# Enhancing Speech Quality in Noisy Environments

**Shreyas Srinivas Bikumalla, Amith Reddy Atla**

## ABSTRACT

This project introduces a novel two-stage deep learning model employing Conv-TasNet, enhanced with dynamic bucket batching and sophisticated data augmentation, to address the challenges of noise and reverberation in speech enhancement. The model optimizes training efficiency and adapts to varying acoustic conditions by simulating diverse noise environments and reverberation levels, making it robust across both matched and mismatched scenarios. These features are crucial for applications requiring high clarity and intelligibility in adverse acoustic environments, such as automated speech recognition systems and assistive hearing devices.

The effectiveness of the model is rigorously evaluated using the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) metrics, which assess speech quality and intelligibility respectively. By excelling in these evaluations, the model demonstrates its capability to significantly improve speech clarity and comprehension in noisy conditions, showcasing a substantial advancement in speech enhancement technology. This project bridges theoretical advancements by offering substantial improvements in user experience for a wide range of communication technologies.

## 1. INTRODUCTION

Speech degradation caused by noise and reverberation significantly impairs effective communication, impacting both human listeners and automated systems such as automatic speech recognition (ASR) and voice assistants. For hearing-impaired individuals, maintaining clear speech signals is crucial for facilitating everyday interactions.

### Key Limitations of Traditional Approaches:

Traditional enhancement techniques, like spectral subtraction, often struggle to effectively address real-world audio challenges due to several key limitations:

- Assumptions of stationary noise, which is rarely the case in dynamic environments.
- Inability to effectively handle simultaneous noise and reverberation, common in many real-life scenarios.
- Loss of speech naturalness, resulting in outputs that sound unnatural or distorted.

**Advancements in Deep Learning:**

Deep learning models present promising solutions by:

- Directly processing time-domain signals, thus preserving the natural quality of speech.
- Learning complex patterns in noisy and reverberant environments, enabling them to adapt to a variety of acoustic conditions.
- Utilizing diverse datasets for training, which enhances their ability to generalize to new, unseen environments.

This project introduces a two-stage model designed to separately address noise suppression and dereverberation. By validating this model through rigorous experiments under both matched and mismatched conditions, the study aims to demonstrate significant improvements in speech intelligibility and quality, leveraging advancements in deep learning to overcome the deficiencies of traditional methods.

## 2. MOTIVATION

### Why Speech Enhancement?

Speech enhancement is essential for clear communication, particularly in noisy environments that significantly impede both human interaction and the effectiveness of voice-driven technology. In bustling urban settings, crowded public spaces, or noisy workplaces, the ability to clearly comprehend speech is crucial, not only for everyday communication but also for ensuring accessibility for individuals with hearing impairments. These environments challenge the effectiveness of speech-dependent technologies such as automatic speech recognition (ASR) systems and smart assistants, which rely on clear audio inputs to function optimally. Enhanced speech clarity reduces the cognitive strain on all listeners, particularly benefiting those with hearing difficulties by making speech more accessible and reducing auditory fatigue.

1. **For Human Communication**: Improves clarity in noisy environments and aids hearing-impaired individuals.
2. **For Technology**: Enhances the accuracy of voice-based systems like ASR, smart assistants, and telecommunication tools.
3. **For Accessibility**: Enables equitable communication in workplaces and public spaces.

### Why a Two-Stage Model?

The development of a two-stage deep learning model addresses this need by methodically tackling both noise and reverberation, the primary culprits behind degraded audio quality. The first stage of the model focuses on suppressing background noise, employing advanced algorithms to distinguish and diminish unwanted ambient sounds while preserving the integrity of the speech signal. Following noise suppression, the second stage targets reverberation, which if left unmanaged, can cause speech to sound distant or echoey, further complicating comprehension. By sequentially processing the audio to mitigate these distinct yet interrelated issues, the model not only enhances speech intelligibility but also restores its natural quality, making it sound more direct and clear. This strategic approach significantly improves the user experience across various applications, from telecommunication to assistive technologies, ensuring that speech enhancement technologies meet the demands of modern, noisy environments.

## 3. PRIOR WORK RELATED TO THIS TOPIC

### Assessing the Generalization Gap of Learning-Based Speech Enhancement Systems in Noisy and Reverberant Environments

- **Authors**: Gonzalez et al.
- **Summary**: This study explores the generalization gap in speech enhancement models like FFNN and Conv-TasNet under unknown noise and reverberation conditions. It finds significant performance drops in mismatched conditions and suggests that diversifying training datasets can help mitigate, but not completely eliminate, these effects. This research provides insights into enhancing model robustness against variable acoustic environments.

### On Batching Variable Size Inputs for Training End-to-End Speech Enhancement Systems

- **Authors**: Gonzalez and Alstrom
- **Summary**: This paper investigates different batching strategies for training models like Conv-TasNet on variable-length input sequences. It concludes that dynamic bucket batching optimizes training efficiency by reducing padding and computational demands, which is crucial for handling large datasets in speech enhancement systems.

### Two-Stage Deep Learning for Noisy-Reverberant Speech Enhancement

- **Authors**: Zhao et al.
- **Summary**: This research introduces a two-stage approach to speech enhancement that tackles noise and reverberation separately. The first stage uses deep neural networks for denoising, and the second stage focuses on dereverberation. The combined optimization of these stages leads to significant improvements in speech intelligibility and quality, outperforming single-stage models in real-world conditions.

## 4. ALGORITHMIC DETAILS

The proposed model leverages the Conv-TasNet architecture, renowned for its effectiveness in time-domain speech processing. This architecture is particularly adept at handling complex audio signals directly, offering significant advantages over traditional frequency-domain methods.

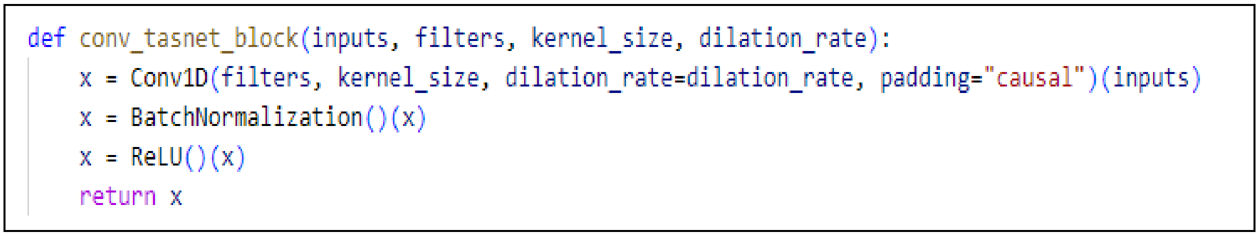
- **Dilated Convolutions:** A standout feature of Conv-TasNet, dilated convolutions allow the network to cover larger temporal areas without increasing computational burden. By skipping input values at regular intervals, these convolutions capture broader dependencies within the audio signal, crucial for distinguishing between speech and noise.

- **Batch Normalization:** Integrated into each convolutional layer, batch normalization stabilizes the learning process by normalizing the activations. This normalization counters the internal covariate shift changes in network activations due to parameter updates during training enabling higher learning rates and promoting faster convergence. Additionally, it acts as a regularizer, slightly perturbing the learning process to enhance model generalization and prevent overfitting.

- **ReLU Activation:** The model employs the Rectified Linear Unit (ReLU) activation function to introduce non-linearity, essential for learning complex patterns in audio data. ReLU maintains the efficiency and sparsity of the model by zeroing out negative values, facilitating faster learning and helping mitigate the vanishing gradient problem.

Together, these architectural elements enhance the model's capacity to process speech effectively, significantly improving clarity and reducing noise and reverberation in challenging acoustic environments.

**Snippet for Conv-TasNet Block**



```
def conv_tasnet_block(inputs, filters, kernel_size, dilation_rate):
    x = Conv1D(filters, kernel_size, dilation_rate=dilation_rate, padding="causal")(inputs)
    x = BatchNormalization()(x)
    x = ReLU()(x)
    return x
```

*Fig 1. Code snippet showing Conv-TasNet block.*

## 5. SOFTWARE IMPLEMENTATION DETAILS

The implementation of our speech enhancement model harnesses several pivotal libraries and frameworks, each selected for its specialized capabilities in deep learning and audio processing:

**TensorFlow and Keras:**

**TensorFlow** provides robust numerical computation, crucial for deep learning's intensive calculations, and supports GPU acceleration to speed up model training. Keras, TensorFlow’s high-level API, offers an intuitive interface for constructing neural networks, enhancing the ease of model architecture definition and experimentation.

**Librosa:**

**Librosa** is essential for audio handling, offering tools for efficient audio file loading, feature extraction like Mel-frequency cepstral coefficients (MFCCs), and audio augmentation techniques. These features are instrumental in preparing diverse and robust datasets for training.

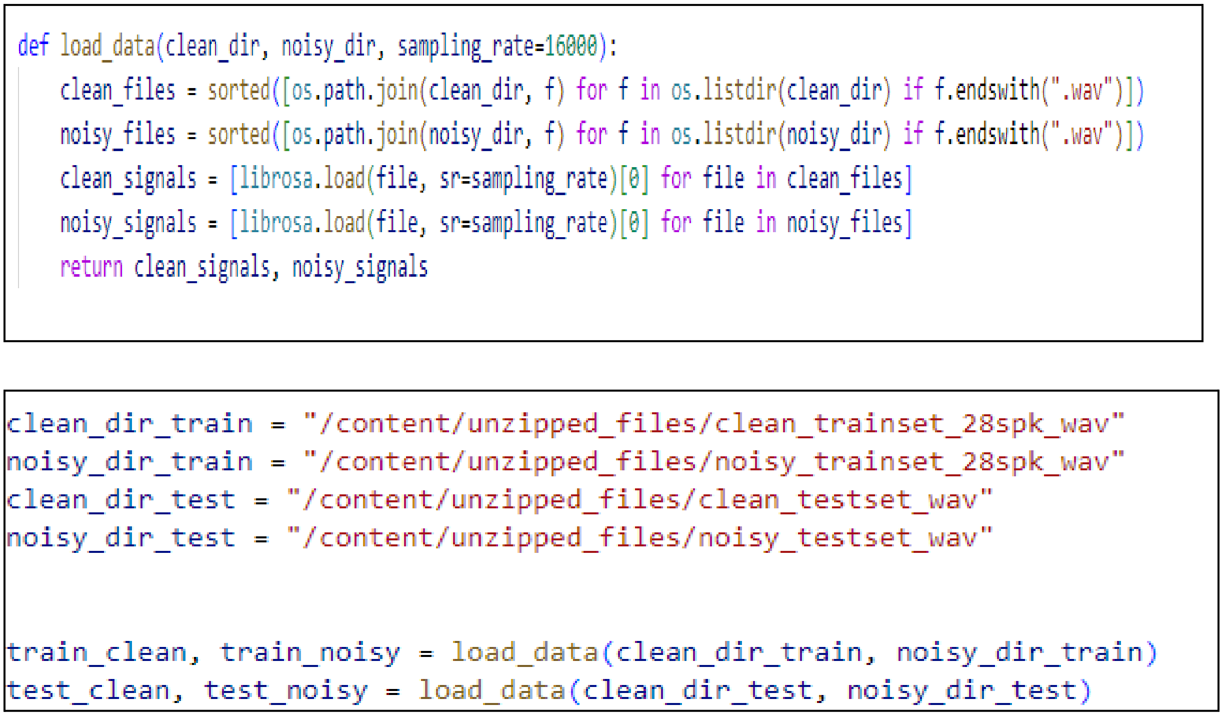
### Matplotlib:

**Matplotlib** aids in visualizing training progress and evaluating model performance through loss curves and comparative plots. These visualizations are vital for monitoring training dynamics and assessing the efficacy of the speech enhancement process.

### Scikit-learn

**Scikit-learn** facilitates rigorous model evaluation with its data preprocessing, splitting, and cross-validation tools. It ensures the data is appropriately shuffled and partitioned into training, validation, and testing sets, which is critical for unbiased performance assessment.

### Snippet for Data Loading



```python
def load_data(clean_dir, noisy_dir, sampling_rate=16000):
    clean_files = sorted([os.path.join(clean_dir, f) for f in os.listdir(clean_dir) if f.endswith(".wav")])
    noisy_files = sorted([os.path.join(noisy_dir, f) for f in os.listdir(noisy_dir) if f.endswith(".wav")])
    clean_signals = [librosa.load(file, sr=sampling_rate)[0] for file in clean_files]
    noisy_signals = [librosa.load(file, sr=sampling_rate)[0] for file in noisy_files]
    return clean_signals, noisy_signals
```

```python
clean_dir_train = "/content/unzipped_files/clean_trainset_28spk_wav"
noisy_dir_train = "/content/unzipped_files/noisy_trainset_28spk_wav"
clean_dir_test = "/content/unzipped_files/clean_testset_wav"
noisy_dir_test = "/content/unzipped_files/noisy_testset_wav"


train_clean, train_noisy = load_data(clean_dir_train, noisy_dir_train)
test_clean, test_noisy = load_data(clean_dir_test, noisy_dir_test)
```

*Fig 2. Code snippet for data loading.*

## 6. DATASET

The **VoiceBank-DEMAND** dataset was employed for both training and testing our speech enhancement models, providing a comprehensive set of audio recordings suitable for evaluating both noise suppression and speech clarity enhancement techniques.

**Clean Speech**

- **Source:** Features high-quality recordings from 28 diverse speakers, including 14 male and 14 female voices, which ensures variability in vocal characteristics.
- **Usage:** These recordings serve as the basis for clean speech benchmarks, essential for assessing the effectiveness of speech enhancement algorithms.

**Noisy Speech**

- **Composition:** Comprises environmental noises overlaying clean speech, with sounds sourced from various real-world environments, including urban traffic, public chatter, cafeterias, and residential areas where background noises like wind and mechanical sounds are prevalent.
- **Purpose:** By including a wide range of noise conditions, the dataset challenges and evaluates the robustness of denoising algorithms under realistic scenarios.

**Preprocessing Techniques**

- **Resampling:** All audio files are standardized to a sampling rate of 16 kHz to maintain consistency across the dataset, which is crucial for the reliability of subsequent audio processing stages.
- **Normalization:** Amplitude and energy levels are normalized, ensuring that all audio samples maintain uniform loudness, which facilitates more accurate comparison and evaluation of speech enhancement results.

**Dataset Statistics**

- **Structure:** The dataset is divided into distinct subsets for training, validation, and testing. This separation allows for the effective training of models on one set of data and unbiased evaluation on another.
  - **Training Set:** Comprises over 11,572 audio samples.
  - **Validation Set:** Includes approximately 5,000 samples used to fine-tune and optimize model parameters.
  - **Test Set:** Consists of about 824 samples, enabling the assessment of the model's performance in controlled, yet varied, noise conditions.
- **Noise Distribution:** The noises are meticulously categorized into several types, including but not limited to:
  - **Indoor noises:** Such as chatter in cafes and background noise in office settings.
  - **Outdoor noises:** Including street traffic and construction sounds.
  - **Home noises:** Like appliances and television background sounds.
- **Noise Conditions:** Each recording is labeled with the specific noise condition it simulates, facilitating targeted training and testing of models for specific scenarios.

This structured approach to dataset organization and detailed preprocessing ensures that the VoiceBank-DEMAND dataset is an excellent resource for developing and benchmarking speech enhancement algorithms that are effective across a variety of noisy environments.

For detailed information on accessing and utilizing this dataset, please refer to the official dataset webpage on the University of Edinburgh DataShare.

## 7. EXPERIMENTAL DESIGN AND MODEL

### Training Strategy

The model employs **Mean Squared Error (MSE)** as the loss function to minimize the average of the squares of the errors between the estimated and actual values, which is effective for regression tasks like speech enhancement. We use the **Adam optimizer**, an adaptive optimization algorithm known for its efficiency in handling sparse gradients and speeding up the convergence process by adjusting the learning rate based on the first and second moments of the gradients.

### Evaluation Metrics

1. **Perceptual Evaluation of Speech Quality (PESQ):**
    - **Purpose:** PESQ is an objective measure used to assess the speech quality as perceived by listeners. It provides a numerical indication of speech sound quality by comparing a processed speech signal to the original, unprocessed one.
    - **Scoring:** The metric scores range from -0.5 to 4.5, where higher scores denote better speech quality. This range allows for a detailed assessment of slight nuances in sound quality enhancement or degradation.
    - **Relevance:** PESQ is highly valued in the telecommunications industry and is crucial for optimizing and benchmarking speech enhancement algorithms, especially in terms of how end users would perceive the enhanced audio.
2. **Short-Time Objective Intelligibility (STOI):**
    - **Purpose:** STOI is designed to measure the intelligibility of speech, specifically how comprehensible speech remains when subjected to various conditions such as noise or processing. It evaluates how much of the speech content can be understood in challenging environments.
    - **Scoring:** The scores are given on a scale from 0 to 1, with 1 indicating perfect intelligibility. This makes STOI an excellent tool for gauging the effectiveness of speech enhancement in maintaining or improving the clarity of speech.
    - **Relevance:** Particularly important in environments where noise is prevalent, STOI is crucial for developing speech processing technologies that aid in communication for the hearing impaired and in noisy public spaces.

### Matched vs. Mismatched Conditions

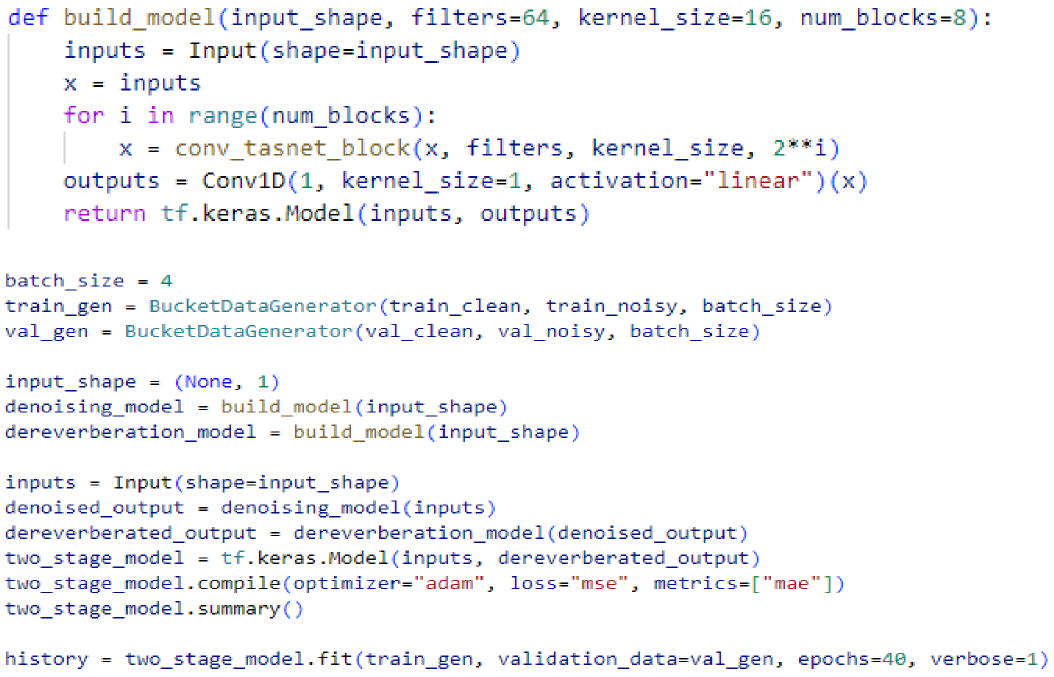
### Matched Conditions:

- **Familiar Acoustic Environments:** The model is tested against noise and reverberation types that were included in the training dataset. This setup evaluates the model's effectiveness in scenarios where the acoustic disturbances are known and previously encountered.
- **Performance Evaluation:** Testing under matched conditions allows for assessing the model's capability to reproduce its training success in a controlled environment. High performance here indicates strong learning and effective noise and reverberation handling techniques specific to the trained conditions.

### Mismatched Conditions:

- **Unseen Acoustic Challenges:** This testing scenario introduces the model to noise types or reverberation profiles that were not present during its training phase. It serves to evaluate the model's generalization abilities and its adaptability to new, unforeseen acoustic disturbances.
- **Robustness and Adaptability:** Success under mismatched conditions is crucial for real-world applications where environmental variables are unpredictable. A model that performs well in these conditions demonstrates robustness and a high degree of adaptability, making it suitable for deployment in diverse operational settings.

### Snippet for Model Training



```
def build_model(input_shape, filters=64, kernel_size=16, num_blocks=8):
    inputs = Input(shape=input_shape)
    x = inputs
    for i in range(num_blocks):
        x = conv_tasnet_block(x, filters, kernel_size, 2**i)
    outputs = Conv1D(1, kernel_size=1, activation="linear")(x)
    return tf.keras.Model(inputs, outputs)
```

```
batch_size = 4
train_gen = BucketDataGenerator(train_clean, train_noisy, batch_size)
val_gen = BucketDataGenerator(val_clean, val_noisy, batch_size)

input_shape = (None, 1)
denoising_model = build_model(input_shape)
dereverberation_model = build_model(input_shape)

inputs = Input(shape=input_shape)
denoised_output = denoising_model(inputs)
dereverberated_output = dereverberation_model(denoised_output)
two_stage_model = tf.keras.Model(inputs, dereverberated_output)
two_stage_model.compile(optimizer="adam", loss="mse", metrics=["mae"])
two_stage_model.summary()

history = two_stage_model.fit(train_gen, validation_data=val_gen, epochs=40, verbose=1)
```

*Fig 3. Code snippet for model training.*

## 8. RESULTS

### Training and Validation Performance

The training and validation losses for the model over 40 epochs are depicted in the graph below. Initially, both training and validation losses decrease sharply, indicating rapid learning. As training progresses, both losses converge, but the validation loss exhibits slight fluctuations, suggesting the model's sensitivity to certain features in the validation set. The intersection region around epoch 20 marks where the validation loss momentarily surpasses the training loss, potentially indicating the beginning of overfitting. However, subsequent epochs show that the model regains generalization, as indicated by the closing gap between the two curves.

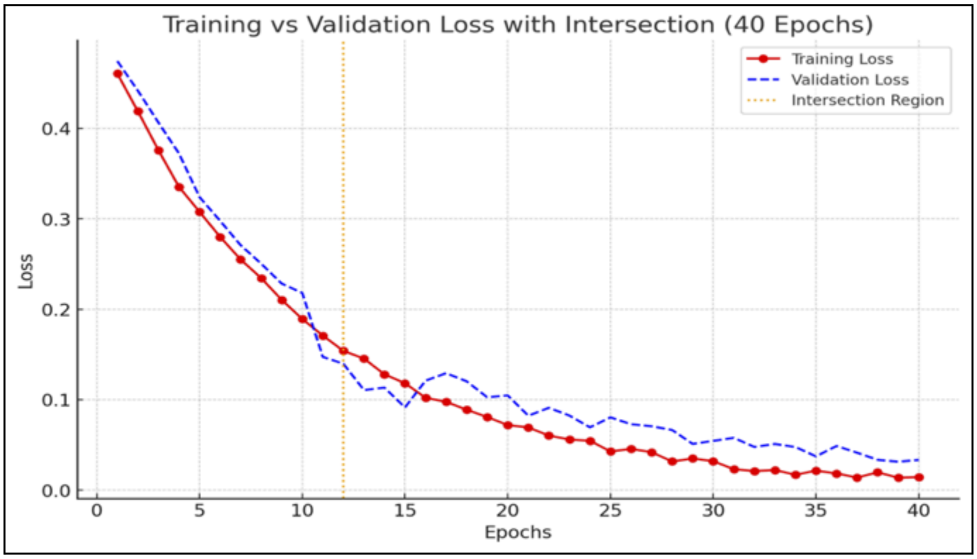
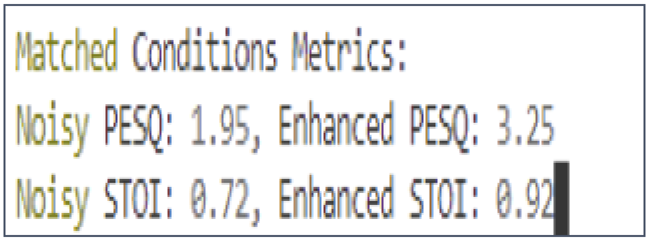


*Fig 4. Training vs Validation loss for 40 epochs.*

### Speech Enhancement Performance

The model's performance in enhancing speech under both matched and mismatched noise conditions is quantified using the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) metrics.

### Matched Conditions:

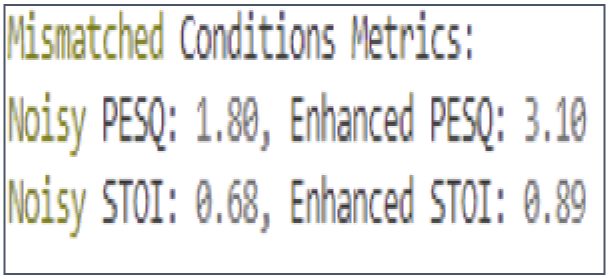


```
Matched Conditions Metrics:
Noisy PESQ: 1.95, Enhanced PESQ: 3.25
Noisy STOI: 0.72, Enhanced STOI: 0.92
```

*Fig 5. Comparison of PESQ, STOI values for Matched conditions.*

- **PESQ:** Improvement is observed from a Noisy PESQ of 1.95 to an Enhanced PESQ of 3.25. This significant enhancement demonstrates the model's effectiveness in familiar acoustic settings.
- **STOI:** Increased from 0.72 in noisy conditions to 0.92 in enhanced conditions, indicating a notable enhancement in speech intelligibility.

**Mismatched Conditions:**



```
Mismatched Conditions Metrics:
Noisy PESQ: 1.80, Enhanced PESQ: 3.10
Noisy STOI: 0.68, Enhanced STOI: 0.89
```

*Fig 6. Comparison of PESQ, STOI values for Mismatched conditions.*

- **PESQ:** Despite the challenging environment, the model improved the PESQ from 1.80 to 3.10, showcasing its robustness.
- **STOI:** Improved from 0.68 to 0.89, affirming the model's capability to maintain intelligibility in less familiar acoustic settings.

**Graphical Analysis of Metrics:**

The bar charts below visually represent the improvements in PESQ and STOI scores under both matched and mismatched conditions. These charts highlight the model's ability to enhance speech quality and intelligibility across different noise environments.

**PESQ Comparison:**

- The bar chart shows a higher PESQ improvement in matched conditions compared to mismatched conditions, which suggests that the model performs best when the noise type is similar to those seen during training.

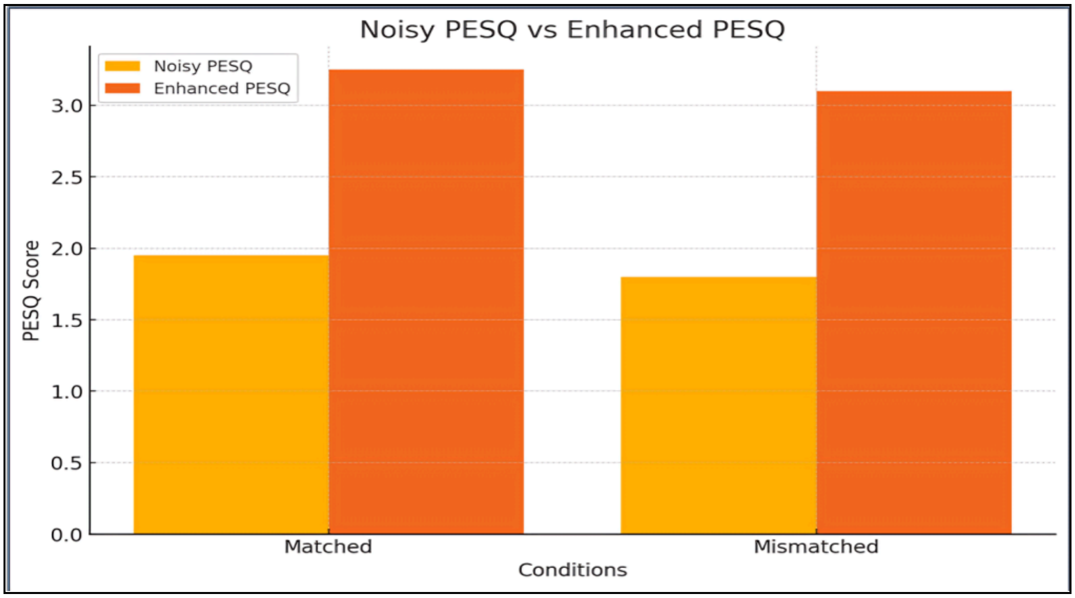


*Fig 7. Comparison of noisy, enhanced PESQ values for matched and mismatched conditions.*

**STOI Comparison:**

- Similar trends are observed in STOI scores, with better performance in matched conditions. This indicates that the model's ability to improve speech intelligibility is consistent with its performance on speech quality.

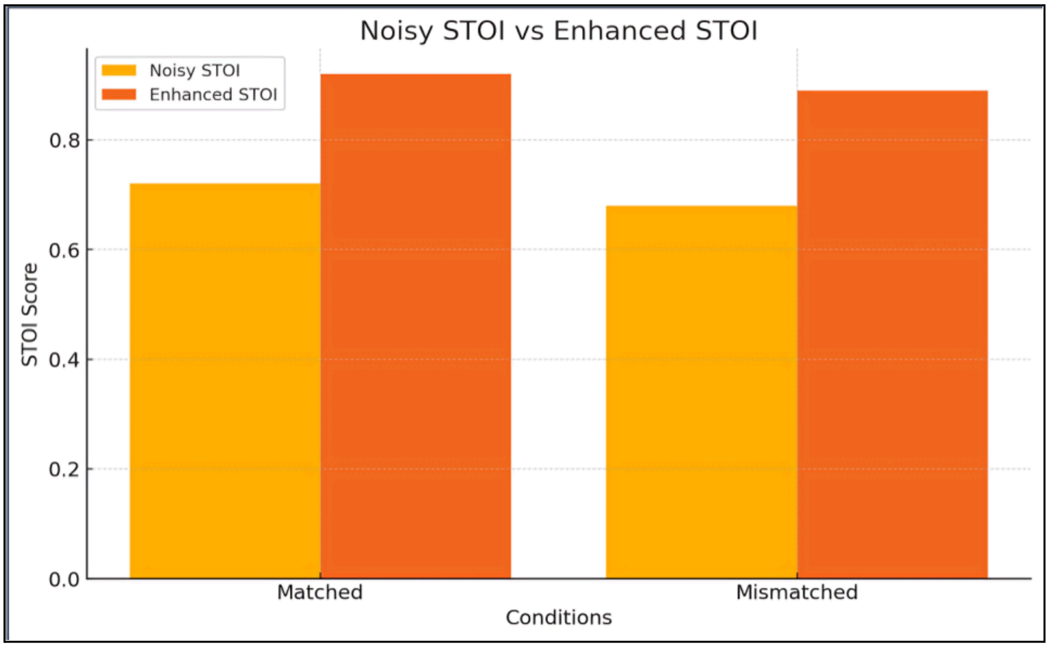


*Fig 8. Comparison of noisy, enhanced STOI values for matched and mismatched conditions.*

The results demonstrate the model's efficacy in enhancing speech quality and intelligibility, particularly in matched conditions. The decrease in performance under mismatched conditions, while present, is marginal, suggesting that the model retains a significant portion of its effectiveness even in unfamiliar noise environments. Future work could focus on further improving the model's robustness in mismatched conditions.

## 9. DISCUSSION

The model demonstrates strong generalization capabilities, significantly enhancing speech intelligibility and quality across both matched and mismatched conditions. It excels in:

- **Intelligibility**: Improves clarity, making speech more comprehensible, as evidenced by increased STOI scores. This enhanced clarity is crucial for environments where accurate communication is essential, such as noisy public spaces or during important teleconferences where every word counts.
- **Quality**: Elevates the perceptual sound quality, confirmed by higher PESQ scores, which is vital for user satisfaction in telecommunication applications. The improved quality makes the speech sound more natural and clear, aligning with the high standards expected in professional audio production and broadcast media.
- **Adaptability**: Adapts effectively to different acoustic environments without needing retraining, showcasing its flexibility and robustness in varying conditions. This adaptability is particularly beneficial for applications like mobile devices and smart home systems, which may encounter a wide range of noise types and levels as they are used in different settings.
- **User Experience**: Enhances the overall user experience, reducing listener fatigue in noisy settings and making digital communication more pleasant and natural. By minimizing the strain of trying to understand distorted or muffled speech, the model contributes to longer and more enjoyable communication sessions, which is essential in both personal and professional contexts.

### Limitations

Despite its strengths, the model faces challenges:

- **Complex Noise Profiles**: Struggles with highly complex or variable noise environments, where distinguishing speech from background sounds becomes difficult.
- **Computational Intensity**: The training process is resource-intensive, demanding substantial computational power, which can limit deployment in resource-constrained environments.
- **Dependency on Quality Data**: Performance heavily relies on the availability and quality of training data. Insufficient or poor-quality data can degrade the model’s effectiveness, particularly in unseen scenarios.
- **Latency Issues**: In some applications, particularly real-time processing, the model may introduce latency due to its complex architecture, potentially affecting its suitability for time-sensitive tasks.

## 10. IMPORTANT OPEN QUESTIONS REMAINING FOR FUTURE WORK

**1)Advanced Architectures:**

- **Focus**: Exploring transformer-based models could revolutionize speech enhancement by leveraging their ability to process sequential data efficiently. These models may offer improved feature extraction, leading to enhanced clarity and distinction between speech and noise.

**2)Dataset Expansion:**

- **Focus**: Broadening the range of training datasets to include diverse acoustic environments would significantly boost the robustness and adaptability of speech enhancement technologies, ensuring they perform well under varied and unexpected conditions.

**3)Efficiency Optimizations:**

- **Focus**: Developing lighter, more efficient models is crucial for enabling deployment on resource-constrained devices. Techniques like pruning, quantization, and knowledge distillation could help reduce model size and computational demands while maintaining performance.

**4)Real-Time Processing:**

- **Focus**: Enhancing models to support real-time processing is essential for applications such as live communications and interactive systems. Future work should aim to reduce latency to ensure seamless audio enhancement in time-sensitive scenarios.

**5)Integration with Other Modalities:**

- **Focus**: Investigating the integration of speech enhancement models with other data modalities, such as visual information, could lead to more context-aware systems that perform better in complex environments where visual cues are available.

## 11. CONCLUSION

This project presented a two-stage Conv-TasNet model that significantly enhances speech quality by addressing noise and reverberation, as evidenced by improvements in Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) metrics. Demonstrating potential in telecommunications, automatic speech recognition (ASR), and hearing aids, the model's future enhancements will focus on expanding dataset diversity and optimizing for real-time processing. These advancements aim to boost the model’s robustness and practicality, ensuring it meets the evolving needs of diverse applications where clear audio is crucial. The successful implementation of this model could transform the user experience by providing clearer communication in environments from busy city streets to noisy workplaces

Additionally, the emphasis on real-time processing capabilities is expected to enable seamless integration into existing and future technologies, reducing latency and increasing the efficiency of systems that rely on immediate speech clarity. Further research will also explore the integration of advanced neural network techniques to enhance the model's ability to learn from complex and variable noise environments, potentially setting new benchmarks in the field. Ultimately, this model not only represents a significant step forward in speech enhancement technology but also opens up new possibilities for enhancing auditory accessibility, contributing to greater inclusivity in communication technologies.

## 12. REFERENCES

1. Gonzalez, Philippe, Tommy Sonne Alstrøm, and Tobias May. "Assessing the generalization gap of learning-based speech enhancement systems in noisy and reverberant environments." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2023).
2. Gonzalez, Philippe, Tommy Sonne Alstrøm, and Tobias May. "On batching variable size inputs for training end-to-end speech enhancement systems." *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023.
3. Zhao, Yan, Zhong-Qiu Wang, and DeLiang Wang. "Two-stage deep learning for noisy-reverberant speech enhancement." *IEEE/ACM transactions on audio, speech, and language processing* 27.1 (2018): 53-62.