DARE-Net: Speech Dereverberation And Room Impulse Response Estimation

Jacob Donley and Paul Calamia
{jdonley, pcalamia}@stanford.edu

1 Introduction

Speech enhancement is a common task for video calling, automatic speech recognition, speech communications, home assistants, and audio forensics [1, 2, 3]. One component of speech enhancement is dereverberation, where the influence of the room on the speech that is captured, in the form of acoustic reflections and scattering from surfaces and objects, is removed to yield a clean, more intelligible version of the spoken words. Deep-learning models for speech dereverberation often rely on synthetic datasets for training in which clean (*i.e.,* anechoic) speech is convolved with room impulse responses (RIRs) to create reverberant speech. Various loss functions and error metrics are used to compare the estimated clean-speech output of the model to the clean-speech input prior to convolution, however the RIR is rarely optimized for or considered as an output. In this project, we have implemented and evaluated deep-learning based speech dereverberation models that estimate both the clean speech and the RIR, the latter of which can be used in downstream tasks such as characterizing the acoustics of a space, or auralizing other sounds to simulate them being in that space.

2 Related Work

Speech *enhancement* has been a popular topic in the research community for decades. More recently there have been advancements in signal processing and deep learning systems, including work that combines multiple systems together to accomplish speech enhancement in an augmented reality application [4]. Further work has looked at hybrid approaches combining signal processing techniques and machine learning methods [5]. Other approaches have tackled the problem using purely deep-learning-based methods, such as with graph neural networks [6] and triple-path attentive recurrent neural networks (RNNs) [7].

Speech *dereverberation* also has seen advances with the use of deep networks such as temporal convolutional networks [8] and, more recently, with generative adversarial networks (GANs) [9]. Room impulse response (RIR) estimation is less common, although deep-learning -based approaches do exist, *e.g.,* using auto-encoders [10] or GANs [11].

3 Approach

We consider two approaches for this project, both based on the U-net architecture [12]. The first model is based on Wave-u-net, shown in Figure 1, a time-domain approach which previously has been applied to acoustic source separation [13] and speech enhancement [14]. We downloaded the model from `https://github.com/f90/Wave-U-Net-Pytorch`, and made some modifications to integrate it into our Pytorch Lightning framework, and to estimate clean speech and an RIR (deconvolution) rather than to separate sources from an additive mixture. For this model the input reverberant speech (~ 4.6s), and the output clean speech and RIR are all in the time domain.

CS230: Deep Learning, Fall 2022, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

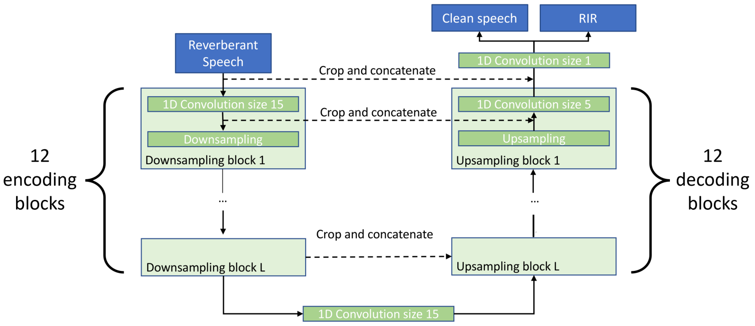

Figure 1: Wave-u-net architecture, adapted from [14]

.

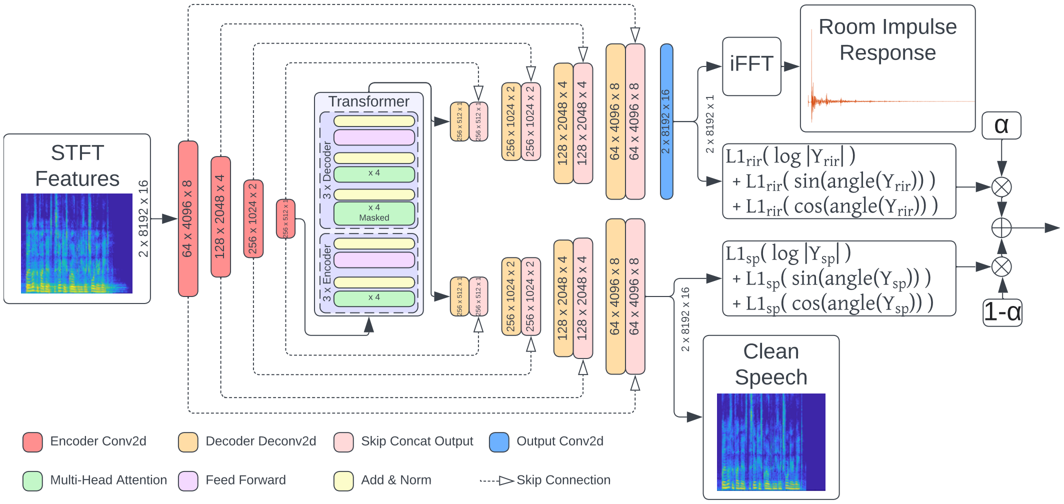

Figure 2: DARE-Net: Our proposed dual-branch UNet with transformer model architecture. The proposed model is trained using a tunable speech and RIR loss.

The second model we call DARE-Net, which significantly extends the model depicted in Figure 2 of [15]. The version from the paper was excessively large, so we reduced both the encoder and decoder segments from 8 layers to 4. We added a second decoder branch to estimate the RIR from the embedding with a hyper-parameter, α, (to trade-off DARE performance) and a transformer with 4-head attention, 3 decoder and 3 encoder layers to act on the UNet embedding layer to learn temporal and spectral relationships. See Figure 2 for the specific architecture. As input we use approximately 2 seconds of reverberant speech in a 2x8192x16 format (channels for magnitude and phase x frequency bins x STFT time). We also modify the training procedure with and without a novel pre-alignment step which significantly improves gradient descent during training. The pre-alignment aligns the target RIR to the early stage output of the network to lower the loss drastically. The ground-truth RIRs must always contain the same initial group delay as there is no way for the network to determine this in a practical scenario.

We try many different loss functions in the time-domain, frequency-domain and time-frequency domain, and find that L1 and MSE losses on the magnitude and phase work best. For the Wave-u-net we investigate L1, MSE and Multi-resolution STFT loss [16, 17] on the whole time domain RIR signal as well as MSE on a 25ms segment around the peak of the RIR. For the Wave-u-net speech loss, MSE on the whole time domain signal was used. The frequency domain DARE-Net model uses L1 loss on the log-magnitude spectrum plus L1 loss on the cos and sin of the phase component to ensure smooth gradients on the circular phase, which was the best performing combination of losses. Other losses that were less effective include L1 and MSE on the unwrapped phase, the mel-scaled magnitude spectrum, the time domain signal, the log absolute time domain signal, the peak value, the peak time delay, the peak position and Kullback–Leibler (KL) divergence on the log absolute time domain signal. However, all metric losses were informative with respect to what they were measuring.

Our code can be accessed from:
`https://github.com/jdonley/Speech-Dereverberation-and-RIR-Estimation`.
The jdonley/dev branch contains our most recent proposed DARE-Net models, and the pcalamia-dev branch contains our most recent Wave-u-net implementation.

4 Datasets

For supervised training, we require that there is a ground truth RIR and a clean speech sample for each reverberant speech example in order for our loss functions to compute the error between the predicted output and the target ground-truth sample for both components. We generate the reverberant speech examples by convolving clean speech with RIRs. For the clean speech, we use the LibriSpeech dataset [18] with a predefined train/dev/test split of 100.6/5.4/5.4 hours of data. For the RIRs we use the MIT IR Survey [19] with a train/dev/test split of 80/10/10% (216/27/27 RIRs from the set of 270). When training and evaluating our models we use a 160k/800/8k split built from the splits of the components.

5 Results

5.1 WaveUNet

Sample results for our wave-u-net model are shown in 3, for which we used a learning rate of 10^{-6}, MSE loss on the speech prediction, and MSE loss on 25ms around the peak of the RIR. The plots of the target and predicted clean speech suggest that the model is able to estimate that component, although listening tests revealed audible distortion artifacts in addition to reduced reverberation. The plots of the target and predicted RIR suggest that this architecture is unable to learn that component. These results were consistent over different loss functions and learning rates. The learning curves for this model (not shown) indicated that the training loss decayed rapidly for approximately half of an epoch (approximately 80k examples) with only marginal improvement after that (up to 70 epochs, our longest training run). The validation loss similarly showed an extremely shallow decay with little improvement over the training period.

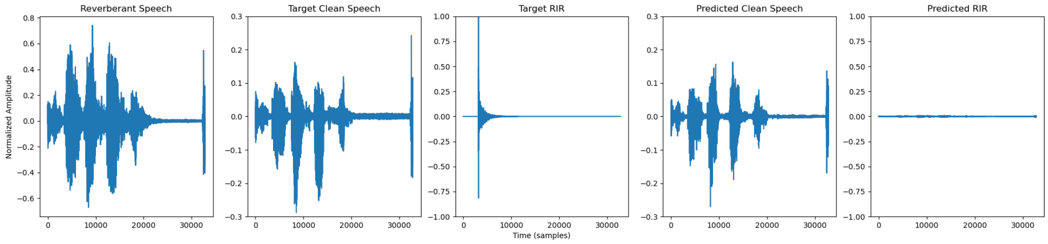

Figure 3: Sample output from our Wave-u-net model after 70 epochs, using an MSE loss on the clean speech, MSE loss on 25ms around the peak of the RIR, and a learning rate of 10^{-6}.

5.2 Proposed DARE-Net

We present results for the proposed DARE-Net architecture as well as an ablation study to show the benefits and trade-offs of a transformer architecture on the embedding space, a joint branched-decoder-loss optimization for dereverberation and RIR estimation and the proposed pre-alignment step described in Section 3.

The validation loss curve for all training experiments using DARE-net followed the training loss curve at slightly larger values due to dropout regularization on the encoder, decoder and transformer layers, which helped prevent model over-fitting. Due to the long training experiments, we ensured that the learning rate was first approximately tuned between 10^{-2} and 10^{-4} and then used an exponentially

Model	α	Epochs	RIR				Speech			
			$\|\|\mathrm{M}\|\|_1$	$\|\|\theta\|\|_1$	$\|\|\mathrm{M}\|\|_2^2$	$\|\|\theta\|\|_2^2$	$\|\|\mathrm{M}\|\|_1$	$\|\|\theta\|\|_1$	$\|\|\mathrm{M}\|\|_2^2$	$\|\|\theta\|\|_2^2$
Wave-u-net	-	70	3.44	1.62	18.3	2.00	-	-	-	-
DARE (IR)	-	60	.883	1.61	1.88	2.00	-	-	-	-
DARE (IR+PA)	-	100	.511	.794	.521	.675	-	-	-	-
DARE (IR+PA+T)	-	100	**.422**	**.721**	**.391**	**.638**	-	-	-	-
DARE (IR+SP+PA+T)	0.1	20	**.461**	.740	**.438**	**.624**	.700	**.920**	.495	**.865**
DARE (IR+SP+PA+T)	0.5	20	.567	.731	.914	.676	.700	1.13	**.493**	1.17
DARE (IR+SP+PA+T)	0.9	20	.490	**.723**	.638	.652	**.696**	1.16	.495	1.21

Table 1: Resulting frequency-domain metric values on the test set. IR refers to the model with an RIR decoder, SP refers to the model with a speech decoder, PA refers to the pre-alignment step, T refers to the embedding transformer. The time-domain metrics for the Wave-u-net were: speech MSE = 0.001, speech L1 = 0.017, RIR MSE = 0.001, RIR L1 = 0.002.

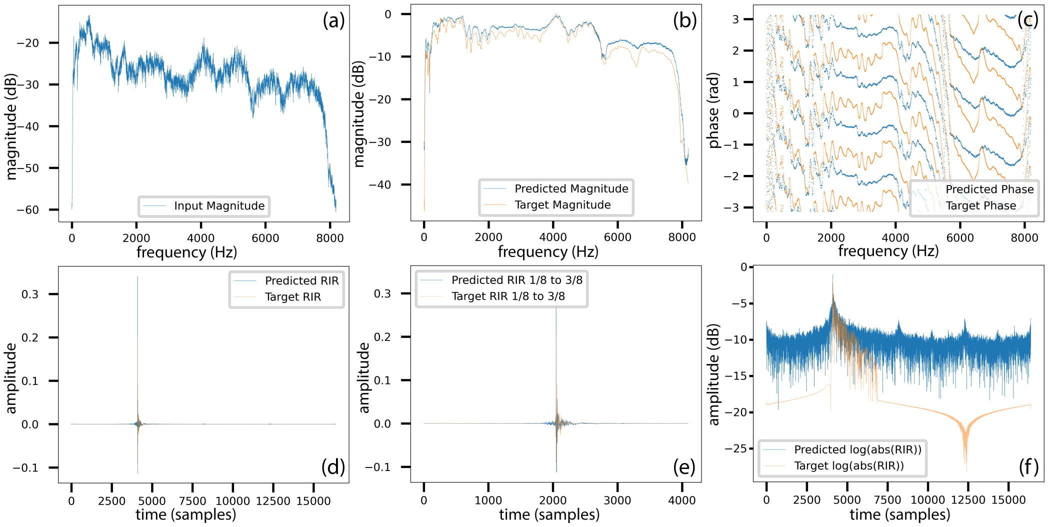

Figure 4: Example output from the 15th epoch of DARE (IR+SP+PA+T) with $\alpha = 0.5$. The following are shown: (a) the noisy input magnitude spectrum, (b) the predicted and target magnitude spectrum, (c) the predicted and target phase, (d) the predicted and target 1second RIR, (e) zoomed-in view of (d), (f) the magnitude of the time domain predicted and target RIR.

decaying learning rate scheduler with $\gamma = 0.9$, which we computed by finding the value that would drop the learning rate by approximately 5 orders of magnitude at 100 epochs.

6 Analysis and Discussion

We found the wave-u-net model to perform poorly regardless of the loss functions, learning rate, and other hyperparameter choices. The predicted clean speech was typically intelligible and noticeably dereverberated but distorted. RIR predictions typically had a small amplitude and a temporal structure more like speech than an RIR, probably due to the skip connections that inject reverberant speech into the late stages of the decoder which could not be suppressed. The model did not appear to be over-fitting based on the learning curves, and because running inference on training data produced similar results to those from the test data.

As mentioned above, we hypothesize that the failure of the Wave-u-net for RIR estimation may be due in part to the skip connections because they add a signal with a very different temporal structure from the RIR to the decoder layers, and the network as implemented does not seem to be able to suppress this. A better network structure might be to have separate decoding blocks, one for the clean speech with skip connections, and one for the RIR without them.

Independent of the RIR estimation, the assessment of speech dereverberation performance with the Wave-u-net would have benefited from blind estimation of reverberation time from speech samples, *e.g.*, as described in [20], to quantify the reduction of reverberation from the input to the output. However, we were unable to find a pre-trained model for this and did not have time to train one ourselves.

We observe that DARE-net outperforms Wave-u-net in all cases. In our ablation study, we find significant improvements are made by performing a pre-alignment step, reducing almost all metrics by over a half. Further improvements are made when using a transformer on the embedding space of the UNet architecture. Our best performing model is DARE-Net with 'IR+PA+T'. This model was also one of the most difficult to train with approximately 15 million parameters and 100 epochs.

We also analyze the effect of changing the α parameter on the dual-branch DARE-Net. We observe that the performance varies across certain metrics when varying α. On average, the network performs best when $\alpha = 0.1$ and is biased towards speech loss. Interestingly, $\alpha = 0.9$ is not the worst performer and outperforms $\alpha = 0.5$, indicating that the network learns best from either RIR loss or speech loss. However, in only 20 epochs, emphasizing the speech loss with lower α values performs almost as well as the best 100-epoch trained RIR loss model, *i.e.* DARE-Net with 'IR+PA+T'.

7 Conclusions and Future Work

In this project, we implemented multiple u-net models in an effort to achieve simultaneous speech dereverberation and RIR estimation from reverberant speech. The time-domain Wave-u-net architecture was moderately successful with the dereverberation task, but failed at the RIR-estimation task. DARE-net, our novel model, which operates in the time/frequency domain, was capable of both tasks. We showed that by tuning a trade-off between speech dereverberation and RIR estimation, we could further optimize performance of both tasks. We introduced two other novel changes to typical u-net architectures with pre-alignment and an embedding transformer, both of which we showed significantly improved performance. Our best model is capable of producing accurate RIRs with only 0.422 and 0.721 in magnitude and phase L1 loss, respectively.

There are multiple avenues for future work. First, we'd like to try estimating clean speech and the RIR, convolving those, and computing the loss between the result and the input reverberant speech. This would impose prior knowledge of the two components' relationship on the model as it's learning. Second, RIR parameter estimation and reconstruction, rather than direct estimation, may help the network learn more efficiently. Finally, we would like to investigate using Perceiver-IO networks [21] on the embedding space or as an end-to-end solution for joint speech DARE.

8 Contributions

Jacob implemented the overall framework for our project in PyTorch Lightning, and implemented and analyzed multiple versions of DARE-Net. Paul implemented the U-net model from [15], and integrated and analyzed the Wave-u-net model within our processing framework. We conceptualized the project together and co-wrote the proposal and reports.

9 Code

Our code can be accessed from:
`https://github.com/jdonley/Speech-Dereverberation-and-RIR-Estimation`.
The jdonley/dev branch contains our final U-net models, and the pcalamia-dev branch contains our final Wave-u-net implementation.

References

[1] Chandan KA Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, et al. The Interspeech 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. *arXiv preprint arXiv:2005.13981*, 2020.

[2] Ahmed Kamil Hasan Al-Ali, David Dean, Bouchra Senadji, and Vinod Chandran. Comparison of speech enhancement algorithms for forensic applications. In *Proceedings of the 16th Australasian International Conference on Speech Science and Technology*, pages 169–172. Australasian Speech Science and Technology Association (ASSTA), 2016.

[3] Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L. Seltzer, Heiga Zen, and Mehrez Souden. Speech processing for digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine*, 36(6):111–124, 2019.

[4] Hanan Beit-On, Moti Lugasi, Lior Madmoni, Anjali Menon, Anurag Kumar, Jacob Donley, Vladimir Tourbabin, and Boaz Rafaely. Audio signal processing for telepresence based on wearable array in noisy and dynamic scenes. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8797–8801, 2022.

[5] Jonah Casebeer, Jacob Donley, Daniel Wong, Buye Xu, and Anurag Kumar. Nice-beam: Neural integrated covariance estimators for time-varying beamformers, 2021.

[6] Panagiotis Tzirakis, Anurag Kumar, and Jacob Donley. Multi-channel speech enhancement using graph neural networks. In *2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3415–3419, 2021.

[7] Ashutosh Pandey, Buye Xu, Anurag Kumar, Jacob Donley, Paul Calamia, and DeLiang Wang. TPARN: Triple-path attentive recurrent network for time-domain multichannel speech enhancement. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6497–6501, 2022.

[8] Yan Zhao, DeLiang Wang, Buye Xu, and Tao Zhang. Monaural speech dereverberation using temporal convolutional networks with self attention. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:1598–1607, 2020.

[9] Vinay Kothapally and John HL Hansen. SkipConvGAN: Monaural speech dereverberation using generative adversarial networks via complex time-frequency masking. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1600–1613, 2022.

[10] Christian J Steinmetz, Vamsi Krishna Ithapu, and Paul Calamia. Filtered noise shaping for time domain room impulse response estimation from reverberant speech. In *2021 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 221–225, 2021.

[11] Anton Ratnarajah, Ishwarya Ananthabhotla, Vamsi Krishna Ithapu, Pablo Hoffmann, Dinesh Manocha, and Paul Calamia. Towards improved room impulse response estimation for speech recognition. *arXiv preprint arXiv:2211.04473*, 2022.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[13] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[14] Craig Macartney and Tillman Weyde. Improved speech enhancement with the wave-u-net. *arXiv preprint arXiv:1811.11307*, 2018.

[15] Ori Ernst, Shlomo E Chazan, Sharon Gannot, and Jacob Goldberger. Speech dereverberation using fully convolutional networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 390–394, 2018.

[16] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6199–6203. IEEE, 2020.

[17] Christian J. Steinmetz and Joshua D. Reiss. auraloss: Audio focused loss functions in PyTorch. In *Digital Music Research Network One-day Workshop (DMRN+15)*, 2020. Available at `https://github.com/csteinmetz1/auraloss`.

[18] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210. IEEE, 2015.

[19] James Traer and Josh H. McDermott. Statistics of natural reverberation enable perceptual separation of sound and space. *Proceedings of the National Academy of Sciences*, 113(48):E7856–E7865, 2016.

[20] Philipp Götz, Cagdas Tuna, Andreas Walther, and Emanuël AP Habets. Blind reverberation time estimation in dynamic acoustic conditions. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 581–585, 2022.

[21] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, et al. Perceiver IO: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021.