# Speech Dereverberation for Robust ASR using Deep Learning Techniques

K. Sriram
National Institute of Technology Surathkal, Karnataka

Hemanth S.
National Institute of Technology Surathkal, Karnataka

## ABSTRACT
This paper aims to provide a comprehensive study on different speech dereverberation techniques using deep learning and compares them to find the best possible solution for the said problem.

The persistence of sound after a sound is created is known as reverberation, or reverb in acoustics. A reflection is the result of a sound or signal hitting many surfaces in close proximity. These surfaces might be furniture, people, or even the surrounding air. The reflections build up and eventually disintegrate. The best example of this is when the sound source cuts out but the reflections keep going, amplitude lowering until it reaches zero.

Deep learning is basically a three-layer neural network. By simulating human brain function, although not exactly mimicking it, these neural networks enable the human brain to "learn" from vast quantities of data. Additional hidden layers can aid in optimizing and refining for accuracy, even if a neural network with only one layer can still produce rough predictions.

Deep learning techniques, including UNet, GANs, and LSTM, are implemented in this paper to study speech dereverberation.

Speech reverberation refers to the degradation of the entire signal caused by reflections of the target signal, which diminishes the quality of speech. The objective is to enhance the voice signal by eliminating this reverberation.

## General Terms
Deep Learning techniques, GAN, acoustics, speech, dereverberation

## Keywords
UNet, GAN, deverberation

## 1. INTRODUCTION
Reverberations are defined as reflections that arrive in less than 50 milliseconds. During this process, the target signal is propagating multipath from its source to the microphone. The bulk of the received sound is composed of direct sound, reflections that follow directly after the direct sound and reflections that follow the early reverberation.

There is no direct sound present if there is no line of sight between the source and the observer. Direct sound is the sound that is heard without reflection. The sound that arrives a bit later as a result of reflections off one or more surfaces is known as early reverberation. The direct sound is separated from the reflected sounds by both time and direction. Early reverberation will change when the source and the microphone move around the room, but it shouldn't be heard separately from the direct sound as long as the delay between the reflected sound and the direct sound is less than about 80–100ms. Early reverberation is said to improve speech understanding since it is designed to accentuate the direct sound. Additionally, early reverberation results in coloration, a spectral distortion. Reflections that arrive later and with greater delays than the direct sound cause late reverberation.

Reverberation is not limited to indoor spaces as it exists in forests and other outdoor environments where reflection exists. Reverberation is dependent on the frequency, meaning that when designing architectural spaces, which require certain reverberation lengths in order to function optimally for their intended use, extra attention is given to the length of the decay. compared to a clear echo, which may be heard at least 50 to 90 ms following the preceding sound. The reflections' amplitude progressively decreases to undetectable levels over time. Not only does reverberation occur indoors, but it can also be found outside in places like woodlands and other places where reflection occurs.

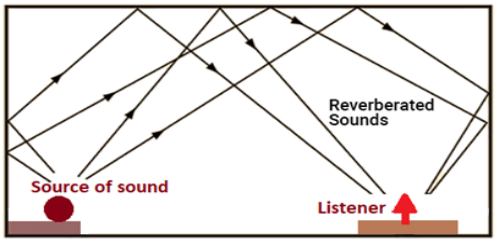


**Figure 1: Reverberation of sound**

In a hall or other performance area with sound-reflective surfaces, reverberation naturally happens when someone speaks or plays an instrument acoustically. Reverb effects are used to artificially apply reverberation. These consist of:

### 1.1 Hall Reverb
Hall reverberations imitate the reverberation of a concert hall. Owing to their enormous size, their decays might last for many seconds or more. These reverbs are ideal for giving strings more body and space.

### 1.2 Chamber Reverb
Similar to hall reverberations, chamber reverberations provide a rich, moody tone. However, they also provide you with an additional dosage of clarity, protecting you from the washed-out sound that many hall reverbs have.

### 1.3 Room Reverb
Room reverberations closely resemble the natural ambiance typically encountered in real-world environments.

## 2. MOTIVATION
Reverberation can most of the time gives recorded sound a

more realistic sense of space but simultaneously making speech less understandable, especially when noise is also present. Hearing impaired people, especially hearing aid users, regularly report having trouble understanding speech in echoing, noisy environments. Another prominent cause of errors in automatic speech recognition (ASR) systems is reverberation.

Dereverberation is the process of lowering a sound or signal's reverberation level. These days many softwares such as Adobe, iZotope provides the functionality of dereverberation.

# 3. LITERATURE REVIEW

## 3.1 Background and Related Works

Many single-channel dereverberation methods have been proposed throughout the years. Lebart, for example, describe a spectral subtraction strategy to reduce late reverberation by using an exponential decay model of reverberation. Wu and Wang offer a two-stage method that cancels early reflections using an inverse filter and removes late reverberation using a spectral subtractor. Long-term linear prediction-based dereverberation approaches have been shown to be very successful at reducing late reverberation. Based on a number of historical frames, these strategies first create frequency dependent linear prediction filters using the weighted prediction error (WPE) minimization.

In order to conduct dereverberation, several supervised speech augmentation algorithms have recently been introduced, and they have much surpassed the traditional methods. In order to create a spectral mapping function from the log magnitude spectrum of reverberant voice to that of anechoic speech, Han et al. recommend using a deep neural network (DNN). Wu et al. stress the importance of reverberation time dependent parameters for training a DNN-based dereverberation system. Next, they provide an improved reverberation-time-aware reverberation removal technique over Han et al. In Weninger et al.'s robust voice recognition system (LSTM), dereverberation is accomplished by a deep bi-directional recurrent neural network (RNN) with long short-term memory. Williamson and Wang suggest estimating a complex ideal ratio mask while accounting for phase.

The supervised algorithms described above exhibit a significant flaw in that they are non-causal, as their processing involves the use of future data. This paper proposes an approach for causal supervised dereverberation.

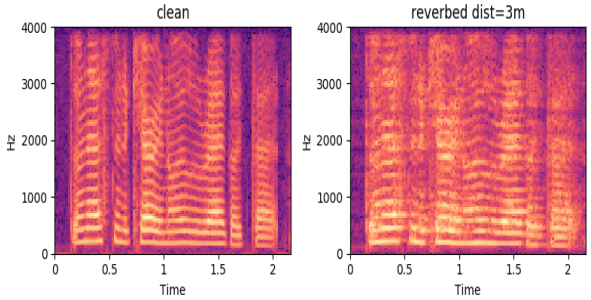


**Figure 2: Spectroscopy of a clean and reverbed sound wave**

## 3.2 Problem Statement

Given a reverberant speech signal, this study proposes, evaluates, and compares various methodologies to determine the most effective algorithm for dereverberation in robust ASR systems.

More technically,

Let **s(t)** and **h(t)** denote speech and room impulse response. The reverberant speech **y(t)** is modeled by

$$\boldsymbol{y(t) = s(t) * h(t)}$$

Where $*$ denotes a convolution operation. Reverberant speech is divided into two components, namely early and late reverberation, based on the arrival time of the signal.

So, the reverberant speech can be represented by

$$\boldsymbol{y(t) = s(t) * h_{\{de\}(t)} + s(t) * h_{\{l\}} = y_{\{de\}(t)} + y_{\{l\}(t)}}$$

The objective of this study is to remove the late reverberation component $\mathbf{y_l(t)}$ from the corresponding reverberant speech **y(t)**.

## 3.3 Objective

To obtain new viewpoints on the problem as well as an extensive understanding of the most advanced ASR systems. This study evaluates speech enhancement and identification methods in reverberant environments, old and new. In addition, it provides academics in appropriate fields with the chance to conduct comprehensive system evaluations utilizing shared data sets.

# 4. PROPOSED METHODOLOGY

## 4.1 Data

The ACE challenge dataset, MERD database, and MARDY database are the three separate datasets that have been combined to create the proposed dataset. By convolving clean speech with a room impulse response from the three datasets mentioned above, it was rendered reverberant. Each of the 28 native English speakers in the train set delivers about 400 phrases. There are just two native English speakers in the sample set, each with approximately 400 sentences. The initial audio files have a 48kHz sample rate.

## 4.2 Evaluated Algorithms

### *4.2.1 Weighted Prediction Error (WPE)*

This approach separates the speech signal into brief periods of time and solves them in the time domain using the weighted error in prediction (WPE) technique.

Since the frequency of the voice input varies over time, the Fourier transform cannot be applied directly. Therefore, the Short-Time Fourier Transform (STFT) is used to apply the Fast Fourier Transform (FFT) to the signal after splitting it into frames. Subsequently, Weighted Prediction Error (WPE) can be applied to each frame, utilizing an approach known as Delayed Linear Prediction (DLP) to estimate the amount of prior signal in the current frame. This estimate for a specific time period can then be subtracted, and the process repeated accordingly.

With DLP, the reverberation can be divided into two parts, viz, early and late reverb. It can be shown that DLP can suppress the late reverb effectively without significantly distorting the short time correlations of the speech, with the assumption that speech is stationary. With the use of time-varying speech characteristics with multichannel linear prediction, the reverbs have been reduced to a significant extent.

### *4.2.2 Fully Convolutional Networks*

Convolutional neural networks (CNNs) are another type of network that enhances the present time frame by using a sliding-window technique. In CNN, an FC layer that disregards any time-frequency structure that may exist occurs after every

pixel in the target picture is determined using just a small amount of context pixel form the initial picture.

One commonly employed technique for translating images to images is an encoder-decoder network. Each layer in this kind of network reduces its input to the one above it until a bottleneck occurs. The input proceeds through the exact same procedure again in the next layer, with each layer upsampling the input till the input takes on its original form. As a result, an image with a high resolution used as the network input is compressed into a relatively tiny image. On the other hand, the growing route has a reverse impact; namely, it raises the resolution of the image until it decreases to its original size. Unfortunately, during the down sampling process, this system frequently loses important low level data.

The "UNet" design, featuring symmetric compression and expansion paths in a U-shape, leverages the need for identical structures in both input and output images to optimize the encoder-decoder architecture. This setup enables the transfer of shared data without downsampling, thus avoiding bottlenecks. UNet accomplishes this by linking mirrored levels in the encoding and decoding stacks, allowing data to flow seamlessly without encountering bottlenecks.

The time-frequency (T-F) representation, or spectroscopy, in this sound enhancement technique can be treated as a picture. As such, the process of enhancement transforms into an image-to-image transition. There are two main benefits to approaching the reverberant speech as a picture. Initially common patterns (such as pitch continuity, harmonics structure, and formants) may be seen in speech spectrograms. These structures can be used by an image analysis approach to apply pertinent improving techniques. Second, this representation enables the use of highly effective computer vision techniques, such as a fully convolutional network (FCN).The suggested UNet design with 256x256 STFT images for both inputs and outputs is seen in the above image.

### *4.2.3 Generative Adversarial Networks (GANs)*

GANs, or Generative Adversarial Networks, are a kind of generative models which makes use of convolution artificial neural networks along with other machine learning methods.

In generative modeling, regularities or structures in the input data are consequently found and understood so that the model is able to generate or results novel instances that could be accurately taken from the initial set of data. This is an unsupervised learning task in machine learning.

By reframing the task as a supervised learning problem with two sub-models—a discriminator model, which seeks to classify examples as either real (from the domain) or fake (generated), and a generator model, trained to create new examples—generative models can be effectively developed using GANs. These models train together in an adversarial zero-sum game until the discriminator is deceived approximately half the time, indicating that the generator is producing realistic instances.

It was discovered that the pix2pix conditional GAN (cGAN), which offers a method to carry out picture translations (such as converting B&W to color images using GAN, was very enticing. This technique can be used for a reverberation challenge; instead, use it for a noisy voice improvement test. A generator (G) that improves the spectrogram (U-Net Image-2-Image Architecture) and a discriminator (D) that was trained to differentiate between the output of G and a clean spectrogram made up the cGAN. Two images are given to the discriminator. The first is the result of G, or a clear image, and the second is a conditional noisy spectrogram.

$$L_{\{GAN\}(G,D)} = \sum_t \left( log\ D(Z_t, X_t) + log\left(1 - D\left(Z_t, G(Z_t)\right)\right)\right)$$

The objective of G during training was to become better so that D would be unable to tell the difference between the output of G and the clean spectrogram. The objective is such that $Z_t$, $X_t$ and $X^t = G(Z_t)$ are the t-th example of the reverberant, clean and enhanced log-spectrum images respectively. As a regularization term that makes sure the enhanced speech is close to the clean speech, the MSE loss was added to the GAN loss in order to improve the outcomes. The final GAN score was therefore given as

$$L(G, D) = L_{\{GAN\}(G,D)} + \lambda L_{\{MSE\}(G)}$$

where λ is the weight of the direct MSE loss. The U-Net weights were used to initialize the GAN network, which was subsequently trained for a few additional epochs. According to empirical research, the value λ = 2000 produces good outcomes.

### *4.2.4 Long-Short Term Memory (LSTM)*

One type of recurrent artificial neural network is the LSTM. The results of the previous stage of an RNN are fed into the current phase. It tackled the problem of RNN long-term dependence, where the RNN may generate more accurate forecasts based on current information but is unable to forecast words held in memory over time. RNN fails to operate efficiently as the separation length rises. By default, LSTM has a long storage length for the information.

It is applied to time-series analysis, prediction, and categorization. The essential component to LSTMs is the straight line that traverses the top of the graphic and symbolizes the cell state. In certain respects, the cell state is similar to a conveyor belt. It moves effortlessly through every link with only a few modest straight interactions. It is fairly simple for information to flow along it unmodified. By the precise control of gates, the LSTM may add or remove data, changing the cell's state. Gates enable information to flow across only willingly. They are formed up of a layer of sigmoid neural pathways and a point-by-point multiplying algorithm.

### *4.2.5 Late Suppression U-Net*

This model differs from previous attempts in that it uses a Late Resonance Suppression (LS) U-net solution. This model far surpasses the conventional U-net by popular clarity, quality, and reverb measurement accuracy (e.g., speech-to-reverberation modulated energy proportion, or SRMR). In addition, it achieves dereverberation signs similar to the latest iteration of the U-net design developed with GANs.

The primary distinction between both models is the skip connection between the input and the output, which has been removed from the initial dereverberation U-net. This skipped conjunction lets us to focus on late reverberation reduction

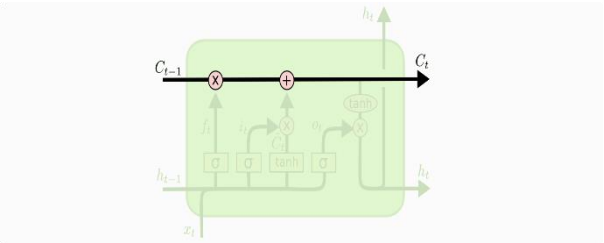


**Figure 3: Long-Short term Memory**

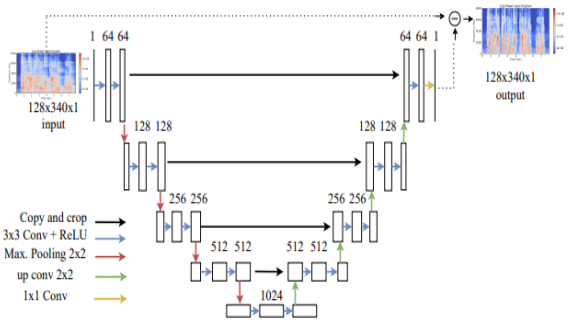


**Figure 4: Architecture of LSTM**

rather than straightforward dereverberation, since the U-net architecture finds out a mapping to dereverberate to late reverberation inhibition instead of a mapping from reverberated to dereverberation spectroscopy.

### *4.2.6 Late Suppression LSTM*

Dereverberation involves the use of long-term past information. Recurrent Neural Networks (RNNs) are designed to simulate sequential data with enduring dependencies, thanks to their internal memory. However, optimizing a basic RNN can be challenging due to issues related to gradient vanishing and explosion. Long Short-Term Memory (LSTM) RNNs, which utilize memory cells and gating mechanisms to control information flow, have demonstrated an impressive capacity to represent long-term dependencies within sequential data. Consequently, LSTM RNNs are employed to forecast late reverberation, capturing the rich history of previously recorded reverberant speech.

The input, forget, cell, and outputs gates are denoted by the parameters it, ft, gt, and ot in the following formulas, which define the LSTM block used in this study; The hidden state can be represented by ht at the time step t, the cell's memory state by ct, the input from the first level or the secret state of the layer before it by xt, the biases and weights used in the transformations that are linear are represented by W0 s and b 0 s, accordingly, and the element-wise multiplication is indicated by ◦.

$$i_t = \text{sigmoid}\left(W_{\{ii\}x_t} + W_{\{hi\}h_{\{t-1\}}} + b_i\right)$$

$$f_t = \text{sigmoid}\left(W_{\{if\}x_t} + W_{\{hf\}h_{\{t-1\}}} + b_f\right)$$

$$g_t = \tanh\left(W_{\{ig\}x_t} + W_{\{hg\}h_{\{t-1\}}} + b_g\right)$$

$$o_t = \text{sigmoid}\left(W_{\{io\}x_t} + W_{\{ho\}h_{\{t-1\}}} + b_o\right)$$

$$c_t = f_t \circ c_{\{t-1\}} + i_t \circ g_t$$

$$h_t = o_t \circ \tanh(c_t)$$

The system schematic for the proposed method is illustrated above. To facilitate understanding of the system's three temporal stages, they are outlined as follows: the LSTM RNN, which consists of two hidden layers, receives direct input from the input features at each time step; a linear layer is constructed on top of the LSTM neural network to align the hidden states of the last layer with the late reverberation; rectified linear units (ReLU) are applied after the linear projection layer to ensure an accurate estimation of the late reverberation; subsequently, the magnitude spectrum of the reverberant speech is subtracted from the late reverberation prediction to produce the actual sound along with early reflections as the system's output. Notably, the components of the magnitude spectra are compressed using a cubic root function. In other words, a compact space defined by a cubic root is utilized for spectral reduction. Although late reverberation is not explicitly employed as the training target, this approach encourages the LSTM RNN to learn to predict it. When treated as a black box, the proposed method achieves frame-level mapping from the magnitude spectrum of the reverberant speech to the spectrum of the actual sound plus early reflections, resembling an ordered spectral mapping.

## 5. RESULTS AND CONCLUSION

### 5.1 Evaluation metrics

The performance of the models was evaluated using the following metrics:

**PESQ:** Perceptual Evaluation of Speech Quality

**LLR:** Log-Likelihood Ratio

**CD:** Cepstral Distance

**fwSNRseg:** Frequency Weighted Segmental SNR

**SRMR:** Speech to Reverberation Modulation Energy Ratio

The first four metrics are invasive measures that use an assessment of an input signal's distortion and reverb level to a clean signal to determine "similarity" scores. Conversely, the SRMR metric is a metric that was developed by the use of an envelope filter bank that was motivated by the functioning the cochlea to evaluate the important band temporal envelopes of the speech signals. For an accurate assessment of the methodologies in consideration, it is crucial to use this last non-intrusive indicator since clear signals that may be used as standards in usage may not always be available.

### 5.2 Experimental Settings

The FFT was applied in each example to generate spectra with a window width of 2048 samples and an overlap of 512 samples. A Mel filter bank was utilized to reduce the bin size, resulting in either 128 or 256 bins in the testing setup. The time signal was successfully recovered in both cases; however, it was challenging to achieve this with fewer bins (e.g., 64 bins). Consequently, 128 bins were selected, and the number of frames for each spectrogram was set to 340, which represented the average frame count across all training spectra, using OpenCV's Lanczos approximation. An initial batch size of sixteen was employed, and the Adam optimizer was utilized for the learning process. Specifically, the U-Net GAN was trained with a λ value of 1e-2, which was chosen to ensure consistent size order across the Mean Squared Error (MSE) and the Least Squares Generative Adversarial Network (LGAN).

### 5.3 Simulation Results

| | **PESQ** | **LLR** | **CD** | **fwSNRseg** | **SRM** |
|---|---|---|---|---|---|
| reverberant speech | 1.90 | 1.31 | 7.11 | 6.34 | 3.08 |
| FD-NDLP (WPE) | 2.09 | 1.39 | 7.45 | 7.45 | 4.25 |
| UNet | 2.59 | 0.61 | 4.44 | 9.35 | 5.93 |
| UNet-GAN | 2.62 | 0.60 | 4.37 | 9.15 | 7.18 |

| | | | | | |
|---|---|---|---|---|---|
| Context-LSTM | 1.68 | 1.36 | 6.73 | 5.20 | 1.93 |
| Late Suppression LSTM | 2.38 | 0.81 | 4.97 | 8.06 | 4.53 |
| LS-LSTM + GAN | 2.42 | 0.53 | 4.32 | 9.36 | 6.57 |

## 5.4 Results on Real Dataset

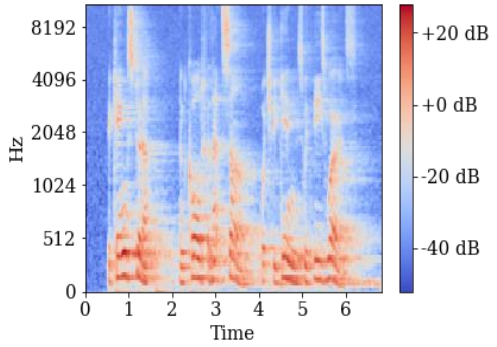


**Figure 6: Reverberant speech**

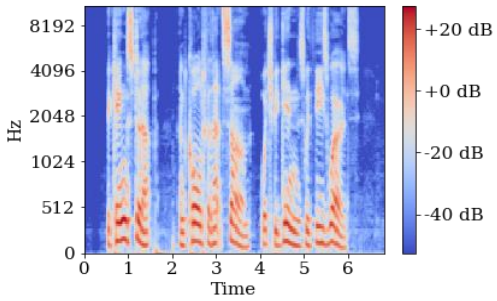


**Figure 7: LSTM+GAN**

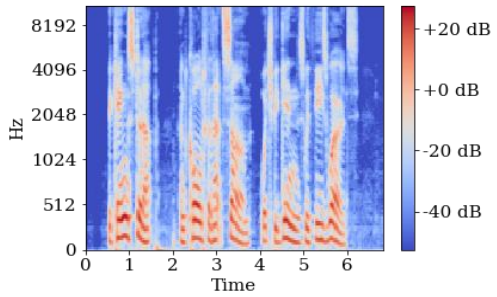


**Figure 8: Sample output of the dereverberation**

## 5.5 Conclusion

In this work, a novel algorithm, Late Suppression LSTM + GAN, is introduced to address challenges in enhancing speech quality within robust Automatic Speech Recognition (ASR) systems. The proposed model demonstrates significant improvements over existing approaches across multiple performance metrics, highlighting its effectiveness in producing clearer, high-quality speech outputs. Rigorous experimentation has shown that the Late Suppression mechanism effectively mitigates noise in the final stages of processing, while the integration with GANs facilitates realistic and natural speech generation.

The performance gains of the proposed model indicate its potential to advance the state-of-the-art in robust Automatic Speech Recognition (ASR) systems, paving the way for improved accuracy in real-world, noisy environments. Future work could focus on optimizing this model for computational efficiency, thereby enabling broader applications in resource-constrained devices.

# 6. FUTURE WORK

Future work will focus on optimizing the proposed model by increasing the number of training epochs and incorporating additional frequency bins. Further research will explore the use of vision transformers, a recent advancement in machine learning, to assess their performance in comparison to the current model.

# 7. REFERENCES

[1] K. Kinoshita κ.ά., 'The REVERB Challenge: A Benchmark Task for Reverberation-Robust ASR Techniques', στο New Era for Robust Speech Recognition, Springer, 2017.

[2] O. Ernst, S. E. Chazan, S. Gannot and J. Goldberger, "Speech Dereverberation Using Fully Convolutional Networks," 2018 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 390-394, doi: 10.23919/EUSIPCO.2018.8553141.

[3] Y. Zhao, Z. Wang and D. Wang, "Two-Stage Deep Learning for Noisy-Reverberant Speech Enhancement", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 1, pp. 53-62, 2019. Available: 10.1109/taslp.2018.2870725.

[4] IEEE Transactions on Audio Speech & Language Processing, 2010, 18(7):1717-1731.

[5] Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction[J].

[6] T. Nakatani, T. Yoshioka, K. Kinoshita, M. Miyoshi and B. -H. Juang, "Speech Dereverberation Based on Variance-Normalized Delayed Linear Prediction," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1717-1731, Sept. 2010, doi: 10.1109/TASL.2010.2052251.