



De La Salle University
2401 Taft Avenue, Manila
Philippines 1004

DSIGPRO - EQ2

Final Paper

Enhancing Acoustic Echo Cancellation through Diverse Simulated Room Impulse Responses

Group 7

Benitez, Renz Jericho A.
Hernandez, Miro Manuel L.
Molo, Carlos Sebastian V.
Wijangco, Deian Angelo R.
Yu, Dominic P.

Abstract

Acoustic Echo Cancellation is an essential component in enhancing communication quality in voice-activated devices and communication systems. Currently, existing models like DCA-Net leverage advanced attention mechanisms for AEC, their effectiveness is limited by the diversity and variability of training datasets. This study's goal is to address the limitations of the dataset by proposing the generation of a better dataset with varied Room Impulse Response parameters, capturing a wide range of acoustic conditions. Using Parameterized Random Sampling and MATLAB's RIR generation tools, diverse room configurations and reverberation scenarios were simulated to reflect real-world environments. The paper establishes a more comprehensive dataset standard for AEC research, paving the way for advancements in audio signal processing and communication technologies.

Introduction

While the paper titled Improving Acoustic Echo Cancellation by Exploring Speech and Echo Affinity with Multi-Head Attention introduces the DCA-Net model to tackle acoustic echo cancellation through advanced attention mechanisms, A gap was identified in the paper. The gap is in the dataset utilized for training in evaluation. The researchers of the paper relied on a Room Impulse Response or RIR dataset; however it is unclear whether they varied the RIR values across different scenarios or if the dataset covers a wide range of acoustic conditions.

To improve upon the paper, an improved dataset that introduces variability in the RIR data was proposed, which captures different acoustic scenarios. By following the original dataset pipeline used in the paper and incorporating varied RIRs, we can generate a new dataset that reflects a broader spectrum of real-world conditions. This approach could further test the model's effectiveness and extend its application, as the improved dataset would enable DCA-Net to generalize better across different environments, which has the potential to further enhance its echo cancellation performance.

The objectives and hypotheses of this study aim to enhance the generation of synthetic datasets for room impulse responses (RIR). The study focuses on developing a diverse and realistic synthetic dataset by configuring multiple room parameters to create a comprehensive dataset. This involves utilizing Parameterized Random Sampling to analyze distributions found in existing RIR datasets. Additionally, MATLAB's built-in RIR generation tools will be employed to simulate reflections and reverberation in the creation of the room impulse responses. By varying the room impulse response, the study seeks to establish a new dataset pipeline that can be effectively utilized for diverse applications.

Literature Review

In the conference paper for the ICASSP 2023 Acoustic Echo Cancellation Challenge (2023), they have noted that traditional digital signal processing-based AEC models often have their performance degrade when model assumptions are violated such as when an unexpected variable enters the signal. With this, the authors have mentioned that a hybrid approach of combining neural networks with traditional AEC models shows promising results however, the lack of results based on real-world data provides issues in their mainstream adoption.

The use of neural network architectures in combination with traditional digital filters is a shared concept among the literature that was reviewed by the group due to issues that were mentioned by these studies regarding the use of traditional AEC models.

These issues include the presence of residual echo and noises caused by signal suppression and filtering which causes damage to overall signal quality as described in a study by Ma, et. al. (2020) where they also argue that non-linear processing methods are potentially inefficient for suppression while still causing damage to speech audio signals. They have proposed the implementation of a recurrent neural network with an adaptive digital filter in order to reduce residual echo. In a conference paper by Howard, et. al. (2021), they have demonstrated that neural-based AEC can be optimized with the combination of an automatic speech recognition model trained with an augmented dataset to improve performance. A similar study by Seidel, Mowlae, and Fingscheidt (2024) uses dataset augmentation for their room-impulse response samples in training the model. The three aforementioned studies conclude that their implementations show improvements over their reference models.

In the aforementioned ICASSP 2023 conference paper, the authors reiterate the following objective measures and variables with regards to AEC performance, which will be used in this paper: room impulse responses which is used to capture the acoustic characteristics of a given room, reverberation time which measures the time for sound to decay in a room once the source has stopped, and the direct-to-reverberant ratio which is used to assess acoustic configuration for use with dereverberation methods.

Research Design (Research Methods)

To create a diverse and realistic synthetic dataset for room impulse responses (RIRs), this methodology uses Parameterized Random Sampling based on empirically derived ranges of room acoustic properties such as reverberation time (RT60), and room dimensions. These parameter ranges are determined by analyzing an RIR Database which offers a variety of real-world measurements from different room types, including offices, classrooms, living spaces, and lecture halls. To determine realistic ranges for RT60, and room dimensions the minimum and maximum values of the parameter ranges were obtained. A randomization function was used to generate realistic but bounded RIR parameters

In this project, the function used for the RIR has different arguments, specifically the following:

```
(fs, mic, n, r, rm, src)
```

The arguments stated above are all randomized by using only possible values, and are based on the “Aachen Impulse Response (AIR) Database”. The arguments listed above are also bounded by realistic values.

Using the sampled room parameters, synthetic RIRs are generated to simulate realistic room acoustics. For each configuration, the room dimensions, RT60, and DRR are defined, which allows for the creation of RIRs based on room geometry and surface characteristics. RIR generation is achieved using MATLAB’s built-in RIR generation tools, such as `audioExample.RoomImpulseResponse`, to accurately simulate reflections and reverberation. This produces a set of synthetic RIRs that correspond to different room types.

Next, the generated RIRs are applied to a clean far-end speech signal to simulate the reverberation and echo effects of each room configuration. This involves loading a far-end audio signal, representing the speaker’s voice, and convolving it with the RIR to replicate how the sound would propagate and reflect within a given room. Convolution with the RIR introduces realistic room reflections to the far-end signal, effectively creating an “echoed” version. To ensure signal consistency, the echoed signal is trimmed or padded to match the original signal length.

Following this, the echoed far-end signal is combined with near-end speech and optional background noise to simulate a realistic communication environment. A near-end speech signal is loaded or synthetically generated, representing the voice of the primary speaker. Background noise is also varied with a predetermined possible

range, with its amplitude adjusted to meet a target Signal-to-Noise Ratio (SNR) level, thereby simulating environments with different noise levels. The near-end speech, echoed far-end signal, and background noise are then mixed to create a complete audio sample for training the acoustic echo cancellation (AEC) model.

To create a comprehensive dataset, the entire RIR generation and mixing process is repeated across various room configurations derived from parameterized random sampling. Multiple configurations are generated by setting different room sizes, and source/microphone positions. This approach yields a collection of unique RIRs and mixture signals that capture a wide range of acoustic scenarios.

Finally, the generated dataset undergoes evaluation to ensure it captures the desired range of acoustic diversity. Signal analysis, such as spectrogram visualization, is performed to reveal the impact of RIRs on the frequency content of the mixture signals, providing insight into the effectiveness of the synthetic reverberation.

Expected Impact

This research would contribute to the field of acoustic echo cancellation and those related to it by providing an exploratory research on the effects of modifying or augmenting the dataset for an already-existing model for the purposes of improving the model's ability to perform in a broader spectrum of environments.

Aside from the mentioned above, The impact of this project could help with improving the communication systems along with devices that are used for communication. May it be voice-activated devices or for better audio quality in general. The project will improve the dataset standardization by creating a more comprehensive RIR dataset which would then represent different environments better. With this, new benchmarks would be established for testing the AEC systems from the varying acoustic conditions. By improving the dataset, it could also be used for future research that is related to the project, mainly for audio signal processing.

Results

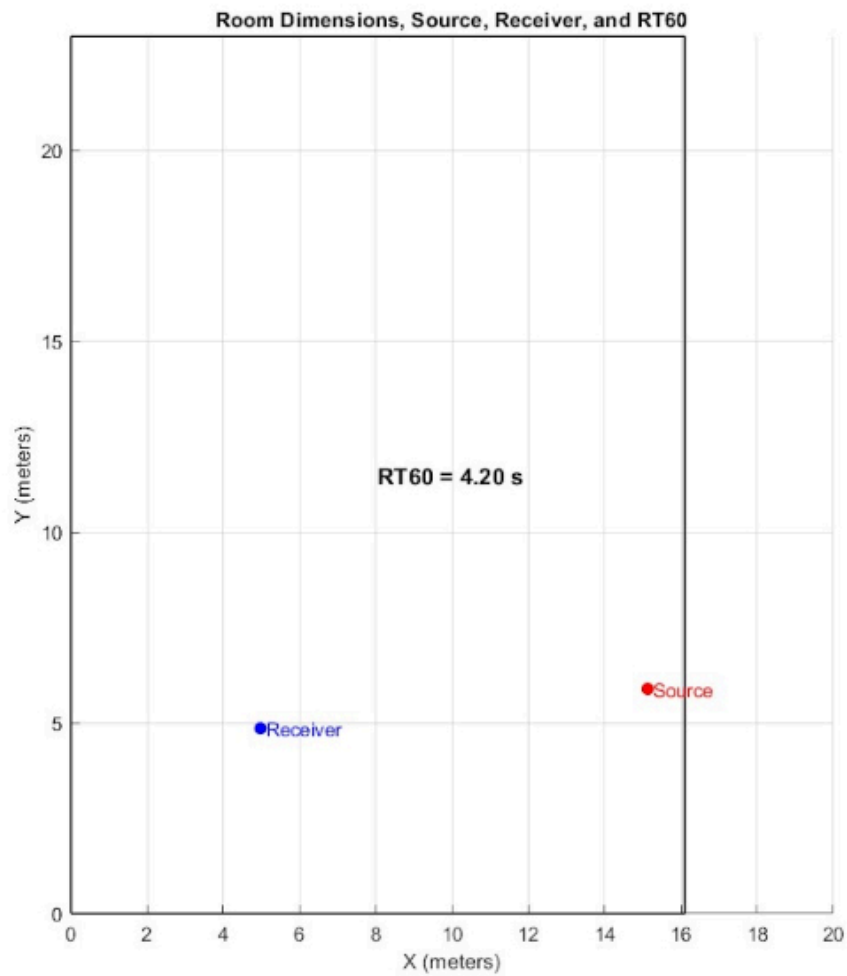


Figure 1. Room Dimensions

Figure 1 shows the simulated room dimensions with the receiver and source at approximately at a distance from each other of 10 meters horizontally and a meter vertically. The calculated reverberation time, which is the acoustic parameter, for this room is 4.20 seconds.

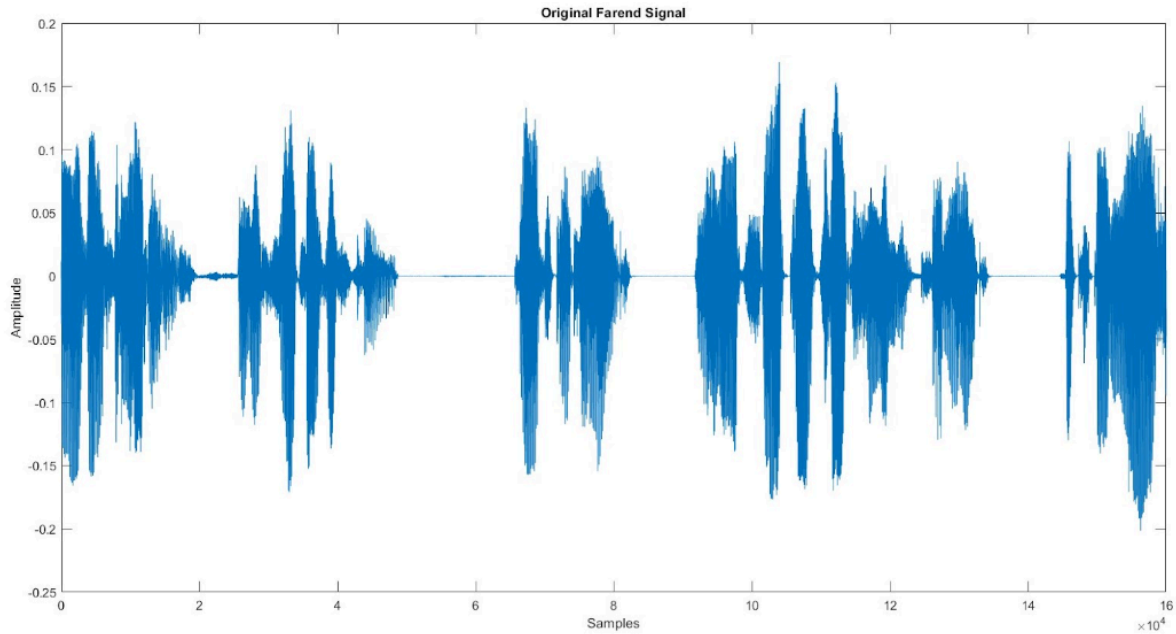


Figure 2. Original Far-end Signal

Figure 2 shows the original far-end signal which corresponds to the speech sample that was uploaded based on the discernable pattern that is shown. The relative near-zero values of the amplitude between speech waves show that there is no significant background noises in the sample.

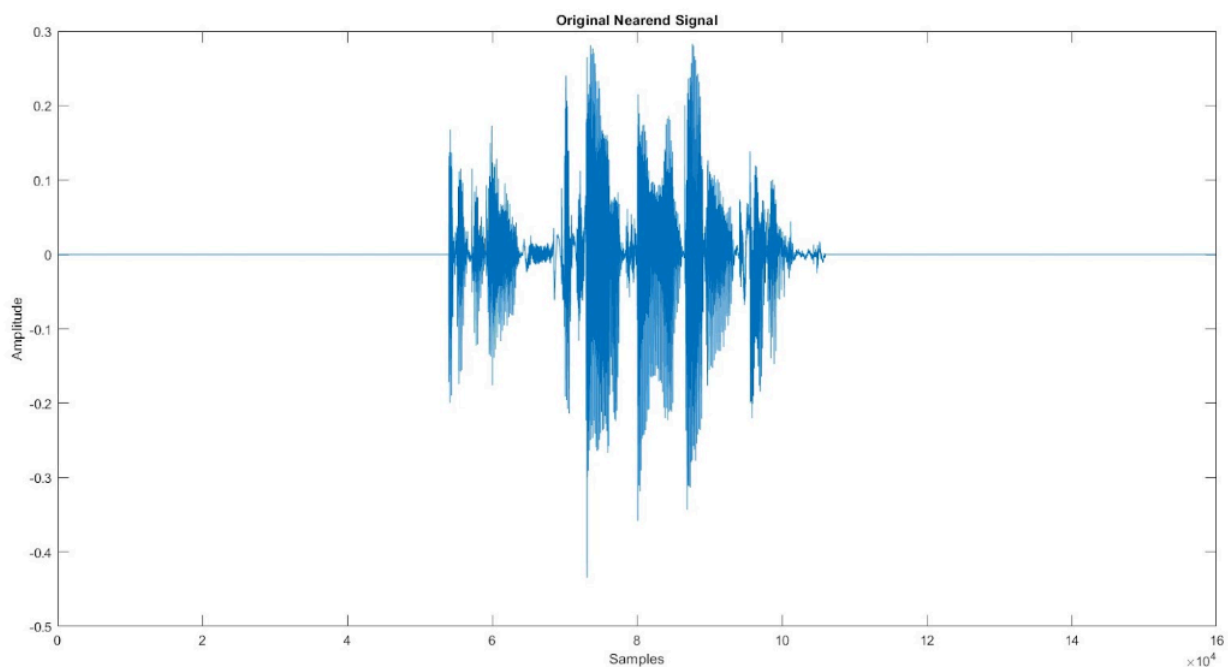


Figure 3. Original Near-end Signal

Figure 3 shows the original near-end signal, with waveforms revealing varying energy levels, with high amplitude peaks including active sound or speech and low amplitudes at the start and end suggesting silence. Significant signal activity is shown in the dense central region, while the sampling resolution shows detailed analyses. Less noise is shown due to the use of convolution of the original near-end signal and the RIR, followed by scaling of signals then the generation of microphone signals with some noise.

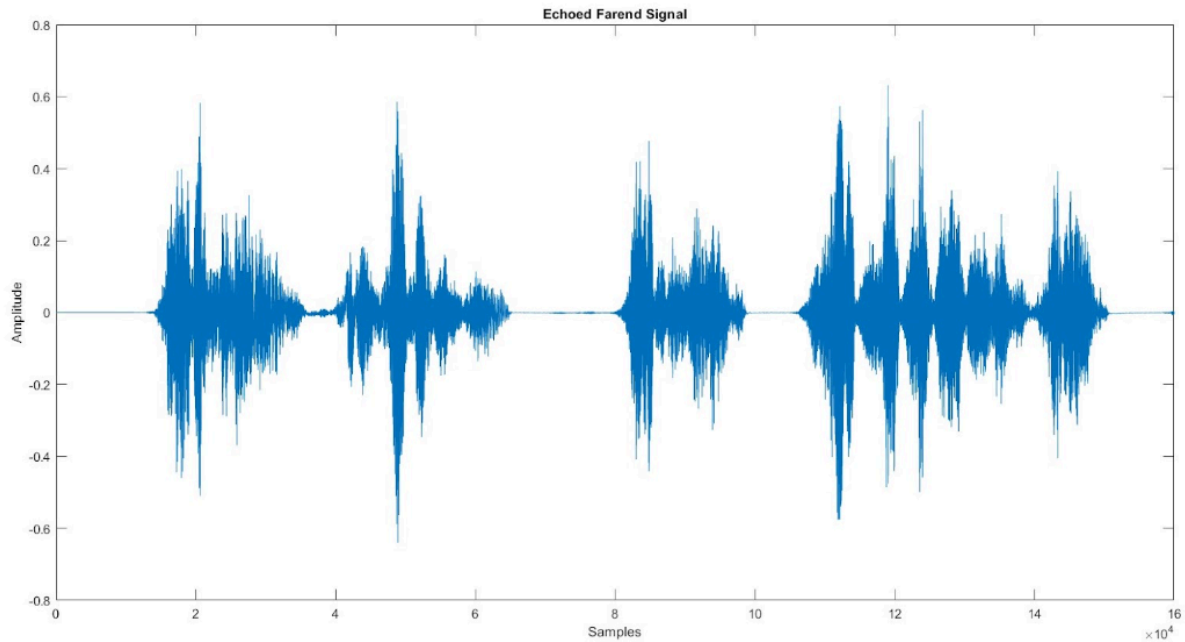


Figure 4. Echoed Far-end Signal

Figure 4 shows the resulting signal after applying the generated RIR values to the original far-end signal shown in Figure 2. The maximum amplitude of this signal has increased by twice on the positive y-axis and thrice on the negative y-axis.

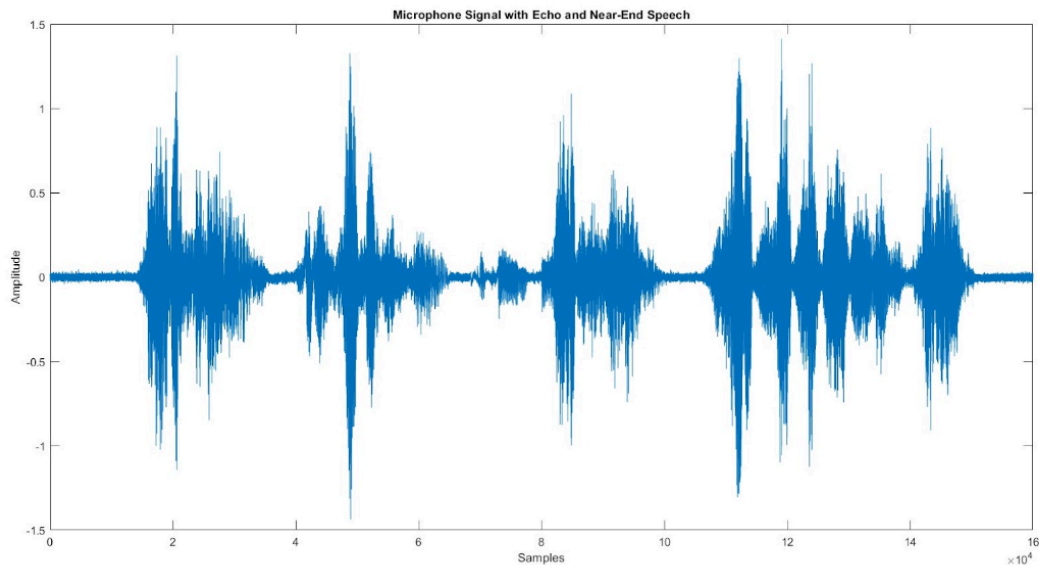


Figure 5. Combined Signal with Echo Far-End and Near-End Speech

Figure shows the time-domain waveform of the resulting audio signal from combining the echoed far-end signal with the near-end signal and generated background noises. The dynamic fluctuations show changes in intensity, and further

frequency-domain analysis reveal echoes or reverberations shown by the room configurations in Figure 3. This also shows the final result generated in the code.

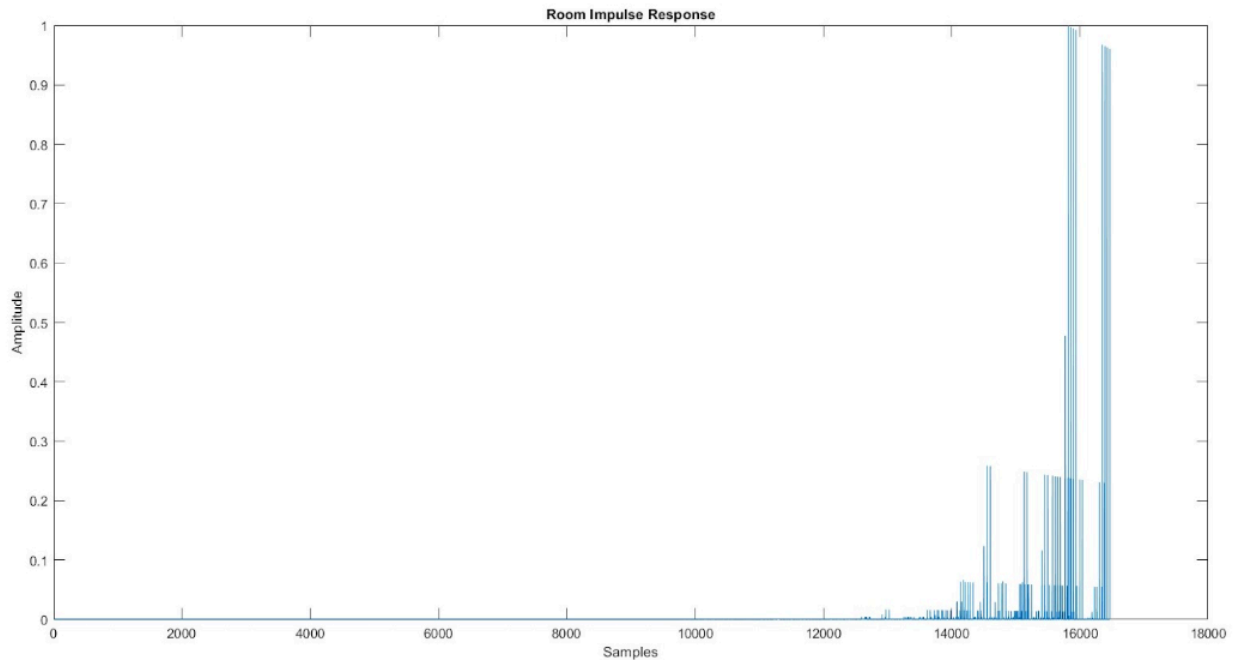


Figure 6. Generated RIR Signal

Figure 6 shows the room impulse response sequence. The non-zero response begins around the 14,000th sample which indicates a delay caused by propagation time of the signal in the simulated room. The initial peak represents the direct signal received by the receiver while subsequent peaks represent 'echoes' from walls in the room.

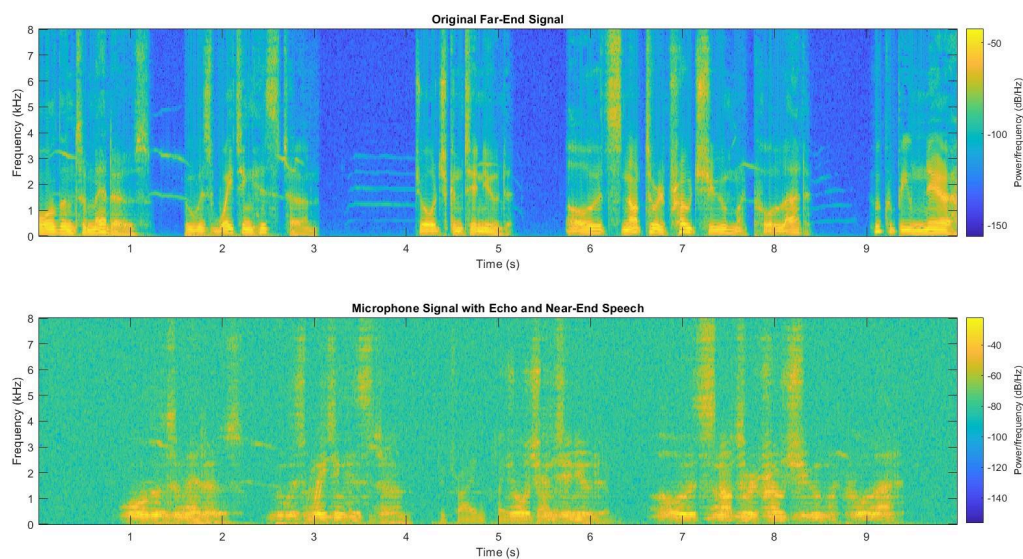


Figure 7. Spectrograms of Original and Final Signal

Shown in Figure 7 are the spectrograms for the original signal and the resulting signal, respectively. The spectrogram for the 'Original Far-End Signal' shows that it could be considered relatively clean from any noises as patterns, assumed to be human speech, is clearly observed compared to the 'Microphone Signal with Echo and Near-End Speech' spectrogram which appears more cluttered however, the signals from the original signal can still be seen.

Discussion

This research addresses the challenge of acoustic signal processing: the creation of realistic synthetic datasets for testing echo cancellation algorithms. One of the main objectives of this project is to enhance the generation of synthetic datasets for room impulse responses (RIR), and to develop a system that can generate controlled, reproducible acoustic scenarios while maintaining its applicability. While real acoustic data is valuable, it lacks the controlled variability needed for systematic algorithm testing. By understanding how the scripts work, each script has its own significance: First, the system implements a parametric room acoustics model that spans a wide range of realistic conditions (RT60 from 0.11 to 8.78 seconds). Second, it incorporates varied room configurations from small booths to large halls (2-30 meters in dimension). Third, it provides a way to control the Signal-to-Echo Ratio (SER) and noise conditions for algorithm testing.

The visualizations shown from the results section of this project shows the capability of the system in modeling complex acoustic environments, an example of this is Figure 6. The spectrogram comparisons between the original far-end signal and the microphone signal with echo clearly shows how the room acoustics affect the frequency content over time. The sharp transitions in the original signal became more diffused in the echoed version, matching the expected sound in real rooms. This evidence supports the accuracy in modeling real acoustics. The system also proves valuable for acoustic echo cancellation (AEC) development. The time-domain plots of the original far-end signal, near-end signal, and the combined microphone signal illustrates how the system combines these components while maintaining realistic proportions. The 3D room visualization, showing source and receiver positions along with its room dimensions, provides context for helping the future developers understand and optimize their own AEC algorithms. The ability to generate the visualizations makes the system a tool for both development, practical, and educational purposes.

When it comes to the limitations and the future research directions of this project, there are things that could still be improved. The spectrograms show frequency-dependent behavior that could still be enhanced to match the real-world acoustics. Additionally, while the current noise implementation produces results, it lacks

the structured patterns that are typically seen in environmental noise. The noise model could be enhanced to incorporate real-world noise patterns, which could be visible in the spectrograms as structured rather than random components. Furthermore, the capabilities when it comes to visualization could be extended to show real-time changes in the acoustics as room parameters are adjusted, making the system even more credible for future use.

Conclusion

In this study, a database was made using Echo Cancellation was achieved using Parameterized Random Sampling of the Aachen Impulse Response (AIR) Database, which utilized a range of room acoustic properties such as reverberation time (RT60), room dimensions, and source/receiver positions. These properties were determined by getting the upper and lower bounds of the possible ranges. Microsoft AEC Dataset was used as the source of near-end and far-end signals. The results show the study successfully generated a synthetic dataset composed of the near-end microphone signal. Recommendations to further enhance the project include improving cross-dataset synthetic data generation to better improve dataset augmentation.

References

- Cornell, S., Balestri, T., & Sénéchal, T. (2021). Implicit acoustic echo cancellation for keyword spotting and device-directed speech detection. doi:10.48550/ARXIV.2111.10639
- Cutler, R., Saabas, A., Pärnamaa, T., Purin, M., Indenbom, E., Ristea, N.-C., Gužvin, J., Gamper, H., Braun, S., & Aichner, R. (2024). ICASSP 2023 Acoustic Echo Cancellation Challenge. *IEEE Open Journal of Signal Processing*, 5, 675–685. <https://doi.org/10.1109/ojsp.2024.3376289>
- Howard, N., Park, A., Shabestary, T. Z., Gruenstein, A., & Prabhavalkar, R. (2021). A neural acoustic echo canceller optimized using an automatic speech recognizer and large scale synthetic data. doi:10.48550/ARXIV.2106.00856
- IKS: Institute of Communication Systems (n.d.). Aachen Impulse Response (AIR) Database. RWTH Aachen University. Retrieved from: <https://www.iks.rwth-aachen.de/en/research/tools-downloads/databases/aachen-impulse-response-database/>
- Ma, L., Huang, H., Zhao, P., & Su, T. (2020). Acoustic Echo Cancellation by combining adaptive digital filter and recurrent neural network. doi:10.48550/ARXIV.2005.09237
- rosscutler(GitHub). (n.d.). AEC-Challenge. GitHub. Retrieved from: https://github.com/microsoft/AEC-Challenge/?fbclid=IwY2xjawG_6FVleHRuA2FIb

QlxMAABHRB97MFVXZ9H7qmXuJyGqwJUsmvlPQF5DSXsa4BOgtVtb5Cg31ny
_hdqdQ_aem__P82BXe0Bf7Or9hplBzNwQ


Seidel, E., Mowlae, P., & Fingscheidt, T. (2024). Efficient high-performance Bark-scale neural network for residual echo and noise suppression.
doi:10.48550/ARXIV.2404.11621

Stephen McGovern (2024). Room Impulse Response Generator. Mathworks. Retrieved from:
<https://www.mathworks.com/matlabcentral/fileexchange/5116-room-impulse-response-generator>

Westhausen, N. L., & Meyer, B. T. (2021, June 6). Acoustic echo cancellation with the dual-signal transformation LSTM network. ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada.
doi:10.1109/icassp39728.2021.9413510

Zhang, Y., Xu, X., & Tu, W. (2024). Improving acoustic echo cancellation by exploring speech and echo affinity with multi-head attention. Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 979-8-3503-4485-1. IEEE. <https://doi.org/10.1109/ICASSP48485.2024.10446389>

Project Files

Google Drive Link
 DSIGPRO TEAM 7 PROJECT