left| \hat{S}_\ell(k) - S_\ell(k) \right|^2. \tag{3}$$

The network is trained in two steps: First, the Y-Net AEC is pretrained separately using its individual loss (2) on $\mathcal{D}_{\text{syn}}^{\text{train}}$. Then, both Y-Net AEC and Y-Net PF are jointly trained on $\mathcal{D}_{\text{syn}}^{\text{train}}$ with the pretrained weights for the Y-Net AEC and a weighted loss function

$$J_\ell = \alpha J_\ell^{\text{AEC}} + (1 - \alpha) J_\ell^{\text{PF}} \tag{4}$$

with $\alpha = 0.25$. This combined loss for both stages ensures that the AEC performance of the first stage is tuned to yield a residual echo characteristic which can be efficiently suppressed by the Y-Net PF. *Note that (4) is a fundamental difference of our work to earlier source separation methods for AEC [12, 13], as we strongly suggest an intermediate echo estimate $\hat{D}_\ell(k)$ being used in losses (2) and (4).*

3.3. Training Parameters

During all stages of the training, the Adam optimizer [29] is used in its standard parameter configuration. The batch size is set to 16 and the backpropagation-through-time [30] (BPTT) unrolling sequence length per batch is set to 50 frames. The initial learning rate is set to $5 \cdot 10^{-3}$. For every 3 epochs without loss improvement on $\mathcal{D}_{\text{syn}}^{\text{val}}$, the learning rate is multiplied with a factor of 0.6. The training is stopped after 100 epochs, if the validation loss does not improve for 10 consecutive epochs, or if the learning rate drops below $5 \cdot 10^{-4}$.

4. Results and Discussion

We use three metrics to measure the performance on our preliminary synthetic testset $\mathcal{D}_{\text{syn}}^{\text{test,pre}}$. Some of these metrics require individual enhanced signal components for NE speech $\tilde{s}(n)$, echo $\tilde{d}(n)$, and noise $\tilde{n}(n)$, which are derived from the enhanced signal $\hat{s}(n) = \tilde{s}(n) + \tilde{d}(n) + \tilde{n}(n)$ and from $s(n)$, $d(n)$, and $n(n)$, using the black-box signal separation approach according to ITU-T Recommendation P.1110 [31], with more details in [32, 33, 34]. All metrics using components obtained by this approach are marked with an index BB.

The quality of the NE speech component is evaluated on the wideband PESQ MOS LQO [35, 36]. The echo reduction is measured by the echo return loss enhancement $\text{ERLE}(n) = 10 \cdot \log \left(d^2(n)/\tilde{d}^2(n) \right)$ [dB]. After applying a first order IIR smoothing filter with factor 0.99, the final ERLE value is computed as an average over the entire dataset. The noise reduction performance of a model is measured as the SNR difference between the signals $y(n)$ and $\hat{s}(n)$ in [dB] and is defined as $\Delta\text{SNR} = 10 \cdot \log_{10} \left(\sum_n \tilde{s}^2(n)/ \sum_n \tilde{n}^2(n) \right) - 10 \cdot \log_{10} \left(\sum_n s^2(n)/ \sum_n n^2(n) \right)$ with sample index n spanning the entire utterance. To give a clear overview of the models' performance, the tests are conducted under several conditions. The four center columns of Table 1 marked as *full mixture* provide results for the preliminary testset $\mathcal{D}_{\text{syn}}^{\text{test,pre}}$. PESQ scores are evaluated on the full processed audio file $\hat{s}(n)$ but also on $\tilde{s}(n)$, the latter labeled as PESQ_{BB}. The three metrics in the columns to the right of the full mixture columns represent special conditions under which only the specified signal component is used as microphone signal $y(n)$. These conditions observe the models' performance beyond double-talk scenarios: How well does the model handle echo or noise if no NE speech is present? And can it pass through clean NE speech without distortions? This metric setup in turn also requires to set $x(n) = 0$ (unless $y(n) = d(n)$) as no echo is present in this case. For these metrics the respective enhanced signal component is equal to the entire enhanced signal, so no black-box approach is needed to compute them. For ΔSNR under the condition of $y(n) = n(n)$, ΔSNR is simplified to $\Delta\text{SNR} = 10 \log_{10} \left(\sum_n n^2(n)/ \sum_n \tilde{n}^2(n) \right)$.

The challenge organizers conducted a subjective evaluation of our models on the final test dataset $\mathcal{D}_{\text{real}}^{\text{test}}$ based on the ITU-T P.808 framework [37]. Four scenarios were evaluated: FE single-talk "echo DMOS" test (P.831 [38]), NE single-talk "MOS" test (P.808), double-talk "echo DMOS" test (P.831), and double-talk "other degradation DMOS" test (P.831). For double-talk, the "echo DMOS" test will be labeled as "FE" in Tables 2 and 3, describing the echo annoyance. The "other degradation DMOS" test is dubbed as "NE", describing quality of NE speech. More details can be found in [27].

The challenge organizers also provided an instrumental measurement algorithm DECMOS, which yields MOS values

Table 1: *Instrumental ratings on the preliminary synthetic testset $\mathcal{D}_{\text{syn}}^{\text{test,pre}}$, and real-time factor (RTF). Y-Net AEC does not perform noise suppression and is therefore separated from the other models. Best results per table segment are marked in bold, second-best results are underlined. Our model proposed for the Interspeech 2021 AEC Challenge is marked in bold.*

| Input (microphone) signals: | full mixture y(n) | | | | $d(n)$ | $n(n)$ | $s(n)$ | |
Method: / Metric:	PESQ	ERLE$_{\text{BB}}$	ΔSNR$_{\text{BB}}$	PESQ$_{\text{BB}}$	ERLE	ΔSNR	PESQ	RTF
Y-Net AEC (EF), pretraining	2.14	**15.86**	2.77	**3.74**	**26.36**	**0.86**	**4.51**	0.42
Y-Net AEC (EF), joint training	**2.16**	15.10	**2.78**	3.73	24.20	0.85	4.48	0.42
Baseline	1.95	27.57	5.31	**3.31**	45.34	1.11	**4.61**	0.05
Y^2-Net (EF/EF) (E, X)	<u>2.46</u>	**32.17**	10.31	3.09	52.19	26.04	<u>4.44</u>	0.95
Y^2-Net (EF/EF) (E, $\hat{\text{D}}$)	2.40	31.55	10.28	3.12	51.98	**29.27**	4.35	0.95
Y^2-Net (LF/EF) (E, X)	<u>2.46</u>	31.72	**10.69**	<u>3.21</u>	**52.45**	28.58	4.31	0.96
Y^2-Net (LF/EF) (E, $\hat{\text{D}}$)	**2.47**	<u>32.13</u>	10.60	3.12	<u>52.35</u>	<u>29.10</u>	4.30	0.96
Y-Net (EF), loss (3) only (as [13])	2.40	27.50	<u>10.62</u>	3.11	38.77	28.83	4.09	0.97

Table 2: *Instrumental DECMOS ratings on the preliminary real testset $\mathcal{D}_{\text{real}}^{\text{test,pre}}$.*

| | single-talk | | double-talk | | |
Method	FE	NE	FE	NE	total
Baseline	3.63	**3.64**	3.01	2.78	3.27
Y^2-Net (EF/EF) (E, X)	3.65	3.55	3.59	3.13	3.48
Y^2-Net (EF/EF) (E, $\hat{\text{D}}$)	3.52	3.58	3.54	3.19	3.46
Y^2-Net (LF/EF) (E, X)	**3.66**	3.63	3.52	3.14	3.49
Y^2-Net (LF/EF) (E, $\hat{\text{D}}$)	**3.66**	3.63	**3.61**	**3.21**	**3.53**

Table 3: *Subjective ratings according to ITU-T P.808 and objective ratings (DECMOS) on the (blind) real testset $\mathcal{D}_{\text{real}}^{\text{test}}$.*

| Metric | Method | single-talk | | double-talk | |
		FE	NE	FE	NE
subj	Baseline	**3.82**	**4.18**	**4.04**	3.45
subj	Y^2-Net (LF/EF) (E, $\hat{\text{D}}$)	3.73	4.16	3.72	**3.53**
obj	Baseline	3.48	3.72	3.02	2.78
obj	Y^2-Net (LF/EF) (E, $\hat{\text{D}}$)	**3.49**	**3.76**	**3.55**	**3.18**

that are claimed to be highly correlated to those of the described subjective listening tests. This is important for the evaluation of our proposed models, as our other instrumental metrics cannot be applied to real datasets because the separate signal components $s(n)$, $d(n)$ and $n(n)$ are not available.

Table 1 shows the results of our proposed model on the preliminary synthetic testset $\mathcal{D}_{\text{syn}}^{\text{test,pre}}$. In the first segment of the table, we report our AEC Y-Net (EF) after pre-training, and after joint training, confirming that joint training hardly compromises echo reduction performance at the output of the first Y-Net AEC. In the second table segment, our Y^2-Net models are tested with two variants of Y-Net PF inputs: the common approach of using intermediate output and reference signal (E, X), and the new approach using intermediate output and estimated echo (E, $\hat{\text{D}}$) as depicted in Figure 1. It can be seen that our Y^2-Net model outperforms the challenge baseline significantly in terms of ERLE, SNR and full mixture PESQ. The baseline can score better values for PESQ$_{\text{BB}}$ due to not performing noise reduction. As the LF/EF Y^2-Net with (E, $\hat{\text{D}}$) shows four 1$^{\text{st}}$ and 2$^{\text{nd}}$ ranks among the Y^2-Net variants, and $s(n)$-only PESQ metrics of ≥ 4.30 points hardly yield audible differences, we select this model as our proposed model (marked in bold). To emphasize the advantage of a two-stage approach, the performance of a single-stage Y-Net, which employs enhanced signal estimation by AEC and noise suppression in one step (similar to [13]), is shown in the last line. The number of feature maps F is set to 100 for this model to achieve a complexity comparable to our Y^2-Net approaches. We observe that our Y^2-Net with intermediate echo estimation not only allows for a considerably better echo suppression, but also delivers better overall PESQ.

Table 1 also reports real-time factors (RTFs), which represent the relative computing time of a time frame compared to the frame shift. Values below 1.0 imply that the inference with this model can be performed in real-time. The RTF was measured on an `Intel i7-10510U` quad-core CPU at 1.8 GHz. All RTFs are below 1.0, therefore the models comply with the challenge rules for the real-time track.

Table 2 shows the results of the proposed model on the preliminary real testset $\mathcal{D}_{\text{real}}^{\text{test,pre}}$ for the DECMOS metric. Our Y^2-Net models perform better than the baseline, with especially notable improvements in the double-talk scenario. Among the tested variants, the Y^2-Net (LF/LE) (E, $\hat{\text{D}}$) maintains the best performance in total and was therefore selected as final model.

As reported in Table 2, we find that both blind real testset $\mathcal{D}_{\text{real}}^{\text{test}}$ single-talk conditions in DECMOS metrics and subjective listening MOS yield similar results for baseline and Y^2-Net (LF/LE) (E, $\hat{\text{D}}$) in Table 3 as well. For double-talk FE, the strong DECMOS score is in line with Table 2. In contrast, the subjective challenge test results show an unexpected higher score of the baseline, while the NE double-talk condition again shows higher scores for our proposed Y^2-Net, subjective and objective. Note that our model shows 0.46 DECMOS points improvement over the baseline in double-talk. Our informal listening supports the DECMOS metric in all conditions.

5. Conclusions

In this work we present a novel Y^2-Net, a two-stage fully convolutional recurrent network model for acoustic echo cancellation and noise suppression, with an echo estimate as output of the first stage. As a further novelty for DNN AEC, we propose to use the estimated echo signal of the first stage as part of the second stage's input, as well as a joint training for both stages still including the intermediate echo estimate loss term. Compared to the baseline approach, the model significantly increases performance during double-talk by an average of 0.46 MOS points in the DECMOS metric.

6. References

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A Practical Approach*. Hoboken, NJ, USA: Wiley-Interscience, 2004.

[2] J. Lee and C. Un, "Block Realization of Multirate Adaptive Digital Filters," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 1, pp. 105–117, Feb. 1986.

[3] H. Shin, A. H. Sayed, and W. Song, "Variable Step-Size NLMS and Affine Projection Algorithms," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 132–135, Feb. 2004.

[4] K. Steinert, M. Schönle, C. Beaugeant, and T. Fingscheidt, "Hands-free System with Low-Delay Subband Acoustic Echo Control and Noise Reduction," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 1521–1524.

[5] R. E. Kalman, "A New Approach to Linear Filtering and Prediction Problems: Journal of Basic Engineering," *Journal of Basic Engineering*, vol. 82, no. 1, pp. 35–45, Mar. 1960.

[6] G. Enzner and P. Vary, "Frequency-Domain Adaptive Kalman Filter for Acoustic Echo Control in Hands-Free Telephones," *Signal Processing (Elsevier)*, vol. 86, no. 6, pp. 1140–1156, Jun. 2006.

[7] J. Franzen and T. Fingscheidt, "A Delay-Flexible Stereo Acoustic Echo Cancellation for DFT-Based In-Car Communication (ICC) Systems," in *Proc. of INTERSPEECH*, Stockholm, Sweden, Aug. 2017, pp. 181–185.

[8] F. Kuech, E. Mabande, and G. Enzner, "State-Space Architecture of the Partitioned-Block-Based Acoustic Echo Controller," in *Proc. of ICASSP*, Florence, Italy, May 2014, pp. 1295–1299.

[9] J. Franzen and T. Fingscheidt, "An Efficient Residual Echo Suppression for Multi-Channel Acoustic Echo Cancellation Based on the Frequency-Domain Adaptive Kalman Filter," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 226–230.

[10] A. Schwarz, C. Hofmann, and W. Kellermann, "Spectral Feature-Based Nonlinear Residual Echo Suppression," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2013, pp. 1–4.

[11] G. Carbajal, R. Serizel, E. Vincent, and E. Humbert, "Multiple-Input Neural Network-Based Residual Echo Suppression," in *Proc. of ICASSP*, Calgary, AB, Canada, Apr. 2018, pp. 231–235.

[12] H. Zhang and D. Wang, "Deep Learning for Acoustic Echo Cancellation in Noisy and Double-Talk Scenarios," in *Proc. of INTERSPEECH*, Hyderabad, India, Sep. 2018, pp. 3239–3243.

[13] H. Zhang, K. Tan, and D. Wang, "Deep Learning for Joint Acoustic Echo and Noise Cancellation with Nonlinear Distortions," in *Proc. of INTERSPEECH*, Graz, Austria, Sep. 2019, pp. 4255–4259.

[14] J. Franzen, E. Seidel, and T. Fingscheidt, "AEC in a Netshell: On Target and Topology Choices for FCRN Acoustic Echo Cancellation," in *Proc. of ICASSP*, Toronto, Canada, Jun. 2021, pp. 156–160.

[15] Z. Zhao, H. Liu, and T. Fingscheidt, "Convolutional Neural Networks to Enhance Coded Speech," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 27, no. 4, pp. 663–678, Apr. 2019.

[16] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Separated Noise Suppression and Speech Restoration: LSTM-Based Speech Enhancement in Two Stages," in *Proc. of WASPAA*, New Paltz, NY, USA, Oct. 2019, pp. 239–243.

[17] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "Fully Convolutional Recurrent Networks for Speech Enhancement," in *Proc. of ICASSP*, Barcelona, Spain, May 2020, pp. 6674–6678.

[18] M. Strake, B. Defraene, K. Fluyt, W. Tirry, and T. Fingscheidt, "INTERSPEECH 2020 Deep Noise Suppression Challenge: A Fully Convolutional Recurrent Network (FCRN) for Joint Dereverberation and Denoising," in *Proc. of INTERSPEECH*, Shanghai, China, Oct. 2020, pp. 2467–2471.

[19] M. Halimeh, T. Haubner, A. Briegleb, A. Schmidt, and W. Kellermann, "Combining Adaptive Filtering and Complex-Valued Deep Postfiltering for Acoustic Echo Cancellation," *ResearchGate: 10.13140/RG.2.2.14083.94241*, Nov. 2020.

[20] V. Dumoulin and F. Visin, "A Guide to Convolution Arithmetic for Deep Learning," *arXiv:1603.07285*, Jan. 2018.

[21] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier Nonlinearities Improve Neural Network Acoustic Models," in *Proc. of ICML Workshop on Deep Learning for Audio, Speech, and Language Processing*, Atlanta, GA, USA, Jun. 2013, pp. 1–6.

[22] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, and W. Woo, "Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting," in *Proc. of NIPS*, Montreal, QC, Canada, Dec. 2015, pp. 802–810.

[23] M. Abadi *et al.*, "TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems," *arXiv:1603.04467*, Mar. 2016.

[24] R. Cutler, A. Saabas, T. Parnamaa, M. Loide, S. Sootla, M. Purin, H. Gamper, S. Braun, K. Sorensen, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Acoustic Echo Cancellation Challenge: Datasets and Testing Framework," in *Proc. of INTERSPEECH* (accepted for publication), Brno, Czech Republic, Sep. 2021.

[25] H. McGuire, "LibriVox: Acoustical liberation of books in the public domain," 2005–present, [accessed 26.03.2021]. [Online]. Available: https://librivox.org/

[26] N. Raghuvanshi and J. Snyder, "Project Triton," 2010–present, [accessed 26.03.2021]. [Online]. Available: https://www.microsoft.com/en-us/research/project/project-triton/

[27] K. Sridhar, R. Cutler, A. Saabas, T. Parnamaa, M. Loide, H. Gamper, S. Braun, R. Aichner, and S. Srinivasan, "ICASSP 2021 Acoustic Echo Cancellation Challenge: Datasets, Testing Framework, and Results," *arXiv:2009.04972*, Oct. 2020.

[28] C. K A Reddy, H. Dubey, K. Koishida, A. Nair, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "INTERSPEECH 2021 Deep Noise Suppression Challenge," *arXiv:2101.01902*, Jan. 2021.

[29] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *Proc. of ICLR*, San Diego, CA, USA, May 2015, pp. 1–15.

[30] R. Pascanu, T. Mikolov, and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks," in *Proc. of ICML*, Atlanta, GA, USA, Jun. 2013, p. 1310–1318.

[31] "ITU-T Recommendation P.1110, Wideband Hands-Free Communication in Motor Vehicles," ITU-T, Mar. 2017.

[32] T. Fingscheidt and S. Suhadi, "Quality Assessment of Speech Enhancement Systems by Separation of Enhanced Speech, Noise, and Echo," in *Proc. of INTERSPEECH*, Antwerp, Belgium, Aug. 2007, pp. 818–821.

[33] T. Fingscheidt, S. Suhadi, and K. Steinert, "Towards Objective Quality Assessment of Speech Enhancement Systems in a Black Box Approach," in *Proc. of ICASSP*, Las Vegas, NV, USA, Apr. 2008, pp. 273–276.

[34] K. Steinert, S. Suhadi, T. Fingscheidt, and M. Schönle, "Instrumental Speech Distortion Assessment of Black Box Speech Enhancement Systems," in *Proc. of IWAENC*, Seattle, WA, USA, Sep. 2008, pp. 273–276.

[35] "ITU-T Recommendation P.862.2, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU-T, Nov. 2007.

[36] "ITU-T Recommendation P.862.2 Corrigendum 1, Wideband Extension to Recommendation P.862 for the Assessment of Wideband Telephone Networks and Speech Codecs," ITU-T, Oct. 2017.

[37] "ITU-T Recommendation P.808, Subjective Evaluation of Speech Quality with a Crowdsourcing Approach," ITU-T, Jun. 2018.

[38] "ITU-T Recommendation P.831, Subjective Performance Evaluation of Network Echo Cancellers," ITU-T, Dec. 1998.