Progressive Speech Enhancement with Residual Connections

Jorge Llombart, Dayana Ribas, Antonio Miguel, Luis Vicente, Alfonso Ortega, Eduardo Lleida
ViV oLab, Aragon Institute for Engineering Research (I3A)
University of Zaragoza, Spain
{jllombg,dribas,amiguel,lvicente,ortega,lleida}@unizar.es

Abstract
This paper studies the Speech Enhancement based on Deep Neural Networks. The proposed architecture gradually follows the signal transformation during enhancement by means of a visualization probe at each network block. Alongside the process, the enhancement performance is visually inspected and evaluated in terms of regression cost. This progressive scheme is based on Residual Networks. During the process, we investigate a residual connection with a constant number of channels, including internal state between blocks, and adding progressive supervision. The insights provided by the interpretation of the network enhancement process leads us to design an improved architecture for the enhancement purpose. Following this strategy, we are able to obtain speech enhancement results beyond the state-of-the-art, achieving a favorable trade-off between dereverberation and the amount of spectral distortion.

Index Terms: progressive speech enhancement, deep learning, interpretability, residual networks, speech quality measures.

1. Introduction
During the last few years, Speech Enhancement (SE) based on Deep Neural Networks (DNN) has emerged and positioned among the most active topics in the speech processing community. Previous work evidences the ability of deep learning approaches for discovering underlying relations between the clean and the corrupted signal [1, 2, 3, 4]. The fact is that beyond what we know about the network main goal of minimizing the error between clean and corrupted signal, we actually are not sure on further “why” and “how” transformations are happening inside the network. This black box effect is probably the major handicap/complaint against deep learning solutions, because it hinders the research process, and gives place to many empirical solutions.

Interpretability of deep neural networks has recently emerged as an area of machine learning research. It aims to provide a better understanding of how models perform feature selection and derive their classification decisions, such that the findings impact DNN solutions. Recently, the top scientific conferences in the field have dedicated special spaces to this aim [5, 6], evidencing the interest of the R&D community.

This paper is motivated to contribute to the accuracy and interpretability of SE solutions. We present a SE architecture following the feature-mapping strategy, where the enhancement process can be followed step by step by means of a visualization probe at each network block. The visualization of the partial enhancement in each step allows us to supervise the process and collect relevant details on how it is performed. This information is useful to detect which steps are meaningful in the enhancement, and which others can be discarded during the evaluation. Therefore, even when the network is already trained, we can select a different grade of enhancement according to the application. This way we have obtained a proper trade-off between accuracy and computational effort.

The architecture proposed is based on the recent and powerful topology of Residual Networks (RN) [7] using one-dimensional convolution layers. We have been able to get close to a performance in the state-of-the-art while exposing the step-by-step network processing at the same time. The high potential of convolutional networks and residual connections have previously improved the performances for the speech enhancement task [8, 9]. Residual connections provide the signal with a linear shortcut, while the non-linear path enhances it in several steps by adding or subtracting corrections. This mechanism could automatically compensate the amount of processing with the level of distortion in the signal. This has also been tested in different more complex architectures using recurrent-based topologies [10] or adversarial networks [11]. Recent work has studied the effect of residual and highway connections in speech enhancement models contributing to a better understanding of such models [12].

Contributions of this paper complement previous work on SE using deep learning and the interpretability of these solutions. We search a better understanding of speech enhancement process by closely following the transformations carried out by the architecture. This way we contribute to improving the accuracy of the SE through a novel architecture based on RN.

In the following, section 2 presents the proposed architecture and its evolution through the study. Section 3 describes the experimental setup designed for testing the SE performance in reverberated speech through speech quality measures. Section 4 discusses results and finally section 5 concludes the paper.

2. Proposed architectures

2.1. Constant Channel Residual Network
In order to progressively enhance the input without losing the spectral representation in each block, we have designed an RN that maintains the same number of channels throughout the residual connection. We will call this architecture Constant Channel Residual Network (CCRN). Figure 1 shows our system which uses multiple input sources to provide a variety of signal representations. We provide multiple representations of the input in order to maintain as much reverberant impulse response inside the preprocessing analysis window as possible, without losing the temporal resolution of the acoustic events. The front-end starts segmenting speech signals in 25, 50, and 75 ms Hamming window frames, every 10 ms. For each frame segment, three types of acoustic feature vectors are computed and stacked depending on the window width, to create a single input feature vector for the network: the logarithm of a 512-dimensional FPT, and 32, 50, 100-dimensional Mel filterbank & cepstral features. So the input feature vector has 876 dimensions. Finally, each feature vector is variance normalized.
The network processes input features with a first convolutional layer followed by $L = 14$ Residual Blocks (RB). This first layer uses the input dimension as the number of input channels, and the dimension of the logarithmic spectrum as the number of output channels. The RB has two stages composed by a Batch Normalization (BN) layer [13], a non-linearity by means of a Parametric Rectified Linear Unit (PReLU) [14], and a 1-dimension convolutional layer with a kernel of $k = 3$ with the same number of channels at the output as at the input. The combination of BN and PReLU provides a smoother representation for regression tasks than the typical ReLU. The residual connection is the addition of the input of the RB and its output. Our goal is to estimate the logarithmic spectrum of the clean signal (i.e. the enhanced $X_{S,L}$) from the logarithmic spectrum of the noisy signal. Based on the experience from previous work [15], we process the full input signal as a sequence, instead of frame by frame. We use a loss function based on Mean Square Error (MSE) among frames (equation: (1)):

$$J(Y, X_{S,L}) = \frac{1}{N} \sum_{n=0}^{N-1} \sum_{t=0}^{T-1} \text{MSE}(y_{n,t}, x_{S,L,n,t})$$

where $N$ is the feature dimension, $T$ is the sequence length, $y_{n,t}$ are $Y$ frames and $x_{S,L,n,t}$ are $X_{S,L}$ frames.

Based on the previous description, we place a probe-output at each block, such that we can inspect the evolution of the enhancement process. This is possible because we maintain unaltered the number of channels for each RB. Figure 2 shows an example of some steps output spectrum obtained.

Note that the standard convolution layer is a linear combination of all input channels with a context for each output channel. Therefore the resulting 512-dimensional matrix does not correspond to a proper spectrum. Also see that some frequency channels seem to group great part of the spectral information, while the rest seems to get blurred. This indicates the network focuses on certain frequencies channels consistently with the findings in [10]. This means the network steps over the spectral time-frequency structure, and only considers the weight changes from the MSE minimization. This could be related to the different level of distortion among the frequency channels.

Furthermore, see that the first and last step of the network processing is quite remarkable. The first step transforms the signal spectrum to strong or blurred frequency channels from the convolution output. While the last step suddenly reorders and recovers the signal with enhancement included.

2.2. Constant Channel Residual Network with State Path

The CCRN architecture was designed to progressively enhance the input signal without losing the log-spectral representation. However, despite we provide a shortcut path to allow the input to pass through with additive modifications, the training of the CCRN makes a disorganized spectral representation. As we show in section 2.1 apparently most part of the information is grouped in some channels. In order to pass the input over the residual path without changing its representation, we add a state path between RBs. In this way, the representation of the signal created by the network has its own path to going on. Moreover, this state path allows having more channels at each layer, while maintaining the same number of channels in the residual path. We call the new architecture in figure 3 Constant Channel Residual Network with State path (CCRN-State).

This architecture stacks the channels of both paths at the input of the block. Inspired on the Wide Residual Networks [16], we increase the number of channels in the first convolution of the block. Then, in order to obtain the residual path and a new
2.3. Progressive Supervision

Finally, to force the networks to provide a proper signal reconstruction at each step we add the MSE cost term at each block output. You can see antecedents of this strategy in classification tasks [17], although we are using it in a regression task. In equation (2), we add to the training cost the MSE between the clean reference and each block output $X_{S,l} \forall l \in [1,L]$. This second part in the cost function is weighted by $\alpha$. In our experiment, we choose $\alpha = 0.1$ from development trials. We call this cost Progressive Supervision, because we take care of how much the network enhances the signal in each block.

$$J_{PS}(Y, X_{S,L}) = J(Y, X_{S,L}) + \alpha \frac{1}{L} \sum_{l=1}^{L} J(Y, X_{S,l}) (2)$$

Figure 4 shows a spectrogram example where we can see that finally we obtained an evolutionary pattern in the enhancement process. Note that both architectures explored in previous sections 2.1 and 2.2 will have the same output representation because throughout the residual path there is a constant number of channels. Anyway, at the end of the processing, enhancement results could be different.

We can see the earlier blocks take care of the more noticeable distorted areas of the spectrum, e.g. look at the trail of the reverberation. Apparently, the network is establishing a pattern of what is distortion and what is not. We also note that the network mainly focuses on the spectrum valleys and gradually the granularity in them is removed. Also, see how the spectral trail effect because of reverberation, is gradually removed. After last blocks, the network starts softening the spectrum in order to produce slow spectral magnitude changes. This avoids undesirable auto-generated distortions such as the annoying musical noise. However, it could also have an over-softening effect that causes an unrealistic effect in the final output (see block 14 output).

So far, the interpretative analysis of the enhancement process does not allow us to be totally sure about the impact on SE performance. In the following sections, we will assess the accuracy of the proposed model in an objective manner. Note that due to space issues the analysis so far used a single signal example. However, the observations discussed are common to different sentences and distortion types. See more examples in https://medium.com/vivolab

3. Experimental setup

Speech enhancement system was developed using the Pytorch toolkit [18]. Input examples for training were generated on-the-fly, distorting contiguous random sequences of 200 samples from Timit[19]. Librispeech[20] and Tedlium[21] databases. AdamW algorithm was used to train the network [22, 23].

Approaches were tested on the official Development and Evaluation sets of the REVERB Challenge [24]. The dataset has simulated speech from the convolution of WSJCAM0 Corpus [25] with three measured Room Impulse Responses (RIR) ($RT_{60} = 0.25, 0.5, 0.7$s) at two speaker-microphone distances: far (2m) and near (0.5m). It was added stationary noise recordings from the same rooms ($SNR = 20$ dB). Besides, it has real recordings, acquired in a reverberant meeting room ($RT_{60} = 0.7s$) at two speaker-microphone distances: far (2.5m) and near (1m) from the MC-WSJ-AV corpus [26]. We also used real speech samples from VoiceHome v0.2 [27] and v1.0 [28]. VoiceHome was recorded in a domestic environment from 3 real homes, such that the background noise is the one typically found in households e.g. dishwasher, vacuum or television.

3.1. Reference and performance assessment

We compare the performance with the state-of-the-art dereverberation method called Weighted Prediction Error (WPE)[29], which is known to effectively reduce reverberation and greatly boosts the speech enhancement performance. We used the more recent version of WPE [30] which is also based on DNN [31]. However, WPE uses a different architecture based on LSTM.

In order to assess the enhancement, we measure the distortion reduction through Log-likelihood ratio (also known as Itakura distance) ($LLR$) [32] computed in the active speech segments. The closer the target feature to the reference, the lower the spectral distortion, therefore smaller values indicate better speech quality. On the other hand, we assess the reverberation level of the signal through Speech-to-Reverberation Modulation energy Ratio (SRMR) [33]. In this case, higher values indicate better speech quality.

4. Results and Discussion

4.1. Speech quality for processing tasks

Table 1 presents speech quality results in terms of distortion. The first row corresponds to the reverberant unprocessed speech compared to the quality achieved using WPE or CCRN-based architectures. All enhancement methods that appear in this paper were able to enhance the corrupted speech data, but our proposed architectures outperform WPE in terms of distortion. In spite of the CCRN-State based architectures have more degrees of freedom note that CCRN + Progressive Supervision reach the best performance.

4.2. Speech quality for dereverberation: Simulated vs. Real

Table 2 shows the average of SRMR results over the evaluated conditions for simulated and real speech samples. The first row corresponds to the unprocessed speech data. Note that
the best performances for all datasets evaluated were achieved by CCRN + Progressive Supervision, consistent with the previous result for LLR. The consistency in performance through different datasets supports the robustness of the method, indicating that its parameters are not adjusted to some specific set of speech signals. Furthermore, positive results beyond simulated reverberated speech encourage the use of this method in realistic scenarios. Moreover, note that all CCRN models were trained with artificially synthesized reverberation, however, they showed to be effectively dealing with a reverberated speech from real-world scenarios.

Table 2: Speech quality through SRMR results for real reverberated speech samples.

<table>
<thead>
<tr>
<th>Methods</th>
<th>REV-Dev</th>
<th>REV-Eval</th>
<th>VH-v0.2</th>
<th>VH-v1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unprocessed</td>
<td>3.79</td>
<td>3.18</td>
<td>3.19</td>
<td>4.51</td>
</tr>
<tr>
<td>CCRN</td>
<td>4.70</td>
<td>4.11</td>
<td>5.14</td>
<td>5.98</td>
</tr>
<tr>
<td>+Prog Sup</td>
<td>5.01</td>
<td>4.44</td>
<td>6.13</td>
<td>7.01</td>
</tr>
<tr>
<td>CCRN-State</td>
<td>4.65</td>
<td>3.87</td>
<td>3.09</td>
<td>6.45</td>
</tr>
<tr>
<td>+Prog Sup</td>
<td>4.88</td>
<td>4.20</td>
<td>5.35</td>
<td>0.62</td>
</tr>
</tbody>
</table>

4.2.1. Reverberation level, Room sizes, and Near & Far field

Figure 5a shows the evolution of SRMR results with the increase of reverberation level for different room sizes and RT60($s$).

Figure 5b shows the variation of speech quality through SRMR measure in simulated reverberated speech samples from REVERB Dev & Eval datasets.

All CCRN-based methods outperform WPE baseline, but CCRN + Progressive Supervision achieves the higher speech quality for all conditions evaluated. Similar behaviour is obtained for far (250m) and near (50m) conditions (Figure 5b). Note that again the CCRN + Progressive Supervision model achieves the best performance, mainly in far-field conditions. The introduction of the Progressive Supervision stimulates the correct performance of the enhancement process beyond the blind modification of frequency channels performed by CCRN and CCRN-State. It contributes to regularize the network parameters and also provides a progressive transformation of the spectrum towards the final enhancement.

4.3. Architecture discussion

During the enhancement process, architectures CCRN and CCRN-State provide a messy spectrum that does not supply clues of what the network is doing. This representation might be a codification of the input, or even it could be learning the training examples. With Progressive Supervision, the networks are able to show the evolution of the predicted signal. Moreover, this cost function regularizes the network because it prevents to learn concrete examples. Finally, we can see how the signal is enhanced at each block.

We can view the progressive enhancement through residual connections as a spectral power subtraction method. Each RB computes the weighted spectral power, while the residual connection adds it to the corrupted signal. In this context, the CCRN + Progressive Supervision method is comparable to the WPE method. Both, estimate the spectral power to make a subtraction. CCRN + Progressive Supervision uses convolutional layers at each of the RB, and WPE uses an LSTM. However, unlike WPE, CCRN + Progressive Supervision applies as many spectrum subtractions as blocks have the architecture.

An additional advantage of the CCRN + Progressive Supervision is that during training we have access to the reconstruction error at each block. This allows us to train a big network with many RB and then only use the number of RB that actually provides significant cost reduction. This is a desirable quality when dealing with a clean signal.

5. Conclusions and Future

This paper proposed a deep learning solution for Speech Enhancement based on residual networks. By means of an interpretative study of the progressive transformations performed by the network, we were able to design an improved architecture for the speech enhancement purpose. The mechanism of Progressive Supervision contributed to regularize the network parameters. It demonstrated to be able to stimulate the correct performance of the enhancement process, beyond the blind modification of frequency channels performed by the other alternative evaluated (CCRN and CCRN-State). The proposal obtained speech enhancement results beyond the state-of-the-art, achieving a favorable trade-off between dereverberation and the amount of spectral distortion.

We showed that the use of interpretative analysis of the process inside the networks can provide useful insights to develop improved solutions. Future work will use the experience at introducing the reconstruction error in each network block for exploring a novel design that prevents an over-enhancement effect. We plan to define an architecture with many RB at training, and during evaluation, to adjust the network size considering the information provided by the reconstruction error.

6. Acknowledgment

Funded by the Government of Aragon (Reference Group T36_17R) and co-financed with Feder 2014-2020 “Building Europe from Aragon”. We thanks the Spanish Ministry of Economy and Competitiveness through the project TIN2017-85854-C4-1-R, and NVIDIA Corporation for the donation of the Titan Xp GPU. This material is based upon work supported by Google Cloud.
7. References


