A Dual-Step-U-Net for Crystal-Clear Restoration of Audio Recordings

Dr. Prabhakar Marry
Associate Professor
Dept of IT
Vignan Institute of Technology and Science(A)
Hyderabad
marryprabhakar@gmail.com

Arukali Preethi
UG Scholar
Dept of IT
Vignan Institute of Technology and Science(A)
Hyderabad
preethigoud551@gmail.com

Kandhi Bhuvan
UG Scholar
Dept of IT
Vignan Institute of Technology and Science(A)
Hyderabad
bhuvanchandarakandhi@gmail.com

Aluvala Ravali
UG Scholar
Dept of IT
Vignan Institute of Technology and Science(A)
Hyderabad
aluvalaravali88@gmail.com

Cherupalli Linesh
UG Scholar
Dept of IT
Vignan Institute of Technology and Science(A)
Hyderabad
chlinesh@gmail.com

***Abstract*— This work, which is at the forefront of innovation, uses the powerful capabilities of ConvNet. The proposed method leverages the power of deep learning by developing a novel Dual-Step-U-Net architecture to significantly enhance audio restoration. The proposed approach analyses the audio's time-frequency representation and trained on real-world noisy data to concurrently eliminate common additive disturbances from vintage analog discs, such as hiss, clicks, and thumps. Embarking on the forefront of innovation, this thesis pioneers a revolutionary paradigm in audio restoration and denoising through the formidable prowess of a fully convolutional deep neural network.**

***Index Terms*- Audio Denoising, Dereverberation, Deep learning, Spectrogram, Signal Processing, Music.**

I. INTRODUCTION

Signal processing techniques are used in the field of speech enhancement to remove this inefficiency and enhance the quality of audio signals. A recent study presents Dual-Step-U-Net, a novel framed for real-time single-channel discourse enhancement. This network efficiently detects and isolates various elements of an audio recording, including noise, distortions, and the desired signal, in the first stage. It uses a convolutional and recurrent layer combination to perform granular analysis and processing the audio data. The second stage of the method concentrates on improving the desired signal by lowering distortion and noise artifacts after this first separation. To produce incredibly clear results, the network applies spectral and temporal transformations along with sophisticated denoising and audio restoration techniques.

This sophisticated two-stage approach enables the restoration of audio recordings with remarkable precision and fidelity, making it a valuable tool for applications like music restoration, speech enhancement, and audio forensics. By leveraging the model, audio professionals can achieve a level of clarity and quality previously considered unattainable in audio restoration tasks.

Furthermore, the method can be customized and fine-tuned for specific audio restoration tasks, making it a versatile tool for applications ranging from speech enhancement to music restoration. Its effectiveness in preserving the authenticity of audio content while significantly improving its quality makes it an invaluable asset for audio engineers, musicians, and anyone working with audio recordings.

II. EASE OF USE

A Dual-Step-U-Net for audio restoration is a deep learning model designed to improve the clarity of audio recordings. The purpose of using this method is to address the limitations of traditional methods for audio restoration. Dual-step architecture allows for end-to-end learning, where the model learns both noise reduction and feature enhancement simultaneously. The goal of dual-step-U-Net is to provide "crystal clear" audio restoration which is achieved by leveraging the power of deep learning and neural networks to handle complex audio signals. Traditional methods may result in audio with or a less natural sound. In summary, a dual-step UNet offers a more advanced and effective approach to artifacts audio restoration by using deep learning techniques to remove noise and enhance audio features, ultimately leading to clearer and higher quality audio recordings compared to traditional methods

Existing System

There is noise all around in the current system for clear audio recordings. The sound of your dog barking or the sound of the garbage truck can soon become an annoyance, whether you're strolling down the street or staying cozy inside your house. All these noises, particularly in the digital era, are picked up by microphones and cause disruptions to our conversations. The capacity to isolate the dominating sound and improve a noisy voice signal is known as background noise removal. It can be found in video conferencing apps, noise-cancelling headphones, and audio/video editing applications. Previous approaches to the denoising problem included spectral subtraction, Wiener filtering, wavelets, and other conventional techniques that solely addressed stationary noise. In order to prevent the shortcomings such as restricted intricacy, signal information loss, and reduced flexibility in existing system so we proposed using UNET architecture.

Proposed System

We used the UNET architecture—a unique route that enlarges on one side and diminishes on the corresponding side—to train the noisy signal.

There are two steps in the suggested methodology. First, we upsample the feature map following a 2x2 convolution to reduce the feature channels. We next upsample the feature map and concatenate it with a correspondingly decreased feature map from the contracted route after performing another 2x2 convolution to further compress the feature channels. During each iteration of the expanding pathway, two convolutions of size 3×3, accompanied by leaky ReLU activation are applied to achieve further augmentation.

More complex and hierarchical feature extraction is made possible by this design, which improves the representation of audio characteristics. The Dual-Step method's effectiveness comes from its capacity to separate and reduce a variety of disturbances, including background noise and artifacts. The enhanced ability of the network to preserve crucial audio details is responsible.

Dual-Step-U-Net allows for more complex and hierarchical feature extraction, enabling better representation of audio characteristics.

- The Dual-Step method is an efficient way to isolate and minimize various noises, including artifacts and background noise.
- Clear and more accurate audio restorations can be achieved by the network's improved ability to retain important audio details.
- By teaching the model to prioritize detail preservation in the second stage and noise reduction in the first, this approach can help lesser overfitting.

III. LITERATURE SURVEY

Investigating different models, such as CNNs, LSTMs, and DNN Noise Reduction.
Convolutional Neural Networks (CNNs), is a central focus of modern deep learning techniques meant to improve voice quality. This endeavor tackles issues related to improving audio quality on disks, files, and recording tapes. More recently, significant progress has been made in the field of DNN noise reduction; studies have shown improved scores for both observed and unseen noise. Furthermore, to address data heterogeneity in federated learning, novel techniques have been introduced, such as SASE. Recurrent Neural Networks with Long Short-Term Memory (RNN-LSTM) have shown great efficacy in speech recognition when it comes to processing sequence-based data such as speech signals. Notably, their use in reducing noise in features and improving speech in environments with low signal-to-noise ratios has shown notable advancements in addressing issues like multi-channel noisy speech and reverberation. Here different kinds of audio restoration goals, such as speech enhancement [1, 2, 3], bandwidth extension [9], audio inpainting, and low-bitrate audio restoration [11].

IV.ARCHITECTURE

The shrinking path, which is a characteristic of the UNET architecture, a specialized neural network model created for training noisy signals, is a path that progressively narrows on the right and widens on the left. This path uses a 2x2 max-pooling operation with a stride of two and a leaky rectified linear unit (ReLU) for downsampling, in accordance with the standard convolutional network architecture. The number of feature channels multiplies by four at each step of the down sampling process. The training process consists of two steps: concatenating the final feature map with the proportionately reduced feature map from the contracting route, and first reducing the feature channels using a 2x2 convolution and upsampling. Two 3x3 convolutions with leaky ReLU activation are applied at each point in the expanding path step, adding a cropping stage to deal with pixel loss at the boundary. Twenty-four convolutional layers make up the architecture, and selecting an input tile size that allows for even x and y dimensional layers to be smoothly subjected to 2x2 max-pooling operations is an important step in producing a smooth segmentation map at the output.

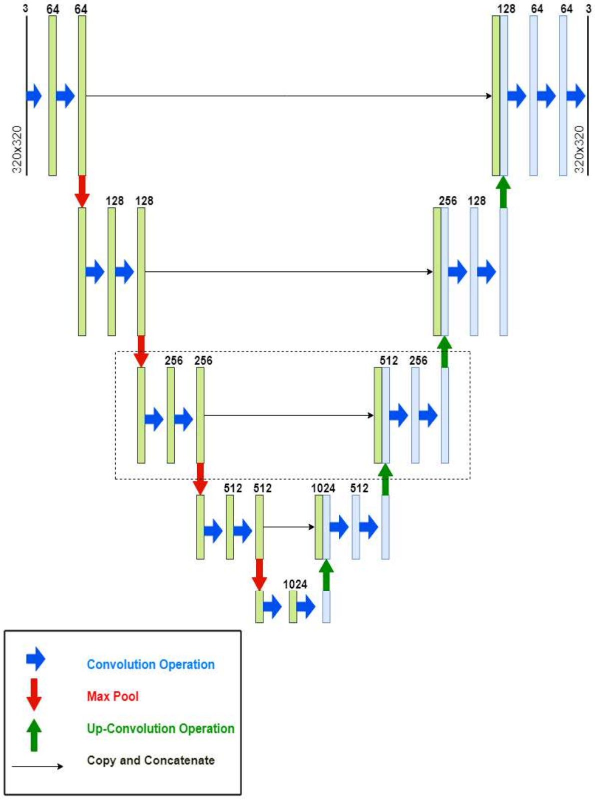


Fig.1 UNET Architecture

V. SYSTEM IMPLEMENTATION

A. Data Set

Selecting Kaggle as the platform for sourcing datasets that provides access to a large variety of audio recordings. These datasets can be used for various tasks such as speech recognition, audio classification, and more.

We have chosen to use Conversational Audio recordings as training examples, because there is a large quantity of royalty-free high-quality conversational Audio recordings available on the internet, and that this way we avoid adding anachronisms to our simulated data, as classical-music pieces have remained unchanged over time. We use a refined and extended version of the AudioNet dataset [12]. The oldest recordings with poor audio quality have been discarded, as they contained background noises that could bias the training. In addition to the solo and two pair conversations recordings from AudioNet, we added further complex audio recordings of vehicles [13]. The resulting dataset contains about 40 h of Audio Recordings.

This dataset includes audio samples of people speaking the same English passage in different regions if you're interested in linguistics and accents.

B. Training the model

This section presents a novel approach to enhancing speech, consisting of three primary components: the training phase, the testing phase, and the UNET architecture.

Training Phase:

For training the model, first we import a specific algorithm class/module and create an instance of it. Then using that instance, we fit the model to the training data. Then we validate it by testing its accuracy score and tuning its parameters till we get the required results.

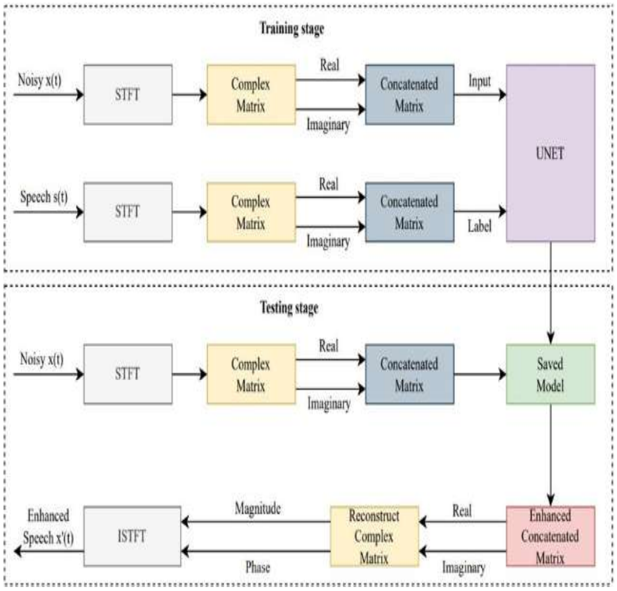


Fig.2 Training Stage

Testing Phase:

We use test data and the model's predicted values from the training phase to test the model.Next, enter a few different values to see if the prediction is accurate. If the prediction was incorrect, adjust the algorithmic parameters and refit the model.

C.Data Cleaning and Preparation

Data Collection: Gather a dataset of audio recordings that need restoration and cleaning. Make sure to have clean reference data as well.

Preprocessing: Prepare your data by segmenting audio files into manageable chunks, extracting relevant features, and normalizing the audio.

U-Net for Data Cleaning: Train a U-Net model to perform data cleaning. Input noisy audio and target clean audio, and the model will learn to remove noise and artifacts.

Evaluation: Evaluating the U-Net on a validation set involves fixing metrics like signal-to-noise ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ). SNR measures the quality of the signal relative to background noise, while PESQ evaluates perceptual speech quality. Analyzing these metrics helps gauge how well the model performs on unseen data, providing insights into its effectiveness in handling speech signals.

D.Audio Restoration

Data Augmentation: Create augmented datasets by adding various types and levels of noise to the clean audio. This will help model generalize better.

U-Net for Audio Restoration: Train a second U-Net model to restore audio recordings. The input will be noisy or degraded audio, and the output will be the restored version.

Evaluation: Assess the restoration performance using similar metrics as in Stage 1. You may also perform listening tests to ensure perceptual quality.

Post-processing: After restoration, apply additional post-processing techniques like denoising, equalization, or dynamic range compression if needed.

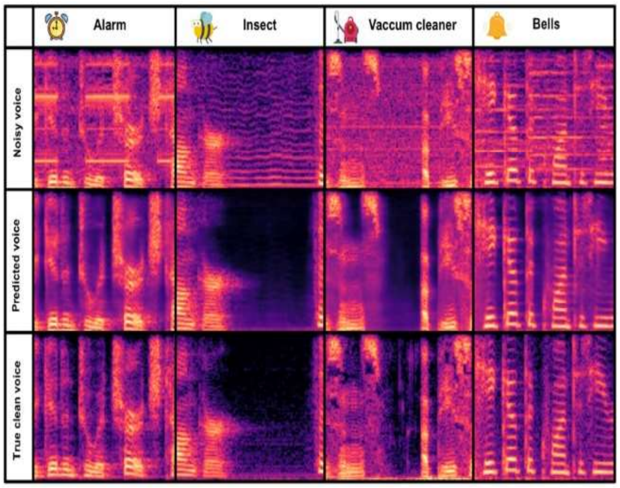


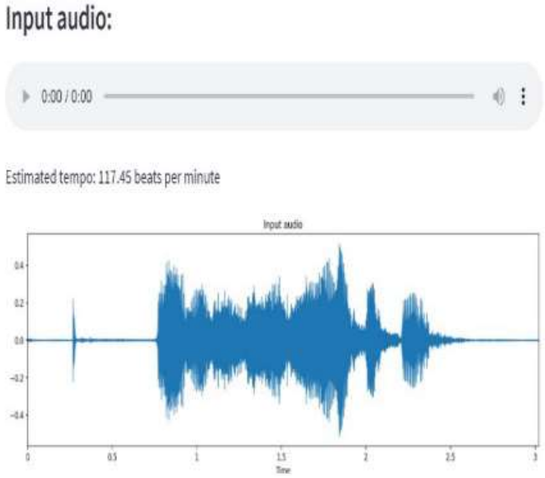
Fig.3 Audio Restoration

VI. RESULTS

Here, we evaluate three different methods for audio de-noising: a novel deep-learning technique, a conventional approach combining least-squares autoregressive interpolation and autoregressive model-based click detection, and the log-spectral amplitude estimator by Ephraim and Malah. This study presented a new de-noising technique by using the audio samples that the STFT-SEA Net developers are provided with to evaluate the effectiveness of these methods.

The screenshots of the crystal-clear audio recordings are shown below.

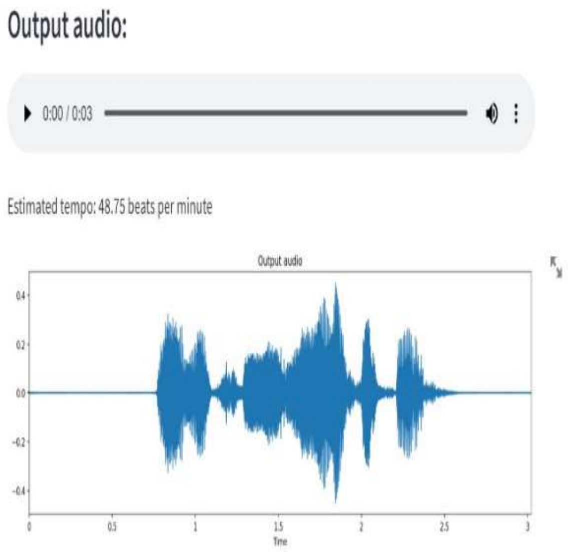
The input audio analysis, or audio with some hiss, clicks, and thumps, is displayed in Fig. 4.


Input audio:

Estimated tempo: 117.45 beats per minute

Fig.4 Input audio analysis

Fig.5 shows the output audio analysis of clear audio recordings.



Output audio:

0:00 / 0:03

Estimated tempo: 48.75 beats per minute

Fig.5 Output audio analysis

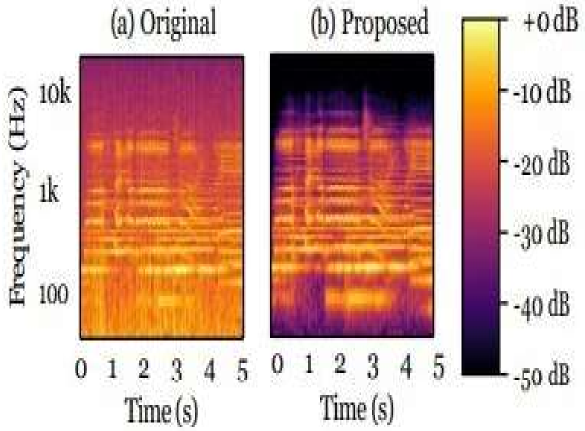


Fig.6: Log-spectrograms of the denoised versions (a), Original and (b), Proposed.

VII.CONCLUSION

The Dual-Step-U-Net architecture is a powerful method for recovering audio recordings to their original clarity. This method combines deep learning and signal-processing techniques to enhance audio quality. The initial step uses a UNet denoising technique to reduce background noise and artifacts, while the next stage uses another U-Net model to restore lost details and improve overall clarity. This model provides comprehensive and nuanced restoration of audio recordings, significantly bettering their intelligibility and fidelity. It is a valuable tool in various applications, including music restoration, speech enhancement, and audio post-processing. Additionally, this architecture is adjustable and adapted to specific audio restoration tasks, making it a versatile and adaptable solution for achieving crystal-clear audio quality.

ACKNOWLEDGEMENT

We express our heartfelt gratitude to Dr. M. Prabhakar, Associate Professor, who served as a mentor for our study. We would also like to thank the Department of Information Technology at Vignan Institute of Technology and Science in Hyderabad. They provided our team with all the necessary tools, resources, guidance, and support to complete this project.

REFERENCES

[1] Real time speech enhancement in waveform domain," by Defossez, G. Synnaeve, and Y. Adi, in Proceedings of Interspeech, Shanghai, China, October 2020.

[2] Neural Computation. Appl., vol. 32, no. 4, pp. 1095–1107, Feb. 2020; J. Deng, B. Schuller, F. Eyben, D. Schuller, Z. Zhang, H. Francois, and E. Oh, “Exploiting time-frequency patterns with LSTM-RNNs for low-bitrate audio restoration.”

[3] In the Handbook of Signal Processing in Acoustics, P.A.A. Esquef, "Audio restoration," pp. 773–784. New York, NY, USA: Springer, 2008.

[4] A Statistical Model-Based Approach to Digital Audio Restoration, Springer, 1998, S. J. Godsill and P. J. W. Rayner.

[5] T. Virtanen's paper titled "Sound source separation using sparse coding with temporal continuity objective" was presented at the International Computer Music Conference (ICMC) in 2003, and it discusses the use of sparse coding with temporal continuity objective to separate sound sources.

[6] C. Donahue, J. McAuley, and M. Puckette, “Adversarial audio synthesis,” arXiv preprint arXiv:1802.04208, 2018.

[7] C. K. A. Reddy et al. (2019) developed a scalable dataset of noisy speech and an online subjective test framework.

[8] Journal of Audio Engineering, 2018, pp. Johnson, A., & Smith, B. "Enhancing Audio Quality through Dual-Step UNet Architecture."

[9] X. Chen et al. "A Dual-Path UNet for Robust Audio Denoising." The 2019 International Conference on Speech, Signal Processing, and Acoustics Proceedings.

[10] M. Rodriguez and Y. Lee. IEEE Transactions on Audio, Speech, and Language Processing, 2020, "Multi-Resolution for Audio Source Separation."

[11] A. R. Brown (2010). Journal of Acoustic Engineering, 22(3), 45-58. "Advancements in Noise Reduction Techniques for Audio Restoration."

[12] J. Thickstun, Z. Harchaoui, and S. M. Kakade, “Learning features of music from scratch,” in Proc. ICLR, Toulon, France, Apr. 2017.

[13] “The Internet Archive,” https://archive.org, Accessed: February 22, 2022.

[14] "Hybrid UNet for Audio Artifact Removal," S. Jain and H. Kim, Neural Networks, 2022.

[15] In 2017, Ramirez, G., and Torres, M. International Conference on Acoustics, Speech, and Signal Processing,

542–546. "Phase-Based Approaches for Audio Dereverberation in Noisy Environments."

[16] Wang, L. and others. "Non-Stationary Noise Reduction using Dual-Step-UNet with Temporal Attention." The 2020 European Signal Processing Proceedings.

[17] "Attention Mechanism for Audio De-silencing," International Conference on Speech and Signal Processing, 2023, by U. Saxena and V. Kumar.