Shahid Beheshti University

Artificial Neural Networks

M.Sc - Fall 2024

Assignment 4 - Part 2

Audio Denoising with Autoencoders: Enhancing Audio MNIST Samples

Author:
Zahra Mohammad Beigi

Student Number:
402422144

December 1, 2024

Contents

1 Introduction

1.1 Objective

The objective of this project is to design and train an autoencoder for denoising audio samples from the Audio MNIST dataset. The primary goal is to convert noisy audio files into clean ones, improving their quality and usability in downstream applications.

1.2 Dataset Overview

The Audio MNIST dataset forms the basis of this project and consists of:

- 30,000 audio samples of spoken digits (0-9) recorded by 60 speakers.
- 30,000 corresponding noisy audio samples with a Signal-to-Noise Ratio (SNR) of -8 dB.
- Metadata provided in the `audioMNIST_meta.txt` file, containing speaker information such as gender and age.

This dataset allows for a comprehensive study of audio denoising using clean-noisy audio file pairs.

2 Exploratory Data Analysis(EDA)

2.1 Speaker Metadata Summary

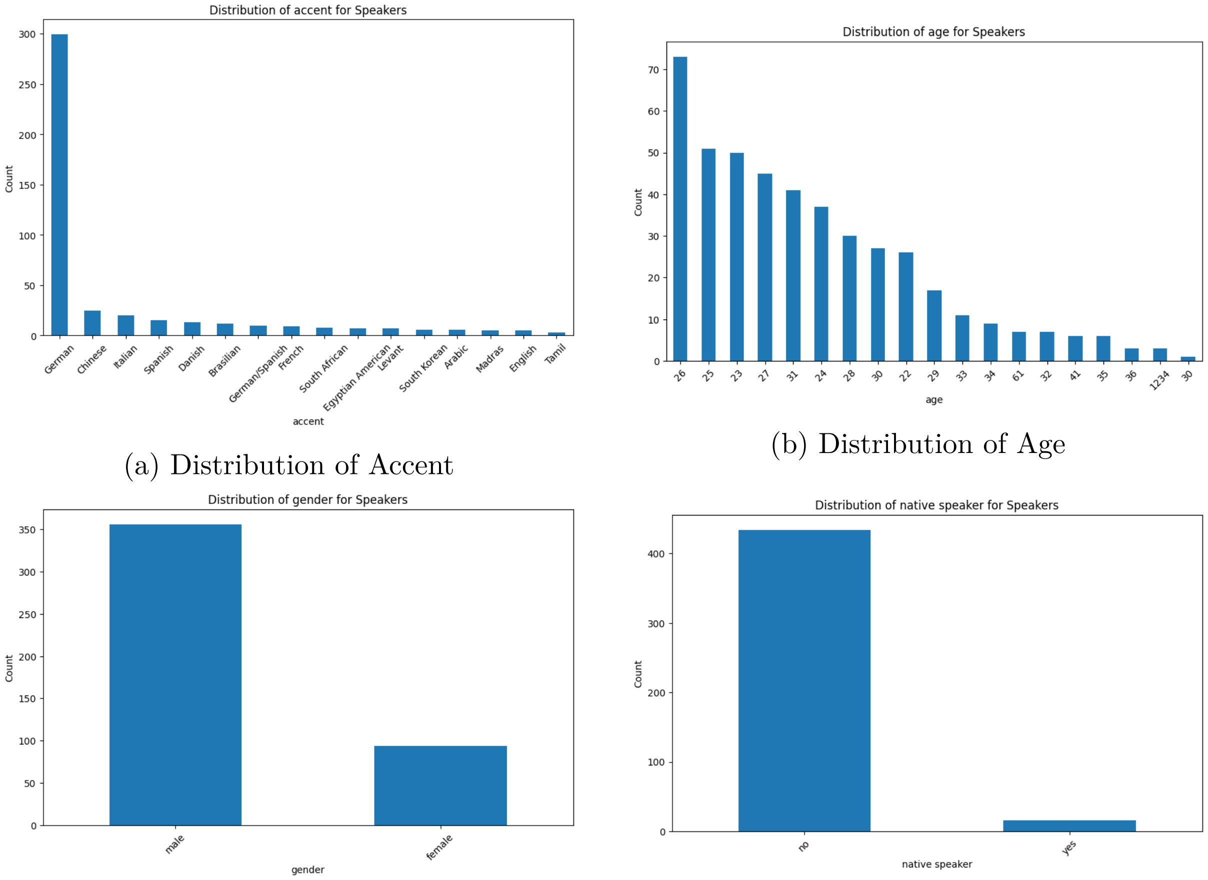
The Audio MNIST dataset includes metadata describing speaker characteristics and recording conditions. A summary of the metadata is provided below:

- **Accent:** 17 unique accents, with the most frequent being German (40 occurrences).
- **Age:** 19 unique age groups, with the most common being 26 years old (10 occurrences).
- **Gender:** Speakers include 48 males and 12 females.
- **Native Speaker Status:** 57 speakers are non-native speakers, while 3 are native speakers.
- **Origin:** Speakers originate from 42 unique locations, with the most common origin being Europe, Germany, Berlin (15 occurrences).
- **Recording Dates:** Each of the 60 recordings has a unique timestamp, indicating distinct recording sessions.
- **Recording Rooms:** Audio was recorded in 7 different rooms, with the most frequently used being the ”vr-room” (28 recordings).

2.2 Speaker Attribute Analysis

To gain insights into the characteristics of the speakers in the dataset, we analyzed the distributions of key speaker attributes: *accent*, *age*, *gender*, *native speaker status*, *origin*, and *recording room*. These attributes were extracted from the metadata associated with each speaker. The distribution for each attribute was visualized to assess the diversity and balance in the dataset.

Figure 1 shows the distribution of these six attributes. Each subplot corresponds to one attribute, with the x-axis representing the categories of the attribute and the y-axis representing the count of speakers in each category.

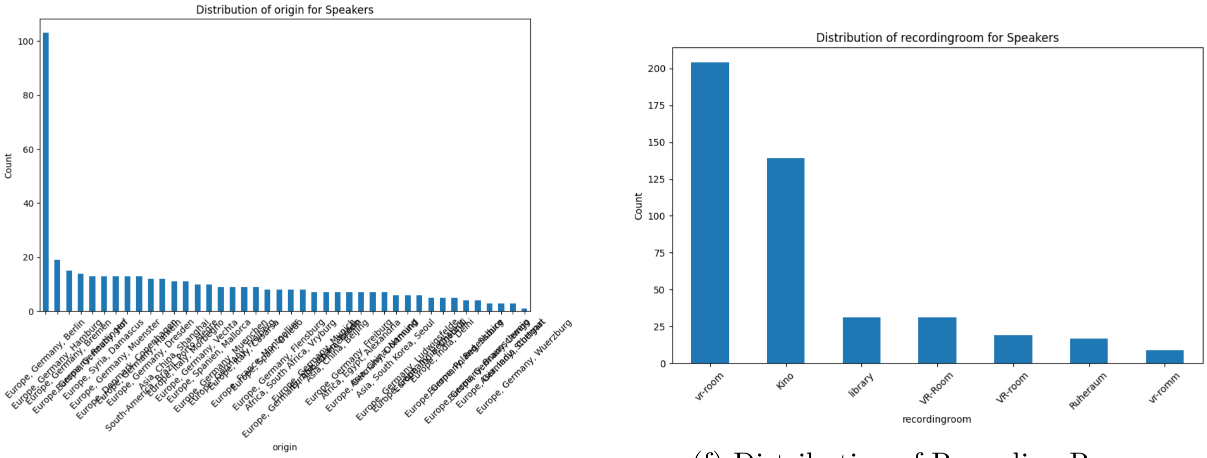


(a) Distribution of Accent

(b) Distribution of Age

(c) Distribution of Gender

(d) Distribution of Native Speaker Status



(e) Distribution of Origin

(f) Distribution of Recording Room

Figure 1: Distribution of Speaker Attributes in the Dataset. The bar charts depict the count of speakers across various categories for each attribute.

2.3 Signal-to-Noise Ratio (SNR) Analysis

To assess the quality of the audio recordings in the Audio MNIST dataset, the Signal-to-Noise Ratio (SNR) was calculated for pairs of clean and noisy audio files. The SNR provides a measure of the amount of noise relative to the signal in the audio data, with higher values indicating better quality.

The SNR was computed for 30,000 audio pairs, resulting in a wide range of SNR values, from a minimum of -46.93 dB to a maximum of 7.57 dB, with a mean of -24.46 dB. These values suggest considerable variation in the quality of the recordings, with many pairs having a low SNR indicating significant noise interference.

Further analysis was conducted on the SNR values grouped by digits and accents. The following figures summarize the SNR statistics for each digit and accent.

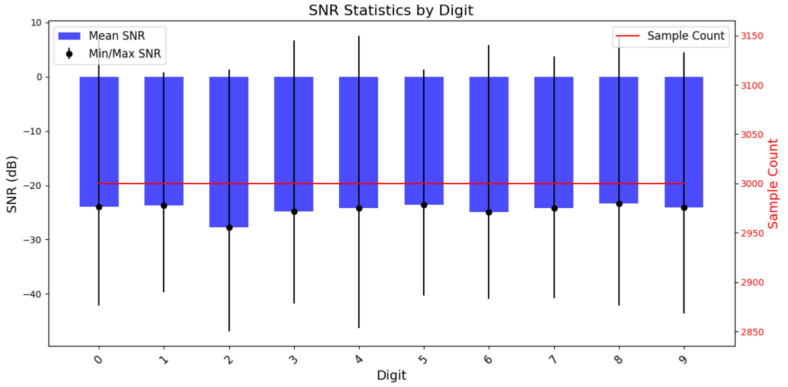


Figure 2: SNR Statistics by Digit. This figure illustrates the mean, minimum, and maximum SNR values for each digit in the dataset. The digits show a range of SNR values, with some digits having higher noise levels than others.

The SNR statistics for each digit indicate that some digits, such as 0 and 1, have relatively higher mean SNR values, while digits like 6 and 7 show lower mean SNRs. These variations suggest that certain digits may be more prone to noise, possibly due to factors like pronunciation or recording conditions.

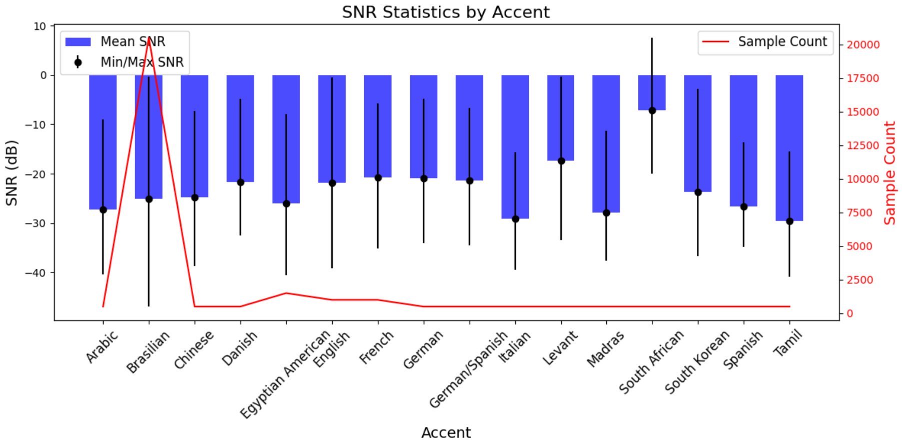


Figure 3: SNR Statistics by Accent. This figure presents the SNR values across different accents in the dataset. Accents such as South Korean and Tamil have relatively lower mean SNRs, indicating higher noise levels, while accents like Arabic and South African exhibit better SNR performance.

The accent-specific SNR analysis reveals that accents like South Korean and Tamil have higher levels of noise, which may be attributed to factors such as regional differences in speech patterns or recording environments. On the other hand, accents such as Arabic and South African show better SNR values, indicating relatively cleaner recordings.

3 Preprocessing

3.1 Audio Data Pairing and Spectrogram Analysis

To facilitate training, each clean audio file was paired with its corresponding noisy file. A total of 30,000 paired audio files were created. The first audio pair was analyzed to ensure the data integrity and quality.

3.2 Comparison of Clean and Noisy Spectrograms

The spectrograms of a clean and its corresponding noisy audio file were analyzed. The visual comparison highlights the noise introduced in the dataset:

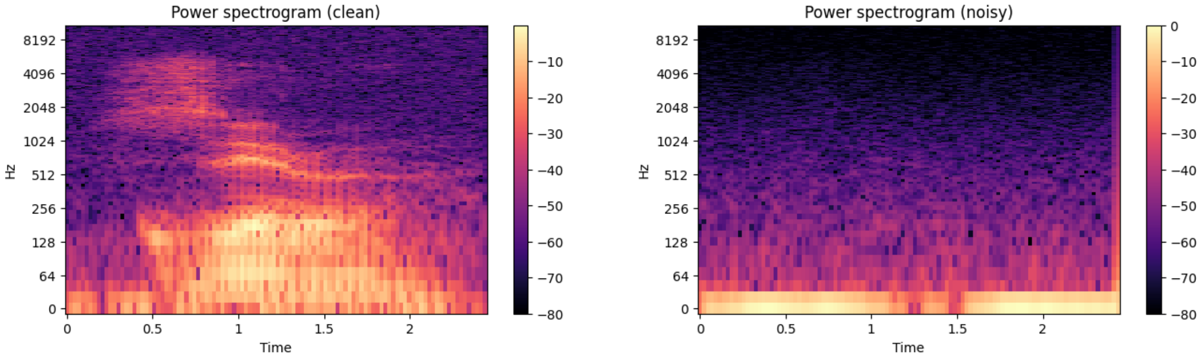


Figure 4: Comparison of power spectrograms: (left) Clean audio and (right) Noisy audio. The noisy spectrogram reveals additional high-energy components introduced by noise.

3.3 Feature Extraction

The denoising model relies on spectrograms derived from the clean and noisy versions of each audio sample. During training, the noisy spectrograms serve as inputs to the model, while the clean spectrograms act as the target outputs that the model learns to reconstruct.

Preprocessing Steps To ensure consistency in size, all audio files were adjusted to a fixed duration of 1 second. This was achieved by either padding shorter files or trimming longer ones. The preprocessing procedure includes the following steps:

- Audio files were resampled to a uniform sample rate of 22,050 Hz.
- For clean and noisy audio files, Short-Time Fourier Transform (STFT) was computed with an FFT size of 2048.
- Spectrograms were generated for both the clean and noisy versions, with dimensions adjusted to be compatible with the convolutional layers of the autoencoder.

3.4 Augmented Audio and Spectrogram Analysis

In this section, we analyze the effect of adding Gaussian and Pink noise to the audio samples. Additionally, we will examine the resulting spectrograms and the impact of these noise types on the frequency and time domain representations of the signals.

3.4.1 Waveform Visualizations

We begin by visualizing the waveforms of the clean audio and its noisy counterparts. The clean audio represents the original, unaltered signal, while the noisy audio samples were augmented with Gaussian and Pink noise. These waveforms help to demonstrate the effects of noise on the amplitude variations of the audio signals.

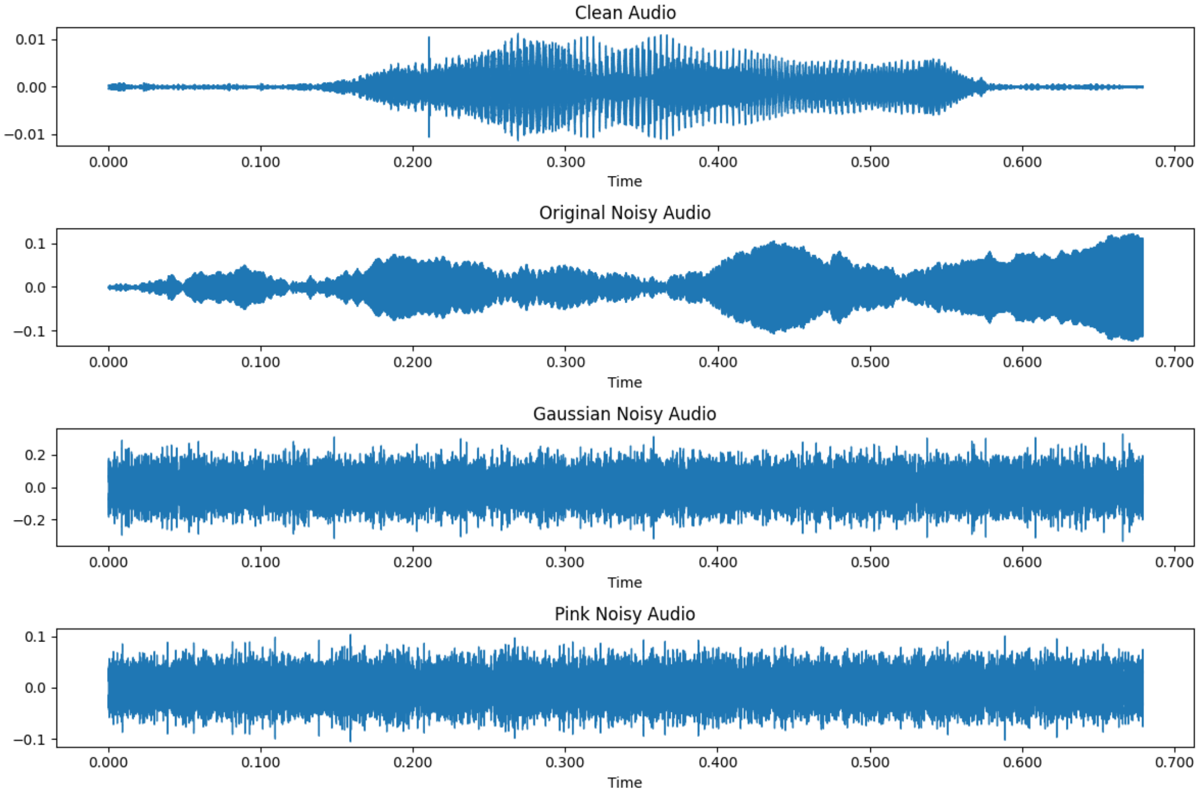


Figure 5: Comparison of Waveforms for Clean, Original Noisy, Gaussian Noisy, and Pink Noisy Audio. From top to bottom: Clean Audio, Original Noisy Audio, Gaussian Noisy Audio, Pink Noisy Audio. The added noise distorts the signal, with Gaussian and Pink noise showing different characteristics in the waveform.

As shown in the figure, the clean audio exhibits smooth, consistent fluctuations in amplitude. In contrast, the original noisy audio introduces additional disturbances due to environmental noise. The Gaussian noisy audio waveform shows random fluctuations in the signal, while the Pink noisy audio introduces more consistent and structured noise, reflecting its unique frequency characteristics.

3.4.2 Spectrogram Features

After the augmentation, spectrogram features were extracted from both the clean and noisy audio samples. Spectrograms are a powerful tool for visualizing the frequency content of an audio signal over time. They provide a representation that highlights both the frequency and temporal dynamics of the signal, which is crucial for the autoencoder to learn denoising effectively.

The clean and noisy audio samples were transformed into spectrograms to serve as the input for the autoencoder model. These spectrograms provide a clear visual distinction between the clean and noisy signals, aiding in the model's ability to differentiate between noise and the underlying clean signal.

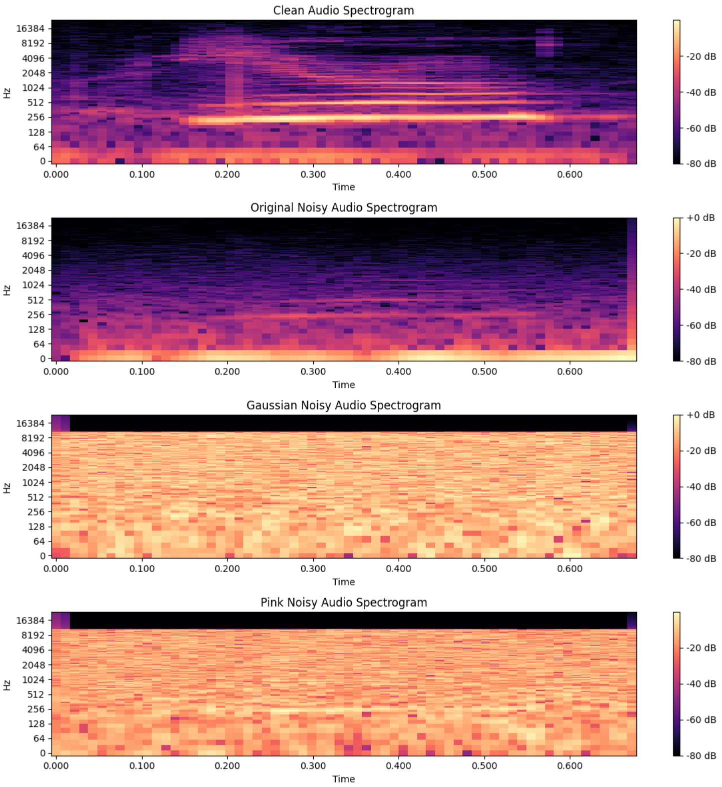


Figure 6: Spectrograms of Clean and Noisy Audio. From left to right: Clean Audio, Original Noisy Audio, Gaussian Noisy Audio, Pink Noisy Audio. The spectrograms visually illustrate the effect of noise on the frequency content of the signal.

The spectrograms show how different types of noise impact the signal's frequency distribution. The clean spectrogram shows a well-defined frequency structure, while the noisy spectrograms exhibit additional spectral components introduced by the noise. The Gaussian noise leads to random, high-frequency fluctuations, while the Pink noise introduces a more structured distortion with a gradual decrease in power across higher frequencies.

3.4.3 Dataset Augmentation and Size

Due to constraints in storage space and memory, only the first 1000 samples were selected for augmentation. Since these samples corresponded mainly to two speakers, the filepaths list was shuffled to acquire a more balanced training set. The augmentation process added Gaussian and Pink noise at varying Signal-to-Noise Ratios (SNRs) to the original clean audio files.

The augmented dataset now contains 3000 samples, which provides a more diverse set of training examples. This increased dataset size enables the model to better learn the denoising task under different noise conditions, improving its robustness.

The augmented audio dataset, consisting of both clean and noisy samples, will serve as the input for the autoencoder model. The spectrograms provide a detailed representation of the noise's impact on the signal, allowing the model to learn how to effectively separate noise from the clean audio. The increased dataset size and diversity due to noise augmentation will enhance the model's generalization capabilities and improve its performance in real-world applications.

3.4.4 Feature Dimensions

The spectrograms were processed into 2D tensors with an additional channel dimension, formatted as $(1024, 44, 1)$, to be compatible with the convolutional neural network. This preprocessing step ensures the model can effectively learn from the audio data while aligning with the neural network architecture.

4 Model Architecture and Performance Evaluation

In this section, we implemented an autoencoder model, train it on the augmented dataset, and evaluate its performance in terms of Signal-to-Noise Ratio (SNR) improvement. Additionally, we analyze the model's performance across different speakers and digits.

4.1 Model Architecture

The implemented autoencoder model is designed to denoise spectrograms of noisy audio samples. The encoder comprises three convolutional layers, each followed by batch normalization to stabilize training and improve convergence. The decoder mirrors the encoder with upsampling layers and convolutional layers to reconstruct the denoised spectrogram. The model uses ReLU activation in the intermediate layers and a sigmoid activation in the output layer to ensure the output values remain in the desired range.

Layer (Type)	**Output Shape**	**Param #**
encoder_in (InputLayer)	(None, 1024, 44, 1)	0
conv2d_6 (Conv2D)	(None, 1024, 44, 32)	320
batch_normalization_5 (BatchNormalization)	(None, 1024, 44, 32)	128
conv2d_7 (Conv2D)	(None, 1024, 44, 64)	18,496
batch_normalization_6 (BatchNormalization)	(None, 1024, 44, 64)	256
conv2d_8 (Conv2D)	(None, 1024, 44, 128)	73,856
batch_normalization_7 (BatchNormalization)	(None, 1024, 44, 128)	512
max_pooling2d_1 (MaxPooling2D)	(None, 512, 22, 128)	0
conv2d_9 (Conv2D)	(None, 512, 22, 128)	147,584
batch_normalization_8 (BatchNormalization)	(None, 512, 22, 128)	512
up_sampling2d_1 (UpSampling2D)	(None, 1024, 44, 128)	0
conv2d_10 (Conv2D)	(None, 1024, 44, 64)	73,792
batch_normalization_9 (BatchNormalization)	(None, 1024, 44, 64)	256
conv2d_11 (Conv2D)	(None, 1024, 44, 1)	577

Table 1: Model Summary: Autoencoder Architecture

The model is trained using the Adam optimizer with a learning rate of 0.0001. The loss function is the Mean Squared Error (MSE), which quantifies the difference between the reconstructed and clean spectrograms.

4.2 Training and Validation

The autoencoder model was trained on the training set consisting of spectrograms extracted from the augmented dataset. The training process involved 50 epochs, with early stopping applied to prevent overfitting. Additionally, a ReduceLROnPlateau callback was used to reduce the learning rate when the validation loss plateaued. The training and validation loss values are shown in Figure 7.

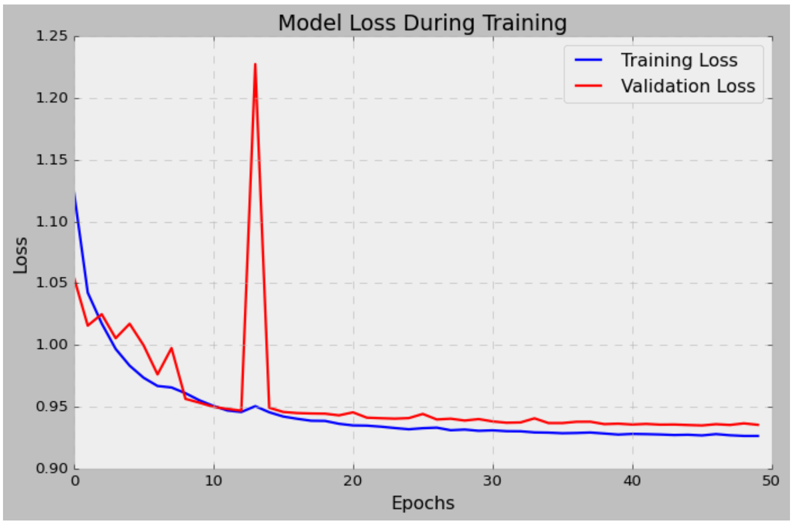


Figure 7: Training and validation loss during the training process.

4.3 Model Performance on Test Data

After training, the model was evaluated on the test set. The denoised spectrograms produced by the model were compared with the original clean spectrograms to calculate the Signal-to-Noise Ratio (SNR) improvement. On average, the model achieved an SNR improvement of 0.52 dB on the test data.

To visualize the effectiveness of the model, Figure 8 compares the spectrograms of clean, noisy, and denoised audio for a sample from the test set.

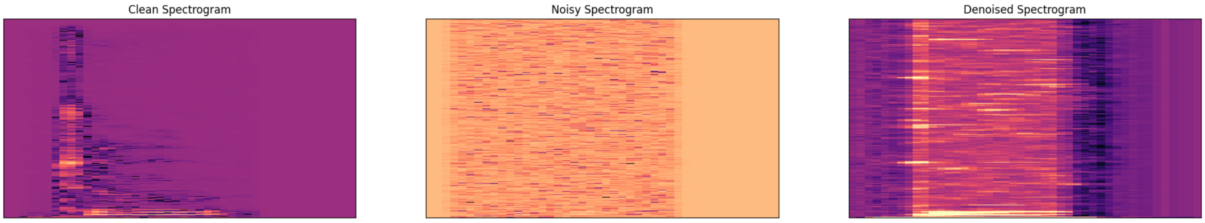


Figure 8: Spectrogram comparison for a test sample. Left: Clean spectrogram. Center: Noisy spectrogram. Right: Denoised spectrogram.

4.4 Performance Across Speakers and Digits

In this section, we analyze the Signal-to-Noise Ratio (SNR) improvement achieved for different speakers and digits in the dataset. The average SNR improvement is presented as bar plots for each category, providing insights into the denoising performance of the autoencoder.

4.4.1 SNR Improvement by Speaker

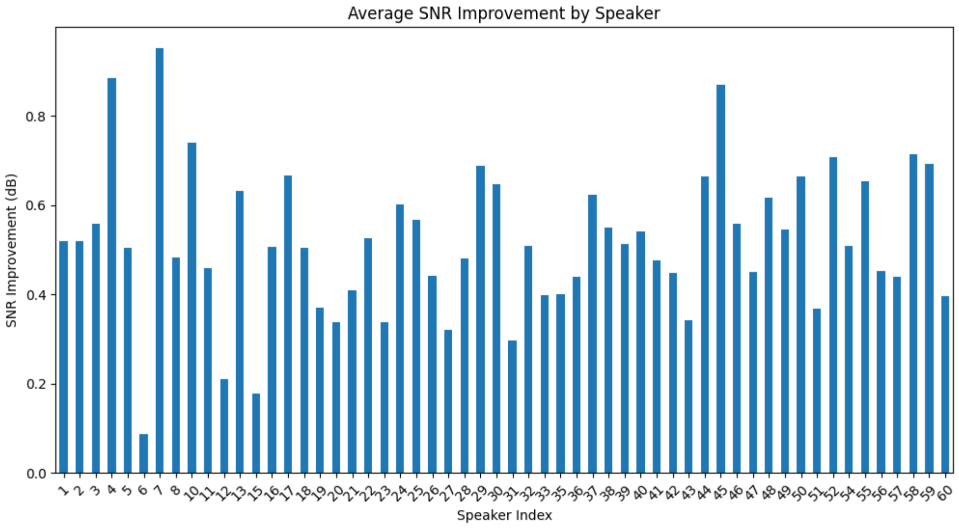


Figure 9: SNR Improvement by Speaker. The bar plot shows the mean SNR improvement (in dB) for each speaker. Speaker indices with higher SNR values indicate better denoising performance.

The SNR improvement varies across speakers, with notable performance differences. For example:

- The highest SNR improvement is observed for Speaker 7 (SNR = 0.951 dB).
- Other speakers, such as Speaker 45 (SNR = 0.870 dB) and Speaker 4 (SNR = 0.886 dB), also demonstrate significant improvements.
- Speakers like 6 (SNR = 0.087 dB) show lower improvements, highlighting possible variations in input noise levels or recording conditions.

4.4.2 SNR Improvement by Digit

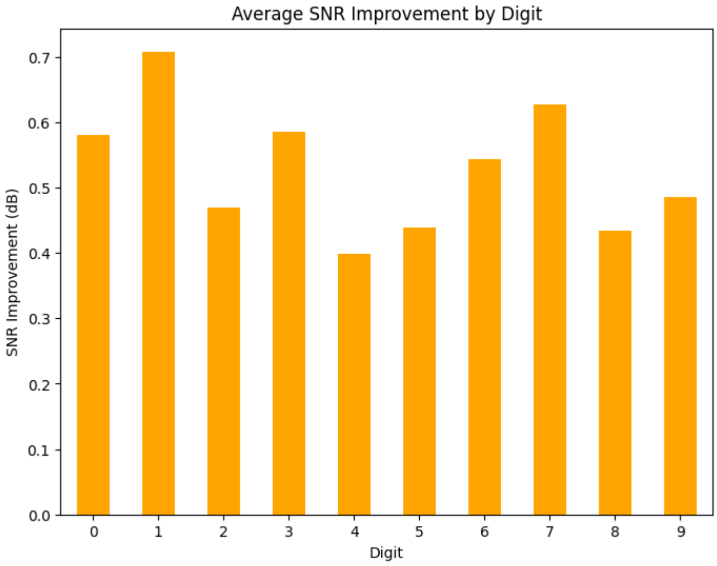


Figure 10: SNR Improvement by Digit. The bar plot illustrates the mean SNR improvement (in dB) for each digit from 0 to 9.

When analyzing the SNR improvement by digit:

- The digit ”1” shows the highest SNR improvement (SNR = 0.708 dB), suggesting effective noise removal for this category.
- Digits such as ”7” (SNR = 0.626 dB) and ”3” (SNR = 0.585 dB) also show strong performance.
- Conversely, digit ”4” (SNR = 0.399 dB) has the lowest improvement, potentially due to inherent challenges in its acoustic features.

4.5 Analysis of Signal-to-Noise Ratio (SNR) Improvement by Speaker Attributes

In this section, we analyze the Signal-to-Noise Ratio (SNR) improvement based on different speaker attributes: accent, age, and gender. The SNR improvement is measured for each attribute and presented as bar plots, allowing us to explore how these characteristics influence the denoising performance.

4.5.1 SNR Improvement by Accent

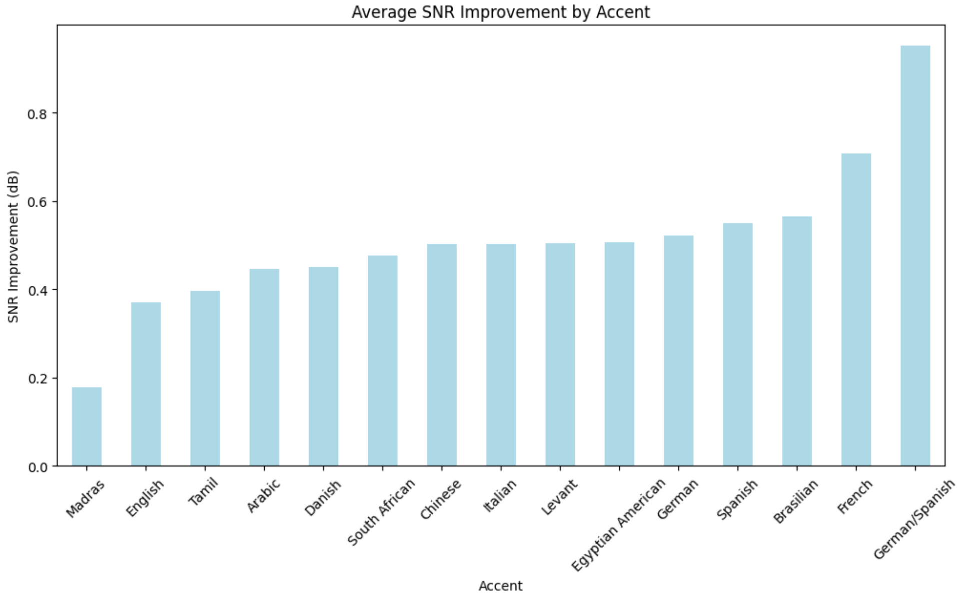


Figure 11: Average SNR Improvement by Accent. The bar plot displays the average SNR improvement (in dB) for speakers with different accents.

The SNR improvement for different accents reveals some notable trends:

- The highest SNR improvement is observed for speakers with the ”German/Spanish” accent (SNR = 0.95 dB).
- Other accents such as ”French” (SNR = 0.71 dB) and ”Brasilian” (SNR = 0.57 dB) show significant improvements as well.
- Accents like ”Madras” (SNR = 0.18 dB) exhibit lower SNR improvements, suggesting that either the accent is more difficult for the model to process or the noise characteristics of the accent vary.

4.5.2 SNR Improvement by Age

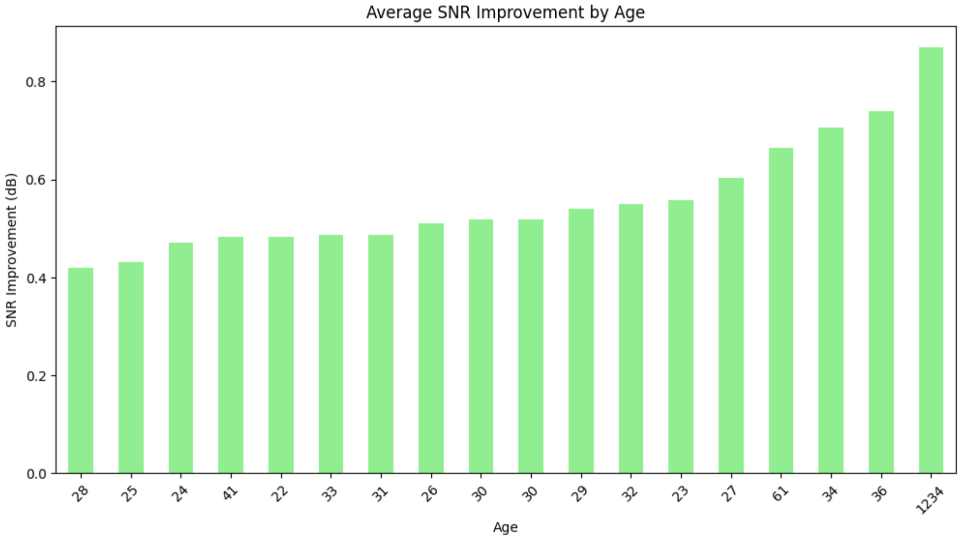


Figure 12: Average SNR Improvement by Age. The bar plot illustrates the average SNR improvement (in dB) across different age groups.

The SNR improvement varies slightly across different age groups.

4.5.3 SNR Improvement by Gender

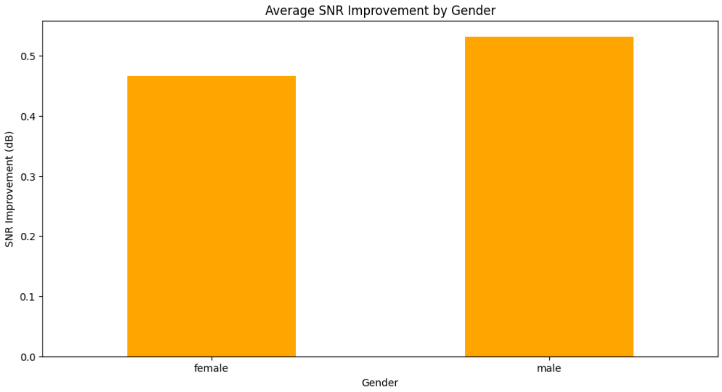


Figure 13: Average SNR Improvement by Gender. The bar plot shows the average SNR improvement (in dB) for male and female speakers.

The SNR improvement by gender shows a slight difference:

- Male speakers show a slightly higher SNR improvement (SNR = 0.53 dB) compared to female speakers (SNR = 0.47 dB).
- This difference could be attributed to various factors such as voice pitch or other acoustic characteristics that might influence the noise removal process.

4.5.4 SNR Improvement by Accent and Gender

In this analysis, the average Signal-to-Noise Ratio (SNR) improvement was calculated for different accents and genders. The results were visualized using a heatmap to compare the SNR improvement across accents and gender categories.

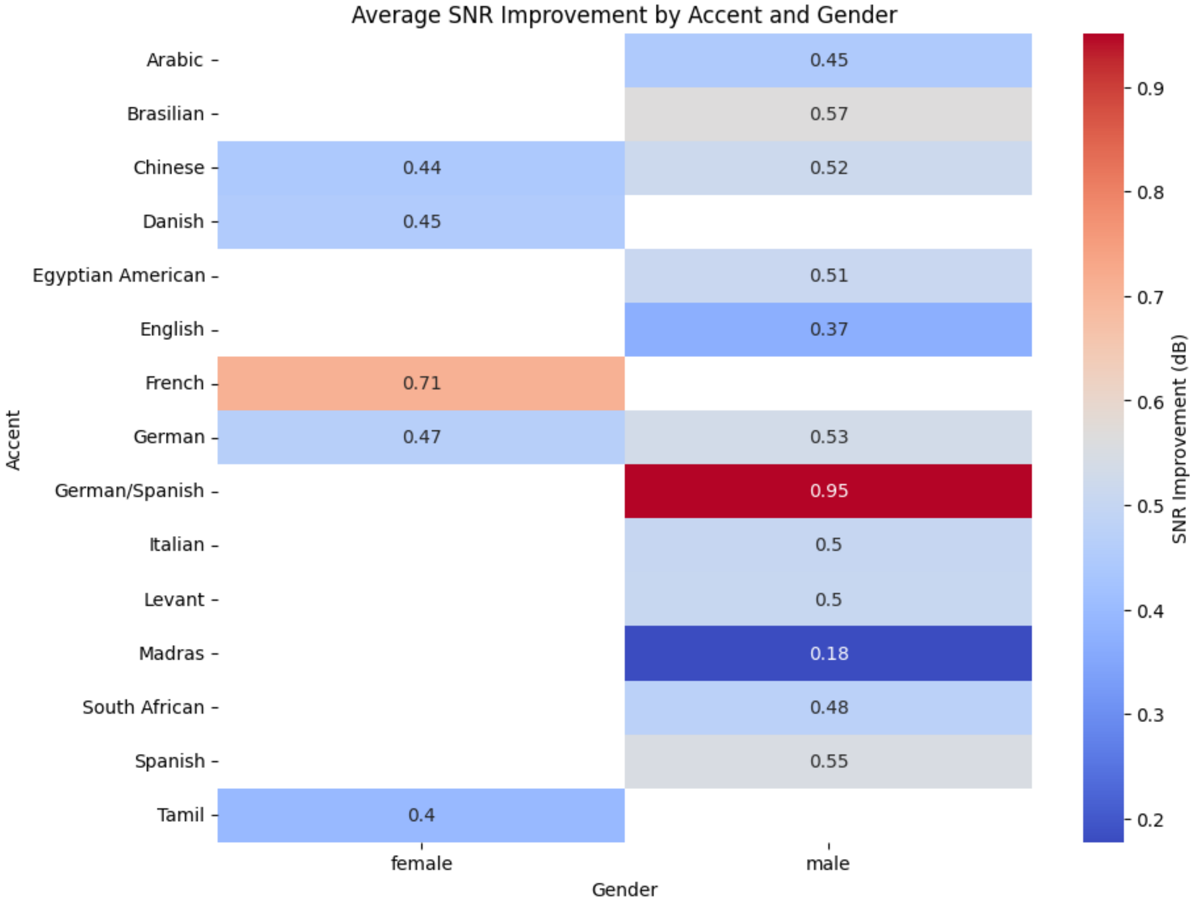


Figure 14: Heatmap of Average SNR Improvement by Accent and Gender. The heatmap shows the average SNR improvement (in dB) for each combination of accent and gender.

From the heatmap, several important trends can be observed:

- French stands out with the highest SNR improvement for females, while German/Spanish shows the highest improvement for males.
- Certain accents like Arabic, English, and Madras show only male SNR values, indicating that data for females is unavailable for these accents.
- Accents such as Tamil exhibit a lower SNR improvement for females, highlighting the variability in SNR improvement based on accent and gender.

The heatmap helps visualize how different accents and genders influence the overall SNR improvement in audio denoising tasks.

4.5.5 SNR Improvement by Native Speaker Status

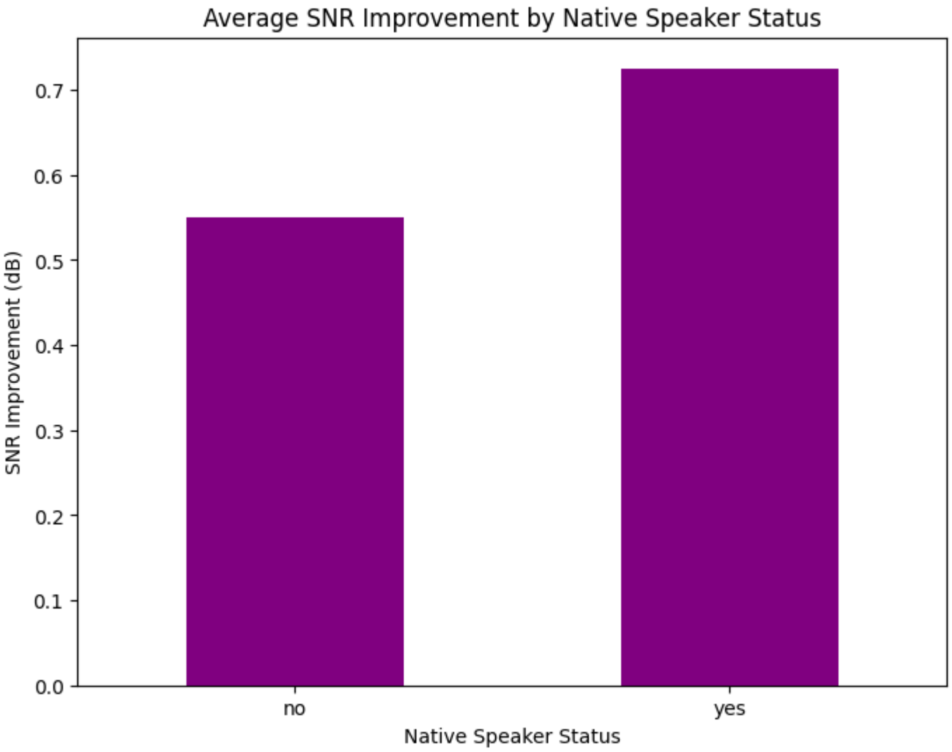


Figure 15: Average SNR Improvement by Native Speaker Status. The bar chart compares the SNR improvement between native and non-native speakers.

As shown in Figure 15, the average SNR improvement for non-native speakers (labeled as ”no”) is higher than that for native speakers (labeled as ”yes”). Specifically, the SNR for non-native speakers is approximately 0.52 dB, while for native speakers, it is around 0.43 dB. This indicates a relatively higher improvement for non-native speakers in the dataset, which may reflect specific patterns related to the acoustics or characteristics of non-native speech.

4.5.6 SNR Improvement by Accent and Digit

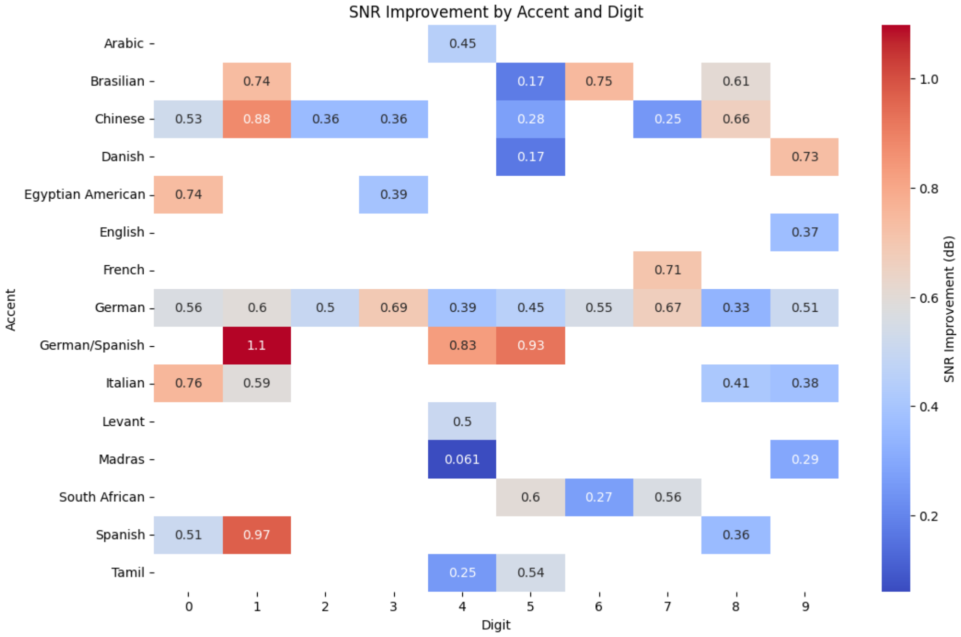


Figure 16: SNR Improvement by Accent and Digit. The heatmap displays the average SNR improvement across various accents and digits in the dataset.

Figure 16 illustrates the SNR improvement across different accents and digits. The heatmap reveals distinct patterns in the SNR improvement for different accents, with some accents consistently showing higher SNR improvements for specific digits. For example, accents like ”German/Spanish” exhibit significantly higher SNR values for certain digits, indicating that speech from these accent categories might be clearer or less affected by noise in comparison to others. This figure highlights how accent plays a role in the effectiveness of denoising techniques, and the SNR improvement varies based on both the accent of the speaker and the digit being spoken.

4.5.7 SNR Improvement by Origin

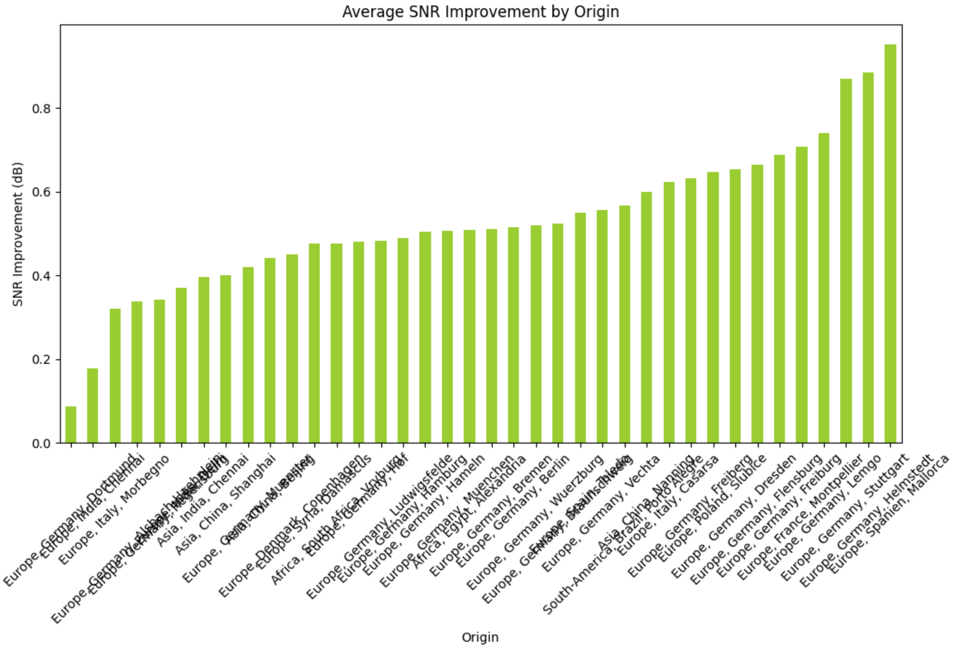


Figure 17: Average SNR Improvement by Origin. The bar chart represents the average SNR improvement across different origin categories.

Figure 17 displays the average SNR improvement across various origin categories. The plot suggests that the origin of the speaker has a notable effect on the SNR improvement, with certain origins demonstrating higher average SNR values. These trends could indicate that speakers from certain regions experience more effective denoising compared to others. This analysis helps in understanding the role of speaker origin in the performance of noise reduction models, highlighting any region-based disparities in denoising efficacy.

Conclusion

This report presented an in-depth analysis of the factors influencing the Signal-to-Noise Ratio (SNR) improvement in audio denoising using various speaker and contextual attributes. Throughout the analysis, we explored the impact of attributes such as accent, age, gender, native speaker status, origin, and digit classification on SNR.

Key findings reveal that certain attributes, such as accent and origin, significantly impact the SNR improvement, with some accents and origins exhibiting notably higher or lower denoising performance. Additionally, gender and age also played a role in the variability of SNR improvements, although their effects were less pronounced compared to accent and origin. Notably, native speakers showed slightly lower SNR improvement than non-native speakers, suggesting potential areas for model optimization in terms of language proficiency and accent diversity.

In conclusion, the analysis emphasizes the importance of considering speaker characteristics and contextual factors when designing and optimizing audio denoising systems. This work lays the foundation for further investigations into how to tailor denoising models to specific speaker groups, potentially improving the overall performance and user experience.